17712 U.S. PTO
101606

PTO/SB/05 (07-06)
Approved for use through 01/31/2007. OMB 0651-0032
U.S. Patent and Trademark Office. U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

UTILITY PATENT APPLICATION TRANSMITTAL

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Attorney Docket No.	090611-0357945
First Inventor	LARRY BALDWIN
Title	See 1 in Addendum
Express Mail Label No.	

APPLICATION ELEMENTS

See MPEP chapter 600 concerning utility patent application contents.

1. [X] **Fee Transmittal Form** (e.g., PTO/SB/17)
   (Submit an original and a duplicate for fee processing)
2. [X] **Applicant claims small entity status.** See 37 CFR 1.27.
3. [X] **Specification** [*Total Pages* 43]
   Both the claims and abstract must start on a new page
   (For information on the preferred arrangement, see MPEP 608.01(a))
4. [X] **Drawing(s)** (*35 U.S.C. 113*) [*Total Sheets* 3]
5. **Oath or Declaration** [*Total Sheets* _____]
   a. [ ] Newly executed (original or copy)
   b. [ ] A copy from a prior application (37 CFR 1.63(d))
   (for continuation/divisional with Box 18 completed)
   i. [ ] DELETION OF INVENTOR(S)
   Signed statement attached deleting inventor(s) name in the prior application, see 37 CFR 1.63(d)(2) and 1.33(b).
6. [X] **Application Data Sheet.** See 37 CFR 1.76
7. [ ] **CD-ROM or CD-R** in duplicate, large table or Computer Program *(Appendix)*
   [ ] Landscape Table on CD
8. **Nucleotide and/or Amino Acid Sequence Submission**
   (if applicable, items a. – c. are required)
   a. [ ] Computer Readable Form (CRF)
   b. Specification Sequence Listing on:
      i. [ ] CD-ROM or CD-R (2 copies); or
      ii. [ ] Paper
   c. [ ] Statements verifying identity of above copies

ADDRESS TO: **Commissioner for Patents**
P.O. Box 1450
Alexandria VA 22313-1450

113268 U.S. PTO
11/580926
101606

ACCOMPANYING APPLICATION PARTS

9. [ ] **Assignment Papers** (cover sheet & document(s))
   Name of Assignee _____
10. [ ] **37 CFR 3.73(b) Statement** *(when there is an assignee)* [ ] **Power of Attorney**
11. [ ] **English Translation Document** *(if applicable)*
12. [ ] **Information Disclosure Statement** (PTO/SB/08 or PTO-1449)
    [ ] Copies of foreign patent documents, publications, & other information
13. [ ] **Preliminary Amendment**
14. [X] **Return Receipt Postcard** (MPEP 503)
    (Should be specifically itemized)
15. [ ] **Certified Copy of Priority Document(s)**
    (if foreign priority is claimed)
16. [ ] **Nonpublication Request** under 35 U.S.C. 122(b)(2)(B)(i). Applicant must attach form PTO/SB/35 or equivalent.
17. [ ] Other: _____

18. If a CONTINUING APPLICATION, *check appropriate box, and supply the requisite information below and in the first sentence of the specification following the title, or in an Application Data Sheet under 37 CFR 1.76:*

[ ] Continuation [ ] Divisional [ ] Continuation-in-part (CIP) of prior application No.: .....

Prior application information: *Examiner* _____ *Art Unit:* _____

19. CORRESPONDENCE ADDRESS

[X] The address associated with Customer Number: 00909 ***OR*** [ ] Correspondence address below

Name					
Address					
City		State		Zip Code	
Country		Telephone		Email	

Signature		Date	October 16, 2006
Name (Print/Type)	Syed Jafar Ali	Registration No. (Attorney/Agent)	58780

This collection of information is required by 37 CFR 1.53(b). The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: **Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Addendum

1. SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

PTO/SB/17 (01-06)
Approved for use through 07/31/2006. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995 no persons are required to respond to a collection of information unless it displays a valid OMB control number

Fees pursuant to the Consolidated Appropriations Act, 2005 (H.R. 4818).

FEE TRANSMITTAL
For FY 2006

☒ Applicant claims small entity status. See 37 CFR 1.27

TOTAL AMOUNT OF PAYMENT ($) 1,050.00

Complete if Known	
Application Number	TBD
Filing Date	October 16, 2006
First Named Inventor	LARRY BALDWIN
Examiner Name	
Art Unit	
Attorney Docket No.	090611-0357945

METHOD OF PAYMENT (check all that apply)

☐ Check ☐ Credit Card ☐ Money Order ☐ None ☐ Other (please identify): _____

☒ Deposit Account Deposit Account Number: 033975 Deposit Account Name: PILLSBURY WINTHROP SHAW PITTMAN LLP

For the above-identified deposit account, the Director is hereby authorized to: (check all that apply)

☒ Charge fee(s) indicated below ☐ Charge fee(s) indicated below, **except for the filing fee**

☒ Charge any additional fee(s) or underpayments of fee(s) under 37 CFR 1.16 and 1.17 ☒ Credit any overpayments

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

FEE CALCULATION (All the fees below are due upon filing or may be subject to a surcharge.)

1. BASIC FILING, SEARCH, AND EXAMINATION FEES

	FILING FEES		SEARCH FEES		EXAMINATION FEES		
		Small Entity		Small Entity		Small Entity	
Application Type	**Fee ($)**	**Fee ($)**	**Fee ($)**	**Fee ($)**	**Fee ($)**	**Fee ($)**	**Fees Paid ($)**
Utility	300	150	500	250	200	100	500.00
Design	200	100	100	50	130	65	
Plant	200	100	300	150	160	80	
Reissue	300	150	500	250	600	300	
Provisional	200	100	0	0	0	0	

2. EXCESS CLAIM FEES

Fee Description	**Fee ($)**	**Small Entity Fee ($)**
Each claim over 20 (including Reissues)	50	25
Each independent claim over 3 (including Reissues)	200	100
Multiple dependent claims	360	180

Total Claims		**Extra Claims**	**Fee ($)**		**Fee Paid ($)**
42	- 20 or HP =	22	x 25.00	=	550.00

HP = highest number of total claims paid for, if greater than 20.

Indep. Claims		**Extra Claims**	**Fee ($)**		**Fee Paid ($)**
3	- 3 or HP =	0	x	=	0.00

HP = highest number of independent claims paid for, if greater than 3.

Multiple Dependent Claims	
Fee ($)	**Fee Paid ($)**

3. APPLICATION SIZE FEE

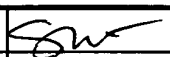
If the specification and drawings exceed 100 sheets of paper (excluding electronically filed sequence or computer listings under 37 CFR 1.52(e)), the application size fee due is $250 ($125 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).

Total Sheets		**Extra Sheets**		**Number of each additional 50 or fraction thereof**	**Fee ($)**		**Fee Paid ($)**
46	- 100 =	-54	/ 50 =	_____ (round **up** to a whole number)	x 125.00	=	0.00

4. OTHER FEE(S)

	Fees Paid ($)
Non-English Specification, $130 fee (no small entity discount)	
Other (e.g., late filing surcharge): _____	

SUBMITTED BY

Signature		Registration No. (Attorney/Agent)	58780	Telephone	703.770.7541
Name (Print/Type)	Syed Jafar Ali			Date	October 16, 2006

This collection of information is required by 37 CFR 1.136. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 30 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

FIELD OF THE INVENTION

(001) The invention relates to a cooperative conversational model for a human to machine voice user interface.

BACKGROUND OF THE INVENTION

(002) Advances in technology, particularly within the convergence space, have resulted in an increase in demand for voice recognition software that can exploit technology in ways that are intuitive to humans. While communication between human beings is most often "cooperative," in that information and/or context is shared to advance mutual conversational goals, existing Human-to-Machine interfaces fail to provide the same level of intuitive interaction. For example, each human participant in a conversation can contribute to an exchange for the benefit of the exchange. This is done through shared assumptions and expectations regarding various aspects of the conversation, such as the topic, participant knowledge about the topic, expectations of the other participant's knowledge about the topic, appropriate word usage for the topic and/or participants, conversational development based on previous utterances, the participants' tone or inflection, the quality and quantity of contribution expected from each participant, and many other factors. Participating in conversations that continually build and draw upon shared information is a natural and intuitive way for humans to converse.

(003) In contrast, complex Human-to-Machine interfaces do not allow users to exploit technology in an intuitive way, which inhibits mass-market adoption for various technologies. Incorporating a speech interface helps to alleviate this burden by making interaction easier and faster, but existing speech interfaces (when they actually work) still require significant learning on the part of the user. That is, existing speech interfaces are unable to bridge the gap between archaic Human-to-Machine interfaces and conversational speech that would

make interaction with systems feel normal. Users should be able to directly request what they want from a system in a normal, conversational fashion, without having to memorize exact words or phrases. Alternatively, when users are uncertain of particular needs, they should be able to engage the system in a productive, cooperative dialogue to resolve their requests. Instead, existing speech interfaces force users to dumb down their requests to match simple sets of instructions in simple languages in order to communicate requests in ways that systems can understand. Using existing speech interfaces, there is virtually no option for dialogue between the user and the system to satisfy mutual goals.

(004) Therefore, existing systems lack a conversational speech model that can provide users with the ability to interact with systems in ways that are inherently intuitive to human beings. Existing systems suffer from these and other problems.

SUMMARY OF THE INVENTION

(005) According to various embodiments and aspects of the invention, a cooperative conversational voice user interface may understand free form human utterances, freeing users from being restricted to a fixed set of commands and/or requests. Rather, users can engage in cooperative conversations with a machine to complete a request or series of requests using a natural, intuitive, free form manner of expression.

(006) According to an aspect of the invention, an exemplary system architecture for implementing a cooperative conversational voice user interface is provided. The system may receive an input, which may include a human utterance received by an input device, where the utterance may include one or more requests. As used herein, an "utterance" may be words, syllables, phonemes, or any other audible sound made by a human being. As used herein, a "request" may be a command, directive, or other instruction for a device, computer, or other machine to retrieve information, perform a task, or take some other action. In one implementation, the input may be a multi-

modal input, where at least part of the multi-modal input is an utterance. The utterance component of the input may be processed by an Automatic Speech Recognizer to generate one or more preliminary interpretations of the utterance. The one or more preliminary interpretations may then be provided to a conversational speech engine for further processing, where the conversational speech engine may communicate with one or more databases to generate an adaptive conversational response, which may be returned to the user as an output. In one implementation, the output may be a multi-modal output. For example, the utterance may include a request to perform an action, and the output may include a conversational response reporting success or failure, as well as an execution of the action.

(007) According to another aspect of the invention, an exemplary conversational speech engine may generate an adaptive conversational response to a request or series of requests. The conversational speech engine may include a free form voice search module that may understand an utterance made using typical, day-to-day language (i.e., in free form), and may account for variations in how humans normally speak, the vocabulary they use, and the conditions in which they speak. To account for intangible variables of human speech, the free form search module may include models of casual human speech. For example, in one implementation, the free form search module may understand specialized jargon and/or slang, tolerate variations in word order, and tolerate verbalized pauses or stuttered speech. For example, formalized English requests, where a verb precedes a noun, may be treated in an equivalent manner to requests where the noun precedes the verb. In another implementation, compound requests and/or compound tasks with multiple variables may be identified in a single utterance. By identifying all relevant information for completing one or more tasks from a single utterance, advantages may be provided over existing voice user interfaces, such as Command and Control systems that use verbal menus to restrict information

that a person can provide at a given point. In another implementation, inferring intended requests from incomplete or ambiguous requests may provide a conversational feel. By modeling what contextual signifiers, qualifiers, or other information may be required to perform a task in an identified context, an adaptive response may be generated, such as prompting a user for missing contextual signifiers, qualifiers, or other information. In one implementation, the response may ask for missing information in a way that most restricts possible interpretations, and the response may be framed to establish a domain for a subsequent user utterance. In another implementation, common alternatives for nouns and verbs may be recognized to reflect variations in usage patterns according to various criteria. Thus, variations in expression may be supported because word order is unimportant or unanticipated, and nouns and/or verbs may be represented in different ways to give simplistic, yet representative, examples. In another implementation, requests may be inferred from contradictory or otherwise inaccurate information, such as when an utterance includes starts and stops, restarts, stutters, run-on sentences, or other imperfect speech. For example, a user may sometimes change their mind, and thus alter the request in mid-utterance, and the imperfect speech feature may nonetheless be able to infer a request based on models of human speech. For example, various models may indicate that a last criterion is most likely to be correct, or intonation, emphasis, stress, use of the word "not," or other models may indicate which criterion is most likely to be correct.

(008) According to another aspect of the invention, the conversational speech engine may include a noise tolerance module that may discard words or noise which has no meaning in a given context to reduce a likelihood of confusion. Moreover, the noise tolerance module may filter out environmental and non-human noise to further reduce a likelihood of confusion. In one implementation, the noise tolerance module may cooperate with other modules and features to filter out words that do not fit into an identified context. For example, the noise

tolerance module may filter other human conversations and/or utterances within a range of one or more microphones. For example, a single device may include multiple microphones, or multiple devices may each include one or more microphones, and the noise tolerance module may collate inputs and cooperatively filter out sound by comparing a speech signal from the various microphones. The noise tolerance module may also filter out non-human environmental noise within range of the microphones, out-of-vocabulary words caused by speaker ambiguity or malapropisms, or other noise that may be unrelated to a target request. Performance benchmarks for the noise tolerance module may be defined by noise models based on human criteria. For example, if a driver of a car is 92% likely to be understood by a passenger when traveling at 65 miles-per-hour with windows cracked, then performance benchmarks for the noise tolerance module may have a similar performance under such conditions.

(009) According to another aspect of the invention, the conversational speech engine may include a context determination process that determines one or more contexts for a request to establish meaning within a conversation. The one or more contexts may be determined by having one or more context domain agents compete to determine a most appropriate domain for a given utterance. Once a given domain "wins" the competition, the winning domain may be responsible for establishing or inferring further contexts and updating short-term and long-term shared knowledge. If there is a deadlock between context domain agents, an adaptive conversational response may prompt the user to assist in disambiguating between the deadlocked agents. Moreover, the context determination process may infer intended operations and/or context based on previous utterances and/or requests, whereas existing systems consider each utterance independently, potentially making the same errors over and over again. For example, if a given interpretation turns out to be incorrect, the incorrect interpretation may be removed as a potential

interpretation from one or more Automatic Speech Recognizer grammars and/or from possible interpretations determined by the conversational speech engine, thereby assuring that a mistake will not be repeated for an identical utterance.

(010) The context determination process may provide advantages over existing voice user interfaces by continually updating one or more models of an existing context and establishing context as a by-product of a conversation, which cannot be established a priori. Rather, the context determination process may track conversation topics and attempt to fit a current utterance into recent contexts, including switching between contexts as tasks are completed, partially completed, requested, etc. The context determination process may identify one or more context domains for an utterance by defining a collection of related functions that may be useful for users in various context domains. Moreover, each context domain may have relevant vocabularies and thought collections to model word groupings, which when evaluated together, may disambiguate one context domain from another. Thus, eliminating out-of-context words and noise words when searching for relevant combinations may enhance accuracy of inferences. This provides advantages over existing systems that attempt to assign meaning to every component of an utterance (i.e., including out-of-context words and noise words), which results in nearly infinite possible combinations and greater likelihood of confusion. The context determination process may also be self-aware, assigning degrees of certainty to one or more generated hypotheses, where a hypothesis may be developed to account for variations in environmental conditions, speaker ambiguity, accents, or other factors. By identifying a context, capabilities within the context, vocabularies within the context, what tasks are done most often historically in the context, what task was just completed, etc., the context determination process may establish intent from rather meager phonetic clues. Moreover, just as in human-to-human conversation, users may switch contexts at any time

without confusion, enabling various context domains to be rapidly selected, without menu-driven dead ends, when an utterance is unambiguous.

(011) According to another aspect of the invention, an exemplary cooperative conversational model may build upon free form voice search, noise tolerance, and context determination to implement a conversational Human-to-Machine interface that reflects human interaction and normal conversational behavior. That is, the cooperative conversational model enables humans and machines to participant in a conversation with an accepted purpose or direction, with each participant contributing to the conversation for the benefit of the conversation. By taking advantage of human presumptions about utterances that humans rely upon, both as speakers and listeners, a Human-to-Machine interface may be analogous to everyday human-to-human conversation. In one implementation, the exemplary cooperative conversation model may take incoming data (shared knowledge) to inform a decision (intelligent hypothesis building), and then may refine the decision and generate a response (adaptive response building).

(012) According to another aspect of the invention, shared knowledge may include both short-term and long-term knowledge. Short-term knowledge may accumulate during a single conversation, where input received during a single conversation may be retained. The shared knowledge may include cross-modality awareness, where in addition to accumulating input relating to user utterances, requests, locations, etc., the shared knowledge may accumulate a current user interface state relating to other modal inputs to further build shared knowledge models. The shared knowledge may be used to build one or more intelligent hypotheses using current and relevant information, build long-term shared knowledge by identifying information with long-term significance, and generate adaptive responses with relevant state and word usage information. Moreover, because cooperative conversations model human conversations, short-term session data may be expired after a psychologically appropriate

amount of time, thereby humanizing system behavior, reducing a likelihood of contextual confusion based on stale data, while also adding relevant information from an expired session context to long-term knowledge models. Long-term shared knowledge may generally be user-centric, rather than session-based, where inputs may be accumulated over time to build user, environmental, cognitive, historical, or other long-term knowledge models. Long-term and short-term shared knowledge may be used simultaneously anytime a user engages in a cooperative conversation. Long-term shared knowledge may include explicit and/or implicit user preferences, a history of recent contexts, requests, tasks, etc., user-specific jargon related to vocabularies and/or capabilities of a context, most often used word choices, or other information. The long-term shared knowledge may be used to build one or more intelligent hypotheses using current and relevant information, generate adaptive responses with appropriate word choices when unavailable via short-term shared knowledge, refine long-term shared knowledge models, identify a frequency of specific tasks, identify tasks a user frequently has difficulty with, or provide other information and/or analysis to generate more accurate conversational responses. Shared knowledge may also be used to adapt a level of unprompted support (e.g., for novices versus experienced users, users who are frequently misrecognized, etc.) Thus, shared knowledge may enable a user and a voice user interface to share assumptions and expectations such as topic knowledge, conversation history, word usage, jargon, tone, or other assumptions and/or expectations that facilitate a cooperative conversation between human users and a system.

(013) According to another aspect of the invention, a conversation type may be identified for any given utterance. Categorizing and developing conceptual models for various types of exchanges may consistently align user expectations and domain capabilities. One or more intelligent hypotheses may be generated as to a conversation type by considering conversational goals,

participant roles, and/or an allocation of information among the participants. Based on the conversational goals, participant roles, and allocation of information, the intelligent hypotheses may consider various factors to classify a conversation (or utterance) into general types of conversations that can interact with one another to form many more variations and permutations of conversation types (e.g., a conversation type may change dynamically as information is reallocated from one participant to another, or as conversational goals change based on the reallocation of information).

(014) According to another aspect of the invention, the intelligent hypotheses may include one or more hypotheses of a user's intent in an utterance. In addition, the intelligent hypotheses may use short-term and/or long-term shared knowledge to proactively build and evaluate interaction with a user as a conversation progresses or over time. The hypotheses may model human-to-human interaction to include a varying degree of certainty for each hypothesis. That is, just as humans rely on knowledge shared by participants to examine how much and what kind of information was available, the intelligent hypotheses may leverage the identified conversation type and shared knowledge to generate a degree of certainty for each hypothesis.

(015) According to another aspect of the invention, syntactically, grammatically, and contextually sensitive "intelligent responses" may be generated from the intelligent hypotheses that can be used to generate a conversational experience for a user, while also guiding the user to reply in a manner favorable for recognition. The intelligent responses may create a conversational feel by adapting to a user's manner of speaking, framing responses appropriately, and having natural variation and/or personality (e.g., by varying tone, pace, timing, inflection, word use, jargon, and other variables in a verbal or audible response).

(016) According to another aspect of the invention, the intelligent responses may adapt to a user's manner of speaking by using contextual signifiers and

grammatical rules to generate one or more sentences that may cooperate with the user. By taking advantage of shared knowledge about how a user utters a request, the responses may be modeled using similar techniques used to recognize requests. The intelligent responses may rate possible responses statistically and/or randomize responses, which creates an opportunity to build an exchange with natural variation and conversational feel. This provides advantages over existing voice user interfaces where input and output is incongruous, as the input is "conversational" and the output is "computerese."

(017) According to another aspect of the invention, the intelligent responses may frame responses to influence a user reply utterance for easy recognition. For example, the responses may be modeled to illicit utterances from the user that may be more likely to result in a completed request. Thus, the responses may conform to a cooperative nature of human dialog and a natural human tendency to "parrot" what was just heard as part of a next utterance. Moreover, knowledge of current context may enhance responses to generate more meaningful conversational responses. Framing the responses may also deal with misrecognitions according to human models. For example, humans frequently remember a number of recent utterances, especially when one or more previous utterances were misrecognized or unrecognized. Another participant in the conversation may limit correction to a part of the utterance that was misrecognized or unrecognized, or over subsequent utterances and/or other interactions, clues may be provided to indicate the initial interpretation was incorrect. Thus, by storing and analyzing multiple utterances, utterances from earlier in a conversation may be corrected as the conversation progresses.

(018) According to another aspect of the invention, the intelligent responses may include multi-modal, or cross-modal, responses to a user. In one implementation, responses may be aware of and control one or more devices and/or interfaces, and users may respond by using whichever input method, or combination of input methods, is most convenient.

(019) According to another aspect of the invention, the intelligent responses may correct a course of a conversation without interrupting conversational flow. That is, even though the intelligent responses may be reasonably "sure," the intelligent responses may nonetheless sometimes be incorrect. While existing voice user interfaces tend to fail on average conversational missteps, normal human interactions may expect missteps and deal with them appropriately. Thus, responses after misrecognitions may be modeled after clarifications, rather than errors, and words may chosen in subsequent responses to move conversation forward and establish an appropriate domain to be explored with the user.

(020) Other objects and advantages of the invention will be apparent to those skilled in the art based on the following drawings and detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

(021) Fig. 1 is an exemplary block diagram of a system architecture according to one aspect of the invention.

(022) Fig. 2 is an exemplary block diagram of a conversational speech engine according to one aspect of the invention.

(023) Fig. 3 is an exemplary block diagram of a cooperative conversational model according to one aspect of the invention.

DETAILED DESCRIPTION

(024) Referring to Fig. 1, an exemplary system architecture for implementing a cooperative conversational voice user interface is illustrated according to one aspect of the invention. The system may receive an input 105 from a user, where in one implementation, input 105 may be an utterance received by an input device (e.g., a microphone), where the utterance may include one or

more requests. Input 105 may also be a multi-modal input, where at least part of the multi-modal input is an utterance. For example, the input device may include a combination of a microphone and a touch-screen device, and input 105 may include an utterance that includes a request relating to a portion of a display on the touch-screen device that the user is touching. For instance, the touch-screen device may be a navigation device, and input 105 may include an utterance of "Give me directions to here," where the user may be requesting directions to a desired destination on the display of the navigation device.

(025) The utterance component of input 105 may be processed by an Automatic Speech Recognizer 110 to generate one or more preliminary interpretations of the utterance. Automatic Speech Recognizer 110 may process the utterance using any suitable technique known in the art. For example, in one implementation, Automatic Speech Recognizer 110 may interpret the utterance using techniques of phonetic dictation to recognize a phoneme stream, as described in copending U.S. Patent Application Ser. No. 11/513,269, entitled "Dynamic Speech Sharpening," which is hereby incorporated by reference in its entirety. The one or more preliminary interpretations generated by Automatic Speech Recognizer 110 may then be provided to a conversational speech engine 115 for further processing. Conversational speech engine 115 may include a conversational language processor 120 and/or a voice search engine 125, described in greater detail in Fig. 2 below. Conversational speech engine 115 may communicate with one or more databases 130 to generate an adaptive conversational response, which may be returned to the user as an output 140. In one implementation, output 140 may be a multi-modal output and/or an interaction with one or more applications 145 to complete the request. For example, output 140 may include a combination of an audible response and a display of a route on a navigation device. For example, the utterance may include a request to perform an action, and output 140 may include a conversational response reporting

success or failure, as well as an execution of the action. In addition, in various implementations, Automatic Speech Recognizer 110, conversational speech engine 115, and/or databases 130 may reside locally (e.g., on a user device), remotely (e.g., on a server), or a hybrid model of local and remote processing may be used (e.g., lightweight applications may be processed locally while computationally intensive applications may be processed remotely).

(026) Referring to Fig. 2, an exemplary block diagram is provided illustrating a conversational speech engine 215 according to one aspect of the invention. Conversational speech engine 215 may include a conversational language processor 220 that generates an adaptive conversational response to a request or series of requests using a free form voice search module 245, a noise tolerance module 250, and/or a context determination process 255. According to one aspect of the invention, modules 245-255 may communicate with a voice search engine 225 that includes one or more context domain agents 230 and/or one or more vocabularies 235 to aid in interpreting utterances and generating responses, as described in "Enhancing the VUE™ (Voce-User-Experience) Through Conversational Speech," by Tom Freeman and Larry Baldwin, which is herein incorporated by reference in its entirety. Conversational speech engine 215 may generate an adaptive conversational response to one or more requests, where the requests may depend on unspoken assumptions, incomplete information, context established by previous utterances, user profiles, historical profiles, environmental profiles, or other information. Moreover, conversational speech engine 215 may track which requests have been completed, which requests are being processed, and/or which requests cannot be processed due to incomplete or inaccurate information, and the response may be generated accordingly.

(027) According to one aspect of the invention, free form voice search module 245 may understand an utterance made using typical, day-to-day language (i.e., in free form), and may account for variations in how humans normally

speak, the vocabulary they use, and the conditions in which they speak. Because variables such as stress, distraction, and serendipity are always different and infinitely varied, free form search module 245 may be designed with a goal of understanding that no human will come to the same Human-to-Machine interface situation in the same way twice. Thus, free form search module 245 may implement one or more features that model casual human speech. In various implementations, free form search module 245 may include, among other things, a free form utterance feature, a one-step access feature, an inferencing intended operations feature, an alternative expression feature, and/or an imperfect speech feature.

(028) The free form utterance feature may understand specialized jargon and/or slang, tolerate variations in word order (e.g., whether a subject of a request comes before or after a verb may be irrelevant), and tolerate verbalized pauses (e.g., "um," "ah," "eh," and other utterances without meaning). For example, the free form utterance feature may treat formalized English verb-before-noun requests in an equivalent manner to free form requests where a noun may precede a verb. For example, user utterances of "Change it to the Squizz" and "You know, um, that Squizz channel, ah, switch it there" may be treated equivalently (where Squizz is a channel on XM Satellite Radio). In either case, the free form utterance feature is able to identify "Squizz" as a subject of the utterance and "Change it" or "switch it" as a verb or request for the utterance (e.g., by cooperating with context determination process 255, or other features, and identifying a relevant context domain agent 230 and/or vocabulary 235 to interpret the utterance).

(029) The one-step access feature may understand utterances that include compound requests with multiple variables. For example, a user utterance may be "What is the forecast for Boston this weekend?" The one-step access feature may identify "weather" as a context (e.g., by cooperating with context determination process 255, or other features, and identifying "forecast" as a

synonym of "weather"), and search for a city equal to "Boston" and a time equal to "weekend." By identifying all relevant information for completing a task from a single utterance, the one-step access feature may overcome drawbacks of existing voice user interfaces, such as Command and Control systems that use verbal menus to restrict information that a person can provide at a given point (e.g., a Command and Control system for a phone directory service may say: "State please," . . . "City please," . . . "What listing," etc.). Moreover, some utterances may include compound requests, and the one-step access feature may decompose the compound requests into sub-tasks. For example, a user utterance of "I need to be at a meeting tomorrow in San Francisco at 8:00 am" may be decomposed into a set of sub-tasks such as (1) checking availability and reserving a flight on an evening before the meeting, (2) checking availability and reserving a hotel, (3) checking availability and reserving a car, etc., where users may further designate preferences for various tasks (e.g., first check availability on an airline for which the user is a frequent flyer). Depending on a level of shared knowledge about a user's preferences and/or historical patterns, the one-step access feature may infer additional tasks from a request. For example, in the above example, the one-step access feature may also check a weather forecast, and if the weather is "nice" (as defined by the user preferences and/or as inferred from historical patterns), the one-step access feature may schedule a tee-time at a preferred golf course in San Francisco.

(030) The inferencing intended operations feature may identify an intended request from incomplete or ambiguous requests. For example, when a user utters "Route <indecipherable> Chicago <indecipherable> here," where the user intended to say "Route calculation to Chicago from here," the inferencing intended operations feature may model what is required to calculate a route (an origination point and a destination point). Because the utterance includes the origination point and the destination point, a request to calculate a route

from the user's present location to Chicago may be inferred. Similarly, when the inferencing intended operations feature does not have sufficient information to infer a complete request, an adaptive conversational response may be generated to prompt the user for missing information. For example, when an utterance includes a request for a stock quote but not a company name (e.g., "Get me the stock price for <indecipherable>"), the response may be "What company's stock quote do you want?" The user may then provide an utterance including the company name, and the request may be completed. In one implementation, the response may ask for missing information in a way that most restricts possible interpretations (e.g., in a request for a task that requires both a city and a state, the state may be asked for first because there are fewer states than cities). Moreover, the inferencing intended operations feature may model compound tasks and/or requests by maintaining context and identifying relevant and/or missing information at both a composite and sub-task level.

(031) The alternative expression feature may recognize common alternatives for nouns and verbs to reflect variations in usage patterns according to various criteria. For example, users may vary expression based on age, socio-economics, ethnicity, user whims, or other factors. Thus, the alternative expression feature may support variations in expression where word order is unimportant or unanticipated. Alternatives in expression based on various criteria or demographics may be loaded into context domain agents 230 and/or vocabularies 235, and the alternative expression feature may update context domain agents 230 and/or vocabularies 235 based on inferred or newly discovered variations. In one implementation, conversational speech engine 215 may include a subscription interface to update changes to context domain agents 230 and/or vocabularies 235 (e.g., a repository may aggregate various user utterances and deploy updates system wide). In operation, the alternative expression feature may allow nouns and/or verbs to be represented in different

ways to give simplistic, yet representative, examples. For example, a user interested in a weather forecast for Washington, D.C. may provide any of the following utterances, each of which are interpreted equivalently: "What's the weather like in DC," "Is it raining inside the Beltway," Gimme the forecast for the capital," etc. Similarly, utterances of "Go to my home," "Go home," "Show route to home," and "I would like to know my way home" may all be interpreted equivalently, where a user profile may include the user's home address and a navigation route to the home address may be calculated.

(032) The imperfect speech feature may be able to infer requests from contradictory or otherwise inaccurate information, such as when an utterance includes starts and stops, restarts, stutters, run-on sentences, or other imperfect speech. For example, a user may sometimes change their mind, and thus alter the request in mid-utterance, and the imperfect speech feature may nonetheless be able to infer a request based on models of human speech. For example, for an utterance of "Well, I wanna . . . Mexi . . . no, steak restaurant please, I'm hungry," existing voice user interfaces make no assumptions regarding models of human speech and would be unable to infer whether the user wanted a Mexican or steak restaurant. The imperfect speech feature overcomes these drawbacks by using various models of human understanding that may indicate that a last criterion is most likely to be correct, or intonation, emphasis, stress, use of the word "not," or other models may indicate which criterion is most likely to be correct. Thus, in the above example, the imperfect speech feature may infer that the user wants a steak restaurant.

(033) According to one aspect of the invention, noise tolerance module 250 may be closely related to the imperfect speech feature, and may operate to discard words or noise that has no meaning in a given context so as not to create confusion. Moreover, noise tolerance module 250 may filter out environmental and non-human noise to further reduce a likelihood of confusion. In one implementation, noise tolerance module 250 may cooperate with other

modules and features to filter out words that do not fit into a context. For example, one or more contexts may be identified, and words that have no meaning with respect to system capabilities, random human utterances without meaning and other noise may be filtered out. Thus, noise tolerance module 250 may model real-world conditions to identify meaningful requests. For example, noise tolerance module 250 may filter other human conversations and/or utterances within a range of one or more microphones, For example, a single device may include multiple microphones, or multiple devices may each include one or more microphones, and the noise tolerance module may collate inputs and cooperatively filter out sound by comparing a speech signal from the various microphones. Noise tolerance module 250 may also filter out non-human environmental noise within the range of the microphones, out-of-vocabulary words, which could be a result of speaker ambiguity or malapropisms, or other noise that may be unrelated to a target request. Noise models in noise tolerance module 250 may define performance benchmarks based on human criteria. For example, if a driver of a car, traveling at 65 miles-per-hour, with windows cracked is 92% likely to be understood by a passenger, then noise tolerance module 250 may have a similar performance under those conditions.

(034) According to one aspect of the invention, conversational speech engine 215 may include a context determination process 255 that determines one or more contexts for a request to establish meaning within a conversation. The one or more contexts may be determined by having one or more context domain agents compete to determine a most appropriate domain for a given utterance, as described in copending U.S. Patent Application Ser. No. 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," and copending U.S. Patent Application Ser. No. 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," both of which are hereby incorporated by reference in

their entirety. Once a given context domain agent "wins" the competition, the winning agent may be responsible for establishing or inferring further contexts and updating short-term and long-term shared knowledge. If there is a deadlock between context domain agents, an adaptive conversational response may prompt the user to assist in disambiguating between the deadlocked agents. For example, a user utterance of "What about traffic?" may have a distinct meaning in various contexts. That is, "traffic" may have a first meaning when the user is querying a system's media player (i.e., "traffic" would be a Rock and Roll band led by singer/songwriter Steve Winwood), a second meaning when the user is querying a search interface regarding Michael Douglas films (i.e., "traffic" would be a film directed by Steven Soderbergh), a third meaning when the user is querying a navigation device for directions to an airport (i.e., "traffic" would be related to conditions on roads along a route to the airport).

(035) Moreover, context determination process 255 may infer intended operations and/or context based on previous utterances and/or requests, whereas existing systems consider each utterance independently, potentially making the same errors over and over again. For example, if a given interpretation turns out to be incorrect, the incorrect interpretation may be removed as a potential interpretation from one or more Automatic Speech Recognizer grammars and/or from possible subsequent interpretations determined by context determination process 255, thereby assuring that a mistake will not be repeated for an identical utterance.

(036) Context determination process 255 may overcome drawbacks of existing systems by continually updating one or more models of an existing context, where establishing context may be a by-product of a conversation, which cannot be established a priori. Context determination process 255 may establish a first context domain, change to a second context domain, change back to the first context domain, and so on, as tasks are completed, partially

completed, requested, etc, and a context stack may track conversation topics and attempt to fit a current utterance into a most-recent context, next-most-recent topic, etc., traversing the context stack until a most likely intent can be established. For example, a user may utter "What's the traffic report," and context determination process 255 may establish Traffic as a context, and return an output including a traffic report, which does not happen to mention traffic on Interstate-5. The user may then utter "What about I-5?" and context determination process 255 may know that the current context is Traffic, a traffic report including information about Interstate-5 may be searched for, and the traffic report indicating that Interstate-5 is crowded may be returned as an output. The user may then utter "Is there a faster way?" and context determination module 255 may know that the current context is still Traffic, and may search for routes to a specified destination with light traffic and avoiding Interstate-5. Moreover, context determination process 255 may build context based on user profiles, environmental profiles, historical profiles, or other information to further refine the context. For example, the profiles may indicate that Interstate-5 is a typical route taken Monday through Friday.

(037) The profiles may be particularly meaningful when attempting to disambiguate between contexts where a word has different meanings in different contexts. For example, a user may utter "What's the weather in Seattle?" and context determination process 255 may establish Weather as a context, as well as establishing Seattle as an environmental context. The user may then utter "and Portland?" and context determination process 255 may return a weather report for Portland, Oregon based on the Weather and an environmental proximity between Portland, Oregon and Seattle, Washington. The user may then ask "What time does the game start?"" and a search for sports events with teams from Seattle and/or Portland may occur, with results presented conversationally according to methods described in greater detail below in Fig. 3. Correlatively, had user originally uttered "What's the weather in

Portsmouth, New Hampshire," in the second utterance, context determination process 255 may instead retrieve a weather report for Portland, Maine based on an environmental proximity to New Hampshire. Moreover, when environmental profiles, contextual shared knowledge, and/or other short-term and/or long-term shared knowledge does not provide enough information to disambiguate between possibilities, responses may prompt the user with a request for further information (e.g., "Did you mean Portland, Maine, or Portland, Oregon?").

(038) Context determination process 255 may cooperate with context domain agents 230, where each context domain agent 230 may define a collection of related functions that may be useful for users. Moreover, each context domain agent 230 may include a relevant vocabulary 235 and thought collections that model word groupings, which when evaluated together, may disambiguate one context domain from another (e.g., a Music context domain agent 230 may include a vocabulary 235 for songs, artists, albums, etc., whereas a Stock context domain agent 230 may include a vocabulary 235 for company names, ticker symbols, financial metrics, etc.). Thus, accuracy in identifying meaning may be enhanced by eliminating out-of-context words and noise words when searching for relevant combinations. In contrast, existing systems attempt to assign meaning to every component of an utterance (e.g., including out-of-context words and noise words), which results in nearly infinite possible combinations and greater likelihood of confusion. Moreover, context domain agents 230 may include metadata for each criteria to further assist in interpreting utterances, inferring intent, completing incomplete requests, etc. (e.g., a Space Needle vocabulary word may include metadata for Seattle, landmark, tourism, Sky City restaurant, etc.). Given a disambiguated criterion, context determination process 255 may thus be able to automatically determine other information needed to complete a request, discard importance of word order, and perform other enhancements for conversational speech.

(039) Context domain agents 230 may also be self-aware, assigning degrees of certainty to one or more generated hypotheses, where a hypothesis may be developed to account for variations in environmental conditions, speaker ambiguity, accents, or other factors. Conceptually, context domain agents 230 may be designed to model utterances as a hard-of-hearing person would at a noisy party. By identifying a context, capabilities within the context, vocabularies within the context, what tasks are done most often historically in the context, what task was just completed, etc., a context domain agent 230 may establish intent from rather meager phonetic clues. Moreover, the context stack may be one of a plurality of components for establishing context, and thus not a constraint upon the user. All context domains may be accessible, allowing the user to switch contexts at any time without confusion. Thus, just as in human-to-human conversation, context domains may be rapidly selected, without menu-driven dead ends, when an utterance is unambiguous. For example, a user may utter, "Please call Rich Kennewick on his cell phone," and a system response of "Do you wish me to call Rich Kennewick on his cell?" may be generated. The user may decide at that point to call Rich Kennewick later, and instead, listen to some music. Thus, the user may then utter, "No, play the Louis Armstrong version of Body and Soul from my iPod," and a system response of "Playing Body and Soul by Louis Armstrong" may be generated as Body and Soul is played through a media player. In this example, the later utterance has no contextual connection to the first utterance, yet because request criteria in the utterances are unambiguous, contexts can be switched easily without relying on the context stack.

(040) Referring to Fig. 3, an exemplary cooperative conversational model 300 is illustrated according to an aspect of the invention. Cooperative conversational model 300 may build upon free form voice search 245, noise tolerance 250, and context determination 255 to implement a conversational Human-to-Machine interface that reflects how humans interact with each other and their normal

behavior in conversation. Simply put, cooperative conversational model 300 enables humans and machines to participate in a conversation with an accepted purpose or direction, with each participant contributing to the conversation for the benefit of the conversation. That is, cooperative conversational model 300 incorporates technology and process-flow that takes advantage of human presumptions about utterances that humans rely upon, both as speakers and listeners, thereby creating a Human-to-Machine interface that is analogous to everyday human-to-human conversation. In one implementation, a cooperative conversation may take incoming data (shared knowledge) 305 to inform a decision (intelligent hypothesis building) 310, and then may refine the decision and generate a response (adaptive response building) 315.

(041) According to one aspect of the invention, shared knowledge 305 includes both short-term and long-term knowledge about incoming data. Short-term knowledge may accumulate during a single conversation, while long-term knowledge may accumulate over time to build user profiles, environmental profiles, historical profiles, cognitive profiles, etc.

(042) Input received during a single conversation may be retained in a Session Input Accumulator. The Session Input Accumulator may include cross-modality awareness, where in addition to accumulating input relating to user utterances, requests, locations, etc., the Session Input Accumulator may accumulate a current user interface state relating to other modal inputs to further build shared knowledge models and more accurate adaptive responses (e.g., when a user utters a request relating to a portion of a touch-screen device, as described above). For example, the Session Input Accumulator may accumulate inputs including recognition text for each utterance, a recorded speech file for each utterance, a list-item selection history, a graphical user interface manipulation history, or other input data. Thus, the Session Input Accumulator may populate Intelligent Hypothesis Builder 310 with current and relevant information, build

long-term shared knowledge by identifying information with long-term significance, provide Adaptive Response Builder 315 with relevant state and word usage information, retain recent contexts for use with Intelligent Hypothesis Builder 310, and/or retain utterances for reprocessing during multi-pass evaluations. Moreover, because cooperative conversations 300 model human conversations, short-term session data may be expired after a psychologically appropriate amount of time, thereby humanizing system behavior. For example, a human is unlikely to recall a context of a conversation from two years ago, but because the context would be identifiable by a machine, session context is expired after a predetermined amount of time to reduce a likelihood of contextual confusion based on stale data. However, relevant information from an expired session context may nonetheless be added to user, historical, environmental, cognitive, or other long-term knowledge models.

(043) Long-term shared knowledge may generally be user-centric, rather than session-based. That is, inputs may be accumulated over time to build user, environmental, cognitive, historical, or other long-term knowledge models. Long-term and short-term shared knowledge (collectively, shared knowledge 305) may be used simultaneously anytime a user engages in a cooperative conversation 300. Long-term shared knowledge may include explicit and/or implicit user preferences, a history of most recently used agents, contexts, requests, tasks, etc., user-specific jargon related to vocabularies and/or capabilities of an agent and/or context, most often used word choices, or other information. The long-term shared knowledge may be used to populate Intelligent Hypothesis Builder 310 with current and relevant information, provide Adaptive Response Builder 315 with appropriate word choices when the appropriate word choices are unavailable via the Session Input Accumulator, refine long-term shared knowledge models, identify a frequency of specific tasks, identify tasks a user frequently has difficulty with, or provide other

information and/or analysis to generate more accurate conversational responses.

(044) As described above, shared knowledge 305 may be used to populate Intelligent Hypothesis Builder 310, such that a user and a voice user interface may share assumptions and expectations such as topic knowledge, conversation history, word usage, jargon, tone (e.g., formal, humorous, terse, etc.), or other assumptions and/or expectations that facilitate interaction at a Human-to-Machine interface.

(045) According to an aspect of the invention, one component of a successful cooperative conversation may be identifying a type of conversation from an utterance. By categorizing and developing conceptual models for various types of exchanges, user expectations and domain capabilities may be consistently aligned. Intelligent Hypothesis Builder 310 may generate a hypothesis as to a conversation type by considering conversational goals, participant roles, and/or an allocation of information among the participants. Conversational goals may broadly include: (1) getting a discrete piece of information or performing a discrete task, (2) gathering related pieces of information to make a decision, and/or (3) disseminating or gathering large amounts of information to build expertise. Participant roles may broadly include: (1) a leader that controls a conversation, (2) a supporter that follows the leader and provides input as requested, and/or (3) a consumer that uses information. Information may be held by one or more of the participants at the outset of a conversation, where a participant may hold most (or all) of the information, little (or none) of the information, or the information may be allocated roughly equally amongst the participants. Based on the conversational goals, participant roles, and allocation of information, Intelligent Hypothesis Builder 310 may consider various factors to classify a conversation (or utterance) into general types of conversations that can interact with one another to form many more variations and permutations of conversation types (e.g., a conversation

type may change dynamically as information is reallocated from one participant to another, or as conversational goals change based on the reallocation of information).

(046) For example, in one implementation, a query conversation may include a conversational goal of getting a discrete piece of information or performing a particular task, where a leader of the query conversation may have a specific goal in mind and may lead the conversation toward achieving the goal. The other participant may hold the information and may support the leader by providing the information. In a didactic conversation, a leader of the conversation may control information desired by a supporter of the conversation. The supporter's role may be limited to regulating an overall progression of the conversation and interjecting queries for clarification. In an exploratory conversation, both participants share leader and supporter roles, and the conversation may have no specific goal, or the goal may be improvised as the conversation progresses. Based on this model, Intelligent Hypothesis Builder 310 may broadly categorize a conversation (or utterance) according to the following diagram:

QUERY		
	Participant A	*Participant B*
	User	Voice User Interface
GOAL	Get information/action	Provide information/action
ROLE	Leader/Consumer	Supporter/Dispenser
INFORMATION ALLOCATION	Less	More

DIDACTIC		
	Participant A	*Participant B*
	User	Voice User Interface
GOAL	Get information	Provide information
ROLE	Follower/Consumer	Leader/Dispenser
INFORMATION ALLOCATION	Less	More

EXPLORATORY		
	Participant A	*Participant B*
	User	Voice User Interface
GOAL	Gather/share information	Gather/share information
ROLE	Follower/Consumer and Leader/Dispenser	Follower/Consumer and Leader/Dispenser
INFORMATION ALLOCATION	Equal or alternating	Equal or alternating

(047) Intelligent Hypothesis Builder 310 may use an identified conversation type to assist in generating a set of hypotheses as to a user's intent in an utterance. In addition, Intelligent Hypothesis Builder 310 may use short-term shared knowledge from the Session Input Accumulator to proactively build and evaluate interaction with a user as a conversation progresses, as well as long-term shared knowledge to proactively build and evaluate interaction with the user over time. Intelligent Hypothesis Builder 310 may thus adaptively arrive at a set of n-best hypotheses about user intent, and the n-best hypotheses may be provided to an Adaptive Response Builder 315. In addition, Intelligent

Hypothesis Builder 310 may model human-to-human interaction by calculating a degree of certainty for each of the hypotheses. That is, just as humans rely on knowledge shared by participants to examine how much and what kind of information was available, Intelligent Hypothesis Builder 310 may leverage the identified conversation type and short-term and long-term shared knowledge to generate a degree of certainty for each hypothesis.

(048) According to another aspect of the invention, Intelligent Hypothesis Builder 310 may generate one or more explicit hypotheses of a user's intent when an utterance contains all information (including qualifiers) needed to complete a request or task. Each hypothesis may have a corresponding degree of certainty, which may be used to determine a level of unprompted support to provide in a response. For example, a response may include a confirmation to ensure the utterance was not misunderstood. or the response may adaptively prompt a user to provide missing information.

(049) According to another aspect of the invention, Intelligent Hypothesis Builder 310 may use short-term knowledge to generate one or more implicit hypotheses of a user's intent when an utterance may be missing required qualifiers or other information needed to complete a request or task. Each hypothesis may have a corresponding degree of certainty. For instance, when a conversation begins, short-term knowledge stored in the Session Input Accumulator may be empty, and as the conversation progresses, the Session Input Accumulator may build a history of the conversation. Intelligent Hypothesis Builder 310 may use data in the Session Input Accumulator to supplement or infer additional information about a current utterance. For example, Intelligent Hypothesis Builder 310 may evaluate a degree of certainty based on a number of previous requests relevant to the current utterance. In another example, when the current utterance contains insufficient information to complete a request or task, data in the Session Input Accumulator may be used to infer missing information so that a hypothesis can be generated. In still

another example, Intelligent Hypothesis Builder 310 may identify syntax and/or grammar to be used by Adaptive Response Builder 315 to formulate personalized and conversational response. In yet another example, when the current utterance contains a threshold amount of information needed to complete a request or task, data in the Session Input Accumulator may be relied upon to tune a degree of certainty.

(050) According to another aspect of the invention, Intelligent Hypothesis Builder 310 may use long-term shared knowledge to generate one or more implicit hypotheses of a user's intent when an utterance is missing qualifiers or other information needed to complete a request or task. Each hypothesis may have a corresponding degree of certainty. Using long-term knowledge may be substantially similar to using short-term shared knowledge, except that information may be unconstrained by a current session, and an input mechanism may include information from additional sources other than conversational sessions. For example, Intelligent Hypothesis Builder 310 may use information from long-term shared knowledge at any time, even when a new conversation is initiated, whereas short-term shared knowledge may be limited to an existing conversation (where no short-term shared knowledge would be available when a new conversation is initiated). Long-term shared knowledge may come from several sources, including user preferences or a plug-in data source (e.g., a subscription interface to a remote database), expertise of a user (e.g., based on a frequency of errors, types of tasks requested, etc., the user may be identified as a novice, intermediate, experienced, or other type of user), agent-specific information and/or language that may also apply to other agents (e.g., by decoupling information from an agent to incorporate the information into other agents), frequently used topics passed in from the Session Input Accumulator, frequently used verbs, nouns, or other parts of speech, and/or other syntax information passed in from the Session Input Accumulator, or other sources of long-term shared knowledge may be used.

(051) According to another aspect of the invention, knowledge-enabled utterances, as generated by Intelligent Hypothesis Builder 310, may include one or more explicit (supplied by a user), and one or more implicit (supplied by Intelligent Hypothesis Builder 310) contextual signifiers, qualifiers, criteria, and other information that can be used to identify and evaluate relevant tasks. At that point, Intelligent Hypothesis Builder 310 may provide an input to Adaptive Response Builder 315. The input received by Adaptive Response Builder 315 may include at least a ranked list of hypotheses, including explicit and/or implicit hypotheses, each of which may have a corresponding degree of certainty. A hypothesis may be assigned one of four degrees of certainty: (1) "sure," where contextual signifiers and qualifiers relate to one task, context and qualifiers relate to one task, and an ASR confidence level exceeds a predetermined threshold; (2) "pretty sure," where contextual signifiers and qualifiers relate to more than one task (select top-ranked task) and criteria relates to one request, and/or the ASR confidence level is below the predetermined threshold; (3) "not sure," where additional contextual signifiers or qualifiers are needed to indicate or rank a task; and (4) "no hypothesis," where little or no information can be deciphered. Each degree of certainty may further be classified as explicit or implicit, which may be used to adjust a response. The input received by Adaptive Response Builder 310 may also include a context, user syntax and/or grammar, context domain agent specific information and/or preferences (e.g., a travel context domain agent may know a user frequently requests information about France, which may be shared with a movie context domain agent so that responses may occasionally include French movies).

(052) According to another aspect of the invention, Adaptive Response Builder 315 may build syntactically, grammatically, and contextually sensitive "intelligent responses" that can be used with one or more agents to generate a conversational experience for a user, while also guiding the user to reply in a manner favorable for recognition. In one implementation, the intelligent

responses may include a verbal or audible reply played through an output device (e.g., a speaker), and/or an action performed by a device, computer, or machine (e.g., downloading a web page, showing a list, executing an application, etc.). In one implementation, an appropriate response may not require conversational adaptation, and default replies and/or randomly selected response sets for a given task may be used.

(053) According to another aspect of the invention, Adaptive Response Builder 310 may draw on information maintained by Intelligence Hypothesis Builder 310 to generate responses that may be sensitive to context, task recognition of a current utterance, what a user already knows about a topic, what an application already knows about the topic, shared knowledge regarding user preferences and/or related topics, appropriate contextual word usage (e.g., jargon), words uttered by the user in recent utterances, conversational development and/or course correction, conversational tone, type of conversation, natural variation in wording of responses, or other information. As a result, Adaptive Response Builder 315 may generate intelligent responses that create conversational feel, adapt to information that accumulates over a duration of a conversation, maintain cross-modal awareness, and keep the conversation on course.

(054) According to another aspect of the invention, Adaptive Response Builder 315 may create a conversational feel by adapting to a user's manner of speaking, framing responses appropriately, and having natural variation and/or personality (e.g., by varying tone, pace, timing, inflection, word use, jargon, and other variables in a verbal or audible response). Adapting to a user's manner of speaking may include using contextual signifiers and grammatical rules to generate one or more sentences for use as response sets that may cooperate with the user. By taking advantage of short-term (from the Session Input Accumulator) and long-term (from one or more profiles) shared knowledge about how a user utters a request, the responses may be modeled using

techniques used to recognize requests. Adaptive Response Builder 315 may rate possible responses statistically and/or randomize responses, which creates an opportunity to build an exchange with natural variation and conversational feel. This may be a significant advantage over existing voice user interfaces with incongruous input and output, where the input is "conversational" and the output is "computerese." The following examples may demonstrate how a response may adapt to a user's input word choices and manner of speaking:

User	**Do you know** (*mumbled words*) Seattle (*more mumbled words*)?
Voice User Interface	Did you want Seattle sports scores, weather, traffic, or news?

User	**Find me** (*mumbled words*) Seattle (*more mumbled words*)?
Voice User Interface	I *found* Seattle, did you want sports scores, weather, traffic, or news?

User	**Get me** (*mumbled words*) Seattle (*more mumbled words*)?
Voice User Interface	I've *got* Seattle, did you want me to *get* sports scores, weather, traffic, or news?

(055) According to another aspect of the invention, Adaptive Response Builder 315 may frame responses to influence a user to reply with an utterance that may be easily recognized. For example, a user may utter, "Get me the news" and a voice user interface response may be "Which of these categories? Top news stories, international news, political news, or sports news?" The response may be likely to illicit utterances from the user, such as "Top news stories" or "International news," which are more likely to result in a completed request. Thus, the responses may conform to a cooperative nature of human dialog, and a natural human tendency to "parrot" what was just heard as part of a next

utterance. Moreover, knowledge of current context may enhance responses to generate more meaningful conversational responses, such as in the following exchange:

User	What's the weather like in Dallas?
Voice User Interface	In Dallas, it's sunny and 90 degrees.
User	What theaters are showing the movie "The Fantastic Four" there?
Voice User Interface	10 theaters in Dallas are showing "The Fantastic Four." Do you want show times for a particular theater?

(056) Framing the responses may also deal with misrecognitions according to human models. For example, humans frequently remember a number of recent utterances, especially when one or more previous utterances were misrecognized or unrecognized. Another participant in the conversation may limit correction to a part of the utterance that was misrecognized or unrecognized, or over subsequent utterances and/or other interactions, clues may be provided to indicate the initial interpretation was incorrect. Thus, by storing and analyzing multiple utterances, utterances from earlier in a conversation may be corrected as the conversation progresses.

(057) According to another aspect of the invention, Adaptive Response Builder 315 may generate multi-modal, or cross-modal, responses to a user. In one implementation, responses may be aware of and control one or more devices and/or interfaces, and users may respond by using whichever input method, or combination of input methods, is most convenient. For example, a response asking the user to direct an utterance with a "Yes" or "No" in a multi-modal environment may also display alternatives visually.

(058) According to another aspect of the invention, Adaptive Response Builder 315 may correct a course of a conversation without interrupting conversational

flow. Adaptive Response Builder 315 may generate intelligent responses based on the ranked list of hypotheses and corresponding degrees of certainty, which may be used to correct a course of a conversation without interrupting conversational flow. That is, even though the intelligent responses may be reasonably "sure," the intelligent responses may nonetheless sometimes be incorrect. While existing voice user interfaces tend to fail on average conversational missteps, normal human interactions may expect missteps and deal with them appropriately. Thus, responses after a misrecognition may be modeled after clarifications, rather than errors, and words may chosen in subsequent responses that move conversation forward and establish an appropriate domain to be explored with the user. For example, course correction may result in the following exchange:

User	Can you **get** (*mumbled words here*) Mariners (*more mumbled words*)?
Voice User Interface	*I've got* the score for the Mariners game? 4-2 Mariners.
User	No, the start time for tomorrow's game.
Voice User Interface	Oh, 7:05.

(059) The above disclosure has been described in terms of specific exemplary aspects, implementations, and embodiments of the invention. However, those skilled in the art will recognize various changes and modifications that may be made without departing from the scope and spirit of the invention. Therefore, the specification and drawings are to be regarded as exemplary only, and the scope of the invention is to be determined solely by the appended claims.

CLAIMS

What is claimed is:

1. A method for providing a voice user interface that generates cooperative conversational responses to user utterances, the responses being based on shared knowledge about a current conversation and/or one or more past conversations, the method comprising:

receiving an input, the input including at least one utterance;

updating the shared knowledge to include knowledge about the at least one utterance;

determining an intended meaning for the at least one utterance based on the updated shared knowledge; and

generating a response based on the determined intended meaning.

2. The method of claim 1, wherein updating the shared knowledge includes populating a short-term context stack with information about the at least one utterance.

3. The method of claim 2, further comprising expiring the information about the at least one utterance in the short-term context stack after a predetermined amount of time.

4. The method of claim 1, wherein updating the shared knowledge includes updating one or more long-term profiles based on information about the at least one utterance.

5. The method of claim 1, wherein determining the intended meaning includes identifying a conversation type for the at least one utterance, the conversation type based on one or more of a goal for the utterance, one or

more participant roles for the utterance, and/or an allocation of information for the utterance.

6. The method of claim 1, wherein determining the intended meaning includes generating at least one hypothesis of an intended meaning for the at least one utterance, the at least one hypothesis including a degree of certainty.

7. The method of claim 6, wherein generating the at least one hypothesis includes identifying contextual signifiers and/or qualifiers in the at least one utterance.

8. The method of claim 7, further comprising inferring missing contextual signifiers and/or qualifiers for the at least one utterance based on the shared knowledge.

9. The method of claim 8, wherein the degree of certainty for the at least one hypothesis is based on one or more of the identified contextual signifiers and/or qualifiers, the inferred contextual signifiers and/or qualifiers, and/or contextual signifiers and/or qualifiers necessary for generating a response.

10. The method of claim 1, wherein generating the response includes adapting contextual signifiers and/or grammar in the response to reflect contextual signifiers and/or grammar in the at least one utterance.

11. The method of claim 1, wherein generating the response includes framing the response to influence a subsequent utterance.

12. The method of claim 1, further comprising:
learning that the determined intended meaning is incorrect;

further updating the shared knowledge to include knowledge that the determined intended meaning is incorrect; and

determining a second intended meaning for the at least one utterance based on the further updated shared knowledge, wherein the incorrect determined intended meaning is excluded as a possible second intended meaning.

13. The method of claim 1, wherein the input is a multi-modal input.

14. The method of claim 1, wherein the response is a multi-modal response.

15. A computer readable medium containing computer-executable instructions for providing a voice user interface that generates cooperative conversational responses to user utterances, the responses being based on shared knowledge about a current conversation and/or one or more past conversations, the computer-executable instructions operable when executed to:

receive an input, the input including at least one utterance;

update the shared knowledge to include knowledge about the at least one utterance;

determine an intended meaning for the at least one utterance based on the updated shared knowledge; and

generate a response based on the determined intended meaning.

16. The computer readable medium of claim 15, wherein the computer-executable instructions are operable to update the shared knowledge by populating a short-term context stack with information about the at least one utterance.

17. The computer readable medium of claim 16, wherein the computer-executable instructions are operable to expire the information about the at least one utterance in the short-term context stack after a predetermined amount of time.

18. The computer readable medium of claim 15, wherein the computer-executable instructions are operable to update the shared knowledge by updating one or more long-term profiles based on information about the at least one utterance.

19. The computer readable medium of claim 15, wherein the computer-executable instructions are operable to determine the intended meaning by identifying a conversation type for the at least one utterance, the conversation type based on one or more of a goal for the utterance, one or more participant roles for the utterance, and/or an allocation of information for the utterance.

20. The computer readable medium of claim 15, wherein the computer-executable instructions are operable to determine the intended meaning by generating at least one hypothesis of an intended meaning for the at least one utterance, the at least one hypothesis including a degree of certainty.

21. The computer-readable medium of claim 20, wherein the computer-executable instructions are operable to generate the at least one hypothesis by identifying contextual signifiers and/or qualifiers in the at least one utterance.

22. The computer readable medium of claim 21, wherein the computer-executable instructions are operable to infer missing contextual signifiers and/or qualifiers for the at least one utterance based on the shared knowledge.

23. The computer readable medium of claim 22, wherein the degree of certainty for the at least one hypothesis is based on one or more of the identified contextual signifiers and/or qualifiers, the inferred contextual signifiers and/or qualifiers, and/or contextual signifiers and/or qualifiers necessary for generating a response.

24. The computer readable medium of claim 15, wherein the computer-executable instructions are operable to generate the response by adapting contextual signifiers and/or grammar in the response to reflect contextual signifiers and/or grammar in the at least one utterance.

25. The computer readable medium of claim 15, wherein the computer-executable instructions are operable to generate the response by framing the response to influence a subsequent utterance.

26. The computer readable medium of claim 15, wherein the computer-executable instructions are further operable when executed to:

learn that the determined intended meaning is incorrect;

further update the shared knowledge to include knowledge that the determined intended meaning is incorrect; and

determine a second intended meaning for the at least one utterance based on the further updated shared knowledge, wherein the incorrect determined intended meaning is excluded as a possible second intended meaning.

27. The computer readable medium of claim 15, wherein the input is a multi-modal input.

28. The computer readable medium of claim 15, wherein the response is a multi-modal response.

29. A system for providing a voice user interface that generates cooperative conversational responses to user utterances, the system including at least one memory for storing shared knowledge about a current conversation and/or one or more past conversations, the system comprising:

at least one input device that receives an input, the input including at least one utterance; and

one or more processors collectively operable to:

update the shared knowledge to include knowledge about the at least one utterance;

determine an intended meaning for the at least one utterance based on the updated shared knowledge; and

generate a response based on the determined intended meaning.

30. The system of claim 29, wherein the one or more processors are operable to update the shared knowledge by populating a short-term context stack with information about the at least one utterance.

31. The system of claim 30, wherein the one or more processors are operable to expire the information about the at least one utterance in the short-term context stack after a predetermined amount of time.

32. The system of claim 29, wherein the one or more processors are operable to update the shared knowledge by updating one or more long-term profiles based on information about the at least one utterance.

33. The system of claim 29, wherein the one or more processors are operable to determine the intended meaning by identifying a conversation type for the at least one utterance, the conversation type based on one or more of a goal for the utterance, one or more participant roles for the utterance, and/or an allocation of information for the utterance.

34. The system of claim 29, wherein the one or more processors are operable to determine the intended meaning by generating at least one hypothesis of an intended meaning for the at least one utterance, the at least one hypothesis including a degree of certainty.

35. The system of claim 34, wherein the one or more processors are operable to generate the at least one hypothesis by identifying contextual signifiers and/or qualifiers in the at least one utterance.

36. The system of claim 35, wherein the one or more processors are operable to infer missing contextual signifiers and/or qualifiers for the at least one utterance based on the shared knowledge.

37. The system of claim 36, wherein the degree of certainty for the at least one hypothesis is based on one or more of the identified contextual signifiers and/or qualifiers, the inferred contextual signifiers and/or qualifiers, and/or contextual signifiers and/or qualifiers necessary for generating a response.

38. The system of claim 29, wherein the one or more processors are operable to generate the response by adapting contextual signifiers and/or grammar in the response to reflect contextual signifiers and/or grammar in the at least one utterance.

39. The system of claim 29, wherein the one or more processors are operable to generate the response by framing the response to influence a subsequent utterance.

40. The system of claim 29, wherein the one or more processors are further operable to:

learn that the determined intended meaning is incorrect;

further update the shared knowledge to include knowledge that the determined intended meaning is incorrect; and

determine a second intended meaning for the at least one utterance based on the further updated shared knowledge, wherein the incorrect determined intended meaning is excluded as a possible second intended meaning.

41. The system of claim 29, wherein the input is a multi-modal input.

42. The system of claim 29, wherein the response is a multi-modal response.

ABSTRACT

A cooperative conversational voice user interface is provided. The cooperative conversational voice user interface may build upon short-term and long-term shared knowledge to generate one or more explicit and/or implicit hypotheses about an intent of a user utterance. The hypotheses may be ranked based on varying degrees of certainty, and an adaptive response may be generated for the user. Responses may be worded based on the degrees of certainty and to frame an appropriate domain for a subsequent utterance. In one implementation, misrecognitions may be tolerated, and conversational course may be corrected based on subsequent utterances and/or responses.

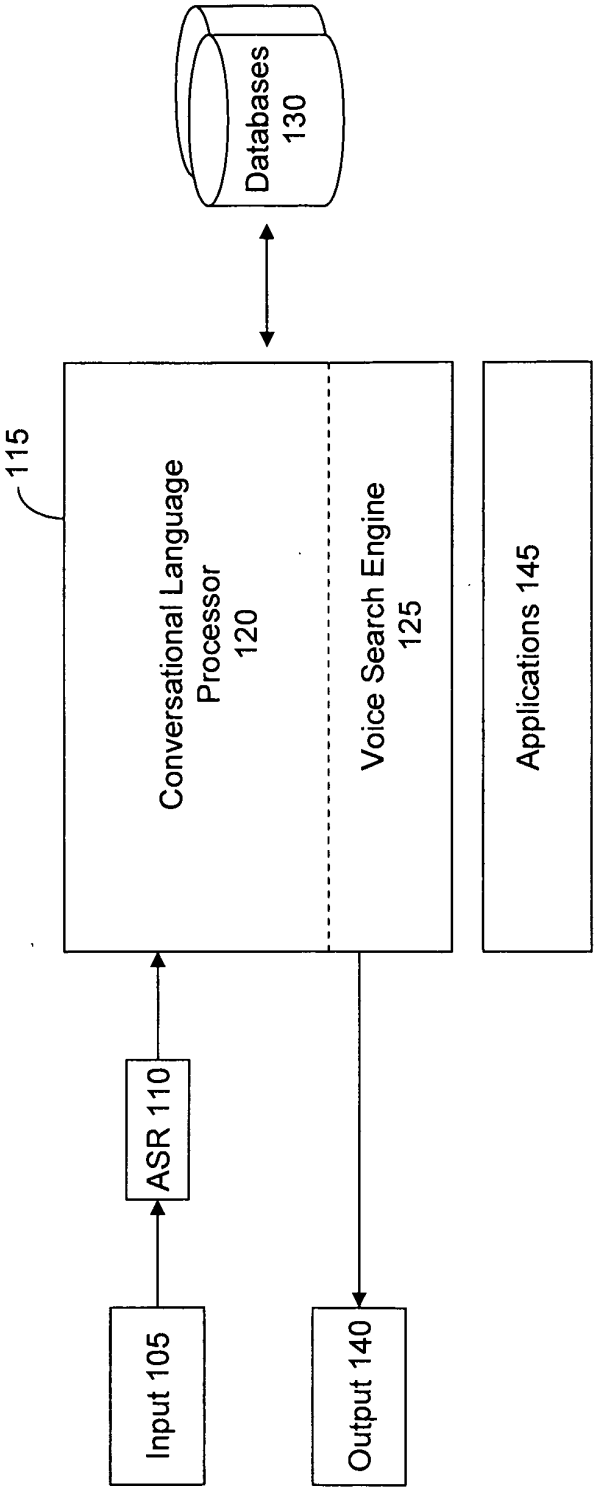

Databases
130
115
Conversational Language
Processor
120
Voice Search Engine
125
Applications 145
ASR 110
Input 105
Output 140

FIG. 1

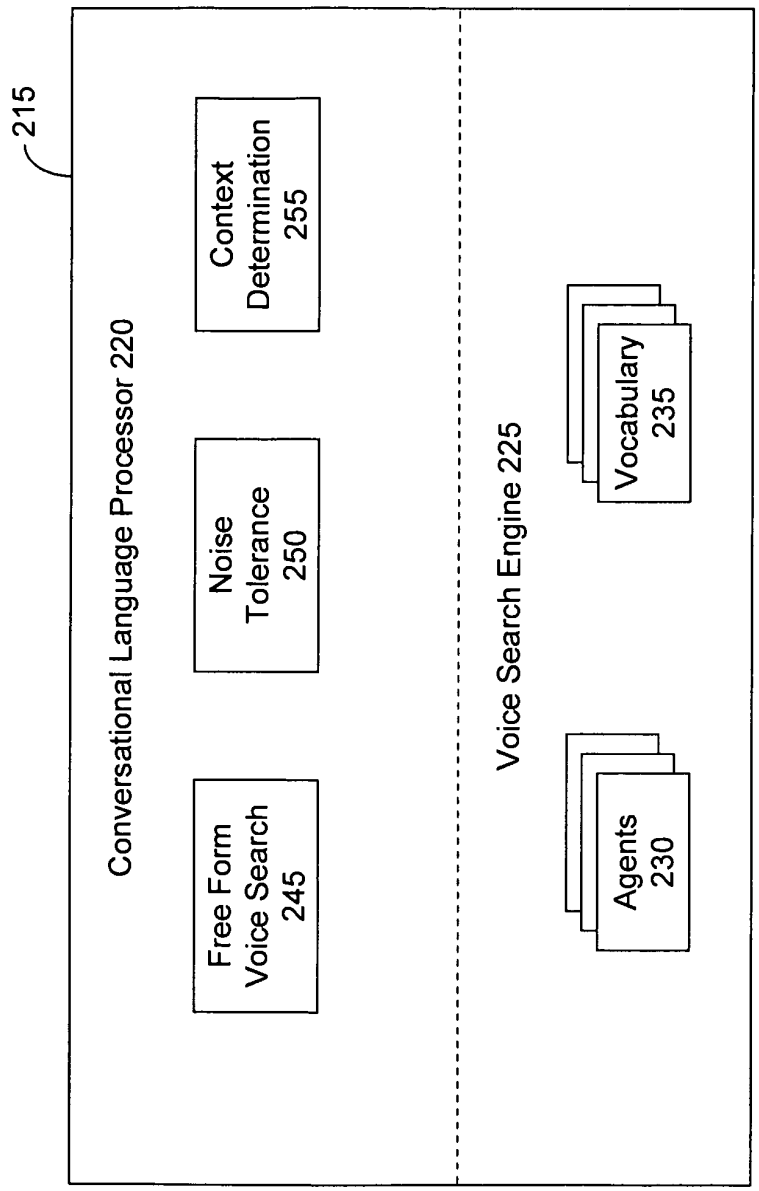

215
Conversational Language Processor 220
Free Form Voice Search 245
Noise Tolerance 250
Context Determination 255
Voice Search Engine 225
Agents 230
Vocabulary 235

FIG. 2

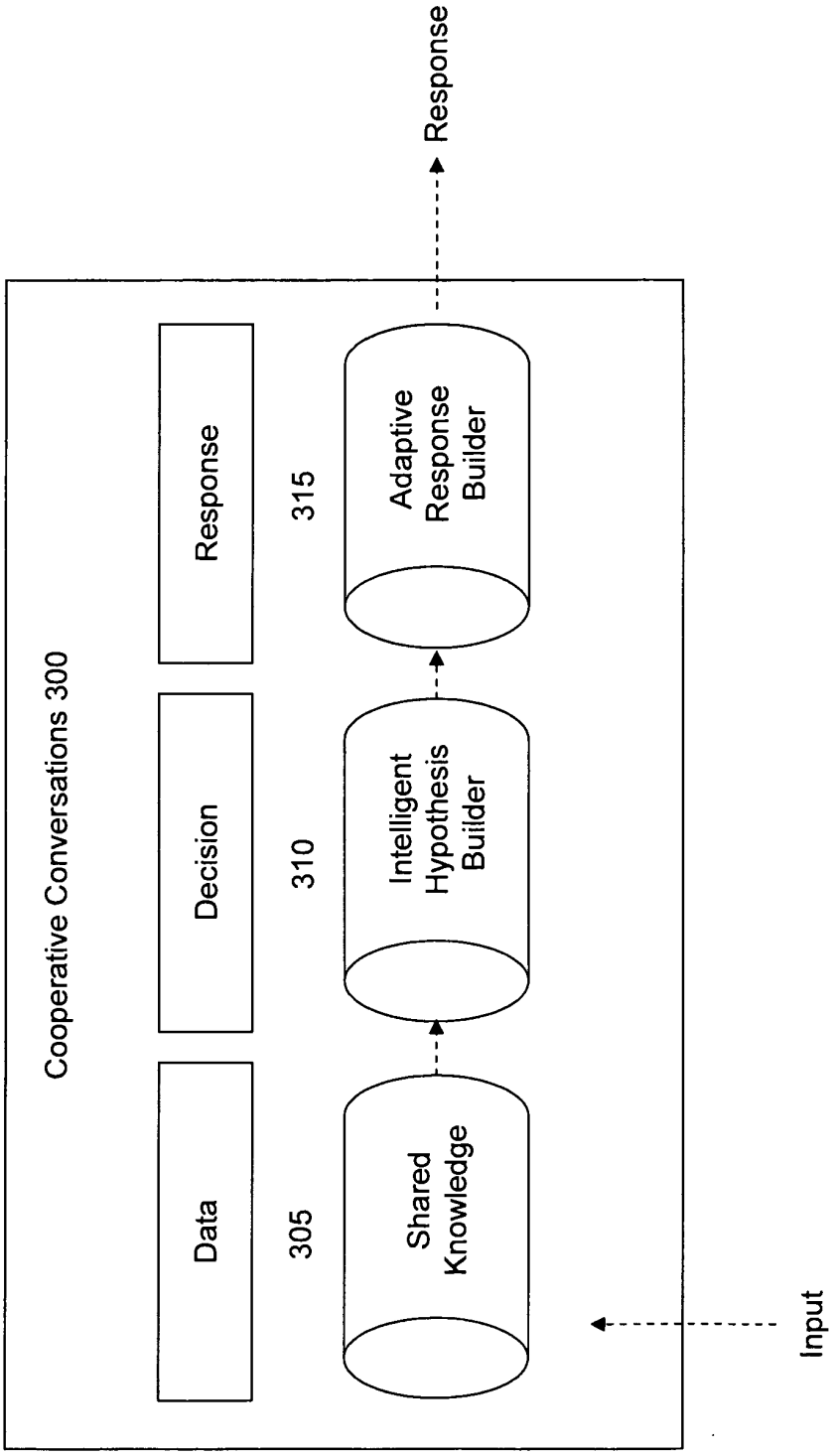

Cooperative Conversations 300
Data
Decision
Response
305
310
315
Shared Knowledge
Intelligent Hypothesis Builder
Adaptive Response Builder
Input
Response

FIG. 3

PATENT APPLICATION SERIAL NO. _____

U.S. DEPARTMENT OF COMMERCE
PATENT AND TRADEMARK OFFICE
FEE RECORD SHEET

10/18/2006 DEMMANU1 00000032 033975 11580926

01 FC:2011 150.00 DA
02 FC:2111 250.00 DA
03 FC:2311 100.00 DA
04 FC:2202 550.00 DA

PTO-1556
(5/87)

*U.S. Government Printing Office: 2002 — 469-267/69033

PTO/SB/06 (12-04)
Approved for use through 7/31/2006. OMB 0651-0032
U.S. Patent and Trademark Office: U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PATENT APPLICATION FEE DETERMINATION RECORD

Substitute for Form PTO-875 Effective December 8, 2004

Application or Docket Number: 11-580-926

APPLICATION AS FILED – PART I

FOR	(Column 1) NUMBER FILED	(Column 2) NUMBER EXTRA	SMALL ENTITY RATE ($)	SMALL ENTITY FEE ($)	OR	OTHER THAN SMALL ENTITY RATE ($)	OTHER THAN SMALL ENTITY FEE ($)
BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A	N/A	150.00		N/A	300.00
SEARCH FEE (37 CFR 1.16(k), (l), or (m))	N/A	N/A	N/A	$250		N/A	$500
EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A	N/A	$100		N/A	$200
TOTAL CLAIMS (37 CFR 1.16(i))	42 minus 20 =	* 22	X$ 25 =	550	OR	X$50 =	
INDEPENDENT CLAIMS (37 CFR 1.16(h))	3 minus 3 =	* 0	X100 =			X200 =	
APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $250 ($125 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).						
MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))			+180=			+360=	
* If the difference in column 1 is less than zero, enter "0" in column 2.			TOTAL	1050		TOTAL	

APPLICATION AS AMENDED – PART II

AMENDMENT A	(Column 1)	CLAIMS REMAINING AFTER AMENDMENT	(Column 2)	HIGHEST NUMBER PREVIOUSLY PAID FOR	(Column 3) PRESENT EXTRA	SMALL ENTITY RATE ($)	SMALL ENTITY ADDITIONAL FEE ($)	OR	OTHER THAN SMALL ENTITY RATE ($)	OTHER THAN SMALL ENTITY ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	*	Minus	**	=	X$ 25 =		OR	X$50 =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	X100 =		OR	X200 =	
	Application Size Fee (37 CFR 1.16(s))									
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))					+180=		OR	+360=	
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

AMENDMENT B	(Column 1)	CLAIMS REMAINING AFTER AMENDMENT	(Column 2)	HIGHEST NUMBER PREVIOUSLY PAID FOR	(Column 3) PRESENT EXTRA	RATE ($)	ADDITIONAL FEE ($)	OR	RATE ($)	ADDITIONAL FEE ($)
	Total (37 CFR 1.16(i))	*	Minus	**	=	X$ 25 =		OR	X$50 =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	X100 =		OR	X200 =	
	Application Size Fee (37 CFR 1.16(s))									
	FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))					+180=		OR	+360=	
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.
** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".
*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".
The "Highest Number Previously Paid For" (Total or Independent) is the highest number found in the appropriate box in column 1.

This collection of information is required by 37 CFR 1.16. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: **Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of: LARRY BALDWIN et al.

Application No.: TBD

Filed: October 16, 2006

For: SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

APPLICATION DATA SHEET
37 C.F.R. § 1.76

BIBLIOGRAPHIC DATA

1. Applicant information

First applicant: LARRY BALDWIN
Citizenship: United States of America
Residence: 25498 S.E. 274th Place, Maple Valley, WA 98038

Second applicant: MICHAEL TJALVE
Citizenship: Denmark
Residence: 14904 SE 47th Court, Bellevue, WA 98006

Third applicant: BLANE EBERSOLD
Citizenship: United States of America
Residence: 4721 Admiral Way S.W., Seattle, WA 98116

Fourth applicant: TOM FREEMAN
Citizenship: United States of America
Residence: 6735 83rd Avenue S.E., Mercer Island, WA 98040

2. Correspondence information

Correspondence for this application should be addressed as follows:

Customer No.: 00909

3. Application information

Title of Invention: SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

Docket number assigned to this application: 090611-0357945

Total number of drawing sheets: 3
Type of application: Utility
Application is to be published.

Secrecy order under § 5.2:

This application does not disclose subject matter of an application which is under a secrecy order pursuant to § 5.2.

4. Representative information

The following have a power of attorney or authorization of agent in this application:

Customer No.: 00909

5. Assignee information

The assignee of this application is:

VoiceBox Technologies, Inc.
10505 NE 38th Place, Building 9
Kirkland, WA 98033
Extent of interest of assignee in application: Entire

Date: October 16, 2006

Syed Jafar Ali
Registration No. 58780
PILLSBURY WINTHROP SHAW PITTMAN LLP
P.O. Box 10500
McLean, VA 22102
703.770.7541
Customer No. 00909

17712 U.S. PTO
101606

PTO/SB/05 (07-06)
Approved for use through 01/31/2007. OMB 0651-0032
U.S. Patent and Trademark Office. U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

UTILITY PATENT APPLICATION TRANSMITTAL

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Attorney Docket No.	090611-0357945
First Inventor	LARRY BALDWIN
Title	See 1 in Addendum
Express Mail Label No.	

APPLICATION ELEMENTS

See MPEP chapter 600 concerning utility patent application contents.

1. [X] **Fee Transmittal Form** (e.g., PTO/SB/17)
   (Submit an original and a duplicate for fee processing)
2. [X] **Applicant claims small entity status.** See 37 CFR 1.27.
3. [X] **Specification** [*Total Pages* 43]
   Both the claims and abstract must start on a new page
   (For information on the preferred arrangement, see MPEP 608.01(a))
4. [X] **Drawing(s)** (*35 U.S.C. 113*) [*Total Sheets* 3]
5. **Oath or Declaration** [*Total Sheets* _____]
   a. [ ] Newly executed (original or copy)
   b. [ ] A copy from a prior application (37 CFR 1.63(d))
   (for continuation/divisional with Box 18 completed)
   i. [ ] DELETION OF INVENTOR(S)
   Signed statement attached deleting inventor(s) name in the prior application, see 37 CFR 1.63(d)(2) and 1.33(b).
6. [X] **Application Data Sheet.** See 37 CFR 1.76
7. [ ] **CD-ROM or CD-R** in duplicate, large table or Computer Program *(Appendix)*
   [ ] Landscape Table on CD
8. **Nucleotide and/or Amino Acid Sequence Submission**
   (if applicable, items a. – c. are required)
   a. [ ] Computer Readable Form (CRF)
   b. Specification Sequence Listing on:
      i. [ ] CD-ROM or CD-R (2 copies); or
      ii. [ ] Paper
   c. [ ] Statements verifying identity of above copies

ADDRESS TO: **Commissioner for Patents**
P.O. Box 1450
Alexandria VA 22313-1450

113268 U.S. PTO
11/580926
101606

ACCOMPANYING APPLICATION PARTS

9. [ ] **Assignment Papers** (cover sheet & document(s))
   Name of Assignee _____
10. [ ] **37 CFR 3.73(b) Statement** *(when there is an assignee)* [ ] **Power of Attorney**
11. [ ] **English Translation Document** *(if applicable)*
12. [ ] **Information Disclosure Statement** (PTO/SB/08 or PTO-1449)
    [ ] Copies of foreign patent documents, publications, & other information
13. [ ] **Preliminary Amendment**
14. [X] **Return Receipt Postcard** (MPEP 503)
    (Should be specifically itemized)
15. [ ] **Certified Copy of Priority Document(s)**
    (if foreign priority is claimed)
16. [ ] **Nonpublication Request** under 35 U.S.C. 122(b)(2)(B)(i). Applicant must attach form PTO/SB/35 or equivalent.
17. [ ] Other: _____

18. If a CONTINUING APPLICATION, *check appropriate box, and supply the requisite information below and in the first sentence of the specification following the title, or in an Application Data Sheet under 37 CFR 1.76:*

[ ] Continuation [ ] Divisional [ ] Continuation-in-part (CIP) of prior application No.: .....

Prior application information: *Examiner* _____ *Art Unit:* _____

19. CORRESPONDENCE ADDRESS

[X] The address associated with Customer Number: 00909 ***OR*** [ ] Correspondence address below

Name					
Address					
City		State		Zip Code	
Country		Telephone		Email	

Signature		Date	October 16, 2006
Name (Print/Type)	Syed Jafar Ali	Registration No. (Attorney/Agent)	58780

This collection of information is required by 37 CFR 1.53(b). The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: **Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**
If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Addendum

1. SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

PTO/SB/17 (01-06)
Approved for use through 07/31/2006. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995 no persons are required to respond to a collection of information unless it displays a valid OMB control number

Fees pursuant to the Consolidated Appropriations Act, 2005 (H.R. 4818).

FEE TRANSMITTAL
For FY 2006

☒ Applicant claims small entity status. See 37 CFR 1.27

TOTAL AMOUNT OF PAYMENT ($) 1,050.00

Complete if Known	
Application Number	TBD
Filing Date	October 16, 2006
First Named Inventor	LARRY BALDWIN
Examiner Name	
Art Unit	
Attorney Docket No.	090611-0357945

METHOD OF PAYMENT (check all that apply)

☐ Check ☐ Credit Card ☐ Money Order ☐ None ☐ Other (please identify): _____

☒ Deposit Account Deposit Account Number: 033975 Deposit Account Name: PILLSBURY WINTHROP SHAW PITTMAN LLP

For the above-identified deposit account, the Director is hereby authorized to: (check all that apply)

☒ Charge fee(s) indicated below ☐ Charge fee(s) indicated below, **except for the filing fee**

☒ Charge any additional fee(s) or underpayments of fee(s) under 37 CFR 1.16 and 1.17 ☒ Credit any overpayments

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

FEE CALCULATION (All the fees below are due upon filing or may be subject to a surcharge.)

1. BASIC FILING, SEARCH, AND EXAMINATION FEES

	FILING FEES		SEARCH FEES		EXAMINATION FEES		
		Small Entity		Small Entity		Small Entity	
Application Type	**Fee ($)**	**Fee ($)**	**Fee ($)**	**Fee ($)**	**Fee ($)**	**Fee ($)**	**Fees Paid ($)**
Utility	300	150	500	250	200	100	500.00
Design	200	100	100	50	130	65	
Plant	200	100	300	150	160	80	
Reissue	300	150	500	250	600	300	
Provisional	200	100	0	0	0	0	

2. EXCESS CLAIM FEES

Fee Description	Fee ($)	Small Entity Fee ($)
Each claim over 20 (including Reissues)	50	25
Each independent claim over 3 (including Reissues)	200	100
Multiple dependent claims	360	180

Total Claims	Extra Claims	Fee ($)	Fee Paid ($)
42 - 20 or HP =	22	x 25.00	= 550.00

HP = highest number of total claims paid for, if greater than 20.

Indep. Claims	Extra Claims	Fee ($)	Fee Paid ($)
3 - 3 or HP =	0	x	= 0.00

HP = highest number of independent claims paid for, if greater than 3.

Multiple Dependent Claims	
Fee ($)	**Fee Paid ($)**

3. APPLICATION SIZE FEE

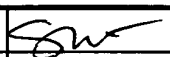
If the specification and drawings exceed 100 sheets of paper (excluding electronically filed sequence or computer listings under 37 CFR 1.52(e)), the application size fee due is $250 ($125 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).

Total Sheets	Extra Sheets	Number of each additional 50 or fraction thereof	Fee ($)	Fee Paid ($)
46 - 100 =	-54	/ 50 = _____ (round up to a whole number) x	125.00 =	0.00

4. OTHER FEE(S)

	Fees Paid ($)
Non-English Specification, $130 fee (no small entity discount)	
Other (e.g., late filing surcharge): _____	

SUBMITTED BY

Signature		Registration No. (Attorney/Agent)	58780	Telephone	703.770.7541
Name (Print/Type)	Syed Jafar Ali			Date	October 16, 2006

This collection of information is required by 37 CFR 1.136. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 30 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING or 371(c) DATE	GRP ART UNIT	FIL FEE REC'D	ATTY.DOCKET.NO	DRAWINGS	TOT CLAIMS	IND CLAIMS
11/580,926	10/16/2006	2626	1050	090611-0357945	3	42	3

CONFIRMATION NO. 6670

00909
PILLSBURY WINTHROP SHAW PITTMAN, LLP
P.O. BOX 10500
MCLEAN, VA22102

FILING RECEIPT

Date Mailed: 11/07/2006

Receipt is acknowledged of this regular Patent Application. It will be considered in its order and you will be notified as to the results of the examination. Be sure to provide the U.S. APPLICATION NUMBER, FILING DATE, NAME OF APPLICANT, and TITLE OF INVENTION when inquiring about this application. Fees transmitted by check or draft are subject to collection. Please verify the accuracy of the data presented on this receipt. **If an error is noted on this Filing Receipt, please mail to the Commissioner for Patents P.O. Box 1450 Alexandria Va 22313-1450. Please provide a copy of this Filing Receipt with the changes noted thereon. If you received a "Notice to File Missing Parts" for this application, please submit any corrections to this Filing Receipt with your reply to the Notice. When the USPTO processes the reply to the Notice, the USPTO will generate another Filing Receipt incorporating the requested corrections (if appropriate).**

Applicant(s)

Larry Baldwin, Maple Valley, WA;
Michael Tjalve, Bellevue, WA;
Blane Ebersold, Seattle, WA;
Tom Freeman, Mercer Island, WA;

Assignment For Published Patent Application

VoiceBox Technologies, Inc., Kirkland, WA

Power of Attorney: None

Domestic Priority data as claimed by applicant

Foreign Applications

If Required, Foreign Filing License Granted: 11/07/2006

The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is **US11/580,926**

Projected Publication Date: To Be Determined - pending completion of Missing Parts

Non-Publication Request: No

Early Publication Request: No

** SMALL ENTITY **

Title

Preliminary Class

704

PROTECTING YOUR INVENTION OUTSIDE THE UNITED STATES

Since the rights granted by a U.S. patent extend only throughout the territory of the United States and have no effect in a foreign country, an inventor who wishes patent protection in another country must apply for a patent in a specific country or in regional patent offices. Applicants may wish to consider the filing of an international application under the Patent Cooperation Treaty (PCT). An international (PCT) application generally has the same effect as a regular national patent application in each PCT-member country. The PCT process **simplifies** the filing of patent applications on the same invention in member countries, but **does not result** in a grant of "an international patent" and does not eliminate the need of applicants to file additional documents and fees in countries where patent protection is desired.

Almost every country has its own patent law, and a person desiring a patent in a particular country must make an application for patent in that country in accordance with its particular laws. Since the laws of many countries differ in various respects from the patent law of the United States, applicants are advised to seek guidance from specific foreign countries to ensure that patent rights are not lost prematurely.

Applicants also are advised that in the case of inventions made in the United States, the Director of the USPTO must issue a license before applicants can apply for a patent in a foreign country. The filing of a U.S. patent application serves as a request for a foreign filing license. The application's filing receipt contains further information and guidance as to the status of applicant's license for foreign filing.

Applicants may wish to consult the USPTO booklet, "General Information Concerning Patents" (specifically, the section entitled "Treaties and Foreign Patents") for more information on timeframes and deadlines for filing foreign patent applications. The guide is available either by contacting the USPTO Contact Center at 800-786-9199, or it can be viewed on the USPTO website at http://www.uspto.gov/web/offices/pac/doc/general/index.html.

For information on preventing theft of your intellectual property (patents, trademarks and copyrights), you may wish to consult the U.S. Government website, http://www.stopfakes.gov. Part of a Department of Commerce initiative, this website includes self-help "toolkits" giving innovators guidance on how to protect intellectual property in specific countries such as China, Korea and Mexico. For questions regarding patent enforcement issues, applicants may call the U.S. Government hotline at 1-866-999-HALT (1-866-999-4158).

LICENSE FOR FOREIGN FILING UNDER

Title 35, United States Code, Section 184

Title 37, Code of Federal Regulations, 5.11 & 5.15

GRANTED

The applicant has been granted a license under 35 U.S.C. 184, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" followed by a date appears on this form. Such licenses are issued in all applications where the conditions for issuance of a license have been met, regardless of whether or not a license may be required as set forth in 37 CFR 5.15. The scope and limitations of this license are set forth in 37 CFR 5.15(a) unless an earlier license has been issued under 37 CFR 5.15(b). The license is subject to revocation upon written notification. The date indicated is the effective date of the license, unless an earlier license of similar scope has been granted under 37 CFR 5.13 or 5.14.

This license is to be retained by the licensee and may be used at any time on or after the effective date thereof unless it is revoked. This license is automatically transferred to any related applications(s) filed under 37 CFR 1.53(d). This license is not retroactive.

The grant of a license does not in any way lessen the responsibility of a licensee for the security of the subject matter as imposed by any Government contract or the provisions of existing laws relating to espionage and the national security or the export of technical data. Licensees should apprise themselves of current regulations especially with respect to certain countries, of other agencies, particularly the Office of Defense Trade Controls, Department of State (with respect to Arms, Munitions and Implements of War (22 CFR 121-128)); the Bureau of Industry and Security, Department of Commerce (15 CFR parts 730-774); the Office of Foreign AssetsControl, Department of Treasury (31 CFR Parts 500+) and the Department of Energy.

NOT GRANTED

No license under 35 U.S.C. 184 has been granted at this time, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" DOES NOT appear on this form. Applicant may still petition for a license under 37 CFR 5.12, if a license is desired before the expiration of 6 months from the filing date of the application. If 6 months has lapsed from the filing date of this application and the licensee has not received any indication of a secrecy order under 35 U.S.C. 181, the licensee may foreign file the application pursuant to 37 CFR 5.15(b).

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371 (c) DATE	FIRST NAMED APPLICANT	ATTORNEY DOCKET NUMBER
11/580,926	10/16/2006	Larry Baldwin	090611-0357945

CONFIRMATION NO. 6670

00909
PILLSBURY WINTHROP SHAW PITTMAN, LLP
P.O. BOX 10500
MCLEAN, VA 22102

FORMALITIES LETTER

Date Mailed: 11/07/2006

NOTICE TO FILE MISSING PARTS OF NONPROVISIONAL APPLICATION

FILED UNDER 37 CFR 1.53(b)

Filing Date Granted

Items Required To Avoid Abandonment:

An application number and filing date have been accorded to this application. The item(s) indicated below, however, are missing. Applicant is given **TWO MONTHS** from the date of this Notice within which to file all required items and pay any fees required below to avoid abandonment. Extensions of time may be obtained by filing a petition accompanied by the extension fee under the provisions of 37 CFR 1.136(a).

- The oath or declaration is missing. *A properly signed oath or declaration in compliance with 37 CFR 1.63, identifying the application by the above Application Number and Filing Date, is required.*
  Note: If a petition under 37 CFR 1.47 is being filed, an oath or declaration in compliance with 37 CFR 1.63 signed by all available joint inventors, or if no inventor is available by a party with sufficient proprietary interest, is required.

The applicant needs to satisfy supplemental fees problems indicated below.

The required item(s) identified below must be timely submitted to avoid abandonment:

• To avoid abandonment, a surcharge (for late submission of filing fee, search fee, examination fee or oath or declaration) as set forth in 37 CFR 1.16(f) of $65 for a small entity in compliance with 37 CFR 1.27, must be submitted with the missing items identified in this letter.

SUMMARY OF FEES DUE:

Total additional fee(s) required for this application is **$65** for a small entity

- **$65** Surcharge.

Replies should be mailed to: Mail Stop Missing Parts
Commissioner for Patents

P.O. Box 1450
Alexandria VA 22313-1450

Registered users of EFS-Web may alternatively submit their reply to this notice via EFS-Web.
https://sportal.uspto.gov/authenticate/AuthenticateUserLocalEPF.html

For more information about EFS-Web please call the USPTO Electronic Business Center at **1-866-217-9197** or visit our website at http://www.uspto.gov/ebc.

If you are not using EFS-Web to submit your reply, you must include a copy of this notice.

Office of Initial Patent Examination (571) 272-4000, or 1-800-PTO-9199, or 1-800-972-6382
PART 3 - OFFICE COPY

Attorney Docket: 090611-0357945

OIPE
JAN 05 2007
PATENT & TRADEMARK OFFICE

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re PATENT APPLICATION of: LARRY BALDWIN *et al.*

Application No.: 11/580,926 Confirmation No: 6670

Filed: October 16, 2006 Group Art Unit: 2626

Title: SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

RESPONSE TO NOTICE TO FILE MISSING PARTS

Mail Stop Missing Parts
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

In response to the Notice to File Missing Parts of Application mailed November 7, 2006, a copy of which is enclosed, Applicants herewith submit the following:

Executed Combined Declaration and Power of Attorney of Inventors
Assignment and Recordation Cover Sheet

Please see attached PTO/SB/17 for payment of fees.

Applicants are entitled to claim small entity status.

January 5, 2007

Respectfully submitted,

PILLSBURY WINTHROP SHAW PITTMAN LLP

By _____
Syed Jafar Ali
Registration No. 58780
Telephone: 703.770.7541

P.O. Box 10500
McLean, VA 22102
Customer No.: 00909

400509482v1

OIPE MPE JAN 05 2007 PATENT & TRADEMARK OFFICE

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371 (c) DATE	FIRST NAMED APPLICANT	ATTORNEY DOCKET NUMBER
11/580,926	10/16/2006	Larry Baldwin	090611-0357945

CONFIRMATION NO. 6670

00909
PILLSBURY WINTHROP SHAW PITTMAN, LLP
P.O. BOX 10500
MCLEAN, VA 22102

FORMALITIES LETTER

Date Mailed: 11/07/2006

NOTICE TO FILE MISSING PARTS OF NONPROVISIONAL APPLICATION

FILED UNDER 37 CFR 1.53(b)

Filing Date Granted

Items Required To Avoid Abandonment:

An application number and filing date have been accorded to this application. The item(s) indicated below, however, are missing. Applicant is given **TWO MONTHS** from the date of this Notice within which to file all required items and pay any fees required below to avoid abandonment. Extensions of time may be obtained by filing a petition accompanied by the extension fee under the provisions of 37 CFR 1.136(a).

- The oath or declaration is missing. *A properly signed oath or declaration in compliance with 37 CFR 1.63, identifying the application by the above Application Number and Filing Date, is required.*
  Note: If a petition under 37 CFR 1.47 is being filed, an oath or declaration in compliance with 37 CFR 1.63 signed by all available joint inventors, or if no inventor is available by a party with sufficient proprietary interest, is required.

The applicant needs to satisfy supplemental fees problems indicated below.

The required item(s) identified below must be timely submitted to avoid abandonment:

- To avoid abandonment, a surcharge (for late submission of filing fee, search fee, examination fee or oath or declaration) as set forth in 37 CFR 1.16(f) of $65 for a small entity in compliance with 37 CFR 1.27, must be submitted with the missing items identified in this letter.

SUMMARY OF FEES DUE:

01/08/2007 KAHMED1 00000143 033975 11580926
01 FC:2051 65.00 DA

Total additional fee(s) required for this application is **$65** for a small entity

- **$65** Surcharge.

Replies should be mailed to: Mail Stop Missing Parts
Commissioner for Patents

P.O. Box 1450
Alexandria VA 22313-1450

Registered users of EFS-Web may alternatively submit their reply to this notice via EFS-Web.
https://sportal.uspto.gov/authenticate/AuthenticateUserLocalEPF.html

For more information about EFS-Web please call the USPTO Electronic Business Center at **1-866-217-9197** or visit our website at http://www.uspto.gov/ebc.

If you are not using EFS-Web to submit your reply, you must include a copy of this notice.

Office of Initial Patent Examination (571) 272-4000, or 1-800-PTO-9199, or 1-800-972-6382
PART 2 - COPY TO BE RETURNED WITH RESPONSE

OIPE JAN 05 2007 PATENT & TRADEMARK OFFICE

PTO/SB/17 (01-06)
Approved for use through 07/31/2006. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995 no persons are required to respond to a collection of information unless it displays a valid OMB control number

Fees pursuant to the Consolidated Appropriations Act, 2005 (H.R. 4818).

FEE TRANSMITTAL
For FY 2006

[X] Applicant claims small entity status. See 37 CFR 1.27

TOTAL AMOUNT OF PAYMENT ($) 65.00

Complete if Known	
Application Number	11/580,926
Filing Date	October 16, 2006
First Named Inventor	LARRY BALDWIN
Examiner Name	
Art Unit	2626
Attorney Docket No.	090611-0357945

METHOD OF PAYMENT (check all that apply)

[ ] Check [ ] Credit Card [ ] Money Order [ ] None [ ] Other (please identify): _____

[X] Deposit Account Deposit Account Number: 033975 Deposit Account Name: PILLSBURY WINTHROP SHAW PITTMAN LLP

For the above-identified deposit account, the Director is hereby authorized to: (check all that apply)

[X] Charge fee(s) indicated below [ ] Charge fee(s) indicated below, **except for the filing fee**

[X] Charge any additional fee(s) or underpayments of fee(s) under 37 CFR 1.16 and 1.17 [X] Credit any overpayments

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

FEE CALCULATION (All the fees below are due upon filing or may be subject to a surcharge.)

1. BASIC FILING, SEARCH, AND EXAMINATION FEES

	FILING FEES		SEARCH FEES		EXAMINATION FEES		
		Small Entity		Small Entity		Small Entity	
Application Type	Fee ($)	Fee ($)	Fee ($)	Fee ($)	Fee ($)	Fee ($)	Fees Paid ($)
Utility	300	150	500	250	200	100	
Design	200	100	100	50	130	65	
Plant	200	100	300	150	160	80	
Reissue	300	150	500	250	600	300	
Provisional	200	100	0	0	0	0	

2. EXCESS CLAIM FEES

Fee Description	Fee ($)	Small Entity Fee ($)
Each claim over 20 (including Reissues)	50	25
Each independent claim over 3 (including Reissues)	200	100
Multiple dependent claims	360	180

Total Claims _____ - 20 or HP = _____ x _____ = _____ Extra Claims / Fee ($) / Fee Paid ($)

HP = highest number of total claims paid for, if greater than 20.

Indep. Claims _____ - 3 or HP = _____ x _____ = _____ Extra Claims / Fee ($) / Fee Paid ($)

HP = highest number of independent claims paid for, if greater than 3.

Multiple Dependent Claims — Fee ($) _____ Fee Paid ($) _____

3. APPLICATION SIZE FEE

If the specification and drawings exceed 100 sheets of paper (excluding electronically filed sequence or computer listings under 37 CFR 1.52(e)), the application size fee due is $250 ($125 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).

Total Sheets _____ - 100 = _____ Extra Sheets / 50 = _____ Number of each additional 50 or fraction thereof (round up to a whole number) x Fee ($) 125.00 = Fee Paid ($) 0.00

4. OTHER FEE(S)

Non-English Specification, $130 fee (no small entity discount) Fees Paid ($) _____

Other (e.g., late filing surcharge): Surcharge for late filing of Declaration 65.00

SUBMITTED BY

Signature	[signature]	Registration No. (Attorney/Agent)	58780	Telephone	703.770.7541
Name (Print/Type)	Syed Jafar Ali			Date	January 5, 2007

This collection of information is required by 37 CFR 1.136. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 30 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

OIPE
JAN 05 2007
PATENT & TRADEMARK OFFICE

FOR UTILITY ORIGINAL DECLARATION

RULE 63 (37 C.F.R. 1.63)
DECLARATION AND POWER OF ATTORNEY
FOR PATENT APPLICATION
IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

As a below named inventor, I hereby declare that my residence, post office address and citizenship are as stated below next to my name, and I believe I am an original, first and joint inventor of the subject matter which is claimed and for which a patent is sought on the invention entitled SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE, the specification of which was filed on October 16, 2006 as U.S. Application No. 11/580,926.

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above. I acknowledge the duty to disclose all information known to me to be material to patentability as defined in 37 C.F.R. 1.56. Except as noted below, I hereby claim foreign priority benefits under 35 U.S.C. 119(a)-(d) or 365(b) of any foreign application(s) for patent or inventor's certificate, or 365(a) of any PCT International Application which designated at least one other country than the United States, listed below and have also identified below any foreign application for patent or inventor's certificate, or PCT International Application, filed by me or my assignee disclosing the subject matter claimed in this application and having a filing date (1) before that of the application on which priority is claimed, or (2) if no priority claimed, before the filing date of this application:

PRIOR FOREIGN APPLICATION(S)

Number	Country	Filed	Date First Laid Open Or Published	Date Patented or Granted	Priority Claimed

Except as noted below, I hereby claim domestic priority benefit under 35 U.S.C. 119(e) or 120 and/or 365(c) of the indicated United States applications listed below and PCT international applications listed above or below and, if this is a continuation-in-part (CIP) application, insofar as the subject matter disclosed and claimed in this application is in addition to that disclosed in such prior applications, I acknowledge the duty to disclose all information known to me to be material to patentability as defined in 37 C.F.R. 1.56 which became available between the filing date of each such prior application and the national or PCT international filing date of this application:

PRIOR U.S. PROVISIONAL, NONPROVISIONAL AND/OR PCT APPLICATION(S)

Application Number	Filed	Status pending, abandoned, patented	Priority Claimed

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

And I hereby appoint Pillsbury Winthrop Shaw Pittman LLP, Intellectual Property Group, (to whom all communications are to be directed), and persons of that firm who are associated with USPTO Customer No. 00909 individually and collectively my attorneys to prosecute this application and to transact all business in the Patent and Trademark Office connected therewith and with the resulting patent, and I hereby authorize them to delete from that Customer No. names of persons no longer with their firm, to add new persons of their Firm to that Customer No., and to act and rely on instructions from and communicate directly with the person/assignee/attorney/firm/organization who/which first sends/sent this case to them and by whom/which I hereby declare that I have consented after full disclosure to be represented unless/until I instruct the above Firm and/or an attorney of that Firm in writing to the contrary.

Power of Attorney to Customer Number 00909

INVENTOR'S SIGNATURE: Larry Baldwin **Date:** 12/28/2006

Name	LARRY		BALDWIN
	First	Middle Initial	Family Name
Residence	Maple Valley	WA	United States of America
	City	State/Foreign Country	Country of Citizenship
Mailing Address	25498 S.E. 274th Place, Maple Valley, WA 98038		

INVENTOR'S SIGNATURE: **Date:** 12/28/2006

Name	TOM		FREEMAN
	First	Middle Initial	Family Name
Residence	Mercer Island	WA	United States of America
	City	State/Foreign Country	Country of Citizenship
Mailing Address	6735 83rd Avenue, S.E., Mercer Island, WA 98040		

400495356v1

INVENTOR'S SIGNATURE: [signature] **Date:** 12/28/2006

Name	MICHAEL		TJALVE
	First	Middle Initial	Family Name
Residence	Bellevue	WA	Denmark
	City	State/Foreign Country	Country of Citizenship
Mailing Address	14904 S.E. 47th Court, Bellevue, WA 98006		

INVENTOR'S SIGNATURE: [signature] **Date:** 12/28/2006

Name	BLANE		EBERSOLD
	First	Middle Initial	Family Name
Residence	Seattle	WA	United States of America
	City	State/Foreign Country	Country of Citizenship
Mailing Address	4721 Admiral Way, S.W., Seattle, WA 98116		

INVENTOR'S SIGNATURE: [signature] **Date:** 12/28/2006

Name	CHRIS		WEIDER
	First	Middle Initial	Family Name
Residence	Everett	WA	United States of America
	City	State/Foreign Country	Country of Citizenship
Mailing Address	2618 Grand Avenue, #B-508, Everett, WA 98201		

Atty. Dkt. No. 090611-0357945

400495356v1

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING or 371(c) DATE	GRP ART UNIT	FIL FEE REC'D	ATTY.DOCKET.NO	DRAWINGS	TOT CLAIMS	IND CLAIMS
11/580,926	10/16/2006	2626	1115	090611-0357945	3	42	3

CONFIRMATION NO. 6670

909
PILLSBURY WINTHROP SHAW PITTMAN, LLP
P.O. BOX 10500
MCLEAN, VA22102

UPDATED FILING RECEIPT

Date Mailed: 01/23/2007

Receipt is acknowledged of this regular Patent Application. It will be considered in its order and you will be notified as to the results of the examination. Be sure to provide the U.S. APPLICATION NUMBER, FILING DATE, NAME OF APPLICANT, and TITLE OF INVENTION when inquiring about this application. Fees transmitted by check or draft are subject to collection. Please verify the accuracy of the data presented on this receipt. **If an error is noted on this Filing Receipt, please mail to the Commissioner for Patents P.O. Box 1450 Alexandria Va 22313-1450. Please provide a copy of this Filing Receipt with the changes noted thereon. If you received a "Notice to File Missing Parts" for this application, please submit any corrections to this Filing Receipt with your reply to the Notice. When the USPTO processes the reply to the Notice, the USPTO will generate another Filing Receipt incorporating the requested corrections (if appropriate).**

Applicant(s)

Larry Baldwin, Maple Valley, WA;
Tom Freeman, Mercer Island, WA;
Michael Tjalve, Bellevue, WA;
Blane Ebersold, Seattle, WA;
Chris Weider, Everett, WA;

Assignment For Published Patent Application

VoiceBox Technologies, Inc., Kirkland, WA

Power of Attorney: The patent practitioners associated with Customer Number 00909

Domestic Priority data as claimed by applicant

Foreign Applications

If Required, Foreign Filing License Granted: 11/07/2006

The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is **US11/580,926**

Projected Publication Date: 04/17/2008

Non-Publication Request: No

Early Publication Request: No

** SMALL ENTITY **

Title

System and method for a cooperative conversational voice user interface

Preliminary Class

704

PROTECTING YOUR INVENTION OUTSIDE THE UNITED STATES

Since the rights granted by a U.S. patent extend only throughout the territory of the United States and have no effect in a foreign country, an inventor who wishes patent protection in another country must apply for a patent in a specific country or in regional patent offices. Applicants may wish to consider the filing of an international application under the Patent Cooperation Treaty (PCT). An international (PCT) application generally has the same effect as a regular national patent application in each PCT-member country. The PCT process **simplifies** the filing of patent applications on the same invention in member countries, but **does not result** in a grant of "an international patent" and does not eliminate the need of applicants to file additional documents and fees in countries where patent protection is desired.

Almost every country has its own patent law, and a person desiring a patent in a particular country must make an application for patent in that country in accordance with its particular laws. Since the laws of many countries differ in various respects from the patent law of the United States, applicants are advised to seek guidance from specific foreign countries to ensure that patent rights are not lost prematurely.

Applicants also are advised that in the case of inventions made in the United States, the Director of the USPTO must issue a license before applicants can apply for a patent in a foreign country. The filing of a U.S. patent application serves as a request for a foreign filing license. The application's filing receipt contains further information and guidance as to the status of applicant's license for foreign filing.

Applicants may wish to consult the USPTO booklet, "General Information Concerning Patents" (specifically, the section entitled "Treaties and Foreign Patents") for more information on timeframes and deadlines for filing foreign patent applications. The guide is available either by contacting the USPTO Contact Center at 800-786-9199, or it can be viewed on the USPTO website at http://www.uspto.gov/web/offices/pac/doc/general/index.html.

For information on preventing theft of your intellectual property (patents, trademarks and copyrights), you may wish to consult the U.S. Government website, http://www.stopfakes.gov. Part of a Department of Commerce initiative, this website includes self-help "toolkits" giving innovators guidance on how to protect intellectual property in specific countries such as China, Korea and Mexico. For questions regarding patent enforcement issues, applicants may call the U.S. Government hotline at 1-866-999-HALT (1-866-999-4158).

LICENSE FOR FOREIGN FILING UNDER

Title 35, United States Code, Section 184

Title 37, Code of Federal Regulations, 5.11 & 5.15

GRANTED

The applicant has been granted a license under 35 U.S.C. 184, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" followed by a date appears on this form. Such licenses are issued in all applications where the conditions for issuance of a license have been met, regardless of whether or not a license may be required as set forth in 37 CFR 5.15. The scope and limitations of this license are set forth in 37 CFR 5.15(a) unless an earlier license has been issued under 37 CFR 5.15(b). The license is subject to revocation upon written notification. The date indicated is the effective date of the license, unless an earlier license of similar scope has been granted under 37 CFR 5.13 or 5.14.

This license is to be retained by the licensee and may be used at any time on or after the effective date thereof unless it is revoked. This license is automatically transferred to any related applications(s) filed under

37 CFR 1.53(d). This license is not retroactive.

The grant of a license does not in any way lessen the responsibility of a licensee for the security of the subject matter as imposed by any Government contract or the provisions of existing laws relating to espionage and the national security or the export of technical data. Licensees should apprise themselves of current regulations especially with respect to certain countries, of other agencies, particularly the Office of Defense Trade Controls, Department of State (with respect to Arms, Munitions and Implements of War (22 CFR 121-128)); the Bureau of Industry and Security, Department of Commerce (15 CFR parts 730-774); the Office of Foreign AssetsControl, Department of Treasury (31 CFR Parts 500+) and the Department of Energy.

NOT GRANTED

No license under 35 U.S.C. 184 has been granted at this time, if the phrase "IF REQUIRED, FOREIGN FILING LICENSE GRANTED" DOES NOT appear on this form. Applicant may still petition for a license under 37 CFR 5.12, if a license is desired before the expiration of 6 months from the filing date of the application. If 6 months has lapsed from the filing date of this application and the licensee has not received any indication of a secrecy order under 35 U.S.C. 181, the licensee may foreign file the application pursuant to 37 CFR 5.15(b).

OIPE IAP97 SEP 27 2007 PATENT & TRADEMARK OFFICE

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANTS :	Larry BALDWIN *et al.*	CONFIRMATION NO.:	6670
SERIAL NUMBER :	11/580,926	EXAMINER:	*Unknown*
FILING DATE :	October 16, 2006	ART UNIT:	2626
FOR :	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE		

INFORMATION DISCLOSURE STATEMENT (IDS)

Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313-1450

Sir:

Pursuant to the duty of disclosure under 37 C.F.R. §§ 1.56, 1.97 and 1.98, Applicants hereby make of record in the above-identified application the documents listed on the attached Form PTO-1449.

Applicants have listed publication dates on the attached PTO-1449 based on information presently available to the undersigned. However, the listed publication dates should not be construed as an admission that the information was actually published on the date indicated. Where the publication date of a listed document does not provide a month of publication, the year of publication of the listed document is sufficiently earlier than the effective U.S. filing date and any foreign priority date so that the month of publication is not an issue.

Applicants reserve the right to establish the patentability of the claimed invention over any of the information provided herewith, and/or to prove that this information may not be

prior art, and/or to prove that this information may not be enabling for the teachings purportedly offered.

This Information Disclosure Statement is being filed within three months of the filing date of a national application other than a continued prosecution application under 37 C.F.R. §1.53(d), or within three months of the date of entry of the national stage as set forth in 37 C.F.R. §1.491 in an international application, or before the mailing date of a first Office Action on the merits, or before the mailing date of a first Office Action after the filing of a request for continued examination under 37 C.F.R. §1.114. No certification or fee is required. 37 C.F.R. §1.97(b).

A copy of each cited U.S. patent is not submitted herewith pursuant to the Official Gazette Notice of August 5, 2003, wherein the Office waived the requirement under 37 C.F.R. §1.98(a)(2)(i) for all U.S. national phase applications filed after June 30, 2003.

It is respectfully requested that the Examiner initial and return a copy of the enclosed PTO-1449.

Respectfully submitted,

PILLSBURY WINTHROP SHAW PITTMAN LLP

By _____
Syed Jafar Ali
Registration No.: 58,780

Customer No.
00909

P. O. Box 10500
McLean, Virginia 22102
Telephone: (703) 770-7900
Telefax: (703) 770-7901

Dated: September 27, 2007

OIPE IAP97 SEP 27 2007 PATENT & TRADEMARK OFFICE

INFORMATION DISCLOSURE CITATION IN AN APPLICATION

Form PTO-1449

Docket Number	Application Number
090611-0357945	11/580,926
Applicant(s)	
Larry BALDWIN *et al.*	
Filing Date	Group Art Unit
October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	1A	5,155,743	10/13/1992	Jacobs	375	28	
	1B	5,386,556	01/31/1995	Hedin et al.	395	600	
	1C	5,424,947	06/13/1995	Nagao et al.	364	419.08	
	1D	5,471,318	11/28/1995	Ahuja et al.	358	400	
	1E	5,475,733	12/12/1995	Eisdorfer et al.	379	52	
	1F	5,499,289	03/12/1996	Bruno et al.	379	220	
	1G	5,517,560	05/14/1996	Greenspan	379	114	
	1H	5,533,108	07/02/1996	Harris et al.	379	201	
	1I	5,537,436	07/16/1996	Bottoms et al.	375	222	
	1J	5,539,744	07/23/1996	Chu et al.	370	60	
	1K	5,557,667	09/17/1996	Bruno et al.	379	201	
	1L	5,563,937	10/08/1996	Bruno et al.	379	201	
	1M	5,590,039	12/31/1996	Ikeda et al.	395	759	
	1N	5,633,922	05/27/1997	August et al.	379	220	
	1O	5,675,629	10/07/1997	Raffel et al.	379	58	
	1P	5,708,422	01/13/1998	Blonder et al.	340	825.34	
	1Q	5,721,938	02/24/1998	Stuckey	395	754	
	1R	5,722,084	02/24/1998	Chakrin et al.	455	551	
	1S	5,742,763	04/21/1998	Jones	395	200.3	
	1T	5,752,052	05/12/1998	Richardson et al.	395	759	
	1U	5,754,784	05/19/1998	Garland et al.	395	200.49	
	1V	5,761,631	06/02/1998	Nasukawa	704	9	
	1W	5,794,050	08/11/1998	Dahlgren et al.	395	708	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO
	1X	WO 01/78065	10/18/2001	International				

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

	1Y	REUTERS, "IBM to Enable Honda Drivers to Talk to Cars", Charles Schwab & Co., Inc., July 28, 2002, 1 page.
	1Z	LIN, Bor-shen, *et al.*, "A Distributed Architecture for Cooperative Spoken Dialogue Agents with Coherent Dialogue State and History", ASRU'99, 1999, 4 pages.
	1AA	KUHN, Thomas, *et al.*, "Hybrid In-Car Speech Recognition for Mobile Multimedia Applications", Vehicular Technology Conference, IEEE, July 1999, pages 2009-2013.

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	2A	5,799,276	08/25/1998	Komissarchik *et al.*	704	251	
	2B	5,802,510	09/01/1998	Jones	707	2	
	2C	5,832,221	11/03/1998	Jones	375	200.36	
	2D	5,878,386	03/02/1999	Coughlin	704	10	
	2E	5,895,466	04/20/1999	Goldberg et al.	707	5	
	2F	5,911,120	06/08/1999	Jarett et al.	455	417	
	2G	5,918,222	06/29/1999	Fukui *et al.*	707	1	
	2H	5,926,784	07/20/1999	Richardson et al.	704	9	
	2I	5,933,822	08/03/1999	Braden-Harder et al.	707	5	
	2J	5,953,393	09/14/1999	Culbreth et al.	379	88.25	
	2K	5,963,894	10/05/1999	Richardson et al.	704	9	
	2L	5,963,940	10/05/1999	Liddy et al.	707	5	
	2M	5,987,404	11/16/1999	Della Pietra et al.	704	9	
	2N	5,991,721	11/23/1999	Asano et al.	704	257	
	2O	5,995,119	11/30/1999	Cosatto et al.	345	473	
	2P	6,009,382	12/28/1999	Martino et al.	704	1	
	2Q	6,014,559	01/11/2000	Amin	455	413	
	2R	6,021,384	02/01/2000	Gorin et al.	704	1	
	2S	6,044,347	03/28/2000	Abella et al.	704	272	
	2T	6,049,602	04/11/2000	Foladare et al.	379	265	
	2U	6,058,187	05/02/2000	Chen	380	21	
	2V	6,078,886	06/20/2000	Dragosh et al.	704	270	
	2W	6,081,774	06/27/2000	De Hita et al.	704	9	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

	2X	BELVIN, Robert, *et al.*, "Development of the HRL Route Navigation Dialogue System", Proceedings of the First International Conference on Human Language Technology Research, San Diego, 2001, pages 1-5.
	2Y	LIND, R., *et al.*, "The Network Vehicle—A Glimpse into the Future of Mobile Multi-Media", IEEE Aerosp. Electron. Systems Magazine, Volume 14, No. 9, September 1999, pages 27-32.

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	3A	6,101,241	08/08/2000	Boyce et al.	379	88.01	
	3B	6,119,087	09/12/2000	Kuhn et al.	704	270	
	3C	6,134,235	10/17/2000	Goldman et al.	370	352	
	3D	6,144,667	11/07/2000	Doshi et al.	370	401	
	3E	6,160,883	12/12/2000	Jackson et al.	379	230	
	3F	6,173,279	01/09/2001	Levin et al.	707	5	
	3G	6,175,858	01/16/2001	Bulfer et al.	709	206	
	3H	6,185,535	02/06/2001	Hedin *et al.*	704	270	
	3I	6,192,110	02/20/2001	Abella et al.	379	88.01	
	3J	6,192,338	02/20/2001	Haszto et al.	704	257	
	3K	6,195,651	02/27/2001	Handel *et al.*	707	2	
	3L	6,208,972	03/27/2001	Grant *et al.*	704	275	
	3M	6,219,346	04/17/2001	Maxemchuk	370	338	
	3N	6,219,643	04/17/2001	Cohen et al.	704	257	
	3O	6,233,556	05/15/2001	Teunen et al.	704	250	
	3P	6,233,559	05/15/2001	Balakrishnan	704	275	
	3Q	6,246,981	06/12/2001	Papineni et al.	704	235	
	3R	6,272,455	08/07/2001	Hoshen et al.	704	1	
	3S	6,292,767	09/18/2001	Jackson et al.	704	1	
	3T	6,314,402	11/06/2001	Monaco et al.	704	275	
	3U	6,366,886	04/02/2002	Dragosh et al.	704	270.1	
	3V	6,385,646	05/07/2002	Brown et al.	709	217	
	3W	6,393,428	05/21/2002	Miller et al.	707	102	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

	3X	ZHAO, Yilin, "Telematics: Safe and Fun Driving", IEEE Intelligent Systems, Volume 17, Issue 1, 2002, pages 10-14.

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	4A	6,404,878	06/11/2002	Jackson et al.	379	221.01	
	4B	6,408,272	06/18/2002	White *et al.*	704	270.1	
	4C	6,411,810	06/25/2002	Maxemchuk	455	453	
	4D	6,418,210	07/09/2002	Sayko	379	142.15	
	4E	6,420,975	07/16/2002	DeLine *et al.*	340	815.4	
	4F	6,430,285	08/06/2002	Bauer et al.	379	265.01	
	4G	6,434,523	08/13/2002	Monaco	704	257	
	4H	6,434,524	08/13/2002	Weber	704	257	
	4I	6,442,522	08/27/2002	Carberry et al.	704	257	
	4J	6,446,114	09/03/2002	Bulfer et al.	709	206	
	4K	6,453,153	09/17/2002	Bowker et al.	455	67.4	
	4L	6,453,292	09/17/2002	Ramaswamy et al.	704	235	
	4M	6,456,711	09/24/2002	Cheung et al.	379	265.09	
	4N	6,498,797	12/24/2002	Anerousis et al.	370	522	
	4O	6,499,013	12/24/2002	Weber	704	257	
	4P	6,501,833	12/31/2002	Phillips *et al.*	379	88.07	
	4Q	6,501,834	12/31/2002	Milewski et al.	379	93.24	
	4R	6,523,061	02/18/2003	Halverson et al.	709	202	
	4S	6,532,444	03/11/2003	Weber	704	257	
	4T	6,539,348	03/25/2003	Bond et al.	704	9	
	4U	6,553,372	04/22/2003	Brassell et al.	707	5	
	4V	6,556,973	04/29/2003	Lewin	704	277	
	4W	6,560,576	05/06/2003	Cohen et al.	704	270	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	5A	6,567,778	05/20/2003	Chao Chang et al.	704	257	
	5B	6,570,964	05/27/2003	Murveit et al.	379	67.1	
	5C	6,574,624	06/03/2003	Johnson et al.	707	5	
	5D	6,594,257	07/15/2003	Doshi et al.	370	352	
	5E	6,604,077	08/05/2003	Dragosh et al.	704	270.1	
	5F	6,611,692	08/26/2003	Raffel et al.	455	552	
	5G	6,614,773	09/02/2003	Maxemchuk	370	337	
	5H	6,615,172	09/02/2003	Bennett *et al.*	704	257	
	5I	6,629,066	09/30/2003	Jackson et al.	704	9	
	5J	6,650,747	11/18/2003	Bala et al.	379	265.06	
	5K	6,691,151	02/10/2004	Cheyer et al.	709	202	
	5L	6,741,931	05/25/2004	Kohut *et al.*	701	209	
	5M	6,937,977	08/30/2005	Gerson	704	201	
	5N	6,980,092	12/27/2005	Turnbull *et al.*	340	425.5	
	5O	6,996,531	02/07/2006	Korall *et al.*	704	270	
	5P	7,020,609	03/28/2003	Thrift *et al.*	704	270.1	
	5Q	7,027,975	04/11/2006	Pazandak *et al.*	704	9	
	5R	7,062,488	06/13/2006	Reisman	707	8	
	5S	7,092,928	08/15/2006	Elad *et al.*	706	60	
	5T	7,137,126	11/14/06	Coffman *et al.*	719	328	
	5U	2002/0143535	10/03/2002	Kist *et al.*	704	251	
	5V	2003/0182132	09/25/2003	Niemoeller	704	275	
	5W	2004/0044516	03/04/2004	Kennewick *et al.*	704	5	
	5X	2004/0193420	09/30/2004	Kennewick *et al.*	704	257	
	5Y	2005/0021826	01/27/2005	Kumar	709	232	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 October 2001 (18.10.2001)

PCT

(10) International Publication Number
WO 01/78065 A1

(51) International Patent Classification[7]: G10L 15/22, 15/26

(21) International Application Number: PCT/US01/11138

(22) International Filing Date: 6 April 2001 (06.04.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/543,790 6 April 2000 (06.04.2000) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 09/543,790 (CON)
Filed on 6 April 2000 (06.04.2000)

(71) Applicant *(for all designated States except US)*: **ONE VOICE TECHNOLOGIES, INC.** [US/US]; Suite 240, 6333 Greenwich Drive, San Diego, CA 92122 (US).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **WEBER, Dean, C.** [US/US]; 11422 Cypress Terrace Place, San Diego, CA 92131 (US).

(74) Agents: **SZE, James, Y.** et al.; Pillsbury Winthrop LLP, P.O. Box 7880, 50 Fremont Street, San Francisco, CA 94120-7880 (US).

(81) Designated States *(national)*: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 01/78065 A1

(54) Title: NATURAL LANGUAGE AND DIALOGUE GENERATION PROCESSING

(57) Abstract: A system and method for interacting with objects, via a computer using utterances, speech processing and natural language processing. A Dialogue Generation Log provides a conversational memory by relating earlier generated utterances with possible new utterances. The Dialogue Generation Log encompasses a memory structure relating the prior spoken utterances with objects, and conversational contexts. The speech processor searches the Dialogue Generation Log for a matching phrase for the proposed utterance, and determines whether the utterance should be repeated. The determination may be made on a case-by-case basis, depending upon the context of the conversational utterance. Custom user information may be stored globally using a global user registry for remote access. The system reduces the repetitive nature of computer generated conversation by not voicing utterances that have been generated recently, and by learning user preferences from already voiced conversations.

NATURAL LANGUAGE AND DIALOGUE GENERATION PROCESSING

BACKGROUND OF THE INVENTION

I. Field of the Invention

The present invention relates to speech recognition for an object-based computer user interface. More specifically, the embodiments relate to a novel method and system for user interaction with a computer using speech recognition and natural language processing.

II. Description of the Related Art

As computers have become more prevalent it has become clear that many people have great difficulty understanding and communicating with computers. A user must often learn archaic commands and non-intuitive procedures in order to operate the computer. For example, most personal computers use windows-based operating systems that are largely menu-driven. This requires that the user learn which menu commands or sequence of commands produce the desired results.

Furthermore, traditional interaction with a computer is often slowed by manual input devices such as keyboards or mice. Many computer users are not fast typists. As a result, much time is spent communicating commands and words to the computer through these manual input devices. It is becoming clear that an easier, faster and more intuitive method of communicating

with computers and networked objects, such as web-sites, is needed.

One proposed method of computer interaction is speech recognition. Speech recognition involves software and hardware that act together to audibly detect human speech and translate the detected speech into a string of words. As is known in the art, speech recognition works by breaking down sounds the hardware detects into smaller, non-divisible sounds called phonemes. Phonemes are distinct units of sound. For example, the word "those" is made up of three phonemes; the first is the "th" sound, the second is the "o" sound, and the third is the "s" sound. The speech recognition software attempts to match the detected phonemes with known words from a stored dictionary. An example of a speech recognition system is given in U.S. Patent No. 4,783,803, entitled "SPEECH RECOGNITION APPARATUS AND METHOD", issued November 8, 1998, assigned to Dragon Systems, Inc., and incorporated herein by reference. Presently, there are many commercially available speech recognition software packages available from such companies as Dragon Systems, Inc. and International Business Machines, Inc.

One limitation of these speech recognition software packages or systems is that they typically only perform command and control or dictation functions. Thus, the user is still required to learn a vocabulary of commands in order to operate the computer.

A proposed enhancement to these speech recognition systems is to process the detected words using a natural language processing system. Natural language processing generally involves determining a conceptual "meaning" (e.g., what meaning the speaker intended to convey) of the detected words by analyzing their grammatical relationship and relative context. For example, U.S. Patent No. 4,887,212, entitled "PARSER FOR NATURAL LANGUAGE TEXT", issued December 12, 1989, assigned to International Business Machines Corporation

and incorporated by reference herein teaches a method of parsing an input stream of words by using word isolation, morphological analysis, dictionary look-up and grammar analysis.

Natural language processing used in concert with speech recognition provides a powerful tool for operating a computer using spoken words rather than manual input such as a keyboard or mouse. However, one drawback of a conventional natural language processing system is that it may fail to determine the correct "meaning" of the words detected by the speech recognition system. In such a case, the user is typically required to recompose or restate the phrase, with the hope that the natural language processing system will determine the correct "meaning" on subsequent attempts. Clearly, this may lead to substantial delays as the user is required to restate the entire sentence or command. Another drawback of conventional systems is that the processing time required for the speech recognition can be prohibitively long. This is primarily due to the finite speed of the processing resources as compared with the large amount of information to be processed. For example, in many conventional speech recognition programs, the time required to recognize the utterance is long due to the size of the dictionary file being searched.

An additional drawback of conventional speech recognition and natural language processing systems is that they are not interactive, and thus are unable to cope with new situations. When a computer system encounters unknown or new networked objects, new relationships between the computer and the objects are formed. Conventional speech recognition and natural language processing systems are unable to cope with the situations that result from the new relationships posed by previously unknown networked objects. As a result, a conversational-style interaction with the computer is not possible. The user is required to

communicate complete concepts to the computer. The user is not able to speak in sentence fragments because the meaning of these sentence fragments (which is dependent on the meaning of previous utterances) will be lost.

Moreover, conventional speech recognition and natural language processing systems fail to generate intelligent conversations because they lack a conversational memory. In other words, conventional systems do not remember the contents of their conversations with users. As a result, conversations with such system are repetitive, and may result in users repeatedly telling the systems the same facts over-and-over, or worse, have systems fail to learn and remember simple concepts.

Another drawback of conventional speech recognition and natural language processing systems is that once a user successfully "trains" a computer system to recognize the user's speech and voice commands, the user cannot easily move to another computer without having to undergo the process of training the new computer. As a result, changing a user's computer workstations or location results in wasted time by users that need to re-train the new computer to the user's speech habits and voice commands.

What is needed is an interactive user interface for a computer that utilizes speech recognition and natural language processing which avoids the drawbacks mentioned above.

BRIEF DESCRIPTION OF THE DRAWINGS

The features, objects and advantages of the embodiments discussed herein will become more apparent from the detailed description set forth below when taken in conjunction with the

drawings in which like reference characters identify correspondingly throughout and wherein:

FIG. 1 is a functional block diagram of an exemplary computer system;

FIG. 2 is an expanded functional block diagram of the CPU and storage medium of the computer system of FIG. 1;

FIGS. 3A-D are a flowchart of the method of providing interactive speech recognition and natural language processing to a computer;

FIG. 4 is a diagram of selected columns of an exemplary natural language processing (NLP) database;

FIG. 5 is a diagram of an exemplary Database Definition File (DDF);

FIG. 6 is a diagram of selected columns of an exemplary object table;

FIGS. 7A-D are a flowchart illustrating the linking of interactive speech recognition and natural language processing to a networked object, such as a web-page;

FIG. 8 is a diagram depicting a computer system connecting to other computers, storage media, and web-sites via the Internet;

FIG. 9 is a diagram of an exemplary user global user registry;

FIGS. 10A-B are flowcharts illustrating alternate embodiments of the retrieval and enabling of an individual's user global user registry during login at a computer workstation; and

FIG. 11 is a diagram depicting an exemplary Dialogue Generation Log.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The embodiments discussed herein will now be disclosed with reference to a functional

block diagram of an exemplary computer system **100** of FIG. 1. In FIG. 1, computer system **100** includes a central processing unit (CPU) **102**. The CPU **102** may be any general purpose microprocessor or microcontroller as is known in the art, appropriately programmed to perform the method described herein with reference to FIGS. 3A-D. The software for programming the CPU can be found at storage medium **108** or alternatively from another location across a computer network. For example, the CPU **102** may be a conventional microprocessor such as the Pentium III™ processor manufactured by Intel Corporation or the like.

CPU **102** communicates with a plurality of peripheral equipment, including a display **104**, manual input **106**, storage medium **108**, microphone **110**, speaker **112**, data input port **114** and network interface **116**. Display **104** may be a visual display such as a CRT, LCD screen, touch-sensitive screen, or other monitors as are known in the art for visually displaying images and text to a user. Manual input **106** may be a conventional keyboard, keypad, mouse, trackball, or other input device as is known in the art for the manual input of data. Storage medium **108** may be a conventional read/write memory such as a magnetic disk drive, floppy disk drive, CD-ROM drive, silicon memory or other memory device as is known in the art for storing and retrieving data. Significantly, storage medium **108** may be remotely located from CPU **102**, and be connected to CPU **102** via a network such as a local area network (LAN), or a wide area network (WAN), or the Internet. Microphone **110** may be any suitable microphone as is known in the art for providing audio signals to CPU **102**. Speaker **112** may be any suitable speaker as is known in the art for reproducing audio signals from CPU **102**. It is understood that microphone **110** and speaker **112** may include appropriate digital-to-analog and analog-to-digital conversion circuitry as appropriate. Data input port **114** may be any data port as is known in the art for

interfacing with an external accessory using a data protocol such as RS-232, Universal Serial Bus, or the like. Network interface **116** may be any interface as known in the art for communicating or transferring files across a computer network, examples of such networks include TCP/IP, ethernet, or token ring networks. In addition, on some systems, a network interface **116** may consist of a modem connected to the data input port **114**.

Thus, FIG. 1 illustrates the functional elements of a computer system **100**. Each of the elements of computer system **100** may be suitable off-the-shelf components as described above. The present embodiment provides a method and system for human interaction with the computer system **100** using speech.

As shown in FIG. 8, the computer system **100** may be connected to the Internet **700**, a collection of computer networks. To connect to the Internet **700**, computer system **100** may use a network interface **116**, a modem connected to the data input port **114**, or any other method known in the art. Web-sites **710**, other computers **720**, and storage media **108** may also be connected to the Internet through such methods known in the art.

Turning now to FIG. 2, an expanded functional block diagram of CPU **102** and storage medium **108** is illustrated. It is understood that the functional elements of FIG. 2 may be embodied entirely in hardware, firmware, or as software instructions encoded on a computer-readable medium. In the case of a software embodiment, the software may be found at storage medium **108** or at an alternate location across a computer network. CPU **102** includes speech recognition processor **200**, data processor **201**, natural language processor **202**, and application interface **220**. The data processor **201** interfaces with the display **104**, storage medium **108**, microphone **110**, speaker **112**, data input port **114**, and network interface **116**. The data

processor **201** allows the CPU to locate and read data from these sources. Natural language processor **202** further includes variable replacer **204**, string formatter **206**, word weighter **208**, boolean tester **210**, pronoun replacer **211**, and search engine **213**. Storage medium **108** includes a plurality of context-specific grammar files **212**, general grammar file **214**, dictation grammar **216**, context-specific dictation model **217**, natural language processor (NLP) database **218**, and a dialogue generation log (DGL) file **219**. In the preferred embodiment, the grammar files **212**, **214**, and **216** are Bakus-Naur Form (BNF) files, which describe the structure of the language spoken by the user. BNF files are well known in the art for describing the structure of language, and details of BNF files will therefore not be discussed herein. One advantage of BNF files is that hierarchical tree-like structures may be used to describe phrases or word sequences, without the need to explicitly recite all combinations of these word sequences. Thus, the use of BNF files in the preferred embodiment minimizes the physical sizes of the files **212**, **214**, and **216** in the storage medium **108**, increasing the speed at which these files can be enabled and searched as described below. However, in alternate embodiments, other file structures are used.

The context-specific dictation model **217** is an optional file that contains specific models to improve dictation accuracy. These models enable users to specify word orders and word models. The models accomplish this by describing words and their relationship to other words, thus determining word meaning by contextual interpretation in a specific field or topic. Take for example, the phrase "650 megahertz microprocessor computer." A context-specific dictation model **217** for computers may indicate the likelihood of the word "microprocessor" being associated with "computer," and that a number, such as "650" is likely to be found near the word "megahertz." By interpreting the context of the words, via a context-specific dictation model

217, a speech recognition processor would analyze the phrase, interpret a single object, i.e. the computer, and realize that "650 megahertz microprocessor" are adjectives or traits describing the type of computer.

Topics for context-specific dictation models **217** vary widely, and may include any topic area of interest to a user—both broad and narrow. Broad topics may include: history, law, medicine, science, technology, or computers. Specialized topics, such as a particular field of literature encountered at a book retailer's web-site are also possible. Such a context-specific dictation model **217** may contain text for author and title information, for example.

The context-specific dictation model **217** format relies upon the underlying speech recognition processor **200**, and is specific to each type of speech recognition processor **200**.

A dialogue generation log **219** is a file, database table, or any other log known in the art that contains a memory of the conversations between a computer or system practicing the present embodiment, and a user of such a system. The purpose behind the dialogue generation log **219** is to create a conversational memory by storing a log of what is spoken. The log is also used to avoid repetitive speech generation.

An exemplary dialogue generation log **219** is shown in FIG. 11. In this example, the dialogue generation log **219** is comprised of individual dialogue generation log entries **1112A-*x***, each entry representing a line of dialogue or statement exchanged between the system and a user. Although the format of each entry may vary from embodiment to embodiment, an example dialogue generation log entry would contain information such as a log entry identifier (ID) **1102**, a user identifier **1104**, a statement **1106**, the context of the statement **1108**, and a date/time stamp **1110**. The log entry ID **1102** uniquely identifies each entry **1112**. A user ID **1104** represents

each user of the system. Statements made by the system itself are also recorded in the log, and a special user ID **1104** is given to the system. In the example illustrated in FIG 11, the system is designated as user "-1," and other users are given positive numbers as their user ID **1104**. Other formats to represent users are certainly possible likely. The important part is that the system is uniquely identified with a user ID **1104** different from system users. Next, each entry contains a statement of what was spoken **1106**, and preferably the context **1108** in which the statement was spoken. For example, the statement "I would like to see movies," would be linked to the context of "movies" or "film." Lastly, a date-time-stamp **1110** may be recorded with each entry to identify the time frame (i.e. how recent) the statement was last made.

The operation and interaction of these functional elements of FIG. 2 will be described with reference to the flowchart of FIGS. 3A-3D. In FIG. 3A, the flow begins at block **300** with the providing of an utterance to speech processor **200**. An utterance is a series of sounds having a beginning and an end, and may include one or more spoken words. A microphone **110** may capture the spoken words in block **300**. Alternately, the utterance may be provided to the speech processor **200** over data input port **114**, or from storage medium **108**. Preferably, the utterance is in a digital format such as the well-known ".wav" audio file format.

The flow proceeds to decision **302** where the speech processor **200** determines whether one of the context-specific grammars **212** has been enabled. If the context-specific grammars **212** are enabled, the context-specific grammars **212** are searched at block **304**. In the preferred embodiment, the context-specific grammars **212** are BNF files that contain words and phrases which are related to a parent context. In general, a context is a subject area. For example, in one embodiment applicable to personal computers, examples of contexts may be "news", or

"weather", or "stocks". In such a case, the context-specific grammars **212** would each contain commands, control words, descriptors, qualifiers, or parameters that correspond to a different one of these contexts. The use of contexts provides a hierarchal structure for types of information. Contexts and their use will be described further below with reference to the NLP database **218**.

If a context-specific grammar **212** has been enabled, the context-specific grammar **212** is searched for a match to the utterance provided at block **300**. However, if a context-specific grammar **212** has not been enabled, the flow proceeds to block **308** where the general grammar **214** is enabled.

In the preferred embodiment, the general grammar **214** is a BNF file which contains words and phrases which do not, themselves, belong to a parent context, but may have an associated context for which a context-specific grammar file **212** exists. In other words, the words and phrases in the general grammar **214** may be at the root of the hierarchal context structure. For example, in one embodiment applicable to personal computers, the general grammar **214** would contain commands and control phrases.

In block **310**, the general grammar **214** is searched for a matching word or phrase for the utterance provided at block **300**. A decision is made, depending on whether the match is found, at block **312**. If a match is not found, then the dictation grammar **216** is enabled at block **314**. In the preferred embodiment, the dictation grammar **216** is a BNF file that contains a list of words that do not, themselves, have either a parent context or an associated context. For example, in one embodiment applicable to a personal computer, the dictation grammar **216** contains a relatively large list of general words similar to a general dictionary.

In block **316** the dictation grammar is searched for matching words for each word of the

utterance provided at block **300**. At decision block **318**, if no matching words are found, any relevant context-specific dictation model **217** is enabled at block **317**.

At decision block **319**, if no matching words are found, a visual error message is optionally displayed at the display **104** or an audible error message is optionally reproduced through speaker **112**, at block **320**. The process ends until another utterance is provided to the speech processor **200** at block **300**.

Thus, as can be seen from the above description, when an utterance is provided to the speech processor **200**, the enabled context-specific grammar **212**, if any, is first searched. If there are no matches in the enabled context-specific grammar **212**, then the general grammar **214** is enabled and searched. If there are no matches in the general grammar **214**, then the dictation grammar **316** is enabled and searched. Finally, if there are no matches in the dictation grammar **316**, a context-specific dictation model **217** is enabled **317** and used to interpret the utterance.

In the preferred embodiment, when the speech recognition processor **200** is searching either the context-specific grammar **212** or the general grammar **214**, it is said to be in the "command and control" mode. In this mode, the speech recognition processor **200** compares the entire utterance as a whole to the entries in the grammar. By contrast, when the speech recognition processor **200** is searching the dictation grammar, it is said to be in the "dictation" mode. In this mode, the speech recognition processor **200** compares the utterance to the entries in the dictation grammar **216** one word at a time. Finally, when the speech recognition processor **200** is matching the utterance with a context-specific dictation model **217**, it is said to be in "model matching" mode. It is expected that searching for a match for an entire utterance in the command and control mode will generally be faster than searching for one word at a time in

dictation or model matching modes.

It is further expected that any individual context-specific grammar **212** will be smaller in size (i.e., fewer total words and phrases) than the general grammar **214**, which in turn will be smaller in size than the dictation grammar **216**. Thus, by searching any enabled context-specific grammar **212** first, it is likely that a match, if any, will be found more quickly, due at least in part to the smaller file size. Likewise, by searching the general grammar **214** before the dictation grammar **216**, it is likely that a match, if any, will be found more quickly.

Additionally, as will be explained further below with regard to the ability of the present embodiment to adaptively add to both the context-specific grammar **212** and the general grammar **214**, they will contain the most common utterances. As such, it is expected that a match is more likely to be found quickly in the context-specific grammar **212** or the general grammar **214** than in the dictation grammar **216**.

Finally, as will be explained further below, the words and phrases in the enabled context-specific grammar **212** are more likely to be uttered by the user because they are words that are highly relevant to the subject matter about which the user was most recently speaking. This also allows the user to speak in a more conversational style, using sentence fragments, with the meaning of his words being interpreted according to the enabled context-specific grammar **212**.

By searching in the above-described sequence, the present embodiment may search more efficiently than if the searching were to occur one entry at a time in a single, large list of all expected words and phrases.

Referring back to decision **306**, if a match is found in the context-specific grammar **212**, then there are three possible next alternatives shown in FIG. 3A. For each matching entry in the

enabled context-specific grammar **212**, there may be an associated action to be taken by the speech recognition processor **200**. Block **322** shows that one action may be to direct application interface **220** to take some action with respect to a separate software application or entity. For example, application interface **220** may use the Speech Application Programming Interface (SAPI) standard by Microsoft to communicate with an external application. The external application may be directed, for example, to access a particular Internet web site URL or to speak a particular phrase by converting text to speech. Other actions may be taken as will be discussed further below with reference to the NLP database **218** of FIG. 4.

Block **324** shows that another action may be to access a row in the natural language processing (NLP) database **218** directly, thereby bypassing the natural language processing method described further below. Block **326** shows that another action may be to prepend a word or phrase for the enabled context to the matching word or phrase found in the context-specific grammar **306**. For example, if the enabled context were "movies" and the matching utterance were "8 o'clock," the word "movies" would be prepended to the phrase "8 o'clock" to form the phrase "movies at 8 o'clock."

Likewise, if a match is found in the general grammar **214**, then the flow may proceed to block **322** where the application interface **220** is directed to take an action as described above, or to block **324** where a row in the NLP database is directly accessed. However, if a match is found in the general grammar **214**, no prepending of a context occurs because, as stated above, the entries in the general grammar **214** do not, themselves, have a parent context.

Alternatively, manually entered words may be captured, at block **301**, and input into the natural language processor.

Finally, with reference to FIG. 3A, words may be entered manually via manual input **106**. In this case, no speech recognition is required, and yet natural language processing of the entered words is still desired. Thus, the flow proceeds to FIG. 3B.

In FIG. 3B, at block **328**, the natural language processor **202** formats the phrase for natural language processing analysis. This formatting is accomplished by string formatter **206** and may include such text processing as removing duplicate spaces between words, making all letters lower case (or upper case), expanding contractions (e.g., changing "it's" to "it is"), and the like. The formatting prepares the phrase for parsing.

The flow continues to block **330** where certain "word-variables" are replaced with an associated wildcard function by variable replacer **204** in preparation for accessing the NLP database **218**. As used herein, the term "word-variables" refers to words or phrases that represent amounts, dates, times, currencies, and the like. For example, in one embodiment the phrase "what movies are playing at 8 o'clock" would be transformed at block **330** to "what movies are playing at $time" where "$time" is a wildcard function used to represent any time value. As another example, in one embodiment the phrase "sell IBM stock at 100 dollars" would be transformed at block **330** to "sell IBM stock at $dollars" where "$dollars" is a wildcard function used to represent any dollar value. This block may be accomplished by a simple loop that searches the phrase for key tokens such as the words "dollar" or "o'clock" and replaces the word-variables with a specified wildcard function. In order to keep track of the location in the phrase where the substitution was made, an array may be used. This allows re-substitution of the original word-variable back into the phrase at the some position after the NLP database **218** has been searched.

The purpose of replacing word-variables with an associated wildcard function at block **330** is to reduce the number of entries that must be present in the NLP database **218**. For example, the NLP database **218** would only contain the phrase "what movies are playing at $time" rather than a separate entry for 8 o'clock, 9 o'clock, 10 o'clock, and so on. The NLP database **218** will be described further below.

At block **332**, pronouns in the phrase are replaced with proper names by pronoun replacer **211**. For example, in one embodiment the pronouns "I," "my," or "mine" would be replaced with the speaker's name. Block **332** allows user-specific facts to be stored and accessed in the NLP database **218**. For example, the sentence "who are my children" would be transformed into "who are Dean's children" where "Dean" is the speaker's proper name. Again, this block may be performed in a simple loop that searches the phrase for pronouns, and replaces the pronouns found with an appropriate proper name. In order to keep track of the locations in the phrase where a substitution was made, an array may be used.

In block **334**, the individual words in the phrase are weighted according to their relative "importance" or "significance" to the overall meaning of the phrase by word weighter **208**. For example, in one embodiment there are three weighting factors assigned. The lowest weighting factor is assigned to words such as "a," "an," "the," and other articles. The highest weighting factor is given to words that are likely to have a significant relation to the meaning of the phrase. For example, these may include all verbs, nouns, adjectives, and proper names in the NLP database **218**. A medium weighting factor is given to all other words in the phrase. The purpose of this weighting is to allow for more powerful searching of the NLP database **218**.

An example of selected columns of the NLP database **218** of one embodiment is shown in

FIG. 4. The NLP database **218** comprises a plurality of columns **400-410**, and a plurality of rows **412A-412N**. In column **400**, the entries represent phrases that are "known" to the NLP database. In column **402**, a number of required words for each entry in column **400** is shown. In column **404**, an associated context or subcontext for each entry in column **400** is shown. In columns **408** and **410**, one or more associated actions are shown for each entry in column **400**. It should be noted that the NLP database **218** shown in FIG. 4 is merely a simplified example for the purpose of teaching the present embodiment. Other embodiments may have more or fewer columns with different entries.

Referring back to FIG. 3B, at block **336**, the NLP database **218** is searched for possible matches to the phrase, based on whether the entry in column **400** of the NLP database **218** contains any of the words in the phrase (or their synonyms), and the relative weights of those words. At block **338**, a confidence value is generated for each of the possible matching entries based on the number of occurrences of each of the words in the phrase and their relative weights. Weighted word searching of a database is well known in the art and may be performed by commercially available search engines such as the product "dtsearch" by DT Software, Inc. of Arlington, Virginia. Likewise, searching using synonyms is well known in the art and may be accomplished using such publicly available tools such as "WordNet," developed by the Cognitive Science Laboratory of Princeton University in Princeton, New Jersey. The search engine may be an integral part of the natural language processor **202**.

At decision **340**, the natural language processor **202** determines whether any of the possible matching entries has a confidence value greater than or equal to some predetermined minimum threshold, T. The threshold T represents the lowest acceptable confidence value for

which a decision can be made as to whether the phrase matched any of the entries in the NLP database **218**. If there is no possible matching entry with a confidence value greater than or equal to T, then the flow proceeds to block **342** where an optional error message is either visually displayed to the user over display **104** or audibly reproduced over speaker **112**. In one embodiment, the type of error message, if any, displayed to the user may depend on how many "hits" (i.e., how many matching words from the phrase) were found in the highest-confidence NLP database entry. A different type of error message would be generated if there was zero or one hits, than if there were two or more hits.

If, however, there is at least one entry in the NLP database **218** for which a confidence value greater than or equal to T exists, then the flow proceeds to block **344** where the "noise" words are discarded from the phrase. The "noise" words include words that do not contribute significantly to the overall meaning of the phrase relative to the other words in the phrase. These may include articles, pronouns, conjunctions, and words of a similar nature. "Non-noise" words would include words that contribute significantly to the overall meaning of the phrase. "Non-noise" words would include verbs, nouns, adjectives, proper names, and words of a similar nature.

The flow proceeds to FIG. 3C where the non-noise word requirement is retrieved from column **402** of the NLP database **218** for the highest-confidence matching entry at block **346**. For example, if the highest-confidence matching phrase was the entry in row **412A**, (e.g., "what movies are playing at $time"), then the number of required non-noise words is 3.

At decision **348**, a test is made to determine whether the number of required non-noise words from the phrase is actually present in the highest-confidence entry retrieved from the NLP

database **218**. This test is a verification of the accuracy of the relevance-style search performed at block **336**, it being understood that an entry may generate a confidence value higher than the minimum threshold, T, without being an acceptable match for the phrase.

The nature of the test performed at decision **348** is a boolean "AND" test performed by boolean tester **210**. The test determines whether each one of the non-noise words in the phrase (or its synonym) is actually present in the highest-confidence entry. If there are a sufficient number of required words actually present in the highest-confidence entry, then the flow proceeds to block **350**, where the natural language processor **202** directs application interface **220** to take an associated action from column **408** or **410**. It is understood that additional action columns may also be present.

For example, if the highest confidence entry was the entry in row **412A**, and the boolean test of decision **348** determined that there actually were 3 non-noise words from the phrase in the entry in column **400**, then the associated action in column **408** (e.g., access movie web site) would be taken. Other entries in the NLP database have other associated actions. For example, if the highest-confidence entry is that in row **412E** (e.g., "what time is it"), the associated action may be for natural language processor **202** to direct a text-to-speech application (not shown) to speak the present time to the user through the speaker **112**. As another example, if the highest-confidence entry is that in row **412F** (e.g., "show me the news"), the first associated action may be to access a predetermined news web site on the Internet, and a second associated action may be to direct an image display application (not shown) to display images associated with the news. Different or additional actions may also be performed.

Also, if the highest-confidence entry contains the required number of non-noise words

from the phrase as determined at decision **348**, the natural language processor **202** instructs the speech recognition processor **200** to enable the context-specific grammar **212** for the associated context of column **404**. Thus, for row **412A**, context-specific grammar **212** for the context "movies" would be enabled. Thus, when the next utterance is provided to the speech recognition processor **200** in block **300** of FIG. 3A, it would search the enabled context-specific grammar **212** for "movies" before searching the general grammar **214**. As previously stated, enabling the appropriate context-specific grammar **212** greatly increases the likelihood of fast, successful speech recognition, and enhances the user's ability to communicate with the computer in a conversational style.

If, however, back at decision **348**, the required number of non-noise words from the phrase is not actually present in the highest-confidence entry retrieved from the NLP database **218**, then the flow proceeds to block **354** where the user is prompted over display **104** or speaker **112** whether the highest-confidence entry was meant, and record the inquiry in the dialogue generation log **219**. For example, if the user uttered "How much is IBM stock selling for today," the highest-confidence entry in the NLP database **218** may be the entry in row **412B**. In this case, although the relevance factor may be high, the number of required words (or their synonyms) may not be sufficient. Thus, the user would be prompted at block **354** whether he meant "what is the price of IBM stock on August 28, 1998."

The user may respond either affirmatively or negatively. If it is determined at decision **356** that the user has responded affirmatively, then the action associated with the highest-confidence entry is taken at block **350**, and the associated context-specific grammar **212** enabled at block **352**. The speech made by the user is also recorded as an entry of the dialogue generation

log **219**.

If, however, it is determined at decision **356** that the user has responded negatively, then the flow proceeds to FIG. 3D where the associated context from column **404** of NLP database **218** is retrieved for the highest-confidence entry, and the user is prompted for information using a context-based interactive dialog at block **360**. For example, if the user uttered "what is the price of XICOR stock today," and the highest confidence entry from the NLP database **218** was row **412B** (e.g., "what is the price of IBM stock on $date"), then the user would be prompted at block **354** whether that was what he meant.

If the user responds negatively, then the context "stock" is retrieved from column **404** at block **358**, and the context-based interactive dialog for the stock context is presented to the user over the display **104** and speaker **112**. Such a context-based interactive dialog may entail prompting the user for the name and stock ticker symbol of XICOR stock. The user may respond by speaking the required information. A different context-based interactive dialog may be used for each of the possible contexts. For example, the "weather" context-based interactive dialog may entail prompting the user for the name of the location (e.g., the city) about which weather information is desired. Also, the "news" context-based interactive dialog may entail prompting the user for types of articles, news source, Internet URL for the news site, or other related information.

Upon completion of the context-based interactive dialog, the NLP database **218**, general grammar **214**, and context-specific grammar **212** are updated to include the new information, at block **362**. In this way, the next time the user asks for that information, a proper match will be found, and the appropriate action taken without prompting the user for more information. Thus,

the present embodiment adaptively "learns" to recognize phrases uttered by the user.

In one embodiment, one or more of the NLP database **218**, context-specific grammar **212**, general grammar **214**, and dictation grammar **216** also contain time-stamp values (not shown) associated with each entry. Each time a matching entry is used, the time-stamp value associated with that entry is updated. At periodic intervals, or when initiated by the user, the entries that have a time-stamp value before a certain date and time are removed from their respective databases/grammars. In this way, the databases/grammars may be kept to an efficient size by "purging" old or out-of-date entries. This also assists in avoiding false matches.

In an alternate embodiment, the updates to the NLP database **218**, general grammar **214**, and context-specific grammar **212** are stored in a user global user registry **800**, shown in FIG. 9. A global user registry entry **800a** would be comprised of any general grammar additions **214a**, context-specific grammar additions **212a**, and NLP database additions **218a** created by the user training. Since each user of the system would have a different global user registry entry **800a**, the embodiment would be flexible enough to allow for special customizations and could adapt to the idiosyncrasies of individual users.

In yet another embodiment, the updates to global user registry entry **800** also contain dialogue generation log additions **219a**, user preferences **802**, and remembered information **804**, as shown in FIG. 9. The dialogue generation log additions **219a** would be comprised of the dialogue exchanged between the system and an individual user. The addition of dialogue generation log additions **219a** allow the system greater intelligence in its conversations with an individual user, to be transparent and mobile across different computers. User preferences **802** represents the stored system preferences of an individual user. Examples of such user preferences

802, include, but are not limited to: window color and placement, system sounds, font preferences, and other user-interface preferences. Remembered information 804 includes. user schedule information, learned facts derived from the dialogue generation log 219, and any other user (but non-user-interface) stored data not included in user preferences. For example, if the system is told, "I own a dog," the remembered information will include that fact. Remembered information may be stored or represented by any information storage format known in the art, including any database format.

Moreover, in some embodiments, the global user registry 800 would be stored locally and mirrored at known server locations. The mirrored copy, referred to as the "travelling" global user registry entry, enables users to access their phrases "adaptively" learned by the embodiment, even when the user is logged into a different location. FIG. 10A illustrates an embodiment that accesses customized global user registry entries 800a at local and remote (travelling) locations. Initially, a valid system user is verified, by any means known in the art, and then the system searches for a locally stored global user registry entry. For example, the system queries users for their login ID and password as shown in block 900. If the password and login ID match, as determined by decision block 905, the user is deemed to be a valid user. It is well understood that this login ID and password are but one of many methods known in the art to verify valid users, and that all such validation systems could be easily substituted. If no local global user registry entry is found, block 910, the system searches for a travelling global user registry entry, block 920. If either search turns up a global user registry entry, the global user registry entry is loaded, blocks 915 and 925, respectively. Provided that the retrieval of the global user registry entry 800a is successful, blocks 930 and 935, the global user registry entry 800a is enabled at block

940 by extracting the general grammar additions **214a**, context-specific grammar additions **212a**, NLP database additions **218a**, dialogue generation log additions **219a**, user preferences **802**, and remembered information **804**. These "learned" adaptations are then used by the system, as discussed earlier with the method of FIGS 3A-3D. If the retrieval of the travelling global user registry entry **800a** is unsuccessful, standard (non-custom user) processing is performed as reflected at block **945**.

Alternatively, in some embodiments, the user only logs into the system when accessing a terminal that does not recognize the user as a local user. This embodiment is shown in FIG. 10B. From this remote system, the user is required to specify an ID and a password. Initially, the system determines whether the user has a locally stored voice profile, block **910**. If the user has a local voice profile, it is loaded at block **915**, and flow continues at decision block **950**. If the user does not have a locally stored global user registry entry, the system queries the user for a login identifier and a password at block **900**. If the user is valid, the travelling global user registry entry is retrieved at block **925**, and flow continues at block **950**. Otherwise, the user is treated as a user with standard (non-custom) processing, at block **945**. Provided that the retrieval of the global user registry entry **800a** is successful, block **950**, the global user registry entry **800a** is enabled at block **940** by extracting the general grammar additions **214a**, context-specific grammar additions **212a**, NLP database additions **218a**, dialogue generation log additions **219a**, user preferences **802**, and remembered information **804**. These "learned" adaptations are then used by the system, as discussed earlier with the method of FIGS 3A-3D.

In one embodiment, speech recognition and natural language processing may be used to interact with objects, such as help files (".hlp" files), World-Wide-Web ("WWW" or "web")

pages, or any other objects that have a context-sensitive voice-based interface.

FIG. 5 illustrates an exemplary Dialog Definition File (DDF) **500** which represents information necessary to associate the speech recognition and natural language processing to an internet object, such as a text or graphics file or, in the preferred embodiment, a web-page or help file. Although in its simplest embodiment the Dialog Definition File **500** consists of an object table **510**, the DDF may also contain additional context-specific grammar files **214** and additional entries for the natural language processing (NLP) database **218**, as illustrated in FIG. 5. The preferred embodiment of the DDF **500** includes an object table **510**, a context-specific grammar file **214**, a context-specific dictation model **217**, and a file containing entries to the natural language processing database **218**. These components may be compressed and combined into the DDF file **500** by any method known in the art, such as through Lempel-Ziv compression. The context-specific specific grammar file **214** and the natural language processing database **218** are as described in earlier sections. The object table **510** is a memory structure, such as a memory tree, chain or table, which associates an address of a resource with various actions, grammars, or entries in the NLP database **218**.

An exemplary embodiment of the object table **510** is illustrated in FIG. 6. FIG. 6 illustrates a memory table which may contain entry columns for: an object **520**, a Text-to-Speech (TTS) flag **522**, a text speech **524**, a use grammar flag **526**, an append grammar flag **528**, an "is yes/no?" flag, and "do yes" **532** and "do no" **534** actions. Each row in the table **540A-540*n*** would represent the grammar and speech related to an individual object. The exemplary embodiment would refer to objects **520** through a Universal Resource Locator (URL). A URL is a standard method of specifying the address of any resource on the Internet that is part of the

World-Wide-Web. As this standard is well known in the art for describing the location of Internet resources and objects, the details of URLs will therefore not be discussed herein. One advantage of URLs is that they can specify information in a large variety of object formats, including hypertext, graphical, database and other files, in addition to a number of object devices and communication protocols. However, as shown in FIG. 6, URLs and other method of specifying objects can be used.

When combined with the text speech **524**, the Text-to-Speech (TTS) flag **522** indicates whether an initial statement should be voiced over speaker **112** when the corresponding object is transferred. For example, when transferring the web page listed in the object column **520** of row **540A** (`http://www.our-pet-store.com`), the TTS flag **522** is marked, indicating the text speech **524**, "Hello, welcome to...," is to be voiced over speaker **112**.

In some embodiments, the utterance of the initial statement may be dependent on the dialogue generation log **219**. In these embodiments, the system first checks to see if the statement was made recently to a user, before re-voicing the statement. There are several advantages in doing this. By consulting the dialogue generation log **219**, a conversation can be made more intelligent by using the dialogue generation log **219** as a "conversational memory." Moreover, a conversational memory formed by the dialogue generation log **219** also acts as a "sanity check" to make sure conversations do not become repetitive.

The use of the dialogue generation log **219** is best illustrated by example. Suppose a user was examining a speech-enhanced pet-store web-site using the present embodiment. Initially, an object table **510** is transferred from the web-site to the user's computer system **100**. When transferring the web page listed in the object column **520**, the system notes that the TTS flag **522**

is marked, indicating the text speech **524**, "Hello, welcome to our pet store, would you like to see our dog specials?" Since this is the user's first visit to the web page, no context or statement-related entries are found in the dialogue generation log **219**, so the statement is to be voiced over speaker **112**. The user responds, "no," and decides to surf a different web page instead. Transparent to the user, the system records both statements in the dialogue generation log **219**. Moments later, the user presses the web-browser "back" key, returning to the original pet-store web page. At this time, the user returns to the web page, and again, the system notes that the TTS flag **522** is marked, indicating the text speech **524**, "Hello, welcome to our pet store, would you like to see our dog specials?" Since this is the user's subsequent visit to the web page, a statement-related entry is found in the dialogue generation log **219**. Since statement was recently voiced over the speaker, it is not voiced on again on a subsequent visit.

The next three object table **510** flags relate to the use of grammars associated with this object. The affirmative marking of the "use grammar" **526** or "append grammar" **528** flags indicate the presence of a content-specific grammar file **214** related to the indicated object. The marking of the "use grammar" flag **526** indicates that the new content-specific grammar file **214** replaces the existing content-specific grammar file, and the existing file is disabled. The "append grammar" flag **528** indicates that the new content-specific grammar file should be enabled concurrently with the existing content-specific grammar file.

Lastly, the remaining column entries relate to a "yes/no" grammar structure. If the "Is yes/no?" flag **530** is marked, then a standard "yes/no" grammar is enabled. When a standard "yes/no" grammar is enabled, affirmative commands spoken to the computer result in the computer executing the command indicated in the "Do Yes" entry **532**. Similarly, a negative

command spoken to the computer results in the computer executing the command indicated in the "Do No" entry **534**. The entries in the "Do Yes" **532** and "Do No" **534** columns may either be commands or pointers to commands imbedded in the NLP Database **218**. For example, as shown in row **540B**, the "Is Yes/No?" flag is marked. An affirmative answer, such as "yes," given to the computer, would result in executing the corresponding command in the "Do Yes" entry **532**; in this specific case, the entry is the number "210," a reference to the 210th command in the NLP database. An answer of "no" would result in the computer executing the 211th command in the NLP database.

Turning now to FIG. 7A, a method and system of providing speech and voice commands to objects, such as a computer reading a help file or browsing the World-Wide-Web, is illustrated. The method of FIGS. 7A-7D may be used in conjunction with the method of FIGS 3A-3D and FIG. 10. In block **602**, an object location is provided to a help file reader or World-Wide-Web browser. A help file reader/browser is a program used to examine hypertext documents that are written to help users accomplish tasks or solve problems, and is well known in the art. The web browser is a program used to navigate through the Internet, and is well known in the art. Block **602**, which providing an object location to the browser, can be as simple as a user clicking on a program "help" menu item, manually typing in a URL, or having a user select a "link" at a chosen web-site. It also may be the result of a voiced command as described earlier with reference to the action associated with each entry in the NLP database **218**. Given the object location, the computer must decide on whether it can resolve object location specified, at block **604**. This resolution process is a process well known in the art. If the computer is unable to resolve the object location or internet address, an error message is displayed in the

browser window, at block **605**, and the system is returned to its initial starting state **600**. If the object location or Internet address is resolved, the computer retrieves the object at block **606**. For a networked object, for example, a web browser sends the web-site a request to for the web page, at block **606**. For a help file application, the help reader reads the help file off of storage media **108**, at block **606**.

A decision is made, depending upon whether the object is retrieved, at block **608**. If the object cannot be retrieved, an error message is displayed in the browser window, at block **605**, and the system is returned to its initial starting state **600**. If the object is retrieved, it is displayed in the help-reader or web-site browser, as appropriate, at block **610**.

In decision block **612**, the computer **100** determines whether the DDF file **500** corresponding to the object is already present on the computer **100**. If the DDF file is present, the flow proceeds to FIG. 7C, if not the flow proceeds to FIG. 7B.

Moving to FIG. 7B, if the DDF file **500** is not present, the computer examines whether the DDF file **500** location is encoded within the object. For example, the DDF file location could be encoded within web page HyperText Markup Language (HTML) as a URL. (Note that HTML is well known in the art, and the details of the language will therefore not be discussed herein.) Encoding DDF file location within HTML code may be done either through listing the DDF file location in an initial HTML meta-tag such as:

```
<meta DDF= "http://www.our-pet-store.com/ConverseIt.ddf">
```

or directly through a scripting tag written into the variation of HTML supported by the browser,

```
<!--
      <DDF= "http://www.our-pet-store.com/ConverseIt.ddf">
-->
```

If the DDF file location information is encoded within the web page, the location's Internet address is resolved, at block **616**, and the computer requests transfer of the DDF file **500**, at block **626**. An equivalent encoding scheme could be used within help file hypertext.

Alternatively, if the DDF file **500** location is not encoded within the object, there are several alternate places that it may be stored. It may be stored in a pre-defined location at a web-site, such as a certain file location in the root directory, or at a different centralized location, such as another Internet server or the storage medium **108** of FIG. 1. Blocks **618** and **620** test for these possibilities. Block **618** determines whether the DDF file is located at the web-site. At block **618**, the computer sends query to the web-site inquiring about the presence of the DDF file **500**. If the DDF file **500** is present at the web-site, the computer requests transfer of the DDF file **500**, at block **626**. If the DDF file **500** is not located at the web-site, the computer queries the centralized location about the presence of a DDF file for the web-site, at block **620**. If the DDF file is present at the web-site, the computer requests transfer of the DDF file, at block **626**. If the DDF file **500** cannot be found, the existing components of any present DDF file, such as the object table **510**, context-specific dictation model **217**, NLP database **218** associated with the object, and context-specific grammar **214** for any previously-viewed object, are deactivated in block **622**. Furthermore, the object is treated as a non-voice-activated object, and only standard grammar files are used, at block **624**. Standard grammar files are the grammar files existing on the system excluding any grammars associated with the content-specific grammar file associated with the object.

If the DDF file **500** is requested at block **626**, and its transfer is unsuccessful, any existing components of any present DDF file **500** are deactivated, at block **622**, and the web-site is treated

as a non-voice-activated object, and only standard grammar files are used, at block **624**.

If the DDF file **500** is requested at block **626** and its transfer is successful at block **628**, it replaces any prior DDF file, at block **630**. Any components of the DDF file **500**, such as the object table **510**, context-specific-grammar files **214**, context-specific-dictation models **217**, and NLP database **218** are extracted at block **632**. A similar technique may be used for obtaining the software necessary to implement the method illustrated in FIGS. 3A-3D, comprising the functional elements of FIG. 2.

The flow moves to FIG. 7C. The object table **510** is read into memory by the computer in block **634**. If the object is present in the site object table **510**, as determined by block **636**, it will be represented by a row **540A-540*n*** of the table, as shown in FIG. 6. Each row of the object table represents the speech-interactions available to a user for that particular object. If no row corresponding to the object exists, then no-speech interaction exists for the web page, and processing ends.

If the object location is present in the site object table **510**, as determined by block **636**, the computer checks if the TTS flag **522** is marked, to determine whether a text speech **524** is associated with the web-page, at block **638**. If there is a text speech **524**, a decision must be made on whether it is appropriate to play the spoken statement **639**. For example, if a user was queried, "Would you like to hear about the history of Mars?" in the past five minutes, it would be inappropriate, and probably annoying, to ask the user the same question again. In making such a determination, the system checks whether the statement has been made to the user recently by examining the dialogue generation log **219**.

An individual user may be matched with a statement or statement context within a recent

interval. Suppose the system asked, "Do you own a cat?" and the user replied, "No, I hate cats." The conversation would be recorded in the dialogue generation log **219**, noting that the context of the conversation is about cats. On a subsequent visit to a pet-store web-site, the system notes that the TTS flag **522** is marked, indicating the text speech **524**, "Hello, welcome to our pet store, would you like to see our cat specials?" Since the text-to-speech statement is an inquiry about cats, the system performs a check in the dialogue generation log 219 to see if there are past conversations with a context **1108** of "cat." The context of "cat" is located, and the system notes that the user responded "no," therefore the user does not own a cat. Consequently, the system realizes that a negative match has been made, and it would be inappropriate to voice the question to the user.

Thus, since dialogue generation log entries **1112** are time dependent, they must be kept up-to-date by the system, and old entries should be culled and removed on a periodic basis. After all, in the case of the above example, it is possible that the user eventually acquires a cat. Old entries may be removed after a certain time-period expires, or after the dialogue generation log **219** reaches a specific size. Dialogue generation log entries **1112** for each user can be removed on a first-in-first-out (FIFO) basis, so that older entries are removed before more recent entries.

If it is appropriate, the statement is voiced at block **640**, and flow continues on to FIG 7D. If there is a context-specific grammar file associated with the object, as determined by decision block **642**, it is enabled at block **644**, and then the NLP database **218** is enabled at block **646**. If no context-specific grammar file is associated with the object, only the NLP database **218** is enabled at block **646**. Once the NLP database is enabled **646**, the system behaves as shown in FIG. 3A-3C, as described above.

In summary, the present embodiment provides a method and system for an object interactive user-interface for a computer. By the use of context-specific grammars that are tied to internet-objects through a Dialog Definition File, the present embodiment decreases speech recognition time and increases the user's ability to communicate with local and networked objects, such as help files or web-pages, in a conversational style. Adaptive updating of the various grammars and the NLP database, the present embodiment further increases interactive efficiency. The adaptive updates can be incorporated into global user registry entries that can be stored locally and remotely, to allow users access to the global user registry entries at various locations.

The previous description of the embodiments is provided to enable any person skilled in the art to make or use the present invention. The various modifications to these embodiments will be readily apparent to those skilled in the art, and the generic principles defined herein may be applied to other embodiments without the use of inventive faculty. Thus, the present invention is not intended to be limited to the embodiments shown herein, but is to be accorded the widest scope consistent with the principles and novel features disclosed herein.

I CLAIM:

CLAIMS

1. A method for interacting with an object via a computer using utterances, the method comprising:

retrieving an object table associated with the object;

searching the object table for a matching entry for the object;

locating an action associated with the matching entry, if the matching entry is found in the object table;

determining whether the action associated with the matching entry is appropriate;

performing the action associated with the matching entry, if the action is deemed appropriate.

2. The method of claim 1 wherein the action associated with the matching entry is voicing a text statement.

3. The method of claim 2 wherein the process of determining whether the action associated with the matching entry is appropriate further comprises:

identifying a context of the voiced text statement associated with the matching entry;

searching a log containing log entries with context and timestamp information to locate matching log entries with the same context as the voiced text statement;

determining whether the voiced text statement is appropriate based on the timestamp of the matching log entries.

4. The method of claim 2 wherein the process of determining whether the action associated with the matching entry is appropriate further comprises:

searching a log containing log entries with statement and timestamp information to locate matching log entries with the same voiced text statement;

determining whether the voiced text statement is appropriate based on the timestamp of the matching log entries.

5. The method of claim 3 or 4 wherein the object table is read from a web-site.

6. The method of claim 3 or 4 wherein the object table is read from a location that stores object tables for multiple objects.

7. A method for interacting with an object via a computer using utterances, the method comprising:

searching a context-specific grammar file for a matching phrase for the utterance;

searching a general grammar file for the matching phrase if the matching phrase is not found in the context-specific grammar file;

searching a dictation grammar for the matching phrase if the matching phrase is not found in the general grammar file;

searching a context-specific dictation model for the matching phrase if the matching phrase is not found in the dictation grammar file;

searching a database for a matching entry for the matching phrase; and

performing an action associated with the matching entry and saving the utterances to a log file if the matching entry is found in the database.

8. The method of claim 7 the log is a dialogue generation log.

9. The method of claim 8 further comprising:

replacing at least one word in the matching phrase prior to the searching the database.

10. The method of claim 9 wherein the replacing the at least one word comprises substituting a wildcard for the at least one word.

11. The method of claim 10 wherein the replacing at least one word comprises substituting a proper name for the at least one word.

12. The method of claim 11 further comprising:

text formatting the matching phrase prior to the searching of the database.

13. The method of claim 12 further comprising:

weighting individual words in the matching phrase according to a relative significance of the individual words prior to the searching of the database.

14. The method of claim 13 further comprising:

updating a global user registry entry with the dialogue generation log with utterances spoken by a user.

15. The method of claim 13 further comprising:

updating the global user registry entry with at least one of the database, the context-specific grammar file and the general grammar file with the matching phrase if the matching entry is not found in the database.

16. The method of claim 15 further comprising:

generating a confidence values for the matching entry;

comparing the confidence value with a threshold value;

determining whether a required number of words from the matching phrase are present in the matching entry;

determining the utterance has been spoken before by examining the dialogue generation

log; and

prompting a user whether the matching entry is a correct interpretation of the utterance if the required number of words from the matching phrase are not present in the matching entry, nor present in the dialogue generation log.

17. The method of claim 16 further comprising:

prompting a user for additional information if the matching entry is not a correct interpretation of the utterance.

18. The method of claim 17 further comprising:

updating at least one of the database, the context-specific grammar file, the general grammar file, and the dialogue generation log with the additional information.

19. A system for interacting with a computer using utterances, the system comprising:

a speech processor for searching a context-specific grammar file for a matching phrase for the utterance, for searching a general grammar file for the matching phrase if the matching phrase is not found in the context-specific grammar file, for searching a dictation grammar for the matching phrase if the matching phrase is not found in the general grammar file, and for searching a context-specific dictation model if the matching phrase is not found in the dictation grammar;

a natural language processor for searching a database for a matching entry for the matching phrase and for searching a dialogue generation log for a matching log entry for the utterance; and

an application interface for performing an action associated with the matching entry if the matching entry is found in the database and the action is appropriate if the matching log entry is found in the dialogue generation log.

20. A computer readable medium encoded with executable program code, the code comprising:

retrieving an object table associated with the object;

searching the object table for a matching entry for the object;

locating an action associated with the matching entry, if the matching entry is found in the object table;

determining whether the action associated with the matching entry is appropriate;

performing the action associated with the matching entry, if the action is deemed appropriate.

21. The computer readable medium of claim 20 wherein the action associated with the matching entry is voicing a text statement.

22. The computer readable medium of claim 21 wherein the process of determining whether the action associated with the matching entry is appropriate further comprises:

identifying a context of the voiced text statement associated with the matching entry;

searching a log containing log entries with context and timestamp information to locate matching log entries with the same context as the voiced text statement;

determining whether the voiced text statement is appropriate based on the timestamp of the matching log entries.

23. The computer readable medium of claim 21 wherein the process of determining whether the action associated with the matching entry is appropriate further comprises:

searching a log containing log entries with statement and timestamp information to locate matching log entries with the same voiced text statement;

determining whether the voiced text statement is appropriate based on the timestamp of the matching log entries.

24. The computer readable medium of claim 22 or 23 wherein the object table is read from a web-site.

25. The computer readable medium of claim 22 or 23 wherein the object table is read from a location that stores object tables for multiple objects.

26. A computer readable medium encoded with executable program code, the code comprising:

searching a context-specific grammar file for a matching phrase for the utterance;

searching a general grammar file for the matching phrase if the matching phrase is not found in the context-specific grammar file;

searching a dictation grammar for the matching phrase if the matching phrase is not found in the general grammar file;

searching a context-specific dictation model for the matching phrase if the matching phrase is not found in the dictation grammar file;

searching a database for a matching entry for the matching phrase; and

performing an action associated with the matching entry and saving the utterances to a log file if the matching entry is found in the database.

27. The computer readable medium of claim 26 wherein the log is a dialogue generation log.

28. The computer readable medium of claim 27 further comprising:

replacing at least one word in the matching phrase prior to the searching the database.

29. The computer readable medium of claim 28 wherein the replacing the at least one word comprises substituting a wildcard for the at least one word.

30. The computer readable medium of claim 29 wherein the replacing at least one word

comprises substituting a proper name for the at least one word.

31. The computer readable medium of claim 30 further comprising:

text formatting the matching phrase prior to the searching of the database.

32. The computer readable medium of claim 31 further comprising:

weighting individual words in the matching phrase according to a relative significance of the individual words prior to the searching of the database.

33. The computer readable medium of claim 32 further comprising:

updating a global user registry entry with the dialogue generation log with utterances spoken by a user.

34. The computer readable medium of claim 32 further comprising:

updating the global user registry entry with at least one of the database, the context-specific grammar file and the general grammar file with the matching phrase if the matching entry is not found in the database.

35. The computer readable medium of claim 34 further comprising:

generating a confidence values for the matching entry;

comparing the confidence value with a threshold value;

determining whether a required number of words from the matching phrase are present in the matching entry;

determining the utterance has been spoken before by examining the dialogue generation log; and

prompting a user whether the matching entry is a correct interpretation of the utterance if the required number of words from the matching phrase are not present in the matching entry, nor

present in the dialogue generation log.

36. The computer readable medium of claim 35 further comprising:

prompting a user for additional information if the matching entry is not a correct interpretation of the utterance.

37. The computer readable medium of claim 36 further comprising:

updating at least one of the database, the context-specific grammar file, the general grammar file, and the dialogue generation log with the additional information.

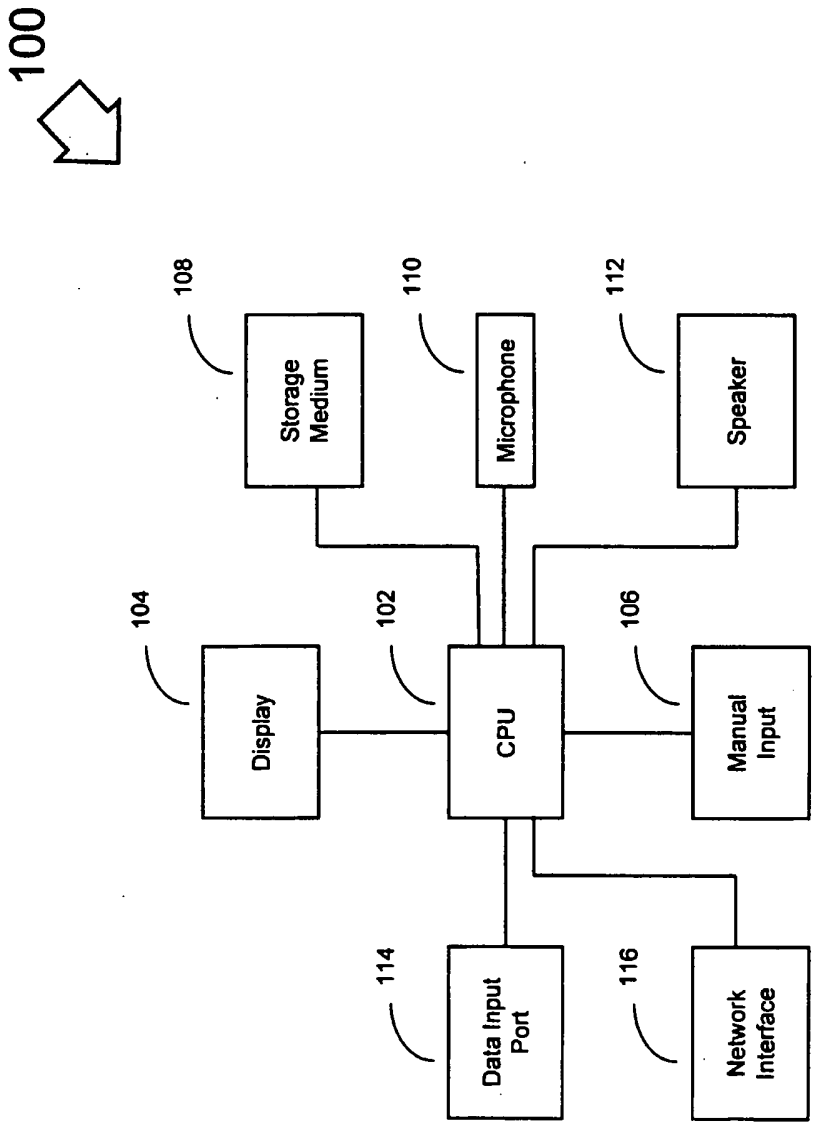

100
108
Storage Medium
110
Microphone
112
Speaker
104
Display
102
CPU
106
Manual Input
114
Data Input Port
116
Network Interface

FIG. 1

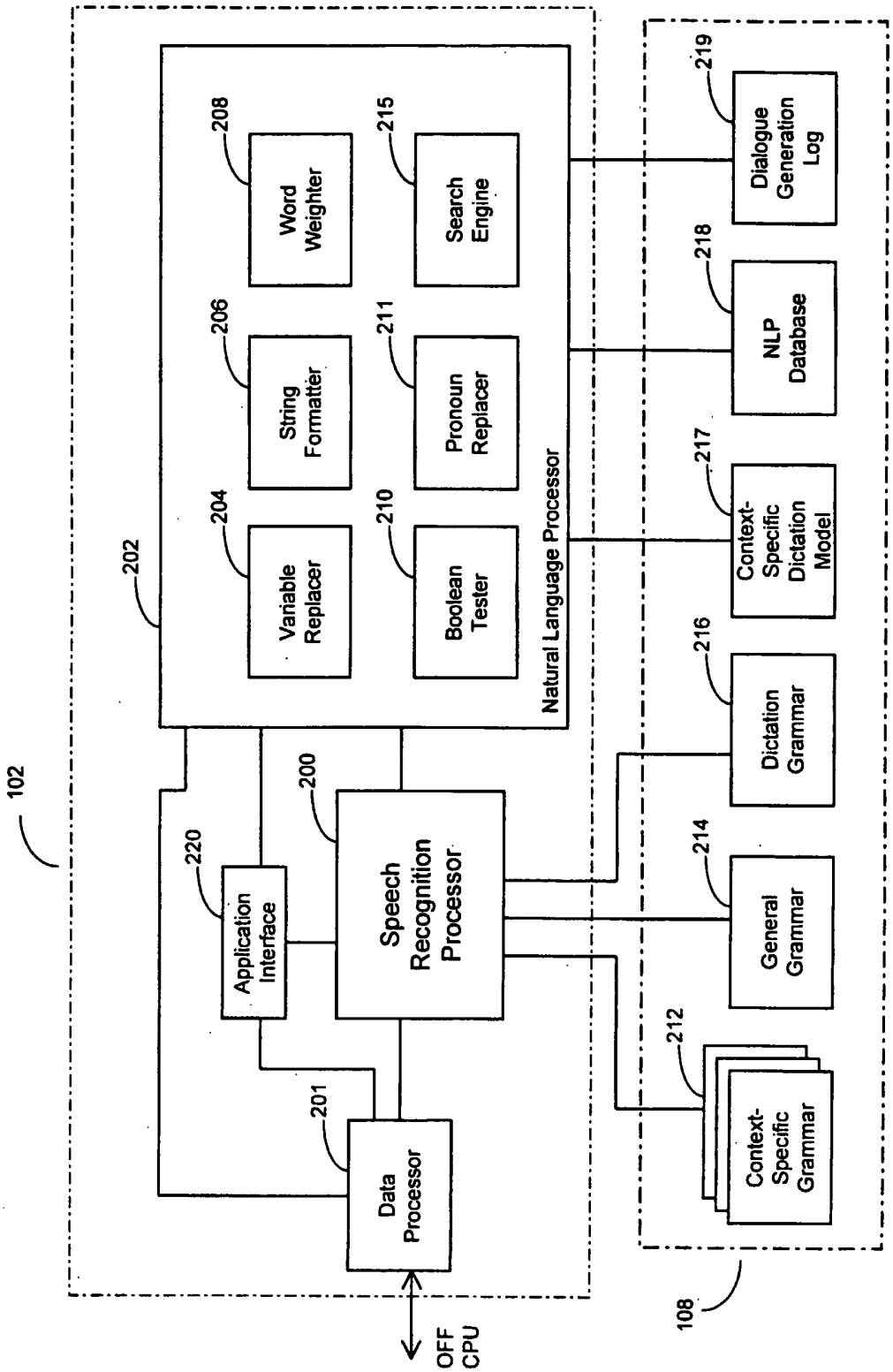

102
202
Natural Language Processor
208
Word Weighter
215
Search Engine
206
String Formatter
211
Pronoun Replacer
204
Variable Replacer
210
Boolean Tester
200
Speech Recognition Processor
220
Application Interface
201
Data Processor
OFF CPU
219
Dialogue Generation Log
218
NLP Database
217
Context-Specific Dictation Model
216
Dictation Grammar
214
General Grammar
212
Context-Specific Grammar
108

FIG. 2

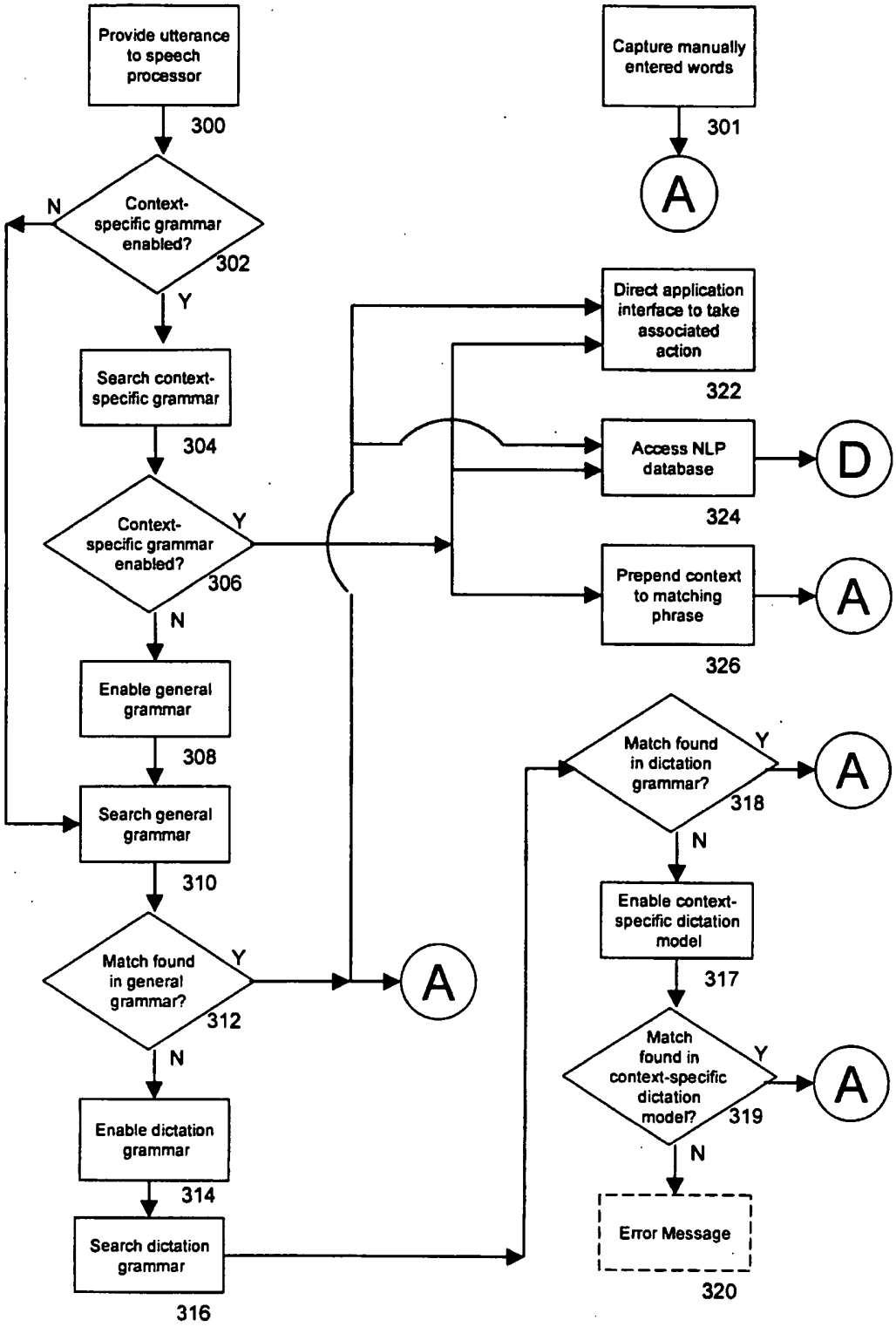

Provide utterance to speech processor
300
Capture manually entered words
301
A
Context-specific grammar enabled?
302
N
Y
Search context-specific grammar
304
Direct application interface to take associated action
322
Access NLP database
324
D
Context-specific grammar enabled?
306
Y
N
Prepend context to matching phrase
326
A
Enable general grammar
308
Match found in dictation grammar?
318
Y
A
N
Search general grammar
310
Enable context-specific dictation model
317
Match found in general grammar?
312
Y
A
N
Match found in context-specific dictation model?
319
Y
A
N
Enable dictation grammar
314
Error Message
320
Search dictation grammar
316

FIG. 3A

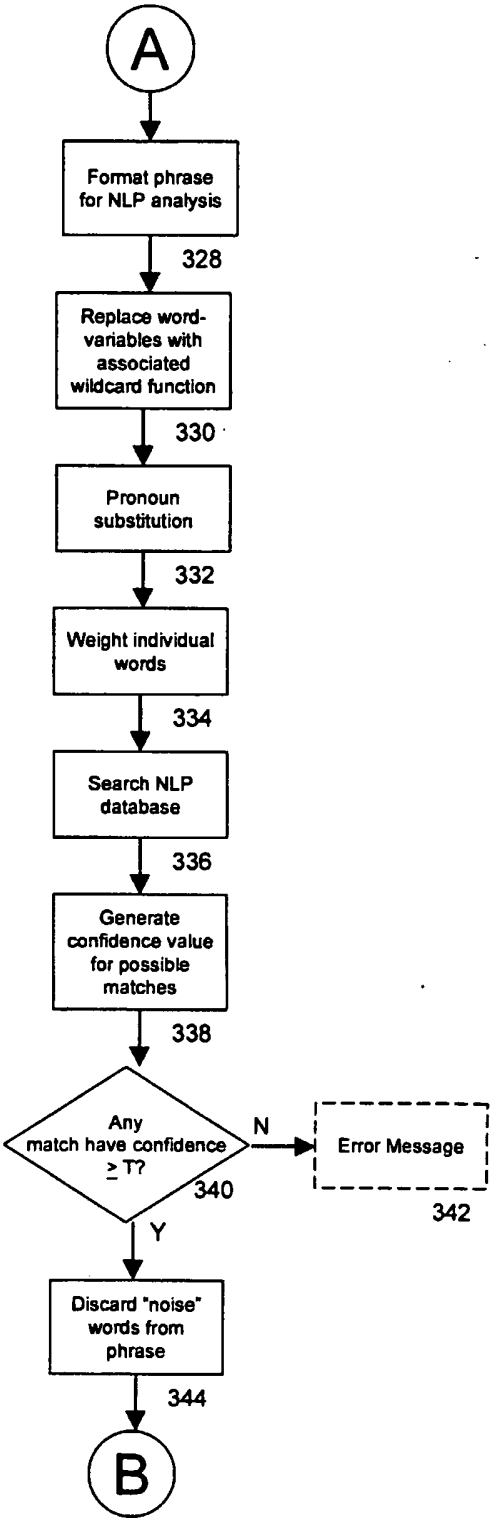

A
Format phrase for NLP analysis
328
Replace word-variables with associated wildcard function
330
Pronoun substitution
332
Weight individual words
334
Search NLP database
336
Generate confidence value for possible matches
338
Any match have confidence ≥ T?
340
N
Error Message
342
Y
Discard "noise" words from phrase
344
B

FIG. 3B

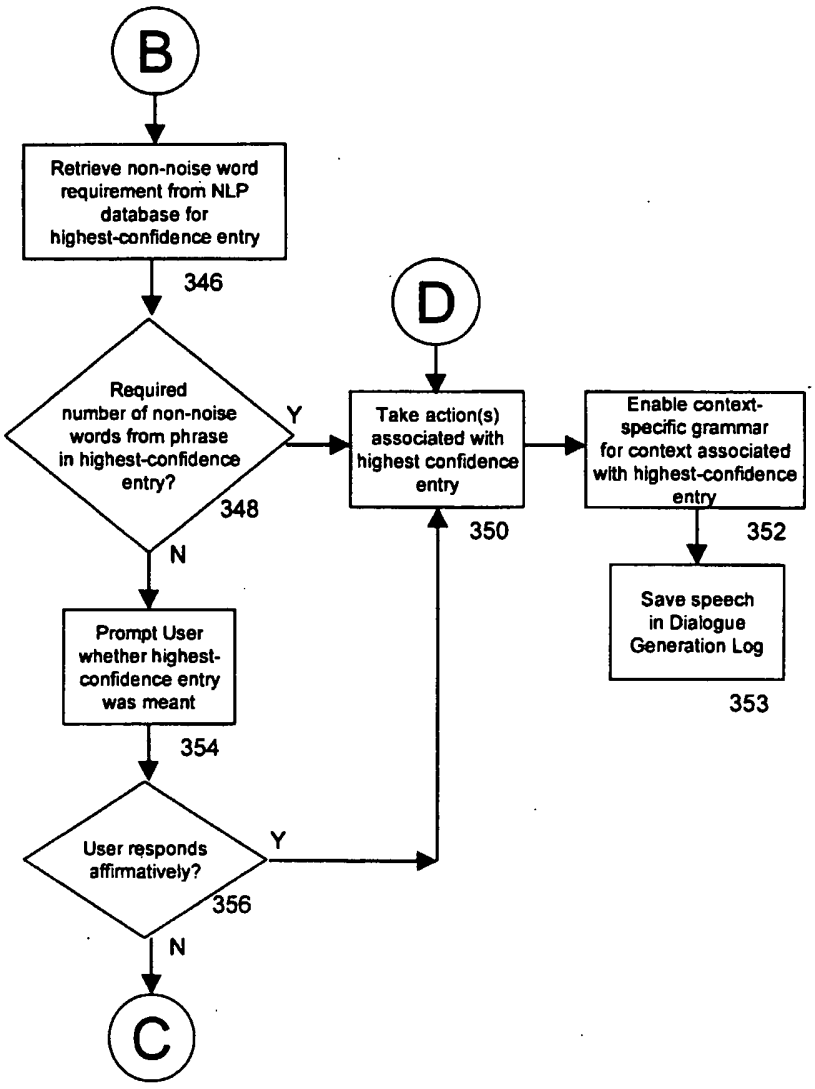

B
Retrieve non-noise word requirement from NLP database for highest-confidence entry
346
Required number of non-noise words from phrase in highest-confidence entry?
348
Y
N
D
Take action(s) associated with highest confidence entry
350
Enable context-specific grammar for context associated with highest-confidence entry
352
Save speech in Dialogue Generation Log
353
Prompt User whether highest-confidence entry was meant
354
User responds affirmatively?
356
Y
N
C

FIG. 3C

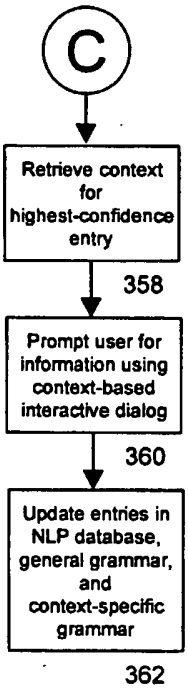

C
Retrieve context for highest-confidence entry
358
Prompt user for information using context-based interactive dialog
360
Update entries in NLP database, general grammar, and context-specific grammar
362

FIG. 3D

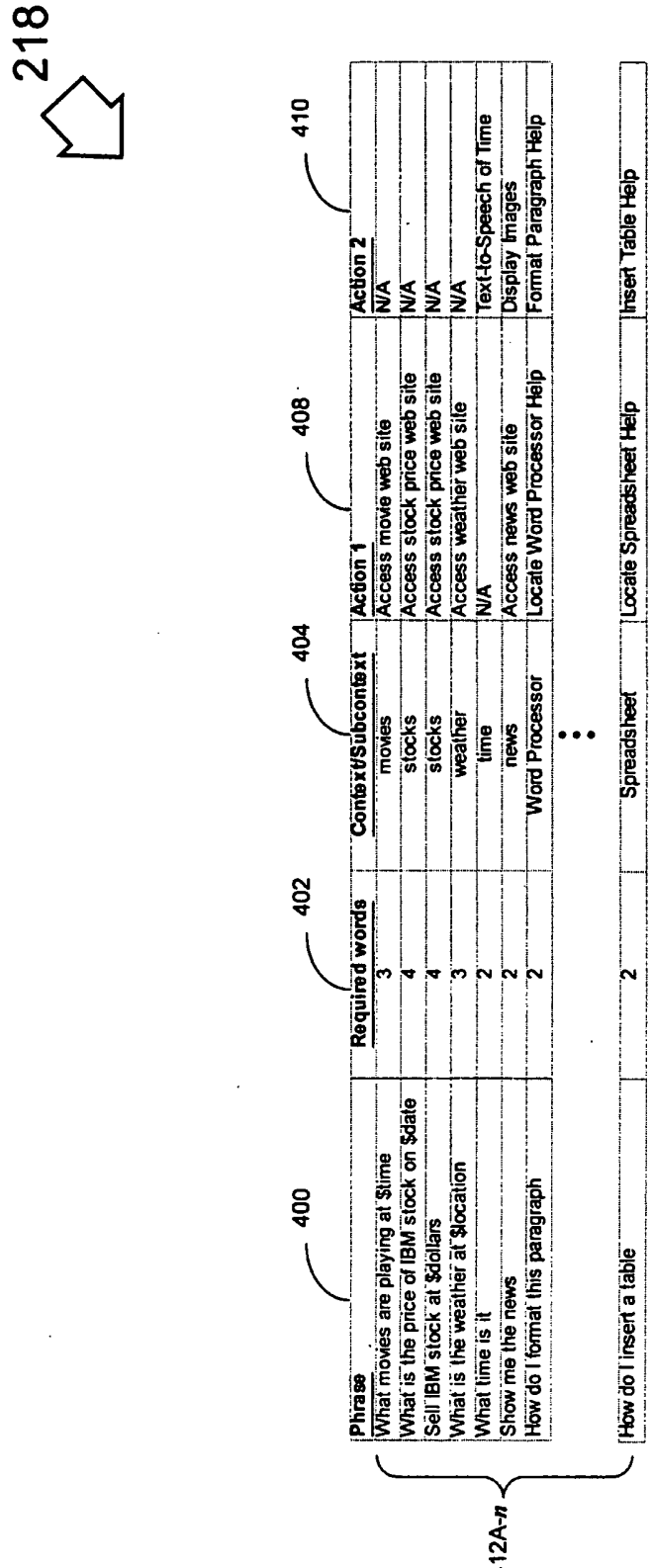

218
400
402
404
408
410
412A-n
Phrase
Required words
Context/Subcontext
Action 1
Action 2
What movies are playing at $time
3
movies
Access movie web site
N/A
What is the price of IBM stock on $date
4
stocks
Access stock price web site
N/A
Sell IBM stock at $dollars
4
stocks
Access stock price web site
N/A
What is the weather at $location
3
weather
Access weather web site
N/A
What time is it
2
time
N/A
Text-to-Speech of Time
Show me the news
2
news
Access news web site
Display Images
How do I format this paragraph
2
Word Processor
Locate Word Processor Help
Format Paragraph Help
How do I insert a table
2
Spreadsheet
Locate Spreadsheet Help
Insert Table Help

FIG. 4

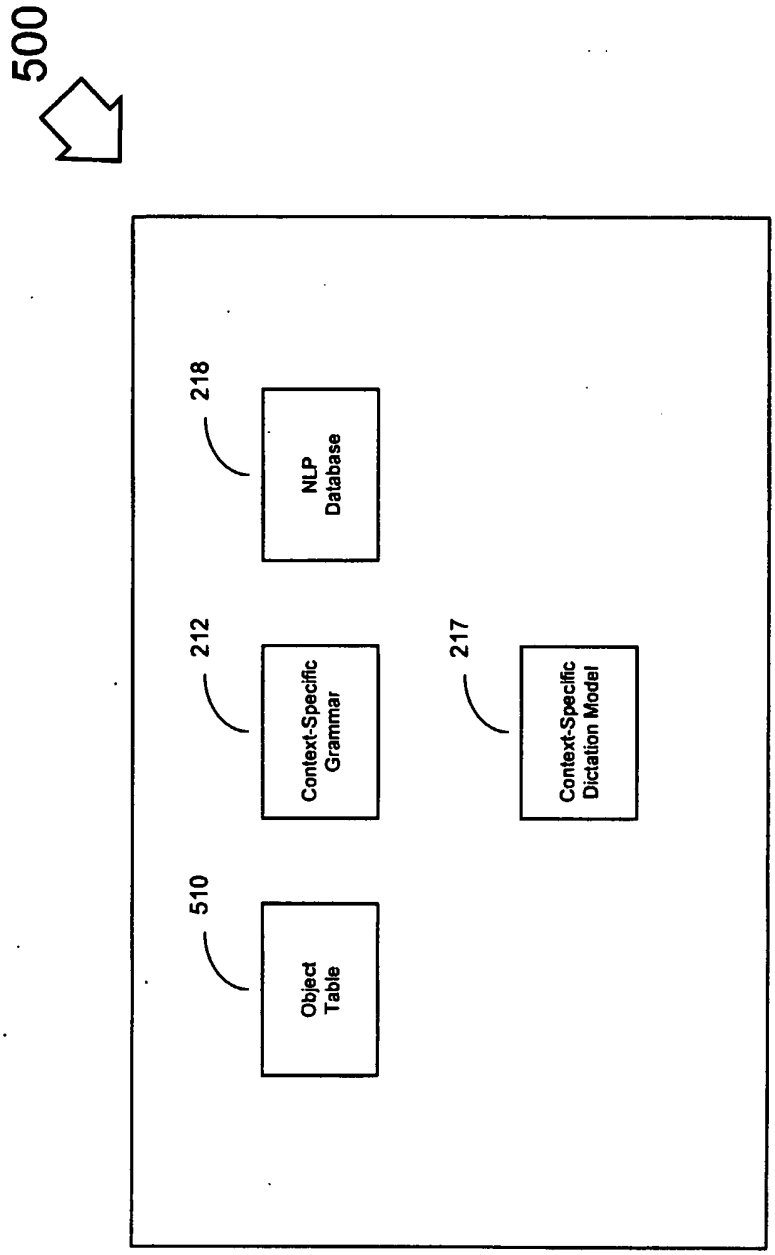

500
218
NLP Database
212
Context-Specific Grammar
217
Context-Specific Dictation Model
510
Object Table

FIG. 5

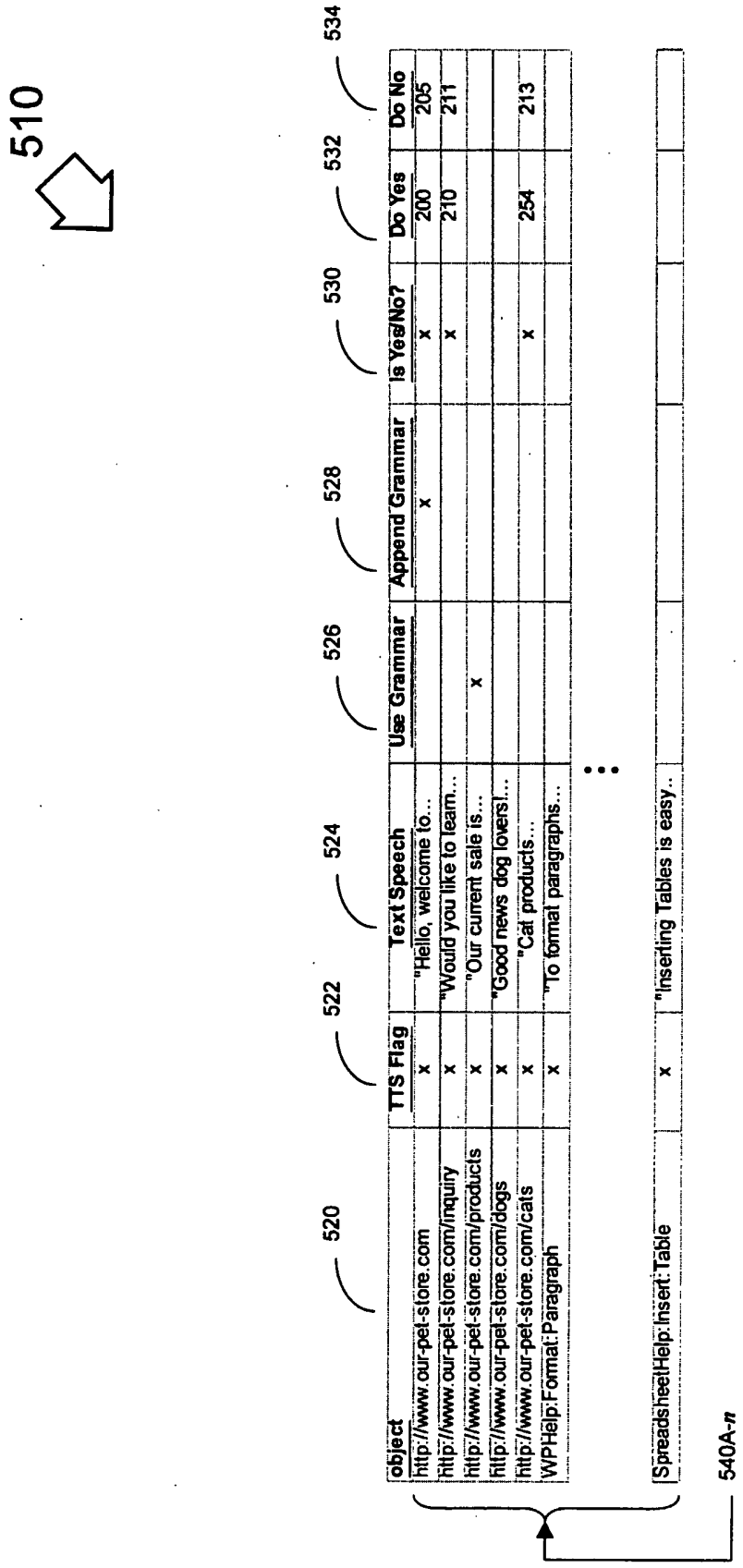

510
520
522
524
526
528
530
532
534
object
TTS Flag
Text Speech
Use Grammar
Append Grammar
Is Yes/No?
Do Yes
Do No
http://www.our-pet-store.com
x
"Hello, welcome to...
200
205
http://www.our-pet-store.com/inquiry
x
"Would you like to learn...
x
210
211
http://www.our-pet-store.com/products
x
"Our current sale is...
x
http://www.our-pet-store.com/dogs
x
"Good news dog lovers!...
http://www.our-pet-store.com/cats
x
"Cat products...
x
254
213
WPHelp:Format:Paragraph
x
"To format paragraphs...
SpreadsheetHelp:Insert:Table
x
"Inserting Tables is easy..
540A-n

FIG. 6

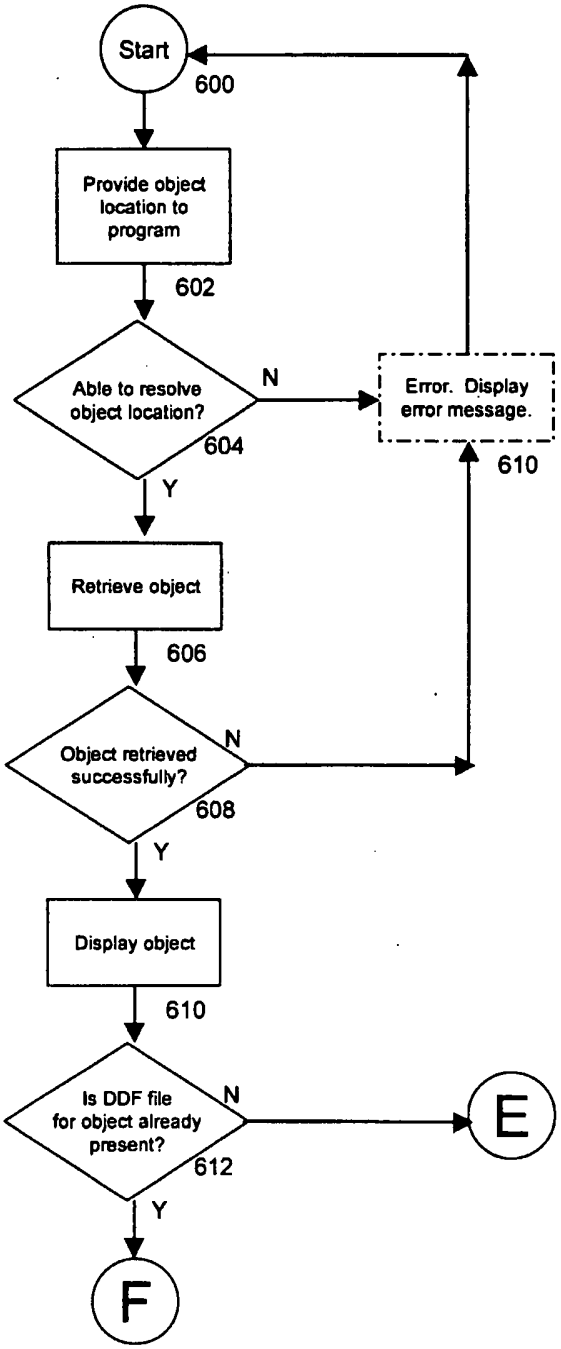

Start
600
Provide object location to program
602
Able to resolve object location?
604
N
Error. Display error message.
610
Y
Retrieve object
606
Object retrieved successfully?
608
N
Y
Display object
610
Is DDF file for object already present?
612
N
E
Y
F

FIG. 7A

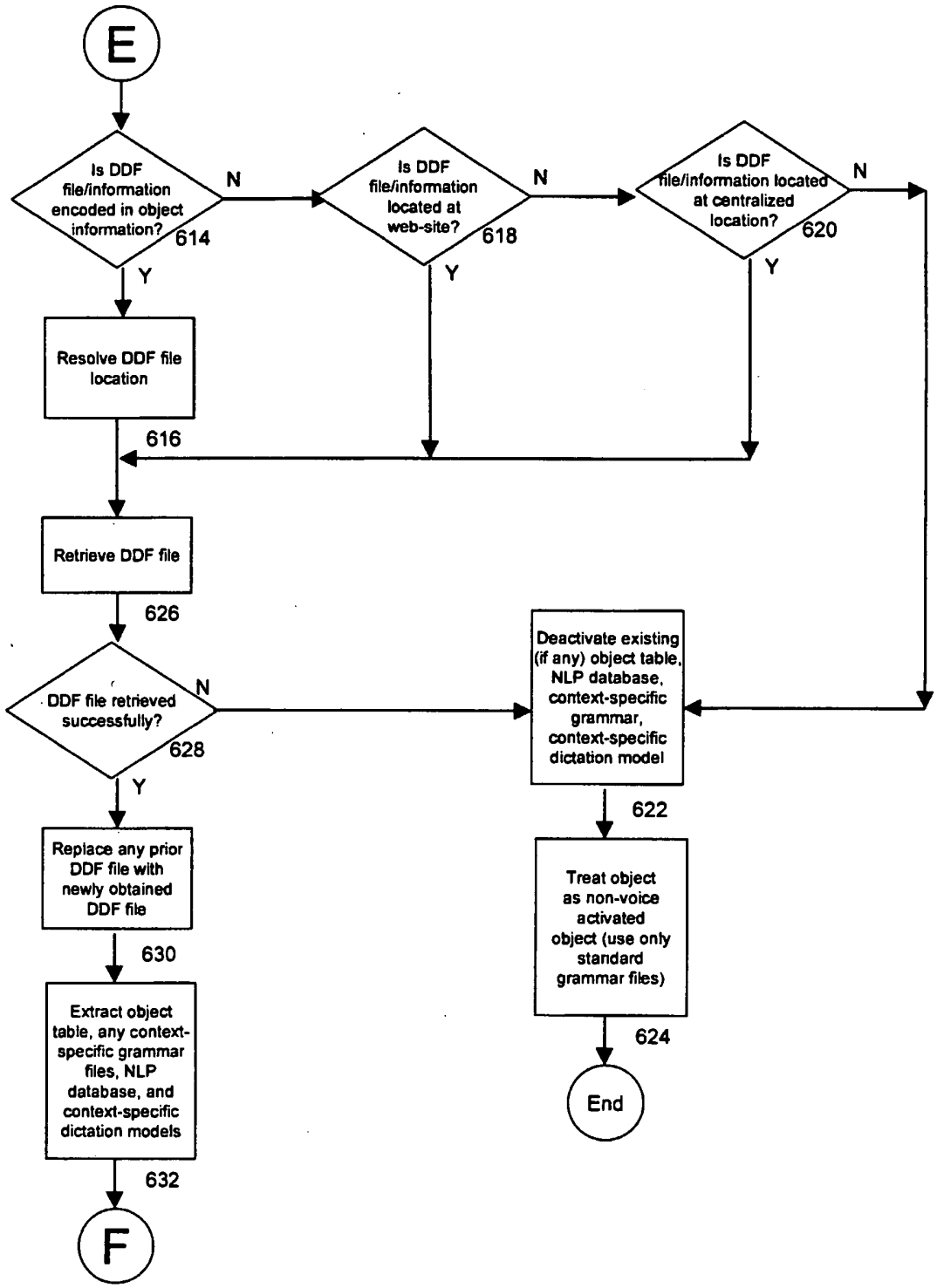

E
Is DDF file/information encoded in object information?
614
N
Y
Is DDF file/information located at web-site?
618
N
Y
Is DDF file/information located at centralized location?
620
N
Y
Resolve DDF file location
616
Retrieve DDF file
626
DDF file retrieved successfully?
628
N
Y
Deactivate existing (if any) object table, NLP database, context-specific grammar, context-specific dictation model
622
Treat object as non-voice activated object (use only standard grammar files)
624
End
Replace any prior DDF file with newly obtained DDF file
630
Extract object table, any context-specific grammar files, NLP database, and context-specific dictation models
632
F

FIG. 7B

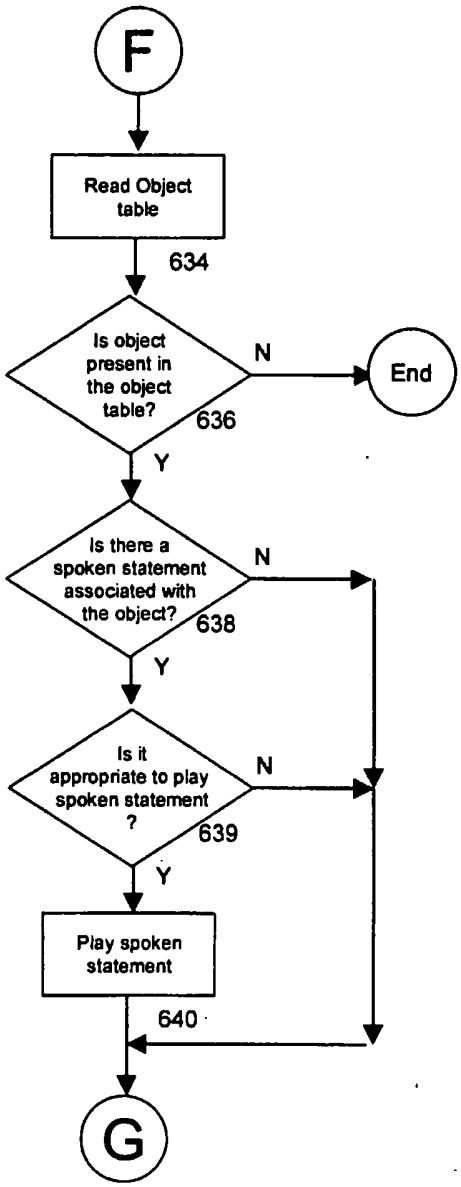

F
Read Object table
634
Is object present in the object table?
636
N
End
Y
Is there a spoken statement associated with the object?
638
N
Y
Is it appropriate to play spoken statement ?
639
N
Y
Play spoken statement
640
G

FIG. 7C

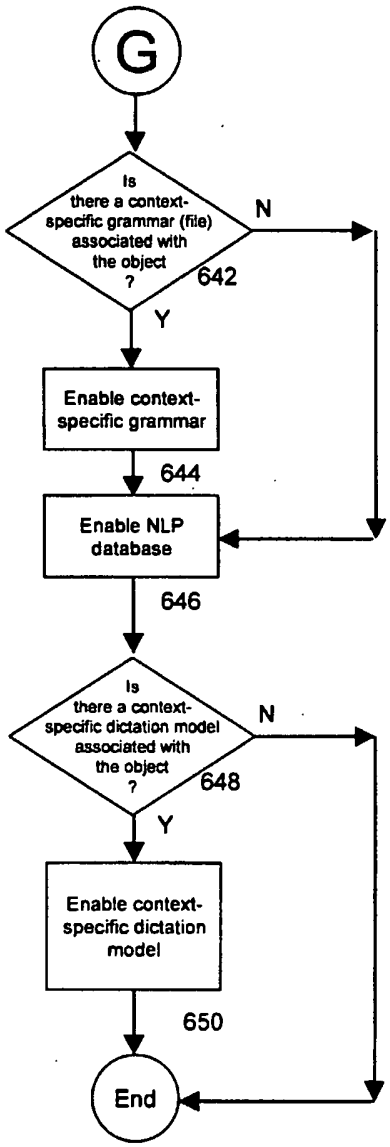

G
Is there a context-specific grammar (file) associated with the object ?
642
N
Y
Enable context-specific grammar
644
Enable NLP database
646
Is there a context-specific dictation model associated with the object ?
648
N
Y
Enable context-specific dictation model
650
End

FIG. 7D

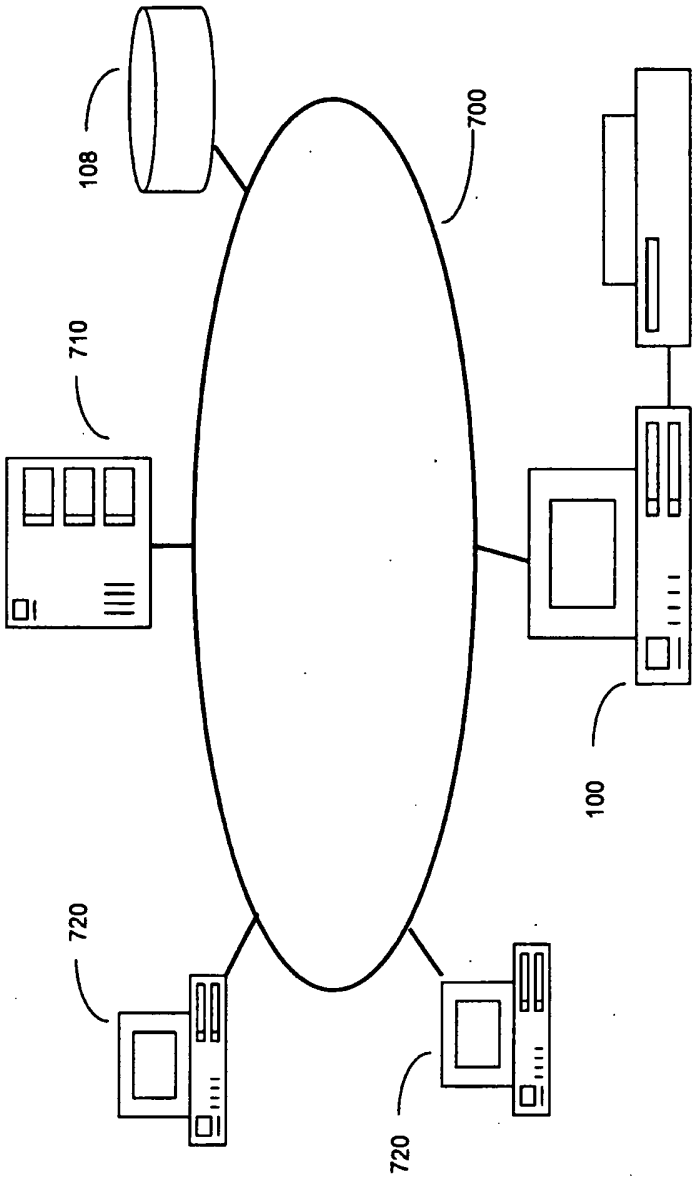


FIG. 8

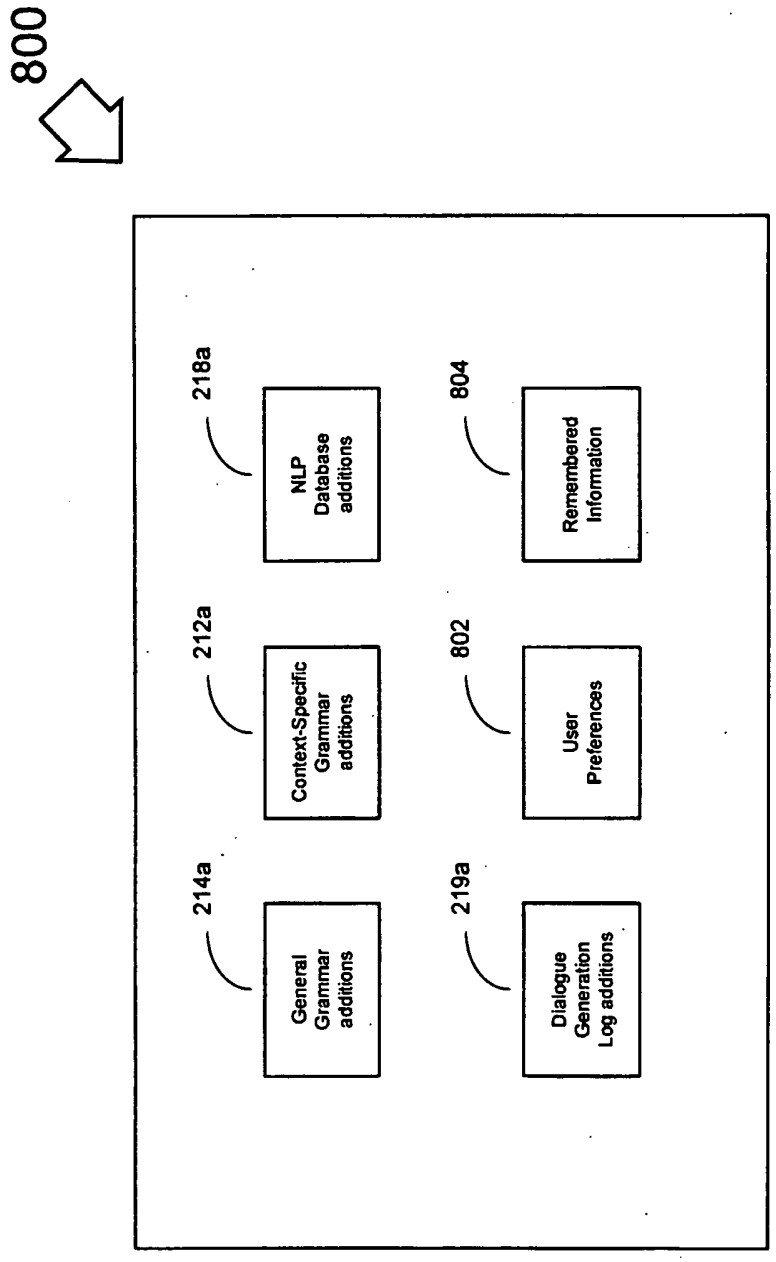

800
218a
NLP Database additions
804
Remembered Information
212a
Context-Specific Grammar additions
802
User Preferences
214a
General Grammar additions
219a
Dialogue Generation Log additions

FIG. 9

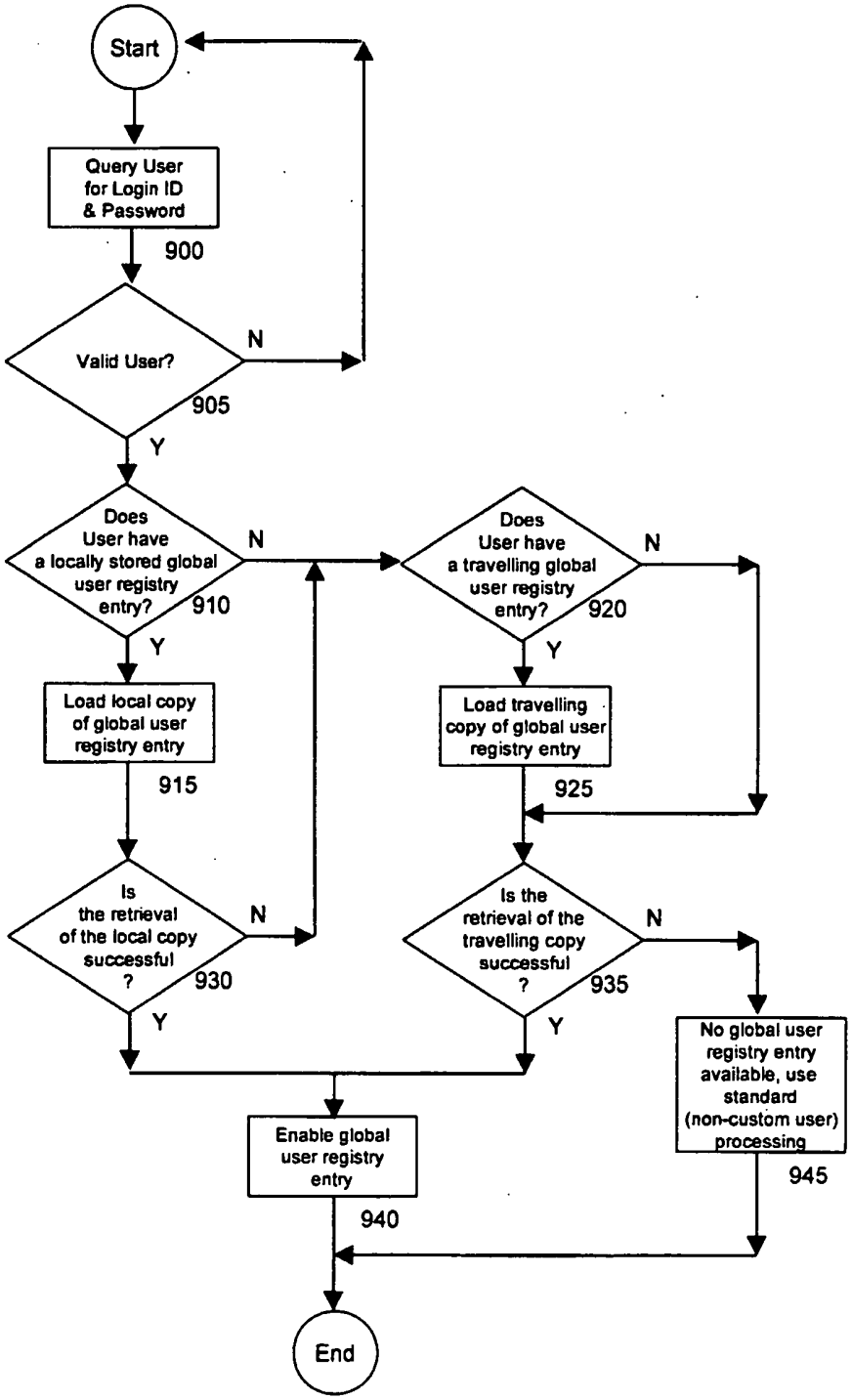

Start
Query User for Login ID & Password
900
Valid User?
N
905
Y
Does User have a locally stored global user registry entry?
N
910
Y
Load local copy of global user registry entry
915
Is the retrieval of the local copy successful ?
N
930
Y
Does User have a travelling global user registry entry?
N
920
Y
Load travelling copy of global user registry entry
925
Is the retrieval of the travelling copy successful ?
N
935
Y
No global user registry entry available, use standard (non-custom user) processing
945
Enable global user registry entry
940
End

FIG. 10A

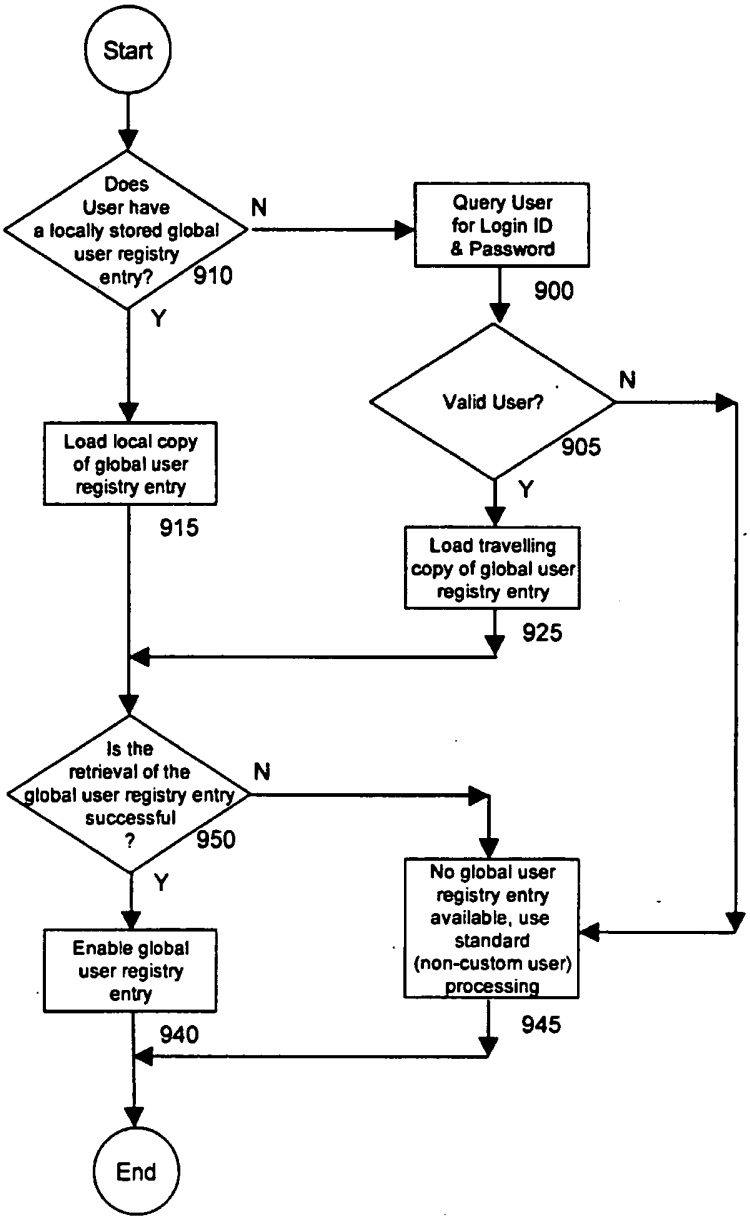

Start
Does User have a locally stored global user registry entry?
910
N
Y
Query User for Login ID & Password
900
Valid User?
905
N
Y
Load local copy of global user registry entry
915
Load travelling copy of global user registry entry
925
Is the retrieval of the global user registry entry successful ?
950
N
Y
No global user registry entry available, use standard (non-custom user) processing
945
Enable global user registry entry
940
End

FIG. 10B

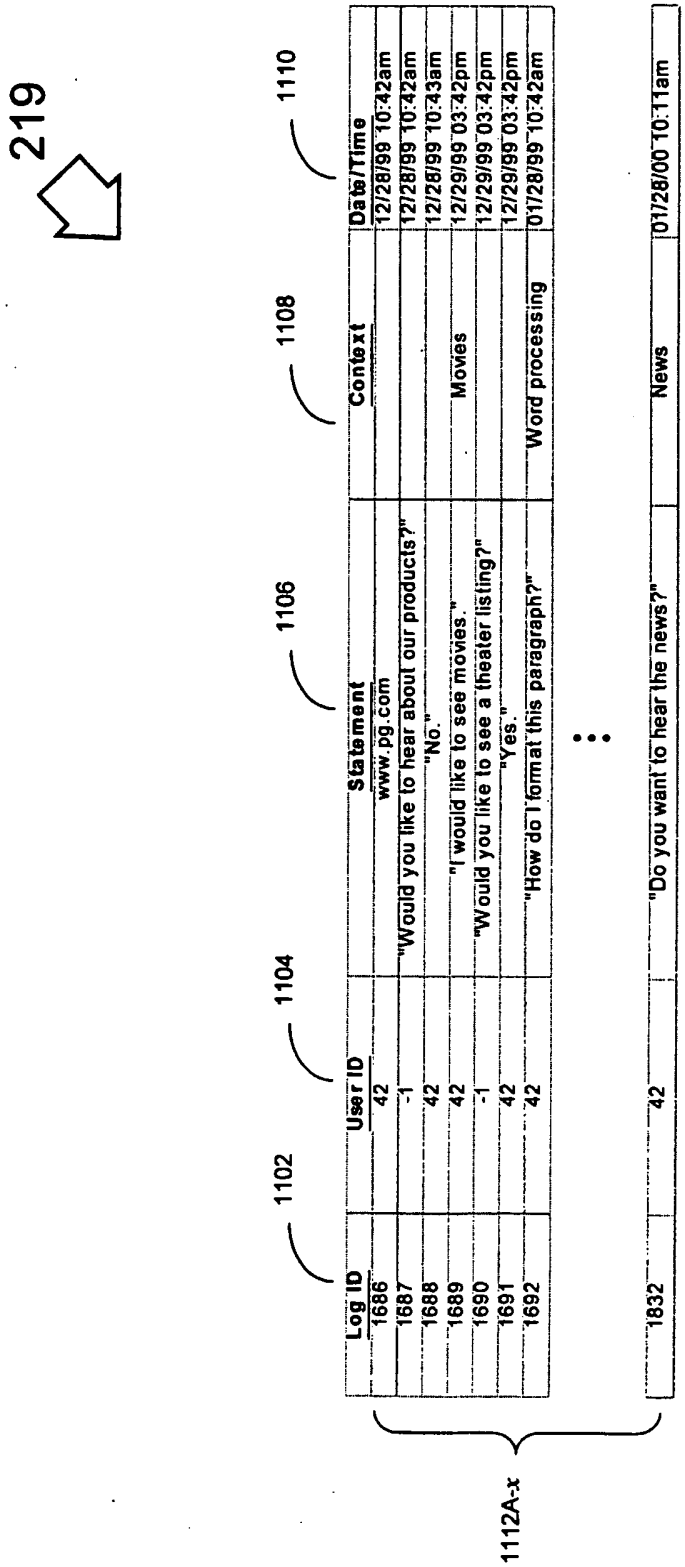

219
1102
1104
1106
1108
1110
Log ID
User ID
Statement
Context
Date/Time
1686
42
www.pg.com
12/28/99 10:42am
1687
-1
"Would you like to hear about our products?"
12/28/99 10:42am
1688
42
"No."
12/28/99 10:43am
1689
42
"I would like to see movies."
Movies
12/29/99 03:42pm
1690
-1
"Would you like to see a theater listing?"
12/29/99 03:42pm
1691
42
"Yes."
12/29/99 03:42pm
1692
42
"How do I format this paragraph?"
Word processing
01/28/99 10:42am
1832
42
"Do you want to hear the news?"
News
01/28/00 10:11am
1112A-x

FIG. 11

INTERNATIONAL SEARCH REPORT

In nal Application No
PCT/US 01/11138

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L15/22 G10L15/26

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 00 14728 A (WEBER DEAN C ;ONE VOICE TECHNOLOGIES INC (US)) 16 March 2000 (2000-03-16) the whole document ---	1-37
X	WO 00 14727 A (WEBER DEAN C ;ONE VOICE TECHNOLOGIES INC (US)) 16 March 2000 (2000-03-16) the whole document ---	1,2, 5-21, 24-37
A	US 5 761 631 A (NASUKAWA) 2 June 1998 (1998-06-02) abstract --- -/--	3-5, 15-18, 22-24, 34-37

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
29 August 2001	06/09/2001

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Quélavoine, R

Form PCT/ISA/210 (second sheet) (July 1992)

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
A	WYARD P J ET AL: "SPOKEN LANGUAGE SYSTEMS - BEYOND PROMPT AND RESPONSE" BT TECHNOLOGY JOURNAL, BT LABORATORIES, GB, vol. 14, no. 1, 1996, pages 187-207, XP000554648 ISSN: 1358-3948 abstract	1-8, 19-27
A	US 5 311 429 A (TOMINAGA) 10 May 1994 (1994-05-10) abstract	15-18, 34-37

INTERNATIONAL SEARCH REPORT

Information on patent family members

Ir ›nal Application No
PCT/US 01/11138

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 0014728 A	16-03-2000	AU 5813099 A	27-03-2000
		AU 6243999 A	27-03-2000
		EP 1110206 A	27-06-2001
		EP 1110205 A	27-06-2001
		WO 0014727 A	16-03-2000
WO 0014727 A	16-03-2000	AU 5813099 A	27-03-2000
		AU 6243999 A	27-03-2000
		EP 1110206 A	27-06-2001
		EP 1110205 A	27-06-2001
		WO 0014728 A	16-03-2000
US 5761631 A	02-06-1998	JP 2855409 B	10-02-1999
		JP 8147299 A	07-06-1996
US 5311429 A	10-05-1994	JP 2301869 A	13-12-1990

Form PCT/ISA/210 (patent family annex) (July 1992)

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(c) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
11/580,926	10/16/2006	Larry Baldwin	090611-0357945

CONFIRMATION NO. 6670

909
PILLSBURY WINTHROP SHAW PITTMAN, LLP
P.O. BOX 10500
MCLEAN, VA22102

Title: System and method for a cooperative conversational voice user interface

Publication No. US-2008-0091406-A1
Publication Date: 04/17/2008

NOTICE OF PUBLICATION OF APPLICATION

The above-identified application will be electronically published as a patent application publication pursuant to 37 CFR 1.211, et seq. The patent application publication number and publication date are set forth above.

The publication may be accessed through the USPTO's publically available Searchable Databases via the Internet at www.uspto.gov. The direct link to access the publication is currently http://www.uspto.gov/patft/.

The publication process established by the Office does not provide for mailing a copy of the publication to applicant. A copy of the publication may be obtained from the Office upon payment of the appropriate fee set forth in 37 CFR 1.19(a)(1). Orders for copies of patent application publications are handled by the USPTO's Office of Public Records. The Office of Public Records can be reached by telephone at (703) 308-9726 or (800) 972-6382, by facsimile at (703) 305-8759, by mail addressed to the United States Patent and Trademark Office, Office of Public Records, Alexandria, VA 22313-1450 or via the Internet.

In addition, information on the status of the application, including the mailing date of Office actions and the dates of receipt of correspondence filed in the Office, may also be accessed via the Internet through the Patent Electronic Business Center at www.uspto.gov using the public side of the Patent Application Information and Retrieval (PAIR) system. The direct link to access this status information is currently http://pair.uspto.gov/. Prior to publication, such status information is confidential and may only be obtained by applicant using the private side of PAIR.

Further assistance in electronically accessing the publication, or about PAIR, is available by calling the Patent Electronic Business Center at 1-866-217-9197.

Pre-Grant Publication Division, 703-605-4283

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	1A	4,430,669	02/07/1984	Cheung	358	122	
	1B	5,274,560	12/28/1993	LaRue	364	444	
	1C	5,377,350	12/27/1994	Skinner	395	600	
	1D	5,500,920	03/19/1996	Kupiec	395	2.79	
	1E	5,696,965	12/09/1997	Dedrick	395	610	
	1F	5,774,859	06/30/1998	Houser *et al.*	704	275	
	1G	5,878,385	03/02/1999	Bralich *et al.*	704	9	
	1H	5,902,347	05/11/1999	Backman *et al.*	701	200	
	1I	6,195,634	02/27/2001	Dudemaine *et al.*	704	231	
	1J	6,415,257	07/02/2002	Junqua *et al.*	704	275	
	1K	6,466,654	10/15/2002	Cooper *et al.*	379	88.01	
	1L	6,513,006	01/28/2003	Howard *et al.*	704	257	
	1M	6,567,797	05/20/2003	Schuetze *et al.*	707	2	
	1N	6,570,555	05/27/2003	Prevost *et al.*	345	156	
	1O	6,574,597	06/03/2003	Mohri *et al.*	704	251	
	1P	6,598,018	07/22/2003	Junqua	704	251	
	1Q	6,631,346	10/07/2003	Karaorman *et al.*	704	9	
	1R	6,643,620	11/04/2003	Contolini *et al.*	704	270	
	1S	6,721,706	04/13/2004	Strubbe *et al.*	704	275	
	1T	6,735,592	05/11/2004	Neumann *et al.*	707	101	
	1U	6,757,718	06/29/2004	Halverson *et al.*	709	218	
	1V	6,795,808	09/21/2004	Strubbe *et al.*	704	275	
	1W	6,901,366	05/31/2005	Kuhn *et al.*	704	275	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

	1X	CHAI *et al.*, "MIND: A Semantics-Based Multimodal Interpretation Framework for Conversational System", Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, June 2002, pages 37-46.
	1Y	CHEYER *et al.*, "Multimodal Maps: An Agent-Based Approach", International Conference on Cooperative Multimodal Communication (CMC/95), May 24-26, 1995, pages 111-121.
	1Z	ELIO *et al.*, "On Abstract Task Models and Conversation Policies" in Workshop on Specifying and Implementing Conversation Policies, Autonomous Agents '99, Seattle, 1999, 10 pages.

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

401346771v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	2A	6,944,594	09/13/2005	Busayapongchai *et al.*	704	275	
	2B	7,127,400	10/24/2006	Koch	704	270.1	
	2C	7,146,319	12/05/2006	Hunt	704	254	
	2D	7,289,606	10/30/2007	Sibal *et al.*	379	52	
	2E	7,376,645	05/20/2008	Bernard	707	3	
	2F	7,398,209	07/08/2008	Kennewick *et al.*	704	255	
	2G	7,424,431	09/09/2008	Greene *et al.*	704	270	
	2H	2001/0041980	11/15/2001	Howard *et al.*	704	270	
	2I	2002/0049805	04/25/2002	Yamada *et al.*	709	202	
	2J	2002/0065568	05/30/2002	Silfvast *et al.*	700	94	
	2K	2002/0082911	06/27/2002	Dunn *et al.*	705	14	
	2L	2002/0120609	08/29/2002	Lang *et al.*	707	1	
	2M	2002/0124050	09/05/2002	Middeljans	709	203	
	2N	2002/0188602	12/12/2002	Stubler *et al.*	707	3	
	2O	2002/0198714	12/26/2002	Zhou	704	252	
	2P	2003/0014261	01/16/2003	Kageyama	704	275	
	2Q	2003/0112267	06/19/2003	Belrose	345	728	
	2R	2004/0025115	02/05/2004	Sienel *et al.*	715	513	
	2S	2004/0166832	08/26/2004	Portman *et al.*	455	412.1	
	2T	2004/0199375	10/07/2004	Ehsani *et al.*	704	4	
	2U	2005/0021334	01/27/2005	Iwahashi	704	240	
	2V	2005/0114116	05/26/2005	Fiedler	704	201	
	2W	2007/0033005	02/08/2007	Cristo *et al.*	704	9	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

	2X	TURUNEN, "Adaptive Interaction Methods in Speech User Interfaces", Conference on Human Factors in Computing Systems, Seattle, Washington, 2001, pages 91-92.
	2Y	MAO, Mark Z., "Automatic Training Set Segmentation for Multi-Pass Speech Recognition", Department of Electrical Engineering, Stanford University, CA, copyright 2005, IEEE, pages I-685 to I-688.

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

401346771v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	3A	2007/0038436	02/15/2007	Cristo *et al.*	704	9	
	3B	2007/0050191	03/01/2007	Weider *et al.*	704	275	
	3C	2008/0091406	04/17/2008	Baldwin *et al.*	704	4	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

	3D	VANHOUCKE, Vincent, "Confidence Scoring and Rejection Using Multi-Pass Speech Recognition", Nuance Communications, Menlo Park, CA, [no date], 4 pages.
	3E	WENG, Fuliang, *et al.*, "Efficient Lattice Representation and Generation", Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, [no date], 4 pages.
	3F	EL MELIANI *et al.*, "A Syllabic-Filler-Based Continuous Speech Recognizer for Unlimited Vocabulary", Canadian Conference on Electrical and Computer Engineering, Volume 2, September 5-8, 1995, pages 1007-1010.

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

401346771v1

Electronic Acknowledgement Receipt

EFS ID:	5634265
Application Number:	11580926
International Application Number:	
Confirmation Number:	6670
Title of Invention:	System and method for a cooperative conversational voice user interface
First Named Inventor/Applicant Name:	Larry Baldwin
Customer Number:	00909
Filer:	Syed Jafar Ali/Susan Trader
Filer Authorized By:	Syed Jafar Ali
Attorney Docket Number:	090611-0357945
Receipt Date:	02-JUL-2009
Filing Date:	16-OCT-2006
Time Stamp:	13:08:38
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		0357945_SIDS.pdf	870548 57bdd06a65124772bb7dc2ab33ca4d64ba1ec197	yes	19

	Multipart Description/PDF files in .zip description		
	Document Description	**Start**	**End**
	Transmittal Letter	1	16
	Information Disclosure Statement (IDS) Filed (SB/08)	17	19

Warnings:					
Information:					
2	NPL Documents	CHAI.pdf	1036235 82e5c44d0c4460a84b5763f6ca9858e2072a4f38	no	10
Warnings:					
Information:					
3	NPL Documents	CHEYER.pdf	136285 5e49cb5ed1eb7cad08bed8e144f0d4292d0d5744	no	11
Warnings:					
Information:					
4	NPL Documents	ELIO.pdf	934896 d46aa8144b37e3a1d338415605eb1066ae6fc010	no	10
Warnings:					
Information:					
5	NPL Documents	TURUNEN.pdf	179163 b47bbd1deaa06420e04206c4216ccee446a68755	no	2
Warnings:					
Information:					
6	NPL Documents	MAO.pdf	87562 1b3ffe999a30b15eea176ac4951c57362d0c3bbb	no	4
Warnings:					
Information:					
7	NPL Documents	VANHOUCKE.pdf	258100 095933ea064f85a3705750030dec1a9690d3ae6d	no	4
Warnings:					
Information:					
8	NPL Documents	WENG.pdf	345795 ab3fae98c6cdea032dd636d4c37094120b0def8c	no	4
Warnings:					
Information:					

9	NPL Documents	EL_MELIANI.pdf	294901 b1c4510f36ba51ee85f20457fdd383145f4e918a	no	4

Warnings:	
Information:	
Total Files Size (in bytes):	4143485

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111
If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371
If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office
If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANTS :	Larry BALDWIN *et al.*	CONFIRMATION NO.:	6670
SERIAL NUMBER :	11/580,926	EXAMINER:	Richemond Dorvil
FILING DATE :	October 16, 2006	ART UNIT:	2626
FOR :	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE		

SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT

Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313-1450

Sir:

Pursuant to the duty of disclosure under 37 C.F.R. §§ 1.56, 1.97 and 1.98, Applicants hereby make of record in the above-identified application the documents listed on the attached Form PTO-1449.

Applicants have listed publication dates on the attached PTO-1449 based on information presently available to the undersigned. However, the listed publication dates should not be construed as an admission that the information was actually published on the date indicated.

Applicants reserve the right to establish the patentability of the claimed invention over any of the information provided herewith, and/or to prove that this information may not be prior art, and/or to prove that this information may not be enabling for the teachings purportedly offered.

This Supplemental Information Disclosure Statement is being filed within three months of the filing date of a national application other than a continued prosecution application under 37 C.F.R. §1.53(d), or within three months of the date of entry of the national stage as set forth in 37 C.F.R. §1.491 in an international application, or before the mailing date of a first Office Action on the merits, or before the mailing date of a first Office Action after the filing of a request for continued examination under 37 C.F.R. §1.114. No certification or fee is required. 37 C.F.R. §1.97(b).

Some of these documents were cited in search reports by a foreign patent office in counterpart foreign applications. Submission of an English language version of each search report that indicates the degree of relevance found by the foreign office is provided.

It is respectfully requested that the Examiner initial and return a copy of the enclosed PTO-1449.

Respectfully submitted,

PILLSBURY WINTHROP SHAW PITTMAN LLP

By_____
Syed Jafar Ali
Registration No.: 58,780

Customer No.
00909

P. O. Box 10500
McLean, Virginia 22102
Telephone: (703) 770-7900
Telefax: (703) 770-7901

Dated: July 2, 2009

PATENT COOPERATION TREATY

PCT

INTERNATIONAL SEARCH REPORT

(PCT Article 18 and Rules 43 and 44)

Applicant's or agent's file reference 090611-0356171	**FOR FURTHER ACTION** see Form PCT/ISA/220 as well as, where applicable, item 5 below.	
International application No. PCT/US06/33378	International filing date (*day/month/year*) 29 August 2006 (29.08.2006)	(Earliest) Priority Date (*day/month/year*) 29 August 2005 (29.08.2005)
Applicant VOICEBOX TECHNOLOGIES, INC.		

This international search report has been prepared by this International Searching Authority and is transmitted to the applicant according to Article 18. A copy is being transmitted to the International Bureau.

This international search report consists of a total of 6 sheets.

☒ It is also accompanied by a copy of each prior art document cited in this report.

1. **Basis of the Report**
   a. With regard to the **language**, the international search was carried out on the basis of:
      ☒ the international application in the language in which it was filed.
      ☐ a translation of the international application into _____, which is the language of a translation furnished for the purposes of international search (Rules 12.3(a) and 23.1(b))
   b. ☐ With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, see Box No. I.
2. ☐ **Certain claims were found unsearchable** (See Box No. II)
3. ☒ **Unity of invention is lacking** (See Box No. III)
4. With regard to the **title**,
   ☒ the text is approved as submitted by the applicant.
   ☐ the text has been established by this Authority to read as follows:

5. With regard to the **abstract**,
   ☐ the text is approved as submitted by the applicant.
   ☒ the text has been established, according to Rule 38.2(b), by this Authority as it appears in Box No. IV. The applicant may, within one month from the date of mailing of this international search report, submit comments to this Authority.

6. With regard to the **drawings**,
   a. the figure of the **drawings** to be published with the abstract is Figure No. 5
      ☐ as suggested by the applicant.
      ☐ as selected by this Authority, because the applicant failed to suggest a figure.
      ☒ as selected by this Authority, because this figure better characterizes the invention.
   b. ☐ none of the figures is to be published with the abstract.

Form PCT/ISA/210 (first sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US06/33378

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:
Please See Continuation Sheet

1. ☒ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of any additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

Form PCT/ISA/210 (continuation of first sheet(2)) (April 2005)

Box IV TEXT OF THE ABSTRACT (Continuation of Item 5 of the first sheet)

A mobile system is provided that includes speech-based (136) and non-speech-based interfaces (114, 155) for telematics applications. The mobile system identifies and uses context, prior information, domain knowledge, and user specific profile data to achieve a natural "environment for users that submit requests and/or commands in multiple domains (106, 108, 110, 112). The invention creates, stores and uses extensive personal profile information for each user (110), thereby improving the reliability of determining the context and presenting the expected results for a particular question or command. The invention may organize domain specific behavior and information into agents (106), that are distributable or updateable over a wide area network.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US06/33378

A. CLASSIFICATION OF SUBJECT MATTER
IPC: **G10L 15/00**(2006.01),**15/26**(2006.01),**21/00**(2006.01);**G06F 17/30**(2006.01)
G10L 13/00(2006.01),19/00(2006.01)

USPC: 704/231,235,270.1,275;704/3,6
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 704/231,235,270.1,275; 704/3,6

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Google Scholar, IEEE Explorer

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 6,567,797 B1 (SCHUETZE et al) 20 May 2003 (20.05.2003), column 20, line 46- column 24, line 10.	1-18 and 31-49
A	US 6,570,555 B1 (PREVOST et al) 27 May 2003 (27.05.2003), abstract.	1-18 and 31-49
Y	US 6,735,592 B1 (NEUMANN et al) 11 May 2004 (11.05.2004), column 19, line 39- column 10, line 56.	1-18 and 31-49
Y	US 6,757,718 B1 (HALVERSON et al) 29 June 2004 (29.06.2004), column 13, lines 15-51.	1-18 and 31-49
Y	US 2002/0188602 A1 (STUBLER et al) 12 December 2002 (12.12.2002), paragraph 0039.	28-30
X --- Y	US 2003/01122267 A1 (BELROSE) 19 June 2003 (19.06.2003), paragraphs 98-114.	19-27 --- 28-30

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
01 August 2007 (01.08.2007)	25 OCT 2007
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US Commissioner for Patents P.O. Box 1450 Alexandria, Virginia 22313-1450 Facsimile No. (571) 273-3201	Authorized officer James S. Wozniak Telephone No. 571-272-7632

Form PCT/ISA/210 (second sheet) (April 2005)

C. (Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2004/0044516 A1 (KENNEWICK et al) 04 March 2004 (04.03.2004), Paragraphs 144-186.	1-18 and 31-49
Y	CHAI et al. MIND: A Semantics-based Multimodal Interpretation Frameworkfor Conversational Systems. Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, June, 2002, pp. 37-46.	1-18 and 31-49
Y	CHEYER etal. Multimodal Maps: An Agent-based Approach. International Conference on Cooperative Multimodal Communication (CMC/95), 24-26 May 1995, pages 111-121.	1-18 and 31-49

BOX III. OBSERVATIONS WHERE UNITY OF INVENTION IS LACKING

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fees must be paid.

Group I, claim(s) 1-18 and 31-49, drawn to a system/method for multimodal query processing and information retrieval.

Group II, claim(s) 19-30, drawn to a system for creating annotated media files using speech and non-speech inputs.

The inventions listed as Groups I and II do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons: Group I is drawn to a mobile system/method for searching for content gathered by domain agents (i.e., process/system for using a product) and Group II is drawn to a separate system for creating annotated media files. The creation of annotated media files is a separate inventive concept because it involves different means/steps and can be applied to other applications such as photo album indexing, captioning, and media library/playlist creation. The invention in Group I would also not require any of the specific annotation means/steps of Group II in order to query content. For example, pre-indexed files or differently indexed content could be utilized. Thus, the inventions of Groups I and II, refer to separate inventions *(i.e., process/system for using a product and system for creating a product- See MPEP 1850 (III)(A))*.

PATENT COOPERATION TREATY

PCT

INTERNATIONAL SEARCH REPORT

(PCT Article 18 and Rules 43 and 44)

Applicant's or agent's file reference 90611-368071	FOR FURTHER ACTION	see Form PCT/ISA/220 as well as, where applicable, item 5 below.
International application No. PCT/US 08/53157	International filing date *(day/month/year)* 06 February 2008 (06.02.2008)	(Earliest) Priority Date *(day/month/year)* 06 February 2007 (06.02.2007)
Applicant VOICEBOX TECHNOLOGIES, INC.		

This international search report has been prepared by this International Searching Authority and is transmitted to the applicant according to Article 18. A copy is being transmitted to the International Bureau.

This international search report consists of a total of 2 sheets.

☐ It is also accompanied by a copy of each prior art document cited in this report.

1. **Basis of the report**

a. With regard to the language, the international search was carried out on the basis of:

☒ the international application in the language in which it was filed.

☐ a translation of the international application into _____ which is the language of a translation furnished for the purposes of international search (Rules 12.3(a) and 23.1(b)).

b. ☐ This international search report has been established taking into account the **rectification of an obvious mistake** authorized by or notified to this Authority under Rule 91 (Rule 43.6*bis*(a)).

c. ☐ With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, see Box No. I.

2. ☐ **Certain claims were found unsearchable** (see Box No. II).

3. ☐ **Unity of invention is lacking** (see Box No. III).

4. With regard to the **title**,

☒ the text is approved as submitted by the applicant.

☐ the text has been established by this Authority to read as follows:

5. With regard to the **abstract**,

☒ the text is approved as submitted by the applicant.

☐ the text has been established, according to Rule 38.2(b), by this Authority as it appears in Box No. IV. The applicant may, within one month from the date of mailing of this international search report, submit comments to this Authority.

6. With regard to the **drawings**,

a. the figure of the drawings to be published with the abstract is Figure No. 1

☒ as suggested by the applicant.

☐ as selected by this Authority, because the applicant failed to suggest a figure.

☐ as selected by this Authority, because this figure better characterizes the invention.

b. ☐ none of the figures is to be published with the abstract.

Form PCT/ISA/210 (first sheet) (April 2007)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 08/53157

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06Q 30/00
USPC - 705/14
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8) - G06Q 30/00 (2008.04)
USPC - 705/14

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
705/1, 7, 8; 725/37, 38, 42

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PubWEST(USPT,PGPB,EPAB,JPAB); Google Scholar; Google
Search Terms Used: method to provide/select/display advertisement, advertisement, voice input, knowledge source, database, track, profile, user profile, advertisement database/source

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,696,965 A (DEDRICK) 09 December 1997 (09.12.1997) entire document, especially col 4, ln 60-66, col 5, ln 10-20, col 6, ln 12-30 and col 8, ln 24 - col 9, ln 10.	1-20

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 April 2008 (24.04.2008)	22 MAY 2008
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (April 2007)

PATENT COOPERATION TREATY

PCT

INTERNATIONAL SEARCH REPORT

(PCT Article 18 and Rules 43 and 44)

Applicant's or agent's file reference 90611-365986	FOR FURTHER ACTION	see Form PCT/ISA/220 as well as, where applicable, item 5 below.
International application No. PCT/US 07/81481	International filing date *(day/month/year)* 16 October 2007 (16.10.2007)	(Earliest) Priority Date *(day/month/year)* 16 October 2006 (16.10.2006)
Applicant VOICEBOX TECHNOLOGIES, INC.		

This international search report has been prepared by this International Searching Authority and is transmitted to the applicant according to Article 18. A copy is being transmitted to the International Bureau.

This international search report consists of a total of 2 sheets.

☐ It is also accompanied by a copy of each prior art document cited in this report.

1. **Basis of the report**

a. With regard to the language, the international search was carried out on the basis of:

☒ the international application in the language in which it was filed.

☐ a translation of the international application into _____ which is the language of a translation furnished for the purposes of international search (Rules 12.3(a) and 23.1(b)).

b. ☐ This international search report has been established taking into account the rectification of an obvious mistake authorized by or notified to this Authority under Rule 91 (Rule 43.6*bis*(a)).

c. ☐ With regard to any nucleotide and/or amino acid sequence disclosed in the international application, see Box No. I.

2. ☐ Certain claims were found unsearchable (see Box No. II).

3. ☐ Unity of invention is lacking (see Box No. III).

4. With regard to the title,

☒ the text is approved as submitted by the applicant.

☐ the text has been established by this Authority to read as follows:

5. With regard to the abstract,

☒ the text is approved as submitted by the applicant.

☐ the text has been established, according to Rule 38.2(b), by this Authority as it appears in Box No. IV. The applicant may, within one month from the date of mailing of this international search report, submit comments to this Authority.

6. With regard to the drawings,

a. the figure of the drawings to be published with the abstract is Figure No. 1

☒ as suggested by the applicant.

☐ as selected by this Authority, because the applicant failed to suggest a figure.

☐ as selected by this Authority, because this figure better characterizes the invention.

b. ☐ none of the figures is to be published with the abstract.

Form PCT/ISA/210 (first sheet) (April 2007)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 07/81481

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 17/28 (2008.04)
USPC - 704/4
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC: 704/4

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
704/1-4;709/204,205,207

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Google; Google Scholar; Google Patents; PubWest(PGPB,USPT,USOC,EPAB,JPAB);
Search Terms Used: voice, user, interface, cooperative conversational, human, machine, voice recognition, free form, natural, intuitive, utterance, shared knowledge, update, response, intended meaning, context, time, infer

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2004/0199375 A1 (EHSANI et al.) 07 October 2004 (07.10.2004), (para [0002], [0168]-[0169],[0210]-[0248])	1-42
Y	US 2005/0021334 A1 (IWAHASHI) 27 January 2005 (27.01.2005), (para [0042]-[0077])	1-42
A	US 5,377,350 A (SKINNER) 27 December 1994 (27.12.1994)	1-42

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 September 2008 (24.09.2008)	01 OCT 2008
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (April 2007)

PATENT COOPERATION TREATY

PCT

INTERNATIONAL SEARCH REPORT

(PCT Article 18 and Rules 43 and 44)

Applicant's or agent's file reference 90611-371246	FOR FURTHER ACTION	see Form PCT/ISA/220 as well as, where applicable, item 5 below.
International application No. PCT/US 08/69508	International filing date *(day/month/year)* 09 July 2008 (09.07.2008)	(Earliest) Priority Date *(day/month/year)* 11 December 2007 (11.12.2007)
Applicant VOICEBOX TECHNOLOGIES, INC.		

This international search report has been prepared by this International Searching Authority and is transmitted to the applicant according to Article 18. A copy is being transmitted to the International Bureau.

This international search report consists of a total of 2 sheets.

☐ It is also accompanied by a copy of each prior art document cited in this report.

1. Basis of the report
   a. With regard to the language, the international search was carried out on the basis of:
      ☒ the international application in the language in which it was filed.
      ☐ a translation of the international application into _____ which is the language of a translation furnished for the purposes of international search (Rules 12.3(a) and 23.1(b)).
   b. ☐ This international search report has been established taking into account the rectification of an obvious mistake authorized by or notified to this Authority under Rule 91 (Rule 43.6*bis*(a)).
   c. ☐ With regard to any nucleotide and/or amino acid sequence disclosed in the international application, see Box No. I.

2. ☐ Certain claims were found unsearchable (see Box No. II).

3. ☐ Unity of invention is lacking (see Box No. III).

4. With regard to the title,
   ☐ the text is approved as submitted by the applicant.
   ☒ the text has been established by this Authority to read as follows:

PROVIDING A NATURAL LANGUAGE VOICE USER INTERFACE IN AN INTEGRATED VOICE NAVIGATION SERVICES ENVIRONMENT

5. With regard to the abstract,
   ☒ the text is approved as submitted by the applicant.
   ☐ the text has been established, according to Rule 38.2(b), by this Authority as it appears in Box No. IV. The applicant may, within one month from the date of mailing of this international search report, submit comments to this Authority.

6. With regard to the drawings,
   a. the figure of the drawings to be published with the abstract is Figure No. 1
      ☒ as suggested by the applicant.
      ☐ as selected by this Authority, because the applicant failed to suggest a figure.
      ☐ as selected by this Authority, because this figure better characterizes the invention.
   b. ☐ none of the figures is to be published with the abstract.

Form PCT/ISA/210 (first sheet) (April 2007)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 08/69508

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 17/27 (2008.04)
USPC - 704/9
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC - 704/9

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC - 704/1, 2, 8, 9, 200, 206, 221, E15.045

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PubWEST (PGPB, USPT, EPAB, JPAB); Google Scholar
Search Terms Used: map, gps, speech, voice, talk, navigation, natural, conversational, grammar, interpret, software, computer, input, interface, user interface, refine, zoom, route, directions

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X -- Y	US 2007/0050191 A1 (WEIDER et al.) 01 March 2007 (01.03.2007), entire document, especially paras. [0061]-[0073], [0100], [0121], [0123], [0126], [0133], [0140], [0142], [0147], [0153], [0167], [0189]-[0194], [0210]-[0211], [0215]	1-42, 45-58 ---------- 43-44
Y	US 5,902,347 A (BACKMAN et al.) 11 May 1999 (11.05.1999), entire document, especially col 1 ln 32-34, col 4 ln 33-44	43-44
A	US 5,274,560 A (LARUE) 28 December 1993 (28.12.1993), entire document	1-58
A	US 6,598,018 B1 (JUNQUA) 22 July 2003 (22.07.2003), entire document	1-58

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 September 2008 (25.09.2008)	02 OCT 2008
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (April 2007)

PATENT COOPERATION TREATY

PCT

INTERNATIONAL SEARCH REPORT

(PCT Article 18 and Rules 43 and 44)

Applicant's or agent's file reference 90611-370865	FOR FURTHER ACTION	see Form PCT/ISA/220 as well as, where applicable, item 5 below.
International application No. PCT/US 08/69524	International filing date *(day/month/year)* 09 July 2008 (09.07.2008)	(Earliest) Priority Date *(day/month/year)* 27 May 2008 (27.05.2008)
Applicant VOICEBOX TECHNOLOGIES, INC.		

This international search report has been prepared by this International Searching Authority and is transmitted to the applicant according to Article 18. A copy is being transmitted to the International Bureau.

This international search report consists of a total of 2 sheets.

☐ It is also accompanied by a copy of each prior art document cited in this report.

1. Basis of the report
   a. With regard to the language, the international search was carried out on the basis of:
      ☒ the international application in the language in which it was filed.
      ☐ a translation of the international application into _____ which is the language of a translation furnished for the purposes of international search (Rules 12.3(a) and 23.1(b)).
   b. ☐ This international search report has been established taking into account the **rectification of an obvious mistake** authorized by or notified to this Authority under Rule 91 (Rule 43.6*bis*(a)).
   c. ☐ With regard to any **nucleotide and/or amino acid sequence** disclosed in the international application, see Box No. I.

2. ☐ Certain claims were found unsearchable (see Box No. II).

3. ☐ Unity of invention is lacking (see Box No. III).

4. With regard to the title,
   ☒ the text is approved as submitted by the applicant.
   ☐ the text has been established by this Authority to read as follows:

5. With regard to the abstract,
   ☒ the text is approved as submitted by the applicant.
   ☐ the text has been established, according to Rule 38.2(b), by this Authority as it appears in Box No. IV. The applicant may, within one month from the date of mailing of this international search report, submit comments to this Authority.

6. With regard to the drawings,
   a. the figure of the drawings to be published with the abstract is Figure No. 1
      ☒ as suggested by the applicant.
      ☐ as selected by this Authority, because the applicant failed to suggest a figure.
      ☐ as selected by this Authority, because this figure better characterizes the invention.
   b. ☐ none of the figures is to be published with the abstract.

Form PCT/ISA/210 (first sheet) (April 2007)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 08/69524

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 13/12 (2008.04)
USPC - 710/65
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC: 710/65

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 710/11, 14, 37, 50, 73; 370/353, 458, 448; 379/907-908; 704/200, 246, 251-252 (see text terms below)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PubWEST (PGPB, USPT, USOC, EPAB, JPAB); DialogPRO; Google (incl. Patents, Scholar) - natural language, multi-modal, central, input, device, client, server, process, distributed, local, distributed system, local, context, speech

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 7,376,645 B2 (BERNARD) 20 May 2008 (20.05.2008), Abstract, FIGS.1-3; and Col 1, ln 19-30; Col 1, ln 35 to Col 2, ln 2; Col 2, ln 21 to Col 4, ln 16; Col 4, ln 21-56, 60; Col 5, ln 10; Col 6, ln 1-5, 7, 22-23; Col 6, ln 37 to Col 7, ln 4; Col 7, ln 13-26; Col 7, ln 52 to Col 10, ln 37	1-27
A	US 2004/0025115 A1 (SIENEL et al.) 5 February 2004 (05.02.2004)	1-27
A	US 2007/0050191 A1 (WEIDER et al.) 1 March 2007 (01.03.2007)	1-27
A	US 2004/0166832 A1 (PORTMAN et al.) 26 August 2004 (26.08.2004)	1-27
A	US 6,615,172 B1 (BENNETT et al.) 2 September 2003 (02.09.2003)	1-27

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
04 October 2008 (04.10.2008)	09 OCT 2008
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (April 2007)

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	A	5,892,813	04/06/1999	Morin *et al.*	379	88.01	
	B	6,721,001	04/13/2004	Berstis	348	231.3	
	C	6,742,021	05/25/2004	Halverson *et al.*	709	218	
	D	6,990,513	01/24/2006	Belfiore *et al.*	709	203	
	E	7,493,559	02/17/2009	Wolff *et al.*	715	727	
	F	2005/0033574	02/10/2005	Kim *et al.*	704	251	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

401497563v1

Electronic Acknowledgement Receipt

EFS ID:	6209526
Application Number:	11580926
International Application Number:	
Confirmation Number:	6670
Title of Invention:	System and method for a cooperative conversational voice user interface
First Named Inventor/Applicant Name:	Larry Baldwin
Customer Number:	00909
Filer:	Syed Jafar Ali/Susan Trader
Filer Authorized By:	Syed Jafar Ali
Attorney Docket Number:	090611-0357945
Receipt Date:	06-OCT-2009
Filing Date:	16-OCT-2006
Time Stamp:	14:04:55
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		0357945_SIDS.pdf	84883 7f9ff1b682b89c0f1ad606b7c062ec1e005aa953	yes	3

	Multipart Description/PDF files in .zip description		
	Document Description	**Start**	**End**
	Transmittal Letter	1	2
	Information Disclosure Statement (IDS) Filed (SB/08)	3	3
Warnings:			
Information:			
	Total Files Size (in bytes):	84883	

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111
If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371
If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office
If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANTS :	Larry BALDWIN *et al.*	CONFIRMATION NO.:	6670
SERIAL NUMBER :	11/580,926	EXAMINER:	Eric L. Yen
FILING DATE :	October 16, 2006	ART UNIT:	2626
FOR :	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE		

SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT

Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313-1450

Sir:

Pursuant to the duty of disclosure under 37 C.F.R. §§ 1.56, 1.97 and 1.98, Applicants hereby make of record in the above-identified application the documents listed on the attached Form PTO-1449.

Applicants reserve the right to establish the patentability of the claimed invention over any of the information provided herewith, and/or to prove that this information may not be prior art, and/or to prove that this information may not be enabling for the teachings purportedly offered.

This Supplemental Information Disclosure Statement is being filed within three months of the filing date of a national application other than a continued prosecution application under 37 C.F.R. §1.53(d), or within three months of the date of entry of the national stage as set forth in 37 C.F.R. §1.491 in an international application, or before the mailing date of a first Office

Action on the merits, or before the mailing date of a first Office Action after the filing of a request for continued examination under 37 C.F.R. §1.114. No certification or fee is required. 37 C.F.R. §1.97(b).

It is respectfully requested that the Examiner initial and return a copy of the enclosed PTO-1449.

Respectfully submitted,

PILLSBURY WINTHROP SHAW PITTMAN LLP

By _____
Syed Jafar Ali
Registration No.: 58,780

Customer No.
00909

P. O. Box 10500
McLean, Virginia 22102
Telephone: (703) 770-7900
Telefax: (703) 770-7901

Dated: October 6, 2009

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
11/580,926	10/16/2006	Larry Baldwin	090611-0357945	6670

909 7590 06/16/2010

PILLSBURY WINTHROP SHAW PITTMAN, LLP
P.O. BOX 10500
MCLEAN, VA 22102

EXAMINER	
YEN, ERIC L	

ART UNIT	PAPER NUMBER
2626	

MAIL DATE	DELIVERY MODE
06/16/2010	PAPER

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Office Action Summary	Application No.	Applicant(s)	
	11/580,926	BALDWIN ET AL.	
	Examiner	Art Unit	
	ERIC YEN	2626	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☐ Responsive to communication(s) filed on _____.

2a)☐ This action is **FINAL**. 2b)☒ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1-42* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1-42* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☒ The drawing(s) filed on *16 October 2006* is/are: a)☒ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☒ Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date _____.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____ .
5) ☐ Notice of Informal Patent Application
6) ☐ Other: _____.

DETAILED ACTION

Claim Rejections - 35 USC § 101

1. 35 U.S.C. 101 reads as follows:

> Whoever invents or discovers any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof, may obtain a patent therefor, subject to the conditions and requirements of this title.

Claims 1-28 are rejected under 35 U.S.C. 101 because the claimed invention is directed to non-statutory subject matter.

Claim 1 and its dependent claims:

are rejected under 35 U.S.C. 101 as not falling within one of the four statutory categories of invention. Supreme Court precedent1[1] and recent Federal Circuit decisions2[2] indicate that a statutory "process" under 35 U.S.C. 101 must (1) be tied to another statutory category (such as a particular apparatus), or (2) transform underlying subject matter (such as an article or material) to a different state or thing. While the instant claim(s) recite a series of steps or acts to be performed, the claim(s) neither transform underlying subject matter nor positively tie to another statutory category that accomplishes the claimed method steps, and therefore do not qualify as a statutory process.

While "voice user interface" is generally directed to man-machine interfaces, given its broadest reasonable interpretation the same steps in the method claim could

---

all be performed by a human communicating in person, where the human's brain stores updated knowledge and interacts/interfaces with a user via voice.

As per Claim 15 and its dependent claims:

Applicant does not define “computer readable medium” in the Specification which under current Office policy means that, given its broadest reasonable interpretation, it includes transitory computer-readable media (signals, carrier waves, etc.) which are non-statutory. An amendment to "non-transitory computer readable medium" is suggested to overcome the rejection.

Claim Rejections - 35 USC § 112

2. The following is a quotation of the first paragraph of 35 U.S.C. 112:

> The specification shall contain a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same and shall set forth the best mode contemplated by the inventor of carrying out his invention.

3. Claims 7-10, 21-24, 35-38 rejected under 35 U.S.C. 112, first paragraph, as failing to comply with the enablement requirement. The claim(s) contains subject matter which was not described in the specification in such a way as to enable one skilled in the art to which it pertains, or with which it is most nearly connected, to make and/or use the invention.

It is not clear what applicant means by “contextual” "signifier" and "qualifier". Applicant’s Specification treats these two terms as having a meaning that is common knowledge and “required to perform a task in an identified context”. While it is common

knowledge that signifiers signify things at least in the sense of indicating something meaningful and that qualifiers modify another term somehow (in the grammatical sense), there is nothing claimed or in the Specification describing what is being qualified/signified, what constitutes a signifier/qualifier (since they are both used in the Specification to describe information needed to complete a task), whether applicant is actually referring to a specific form of signifier/qualifier, whether signifiers/qualifiers are words in the utterances or other grammatical features (e.g. parts of speech), or how these entities (whatever they may be) are used to perform the function of completing a request.

Since the scope of these terms is unclear, the claims are vague and indefinite, and by consequence not enabled because one of ordinary skill in the art logically can't make and use something when he doesn't know what some of the parts of the invention are.

4.

1. The following is a quotation of the second paragraph of 35 U.S.C. 112:

> The specification shall conclude with one or more claims particularly pointing out and distinctly claiming the subject matter which the applicant regards as his invention.

2. Claims 1, 15, 29, and 7-10, 21-24, 35-38, are rejected under 35 U.S.C. 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

Applicant recites the term “cooperative” in the independent claims and provides applicant’s interpretation in the Specification (paragraph 2) as “information and/or

context is shared to advance mutual conversational goals", but this interpretation is ambiguous.

For example, the interpretation does not specify if someone (whether machine or human) is sharing the information/context during the course of the dialog, or if the information is merely "shared" in the sense that both machine and user know the same information (i.e. without one of the machine/user specifically imparting/sharing the information on the other entity).

The phrase "shared to advance mutual conversational goals" seems to imply an actual sharing/imparting by one to another because it grammatically describes an action as opposed to a description of the operating environment.

Applicant's description, however, seems to suggest that it is the other interpretation, and therefore, it is not clear what applicant means.

Since this paragraph may (though it conceivably may not) impact the claim scope (as an explicit definition of a term), the claims are rejected because the interpretation is unclear.

As per Claims 7-10, 21-24, 35-38:

It is not clear what applicant means by "contextual" "signifier" and "qualifier". Applicant's Specification treats these two terms as having a meaning that is common knowledge and "required to perform a task in an identified context". While it is common knowledge that signifiers signify things at least in the sense of indicating something meaningful and that qualifiers modify another term somehow (in the grammatical

sense), there is nothing claimed or in the Specification describing what is being qualified/signified, what constitutes a signifier/qualifier (since they are both used in the Specification to describe information needed to complete a task), whether applicant is actually referring to a specific form of signifier/qualifier, whether signifiers/qualifiers are words in the utterances or other grammatical features (e.g. parts of speech), or how these entities (whatever they may be) are used to perform the function of completing a request.

Since the scope of these terms is unclear, the claims are vague and indefinite.

Claim Rejections - 35 USC § 103

1. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

2. Claims 1, 4-11, 13-15, 18-25, 27-29, 32-39, 41-42 are rejected under 35 U.S.C. 103(a) as being unpatentable over Kargman (US 2005/0015256).

As per Claims 1, 15, 29, Kargman teaches a method (and corresponding computer-readable medium of Claim 15, and system of Claim 29) for providing a voice user interface that generates cooperative conversational responses to user utterances ("customer repeatedly orders… initial or supplemental prompt… individually tailored", paragraph 50; where the consistent ordering of a particular item is based on information

shared between the user and system [i.e. what does the customer frequently order] to advance the mutual goal of taking down the customer's food order [see applicant's interpretation of "cooperative" in Specification, paragraph 2])

the responses being based on shared knowledge about a current conversation and/or one or more past conversations ("order history", paragraph 50; where prior orders are past conversations between the system and customer, and also the system and user both know that the user is ordering food in a current conversation], which is "knowledge about a current conversation" that the responses are tailored to; "mushroom… shrimp", paragraph 48; "whether they wish to order 'pizza'", paragraph 44; where the statistical response processing is done for a current pizza ordering part, where paragraph 44 describes a series of steps, the initial step of selecting pizza/side/drink not requiring the statistical/prior-order information, and so the response is based on knowledge about a current conversation because the system knows it's at a stage where it needs the history information)

the method comprising:

receiving an input, the input including at least one utterance ("calls to place an order", paragraph 50; "speaking the word 'pizza", paragraph 44; "statistically… mushroom… shrimp", paragraph 48; speaking is uttering an input)

updating the shared knowledge to include knowledge about the at least one utterance ("prior orders", paragraph 50; "speaking the word 'pizza", paragraph 44; "statistically… mushroom… shrimp", paragraph 47-48; where in Kargman the example for where statistical information is at least obviously accumulated and updated over time

[since if people start ordering shrimp consistently it would be frustrating to the users to be constantly overridden by mushroom], and so the statistics in the statistical database were updated shared knowledge [whether shared between system and user, since both a user and system know that people don't really order pizzas with seafood very frequently or shared in the sense that it was collected from among a large number of people]. This information is information about the at least one utterance because it contains the disambiguation for determining the meaning of the utterance)

determining an intended meaning for the at least one utterance based on the updated shared knowledge ("prior orders", paragraph 50; "speaking the word 'pizza", paragraph 44; "statistically… mushroom… shrimp", paragraph 48; "statistical database", paragraph 47; where the statistics updated at some point in time with statistical information [updated shared knowledge] about mushrooms and shrimp, for example, are used to determine that the user meant mushroom [intended meaning] instead of shrimp [less likely meaning])

generating a response based on the determined intended meaning ("speak the order back… for their confirmation", paragraph 49; where speaking the order back is a system response and depends on what the system interpreted the order to be [determined intended meaning]).

As per Claims 4, 18, 32, Kargman teaches wherein updating the shared knowledge includes updating one or more long-term profiles based on information about the at least one utterance ("statistical database", paragraph 47; "prior orders",

paragraph 50; where prior order information was updated at som“ point and belongs to a long-term customer profile).

As per Claims 5, 19, 33, Kargman teaches/suggests wherein determining the intended meaning includes identifying a conversation type for the at least one utterance (“pizza”, paragraph 44; where the selection of pizza defines a conversation type to be a pizza conversation instead of a side item conversation; “mushroom”, paragraph 48; where the determination of pizza at least obviously affects which statistics the system looks up to determine the intended meaning to be mushroom instead of shrimp [e.g. there would likely be a different set of statistics for side items since people probably order popcorn shrimp more than they order mushrooms as a side])

the conversation type based on one or more of a goal for the utterance, one or more participant roles for the utterance, and/or an allocation of information for the utterance (“mushroom”, paragraph 48; where goal of ordering a pizza with a topping is one of the listed items [“one or more”]).

As per Claims 6, 20, 34, Kargman teaches wherein determining the intended meaning includes generating at least one hypothesis of an intended meaning for the at least one utterance, the at least one hypothesis including a degree of certainty (“mushroom… shrimp”, paragraph 48; “large thin crust pepperoni pizza”, paragraph 46; where mushroom and shrimp are two hypotheses and mushroom is statistically more

likely/certain, and it is at least obvious that the system eventually figures out precisely what the user wants in the example).

As per Claims 7, 21, 35, Kargman teaches/suggests wherein generating the at least one hypothesis includes identifying contextual signifiers and/or qualifiers in the at least one utterance (“one… large… thin…pizza”, paragraph 46; where, for example, pizza signifies that the user wants a pizza and so can be interpreted as a signifier, and large/thin/etc. "qualify"/limits what kind of pizza the user wants it to be, and it is obvious that all of this information is eventually incorporated into the order hypothesized by the system, and these all apply to the pizza context and so are contextual in that sense).

As per Claims 8, 22, 36, Kargman suggests inferring missing contextual signifiers and/or qualifiers for the at least one utterance based on the shared knowledge ("pizza, drinks or side items", paragraph 44; "repeatedly orders a large pepperoni pizza... individually tailored", paragraph 50; where the initial tailored prompt is presented at a time before the user fills in the "blanks” of toppings-signifier/size-qualifier/etc. [missing contextual signifiers/qualifiers] and is based on the shared knowledge of the customer repeatedly ordering the same thing [i.e. at the very least identifying a specific customer makes the system infer all of the missing, or not-yet-provided information needed for the request for food]).

As per Claims 9, 23, 37, Kargman suggests wherein the degree of certainty for the at least one hypothesis is based on one or more of the identified contextual signifiers and/or qualifiers, the inferred contextual signifiers and/or qualifiers, and/or contextual signifiers and/or qualifiers necessary for generating a response (“mushroom”, paragraph 48; “mushroom… shrimp”, paragraph 48; “large thin crust pepperoni pizza”, paragraph 46; where, in a situation where there are no confusable words like in “large pepperoni pizza”, the system is certain [degree of certainty] that the hypothesis is correct because the qualifier/size and topping/item/signifier cannot be confused like the words shrimp/mushroom. where, in an at least obvious example, requesting “a large mushroom pizza” would be less statistically certain [since shrimp is only “much less likely” instead of “not possible” than “large pepperoni pizza” which does not have a confusable word [and therefore is dead certain]. "Mushroom" signifies what kind of pizza the customer wants. This concept at least obviously can apply to qualifiers like size as well because adjectives have phonetically similar words too [e.g. it is more certain that someone said "medium" which doesn't have a confusable size word than the certainty of “small” which can be confused with the word “tall”, where the confusion would be far more disastrous when someone is ordering “a tall/small glass of water"]. It is also obvious that there is a situation where a confusable adjective and noun both exist in the same utterance. Basing the certainty on identified signifiers/qualifiers is one of the items in the claimed list).

As per Claims 10, 24, 38, Kargman teaches/suggests wherein the one or more processors are operable to generate the response by adapting contextual signifiers and/or grammar in the response to reflect contextual signifiers and/or grammar in the at least one utterance ("speak the order back", paragraph 49; "one large pepperoni pizza, thin crust", paragraph 46; where at the very least speaking/repeating the order "one large pepperoni pizza, thin crust" back to the user reflects the actual items like pizza and pepperoni which signify what the order is in the pizza context [contextual signifiers] and also the grammar characteristics of the input order that make it grammatically correct [more clearly so in an example where the spoken input is "one large thin [crust] pizza]).

As per Claims 11, 25, 39, Kargman teaches wherein generating the response includes framing the response to influence a subsequent utterance ("confirmation", paragraph 49; where prompting for a confirmation influences the user to either confirm or deny the order).

As per Claims 13, 27, 41, Kargman teaches wherein the input is a multi-modal input ("address", paragraph 6; "buttons… user may press", paragraph 32; "speaking", paragraph 44; where applicant claims that the input includes at least one utterance and therefore does not limit it to only one utterance in particular, and since applicant does not limit the input to an input at any particular instantaneous time nor does applicant limit the input to the input for a single dialog turn which is a combination of multiple

mode inputs, the collective information of address [at least obviously typed in via display] and spoken pizza ordering information over time constitute a collective input to the system for completing an order for that particular dialog. Since they are entered in different modes [manual and spoken] the input made of all information entered manually and by speech over the course of a dialogue is therefore multi-modal).

As per Claim 14, 28, 42, Kargman teaches/suggests wherein the response is a multi-modal response ("displayed to permit the customer to confirm", Abstract; "display… pizza or food item ordered", paragraph 32; "speak the order back… for their confirmation", paragraph 49; where both the speaking audibly by the system and the display are described as being used for confirmation and so Kargman at least obviously suggests the use of both display and audible output [multiple modes of communication] to communicate with the user).

3.

4. Claims 2-3, 16-17, 30-31, are rejected under 35 U.S.C. 103(a) as being unpatentable over Kargman, as applied to claims 1, 15, and 29, above, and further in view of Morin et al. (US 5,748,841), hereafter Morin.

As per Claim 2, 16, 30, Kargman fails to teach wherein updating the shared knowledge includes populating a short-term context stack with information about the at least one utterance.

Morin teaches/suggests populating a short-term context stack with information about the at least one utterance ("dialogue being run…dialogue history… dialogue turns or exchanges already made and the context in which they occur", col. 3, lines 6-24; "immediate context", col. 15, lines 37-47; "knowledge of the ongoing dialogue... dialogue history and immediate dialogue", col. 4, lines 5-16; "short-term... immediate focus of the dialogue", col. 10, lines 1-15; where having an immediate context means the system at least obviously stores information about an immediate/short-term context at some point, and this memory location can be interpreted as a short-term context memory/stack [where stack is, in computer contexts, a memory location populated by information to be processed]; as described in Morin, the context and dialog history is derived from utterances in the current dialog).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of invention to **use a known technique to improve similar devices in the same way** because Kargman teaches a **base** machine-user dialog device **upon which the claimed** contextual stack **can be seen as an improvement** by incorporating more information into facilitating a dialog. Morin teaches **a comparable** machine-user dialog device **that has been improved in the same way as the claimed invention** by having a short-term context [immediate context in Morin] to facilitate the dialog. **One of ordinary skill in the art could have applied the known improvement technique in the same way to the base device** by adding a contextual information/dialogue history to the information already used in Kargman to facilitate improved dialog using information derived from past communications, **to obtain the predictable results** of a

machine–user dialog system that uses information from past utterances to determine future dialog behavior.

As per Claims 3, 17, 31, Kargman fails to teach expiring the information about the at least one utterance in the short-term context stack after a predetermined amount of time.

Morin teaches/suggests expiring the information about the at least one utterance in the short-term context stack after a predetermined amount of time (“dialogue being run…dialogue history… dialogue turns or exchanges already made and the context in which they occur”, col. 3, lines 6-24; “immediate context”, col. 15, lines 37-47; “knowledge of the ongoing dialogue... dialogue history and immediate dialogue", col. 4, lines 5-16; "short-term... immediate focus of the dialogue”, col. 10, lines 1-15; where the predetermined time is the amount of time it takes, once the system realizes the immediate context changed, to change the previous immediate context [short-term context stack] to the new immediate context; “immediate context refers to what is logically expected from the user at this time”, col. 3, lines 6-24; “time-out”, col. 9, lines 34-57; where the time-out at least obviously changes the context and one obvious result of a timeout is termination [e.g. if the user just left], which would expire the context and all other current-dialog information upon termination).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of invention to **use a known technique to improve similar devices in the same way** because Kargman teaches a **base** machine-user dialog device **upon which the**

claimed contextual stack **can be seen as an improvement** by incorporating more information into facilitating a dialog. Morin teaches **a comparable** machine-user dialog device **that has been improved in the same way as the claimed invention** by having a short-term context [immediate context in Morin] to facilitate the dialog. **One of ordinary skill in the art could have applied the known improvement technique in the same way to the base device** by adding a contextual information/dialogue history to the information already used in Kargman to facilitate improved dialog using information derived from past communications, **to obtain the predictable results** of a machine–user dialog system that uses information from past utterances to determine future dialog behavior.

5. Claims 12, 26, 40 are rejected under 35 U.S.C. 103(a) as being unpatentable over Kargman, as applied to Claims 1, 15, and 29, above, and further in view of Bielby et al. (US 5,488,652), hereafter Bielby.

As per Claim 12, 26, 40, Kargman fails to teach learning that the determined intended meaning is incorrect, further updating the shared knowledge to include knowledge that the determined intended meaning is incorrect, and determining a second intended meaning for the at least one utterance based on the further updated shared knowledge, wherein the incorrect determined intended meaning is excluded as a possible second intended meaning.

Bielby suggests learning that the determined intended meaning is incorrect (“called entity name corresponding to a frequently called listing is recognized", col. 18, lines 1-30; “recognized called entity name is incorrect”, col. 18, lines 31-44; where the intended meaning of a particular name is noted as incorrect)

further updating the shared knowledge to include knowledge that the determined intended meaning is incorrect (“called entity name corresponding to a frequently called listing is recognized", col. 18, lines 1-30; “announce the next best candidate”, col. 18, lines 31-44; where “shared knowledge”, as claimed, is just generally any information the system has that the user has as well, and so the system learning that one interpretation is wrong updates the systems knowledge to share the user’s knowledge that the recognition was incorrectly performed)

and determining a second intended meaning for the at least one utterance based on the further updated shared knowledge (“called entity name corresponding to a frequently called listing is recognized", col. 18, lines 1-30; “recognized called entity name is incorrect”, col. 18, lines 31-44; where moving on to a next best candidate [second intended meaning] is based on the knowledge that the first candidate was wrong)

wherein the incorrect determined intended meaning is excluded as a possible second intended meaning (“called entity name corresponding to a frequently called listing is recognized", col. 18, lines 1-30; “recognized called entity name is incorrect”, col. 18, lines 31-44; “next” implies that the previous interpretation was excluded and is obviously so since it was already indicated as wrong).

Therefore, it would have been obvious to one of ordinary skill in the art to **use a known technique of** moving onto the next possible candidate **to improve similar devices** because Kargman teaches a base speech recognition device that can have multiple interpretations (shrimp and mushroom, as discussed above) **upon which the claimed** determination of a second interpretation **could be seen as an improvement** by allowing the system to adjust for recognition errors. Bielby **contained a comparable** user-machine interaction device where multiple interpretations of input speech can be made **and has been improved in the same way as the claimed invention** by incorporating a process to accommodate recognition-errors. **One of ordinary skill in the art could have applied the known 'improvement technique' in the same way to the 'base' device** of Kargman by asking the user if he/she meant shrimp if he/she says the system did not recognize mushroom properly, for example, **which would obtain the predictable results** of a machine-user voice interactive system which has a process that can adapt for recognition errors as explicitly described in Bielby.

Double Patenting

6. The nonstatutory double patenting rejection is based on a judicially created doctrine grounded in public policy (a policy reflected in the statute) so as to prevent the unjustified or improper timewise extension of the "right to exclude" granted by a patent and to prevent possible harassment by multiple assignees. A nonstatutory obviousness-type double patenting rejection is appropriate where the conflicting claims

are not identical, but at least one examined application claim is not patentably distinct from the reference claim(s) because the examined application claim is either anticipated by, or would have been obvious over, the reference claim(s). See, e.g., *In re Berg*, 140 F.3d 1428, 46 USPQ2d 1226 (Fed. Cir. 1998); *In re Goodman*, 11 F.3d 1046, 29 USPQ2d 2010 (Fed. Cir. 1993); *In re Longi*, 759 F.2d 887, 225 USPQ 645 (Fed. Cir. 1985); *In re Van Ornum*, 686 F.2d 937, 214 USPQ 761 (CCPA 1982); *In re Vogel*, 422 F.2d 438, 164 USPQ 619 (CCPA 1970); and *In re Thorington*, 418 F.2d 528, 163 USPQ 644 (CCPA 1969).

A timely filed terminal disclaimer in compliance with 37 CFR 1.321(c) or 1.321(d) may be used to overcome an actual or provisional rejection based on a nonstatutory double patenting ground provided the conflicting application or patent either is shown to be commonly owned with this application, or claims an invention made as a result of activities undertaken within the scope of a joint research agreement.

Effective January 1, 1994, a registered attorney or agent of record may sign a terminal disclaimer. A terminal disclaimer signed by the assignee must fully comply with 37 CFR 3.73(b).

7. Claims 1, 12, 15, 26, 29, 40, rejected on the ground of nonstatutory obviousness-type double patenting as being unpatentable over claim 1 of U.S. Patent No.7,620,549. Although the conflicting claims are not identical, they are not patentably distinct from each other because they are both directed to systems that adapt user-machine voice-interface systems based on past interactions including tuning [i.e. updating shared

information] based on previous interactions where the system is determining the interpretation/meaning of the spoken input.

8. Claims 1, 15, and 29, are provisionally rejected on the ground of nonstatutory obviousness-type double patenting as being unpatentable over claim 1 of copending Application No. 11/954,064. Although the conflicting claims are not identical, they are not patentably distinct from each other because they are both directed to interacting with a user using "shared knowledge" to facilitate the dialog.

9. Claims 1, 13, 15, 27, 29, 41, are provisionally rejected on the ground of nonstatutory obviousness-type double patenting as being unpatentable over claim 1 and 6 of copending Application No. 12/127,343. Although the conflicting claims are not identical, they are not patentably distinct from each other because they are both directed to interacting with a user using "shared knowledge" to facilitate the dialog with multi-modal input.

10. Claims 1, 15, 29, are provisionally rejected on the ground of nonstatutory obviousness-type double patenting as being unpatentable over claim 69, 84 of copending Application No. 12/705,311. Although the conflicting claims are not identical, they are not patentably distinct from each other because they are both directed to interacting with a user using shared updated knowledge [from current/prior dialogs] to determine meaning.

This is a provisional obviousness-type double patenting rejection because the conflicting claims have not in fact been patented.

Conclusion

11. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure. See PTO-892.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to ERIC YEN whose telephone number is (571)272-4249. The examiner can normally be reached on M-F 7:30-4:00.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Richemond Dorvil can be reached on 571-272-7602. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

EY 6/15/10
/Eric Yen/
Examiner, Art Unit 2626

Notice of References Cited	Application/Control No. 11/580,926	Applicant(s)/Patent Under Reexamination BALDWIN ET AL.	
	Examiner ERIC YEN	Art Unit 2626	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
*	A	US-2005/0015256	01-2005	Kargman, James B.	704/272
*	B	US-5,748,841	05-1998	Morin et al.	704/257
*	C	US-5,488,652	01-1996	Bielby et al.	379/88.03
*	D	US-2003/0097249	05-2003	Walker et al.	704/1
*	E	US-7,487,088	02-2009	Gorin et al.	704/240
*	F	US-6,374,214	04-2002	Friedland et al.	704/235
*	G	US-6,154,526	11-2000	Dahlke et al.	379/88.03
	H	US-			
	I	US-			
	J	US-			
	K	US-			
	L	US-			
	M	US-			

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

BIB DATA SHEET

CONFIRMATION NO. 6670

SERIAL NUMBER	FILING or 371(c) DATE	CLASS	GROUP ART UNIT	ATTORNEY DOCKET NO.
11/580,926	10/16/2006 RULE	704	2626	090611-0357945

APPLICANTS

Larry Baldwin, Maple Valley, WA;
Tom Freeman, Mercer Island, WA;
Michael Tjalve, Bellevue, WA;
Blane Ebersold, Seattle, WA;
Chris Weider, Everett, WA;

**** CONTINUING DATA ***************************

**** FOREIGN APPLICATIONS ***************************

**** IF REQUIRED, FOREIGN FILING LICENSE GRANTED **** ** SMALL ENTITY **
11/07/2006

Foreign Priority claimed ☐ Yes ☑ No 35 USC 119(a-d) conditions met ☐ Yes ☑ No Verified and Acknowledged /ERIC L YEN/ Examiner's Signature	☐ Met after Allowance Initials	STATE OR COUNTRY	SHEETS DRAWINGS	TOTAL CLAIMS	INDEPENDENT CLAIMS
		WA	3	42	3

ADDRESS

PILLSBURY WINTHROP SHAW PITTMAN, LLP
P.O. BOX 10500
MCLEAN, VA 22102
UNITED STATES

TITLE

System and method for a cooperative conversational voice user interface

FILING FEE RECEIVED		
1115	FEES: Authority has been given in Paper No._____ to charge/credit DEPOSIT ACCOUNT No._____ for following:	☐ All Fees ☐ 1.16 Fees (Filing) ☐ 1.17 Fees (Processing Ext. of time) ☐ 1.18 Fees (Issue) ☐ Other _____ ☐ Credit

BIB (Rev. 05/07).

EAST Search History

EAST Search History (Prior Art)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
L1	7	"20050015256".pn. "5748841".pn. "5488652".pn. "20030097249".pn. "7487088".pn. "6374215".pn. "6154526".pn.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 16:05
S12	11	704/270-275.ccls. and (discourse adj history)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 08:52
S13	139	704/270-275.ccls. and (dialog same history)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 08:56
S14	97	704/270-275.ccls. and (dialog$1 near5 history)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 08:56
S15	1	10/856234	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 09:05
S16	4	11/580926 and cooperat$4	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 09:11
S17	4	11/580926	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 11:00
S18	3	((larry and baldwin) (tom and freeman) (michael and tjalve) (blane and ebersold) (chris and weider)).in. and (cooperative (shared adj knowledge $1)).clm.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 11:02

S19	14	((larry and baldwin) (tom and freeman) (michael and tjalve) (blane and ebersold) (chris and weider)).in. and (updat$4).clm.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 11:10
S20	4	11/580926 and medium	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 11:15
S21	1	"7487088".pn.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 11:38
S22	2	(discourse adj history) same time$1	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 11:38
S23	43	(discourse adj history)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 11:40
S24	5	(discourse adj history) and expir$4	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 11:40
S25	7	(discourse adj history) and gorin.in.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 11:41
S26	10	(dialog$4 adj history) and gorin.in.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 11:45
S27	217	(dialog$4 adj history)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 11:53
S28	148	(dialog$4 adj history) and "704".clas.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 11:53
S29	56	(context$4 same (dialog$4 adj history)) and "704".clas.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 11:53

S30	1	"5748841".pn. and short-term	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 12:09
S31	0	short-term-context and time	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 12:25
S32	0	short-term-context and expir$4	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 12:25
S33	0	short-term-context	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 12:25
S34	7	short-term adj context	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 12:25
S35	7	(short-term adj context) and time	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 12:25
S36	1	10/856234 and display $4	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 12:41
S37	1	"5748841".pn. and multi-modal	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 12:50
S38	0	704/270.ccls. and (multi-modal same this same here)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 12:52
S39	0	704/270-275.ccls. and (multi-modal same this same here)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 12:52
S40	45	704/270-275.ccls. and (multi-modal same order)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 12:52

S41	12	((dialog$4 discourse) adj history) same statistic$4	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:10
S42	0	((dialog$4 discourse) adj history) same (updat$4 near5 statistic $4)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:13
S43	9	((dialog$4 discourse) adj history) and (updat $4 near5 statistic$4)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:13
S44	0	((dialog$4 discourse) adj history) and (confus$4 near5 statistic$4)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:19
S45	0	((dialog$4 discourse) adj history) and (confus$6 same statistic$4)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:19
S46	27	704/270-275.ccls. and (confus$6 same statistic$4)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:19
S47	12	704/270-275.ccls. and (misrecogni$4 same statistic$4)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:21
S48	1	704/270-275.ccls. and ((next adj candidate $1) same (incorrect$2 misrecogniz$4 misunderst$4))	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:29
S49	2	704/270-275.ccls. and ((next adj3 candidate $1) same (incorrect$2 misrecogniz$4 misunderst$4))	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:29
S50	14	"704".clas. and ((next adj3 candidate$1) same (incorrect$2 misrecogniz$4 misunderst$4))	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:31

S51	21	"704".clas. and ((next adj3 best) same (incorrect$2 misrecogniz$4 misunderst$4))	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:43
S52	18	("704".clas. and ((next adj3 best) same (incorrect$2 misrecogniz$4 misunderst$4))) not S50	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:43
S53	13	"704".clas. and (misrecogni$4 near5 updat$4)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:46
S54	0	"704".clas. and ((misrecogni$4 near5 updat$4) with next)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:48
S55	0	"704".clas. and ((misrecogni$4 near5 updat$4) same next)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:48
S56	1	"6374215".pn.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:58
S57	1	"6734215".pn.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:58
S58	1	S21	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:58
S59	1	"6374251".pn.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:58
S60	1	"6374215".pn.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 13:58
S61	0	"704".clas. and ((next adj meaning$1) same updat$4)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 14:02

S63	1	"704".clas. and (((next adj3 (best candidate$1 alternative$1 interpretation$1 meaning$1)) same (misrecogni$4 error $1)) same updat$4)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 14:04
S64	32	"704".clas. and ((next adj3 (best candidate $1)) and re-recogni$4)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 14:05
S65	22	"704".clas. and ((next adj3 (best candidate $1)) and (dialog$4 near5 history))	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 14:09
S66	1	"704".clas. and ((next near5 (error$1 misrecogni$4)) and (dialog$4 near5 history))	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 14:12
S67	0	"704".clas. and ((next near5 (error$1 misrecogni$4)) and (discourse near5 history))	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 14:12
S68	3	11/580926 and (qualifier$1 signifier $1)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 14:40
S69	1	"704".clas. and (signifier$1 same qualifier$1 same context$4)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 14:45
S70	1	"704".clas. and (signifier$1 same qualifier$1)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 14:46
S71	9	"704".clas. and (signifier$1 and qualifier$1)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/06/15 14:46

EAST Search History (Interference)

<This search history is empty>

6/15/2010 4:38:16 PM

C:\ Documents and Settings\ eyen\ My Documents\ EAST\ Workspaces\ 11580926.wsp

Index of Claims	Application/Control No.	Applicant(s)/Patent Under Reexamination
	11580926	BALDWIN ET AL.
	Examiner	Art Unit
	ERIC YEN	2626

✓	Rejected	-	Cancelled	N	Non-Elected	A	Appeal
=	Allowed	÷	Restricted	I	Interference	O	Objected

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIM		DATE								
Final	Original	06/15/2010								
	1	✓								
	2	✓								
	3	✓								
	4	✓								
	5	✓								
	6	✓								
	7	✓								
	8	✓								
	9	✓								
	10	✓								
	11	✓								
	12	✓								
	13	✓								
	14	✓								
	15	✓								
	16	✓								
	17	✓								
	18	✓								
	19	✓								
	20	✓								
	21	✓								
	22	✓								
	23	✓								
	24	✓								
	25	✓								
	26	✓								
	27	✓								
	28	✓								
	29	✓								
	30	✓								
	31	✓								
	32	✓								
	33	✓								
	34	✓								
	35	✓								
	36	✓								

Index of Claims	Application/Control No.	Applicant(s)/Patent Under Reexamination
	11580926	BALDWIN ET AL.
	Examiner	Art Unit
	ERIC YEN	2626

✓	Rejected	-	Cancelled	N	Non-Elected	A	Appeal
=	Allowed	÷	Restricted	I	Interference	O	Objected

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIM		DATE								
Final	Original	06/15/2010								
	37	✓								
	38	✓								
	39	✓								
	40	✓								
	41	✓								
	42	✓								

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	11580926	BALDWIN ET AL.
	Examiner	Art Unit
	ERIC YEN	2626

SEARCHED			
Class	Subclass	Date	Examiner
704	270-275	6/15/10	EY

SEARCH NOTES		
Search Notes	Date	Examiner
EAST search (attached)	6/15/10	EY

INTERFERENCE SEARCH			
Class	Subclass	Date	Examiner

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
/E.Y./	1A	4,430,669	02/07/1984	Cheung	358	122	
	1B	5,274,560	12/28/1993	LaRue	364	444	
	1C	5,377,350	12/27/1994	Skinner	395	600	
	1D	5,500,920	03/19/1996	Kupiec	395	2.79	
	1E	5,696,965	12/09/1997	Dedrick	395	610	
	1F	5,774,859	06/30/1998	Houser *et al.*	704	275	
	1G	5,878,385	03/02/1999	Bralich *et al.*	704	9	
	1H	5,902,347	05/11/1999	Backman *et al.*	701	200	
	1I	6,195,634	02/27/2001	Dudemaine *et al.*	704	231	
	1J	6,415,257	07/02/2002	Junqua *et al.*	704	275	
	1K	6,466,654	10/15/2002	Cooper *et al.*	379	88.01	
	1L	6,513,006	01/28/2003	Howard *et al.*	704	257	
	1M	6,567,797	05/20/2003	Schuetze *et al.*	707	2	
	1N	6,570,555	05/27/2003	Prevost *et al.*	345	156	
	1O	6,574,597	06/03/2003	Mohri *et al.*	704	251	
	1P	6,598,018	07/22/2003	Junqua	704	251	
	1Q	6,631,346	10/07/2003	Karaorman *et al.*	704	9	
	1R	6,643,620	11/04/2003	Contolini *et al.*	704	270	
	1S	6,721,706	04/13/2004	Strubbe *et al.*	704	275	
	1T	6,735,592	05/11/2004	Neumann *et al.*	707	101	
	1U	6,757,718	06/29/2004	Halverson *et al.*	709	218	
	1V	6,795,808	09/21/2004	Strubbe *et al.*	704	275	
/E.Y./	1W	6,901,366	05/31/2005	Kuhn *et al.*	704	275	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

/E.Y./	1X	CHAI *et al.*, "MIND: A Semantics-Based Multimodal Interpretation Framework for Conversational System", Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, June 2002, pages 37-46.
/E.Y./	1Y	CHEYER *et al.*, "Multimodal Maps: An Agent-Based Approach", International Conference on Cooperative Multimodal Communication (CMC/95), May 24-26, 1995, pages 111-121.
/E.Y./	1Z	ELIO *et al.*, "On Abstract Task Models and Conversation Policies" in Workshop on Specifying and Implementing Conversation Policies, Autonomous Agents '99, Seattle, 1999, 10 pages.

EXAMINER	DATE CONSIDERED
/E. Yen/	06/15/2010

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

401346771v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
/E.Y./	2A	6,944,594	09/13/2005	Busayapongchai *et al.*	704	275	
	2B	7,127,400	10/24/2006	Koch	704	270.1	
	2C	7,146,319	12/05/2006	Hunt	704	254	
	2D	7,289,606	10/30/2007	Sibal *et al.*	379	52	
	2E	7,376,645	05/20/2008	Bernard	707	3	
	2F	7,398,209	07/08/2008	Kennewick *et al.*	704	255	
	2G	7,424,431	09/09/2008	Greene *et al.*	704	270	
	2H	2001/0041980	11/15/2001	Howard *et al.*	704	270	
	2I	2002/0049805	04/25/2002	Yamada *et al.*	709	202	
	2J	2002/0065568	05/30/2002	Silfvast *et al.*	700	94	
	2K	2002/0082911	06/27/2002	Dunn *et al.*	705	14	
	2L	2002/0120609	08/29/2002	Lang *et al.*	707	1	
	2M	2002/0124050	09/05/2002	Middeljans	709	203	
	2N	2002/0188602	12/12/2002	Stubler *et al.*	707	3	
	2O	2002/0198714	12/26/2002	Zhou	704	252	
	2P	2003/0014261	01/16/2003	Kageyama	704	275	
	2Q	2003/0112267	06/19/2003	Belrose	345	728	
	2R	2004/0025115	02/05/2004	Sienel *et al.*	715	513	
	2S	2004/0166832	08/26/2004	Portman *et al.*	455	412.1	
	2T	2004/0199375	10/07/2004	Ehsani *et al.*	704	4	
	2U	2005/0021334	01/27/2005	Iwahashi	704	240	
	2V	2005/0114116	05/26/2005	Fiedler	704	201	
/E.Y./	2W	2007/0033005	02/08/2007	Cristo *et al.*	704	9	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

/E.Y./	2X	TURUNEN, "Adaptive Interaction Methods in Speech User Interfaces", Conference on Human Factors in Computing Systems, Seattle, Washington, 2001, pages 91-92.
/E.Y./	2Y	MAO, Mark Z., "Automatic Training Set Segmentation for Multi-Pass Speech Recognition", Department of Electrical Engineering, Stanford University, CA, copyright 2005, IEEE, pages I-685 to I-688.

EXAMINER	DATE CONSIDERED
/E. Yen/	06/15/2010

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

401346771v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
/E.Y./	3A	2007/0038436	02/15/2007	Cristo *et al.*	704	9	
/E.Y./	3B	2007/0050191	03/01/2007	Weider *et al.*	704	275	
/E.Y./	3C	2008/0091406	04/17/2008	Baldwin *et al.*	704	4	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

/E.Y./	3D	VANHOUCKE, Vincent, "Confidence Scoring and Rejection Using Multi-Pass Speech Recognition", Nuance Communications, Menlo Park, CA, [no date], 4 pages.
/E.Y./	3E	WENG, Fuliang, *et al.*, "Efficient Lattice Representation and Generation", Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, [no date], 4 pages.
/E.Y./	3F	EL MELIANI *et al.*, "A Syllabic-Filler-Based Continuous Speech Recognizer for Unlimited Vocabulary", Canadian Conference on Electrical and Computer Engineering, Volume 2, September 5-8, 1995, pages 1007-1010.

EXAMINER	DATE CONSIDERED
/E. Yen/	06/15/2010

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

401346771v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
/E.Y./	A	5,892,813	04/06/1999	Morin *et al.*	379	88.01	
/E.Y./	B	6,721,001	04/13/2004	Berstis	348	231.3	
/E.Y./	C	6,742,021	05/25/2004	Halverson *et al.*	709	218	
/E.Y./	D	6,990,513	01/24/2006	Belfiore *et al.*	709	203	
/E.Y./	E	7,493,559	02/17/2009	Wolff *et al.*	715	727	
/E.Y./	F	2005/0033574	02/10/2005	Kim *et al.*	704	251	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED
/E. Yen/	06/15/2010

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

401497563v1

OIPE IAP07 SEP 27 2007 PATENT & TRADEMARK OFFICE

INFORMATION DISCLOSURE CITATION IN AN APPLICATION

Form PTO-1449

Docket Number	Application Number
090611-0357945	11/580,926
Applicant(s)	
Larry BALDWIN *et al.*	
Filing Date	Group Art Unit
October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
EY	1A	5,155,743	10/13/1992	Jacobs	375	28	
	1B	5,386,556	01/31/1995	Hedin et al.	395	600	
	1C	5,424,947	06/13/1995	Nagao et al.	364	419.08	
	1D	5,471,318	11/28/1995	Ahuja et al.	358	400	
	1E	5,475,733	12/12/1995	Eisdorfer et al.	379	52	
	1F	5,499,289	03/12/1996	Bruno et al.	379	220	
	1G	5,517,560	05/14/1996	Greenspan	379	114	
	1H	5,533,108	07/02/1996	Harris et al.	379	201	
	1I	5,537,436	07/16/1996	Bottoms et al.	375	222	
	1J	5,539,744	07/23/1996	Chu et al.	370	60	
	1K	5,557,667	09/17/1996	Bruno et al.	379	201	
	1L	5,563,937	10/08/1996	Bruno et al.	379	201	
	1M	5,590,039	12/31/1996	Ikeda et al.	395	759	
	1N	5,633,922	05/27/1997	August et al.	379	220	
	1O	5,675,629	10/07/1997	Raffel et al.	379	58	
	1P	5,708,422	01/13/1998	Blonder et al.	340	825.34	
	1Q	5,721,938	02/24/1998	Stuckey	395	754	
	1R	5,722,084	02/24/1998	Chakrin et al.	455	551	
	1S	5,742,763	04/21/1998	Jones	395	200.3	
	1T	5,752,052	05/12/1998	Richardson et al.	395	759	
	1U	5,754,784	05/19/1998	Garland et al.	395	200.49	
	1V	5,761,631	06/02/1998	Nasukawa	704	9	
EY	1W	5,794,050	08/11/1998	Dahlgren et al.	395	708	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO
	1X	WO 01/78065	10/18/2001	International				

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EY	1Y	REUTERS, "IBM to Enable Honda Drivers to Talk to Cars", Charles Schwab & Co., Inc., July 28, 2002, 1 page.
EY	1Z	LIN, Bor-shen, *et al.*, "A Distributed Architecture for Cooperative Spoken Dialogue Agents with Coherent Dialogue State and History", ASRU'99, 1999, 4 pages.
EY	1AA	KUHN, Thomas, *et al.*, "Hybrid In-Car Speech Recognition for Mobile Multimedia Applications", Vehicular Technology Conference, IEEE, July 1999, pages 2009-2013.

EXAMINER	DATE CONSIDERED
/E. Yen/	6/15/10

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
EY	2A	5,799,276	08/25/1998	Komissarchik *et al.*	704	251	
	2B	5,802,510	09/01/1998	Jones	707	2	
	2C	5,832,221	11/03/1998	Jones	375	200.36	
	2D	5,878,386	03/02/1999	Coughlin	704	10	
	2E	5,895,466	04/20/1999	Goldberg et al.	707	5	
	2F	5,911,120	06/08/1999	Jarett et al.	455	417	
	2G	5,918,222	06/29/1999	Fukui *et al.*	707	1	
	2H	5,926,784	07/20/1999	Richardson et al.	704	9	
	2I	5,933,822	08/03/1999	Braden-Harder et al.	707	5	
	2J	5,953,393	09/14/1999	Culbreth et al.	379	88.25	
	2K	5,963,894	10/05/1999	Richardson et al.	704	9	
	2L	5,963,940	10/05/1999	Liddy et al.	707	5	
	2M	5,987,404	11/16/1999	Della Pietra et al.	704	9	
	2N	5,991,721	11/23/1999	Asano et al.	704	257	
	2O	5,995,119	11/30/1999	Cosatto et al.	345	473	
	2P	6,009,382	12/28/1999	Martino et al.	704	1	
	2Q	6,014,559	01/11/2000	Amin	455	413	
	2R	6,021,384	02/01/2000	Gorin et al.	704	1	
	2S	6,044,347	03/28/2000	Abella et al.	704	272	
	2T	6,049,602	04/11/2000	Foladare et al.	379	265	
	2U	6,058,187	05/02/2000	Chen	380	21	
	2V	6,078,886	06/20/2000	Dragosh et al.	704	270	
EY	2W	6,081,774	06/27/2000	De Hita et al.	704	9	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EY	2X	BELVIN, Robert, *et al.*, "Development of the HRL Route Navigation Dialogue System", Proceedings of the First International Conference on Human Language Technology Research, San Diego, 2001, pages 1-5.
EY	2Y	LIND, R., *et al.*, "The Network Vehicle—A Glimpse into the Future of Mobile Multi-Media", IEEE Aerosp. Electron. Systems Magazine, Volume 14, No. 9, September 1999, pages 27-32.

EXAMINER	DATE CONSIDERED
/E. Yen/	6/15/10

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
EY	3A	6,101,241	08/08/2000	Boyce et al.	379	88.01	
	3B	6,119,087	09/12/2000	Kuhn et al.	704	270	
	3C	6,134,235	10/17/2000	Goldman et al.	370	352	
	3D	6,144,667	11/07/2000	Doshi et al.	370	401	
	3E	6,160,883	12/12/2000	Jackson et al.	379	230	
	3F	6,173,279	01/09/2001	Levin et al.	707	5	
	3G	6,175,858	01/16/2001	Bulfer et al.	709	206	
	3H	6,185,535	02/06/2001	Hedin *et al.*	704	270	
	3I	6,192,110	02/20/2001	Abella et al.	379	88.01	
	3J	6,192,338	02/20/2001	Haszto et al.	704	257	
	3K	6,195,651	02/27/2001	Handel *et al.*	707	2	
	3L	6,208,972	03/27/2001	Grant *et al.*	704	275	
	3M	6,219,346	04/17/2001	Maxemchuk	370	338	
	3N	6,219,643	04/17/2001	Cohen et al.	704	257	
	3O	6,233,556	05/15/2001	Teunen et al.	704	250	
	3P	6,233,559	05/15/2001	Balakrishnan	704	275	
	3Q	6,246,981	06/12/2001	Papineni et al.	704	235	
	3R	6,272,455	08/07/2001	Hoshen et al.	704	1	
	3S	6,292,767	09/18/2001	Jackson et al.	704	1	
	3T	6,314,402	11/06/2001	Monaco et al.	704	275	
	3U	6,366,886	04/02/2002	Dragosh et al.	704	270.1	
	3V	6,385,646	05/07/2002	Brown et al.	709	217	
EY	3W	6,393,428	05/21/2002	Miller et al.	707	102	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EY	3X	ZHAO, Yilin, "Telematics: Safe and Fun Driving", IEEE Intelligent Systems, Volume 17, Issue 1, 2002, pages 10-14.

EXAMINER	DATE CONSIDERED
/E. Yen/	6/15/10

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
EY	4A	6,404,878	06/11/2002	Jackson et al.	379	221.01	
	4B	6,408,272	06/18/2002	White *et al.*	704	270.1	
	4C	6,411,810	06/25/2002	Maxemchuk	455	453	
	4D	6,418,210	07/09/2002	Sayko	379	142.15	
	4E	6,420,975	07/16/2002	DeLine *et al.*	340	815.4	
	4F	6,430,285	08/06/2002	Bauer et al.	379	265.01	
	4G	6,434,523	08/13/2002	Monaco	704	257	
	4H	6,434,524	08/13/2002	Weber	704	257	
	4I	6,442,522	08/27/2002	Carberry et al.	704	257	
	4J	6,446,114	09/03/2002	Bulfer et al.	709	206	
	4K	6,453,153	09/17/2002	Bowker et al.	455	67.4	
	4L	6,453,292	09/17/2002	Ramaswamy et al.	704	235	
	4M	6,456,711	09/24/2002	Cheung et al.	379	265.09	
	4N	6,498,797	12/24/2002	Anerousis et al.	370	522	
	4O	6,499,013	12/24/2002	Weber	704	257	
	4P	6,501,833	12/31/2002	Phillips *et al.*	379	88.07	
	4Q	6,501,834	12/31/2002	Milewski et al.	379	93.24	
	4R	6,523,061	02/18/2003	Halverson et al.	709	202	
	4S	6,532,444	03/11/2003	Weber	704	257	
	4T	6,539,348	03/25/2003	Bond et al.	704	9	
	4U	6,553,372	04/22/2003	Brassell et al.	707	5	
	4V	6,556,973	04/29/2003	Lewin	704	277	
EY	4W	6,560,576	05/06/2003	Cohen et al.	704	270	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED
/E. Yen/	6/15/10

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
EY	5A	6,567,778	05/20/2003	Chao Chang et al.	704	257	
	5B	6,570,964	05/27/2003	Murveit et al.	379	67.1	
	5C	6,574,624	06/03/2003	Johnson et al.	707	5	
	5D	6,594,257	07/15/2003	Doshi et al.	370	352	
	5E	6,604,077	08/05/2003	Dragosh et al.	704	270.1	
	5F	6,611,692	08/26/2003	Raffel et al.	455	552	
	5G	6,614,773	09/02/2003	Maxemchuk	370	337	
	5H	6,615,172	09/02/2003	Bennett *et al.*	704	257	
	5I	6,629,066	09/30/2003	Jackson et al.	704	9	
	5J	6,650,747	11/18/2003	Bala et al.	379	265.06	
	5K	6,691,151	02/10/2004	Cheyer et al.	709	202	
	5L	6,741,931	05/25/2004	Kohut *et al.*	701	209	
	5M	6,937,977	08/30/2005	Gerson	704	201	
	5N	6,980,092	12/27/2005	Turnbull *et al.*	340	425.5	
	5O	6,996,531	02/07/2006	Korall *et al.*	704	270	
	5P	7,020,609	03/28/2003	Thrift *et al.*	704	270.1	
	5Q	7,027,975	04/11/2006	Pazandak *et al.*	704	9	
	5R	7,062,488	06/13/2006	Reisman	707	8	
	5S	7,092,928	08/15/2006	Elad *et al.*	706	60	
	5T	7,137,126	11/14/06	Coffman *et al.*	719	328	
	5U	2002/0143535	10/03/2002	Kist *et al.*	704	251	
	5V	2003/0182132	09/25/2003	Niemoeller	704	275	
	5W	2004/0044516	03/04/2004	Kennewick *et al.*	704	5	
	5X	2004/0193420	09/30/2004	Kennewick *et al.*	704	257	
EY	5Y	2005/0021826	01/27/2005	Kumar	709	232	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED
/E. Yen/	6/15/10

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANTS :	Larry BALDWIN et al.	CONFIRMATION NO. :	6670
SERIAL NUMBER :	11/580,926	EXAMINER :	Eric L. Yen
FILING DATE :	October 16, 2006	ART UNIT :	2626
FOR :	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE		

RESPONSE TO NON-FINAL OFFICE ACTION

Mail Stop Amendment
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

In response to the Non-Final Office Action mailed **June 16, 2010**, please amend the above-identified application as follows:

Amendments to the Specification begin on page **2** of this paper.

Amendments to the Claims begin on page **5** of this paper.

Remarks/Arguments begin on page **18** of this paper.

It is believed that no extensions of time or fees for net addition of claims are required beyond those that may otherwise be provided for in documents accompanying this paper. However, if additional extensions of time are necessary to prevent abandonment of this application, then such extensions of time are hereby petitioned for under 37 C.F.R. § 1.136(a), and any fees required therefore (including fees for net addition of claims) are hereby authorized to be charged to our Deposit Account No. 033975 (**Ref. No. 090611-0357945**).

AMENDMENT

Amendments to the Specification: (identifying insertions and ~~deletions~~).

- Please amend paragraph **[025]**, which begins on page 10, as follows:

[025] The utterance component of input 105 may be processed by an Automatic Speech Recognizer 110 to generate one or more preliminary interpretations of the utterance. Automatic Speech Recognizer 110 may process the utterance using any suitable technique known in the art. For example, in one implementation, Automatic Speech Recognizer 110 may interpret the utterance using techniques of phonetic dictation to recognize a phoneme stream, as described in ~~copending~~ U.S. Patent Application ~~Ser.~~ Serial No. 11/513,269, entitled "Dynamic Speech Sharpening," filed August 31, 2006, which issued as U.S. Patent No. 7,634,409 on December 15, 2009, and which is hereby incorporated by reference in its entirety. The one or more preliminary interpretations generated by Automatic Speech Recognizer 110 may then be provided to a conversational speech engine 115 for further processing. Conversational speech engine 115 may include a conversational language processor 120 and/or a voice search engine 125, described in greater detail in Fig. 2 below. Conversational speech engine 115 may communicate with one or more databases 130 to generate an adaptive conversational response, which may be returned to the user as an output 140. In one implementation, output 140 may be a multi-modal output and/or an interaction with one or more applications 145 to complete the request. For example, output 140 may include a combination of an audible response and a display of a route on a navigation device. For example, the utterance may include a request to perform an action, and output 140 may include a conversational response reporting success or failure, as well as an execution of the action. In addition, in various implementations, Automatic Speech Recognizer 110, conversational speech engine 115, and/or databases 130 may reside locally (e.g., on a user device), remotely (e.g., on a server), or a hybrid model of local and remote processing may be used (e.g., lightweight applications may be processed locally while computationally intensive applications may be processed remotely).

- Please amend paragraph **[034]**, which begins on page 16, as follows:

[034] According to one aspect of the invention, conversational speech engine 215 may include a context determination process 255 that determines one or more contexts for a request to establish meaning within a conversation. The one or more contexts may be determined by having one or more context domain agents compete to determine a most appropriate domain for a given utterance, as described in ~~copending~~ U.S. Patent Application ~~Ser.~~ Serial No. 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed August 5, 2005, which issued as U.S. Patent No. 7,640,160 on December 29, 2009 and ~~copending~~ U.S. Patent Application ~~Ser.~~ Serial No. 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," filed August 29, 2005, both of which are hereby incorporated by reference in their entirety. Once a given context domain agent "wins" the competition, the winning agent may be responsible for establishing or inferring further contexts and updating short-term and long-term shared knowledge. If there is a deadlock between context domain agents, an adaptive conversational response may prompt the user to assist in disambiguating between the deadlocked agents. For example, a user utterance of "What about traffic?" may have a distinct meaning in various contexts. That is, "traffic" may have a first meaning when the user is querying a system's media player (i.e., "traffic" would be a Rock and Roll band led by singer/songwriter Steve Winwood), a second meaning when the user is querying a search interface regarding Michael Douglas films (i.e., "traffic" would be a film directed by Steven Soderbergh), a third meaning when the user is querying a navigation device for directions to an airport (i.e., "traffic" would be related to conditions on roads along a route to the airport).

- Please amend paragraph **[048]**, which begins on page 24, as follows:

[048] According to another aspect of the invention, Intelligent Hypothesis Builder 310 may generate one or more explicit hypotheses of a user's intent when an utterance contains all information (including qualifiers) needed to complete a request or task. Each hypothesis may have a corresponding degree of certainty, which may be used to determine a level of unprompted support to provide in a response. For example, a response may include a confirmation to ensure the utterance was not misunderstood[[.]] or the response may adaptively prompt a user to provide missing information.

Amendments to the Claims: Please replace all prior versions and listings of claims with the following listing of claims.

LISTING OF CLAIMS:

1. (**Currently Amended**) A method for providing a cooperative conversational voice user interface ~~that generates cooperative conversational responses to user utterances, the responses being based on shared knowledge about a current conversation and/or one or more past conversations~~, ~~the method~~ comprising:

receiving an utterance at a voice input device, wherein the voice input device receives the ~~including at least one~~ utterance in a current conversation with a user;

accumulating short-term ~~updating the~~ shared knowledge about the current conversation, wherein the short-term shared knowledge includes ~~to include~~ knowledge about the ~~at least one~~ utterance received at the voice input device in the current conversation;

accumulating long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

determining an intended meaning for the ~~at least one~~ utterance, wherein a conversational speech engine determines the intended meaning for the utterance from ~~based on~~ the ~~updated~~ short-term shared knowledge and the long-term shared knowledge; and

generating a response to the utterance, wherein the conversational speech engine generates the response to the utterance based on the ~~determined~~ intended meaning determined from the short-term shared knowledge and the long-term shared knowledge.

2. (**Currently Amended**) The method of claim 1, wherein accumulating ~~updating~~ the short-term shared knowledge about the current conversation includes populating a short-term context stack with information about the ~~at least one~~ utterance received in the current conversation.

3. (**Currently Amended**) The method of claim 2, wherein accumulating the short-term shared knowledge about the current conversation further includes ~~comprising~~ expiring the information about the ~~at least one~~ utterance [[in]] from the short-term context stack after a predetermined amount of time.

4. (**Currently Amended**) The method of claim 1, wherein accumulating ~~updating~~ the long-term shared knowledge about the user includes updating one or more long-term profiles associated with the user to include ~~based on~~ information about the ~~at least one~~ utterance received in the current conversation.

5. (**Currently Amended**) The method of claim 1, wherein determining the intended meaning for the utterance includes identifying a conversation type for the current ~~at least one utterance, the~~ conversation ~~type~~ based on one or more of a conversational goal for the utterance, ~~one or more participant~~ roles for one or more participants in the ~~utterance~~ current conversation, or ~~and/or~~ an allocation of information among ~~for~~ the ~~utterance~~ one or more participants in the current conversation.

6. (**Currently Amended**) The method of claim 1, further comprising generating a ~~wherein determining the intended meaning includes generating at least one~~ hypothesis having a degree of certainty about [[of]] an intent that ~~intended meaning for~~ the user had in speaking the ~~at least one~~ utterance, wherein conversational speech engine further determines the intended meaning for the utterance from the ~~at least one~~ hypothesis ~~including a degree of certainty~~ about the intent that the user had in speaking the utterance.

7. (**Currently Amended**) The method of claim 6, wherein ~~generating~~ the conversational speech engine assigns the degree of certainty to the ~~at least one~~ hypothesis based on a context that the conversational speech engine identifies for the current conversation from one or more ~~includes identifying~~ contextual signifiers ~~and/or~~ or qualifiers recognized in the ~~at least one~~ utterance.

8. (**Currently Amended**) The method of claim [[7]] 1, ~~further comprising~~ wherein determining the intended meaning for the utterance includes:

inferring one or more ~~missing~~ contextual signifiers ~~and/or~~ or qualifiers missing from [[for]] the ~~at least one~~ utterance in response to determining that ~~based on~~ the ~~shared knowledge~~ utterance contains insufficient information to generate the response; and

identifying a context for the current conversation from the one or more inferred contextual signifiers or qualifiers.

9. (**Currently Amended**) The method of claim 8, wherein the conversational speech engine assigns an implicit degree of certainty to ~~for~~ the intended meaning determined for the utterance in response to determining that ~~at least one hypothesis is based on one or more of~~ the utterance contains insufficient information to generate the ~~identified contextual signifiers and/or qualifiers, the inferred contextual signifiers and/or qualifiers, and/or contextual signifiers and/or qualifiers necessary for generating a~~ response.

10. (**Currently Amended**) The method of claim 1, wherein ~~generating~~ the response to the utterance includes ~~adapting~~ one or more words having contextual signifiers, qualifiers, or ~~and/or~~ grammar syntax adapted ~~in the response~~ to reflect contextual signifiers, qualifiers, or ~~and/or~~ grammar syntax recognized in or inferred from the ~~at least one~~ utterance.

11. (**Currently Amended**) The method of claim 1, wherein the one or more words having the adapted contextual signifiers, qualifiers, or grammar syntax frame ~~generating~~ the response ~~includes framing the response~~ to influence a subsequent utterance in the current conversation.

12. (**Currently Amended**) The method of claim 1, ~~further comprising~~ wherein determining the intended meaning for the utterance includes:

~~learning that the determined intended meaning is incorrect;~~

generating multiple possible interpretations for ~~further updating~~ the utterance, wherein an initial interpretation for ~~shared knowledge to include knowledge that~~ the ~~determined intended meaning is incorrect~~ utterance comprises one of the multiple possible interpretations having a highest degree of certainty; and

updating ~~determining a second intended meaning for the at least one utterance based on~~ the short-term ~~further updated~~ shared knowledge about the current conversation to remove the initial interpretation from the multiple possible interpretation in response to determining that the initial interpretation was incorrect, wherein the ~~incorrect determined~~ intended meaning ~~is excluded as a~~ determined from the short-term shared knowledge and the long-term shared knowledge comprises one of the multiple possible ~~second intended meaning~~ interpretations having a next highest degree of certainty.

13. (**Currently Amended**) The method of claim 1, wherein the user speaks the utterance in ~~input is~~ a multi-modal input that further includes one or more non-voice inputs relating to the utterance.

14. (**Currently Amended**) The method of claim 1, wherein the conversational speech engine generates the response [[is]] in a multi-modal output that includes one or more non-voice outputs presented to the user in response to the utterance or one or more tasks executed to process a request identified from the intended meaning for the utterance.

15. (**Currently Amended**) A non-transitory computer readable medium containing computer-executable instructions for providing a cooperative conversational voice user interface ~~that generates cooperative conversational responses to user utterances, the responses being based on shared knowledge about a current conversation and/or one or more past conversations~~, the computer-executable instructions operable when executed to:

receive an utterance at a voice input device, wherein the voice input device receives the ~~including at least one~~ utterance in a current conversation with a user;

accumulate short-term ~~update the~~ shared knowledge about the current conversation, wherein the short-term shared knowledge includes ~~to include~~ knowledge about the ~~at least one~~ utterance received at the voice input device in the current conversation;

accumulate long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

determine an intended meaning for the ~~at least one~~ utterance, wherein a conversational speech engine determines the intended meaning for the utterance from ~~based on~~ the ~~updated~~ short-term shared knowledge and the long-term shared knowledge; and

generate a response to the utterance, wherein the conversational speech engine generates the response to the utterance based on the ~~determined~~ intended meaning determined from the short-term shared knowledge and the long-term shared knowledge.

16. (**Currently Amended**) The non-transitory computer readable medium of claim 15, wherein to accumulate the short-term shared knowledge about the current conversation, the computer-executable instructions are further operable when executed to populate ~~update the shared knowledge by populating~~ a short-term context stack with information about the ~~at least one~~ utterance received in the current conversation.

17. (**Currently Amended**) The non-transitory computer readable medium of claim 16, wherein to accumulate the short-term shared knowledge about the current conversation, the computer-executable instructions are further operable when executed to expire the information about the ~~at least one~~ utterance [[in]] from the short-term context stack after a predetermined amount of time.

18. (**Currently Amended**) The non-transitory computer readable medium of claim 15, wherein to accumulate the long-term shared knowledge about the user, the computer-executable instructions are further operable when executed to update ~~the shared knowledge by updating~~ one or more long-term profiles associated with the user to include ~~based on~~ information about the ~~at least one~~ utterance received in the current conversation.

19. (**Currently Amended**) The non-transitory computer readable medium of claim 15, wherein to determine the intended meaning for the utterance, the computer-executable instructions are further operable when executed to identify ~~determine the intended meaning by identifying~~ a conversation type for the current ~~at least one utterance, the~~ conversation ~~type~~ based on one or more of a conversational goal for the utterance, ~~one or more participant~~ roles for one or more participants in the ~~utterance~~ current conversation, or ~~and/or~~ an allocation of information among ~~for~~ the ~~utterance~~ one or more participants in the current conversation.

20. (**Currently Amended**) The non-transitory computer readable medium of claim 15, wherein the computer-executable instructions are further operable when executed to generate a ~~determine the intended meaning by generating at least one~~ hypothesis having a degree of certainty about [[of]] an intent that ~~intended meaning for~~ the user had in speaking the ~~at least one~~ utterance, wherein conversational speech engine further determines the intended meaning for the utterance from the ~~at least one~~ hypothesis ~~including a degree of certainty~~ about the intent that the user had in speaking the utterance.

21. (**Currently Amended**) The computer-readable medium of claim 20, wherein the conversational speech engine assigns the degree of certainty ~~computer-executable instructions are operable~~ to ~~generate~~ the ~~at least one~~ hypothesis based on a context that the conversational speech engine identifies for the current conversation from one or more ~~by identifying~~ contextual signifiers ~~and/or~~ or qualifiers recognized in the ~~at least one~~ utterance.

22. (**Currently Amended**) The non-transitory computer readable medium of claim [[21]] 15, wherein to determine the intended meaning for the utterance, the computer-executable instructions are further operable when executed to:

infer one or more ~~missing~~ contextual signifiers ~~and/or~~ or qualifiers missing from [[for]] the ~~at least one~~ utterance in response to determining that ~~based on~~ the ~~shared knowledge~~ utterance contains insufficient information to generate the response; and

identify a context for the current conversation from the one or more inferred contextual signifiers or qualifiers.

23. (**Currently Amended**) The non-transitory computer readable medium of claim 22, wherein the conversational speech engine assigns an implicit degree of certainty to ~~for~~ the intended meaning determined for the utterance in response to determining that ~~at least one hypothesis is based on one or more of~~ the utterance contains insufficient information to generate the ~~identified contextual signifiers and/or qualifiers, the inferred contextual signifiers and/or qualifiers, and/or contextual signifiers and/or qualifiers necessary for generating a~~ response.

24. (**Currently Amended**) The non-transitory computer readable medium of claim 15, wherein the ~~computer-executable instructions are operable to generate the~~ response to the utterance includes one or more words having ~~by adapting~~ contextual signifiers, qualifiers, or ~~and/or~~ grammar syntax adapted ~~in the response~~ to reflect contextual signifiers, qualifiers, or ~~and/or~~ grammar syntax recognized in or inferred from the ~~at least one~~ utterance.

25. (**Currently Amended**) The non-transitory computer readable medium of claim 15, wherein the one or more words having the adapted contextual signifiers, qualifiers, or grammar syntax frame ~~computer-executable instructions are operable to generate the response by framing~~ the response to influence a subsequent utterance in the current conversation.

26. (**Currently Amended**) The non-transitory computer readable medium of claim 15, wherein to determine the intended meaning for the utterance, the computer-executable instructions are further operable when executed to:

~~learn that the determined intended meaning is incorrect;~~

generate multiple possible interpretations for ~~further update~~ the utterance, wherein an initial interpretation for ~~shared knowledge to include knowledge that~~ the ~~determined intended meaning is incorrect~~ utterance comprises one of the multiple possible interpretations having a highest degree of certainty; and

update ~~determine a second intended meaning for the at least one utterance based on~~ the short-term ~~further updated~~ shared knowledge about the current conversation to remove the initial interpretation from the multiple possible interpretation in response to determining that the initial interpretation was incorrect, wherein the ~~incorrect determined~~ intended meaning ~~is excluded as a~~ determined from the short-term shared knowledge and the long-term shared knowledge comprises one of the multiple possible ~~second intended meaning~~ interpretations having a next highest degree of certainty.

27. (**Currently Amended**) The non-transitory computer readable medium of claim 15, wherein the user speaks the utterance in ~~input is~~ a multi-modal input that further includes one or more non-voice inputs relating to the utterance.

28. (**Currently Amended**) The non-transitory computer readable medium of claim 15, wherein the conversational speech engine generates the response [[is]] in a multi-modal output that includes one or more non-voice outputs presented to the user in response to the utterance or one or more tasks executed to process a request identified from the intended meaning for the utterance.

29. (**Currently Amended**) A system for providing a cooperative conversational voice user interface ~~that generates cooperative conversational responses to user utterances, the system including at least one memory for storing shared knowledge about a current conversation and/or one or more past conversations~~, ~~the system~~ comprising:

a voice ~~at least one~~ input device configured to receive ~~that receives~~ an ~~input, the input including at least one~~ utterance in a current conversation with a user; and

a conversational speech engine, wherein the conversational speech engine includes one or more processors ~~collectively operable~~ configured to:

accumulate short-term ~~update the~~ shared knowledge about the current conversation, wherein the short-term shared knowledge includes ~~to include~~ knowledge about the ~~at least one~~ utterance received at the voice input device in the current conversation;

accumulate long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

determine an intended meaning for the ~~at least one~~ utterance, wherein the conversational speech engine determines the intended meaning for the utterance from ~~based on~~ the ~~updated~~ short-term shared knowledge and the long-term shared knowledge; and

generate a response to the utterance, wherein the conversational speech engine generates the response to the utterance based on the ~~determined~~ intended meaning determined from the short-term shared knowledge and the long-term shared knowledge.

30. (**Currently Amended**) The system of claim 29, wherein to accumulate the short-term shared knowledge about the current conversation, the one or more processors are ~~operable~~ further configured to populate ~~update the shared knowledge by populating~~ a short-term context stack with information about the ~~at least one~~ utterance received in the current conversation.

31. (**Currently Amended**) The system of claim 30, wherein to accumulate the short-term shared knowledge about the current conversation, the one or more processors are ~~operable~~ further configured to expire the information about the ~~at least one~~ utterance [[in]] from the short-term context stack after a predetermined amount of time.

32. (**Currently Amended**) The system of claim 29, wherein to accumulate the long-term shared knowledge about the user, the one or more processors are further configured ~~operable~~ to update ~~the shared knowledge by updating~~ one or more long-term profiles associated with the user to include ~~based on~~ information about the ~~at least one~~ utterance received in the current conversation.

33. (**Currently Amended**) The system of claim 29, wherein to determine the intended meaning for the utterance, the one or more processors are further configured ~~operable~~ to identify ~~determine the intended meaning by identifying~~ a conversation type for the current ~~at least one utterance, the~~ conversation ~~type~~ based on one or more of a conversational goal for the utterance, ~~one or more participant~~ roles for one or more participants in the ~~utterance~~ current conversation, or ~~and/or~~ an allocation of information among ~~for~~ the ~~utterance~~ one or more participants in the current conversation.

34. (**Currently Amended**) The system of claim 29, wherein the one or more processors are further configured ~~operable~~ to generate a ~~determine the intended meaning by generating at least one~~ hypothesis having a degree of certainty about [[of]] an intent that ~~intended meaning for~~ the user had in speaking the ~~at least one~~ utterance, wherein conversational speech engine further determines the intended meaning for the utterance from the ~~at least one~~ hypothesis ~~including a degree of certainty~~ about the intent that the user had in speaking the utterance.

35. (**Currently Amended**) The system of claim 34, wherein the conversational speech engine assigns the degree of certainty ~~one or more processors are operable~~ to ~~generate~~ the ~~at least one~~ hypothesis based on a context that the conversational speech engine identifies for the current conversation from one or more ~~by identifying~~ contextual signifiers ~~and/or~~ or qualifiers recognized in the ~~at least one~~ utterance.

36. (**Currently Amended**) The system of claim [[35]] 29, wherein to determine the intended meaning for the utterance, the one or more processors are further configured ~~operable~~ to:

infer one or more ~~missing~~ contextual signifiers ~~and/or~~ or qualifiers missing from [[for]] the ~~at least one~~ utterance in response to determining that ~~based on~~ the ~~shared knowledge~~ utterance contains insufficient information to generate the response; and

identify a context for the current conversation from the one or more inferred contextual signifiers or qualifiers.

37. (**Currently Amended**) The system of claim 36, wherein the conversational speech engine assigns an implicit degree of certainty to ~~for~~ the intended meaning determined for the utterance in response to determining that ~~at least one hypothesis is based on one or more of~~ the utterance contains insufficient information to generate the ~~identified contextual signifiers and/or qualifiers, the inferred contextual signifiers and/or qualifiers, and/or contextual signifiers and/or qualifiers necessary for generating a~~ response.

38. (**Currently Amended**) The system of claim 29, wherein the ~~one or more processors are operable to generate the~~ response to the utterance includes one or more words having ~~by adapting~~ contextual signifiers, qualifiers, or ~~and/or~~ grammar syntax adapted ~~in the response~~ to reflect contextual signifiers, qualifiers, or ~~and/or~~ grammar syntax recognized in or inferred from the ~~at least one~~ utterance.

39. (**Currently Amended**) The system of claim 29, wherein the one or more words having the adapted contextual signifiers, qualifiers, or grammar syntax frame ~~one or more processors are operable to generate the response by framing~~ the response to influence a subsequent utterance in the current conversation.

40. (**Currently Amended**) The system of claim 29, wherein to determine the intended meaning for the utterance, the one or more processors are further configured ~~operable~~ to:

~~learn that the determined intended meaning is incorrect;~~

generate multiple possible interpretations for ~~further update~~ the utterance, wherein an initial interpretation for ~~shared knowledge to include knowledge that~~ the ~~determined intended meaning is incorrect~~ utterance comprises one of the multiple possible interpretations having a highest degree of certainty; and

update ~~determine a second intended meaning for the at least one utterance based on~~ the short-term ~~further updated~~ shared knowledge about the current conversation to remove the initial interpretation from the multiple possible interpretation in response to determining that the initial interpretation was incorrect, wherein the ~~incorrect determined~~ intended meaning ~~is excluded as a~~ determined from the short-term shared knowledge and the long-term shared knowledge comprises one of the multiple possible ~~second intended meaning~~ interpretations having a next highest degree of certainty.

41. (**Currently Amended**) The system of claim 29, wherein the user speaks the utterance in ~~input is~~ a multi-modal input that further includes one or more non-voice inputs relating to the utterance.

42. (**Currently Amended**) The system of claim 29, wherein the conversational speech engine generates the response [[is]] in a multi-modal output that includes one or more non-voice outputs presented to the user in response to the utterance or one or more tasks executed to process a request identified from the intended meaning for the utterance.

REMARKS

Upon entry of the foregoing Amendment, claims 1-42 are pending in the application. Claims 1-42 have been amended. No claims have been cancelled or newly added. Applicants believe that this Amendment does not add new matter. In view of the foregoing Amendment and the following Remarks, allowance of all the pending claims is requested.

INFORMATION DISCLOSURE STATEMENT

Applicants note that the Examiner did not initial one of the references cited in the Information Disclosure Statement that Applicants filed on September 27, 2007 (i.e., Reference 1X, Document No. WO 01/78065). As such, Applicants respectfully request that the Examiner consider the non-initialed reference cited in the Information Disclosure Statement that Applicants filed on September 27, 2007 in the next Office Action.

REJECTION UNDER 35 U.S.C. § 101

The Examiner has rejected claims 1-28 under 35 U.S.C. § 101 as allegedly being directed to non-statutory subject matter. In particular, the Examiner alleges that the features recited in independent claim 1 "neither transform underlying subject matter nor positively tie to another statutory class that accomplishes the claimed method steps, and therefore do not qualify as statutory subject matter."

Solely to expedite prosecution of this application, and without acknowledging the propriety of the alleged basis for the rejection, Applicants have amended independent claim 1 to recite, among other things, "receiving an utterance at ***a voice input device***" and having "***a conversational speech engine*** determine[] the intended meaning for the utterance from the short-term shared knowledge and the long-term shared knowledge" and generate a "response to the utterance based on the intended meaning."

Accordingly, for at least the reason that the method recited in amended independent claim 1 includes "a voice input device" and "a conversational speech engine" performing various features further recited, amended independent claim 1 recites statutory subject matter. Applicants therefore request that the Examiner withdraw this rejection of the claims.

Furthermore, the Examiner alleges that, "given its broadest reasonable interpretation," the "computer readable medium" recited in independent claim 15 "includes transitory computer-readable media (signals, carrier waves, etc.) which are non-statutory." Thus, the Examiner has recommended amending independent claim 15 to recite a "non-transitory computer readable medium" to over the rejection. Applicants therefore request that the Examiner withdraw this rejection of the claims.

REJECTION UNDER 35 U.S.C. § 112, FIRST PARAGRAPH

The Examiner has rejected claims 7-10, 21-24, and 35-38 under 35 U.S.C. § 112, first paragraph, as allegedly failing to comply with the enablement requirement. In particular, the Examiner alleges that that the terms "contextual," "signifier," and "qualifier" have an unclear scope because the disclosure allegedly does not describe what the signifiers and qualifiers are.

Solely to expedite prosecution of this application, and without acknowledging the propriety of the alleged basis for the rejection, Applicants have amended claims 7-10, 21-24, and 35-38 to further clarify the various features recited therein. For example, claim 7 has been amended to further clarify that a conversational speech engine identifies a context for a current conversation based on "contextual signifiers or qualifiers ***recognized in the utterance***." Similarly, claim 8 has been amended to further clarify that "contextual signifiers or qualifiers ***missing from the utterance***" may be inferred to identify a context for the current conversation, while claims 10 and 11 have been amended to further clarify that a response to the utterance may include "***one or more words*** having adapted contextual signifiers, qualifiers, or grammar syntax adapted to reflect contextual signifiers, qualifiers, or grammar syntax ***recognized in or inferred from the utterance***."

Accordingly, for at least the reason that a person having ordinary skill in the art would understand the meaning of the terms "contextual signifiers or qualifiers" recited in amended claims 7-10, 21-24, and 35-38, Applicants submit that amended claims 7-10, 21-24, and 35-38 satisfy the enablement requirement. Applicants therefore request that the Examiner withdraw this rejection of the claims.

REJECTION UNDER 35 U.S.C. § 112, SECOND PARAGRAPH

The Examiner has rejected claims 1, 7-10, 15, 21-24, 29, and 35-38 under 35 U.S.C. § 112, second paragraph, as allegedly being indefinite for failing to particularly point out and distinctly claim the subject matter which Applicants regard as the invention. In particular, the Examiner alleges that the term "cooperative" has an ambiguous meaning and therefore renders the independent claims indefinite. In addition, the Examiner alleges that the terms "contextual," "signifier," and "qualifier" have an unclear scope because the terms are allegedly vague and indefinite.

Solely to expedite prosecution of this application, and without acknowledging the propriety of the alleged basis for the rejection, Applicants have amended independent claims 1, 15, and 29 to further clarify the various features recited therein. In particular, the independent claims have been amended to further clarify that the various features recited therein provide the "cooperative conversational voice user interface," wherein the body of the independent claims do not include the allegedly vague term "cooperative."

Accordingly, for at least the reason that a person having ordinary skill in the art would understand that the various features recited in the body of the independent claims provide the "cooperative conversational voice user interface," independent claims 1, 15, and 29 are clear and definite. Thus, Applicants request that the Examiner withdraw this rejection of the claims.

Additionally, as discussed in further above, Applicants have amended claims 7-10, 21-24, and 35-38 to further clarify the various features recited therein that relate to the nature of the "contextual signifiers or qualifiers." Accordingly, for at least the reason that a person having ordinary skill in the art would understand the meaning of the claimed terms "contextual signifiers or qualifiers," amended claims 7-10, 21-24, and 35-38 are clear and definite. Thus, Applicants request that the Examiner withdraw this rejection of the claims.

REJECTION UNDER 35 U.S.C. § 103

A. CLAIMS 1, 4-11, 13-15, 18-25, 27-29, 32-39, AND 41-42

The Examiner has rejected claims 1, 4-11, 13-15, 18-25, 27-29, 32-39, and 41-42 under 35 U.S.C. § 103 as allegedly being unpatentable over U.S. Patent Application Pub. No. 2005/0015256 to Kargman ("Kargman"). This rejection is improper and must be withdrawn because the Examiner for at least the reason that Kargman fails to disclose, teach, or suggest each and every feature of the claimed invention.

More particularly, Kargman fails to disclose, teach, or suggest at least the combination of features that include "receiving an utterance at a voice input device, wherein the voice input device receives the utterance in a current conversation with a user" and "***determining an intended meaning for the utterance***," wherein a conversational speech engine determines the intended meaning for the utterance from short-term shared knowledge about the current conversation and long-term shared knowledge about the user, including one or more past conversations with the user, as recited in amended independent claim 1, for example. The Examiner alleges that Kargman teaches a system that can generate responses to utterances received from a user based on shared knowledge about a current conversation and/or one or more past conversations. Office Action, pages 6-8.

Applicants disagree with the Examiner's assessment for at least the reason that although Kargman generally describes a system that can "generate an initial or supplemental prompt which is individually tailored for [a] particular customer" and "tailor the voice recognition 'grammar' to that particular customer" based on the customer's order history (¶ [0050]), suggesting certain words or tailoring voice recognition components to improve the likelihood that a particular utterance will be correctly recognized does not fairly disclose, teach, or suggest determining an "intended meaning" for the utterance. In particular, Applicants' disclosure includes various passages that describe techniques that may be used to determine or otherwise establish an intended meaning for an utterance, including a conversational speech engine "that determines one or more contexts for a request to establish meaning within a conversation" (¶ [034]). For example, the word "traffic" may have distinct meanings

in different contexts (e.g., the "Traffic" rock and roll band in a music context, the film "Traffic" in a movie context, road conditions in a navigation context, etc.) (¶ [034]).

In contrast, the passages in Kargman that the Examiner cites as allegedly teaching features that relate to determining an intended meaning for an utterance only generally relate to providing prompts to suggest certain choices or tailoring voice recognition grammars to prevent mismatches when recognizing an utterance from a user. However, pre-configuring a voice recognition system based on a history associated with a user does not disclose, teach, or suggest determining "an intended meaning" for what the user then says in an utterance. In other words, where amended independent claim 1 recites "determining ***an intended meaning for the utterance***" from short-term shared knowledge and long-term shared knowledge, the short-term shared knowledge and long-term shared knowledge provide "profiles [that] may be particularly meaningful when attempting to disambiguate between contexts where a word has different meanings in different contexts" (¶ [037]). On the other hand, the voice recognition techniques described in Kargman only relate to making statistical guesses or suggestions to avoid potential "mismatches," which does not disclose, teach, or suggest disambiguating, establishing, or otherwise determining the intended meaning that the user had in speaking any words recognized in the utterance.

Accordingly, for at least the reason that Kargman does not describe or otherwise indicate that the system described therein can distinguish or otherwise determine an intended meaning underlying an utterance that a user speaks in a current conversation, much less determining an intended meaning for an utterance from short-term shared knowledge about the current conversation and long-term shared knowledge about past conversations with the user, Kargman fails to disclose, teach, or suggest at least the foregoing features recited in amended independent claim 1. The rejection is therefore improper and must be withdrawn.

Independent claims 15 and 29 have been amended to include features similar to those set forth in amended independent claim 1. Claims 4-11, 13-14, 18-25, 27-28, 32-39, and 41-42 depend from and add features to one of amended independent claims 1, 15, and 29. Thus, the rejection of these claims is likewise improper and must be withdrawn for at least the same reasons.

B. CLAIMS 2-3, 12, 16-17, 26, 30-31, AND 40

The Examiner has rejected each of claims 2-3, 12, 16-17, 26, 30-31, and 40 under 35 U.S.C. § 103 as allegedly being unpatentable. In particular, the Examiner has rejected claims 2-3, 16-17, and 30-31 as allegedly being unpatentable over Kargman in view of U.S. Patent No. 5,748,841 to Morin et al. ("Morin") and claims 12, 26, and 40 as allegedly being unpatentable over Kargman in view of U.S. Patent No. 5,488,652 to Bielbly et al. ("Bielbly"). These rejections are improper and must be withdrawn for at least the reason that the references relied upon, either alone or in combination, fail to disclose, teach, or suggest each and every feature of the claimed invention.

More particularly, for at least the reasons discussed above, Kargman fails to disclose, teach, or suggest at least the combination of features that include "receiving an utterance at a voice input device, wherein the voice input device receives the utterance in a current conversation with a user" and "***determining an intended meaning for the utterance***," wherein a conversational speech engine determines the intended meaning for the utterance from short-term shared knowledge about the current conversation and long-term shared knowledge about the user, including one or more past conversations with the user, as recited in amended independent claim 1, for example.

Morin fails to cure the foregoing deficiencies of Kargman for at least the reason that although Morin generally describes a system that uses "dialogue context, a dialogue model and syntactic-semantic grammar to progressively build commands . . . [that] are syntactically and semantically correct in the current context" (Abstract), Morin does not describe further using the dialogue context, a dialogue model, and syntactic-semantic grammar to determine the intended meaning for an utterance received from a user. In particular, building commands that have correct syntax and semantics relates, at best, to instructions that may be generated to process a request contained in an utterance, but Morin does not disclose, teach, or suggest that the system builds the commands based on an "intended meaning" previously determined for the utterance. As such, Morin fails to cure the foregoing deficiencies of Kargman for at least these reasons.

Bielbly further fails to cure the foregoing deficiencies of Kargman for at least the reason that although Bielbly generally describes a system that can ask a caller about whether a call entity recognized in an utterance is incorrect (col. 18, lines 1-30), Bielbly does not indicate that "training speech recognition algorithms" in the system described therein results in the system using the trained speech recognition algorithms to determine the intended meaning for the utterance. For example, rather than attempting to independently disambiguate or otherwise determine the intended meaning for the utterance, the system described in Bielbly prompts the user to confirm or disconfirm whether certain words were correctly recognized. Thus, because the system described in Bielbly relies upon user feedback to determine whether an utterance was correctly recognized, Bielbly does not disclose, teach, or suggest that the system determines an "intended meaning" for an utterance. As such, Bielbly further fails to cure the foregoing deficiencies of Kargman for at least these reasons.

Accordingly, for at least the foregoing reasons, the references relied upon, either alone or in combination, fail to disclose, teach, or suggest each and every feature of amended independent claim 1. Independent claims 15 and 29 have been amended to include features similar to those set forth in amended independent claim 1. Claims 2-3, 12, 16-17, 26, 30-31, and 40 depend from and add features to one of amended independent claims 1, 15, and 29. Thus, the rejections of these claims are improper and must be withdrawn for at least the foregoing reasons.

NON-STATUTORY DOUBLE PATENTING REJECTION

The Examiner has rejected claims 1, 12, 15, 26, 29, and 40 under the judicially created doctrine of obviousness-type double patenting, and has provisionally rejected claims 1, 13, 15, 27, 29, and 41 under the judicially created doctrine of obviousness-type double patenting. In particular, the Examiner alleges that claims 1, 12, 15, 26, 29, and 40 are unpatentable over claim 1 of U.S. Patent No. 7,620,549, that claims 1, 15, and 29 are unpatentable over claim 1 of co-pending U.S. Patent Application Serial No. 11/954,064, that claims 1, 13, 15, 27, 29, and 41 are unpatentable over claims 1 and 6 of co-pending U.S. Patent Application Serial No.

12/127,343, and that claims 1, 15, and 29 are unpatentable over claims 69 and 84 of co-pending U.S. Patent Application Serial No. 12/705,311.

Applicants will consider filing Terminal Disclaimers to overcome the double-patenting rejection and the provisional double-patenting rejections once the remaining claim rejections have been withdrawn. Applicants further note that filing a Terminal Disclaimer to obviate a non-statutory double patenting rejection does not constitute an admission that the rejection is proper. *See Quad Envtl. Tech. Corp. v. Union Sanitary Dist.*, 946 F.2d 870 (Fed. Cir. 1991).

CONCLUSION

Having addressed the outstanding rejections, Applicant respectfully submits that a full and complete response has been made to the Office Action. As such, the application is in condition for allowance. Notice to that effect is respectfully requested.

If the Examiner believes, for any reason, that personal communication will expedite prosecution of this application, the Examiner is invited to telephone the undersigned at the number provided.

Date: **September 2, 2010**

Respectfully submitted,

By: _____

Syed Jafar Ali
Registration No. 58,780

PILLSBURY WINTHROP SHAW PITTMAN LLP
P.O. Box 10500
McLean, Virginia 22102
Main: 703-770-7900
Direct: 703-770-7540
Fax: 703-770-7901

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANTS :	Larry BALDWIN *et al.*	CONFIRMATION NO.:	6670
SERIAL NUMBER :	11/580,926	EXAMINER:	Eric L. Yen
FILING DATE :	October 16, 2006	ART UNIT:	2626
FOR :	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE		

SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT

Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313-1450

Sir:

Pursuant to the duty of disclosure under 37 C.F.R. §§ 1.56, 1.97 and 1.98, Applicants hereby make of record in the above-identified application the documents listed on the attached Form PTO-1449.

Applicants have listed a publication date on the attached PTO-1449 based on information presently available to the undersigned. However, the listed publication date should not be construed as an admission that the information was actually published on the date indicated.

Applicants reserve the right to establish the patentability of the claimed invention over any of the information provided herewith, and/or to prove that this information may not be prior art, and/or to prove that this information may not be enabling for the teachings purportedly offered.

This Supplemental Information Disclosure Statement is being filed more than three months after the filing date of a national application other than a continued prosecution application under 37 C.F.R. § 1.53(d), or more than three months after the date of entry of the national stage as set forth in 37 C.F.R. § 1.491 in an international application, and after the mailing date of the first Office Action on the merits, but before the mailing date of any of a Final Action under 37 C.F.R. § 1.113, a Notice of Allowance under 37 C.F.R. § 1.311, or an action that otherwise closes prosecution in the application. 37 C.F.R. § 1.97(c).

Please charge the $180.00 fee [37 C.F.R. § 1.17(p)] to Deposit Account No. 03-3975. In addition, please charge any fee deficiencies, or credit any overpayment, to Deposit Account No. 03-3975 as needed to ensure consideration of the Supplemental Information Disclosure Statement.

It is respectfully requested that the Examiner initial and return a copy of the enclosed PTO-1449.

Respectfully submitted,

PILLSBURY WINTHROP SHAW PITTMAN LLP

By_____
Syed Jafar Ali
Registration No.: 58,780

Customer No.
00909

P. O. Box 10500
McLean, Virginia 22102
Telephone: (703) 770-7900
Telefax: (703) 770-7901

Dated: September 2, 2010

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	1A	5,617,407	04/01/1997	Bareis	369	275.3	
	1B	5,797,112	08/18/1998	Komatsu *et al.*	701	201	
	1C	5,895,464	04/20/1999	Bhandari *et al.*	707	3	
	1D	6,144,938	11/07/2000	Surace *et al.*	704	257	
	1E	6,226,612	05/01/2001	Srenger *et al.*	704	256	
	1F	6,288,319	09/11/2001	Catona	84	609	
	1G	6,308,151	10/23/2001	Smith	704	235	
	1H	6,381,535	04/30/2002	Durocher *et al.*	701	202	
	1I	6,385,596	05/07/2002	Wiser *et al.*	705	51	
	1J	6,430,531	08/06/2002	Polish	704	257	
	1K	6,466,899	10/15/2002	Yano *et al.*	704	1	
	1L	6,556,970	04/29/2003	Sasaki *et al.*	704	257	
	1M	6,581,103	06/17/2003	Dengler	709	231	
	1N	6,678,680	01/13/2004	Woo	707	6	
	1O	6,801,604	10/05/2004	Maes *et al.*	379	88.17	
	1P	6,801,893	10/05/2004	Backfried *et al.*	704	257	
	1Q	6,833,848	12/21/2004	Wolff *et al.*	345	719	
	1R	6,856,990	02/15/2005	Barile *et al.*	707	10	
	1S	6,865,481	03/08/2005	Kawazoe *et al.*	701	211	
	1T	6,877,134	04/05/2005	Fuller *et al.*	715	500.1	
	1U	6,950,821	09/27/2005	Faybishenko *et al.*	707	10	
	1V	6,973,387	12/06/2005	Masclet *et al.*	701	211	
	1W	7,016,849	03/21/2006	Arnold *et al.*	704	275	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

	1X	ARRINGTON, Michael, "Google Redefines GPS Navigation Landscape: Google Maps Navigation for Android 2.0", TechCrunch, printed from the Internet <http://www.techcrunch.com/2009/10/28/google-redefines-car-gps-navigation-google-maps-navigation-android/>, October 28, 2009, 4 pages.

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

402268606v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	2A	7,082,469	07/25/2006	Gold *et al.*	709	231	
	2B	7,136,875	11/14/2006	Anderson *et al.*	707	104.1	
	2C	7,203,644	04/10/2007	Anderson *et al.*	704	246	
	2D	7,228,276	06/05/2007	Omote *et al.*	704	243	
	2E	7,301,093	11/27/2007	Sater *et al.*	84	615	
	2F	7,340,040	03/04/2008	Saylor *et al.*	379	67.1	
	2G	7,461,059	12/02/2008	Richardson *et al.*	707	5	
	2H	7,502,738	03/10/2009	Kennewick *et al.*	704	257	
	2I	7,620,549	11/17/2009	Di Cristo *et al.*	704	257	
	2J	7,634,409	12/15/2009	Kennewick *et al.*	704	257	
	2K	7,640,160	12/29/2009	Di Cristo *et al.*	704	257	
	2L	7,693,720	04/06/2010	Kennewick *et al.*	704	275	
	2M	2002/0069059	06/06/2002	Smith	704	257	
	2N	2002/0138248	09/26/2002	Corston-Oliver *et al.*	704	1	
	2O	2003/0088421	05/08/2003	Maes *et al.*	704	270.1	
	2P	2003/0135488	07/17/2003	Amir *et al.*	707	3	
	2Q	2005/0246174	11/03/2005	DeGolia	704	270	
	2R	2006/0206310	09/14/2006	Ravikumar *et al.*	704	9	
	2S	2007/0043574	02/22/2007	Coffman *et al.*	704	275	
	2T	2007/0055525	03/08/2007	Kennewick *et al.*	704	257	
	2U	2007/0186165	08/09/2007	Maislos *et al.*	715	728	
	2V	2007/0214182	09/13/2007	Rosenberg	707	104.1	
	2W	2007/0250901	10/25/2007	McIntire *et al.*	725	146	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

402268606v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	3A	2007/0265850	11/15/2007	Kennewick *et al.*	704	257	
	3B	2007/0299824	12/27/2007	Pan *et al.*	707	3	
	3C	2008/0103761	05/01/2008	Printz *et al.*	704	9	
	3D	2008/0115163	05/15/2008	Gilboa *et al.*	725	34	
	3E	2008/0133215	06/05/2008	Sarukkai	704	2	
	3F	2008/0189110	08/07/2008	Freeman *et al.*	704	251	
	3G	2008/0235023	09/25/2008	Kennewick *et al.*	704	257	
	3H	2008/0235027	09/25/2008	Cross	704	270.1	
	3I	2008/0319751	12/25/2008	Kennewick *et al.*	704	257	
	3J	2009/0117885	05/07/2009	Roth	455	414.3	
	3K	2009/0144271	06/04/2009	Richardson *et al.*	707	5	
	3L	2009/0150156	06/11/2009	Kennewick *et al.*	704	257	
	3M	2009/0171664	07/02/2009	Kennewick *et al.*	704	257	
	3N	2009/0299745	12/03/2009	Kennewick *et al.*	704	257	
	3O	2010/0023320	01/28/2010	Di Cristo *et al.*	704	9	
	3P	2010/0049501	02/25/2010	Kennewick *et al.*	704	9	
	3Q	2010/0049514	02/25/2010	Kennewick *et al.*	704	233	
	3R	2010/0057443	03/04/2010	Di Cristo *et al.*	704	9	
	3S	2010/0063880	03/11/2010	Atsmon *et al.*	705	14.53	
	3T	2010/0145700	06/10/2010	Kennewick *et al.*	704	257	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

402268606v1

Electronic Patent Application Fee Transmittal

Application Number:	11580926
Filing Date:	16-Oct-2006
Title of Invention:	System and method for a cooperative conversational voice user interface
First Named Inventor/Applicant Name:	Larry Baldwin
Filer:	Syed Jafar Ali/Miriam E. Nelsen
Attorney Docket Number:	090611-0357945

Filed as Small Entity

Utility under 35 USC 111(a) Filing Fees

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Miscellaneous:				
Submission- Information Disclosure Stmt	1806	1	180	180
		Total in USD ($)		**180**

Electronic Acknowledgement Receipt

EFS ID:	8349376
Application Number:	11580926
International Application Number:	
Confirmation Number:	6670
Title of Invention:	System and method for a cooperative conversational voice user interface
First Named Inventor/Applicant Name:	Larry Baldwin
Customer Number:	00909
Filer:	Syed Jafar Ali/Miriam E. Nelsen
Filer Authorized By:	Syed Jafar Ali
Attorney Docket Number:	090611-0357945
Receipt Date:	02-SEP-2010
Filing Date:	16-OCT-2006
Time Stamp:	18:19:26
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	Deposit Account
Payment was successfully received in RAM	$180
RAM confirmation Number	4695
Deposit Account	033975
Authorized User	
The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows: Charge any Additional Fees required under 37 C.F.R. Section 1.21 (Miscellaneous fees and charges)	

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		090611-0357945_Response_NFOA_IDS.pdf	1340478 a6f88375699e78d238b671a04e8dead492590a4d	yes	31

Multipart Description/PDF files in .zip description

Document Description	Start	End
Amendment/Req. Reconsideration-After Non-Final Reject	1	1
Specification	2	4
Claims	5	17
Applicant Arguments/Remarks Made in an Amendment	18	26
Transmittal Letter	27	28
Information Disclosure Statement (IDS) Filed (SB/08)	29	31

Warnings:

Information:

2	NPL Documents	090611-0357945_ARRINGTON.pdf	628568 eba655854391 44f0247bc17ca53fadcbf03e8c83	no	4

Warnings:

Information:

3	Fee Worksheet (PTO-875)	fee-info.pdf	30095 7f972713f36777cbd60c6bcdb635292880d718eb	no	2

Warnings:

Information:

Total Files Size (in bytes):	1999141

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

PTO/SB/06 (07-06)
Approved for use through 1/31/2007. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PATENT APPLICATION FEE DETERMINATION RECORD Substitute for Form PTO-875	Application or Docket Number 11/580,926	Filing Date 10/16/2006	☐ To be Mailed

APPLICATION AS FILED – PART I

	(Column 1)	(Column 2)	SMALL ENTITY ☒		OR	OTHER THAN SMALL ENTITY	
FOR	NUMBER FILED	NUMBER EXTRA	RATE ($)	FEE ($)		RATE ($)	FEE ($)
☐ BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A	N/A			N/A	
☐ SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A	N/A			N/A	
☐ EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A	N/A			N/A	
TOTAL CLAIMS (37 CFR 1.16(i))	minus 20 =	*	X $ =		OR	X $ =	
INDEPENDENT CLAIMS (37 CFR 1.16(h))	minus 3 =	*	X $ =			X $ =	
☐ APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $250 ($125 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).						
☐ MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))							
* If the difference in column 1 is less than zero, enter "0" in column 2.			TOTAL			TOTAL	

APPLICATION AS AMENDED – PART II

		(Column 1)		(Column 2)	(Column 3)	SMALL ENTITY		OR	OTHER THAN SMALL ENTITY	
	09/02/2010	CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA	RATE ($)	ADDITIONAL FEE ($)		RATE ($)	ADDITIONAL FEE ($)
AMENDMENT	Total (37 CFR 1.16(i))	* 42	Minus	** 42	= 0	X $26 =	0	OR	X $ =	
	Independent (37 CFR 1.16(h))	* 3	Minus	***3	= 0	X $110 =	0	OR	X $ =	
	☐ Application Size Fee (37 CFR 1.16(s))									
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE	0	OR	TOTAL ADD'L FEE	

		(Column 1)		(Column 2)	(Column 3)					
		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA	RATE ($)	ADDITIONAL FEE ($)		RATE ($)	ADDITIONAL FEE ($)
AMENDMENT	Total (37 CFR 1.16(i))	*	Minus	**	=	X $ =		OR	X $ =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	X $ =		OR	X $ =	
	☐ Application Size Fee (37 CFR 1.16(s))									
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.

** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".

*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".

The "Highest Number Previously Paid For" (Total or Independent) is the highest number found in the appropriate box in column 1.

Legal Instrument Examiner:
/TYWANA P. LOVELACE/

This collection of information is required by 37 CFR 1.16. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
11/580,926	10/16/2006	Larry Baldwin	090611-0357945	6670

909 7590 11/01/2010

PILLSBURY WINTHROP SHAW PITTMAN, LLP
P.O. BOX 10500
MCLEAN, VA 22102

EXAMINER	
YEN, ERIC L	

ART UNIT	PAPER NUMBER
2626	

NOTIFICATION DATE	DELIVERY MODE
11/01/2010	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

docket_ip@pillsburylaw.com
margaret.drosos@pillsburylaw.com

Office Action Summary	Application No.	Applicant(s)	
	11/580,926	BALDWIN ET AL.	
	Examiner	Art Unit	
	ERIC YEN	2626	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address --

Period for Reply

A SHORTENED STATUTORY PERIOD FOR REPLY IS SET TO EXPIRE 3 MONTH(S) OR THIRTY (30) DAYS, WHICHEVER IS LONGER, FROM THE MAILING DATE OF THIS COMMUNICATION.

- Extensions of time may be available under the provisions of 37 CFR 1.136(a). In no event, however, may a reply be timely filed after SIX (6) MONTHS from the mailing date of this communication.
- If NO period for reply is specified above, the maximum statutory period will apply and will expire SIX (6) MONTHS from the mailing date of this communication.
- Failure to reply within the set or extended period for reply will, by statute, cause the application to become ABANDONED (35 U.S.C. § 133). Any reply received by the Office later than three months after the mailing date of this communication, even if timely filed, may reduce any earned patent term adjustment. See 37 CFR 1.704(b).

Status

1)☒ Responsive to communication(s) filed on *02 September 2010*.

2a)☒ This action is **FINAL**. 2b)☐ This action is non-final.

3)☐ Since this application is in condition for allowance except for formal matters, prosecution as to the merits is closed in accordance with the practice under *Ex parte Quayle*, 1935 C.D. 11, 453 O.G. 213.

Disposition of Claims

4)☒ Claim(s) *1-42* is/are pending in the application.

4a) Of the above claim(s) _____ is/are withdrawn from consideration.

5)☐ Claim(s) _____ is/are allowed.

6)☒ Claim(s) *1-42* is/are rejected.

7)☐ Claim(s) _____ is/are objected to.

8)☐ Claim(s) _____ are subject to restriction and/or election requirement.

Application Papers

9)☐ The specification is objected to by the Examiner.

10)☐ The drawing(s) filed on _____ is/are: a)☐ accepted or b)☐ objected to by the Examiner.

Applicant may not request that any objection to the drawing(s) be held in abeyance. See 37 CFR 1.85(a).

Replacement drawing sheet(s) including the correction is required if the drawing(s) is objected to. See 37 CFR 1.121(d).

11)☐ The oath or declaration is objected to by the Examiner. Note the attached Office Action or form PTO-152.

Priority under 35 U.S.C. § 119

12)☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).

a)☐ All b)☐ Some * c)☐ None of:

1.☐ Certified copies of the priority documents have been received.

2.☐ Certified copies of the priority documents have been received in Application No. _____.

3.☐ Copies of the certified copies of the priority documents have been received in this National Stage application from the International Bureau (PCT Rule 17.2(a)).

* See the attached detailed Office action for a list of the certified copies not received.

Attachment(s)

1) ☒ Notice of References Cited (PTO-892)
2) ☐ Notice of Draftsperson's Patent Drawing Review (PTO-948)
3) ☒ Information Disclosure Statement(s) (PTO/SB/08) Paper No(s)/Mail Date _____.
4) ☐ Interview Summary (PTO-413) Paper No(s)/Mail Date. _____ .
5) ☐ Notice of Informal Patent Application
6) ☐ Other: _____.

DETAILED ACTION

Response to Amendment

1. In response to the Office Action mailed 6/16/10, applicant has submitted an amendment filed 9/2/10.

Claims 1-42 have been amended.

Response to Arguments

1. Applicant's arguments filed 9/2/10 have been fully considered but they are not persuasive.

Double patenting rejections are withdrawn because the previously referenced applications and patent do not include short-term information and long-term information where the short-term information includes knowledge about the utterance received at the voice input device in the current conversation.

Applicant's amendment to the independent claims do not overcome the previous 35 USC 112 1st and 2nd paragraph rejections because the rejections the amended claim language does not clarify what a "signifier" or "qualifier" is. The amended claim language only clarifies where a signifier/qualifier exists.

While it is clear that the signifiers and qualifiers are parts of an utterance (or SHOULD BE parts of an utterance, see Amendment, page 19) and are used to provide a "cooperative conversational voice user interface" (Amendment, page 20), this does not clarify or give any sort of idea as to what constitutes a "signifier" or "qualifier".

To repeat the relevant portion of the rejection:

It is not clear what applicant means by “contextual” "signifier" and "qualifier". Applicant’s Specification treats these two terms as having a meaning that is common knowledge and “required to perform a task in an identified context”. While it is common knowledge that signifiers signify things at least in the sense of indicating something meaningful and that qualifiers modify another term somehow (in the grammatical sense), there is nothing claimed or in the Specification describing what is being qualified/signified, what constitutes a signifier/qualifier (since they are both used in the Specification to describe information needed to complete a task), whether applicant is actually referring to a specific form of signifier/qualifier, whether signifiers/qualifiers are words in the utterances or other grammatical features (e.g. parts of speech), or how these entities (whatever they may be) are used to perform the function of completing a request.

Utterances can include any or all of the possible interpretations/characteristics in the previous paragraph. As currently claimed and describe in the Specification, even if “utterance” is used to clarify that signifiers and qualifiers are WORDS, it is not clear which words constitute signifiers and which words constitute qualifiers, since “signifier” and “qualifier” is not generally known and/or commonly used terms in language analysis. More importantly any number of grammatical characteristics and not just words can be “recognized in or inferred from the utterance” (e.g. that a particular word is in the utterance or that a particular word has a particular part of speech).

Therefore, the examiner maintains the previous 35 USC 112, 1st and 2nd paragraphs previously presented because the amended claim language still does not clarify what applicant means by “signifier” and “qualifier”, which are not terms conventionally used in language analysis.

Applicant argues that Kargman fails to disclose, teach or suggest at least the combination of features that include “receiving an utterance at a voice input device, wherein the voice input device receives the utterance in a current conversation with a user” and “determining an intended meaning for the utterance” wherein a conversational speech engine determines the intended meaning for the utterance from short-term shared knowledge about the current conversation and long-term shared knowledge about the user, including one or more past conversations with the user (Amendment, page 21). Applicant argues that this is because “although Kargman generally describes a system that can ‘generate an initial or supplemental prompt which is individually tailored for [a] particular customer’ and ‘tailor the voice recognition ‘grammar’ to that particular customer’ based on the customer’s order history,' suggesting certain words or tailoring voice recognition components to improve the likelihood that a particular utterance will be correctly recognized does not fairly disclose, teach, or suggest determining an 'intended meaning' for the utterance" (Amendment, page 21). Applicant provides an example of how the word traffic may have distinct meanings given different contexts, particularly in a movie context (the movie "Traffic") or a navigation context ("large quantity of cars moving slowly") (Amendment, pages 21-22).

Applicant, however, does not claim meaning of a single word that has different meanings in different contexts. This is not the only interpretation of the claim language and Kargman teaches more than just generally that prompts are tailored and that speech recognizers are adapted. Particularly, Kargman teaches determining what the utterance means. This is distinct from what a given word means because utterances are not words. They are spoken signals and/or sentences that contain words.

To repeat the relevant portion of the previous prior art rejection:

Kargman teaches:

determining an intended meaning for the at least one utterance based on the updated shared knowledge ("prior orders", paragraph 50; "speaking the word 'pizza", paragraph 44; "statistically… mushroom… shrimp", paragraph 48; "statistical database", paragraph 47; where the statistics updated at some point in time with statistical information [updated shared knowledge] about mushrooms and shrimp, for example, are used to determine that the user meant mushroom [intended meaning] instead of shrimp [less likely meaning])

In Kargman, an example is provided where the utterance (i.e. the spoken signal representing the speech which is to be recognized) could correspond to either "shrimp" or "mushroom" due to phonetic/acoustic similarity (e.g. if everything was muffled such that the system only noticed the "shr" sound). Shrimp and mushroom obviously have different meanings because one word refers to a fungus while another refers to a species of seafood. This is determined "statistically" (paragraph 48) which is based on past information because statistics are accumulated over time (i.e. someone wouldn't

know that the odds of a coin landing heads or tails was 50/50 if the coin was never tossed). Therefore, Kargman determines that a user meant to say (i.e. the UTTERANCE has an "intended meaning" of) shrimp, which has one meaning, OR mushroom, which has another meaning depending on statistical context information (i.e. that the person's speech had an intended meaning of "mushroom" because many people have ordered "mushroom pizza" before).

Therefore, even though the determination of intended meaning is not determination which meaning of a single word which has multiple meanings in multiple contexts is the speaker's intended meaning, as applicant intends, the teachings of Kargman still read on an alternative interpretation of the claim language (i.e. that the input speech signal really meant to say mushroom and not shrimp, which are two different concepts with different meanings).

Therefore, since applicant does not claim "determining the intended meaning that the user had in speaking any words recognized in the utterance" or more specifically "determination which meaning of a single word which has multiple meanings in multiple contexts is the speaker's intended meaning", even if Kargman makes "statistical guesses or suggestions to avoid mismatches" by "pre-configuring a voice recognition system based on a history" (Amendment, page 22), Kargman still teaches the claim limitation under a different interpretation where ambiguity is resolved and one of at least two confusable words EACH WITH DIFFERENT MEANINGS is selected as the word that the system thinks the user meant/intended to say.

Applicant's remaining arguments are directed to the secondary references not teaching "determining the intended meaning". Since the secondary references are not applied to teach this limitation, the examiner maintains the previous prior art rejections under the secondary references, directed to what the secondary references were previously applied to teach.

Claim Objections

2. Claims 6, 11, 12, 20, 25, 26, 34, 39, 40, are objected to because of the following informalities:

Claims 6, 20, 34, recite "wherein conversational speech engine further determines" which presumably refers back to the conversational speech engine.

As per Claims 11, 25, 39, they recite "adapted contextual signifiers, qualifiers, or grammar syntax" which are limitations in claims 10, 24, and 38, but are not present in the independent claims.

As per Claim 12, 26, 40, the claims recite "remove the initial interpretation from the multiple possible interpretation" which should recites "multiple possible interpretations".

3. Appropriate correction is required.

Claim Rejections - 35 USC § 112

1. The following is a quotation of the first paragraph of 35 U.S.C. 112:

> The specification shall contain a written description of the invention, and of the manner and process of making and using it, in such full, clear, concise, and exact terms as to enable any person skilled in the art to which it pertains, or with which it is most nearly connected, to make and use the same and shall set forth the best mode contemplated by the inventor of carrying out his invention.

2. Claims 8-9, 22-23, 36-37 are rejected under 35 U.S.C. 112, first paragraph, as failing to comply with the written description requirement. The claim(s) contains subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention.

As per Claims 8, 22, 36, Applicant's amended claim language recites "identifying a context for the current conversation from the one or more INFERRED contextual signifiers or qualifiers", where the inferred contextual signifiers or qualifiers are missing from the utterance.

Applicant's Specification does recite determining contexts based on updated short-term and long-term updated knowledge (paragraph 33) and previous utterances and/or requests (paragraph 35), but the Specification does not teach where context is determined from INFERRED MISSING information.

Paragraph 7 also seems to suggest that the inference is based on context, rather than the context being based on the inferences.

Therefore, identifying context from inferred previously-missing contextual signifiers/qualifiers is not supported by the original Specification.

Claims 9, 23, 37, recite assigning "an implicit degree of certainty to the intended meaning determined for the utterance in response to determining that the utterance contains insufficient information to generate the response" which are not supported by the original Specification.

Applicant's specification describes bits and pieces of the language but not combined in the manner claimed and there are certain elements of the claim language missing from the Specification.

Paragraph 7 describes determining missing information which suggests the determination that there is insufficient information.

Paragraph 49 describes "when the current utterance contains insufficient information to complete a request or task", which describes "in response to determining that the utterance contains insufficient information to generate the response".

Paragraph 50 also describes implicit hypotheses each having a corresponding degree of certainty, but this does not support assigning an implicit DEGREE OF CERTAINTY because degrees of certainty are not hypotheses. As claimed, there is something about the degree of certainty which is implicit (it is not clear how this is the case). This also seems to support the assigning of a degree of certainty to the intended meaning since one of the hypotheses is eventually determined as the intended meaning and the claim language does not say ONLY the intended meaning is assigned the degree of certainty, however the claim language's intended interpretation appears to be assigning the implicit degree of certainty AFTER a meaning is determined to be the intended meaning (since "determined for the utterance" heavily suggests that the

determining is already performed). This interpretation is not supported by the Specification because the Specification describes assigning degrees of certainty to hypotheses before determining the intended meaning.

3. Claims 7-8, 10, 21-22, 24, 35-36, 38 rejected under 35 U.S.C. 112, first paragraph, as failing to comply with the enablement requirement. The claim(s) contains subject matter which was not described in the specification in such a way as to enable one skilled in the art to which it pertains, or with which it is most nearly connected, to make and/or use the invention.

It is not clear what applicant means by “contextual” "signifier" and "qualifier". Applicant’s Specification treats these two terms as having a meaning that is common knowledge and “required to perform a task in an identified context”. While it is common knowledge that signifiers signify things at least in the sense of indicating something meaningful and that qualifiers modify another term somehow (in the grammatical sense), there is nothing claimed or in the Specification describing what is being qualified/signified, what constitutes a signifier/qualifier (since they are both used in the Specification to describe information needed to complete a task), whether applicant is actually referring to a specific form of signifier/qualifier, whether signifiers/qualifiers are words in the utterances or other grammatical features (e.g. parts of speech), or how these entities (whatever they may be) are used to perform the function of completing a request.

Since the scope of these terms is unclear, the claims are vague and indefinite, and by consequence not enabled because one of ordinary skill in the art logically can't make and use something when he doesn't know what some of the parts of the invention are.

4.

1. The following is a quotation of the second paragraph of 35 U.S.C. 112:

> The specification shall conclude with one or more claims particularly pointing out and distinctly claiming the subject matter which the applicant regards as his invention.

2. Claims 7-10, 21-24, 35-38, are rejected under 35 U.S.C. 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention.

As per Claims 7-8, 10, 21-88, 24, 35-36, 38:

It is not clear what applicant means by “contextual” "signifier" and "qualifier". Applicant’s Specification treats these two terms as having a meaning that is common knowledge and “required to perform a task in an identified context”. While it is common knowledge that signifiers signify things at least in the sense of indicating something meaningful and that qualifiers modify another term somehow (in the grammatical sense), there is nothing claimed or in the Specification describing what is being qualified/signified, what constitutes a signifier/qualifier (since they are both used in the Specification to describe information needed to complete a task), whether applicant is actually referring to a specific form of signifier/qualifier, whether signifiers/qualifiers are words in the utterances or other grammatical features (e.g. parts of speech), or how

these entities (whatever they may be) are used to perform the function of completing a request.

Since the scope of these terms is unclear, the claims are vague and indefinite.

As per Claims 8, 22, 36, as discussed above in the new matter rejection, Applicant's amended claim language recites “identifying a context for the current conversation from the one or more INFERRED contextual signifiers or qualifiers”, where the inferred contextual signifiers or qualifiers are missing from the utterance.

It is not clear if applicant actually intended to claim this limitation since it is not supported and constitutes new matter.

Also the order of the limitations heavily suggests that the inferences are made without determining the context, which does not make sense because inferences typically require some sort of guide, or otherwise the inferences would be made at random and have a high likelihood of being incorrect.

As discussed above in the new matter rejection:

Claims 9, 23, 37, recite assigning "an implicit degree of certainty to the intended meaning determined for the utterance in response to determining that the utterance contains insufficient information to generate the response" which are not supported by the original Specification.

As claimed, there is something about the degree of certainty which is implicit (it is not clear how this is the case). This also seems to support the assigning of a degree of

certainty to the intended meaning since one of the hypotheses is eventually determined as the intended meaning and the claim language does not say ONLY the intended meaning is assigned the degree of certainty, however the claim language's intended interpretation appears to be assigning the implicit degree of certainty AFTER a meaning is determined to be the intended meaning (since "determined for the utterance" heavily suggests that the determining is already performed). This interpretation is not supported by the Specification because the Specification describes assigning degrees of certainty to hypotheses before determining the intended meaning.

Since it is not clear what constitutes an “implicit degree of certainty” and whether the claim language intends to assign this degree of certainty to the intended meaning before or after the intended meaning is determined, the amended claim language is not clear.

Claim Rejections - 35 USC § 103

1. The following is a quotation of 35 U.S.C. 103(a) which forms the basis for all obviousness rejections set forth in this Office action:

> (a) A patent may not be obtained though the invention is not identically disclosed or described as set forth in section 102 of this title, if the differences between the subject matter sought to be patented and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which said subject matter pertains. Patentability shall not be negatived by the manner in which the invention was made.

2. Claims 1, 4-7, 10-11, 14-15, 18-21, 24-25, 28-29, 32-35, 38-39, 42 are rejected under 35 U.S.C. 103(a) as being unpatentable over Kargman (US 2005/0015256).

As per Claims 1, 15, 29, Kargman teaches a method (and corresponding computer-readable medium of Claim 15, and system of Claim 29) for providing a cooperativevoice user interface (“customer repeatedly orders… initial or supplemental prompt… individually tailored”, paragraph 50; where the consistent ordering of a particular item is based on information shared between the user and system [i.e. what does the customer frequently order] to advance the mutual goal of taking down the customer’s food order [see applicant's interpretation of "cooperative" in Specification, paragraph 2])

comprising:

receiving an utterance at a voice input device, wherein the voice input device receives the utterance in a current conversation with a user (“calls to place an order”, paragraph 50; “speaking the word ‘pizza”, paragraph 44; “statistically… mushroom… shrimp”, paragraph 48; speaking is uttering an input, and the speech itself is an utterance, which occurs during a current conversation when a user speaks with a machine to obtain a pizza order)

accumulating short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received at the voice input device in the current conversation (“whether they wish to order ‘pizza’”, paragraph 44; wishing to order pizza is knowledge about an utterance in the current conversation where the user speaks that he wants to order a pizza [i.e. short-term because the conversation occurs in the "short-term", i.e. that the conversation is current and not occurring a distant “long-term” time] and this knowledge

is “shared” by the user to the machine, and is then information that both the user and machine possesses [i.e. it is “shared])

accumulating long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user (“whether they wish to order ‘pizza’”, paragraph 44; “order history", paragraph 50; “prior orders”, paragraph 50; “speaking the word ‘pizza”, paragraph 44; “statistically… mushroom… shrimp”, paragraph 47-48; where prior orders are past conversations between the system and customer obtained in the “long-term” [i.e. at a distant/not-current time in the past]; where in Kargman the example for where statistical information is at least obviously accumulated and updated over time [since if people start ordering shrimp consistently it would be frustrating to the users to be constantly overridden by mushroom])

determining an intended meaning for the utterance, wherein a conversational speech engine determines the intended meaning for the utterance from the short-term shared knowledge and the long-term shared knowledge (“prior orders”, paragraph 50; “speaking the word ‘pizza”, paragraph 44; “statistically… mushroom… shrimp”, paragraph 48; “statistical database”, paragraph 47; where the short term knowledge is the knowledge that the user is ordering some sort of pizza, and using the long-term information can be using the prior order to determine that the user "meant" pepperoni pizza when he said "pizza" because he repeatedly ordered pepperoni. Alternatively since it is obvious to continue to accumulate statistical information just in case things change and people do start ordering shrimp pizza frequently, it is obvious to use the

statistics that obviously contain prior user orders as part of the statistics [since information about prior orders are the information represented in the statistics] to determine whether a user meant shrimp or mushroom)

generating a response to the utterance, wherein the conversational speech engine generates the response to the utterance based on the intended meaning determined from the short-term shared knowledge and the long-term shared knowledge ("speak the order back… for their confirmation", paragraph 49; "tailored", paragraph 50; where the tailored prompt is based on the system determining that the user meant "pepperoni pizza" when the user said "pizza", or alternatively, where speaking the order back is a system response and depends on what the system interpreted the order to be [determined intended meaning] based on the statistics which as discussed above obviously include what the user says over time [to accurately represent the rate at which people order mushroom with pizza]).

As per Claims 4, 18, 32, Kargman teaches wherein accumulating the long-term shared knowledge about the user includes updating one or more long-term profiles associated with the user to include information about the utterance received in the current conversation ("statistical database", paragraphs 47-48; "prior orders", paragraph 50; where the customer's current order becomes a "prior order" after it is completed).

As per Claims 5, 19, 33, Kargman teaches/suggests wherein determining the intended meaning for the utterance includes identifying a conversation type for the

current conversation (“pizza”, paragraph 44; where the selection of pizza defines a conversation type to be a pizza conversation instead of a side item conversation; “mushroom”, paragraph 48; where the determination of pizza [conversation type] at least obviously affects which statistics the system looks up to determine the intended meaning to be mushroom instead of shrimp [e.g. there would likely be a different set of statistics for side items since people probably order popcorn shrimp more than they order mushrooms as a side])

based on one or more of a conversational goal for the utterance, roles for one or more participants in the current conversation, or an allocation of information among one or more participants in the current conversation (“mushroom”, paragraph 48; where goal of ordering a pizza with a topping is one of the listed items [“one or more”]).

As per Claims 6, 20, 34, Kargman teaches generating a hypothesis having a degree of certainty about an intent that the user had in speaking the utterance, wherein conversational speech engine further determines the intended meaning for the utterance from the hypothesis about the intent that the user had in speaking the utterance (“mushroom… shrimp”, paragraph 48; “large thin crust pepperoni pizza”, paragraph 46; where mushroom and shrimp are two hypotheses and mushroom is statistically more likely/certain, and it is at least obvious that the system eventually figures out precisely what the user wants in the example).

As per Claims 7, 21, 35, Kargman teaches/suggests wherein the conversational speech engine assigns the degree of certainty to the hypothesis based on a context that the conversational speech engine identifies for the current conversation from one or more contextual signifiers or qualifiers recognized in the utterance (“one… large… thin…pizza”, paragraph 46; “statistics… when a customer orders a pizza…much more likely that the second ingredient is 'mushroom'", paragraphs 47-48; where, for example, pizza signifies that the user wants a pizza and obviously assigns high probability [degree of certainty] to mushroom based on the fact that the user is talking about pizza [the pizza context versus the seafood context] identified based on the fact that the user spoke/uttered “pizza”).

As per Claims 10, 24, 38, Kargman teaches/suggests wherein the response to the utterance includes one or more words having contextual signifiers, qualifiers, or grammar syntax adapted to reflect contextual signifiers, qualifiers, or grammar syntax in or inferred from the utterance (“speak the order back”, paragraph 49; “one large pepperoni pizza, thin crust”, paragraph 46; where at the very least speaking/repeating the order “one large pepperoni pizza, thin crust” back to the user oncorporates words in the response that reflect the actual items like pizza and pepperoni which signify what the order is in the pizza context [contextual signifiers] and also the grammar characteristics of the input order that make it grammatically correct [more clearly so in an example where the spoken input is "one large thin [crust] pizza]).

As per Claims 11, 25, 39, Kargman teaches wherein the one or more words having the adapted contextual signifiers, qualifiers, or grammar syntax frame the response to influence a subsequent utterance in the current conversation ("confirmation", paragraph 49; where prompting for a confirmation influences the user to either confirm or deny the order).

As per Claim 14, 28, 42, Kargman teaches/suggests wherein the conversational speech engine generates the response in a multi-modal output that includes one or more non-voice outputs presented to the user in response to the utterance or one or more tasks executed to process a request identified from the intended meaning for the utterance ("displayed to permit the customer to confirm", Abstract; "display… pizza or food item ordered", paragraph 32; "speak the order back… for their confirmation", paragraph 49; where both the speaking audibly by the system and the display [non-voice] are described as being used for confirmation and so Kargman at least obviously suggests the use of both display and audible output [multiple modes of communication] to communicate with the user).

3. Claims 2-3, 16-17, 30-31, are rejected under 35 U.S.C. 103(a) as being unpatentable over Kargman, as applied to claims 1, 15, and 29, above, and further in view of Morin et al. (US 5,748,841), hereafter Morin.

As per Claim 2, 16, 30, Kargman fails to teach wherein accumulating short-term shared knowledge about the current conversation includes populating a short-term context stack with information about the utterance received in the current conversation.

Morin teaches/suggests wherein accumulating short-term shared knowledge about the current conversation includes populating a short-term context stack with information about the utterance received in the current conversation ("dialogue being run…dialogue history… dialogue turns or exchanges already made and the context in which they occur", col. 3, lines 6-24; "immediate context", col. 15, lines 37-47; "knowledge of the ongoing dialogue... dialogue history and immediate dialogue", col. 4, lines 5-16; "short-term... immediate focus of the dialogue", col. 10, lines 1-15; where having an immediate context means the system at least obviously stores information about an immediate/short-term context at some point, and this memory location can be interpreted as a short-term context memory/stack [where stack is, in computer contexts, a memory location populated by information to be processed]; as described in Morin, the context and dialog history is derived from utterances in the current dialog).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of invention to **use a known technique to improve similar devices in the same way** because Kargman teaches a **base** machine-user dialog device **upon which the claimed** contextual stack **can be seen as an improvement** by incorporating more information into facilitating a dialog. Morin teaches **a comparable** machine-user dialog device **that has been improved in the same way as the claimed invention** by having a short-term context [immediate context in Morin] to facilitate the dialog. **One of**

ordinary skill in the art could have applied the known improvement technique in the same way to the base device by adding a contextual information/dialogue history to the information already used in Kargman to facilitate improved dialog using information derived from past communications, **to obtain the predictable results** of a machine–user dialog system that uses information from past utterances to determine future dialog behavior.

As per Claims 3, 17, 31, Kargman fails to teach wherein accumulating the short-term shared knowledge about the current conversation further includes expiring the information about the utterance from the short-term context stack after a predetermined amount of time.

Morin teaches/suggests wherein accumulating the short-term shared knowledge about the current conversation further includes expiring the information about the utterance from the short-term context stack after a predetermined amount of time ("dialogue being run…dialogue history… dialogue turns or exchanges already made and the context in which they occur", col. 3, lines 6-24; "immediate context", col. 15, lines 37-47; "knowledge of the ongoing dialogue... dialogue history and immediate dialogue", col. 4, lines 5-16; "short-term... immediate focus of the dialogue", col. 10, lines 1-15; where the predetermined time is the amount of time it takes, once the system realizes the immediate context changed, to change the previous immediate context [short-term context stack] to the new immediate context [which can be interpreted as new accumulated short-term shared knowledge]; "immediate context

refers to what is logically expected from the user at this time", col. 3, lines 6-24; "time-out", col. 9, lines 34-57; where the time-out at least obviously changes the context and one obvious result of a timeout is termination [e.g. if the user just left], which would expire the context and all other current-dialog information upon termination).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of invention to **use a known technique to improve similar devices in the same way** because Kargman teaches a **base** machine-user dialog device **upon which the claimed** contextual stack **can be seen as an improvement** by incorporating more information into facilitating a dialog. Morin teaches **a comparable** machine-user dialog device **that has been improved in the same way as the claimed invention** by having a short-term context [immediate context in Morin] to facilitate the dialog. **One of ordinary skill in the art could have applied the known improvement technique in the same way to the base device** by adding a contextual information/dialogue history to the information already used in Kargman to facilitate improved dialog using information derived from past communications, **to obtain the predictable results** of a machine–user dialog system that uses information from past utterances to determine future dialog behavior.

4. Claims 12, 26, 40 are rejected under 35 U.S.C. 103(a) as being unpatentable over Kargman, as applied to Claims 1, 15, and 29, above, and further in view of Bielby et al. (US 5,488,652), hereafter Bielby.

As per Claim 12, 26, 40, Kargman fails to teach wherein determining the intended meaning for the utterance includes generating multiple possible interpretations for the utterance, wherein an initial interpretation for the utterance comprises one of the multiple possible interpretations having a highest degree of certainty, and updating the short-term shared knowledge about the current conversation to remove the initial interpretation from the multiple possible interpretation in response to determining that the initial interpretation was incorrect, wherein the intended meaning determined from the short-term shared knowledge and the long-term shared knowledge comprises one of the multiple possible interpretations having a next highest degree of certainty.

Bielby suggests wherein determining the intended meaning for the utterance includes generating multiple possible interpretations for the utterance, wherein an initial interpretation for the utterance comprises one of the multiple possible interpretations having a highest degree of certainty (“called entity name corresponding to a frequently called listing is recognized", col. 18, lines 1-30; “recognized called entity name is incorrect”, col. 18, lines 31-44; where the intended meaning of a particular name is noted as incorrect)

and updating the short-term shared knowledge about the current conversation to remove the initial interpretation from the multiple possible interpretation in response to determining that the initial interpretation was incorrect, wherein the intended meaning determined from the short-term shared knowledge and the long-term shared knowledge comprises one of the multiple possible interpretations having a next highest degree of certainty (“called entity name corresponding to a frequently called listing is recognized",

col. 18, lines 1-30; “announce the next best candidate”, col. 18, lines 31-44; moving on to the next best candidate removes the previous candidate from consideration and thus the output candidate [with its corresponding intended meaning] comes form one of the remaining candidates which are most frequently [highest degree of certainty] called people)

Therefore, it would have been obvious to one of ordinary skill in the art to **use a known technique of** moving onto the next possible candidate **to improve similar devices** because Kargman teaches a base speech recognition device that can have multiple interpretations (shrimp and mushroom, as discussed above) **upon which the claimed** determination of a second interpretation **could be seen as an improvement** by allowing the system to adjust for recognition errors. Bielby **contained a comparable** user-machine interaction device where multiple interpretations of input speech can be made **and has been improved in the same way as the claimed invention** by incorporating a process to accommodate recognition-errors. **One of ordinary skill in the art could have applied the known ‘improvement technique' in the same way to the 'base' device** of Kargman by asking the user if he/she meant shrimp if he/she says the system did not recognize mushroom properly, for example, **which would obtain the predictable results** of a machine-user voice interactive system which has a process that can adapt for recognition errors as explicitly described in Bielby.

5. Claims 13, 27, 41, rejected under 35 U.S.C. 103(a) as being unpatentable over Kargman, as applied to Claims 1, 15, 29, above, and further in view of Gorin et al. (US 6,751,591), hereafter Gorin.

As per Claims 13, 27, 41, Kargman fails to teach wherein the user speaks the utterance in a multi-modal input that further includes one or more non-voice inputs relating to the utterance.

Gorin teaches/suggests wherein the user speaks the utterance in a multi-modal input that further includes one or more non-voice inputs relating to the utterance ("verbally, nonverbally, multimodally", col. 3, lines 20-33; "recognizer… user's input communication", col. 3, lines 47-60; "additional multimodal input such as touchtone", col. 6, lines 15-27; where Gorin teaches verbal [voice utterance] and DTMF [non-voice] combinations).

Therefore, it would have been obvious to one of ordinary skill in the art at the time of invention to modify Kargman to include the teaching of Gorin of wherein the user speaks the utterance in a multi-modal input that further includes one or more non-voice inputs relating to the utterance, in order to increase likelihood of understanding, as described by Gorin (col. 6, lines 15-27).

Allowable Subject Matter

6. The following is a statement of reasons for the indication of allowable subject matter:

The prior art of record does not teach identifying a context for the current conversation from one or more inferred contextual signifiers or qualifiers missing from the utterance, inferred in response to determining that the utterance contains insufficient information to generate the response, as in Claims 8, 22, 36, and also does not teach these limitations in Claims 9, 23, and 37, which incorporate the limitations based on their dependence on Claims 8, 22, and 36.

Chai et al. (US 6,829,603) suggests wherein determining the intended meaning for the utterance includes inferring one or more contextual signifiers or qualifiers missing from the utterance in response to determining that the utterance contains insufficient information to generate the response ("instantiated if the values of slots… item name… quantity… model… missing slot 807 values… inferred from context", col. 11, line 57 - col. 12, line 27; where determining there are missing slots determines that there is insufficient information in the utterance when only one utterance has been spoken at the time [thus filling only a limited number of slots] and the slots signify certain types of information and are "contextual" in the sense that they are inferred from context and thus have some attribute associated with that context).

However, Chai does not teach where a context is identified from the inferred missing information.

Conclusion

7. Applicant's amendment necessitated the new ground(s) of rejection presented in this Office action. Accordingly, **THIS ACTION IS MADE FINAL**. See MPEP

§ 706.07(a). Applicant is reminded of the extension of time policy as set forth in 37 CFR 1.136(a).

A shortened statutory period for reply to this final action is set to expire THREE MONTHS from the mailing date of this action. In the event a first reply is filed within TWO MONTHS of the mailing date of this final action and the advisory action is not mailed until after the end of the THREE-MONTH shortened statutory period, then the shortened statutory period will expire on the date the advisory action is mailed, and any extension fee pursuant to 37 CFR 1.136(a) will be calculated from the mailing date of the advisory action. In no event, however, will the statutory period for reply expire later than SIX MONTHS from the date of this final action.

Any inquiry concerning this communication or earlier communications from the examiner should be directed to ERIC YEN whose telephone number is (571)272-4249. The examiner can normally be reached on M-F 7:30-4:00.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Richemond Dorvil can be reached on 571-272-7602. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

EY 10/27/10
/Eric Yen/
Examiner, Art Unit 2626

Notice of References Cited	Application/Control No. 11/580,926	Applicant(s)/Patent Under Reexamination BALDWIN ET AL.	
	Examiner ERIC YEN	Art Unit 2626	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
*	A	US-6,829,603	12-2004	Chai et al.	1/1
*	B	US-6,751,591	06-2004	Gorin et al.	704/257
	C	US-			
	D	US-			
	E	US-			
	F	US-			
	G	US-			
	H	US-			
	I	US-			
	J	US-			
	K	US-			
	L	US-			
	M	US-			

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Index of Claims	Application/Control No.	Applicant(s)/Patent Under Reexamination
	11580926	BALDWIN ET AL.
	Examiner	Art Unit
	ERIC YEN	2626

✓	Rejected	-	Cancelled	N	Non-Elected	A	Appeal
=	Allowed	÷	Restricted	I	Interference	O	Objected

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIM		DATE								
Final	Original	06/15/2010	10/27/2010							
	1	✓	✓							
	2	✓	✓							
	3	✓	✓							
	4	✓	✓							
	5	✓	✓							
	6	✓	✓							
	7	✓	✓							
	8	✓	✓							
	9	✓	✓							
	10	✓	✓							
	11	✓	✓							
	12	✓	✓							
	13	✓	✓							
	14	✓	✓							
	15	✓	✓							
	16	✓	✓							
	17	✓	✓							
	18	✓	✓							
	19	✓	✓							
	20	✓	✓							
	21	✓	✓							
	22	✓	✓							
	23	✓	✓							
	24	✓	✓							
	25	✓	✓							
	26	✓	✓							
	27	✓	✓							
	28	✓	✓							
	29	✓	✓							
	30	✓	✓							
	31	✓	✓							
	32	✓	✓							
	33	✓	✓							
	34	✓	✓							
	35	✓	✓							
	36	✓	✓							

Index of Claims	Application/Control No.	Applicant(s)/Patent Under Reexamination
	11580926	BALDWIN ET AL.
	Examiner	Art Unit
	ERIC YEN	2626

✓	Rejected	-	Cancelled	N	Non-Elected	A	Appeal
=	Allowed	÷	Restricted	I	Interference	O	Objected

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIM		DATE								
Final	Original	06/15/2010	10/27/2010							
	37	✓	✓							
	38	✓	✓							
	39	✓	✓							
	40	✓	✓							
	41	✓	✓							
	42	✓	✓							

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	11580926	BALDWIN ET AL.
	Examiner	Art Unit
	ERIC YEN	2626

SEARCHED			
Class	Subclass	Date	Examiner
704	270-275	6/15/10	EY

SEARCH NOTES		
Search Notes	Date	Examiner
EAST search (attached)	6/15/10	EY
EAST Search (attached)	10/27/10	EY

INTERFERENCE SEARCH			
Class	Subclass	Date	Examiner

EAST Search History

EAST Search History (Prior Art)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
L1	6	"704".clas. and (multi-modal multimodal) and (order$4 near5 food)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 17:09
L2	34	"704".clas. and ((multi-modal multimodal) near5 non-voice)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 17:19
L3	9	"704".clas. and ((multi-modal multimodal) near5 press$4)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 17:21
L4	25	"704".clas. and ((multi-modal multimodal) near5 button$1)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 17:21
L5	354	"704".clas. and ((multi-modal multimodal) near5 input$1)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 17:23
L6	32	"704".clas. and ((multi-modal multimodal) near5 input$1 near5 combination)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 17:23
L7	2	"704".clas. and (infer$4 near5 (missing adj information))	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 17:40
L8	1	"20030216919".pn. and infer$4	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 17:43
L9	1	"20030216919".pn. and missing	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 17:43
L10	1	11/580926 and infer$4 and implicit$2	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 17:46

L11	1	11/580926 and infer$4 and implicit$2 and missing and insufficient	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 17:48
L12	1	11/580926 and infer$4 and implicit$2 and missing and insufficient and certainty	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 17:49
L13	168	"704".clas. and (infer$4 near5 context$1)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 18:27
L14	1	"704".clas. and (infer$4 near5 context$1 near5 missing)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 18:29
L15	1	"6829603".pn. and context	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 18:48
L16	1	"704".clas. and (missing near5 determin$4 near5 context$1)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 19:01
L17	1	"704".clas. and (incomplete near5 determin$4 near5 context$1)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 19:02
L18	4	11/580926 and (infer$4 near5 context)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 19:05
L19	4	11/580926 and (infer$4 same context$4)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 19:07
L20	1	"6829603".pn.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 19:42
L21	1	"7620549".pn.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 19:45
L22	0	"7620549".pn. and term	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 19:46

L23	3	11/954064 and term	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 19:46
L24	1	11/954064 and term.clm.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 19:48
L25	0	11/954064 and (conversationterm).clm.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 19:49
L26	1	11/954064 and (conversation term).clm.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 19:49
L27	0	"7620549".pn. and (term conversation).in.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 19:49
L28	0	12/127343 and (term conversation).in.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 19:50
L29	0	12/705311 and (term conversation).in.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 19:50
S1	1	"20050015256".pn.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 09:21
S2	4	11/580926 and infer$4	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 10:31
S3	1	"5488652".pn.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2010/10/27 11:00

EAST Search History (Interference)

< This search history is empty>

10/27/2010 8:18:04 PM

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
/E.Y./	1A	5,617,407	04/01/1997	Bareis	369	275.3	
	1B	5,797,112	08/18/1998	Komatsu *et al.*	701	201	
	1C	5,895,464	04/20/1999	Bhandari *et al.*	707	3	
	1D	6,144,938	11/07/2000	Surace *et al.*	704	257	
	1E	6,226,612	05/01/2001	Srenger *et al.*	704	256	
	1F	6,288,319	09/11/2001	Catona	84	609	
	1G	6,308,151	10/23/2001	Smith	704	235	
	1H	6,381,535	04/30/2002	Durocher *et al.*	701	202	
	1I	6,385,596	05/07/2002	Wiser *et al.*	705	51	
	1J	6,430,531	08/06/2002	Polish	704	257	
	1K	6,466,899	10/15/2002	Yano *et al.*	704	1	
	1L	6,556,970	04/29/2003	Sasaki *et al.*	704	257	
	1M	6,581,103	06/17/2003	Dengler	709	231	
	1N	6,678,680	01/13/2004	Woo	707	6	
	1O	6,801,604	10/05/2004	Maes *et al.*	379	88.17	
	1P	6,801,893	10/05/2004	Backfried *et al.*	704	257	
	1Q	6,833,848	12/21/2004	Wolff *et al.*	345	719	
	1R	6,856,990	02/15/2005	Barile *et al.*	707	10	
	1S	6,865,481	03/08/2005	Kawazoe *et al.*	701	211	
	1T	6,877,134	04/05/2005	Fuller *et al.*	715	500.1	
	1U	6,950,821	09/27/2005	Faybishenko *et al.*	707	10	
	1V	6,973,387	12/06/2005	Masclet *et al.*	701	211	
/E.Y./	1W	7,016,849	03/21/2006	Arnold *et al.*	704	275	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

/E.Y./	1X	ARRINGTON, Michael, "Google Redefines GPS Navigation Landscape: Google Maps Navigation for Android 2.0", TechCrunch, printed from the Internet <http://www.techcrunch.com/2009/10/28/google-redefines-car-gps-navigation-google-maps-navigation-android/>, October 28, 2009, 4 pages.

EXAMINER	DATE CONSIDERED
/E. Yen/	10/27/2010

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

402268606v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
/E.Y./	2A	7,082,469	07/25/2006	Gold *et al.*	709	231	
	2B	7,136,875	11/14/2006	Anderson *et al.*	707	104.1	
	2C	7,203,644	04/10/2007	Anderson *et al.*	704	246	
	2D	7,228,276	06/05/2007	Omote *et al.*	704	243	
	2E	7,301,093	11/27/2007	Sater *et al.*	84	615	
	2F	7,340,040	03/04/2008	Saylor *et al.*	379	67.1	
	2G	7,461,059	12/02/2008	Richardson *et al.*	707	5	
	2H	7,502,738	03/10/2009	Kennewick *et al.*	704	257	
	2I	7,620,549	11/17/2009	Di Cristo *et al.*	704	257	
	2J	7,634,409	12/15/2009	Kennewick *et al.*	704	257	
	2K	7,640,160	12/29/2009	Di Cristo *et al.*	704	257	
	2L	7,693,720	04/06/2010	Kennewick *et al.*	704	275	
	2M	2002/0069059	06/06/2002	Smith	704	257	
	2N	2002/0138248	09/26/2002	Corston-Oliver *et al.*	704	1	
	2O	2003/0088421	05/08/2003	Maes *et al.*	704	270.1	
	2P	2003/0135488	07/17/2003	Amir *et al.*	707	3	
	2Q	2005/0246174	11/03/2005	DeGolia	704	270	
	2R	2006/0206310	09/14/2006	Ravikumar *et al.*	704	9	
	2S	2007/0043574	02/22/2007	Coffman *et al.*	704	275	
	2T	2007/0055525	03/08/2007	Kennewick *et al.*	704	257	
	2U	2007/0186165	08/09/2007	Maislos *et al.*	715	728	
	2V	2007/0214182	09/13/2007	Rosenberg	707	104.1	
/E.Y./	2W	2007/0250901	10/25/2007	McIntire *et al.*	725	146	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED
/E. Yen/	10/27/2010

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

402268606v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
/E.Y./	3A	2007/0265850	11/15/2007	Kennewick *et al.*	704	257	
	3B	2007/0299824	12/27/2007	Pan *et al.*	707	3	
	3C	2008/0103761	05/01/2008	Printz *et al.*	704	9	
	3D	2008/0115163	05/15/2008	Gilboa *et al.*	725	34	
	3E	2008/0133215	06/05/2008	Sarukkai	704	2	
	3F	2008/0189110	08/07/2008	Freeman *et al.*	704	251	
	3G	2008/0235023	09/25/2008	Kennewick *et al.*	704	257	
	3H	2008/0235027	09/25/2008	Cross	704	270.1	
	3I	2008/0319751	12/25/2008	Kennewick *et al.*	704	257	
	3J	2009/0117885	05/07/2009	Roth	455	414.3	
	3K	2009/0144271	06/04/2009	Richardson *et al.*	707	5	
	3L	2009/0150156	06/11/2009	Kennewick *et al.*	704	257	
	3M	2009/0171664	07/02/2009	Kennewick *et al.*	704	257	
	3N	2009/0299745	12/03/2009	Kennewick *et al.*	704	257	
	3O	2010/0023320	01/28/2010	Di Cristo *et al.*	704	9	
	3P	2010/0049501	02/25/2010	Kennewick *et al.*	704	9	
	3Q	2010/0049514	02/25/2010	Kennewick *et al.*	704	233	
	3R	2010/0057443	03/04/2010	Di Cristo *et al.*	704	9	
	3S	2010/0063880	03/11/2010	Atsmon *et al.*	705	14.53	
/E.Y./	3T	2010/0145700	06/10/2010	Kennewick *et al.*	704	257	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED
/E. Yen/	10/27/2010

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

402268606v1

OIPE IAP07 SEP 27 2007 PATENT & TRADEMARK OFFICE

INFORMATION DISCLOSURE CITATION IN AN APPLICATION

Form PTO-1449

Docket Number	Application Number
090611-0357945	11/580,926
Applicant(s)	
Larry BALDWIN *et al.*	
Filing Date	Group Art Unit
October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
EY	1A	5,155,743	10/13/1992	Jacobs	375	28	
	1B	5,386,556	01/31/1995	Hedin et al.	395	600	
	1C	5,424,947	06/13/1995	Nagao et al.	364	419.08	
	1D	5,471,318	11/28/1995	Ahuja et al.	358	400	
	1E	5,475,733	12/12/1995	Eisdorfer et al.	379	52	
	1F	5,499,289	03/12/1996	Bruno et al.	379	220	
	1G	5,517,560	05/14/1996	Greenspan	379	114	
	1H	5,533,108	07/02/1996	Harris et al.	379	201	
	1I	5,537,436	07/16/1996	Bottoms et al.	375	222	
	1J	5,539,744	07/23/1996	Chu et al.	370	60	
	1K	5,557,667	09/17/1996	Bruno et al.	379	201	
	1L	5,563,937	10/08/1996	Bruno et al.	379	201	
	1M	5,590,039	12/31/1996	Ikeda et al.	395	759	
	1N	5,633,922	05/27/1997	August et al.	379	220	
	1O	5,675,629	10/07/1997	Raffel et al.	379	58	
	1P	5,708,422	01/13/1998	Blonder et al.	340	825.34	
	1Q	5,721,938	02/24/1998	Stuckey	395	754	
	1R	5,722,084	02/24/1998	Chakrin et al.	455	551	
	1S	5,742,763	04/21/1998	Jones	395	200.3	
	1T	5,752,052	05/12/1998	Richardson et al.	395	759	
	1U	5,754,784	05/19/1998	Garland et al.	395	200.49	
	1V	5,761,631	06/02/1998	Nasukawa	704	9	
EY	1W	5,794,050	08/11/1998	Dahlgren et al.	395	708	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO
EY	1X	WO 01/78065	10/18/2001	International				

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EY	1Y	REUTERS, "IBM to Enable Honda Drivers to Talk to Cars", Charles Schwab & Co., Inc., July 28, 2002, 1 page.
EY	1Z	LIN, Bor-shen, *et al.*, "A Distributed Architecture for Cooperative Spoken Dialogue Agents with Coherent Dialogue State and History", ASRU'99, 1999, 4 pages.
EY	1AA	KUHN, Thomas, *et al.*, "Hybrid In-Car Speech Recognition for Mobile Multimedia Applications", Vehicular Technology Conference, IEEE, July 1999, pages 2009-2013.

EXAMINER	DATE CONSIDERED
/Eric Yen/	10/26/10

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
EY	2A	5,799,276	08/25/1998	Komissarchik *et al.*	704	251	
	2B	5,802,510	09/01/1998	Jones	707	2	
	2C	5,832,221	11/03/1998	Jones	375	200.36	
	2D	5,878,386	03/02/1999	Coughlin	704	10	
	2E	5,895,466	04/20/1999	Goldberg et al.	707	5	
	2F	5,911,120	06/08/1999	Jarett et al.	455	417	
	2G	5,918,222	06/29/1999	Fukui *et al.*	707	1	
	2H	5,926,784	07/20/1999	Richardson et al.	704	9	
	2I	5,933,822	08/03/1999	Braden-Harder et al.	707	5	
	2J	5,953,393	09/14/1999	Culbreth et al.	379	88.25	
	2K	5,963,894	10/05/1999	Richardson et al.	704	9	
	2L	5,963,940	10/05/1999	Liddy et al.	707	5	
	2M	5,987,404	11/16/1999	Della Pietra et al.	704	9	
	2N	5,991,721	11/23/1999	Asano et al.	704	257	
	2O	5,995,119	11/30/1999	Cosatto et al.	345	473	
	2P	6,009,382	12/28/1999	Martino et al.	704	1	
	2Q	6,014,559	01/11/2000	Amin	455	413	
	2R	6,021,384	02/01/2000	Gorin et al.	704	1	
	2S	6,044,347	03/28/2000	Abella et al.	704	272	
	2T	6,049,602	04/11/2000	Foladare et al.	379	265	
	2U	6,058,187	05/02/2000	Chen	380	21	
	2V	6,078,886	06/20/2000	Dragosh et al.	704	270	
EY	2W	6,081,774	06/27/2000	De Hita et al.	704	9	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EY	2X	BELVIN, Robert, *et al.*, "Development of the HRL Route Navigation Dialogue System", Proceedings of the First International Conference on Human Language Technology Research, San Diego, 2001, pages 1-5.
EY	2Y	LIND, R., *et al.*, "The Network Vehicle—A Glimpse into the Future of Mobile Multi-Media", IEEE Aerosp. Electron. Systems Magazine, Volume 14, No. 9, September 1999, pages 27-32.

EXAMINER	/Eric Yen/	DATE CONSIDERED	10/26/10

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
EY	3A	6,101,241	08/08/2000	Boyce et al.	379	88.01	
	3B	6,119,087	09/12/2000	Kuhn et al.	704	270	
	3C	6,134,235	10/17/2000	Goldman et al.	370	352	
	3D	6,144,667	11/07/2000	Doshi et al.	370	401	
	3E	6,160,883	12/12/2000	Jackson et al.	379	230	
	3F	6,173,279	01/09/2001	Levin et al.	707	5	
	3G	6,175,858	01/16/2001	Bulfer et al.	709	206	
	3H	6,185,535	02/06/2001	Hedin *et al.*	704	270	
	3I	6,192,110	02/20/2001	Abella et al.	379	88.01	
	3J	6,192,338	02/20/2001	Haszto et al.	704	257	
	3K	6,195,651	02/27/2001	Handel *et al.*	707	2	
	3L	6,208,972	03/27/2001	Grant *et al.*	704	275	
	3M	6,219,346	04/17/2001	Maxemchuk	370	338	
	3N	6,219,643	04/17/2001	Cohen et al.	704	257	
	3O	6,233,556	05/15/2001	Teunen et al.	704	250	
	3P	6,233,559	05/15/2001	Balakrishnan	704	275	
	3Q	6,246,981	06/12/2001	Papineni et al.	704	235	
	3R	6,272,455	08/07/2001	Hoshen et al.	704	1	
	3S	6,292,767	09/18/2001	Jackson et al.	704	1	
	3T	6,314,402	11/06/2001	Monaco et al.	704	275	
	3U	6,366,886	04/02/2002	Dragosh et al.	704	270.1	
	3V	6,385,646	05/07/2002	Brown et al.	709	217	
EY	3W	6,393,428	05/21/2002	Miller et al.	707	102	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EY	3X	ZHAO, Yilin, "Telematics: Safe and Fun Driving", IEEE Intelligent Systems, Volume 17, Issue 1, 2002, pages 10-14.

EXAMINER	DATE CONSIDERED
/Eric Yen/	10/26/10

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
EY	4A	6,404,878	06/11/2002	Jackson et al.	379	221.01	
	4B	6,408,272	06/18/2002	White *et al.*	704	270.1	
	4C	6,411,810	06/25/2002	Maxemchuk	455	453	
	4D	6,418,210	07/09/2002	Sayko	379	142.15	
	4E	6,420,975	07/16/2002	DeLine *et al.*	340	815.4	
	4F	6,430,285	08/06/2002	Bauer et al.	379	265.01	
	4G	6,434,523	08/13/2002	Monaco	704	257	
	4H	6,434,524	08/13/2002	Weber	704	257	
	4I	6,442,522	08/27/2002	Carberry et al.	704	257	
	4J	6,446,114	09/03/2002	Bulfer et al.	709	206	
	4K	6,453,153	09/17/2002	Bowker et al.	455	67.4	
	4L	6,453,292	09/17/2002	Ramaswamy et al.	704	235	
	4M	6,456,711	09/24/2002	Cheung et al.	379	265.09	
	4N	6,498,797	12/24/2002	Anerousis et al.	370	522	
	4O	6,499,013	12/24/2002	Weber	704	257	
	4P	6,501,833	12/31/2002	Phillips *et al.*	379	88.07	
	4Q	6,501,834	12/31/2002	Milewski et al.	379	93.24	
	4R	6,523,061	02/18/2003	Halverson et al.	709	202	
	4S	6,532,444	03/11/2003	Weber	704	257	
	4T	6,539,348	03/25/2003	Bond et al.	704	9	
	4U	6,553,372	04/22/2003	Brassell et al.	707	5	
	4V	6,556,973	04/29/2003	Lewin	704	277	
EY	4W	6,560,576	05/06/2003	Cohen et al.	704	270	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED
/Eric Yen/	10/26/10

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
EY	5A	6,567,778	05/20/2003	Chao Chang et al.	704	257	
	5B	6,570,964	05/27/2003	Murveit et al.	379	67.1	
	5C	6,574,624	06/03/2003	Johnson et al.	707	5	
	5D	6,594,257	07/15/2003	Doshi et al.	370	352	
	5E	6,604,077	08/05/2003	Dragosh et al.	704	270.1	
	5F	6,611,692	08/26/2003	Raffel et al.	455	552	
	5G	6,614,773	09/02/2003	Maxemchuk	370	337	
	5H	6,615,172	09/02/2003	Bennett *et al.*	704	257	
	5I	6,629,066	09/30/2003	Jackson et al.	704	9	
	5J	6,650,747	11/18/2003	Bala et al.	379	265.06	
	5K	6,691,151	02/10/2004	Cheyer et al.	709	202	
	5L	6,741,931	05/25/2004	Kohut *et al.*	701	209	
	5M	6,937,977	08/30/2005	Gerson	704	201	
	5N	6,980,092	12/27/2005	Turnbull *et al.*	340	425.5	
	5O	6,996,531	02/07/2006	Korall *et al.*	704	270	
	5P	7,020,609	03/28/2003	Thrift *et al.*	704	270.1	
	5Q	7,027,975	04/11/2006	Pazandak *et al.*	704	9	
	5R	7,062,488	06/13/2006	Reisman	707	8	
	5S	7,092,928	08/15/2006	Elad *et al.*	706	60	
	5T	7,137,126	11/14/06	Coffman *et al.*	719	328	
	5U	2002/0143535	10/03/2002	Kist *et al.*	704	251	
	5V	2003/0182132	09/25/2003	Niemoeller	704	275	
	5W	2004/0044516	03/04/2004	Kennewick *et al.*	704	5	
	5X	2004/0193420	09/30/2004	Kennewick *et al.*	704	257	
EY	5Y	2005/0021826	01/27/2005	Kumar	709	232	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED
/Eric Yen/	10/26/10

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

PTO/SB/30 (07-09)
Approved for use through 07/31/2012. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995. no persons are required to respond to a collection of information unless it contains a valid OMB control number.

Request for Continued Examination (RCE) Transmittal

Address to:
Mail Stop RCE
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Application Number	11/580,926
Filing Date	October 16, 2006
First Named Inventor	LARRY BALDWIN
Art Unit	2626
Examiner Name	Eric L. Yen
Attorney Docket Number	090611-0357945

This is a Request for Continued Examination (RCE) under 37 CFR 1 .114 of the above-identified application.
Request for Continued Examination (RCE) practice under 37 CFR 1.114 does not apply to any utility or plant application filed prior to June 8, 1995, or to any design application. See Instruction Sheet for RCEs (not to be submitted to the USPTO) on page 2.

1. Submission required under 37 CFR 1.114 Note: If the RCE is proper, any previously filed unentered amendments and amendments enclosed with the RCE will be entered in the order in which they were filed unless applicant instructs otherwise. If applicant does not wish to have any previously filed unentered amendment(s) entered, applicant must request non-entry of such amendment(s).

   a. ☐ Previously submitted. If a final Office action is outstanding, any amendments filed after the final Office action may be considered as a submission even if this box is not checked.

      i. ☐ Consider the arguments in the Appeal Brief or Reply Brief previously filed on _____

      ii. ☐ Other _____

   b. ☒ Enclosed

      i. ☒ Amendment/Reply

      ii. ☐ Affidavit(s)/ Declaration(s)

      iii. ☒ Information Disclosure Statement (IDS)

      iv. ☐ Other _____

2. Miscellaneous

   Suspension of action on the above-identified application is requested under 37 CFR 1.103(c) for a

   a. ☐ period of _____ months. (Period of suspension shall not exceed 3 months; Fee under 37 CFR 1.17(i) required)

   b. ☐ Other _____

3. Fees The RCE fee under 37 CFR 1.17(e) is required by 37 CFR 1.114 when the RCE is filed.
   The Director is hereby authorized to charge the following fees any underpayment of fees or credit any overpayments to

   a. ☒ Deposit Account No. 033975 .

      i. ☒ RCE fee required under 37 CFR 1.17(e)

      ii ☐ Extension of time fee (37 CFR 1.136 and 1.17)

      iii ☐ Other _____

   b. ☐ Check in the amount of $ _____ enclosed

   c. ☐ Payment by credit card (Form PTO-2038 enclosed)

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

SIGNATURE OF APPLICANT, ATTORNEY, OR AGENT REQUIRED

Signature		Date	January 21, 2011
Name (Print/Type)	Syed Jafar Ali	Registration No.	58780

CERTIFICATE OF MAILING OR TRANSMISSION

I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as first class mail in an envelope addressed to: Mail Stop RCE, Commissioner for Patents, P. O. Box 1450, Alexandria, VA 22313-1450 or facsimile transmitted to the U.S. Patent and Trademark Office on the date shown below.

Signature			
Name (Print/Type)		Date	

This collection of information is required by 37 CFR 1.114. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Mail Stop RCE, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

American LegalNet, Inc.
www.FormsWorkFlow.com

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANTS :	Larry BALDWIN et al.	CONFIRMATION NO. :	6670
SERIAL NUMBER :	11/580,926	EXAMINER :	Eric L. Yen
FILING DATE :	October 16, 2006	ART UNIT :	2626
FOR :	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE		

REPLY AND AMENDMENT ACCOMPANYING REQUEST FOR CONTINUED EXAMINATION

Mail Stop RCE
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

In response to the Final Office Action mailed **November 1, 2010**, please amend the above-identified application as follows:

Amendments to the Specification begin on page **2** of this paper.

Amendments to the Claims begin on page **6** of this paper.

Remarks/Arguments begin on page **21** of this paper.

It is believed that no extensions of time or fees for net addition of claims are required beyond those that may otherwise be provided for in documents accompanying this paper. However, if additional extensions of time are necessary to prevent abandonment of this application, then such extensions of time are hereby petitioned for under 37 C.F.R. § 1.136(a), and any fees required therefore (including fees for net addition of claims) are hereby authorized to be charged to our Deposit Account No. 033975 (**Ref. No. 090611-0357945**).

AMENDMENT

Amendments to the Specification: (identifying insertions and ~~deletions~~).

- Please amend paragraph **[006]**, which begins on page 2, as follows:

[006] According to an aspect of the invention, an exemplary system architecture for implementing a cooperative conversational voice user interface is provided. The system may receive an input, which may include a human utterance received by an input device, where the utterance may include one or more requests. As used herein, an "utterance" may be words, syllables, phonemes, or any other audible sound made by a human being. As used herein, a "request" may be a command, directive, or other instruction for a device, computer, or other machine to retrieve information, perform a task, or take some other action. In one implementation, the input may be a multi-modal input, where at least part of the multi-modal input is an utterance. The utterance component of the input may be processed by a speech recognition engine (which may alternatively be referred to as an Automatic Speech Recognizer or ASR) to generate one or more preliminary interpretations of the utterance. The one or more preliminary interpretations may then be provided to a conversational speech engine for further processing, where the conversational speech engine may communicate with one or more databases to generate an adaptive conversational response, which may be returned to the user as an output. In one implementation, the output may be a multi-modal output. For example, the utterance may include a request to perform an action, and the output may include a conversational response reporting success or failure, as well as an execution of the action.

- Please amend paragraph **[009]**, which begins on page 5, as follows:

[009] According to another aspect of the invention, the conversational speech engine may include a context determination process that determines one or more contexts for a request to establish meaning within a conversation. The one or more contexts may be determined by having one or more context domain agents compete to determine a most appropriate domain for a given utterance. Once a given domain "wins" the competition, the winning domain may be responsible for establishing or inferring further contexts and updating short-term and long-term shared knowledge. If there is a deadlock between context domain agents, an adaptive

conversational response may prompt the user to assist in disambiguating between the deadlocked agents. Moreover, the context determination process may infer intended operations and/or context based on previous utterances and/or requests, whereas existing systems consider each utterance independently, potentially making the same errors over and over again. For example, if a given interpretation turns out to be incorrect, the incorrect interpretation may be removed as a potential interpretation from one or more ~~Automatic Speech Recognizer~~ grammars associated with the speech recognition engine and/or from possible interpretations determined by the conversational speech engine, thereby assuring that a mistake will not be repeated for an identical utterance.

- Please amend paragraph **[025]**, which begins on page 12, as follows:

[025] The utterance component of input 105 may be processed by a speech recognition engine 110 (which may alternatively be referred to herein as [[an]] Automatic Speech Recognizer 110, or as shown in Fig. 1, ASR 110) to generate one or more preliminary interpretations of the utterance. The speech recognition engine ~~Automatic Speech Recognizer~~ 110 may process the utterance using any suitable technique known in the art. For example, in one implementation, the speech recognition engine ~~Automatic Speech Recognizer~~ 110 may interpret the utterance using techniques of phonetic dictation to recognize a phoneme stream, as described in U.S. Patent Application Serial No. 11/513,269, entitled "Dynamic Speech Sharpening," filed August 31, 2006, which issued as U.S. Patent No. 7,634,409 on December 15, 2009, and which is hereby incorporated by reference in its entirety. The one or more preliminary interpretations generated by the speech recognition engine ~~Automatic Speech Recognizer~~ 110 may then be provided to a conversational speech engine 115 for further processing. Conversational speech engine 115 may include a conversational language processor 120 and/or a voice search engine 125, described in greater detail in Fig. 2 below. Conversational speech engine 115 may communicate with one or more databases 130 to generate an adaptive conversational response, which may be returned to the user as an output 140. In one implementation, output 140 may be a multi-modal output and/or an interaction with one or more applications 145 to complete the request. For example, output 140 may

include a combination of an audible response and a display of a route on a navigation device. For example, the utterance may include a request to perform an action, and output 140 may include a conversational response reporting success or failure, as well as an execution of the action. In addition, in various implementations, the speech recognition engine ~~Automatic Speech Recognizer~~ 110, conversational speech engine 115, and/or databases 130 may reside locally (e.g., on a user device), remotely (e.g., on a server), or a hybrid model of local and remote processing may be used (e.g., lightweight applications may be processed locally while computationally intensive applications may be processed remotely).

- Please amend paragraph **[035]**, which begins on page 19, as follows:

[035] Moreover, context determination process 255 may infer intended operations and/or context based on previous utterances and/or requests, whereas existing systems consider each utterance independently, potentially making the same errors over and over again. For example, if a given interpretation turns out to be incorrect, the incorrect interpretation may be removed as a potential interpretation from one or more ~~Automatic Speech Recognizer~~ grammars associated with the speech recognition engine and/or from possible subsequent interpretations determined by context determination process 255, thereby assuring that a mistake will not be repeated for an identical utterance.

- Please amend paragraph **[051]**, which begins on page 30, as follows:

[051] According to another aspect of the invention, knowledge-enabled utterances, as generated by Intelligent Hypothesis Builder 310, may include one or more explicit (supplied by a user), and one or more implicit (supplied by Intelligent Hypothesis Builder 310) contextual signifiers, qualifiers, criteria, and other information that can be used to identify and evaluate relevant tasks. At that point, Intelligent Hypothesis Builder 310 may provide an input to Adaptive Response Builder 315. The input received by Adaptive Response Builder 315 may include at least a ranked list of hypotheses, including explicit and/or implicit hypotheses, each of which may have a corresponding degree of certainty. A hypothesis may be assigned one of four degrees of certainty: (1) "sure," where contextual signifiers and qualifiers relate to one task, context and qualifiers relate to one task, and [[an ASR]] a confidence level associated with

a preliminary interpretation generated at the speech recognition engine exceeds a predetermined threshold; (2) "pretty sure," where contextual signifiers and qualifiers relate to more than one task (select top-ranked task) and criteria relates to one request, and/or the [[ASR]] confidence level associated with the preliminary interpretation generated at the speech recognition engine is below the predetermined threshold; (3) "not sure," where additional contextual signifiers or qualifiers are needed to indicate or rank a task; and (4) "no hypothesis," where little or no information can be deciphered. Each degree of certainty may further be classified as explicit or implicit, which may be used to adjust a response. The input received by Adaptive Response Builder 310 may also include a context, user syntax and/or grammar, context domain agent specific information and/or preferences (e.g., a travel context domain agent may know a user frequently requests information about France, which may be shared with a movie context domain agent so that responses may occasionally include French movies).

Amendments to the Claims: Please replace all prior versions and listings of claims with the following listing of claims.

LISTING OF CLAIMS:

1. **(Currently Amended)** A method for providing a cooperative conversational voice user interface, comprising:

receiving an utterance at a voice input device~~, wherein the voice input device receives the utterance in~~ during a current conversation with a user, wherein the utterance includes one or more words that have different meanings in different contexts;

accumulating short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received ~~at the voice input device in~~ during the current conversation;

accumulating long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

determining an intended meaning for the utterance, wherein ~~a conversational speech engine determines~~ determining the intended meaning for the utterance includes:

identifying, at a conversational speech engine, a context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge; and

establishing the intended meaning within the identified context, wherein the conversational speech engine establishes the intended meaning within the identified context to disambiguate an intent that the user had in speaking the one or more words that have the different meanings in the different contexts; and

generating a response to the utterance, wherein the conversational speech engine grammatically or syntactically adapts ~~generates~~ the response ~~to the utterance~~ based on the intended meaning established within ~~determined from the short-term shared knowledge and~~ the ~~long-term shared knowledge~~ identified context.

2. (**Currently Amended**) The method of claim 1, wherein accumulating the short-term shared knowledge about the current conversation includes populating a short-term context stack with information about the utterance received [[in]] during the current conversation.

3. (**Currently Amended**) The method of claim 2, wherein accumulating the short-term shared knowledge about the current conversation further includes expiring the information about the utterance from the short-term context stack after a psychologically appropriate ~~predetermined~~ amount of time.

4. (**Currently Amended**) The method of claim [[1]] 3, wherein accumulating the long-term shared knowledge about the user includes updating one or more long-term profiles associated with the user to include information about the utterance received [[in]] during the current conversation and relevant data associated with the information expired from the short term context stack.

5. (**Currently Amended**) The method of claim 1, wherein determining the intended meaning for the utterance further includes:

identifying ~~a conversation type for the current conversation based on one or more of~~ a conversational goal [[for]] associated with the utterance, roles [[for]] associated with the user and one or more other participants in the current conversation, [[or]] and an information allocation ~~of information~~ among the user and the one or more other participants in the current conversation; and

classifying one or more of the utterance or the current conversation into a conversation type based on one or more of the identified conversational goal, the identified roles, or the identified information allocation, wherein the conversational speech engine further establishes the intended meaning based on the conversation type.

6. (**Currently Amended**) The method of claim [[1]] 5, wherein the established intended meaning comprises ~~further comprising generating~~ a hypothesis having a degree of certainty about ~~an intent that the user had in speaking the utterance, wherein conversational speech engine further determines the intended meaning for the utterance from the hypothesis about~~ the intent that the user had in speaking the one or more words in the utterance.

7. (**Currently Amended**) The method of claim 6, further comprising generating a preliminary interpretation of the utterance at a speech recognition engine coupled to the voice input device and the conversational speech engine, wherein the conversational speech engine assigns the degree of certainty to the hypothesis based on [[a]] one or more of the conversation type, information associated with the identified context, ~~that the conversational speech engine identifies for the current conversation from one~~ or ~~more contextual signifiers or qualifiers recognized in~~ a confidence level associated with the ~~utterance~~ preliminary interpretation generated at the speech recognition engine.

8. (**Currently Amended**) A [[The]] method ~~of claim 1~~ for providing a cooperative conversational voice user interface, comprising:

receiving an utterance at a voice input device during a current conversation with a user;

accumulating short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received during the current conversation;

accumulating long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

determining an intended meaning for the utterance, wherein determining the intended meaning for the utterance includes:

identifying, at a conversational speech engine, a context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge;

inferring additional information about ~~one or more contextual signifiers or qualifiers missing from~~ the utterance from the short-term shared knowledge and the

long-term shared knowledge in response to determining that the utterance contains insufficient information to complete a request in ~~generate~~ the ~~response~~ identified context; and

establishing the intended meaning within the identified ~~identifying a~~ context [[for]] based on the ~~current conversation from the one or more~~ additional information inferred ~~contextual signifiers or qualifiers~~ about the utterance; and

generating a response to the utterance based on the intended meaning established within the identified context.

9. (**Currently Amended**) The method of claim 8, wherein the established ~~conversational speech engine assigns an implicit degree of certainty to the~~ intended meaning ~~determined for the utterance in response to determining~~ comprises an implicit hypothesis having a corresponding degree of certainty about an intent that the ~~utterance contains insufficient information to generate~~ user had in speaking the ~~response~~ utterance.

10. (**Currently Amended**) The method of claim [[1]] 5, wherein the conversational speech engine further grammatically or syntactically adapts the response [[to]] based on the ~~utterance includes one or more words having contextual signifiers, qualifiers, or grammar syntax adapted to reflect contextual signifiers, qualifiers, or grammar syntax recognized in or inferred from the utterance~~ conversation type.

11. (**Currently Amended**) The method of claim 1, wherein the conversational speech engine grammatically [[one]] or syntactically adapts ~~more words having the adapted contextual signifiers, qualifiers, or grammar syntax frame~~ the response to influence a subsequent reply utterance [[in]] that the conversational speech engine expects from the user during the current conversation.

12. (**Currently Amended**) The method of claim 1, ~~wherein determining the intended meaning for the utterance includes~~ further comprising:

generating multiple ~~possible~~ preliminary interpretations [[for]] of the utterance at a speech recognition engine coupled to the voice input device and the conversational speech engine, wherein an initial interpretation [[for]] of the utterance comprises one of the multiple ~~possible~~ preliminary interpretations having a highest ~~degree of certainty~~ confidence level; and

updating the short-term shared knowledge about the current conversation to remove the initial interpretation from the multiple ~~possible interpretation~~ preliminary interpretations in response to determining that the initial interpretation was incorrect, wherein the conversational speech engine determines the intended meaning based on ~~determined from the short-term shared knowledge and the long-term shared knowledge comprises~~ one of the multiple ~~possible~~ preliminary interpretations having a next highest ~~degree of certainty~~ confidence level.

13. (**Previously Presented**) The method of claim 1, wherein the user speaks the utterance in a multi-modal input that further includes one or more non-voice inputs relating to the utterance.

14. (**Currently Amended**) The method of claim 1, wherein the conversational speech engine generates the response in a multi-modal output that includes one or more non-voice outputs ~~presented to the user in response~~ that relate to the utterance or one or more tasks executed to process a request identified from the intended meaning ~~for the utterance~~.

15. (**Currently Amended**) A non-transitory computer readable medium containing computer-executable instructions for providing a cooperative conversational voice user interface, the computer-executable instructions operable when executed to:

receive an utterance at a voice input device~~, wherein the voice input device receives the utterance in~~ during a current conversation with a user, wherein the utterance includes one or more words that have different meanings in different contexts;

accumulate short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received ~~at the voice input device in~~ during the current conversation;

accumulate long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

identify a context associated with the utterance, wherein a conversational speech engine identifies the context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge;

establish ~~determine~~ an intended meaning for the utterance within the identified context, wherein [[a]] the conversational speech engine ~~determines~~ establishes the intended meaning within [[for]] the identified context to disambiguate an intent that ~~utterance from~~ the user had in speaking ~~short-term shared knowledge and~~ the ~~long-term shared knowledge~~ one or more words that have the different meanings in the different contexts; and

generate a response to the utterance, wherein the conversational speech engine grammatically or syntactically adapts ~~generates~~ the response ~~to the utterance~~ based on the intended meaning established within ~~determined from the short-term shared knowledge and~~ the ~~long-term shared knowledge~~ identified context.

16. (**Currently Amended**) The non-transitory computer readable medium of claim 15, wherein to accumulate the short-term shared knowledge about the current conversation, the computer-executable instructions are further operable when executed to populate a short-term context stack with information about the utterance received [[in]] during the current conversation.

17. (**Currently Amended**) The non-transitory computer readable medium of claim 16, wherein to accumulate the short-term shared knowledge about the current conversation, the computer-executable instructions are further operable when executed to expire the information about the utterance from the short-term context stack after a psychologically appropriate ~~predetermined~~ amount of time.

18. (**Currently Amended**) The non-transitory computer readable medium of claim [[15]] 17, wherein to accumulate the long-term shared knowledge about the user, the computer-executable instructions are further operable when executed to update one or more long-term profiles associated with the user to include information about the utterance received [[in]] during the current conversation and relevant data associated with the information expired from the short term context stack.

19. (**Currently Amended**) The non-transitory computer readable medium of claim 15, wherein ~~to determine the intended meaning for the utterance,~~ the computer-executable instructions are further operable when executed to:

identify ~~a conversation type for the current conversation based on one or more of~~ a conversational goal [[for]] associated with the utterance, roles [[for]] associated with the user and one or more other participants in the current conversation, [[or]] and an information allocation ~~of information~~ among the user and the one or more other participants in the current conversation; and

classify one or more of the utterance or the current conversation into a conversation type based on one or more of the identified conversational goal, the identified roles, or the identified information allocation, wherein the conversational speech engine further establishes the intended meaning based on the conversation type.

20. (**Currently Amended**) The non-transitory computer readable medium of claim [[15]] 19, wherein the established intended meaning comprises ~~computer-executable instructions are further operable when executed to generate~~ a hypothesis having a degree of certainty about ~~an intent that the user had in speaking the utterance, wherein conversational speech engine further determines the intended meaning for the utterance from the hypothesis about~~ the intent that the user had in speaking the one or more words in the utterance.

21. (**Currently Amended**) The computer-readable medium of claim 20, wherein the computer-executable instructions are further operable when executed to generate a preliminary interpretation of the utterance at a speech recognition engine, wherein the conversational speech engine assigns the degree of certainty to the hypothesis based on [[a]] one or more of the conversation type, information associated with the identified context, ~~that the conversational speech engine identifies for the current conversation from one~~ or ~~more contextual signifiers or qualifiers recognized in~~ a confidence level associated with the ~~utterance~~ preliminary interpretation generated at the speech recognition engine.

22. (**Currently Amended**) A [[The]] non-transitory computer readable medium ~~of claim 15~~ containing computer-executable instructions for providing a cooperative conversational voice user interface, ~~wherein to determine the intended meaning for the utterance,~~ the computer-executable instructions ~~are further~~ operable when executed to:

receive an utterance at a voice input device during a current conversation with a user;

accumulate short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received during the current conversation;

accumulate long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

identify a context associated with the utterance, wherein a conversational speech engine identifies the context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge;

infer additional information about ~~one or more contextual signifiers or qualifiers missing from~~ the utterance from the short-term shared knowledge and the long-term shared knowledge in response to determining that the utterance contains insufficient information to complete a request in ~~generate~~ the ~~response~~ identified context; ~~and~~

establish an intended meaning for the utterance within the identified ~~identify a~~ context [[for]] based on the ~~current conversation from the one or more~~ additional information inferred ~~contextual signifiers or qualifiers~~ about the utterance; and

generate a response to the utterance based on the intended meaning established within the identified context.

23. (**Currently Amended**) The non-transitory computer readable medium of claim 22, wherein the established ~~conversational speech engine assigns an implicit degree of certainty to the~~ intended meaning ~~determined for the utterance in response to determining~~ comprises an implicit hypothesis having a corresponding degree of certainty about an intent that the ~~utterance contains insufficient information to generate~~ user had in speaking the ~~response~~ utterance.

24. (**Currently Amended**) The non-transitory computer readable medium of claim [[15]] 19, wherein the conversational speech engine further grammatically or syntactically adapts the response [[to]] based on the ~~utterance includes one or more words having contextual signifiers, qualifiers, or grammar syntax adapted to reflect contextual signifiers, qualifiers, or grammar syntax recognized in or inferred from the utterance~~ conversation type.

25. (**Currently Amended**) The non-transitory computer readable medium of claim 15, wherein the conversational speech engine grammatically [[one]] or syntactically adapts ~~more words having the adapted contextual signifiers, qualifiers, or grammar syntax frame~~ the response to influence a subsequent reply utterance [[in]] that the conversational speech engine expects from the user during the current conversation.

26. (**Currently Amended**) The non-transitory computer readable medium of claim 15, wherein ~~to determine the intended meaning for the utterance,~~ the computer-executable instructions are further operable when executed to:

generate multiple ~~possible~~ preliminary interpretations [[for]] of the utterance at a speech recognition engine, wherein an initial interpretation [[for]] of the utterance comprises one of the multiple ~~possible~~ preliminary interpretations having a highest ~~degree of certainty~~ confidence level; and

update the short-term shared knowledge about the current conversation to remove the initial interpretation from the multiple ~~possible interpretation~~ preliminary interpretations in response to determining that the initial interpretation was incorrect, wherein the conversational speech engine identifies the context associated with the utterance and establishes the intended meaning for ~~determined from~~ the utterance based on ~~short-term shared knowledge and the long-term shared knowledge comprises~~ one of the multiple ~~possible~~ preliminary interpretations having a next highest ~~degree of certainty~~ confidence level.

27. (**Previously Presented**) The non-transitory computer readable medium of claim 15, wherein the user speaks the utterance in a multi-modal input that further includes one or more non-voice inputs relating to the utterance.

28. (**Currently Amended**) The non-transitory computer readable medium of claim 15, wherein the conversational speech engine generates the response in a multi-modal output that includes one or more non-voice outputs ~~presented to the user in response~~ that relate to the utterance or one or more tasks executed to process a request identified from the intended meaning ~~for the utterance~~.

29. (**Currently Amended**) A system for providing a cooperative conversational voice user interface, comprising:

a voice input device configured to receive an utterance [[in]] during a current conversation with a user, wherein the utterance includes one or more words that have different meanings in different contexts; and

a conversational speech engine, wherein the conversational speech engine includes one or more processors configured to:

accumulate short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received ~~at the voice input device in~~ during the current conversation;

accumulate long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

identify a context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge;

establish ~~determine~~ an intended meaning for the utterance~~, wherein~~ within the ~~conversational speech engine determines~~ identified context to disambiguate an intent that the ~~intended meaning for~~ user had in speaking the ~~utterance from~~ one or more words that have the ~~short-term shared knowledge and~~ different meanings in the ~~long-term shared knowledge~~ different contexts; and

generate a grammatically or syntactically adapted response ~~to the utterance, wherein the conversational speech engine generates the response~~ to the utterance based on the intended meaning established within ~~determined from the short-term shared knowledge and~~ the ~~long-term shared knowledge~~ identified context.

30. (**Currently Amended**) The system of claim 29, wherein to accumulate the short-term shared knowledge about the current conversation, the one or more processors are further configured to populate a short-term context stack with information about the utterance received [[in]] during the current conversation.

31. **(Currently Amended)** The system of claim 30, wherein to accumulate the short-term shared knowledge about the current conversation, the one or more processors are further configured to expire the information about the utterance from the short-term context stack after a psychologically appropriate ~~predetermined~~ amount of time.

32. **(Currently Amended)** The system of claim [[29]] 31, wherein to accumulate the long-term shared knowledge about the user, the one or more processors are further configured to update one or more long-term profiles associated with the user to include information about the utterance received [[in]] during the current conversation and relevant data associated with the information expired from the short term context stack.

33. **(Currently Amended)** The system of claim 29, wherein ~~to determine the intended meaning for the utterance,~~ the one or more processors are further configured to:

identify ~~a conversation type for the current conversation based on one or more of~~ a conversational goal [[for]] associated with the utterance, roles [[for]] associated with the user and one or more other participants in the current conversation, [[or]] and an information allocation ~~of information~~ among the user and the one or more other participants in the current conversation; and

classify one or more of the utterance or the current conversation into a conversation type based on one or more of the identified conversational goal, the identified roles, or the identified information allocation, wherein the one or more processors are further configured to establish the intended meaning based on the conversation type.

34. **(Currently Amended)** The system of claim [[29]] 33, wherein the established intended meaning comprises ~~one or more processors are further configured to generate~~ a hypothesis having a degree of certainty about ~~an intent that the user had in speaking the utterance, wherein conversational speech engine further determines the intended meaning for the utterance from the hypothesis about~~ the intent that the user had in speaking the one or more words in the utterance.

35. (**Currently Amended**) The system of claim 34, further comprising a speech recognition engine configured to generate a preliminary interpretation of the utterance, wherein the one or more processors are further configured to assign ~~conversational speech engine assigns~~ the degree of certainty to the hypothesis based on [[a]] one or more of the conversation type, information associated with the identified context, ~~that the conversational speech engine identifies for the current conversation from one~~ or ~~more contextual signifiers or qualifiers recognized in~~ a confidence level associated with the ~~utterance~~ preliminary interpretation generated at the speech recognition engine.

36. (**Currently Amended**) A [[The]] system ~~of claim 29~~ for providing a cooperative conversational voice user interface, comprising:

a voice input device configured to receive an utterance during a current conversation with a user; and

a conversational speech engine, wherein ~~to determine~~ the ~~intended meaning for the utterance, the~~ conversational speech engine includes one or more processors ~~are further~~ configured to:

accumulate short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received during the current conversation;

accumulate long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

identify a context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge;

infer additional information about ~~one or more contextual signifiers or qualifiers missing from~~ the utterance from the short-term shared knowledge and the long-term shared knowledge in response to determining that the utterance contains insufficient information to complete a request in ~~generate~~ the ~~response~~ identified context; ~~and~~

establish an intended meaning for the utterance within the identified ~~identify a~~ context [[for]] based on the ~~current conversation from the one or more~~ additional information inferred ~~contextual signifiers or qualifiers~~ about the utterance; and

generate a response to the utterance based on the intended meaning established within the identified context.

37. (**Currently Amended**) The system of claim 36, wherein the established ~~conversational speech engine assigns an implicit degree of certainty to the~~ intended meaning ~~determined for the utterance in response to determining~~ comprises an implicit hypothesis having a corresponding degree of certainty about an intent that the ~~utterance contains insufficient information to generate~~ user had in speaking the ~~response~~ utterance.

38. (**Currently Amended**) The system of claim [[29]] 33, wherein the one or more processors are further configured ~~response~~ to generate the grammatically ~~utterance includes one or more words having contextual signifiers, qualifiers, or grammar syntax adapted to reflect contextual signifiers, qualifiers,~~ or syntactically adapted response ~~grammar syntax recognized in or inferred from~~ based on the ~~utterance~~ conversation type.

39. (**Currently Amended**) The system of claim 29, wherein the one or more processors are further configured to generate ~~words having~~ the grammatically or syntactically adapted ~~contextual signifiers, qualifiers, or grammar syntax frame the~~ response to influence a subsequent reply utterance [[in]] expected from the user during the current conversation.

40. (**Currently Amended**) The system of claim 29, ~~wherein to determine the intended meaning for the utterance, the one or more processors are~~ further ~~configured to~~ comprising a speech recognition engine configured to:

generate multiple ~~possible~~ preliminary interpretations [[for]] of the utterance, wherein an initial interpretation [[for]] of the utterance comprises one of the multiple ~~possible~~ preliminary interpretations having a highest ~~degree of certainty~~ confidence level; and

update the short-term shared knowledge about the current conversation to remove the initial interpretation from the multiple ~~possible interpretation~~ preliminary interpretations in response to determining that the initial interpretation was incorrect, wherein the one or more processors are configured to identify the context associated with the utterance and establish the intended meaning for ~~determined from~~ the utterance based on ~~short-term shared knowledge and the long-term shared knowledge comprises~~ one of the multiple ~~possible~~ preliminary interpretations having a next highest ~~degree of certainty~~ confidence level.

41. (**Previously Presented**) The system of claim 29, wherein the user speaks the utterance in a multi-modal input that further includes one or more non-voice inputs relating to the utterance.

42. (**Currently Amended**) The system of claim 29, wherein the grammatically or syntactically adapted ~~conversational speech engine generates the~~ response comprises [[in]] a multi-modal output that includes one or more non-voice outputs ~~presented to the user in response~~ that relate to the utterance or one or more tasks executed to process a request identified from the intended meaning ~~for the utterance~~.

REMARKS

Upon entry of the foregoing Amendment, claims 1-42 are pending in the application. Claims 1-12, 14-26, 28-40, and 42 have been amended. No claims have been cancelled or newly added. Applicants believe that this Amendment does not add new matter. In view of the foregoing Amendment and the following Remarks, allowance of all the pending claims is requested.

CLAIM OBJECTIONS

The Examiner has objected to claims 6, 11-12, 20, 25-26, 34, and 39-40 because of alleged informalities. Solely to expedite prosecution of this application, and without acknowledging the propriety of the alleged basis for the objection, Applicants note that claims 6, 11-12, 20, 25-26, 34, and 39-40 have been amended as indicated above and submit that the amended claims fully address the objections that the Examiner has raised. Accordingly, Applicants request that the Examiner withdraw the claim objections.

REJECTIONS UNDER 35 U.S.C. § 112, FIRST PARAGRAPH

The Examiner has rejected claims 7-10, 21-24, and 35-38 under 35 U.S.C. § 112, first paragraph. In particular, the Examiner alleges that various features previously recited in claims 8-9, 22-23, and 36-37 fail to comply with the written description requirement, and that various features previously recited in claims 7-8, 10, 21-22, 24, 35-36, and 38 fail to comply with the enablement requirement.

Solely to expedite prosecution of this application, and without acknowledging the propriety of the alleged basis for the rejection, Applicants note that claims 7-10, 21-24, and 35-38 have been amended as indicated above to further clarify the various features recited therein and submit that the amended claims fully address the Examiner's alleged basis for the rejections under 35 U.S.C. § 112, first paragraph.

For example, with respect to the features relating to inferring information associated with an utterance that lacks sufficient information to generate a suitable response, Applicants direct the Examiner's attention to ¶¶ [007], [009], [030], and [049] in Applicants' disclosure.

With respect to the features relating to degrees of certainty associated with an intended meaning determined for an utterance, Applicants direct the Examiner's attention to ¶¶ [010], [014], [039], and [047]-[051] in Applicants' disclosure. With respect to the previously recited features relating to contextual signifiers and qualifiers, Applicants note that various passages in Applicants' disclosure describe how contextual signifiers and qualifiers may be used to model information needed to perform tasks in certain contexts and generate intelligent hypotheses and/or adaptive responses to identify and evaluate relevant tasks, among other things. Nonetheless, solely to expedite prosecution, Applicants note that the claims have been amended to remove the terms "contextual signifiers" and "qualifiers" and to further clarify the various features recited therein.

Accordingly, because the aforementioned passages in Applicants' disclosure support the features recited in amended claims 8-9, 22-23, and 36-37 and describe the features recited in amended claims 8-9, 22-23, and 36-37 in such a way that would enable a person having ordinary skill in the art to make and use the claimed subject matter, claims 7-10, 21-24, and 35-38 satisfy the requirements associated with 35 U.S.C. § 112, first paragraph. Applicants therefore respectfully request that the Examiner withdraw this rejection. Furthermore, Applicants note that the passages in Applicants' disclosure referred to above are to be regarded as exemplary only, as various other passages in Applicants' disclosure support and enable the claimed subject matter.

REJECTION UNDER 35 U.S.C. § 112, SECOND PARAGRAPH

The Examiner has rejected claims 7-10, 21-24, and 35-38 under 35 U.S.C. § 112, second paragraph. In particular, the Examiner alleges that various features previously recited in claims 7-10, 21-24, and 35-38 are indefinite for allegedly failing to particularly point out and distinctly claim the subject matter that Applicants regard as the invention.

Solely to expedite prosecution of this application, and without acknowledging the propriety of the alleged basis for the rejection, Applicants note that claims 7-10, 21-24, and 35-38 have been amended as indicated above to further clarify the various features recited therein and submit that the amended claims fully address the Examiner's alleged basis for the

rejections under 35 U.S.C. § 112, second paragraph. In addition, Applicants direct the Examiner's attention to the various exemplary passages in Applicants' disclosure referred to above in connection with the alleged rejections under 35 U.S.C. § 112, first paragraph, as the features that the Examiner alleges to be indefinite substantially correspond to the features that the Examiner refers to in the alleged written description and enablement rejections.

Accordingly, for at least the reason that the features recited in amended claims 7-10, 21-24, and 35-38 are clear and definite, Applicants respectfully request that the Examiner withdraw this rejection.

REJECTION UNDER 35 U.S.C. § 103

A. CLAIMS 1, 4-7, 10-11, 14-15, 18-21, 24-25, 28-29, 32-35, 38-39, AND 42

The Examiner has rejected claims 1, 4-7, 10-11, 14-15, 18-21, 24-25, 28-29, 32-35, 38-39, and 42 under 35 U.S.C. § 103 as allegedly being unpatentable over U.S. Patent Application Pub. No. 2005/0015256 to Kargman ("Kargman"). This rejection is improper and must be withdrawn because the Examiner for at least the reason that Kargman fails to disclose, teach, or suggest each and every feature of the claimed invention.

More particularly, Kargman fails to disclose, teach, or suggest at least the combined features that include "receiving an utterance at a voice input device during a current conversation with a user, wherein the utterance includes one or more words that have different meanings in different contexts," "identifying . . . a context associated with the utterance from [] short-term shared knowledge [about the current conversation] and [] long-term shared knowledge [about the user and one or more past conversations with the user]," and "establishing [an] intended meaning [for the utterance] within the identified context, wherein [a] conversational speech engine establishes the intended meaning within the identified context to disambiguate an intent that the user had in speaking the one or more words that have the different meanings in the different contexts," as recited in amended independent claim 1, for example.

Applicants note that, to the extent the Examiner alleges that Kargman describes any features relating to context, the Examiner cites the description in ¶¶ [0046]-[0048] that relates

to minimizing mistakes in recognizing certain words based on various statistics associated with a "pizza" context. For example, in ¶¶ [0046]-[0048], Kargman notes that the "words 'mushroom' and 'shrimp' may indeed sound alike to the system, . . . [yet] when a customer orders a pizza with pepperoni, it is much more likely that the second ingredient is 'mushroom' and it is much less likely that it is 'shrimp.'" However, even assuming *arguendo* that the cited passages in Kargman relate to interpreting the meaning associated with a certain word based on the context associated therewith, which Applicants do not concede, the cited passages nonetheless fail to disclose, teach, or suggest "establishing [an] intended meaning . . . within the identified context to disambiguate an intent that the user had in speaking the one or more words that have the different meanings in the different contexts." In particular, determining whether a particular spoken word was "mushroom" or "shrimp" relates, at best, to establishing what word was actually **said** rather than what a particular word **means** in a particular context.

For example, Applicants direct the Examiner's attention to ¶ [034] in Applicants' disclosure, which describes how any one particular word may have various different meanings depending on the context in which the word was said (e.g., "traffic" could refer to road conditions in a navigation context, the rock-and-roll band in a music context, or the Academy Award winning film in a movie context). Thus, amended independent claim 1 recites various features that relate to disambiguating an intent that a user had in speaking "one or more words that have different meanings in different contexts," the cited passages in Kargman relate, at best, to determining what word or words the user actually said. As such, for at least these reasons, Kargman fails to disclose, teach, or suggest disambiguating, establishing, or otherwise determining the intended meaning that the user had in speaking certain words that have different meanings depending on the context in which the words were said..

Accordingly, for at least the foregoing reasons, Kargman fails to disclose, teach, or suggest each and every feature recited in amended independent claim 1. Independent claims 15 and 29 have been amended to include features similar to those set forth in amended independent claim 1. Claims 4-7, 10-11, 14, 18-21, 24-25, 28, 32-35, 38-39, and 42 depend from and add features to one of amended independent claims 1, 15, and 29. Thus, the rejection is improper and must be withdrawn for at least the foregoing reasons.

B. CLAIMS 2-3, 12-13, 16-17, 26-27, 30-31, AND 40-41

The Examiner has rejected each of claims 2-3, 12-13, 16-17, 26-27, 30-31, and 40-41 under 35 U.S.C. § 103 as allegedly being unpatentable. In particular, the Examiner has rejected claims 2-3, 16-17, and 30-31 as allegedly being unpatentable over Kargman in view of U.S. Patent No. 5,748,841 to Morin et al. ("Morin"), claims 12, 26, and 40 as allegedly being unpatentable over Kargman in view of U.S. Patent No. 5,488,652 to Bielbly et al. ("Bielbly"), and claims 13, 27, and 41 as allegedly being unpatentable over Kargman in view of U.S. Patent No. 6,751,591 to Gorin et al. ("Gorin"). These rejections are improper and must be withdrawn for at least the reason that the references relied upon, either alone or in combination, fail to disclose, teach, or suggest each and every feature of the claimed invention.

More particularly, for at least the reasons discussed above, Kargman fails to disclose, teach, or suggest at least the combined features that include "receiving an utterance at a voice input device during a current conversation with a user, wherein the utterance includes one or more words that have different meanings in different contexts," "identifying . . . a context associated with the utterance from [] short-term shared knowledge [about the current conversation] and [] long-term shared knowledge [about the user and one or more past conversations with the user]," and "establishing [an] intended meaning [for the utterance] within the identified context, wherein [a] conversational speech engine establishes the intended meaning within the identified context to disambiguate an intent that the user had in speaking the one or more words that have the different meanings in the different contexts," as recited in amended independent claim 1, for example.

Morin fails to cure the foregoing deficiencies of Kargman for at least the reason that although Morin generally describes a system that uses "dialogue context, a dialogue model and syntactic-semantic grammar to progressively build commands . . . [that] are syntactically and semantically correct in the current context" (Abstract), Morin only describes tracking current context or other context-related information associated within a particular dialogue (or conversation). As such, even assuming *arguendo* that the Examiner has correctly characterized Morin as allegedly describing features that relate to a short-term context stack, which

Applicants do not concede, Morin fails to further disclose, teach, or suggest identifying a context associated with a current utterance from accumulated "short-term shared knowledge about the current conversation" ***and*** accumulated "long-term shared knowledge about the user . . . [and] one or more past conversations with the user." As such, for at least this reason, Morin fails to cure the foregoing deficiencies of Kargman.

Bielbly further fails to cure the foregoing deficiencies of Kargman for at least the reason that although Bielbly generally describes a system that can ask a caller about whether a call entity recognized in an utterance is incorrect (col. 18, lines 1-30), Bielbly does not describe using results from the "speech recognition algorithms" trained in the system described therein to determine or otherwise disambiguate an intended meaning or intent associated with certain words that have different meanings in different contexts. Rather, the Examiner alleges that Bielbly relates to removing an initial interpretation from a set of possible interpretations in response to the initial interpretation being incorrect (e.g., because a user says "incorrect" in response to an output from the speech recognition systems). Thus, Bielbly relates, at best, to similar speech recognition techniques as those described in Kargman (i.e., determining what word or words were actually said), but does not disclose, teach, or suggest establishing or otherwise disambiguating the meaning associated with certain words that were presumably recognized correctly based on the context in which the words were said. As such, for at least this reason, Bielbly further fails to cure the foregoing deficiencies of Kargman.

Gorin further fails to cure the foregoing deficiencies of Kargman for at least the reason that the Examiner has only relied upon Gorin as allegedly describing features that relate to multi-modal inputs that include non-voice inputs. However, even assuming *arguendo* that the Examiner has correctly characterized Gorin, which Applicants do not concede, supporting multi-modal inputs alone does not fairly disclose, teach, or suggest establishing the intended meaning associated with certain words that were spoken in the multi-modal inputs based on a context in which the words were said. As such, for at least this reason, Gorin further fails to cure the foregoing deficiencies of Kargman.

Accordingly, for at least the foregoing reasons, the references relied upon, either alone or in combination, fail to disclose, teach, or suggest each and every feature of amended

independent claim 1. Independent claims 15 and 29 have been amended to include features similar to those set forth in amended independent claim 1. Claims 2-3, 12-13, 16-17, 26-27, 30-31, and 40-41 depend from and add features to one of amended independent claims 1, 15, and 29. Thus, the rejections of these claims are improper and must be withdrawn for at least the foregoing reasons.

ALLOWABLE SUBJECT MATTER

Applicants thank the Examiner for indicating that claims 8-9, 22-23, and 36-37 would be allowable if amended to overcome the rejections under 35 U.S.C. § 112, first and second paragraph. As such, Applicants note that claims 8, 22, and 36 have been rewritten in an independent form, and further that the amendments to claims 8-9, 22-23, and 36-37 overcome the alleged rejections under 35 U.S.C. § 112, first and second paragraph.

As such, because amended claims 8-9, 22-23, and 36-37 recite subject matter that the Examiner has identified as allowable and fully address the Examiner's alleged basis for the rejections under 35 U.S.C. § 112, first and second paragraph, amended claims 8-9, 22-23, and 36-37 are in condition for allowance. Applicants respectfully request notice to that effect.

CONCLUSION

Having addressed the outstanding rejections, Applicants respectfully submit that a full and complete response has been made to the Office Action. As such, the application is in condition for allowance. Notice to that effect is respectfully requested.

If the Examiner believes, for any reason, that personal communication will expedite prosecution of this application, the Examiner is invited to telephone the undersigned at the number provided.

Date: **January 21, 2011**

Respectfully submitted,

By: _____

Syed Jafar Ali
Registration No. 58,780

PILLSBURY WINTHROP SHAW PITTMAN LLP
P.O. Box 10500
McLean, Virginia 22102
Main: 703-770-7900
Direct: 703-770-7540
Fax: 703-770-7901

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANTS :	Larry BALDWIN *et al.*	CONFIRMATION NO.:	6670
SERIAL NUMBER :	11/580,926	EXAMINER:	Eric L. Yen
FILING DATE :	October 16, 2006	ART UNIT:	2626
FOR :	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE		

SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT

Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313-1450

Sir:

Pursuant to the duty of disclosure under 37 C.F.R. §§ 1.56, 1.97 and 1.98, Applicants hereby make of record in the above-identified application the documents listed on the attached Form PTO-1449.

Applicants have listed publication dates on the attached PTO-1449 based on information presently available to the undersigned. However, the listed publication dates should not be construed as an admission that the information was actually published on the date indicated.

Applicants reserve the right to establish the patentability of the claimed invention over any of the information provided herewith, and/or to prove that this information may not be prior art, and/or to prove that this information may not be enabling for the teachings purportedly offered.

This Supplemental Information Disclosure Statement is being filed within three months of the filing date of a national application other than a continued prosecution application under 37 C.F.R. § 1.53(d), or within three months of the date of entry of the national stage as set forth in 37 C.F.R. § 1.491 in an international application, or before the mailing date of a first Office Action on the merits, or before the mailing date of a first Office Action after the filing of a request for continued examination under 37 C.F.R. § 1.114. No certification or fee is required. 37 C.F.R. § 1.97(b).

It is respectfully requested that the Examiner initial and return a copy of the enclosed PTO-1449.

Respectfully submitted,

PILLSBURY WINTHROP SHAW PITTMAN LLP

By _____
Syed Jafar Ali
Registration No.: 58,780

P. O. Box 10500
McLean, Virginia 22102
Telephone: (703) 770-7900
Telefax: (703) 770-7901

Customer No.
00909

Dated: January 21, 2011

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	1A	5,794,196	08/11/1998	Yegnanarayanan *et al.*	704	255	
	1B	5,867,817	02/02/1999	Catallo *et al.*	704	255	
	1C	6,470,315	10/22/2002	Netsch *et al.*	704	256	
	1D	6,604,075	08/05/2003	Brown *et al.*	704	270.1	
	1E	6,622,119	09/16/2003	Ramaswamy *et al.*	704	9	
	1F	6,658,388	12/02/2003	Kleindienst *et al.*	704	275	
	1G	6,708,150	03/16/2004	Hirayama *et al.*	704	243	
	1H	7,231,343	06/12/2007	Treadgold *et al.*	704	9	
	1I	7,386,443	06/10/2008	Parthasarathy *et al.*	704	201	
	1J	7,415,414	08/19/2008	Azara *et al.*	704	270	
	1K	7,809,570	10/05/2010	Kennewick *et al.*	704	257	
	1L	7,818,176	10/19/2010	Freeman *et al.*	704	270	
	1M	2002/0087525	07/04/2002	Abbott *et al.*	707	3	
	1N	2003/0046346	03/06/2003	Mumick *et al.*	709	205	
	1O	2003/0064709	04/03/2003	Gailey *et al.*	455	412	
	1P	2003/0236664	12/25/2003	Sharma	704	251	
	1Q	2004/0006475	01/08/2004	Ehlen *et al.*	704	270.1	
	1R	2004/0193408	09/30/2004	Hunt	704	209	
	1S	2004/0243417	12/02/2004	Pitts, III *et al.*	704	276	
	1T	2005/0043940	02/24/2005	Elder	704	9	
	1U	2005/0137850	06/23/2005	Odell	704	4	
	1V	2007/0179778	08/02/2007	Gong *et al.*	704	9	
	1W	2008/0140385	06/12/2008	Mahajan *et al.*	704	9	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO
	1X	WO 99/46763	09/16/1999	International				
	1Y	WO 00/21232	04/13/2000	International				
	1Z	WO 00/46792	08/10/2000	International				
	1AA	WO 2004/072954	08/26/2001	International				
	1BB	WO 2007/019318	02/15/2007	International				
	1CC	WO 2007/021587	02/22/2007	International				
	1DD	WO 2007/027546	03/08/2007	International				

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

	1EE	BAZZI, Issam *et al.*, "Heterogeneous Lexical Units for Automatic Speech Recognition: Preliminary Investigations", Processing of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 3, June 5-9, 2000, XP010507574, pages 1257-1260.

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

402593690v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	2A	2008/0177530	07/24/2008	Cross *et al.*	704	4	
	2B	2009/0216540	08/27/2009	Tessel *et al.*	704	275	
	2C	2010/0204986	08/12/2010	Kennewick *et al.*	704	226	
	2D	2010/0204994	08/12/2010	Kennewick *et al.*	704	257	
	2E	2010/0217604	08/26/2010	Baldwin *et al.*	704	275	
	2F	2010/0286985	11/11/2010	Kennewick *et al.*	704	257	
	2G	2010/0299142	11/25/2010	Freeman *et al.*	704	9	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO
	2H	WO 2007/027989	03/08/2007	International				
	2I	WO 2008/098039	08/14/2008	International				
	2J	WO 2008/118195	10/02/2008	International				
	2K	WO 2009/075912	06/18/2009	International				
	2L	WO 2009/145796	12/03/2009	International				
	2M	WO 2010/096752	08/26/2010	International				

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

	2N	O'SHAUGHNESSY, Douglas, "Interacting with Computers by Voice: Automatic Speech Recognition and Synthesis", Proceedings of the IEEE, Volume 91, No. 9, September 1, 2003, XP011100665, pages 1272-1305.

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

402593690v1

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :		(11) International Publication Number: **WO 99/46763**
G10L 5/06	A1	(43) International Publication Date: 16 September 1999 (16.09.99)

(21) International Application Number: PCT/US99/05090

(22) International Filing Date: 9 March 1999 (09.03.99)

(30) Priority Data:

60/077,337	9 March 1998 (09.03.98)	US
60/077,738	12 March 1998 (12.03.98)	US
60/077,922	13 March 1998 (13.03.98)	US

(71) Applicant: LERNOUT & HAUSPIE SPEECH PRODUCTS N.V. [BE/BE]; Flanders Language Valley 50, B–8900 Ieper (BE).

(71)(72) Applicant and Inventor: WILSON, Brian [US/US]; 125 Newport Street, Arlington, MA 02174 (US).

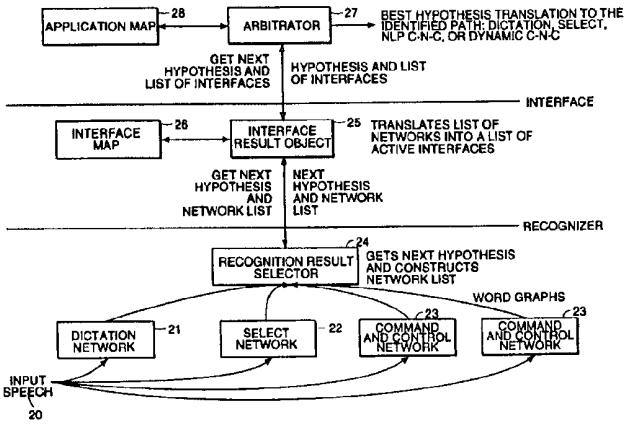
(72) Inventors: GRABHERR, Manfred; 42 Electric Avenue, Somerville, MA 02144 (US). SARUKKAI, Ramesh; Apartment #K–19, 158 Concord Road, Billerica, MA 01821 (US). GANONG, William, F., III; 128 Davis Avenue, Brookline, MA 02146 (US).

(74) Agents: SUNSTEIN, Bruce, D. et al.; Bromberg & Sunstein LLP, 125 Summer Street, Boston, MA 02110–1618 (US).

(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: APPARATUS AND METHOD FOR SIMULTANEOUS MULTIMODE DICTATION



(57) Abstract

A modeless large vocabulary continuous speech recognition system is of the type that represents an input utterance as a sequence of input vectors. The system includes a common library of acoustic model states for arrangement in sequences that form acoustic models, wherein each acoustic model is composed of a sequence of segments modeles and each segment model is composed of a sequence of model states; an input processor that compares each vector in a sequence of input vectors to a set of model states in the common library to produce a match score for each model state in the set reflecting the likelihood that such state is represented by such vector; a plurality of recognition modules and associated recognition grammars, the modules operating in parallel and using the match scores with the acoustic models to determine at least one recognition result in each of the recognition modules, wherein the plurality of recognition modules includes a dictation module for producing at least one probable dictation recognition result, a select module for recognizing a portion of visually displayed text for processing with a command, and a command module for producing at least one probable command recognition result; and an arbitrator that uses an arbitration algorithm and a score ordered queue of recognition results together with their associated recognition modules to compare the recognition results of the recognition modules to select at least one system recognition result.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxemhourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

Apparatus and Method for Simultaneous Multimode Dictation

Field of the Invention

The present invention relates to a speech recognition system, and more particularly to a flexible speech recognition system for large vocabulary continuous speech dictation which also recognizes and acts upon command and control phrases embedded in a user provided dictation stream.

Background Art

Speech recognition systems allow a user to operate and control other applications such as word processors, spreadsheets, databases, etc. Accordingly, a useful speech recognition system allows a user to perform to broad functions: (1) dictate input to an application, and (2) control the input and the application. One approach of prior art systems has been to provide separate dictation processing and control processing modes and require the user to switch between the two modes. Thus, operating mode would be definitely known by the system, since positive direction by the user was necessary to change processing modes.

Another approach was described by Hsu in United States Patent No. 5,677,991 and Yegnanarayanan in United States Patent No. 5,794,196, both of which are incorporated herein by reference in their entirety, in which input speech was parsed by both a large vocabulary isolated word recognition module and a small vocabulary continuous speech recognition module each having an associated application context. Hypotheses produced by the large vocabulary isolated word speech recognition module would correspond to dictated text while hypotheses produced by the small vocabulary continuous speech recognition module would correspond to short application specific command and control sequences. Each recognition module would produce hypotheses corresponding to the input speech and an associated recognition probability or

score. An arbitration algorithm would then select the better scoring hypothesis as a recognition result and direct the result to the associated context.

The approach of Hsu and Yegnanarayanan represented an advance in that a user of the speech recognition system no longer needed to toggle between dictation mode and command mode, rather the system automatically determined whether a given portion of an input utterance should be treated as dictated text or as application related command directives. However, Hsu and Yegnanarayanan explicitly limit the large vocabulary speech recognition module to an isolated word approach which requires a user to pause unnaturally between each word of dictated text.

Summary of the Invention

A preferred embodiment of the present invention represents a method for operating a modeless large vocabulary continuous speech recognition system of the type that represents an input utterance as a sequence of input vectors. The method includes:

(a) providing a common library of acoustic model states for arrangement in sequences that form acoustic models;

(b) comparing each vector in a sequence of input vectors to a set of model states in the common library to produce a match score for each model state in the set reflecting the likelihood that such state is represented by such vector; and

(c) using, in a plurality of recognition modules operating in parallel, the match scores with the acoustic models to determine at least one recognition result in each of the recognition modules.

In a further embodiment, each acoustic model may be composed of a sequence of segment models and each segment model may be composed of a sequence of model states. The match score may be a probability calculation or a distance measure calculation. Each recognition module may include a recognition grammar used with the acoustic models to determine the at least one

recognition result. The recognition grammar may be a context-free grammar, a natural language grammar, or a dynamic command grammar. In addition, or alternatively, the method may further include comparing the recognition results of the recognition modules to select at least one system recognition result. The step of comparing may use an arbitration algorithm and a score ordered queue of recognition results and associated recognition modules. The plurality of recognition modules may include one or more of a dictation module for producing at least one probable dictation recognition result, a select module for recognizing a portion of visually displayed text for processing with a command, and a command module for producing at least one probable command recognition result.

A related embodiment provides a modeless large vocabulary continuous speech recognition system of the type that represents an input utterance as a sequence of input vectors. The system includes a common library of acoustic model states for arrangement in sequences that form acoustic models; an input processor that compares each vector in a sequence of input vectors to a set of model states in the common library to produce a match score for each model state in the set reflecting the likelihood that such state is represented by such vector; and a plurality of recognition modules operating in parallel that use the match scores with the acoustic models to determine at least one recognition result in each of the recognition modules.

In a further embodiment, each acoustic model may be composed of a sequence of segment models and each segment model may be composed of a sequence of model states. The match score may be a probability calculation or a distance measure calculation. Each recognition module may include a recognition grammar used with the acoustic models to determine the at least one recognition result. The recognition grammar may be a context-free grammar, a natural language grammar, or a dynamic command grammar. In addition, or alternatively, the system may further include an arbitrator that compares the recognition results of the recognition modules to select at least one system

recognition result. The arbitrator may include an arbitration algorithm and a score ordered queue of recognition results and associated recognition modules. The plurality of recognition modules may include one or more of a dictation module for producing at least one probable dictation recognition result, a select module for recognizing a portion of visually displayed text for processing with a command, and a command module for producing at least one probable command recognition result.

Brief Description of the Drawings

The present invention will be more readily understood by reference to the following detailed description taken with the accompanying drawings, in which:

Fig. 1 illustrates a block diagram of simultaneous recognition networks established for active applications in accordance with a preferred embodiment of the invention.

Fig. 2 illustrates a block diagram of operation of an arbitration algorithm in accordance with a preferred embodiment of the invention.

Detailed Description of Specific Embodiments

Large vocabulary speech recognition dictation systems heretofore known require an extended sequence of correctly formatted spoken commands even to perform relatively simple command and control functions. See, for example, application number 08/669,242, filed June 24, 1996, allowed for issuance on 2/4/98, a continuation-in-part of application number 08/496,979, filed June 30, 1995, now issued as United States Patent No. 5,677,991; such patent and such applications are hereby incorporated herein by reference. Moreover, some relatively common features are not even supported in existing speech recognition dictation systems. For instance, to create a table in some existing applications, the system will not recognize spoken commands to create a table, but requires use of a mouse or keyboard. Even in command sequences which

are recognized, the format and sequence of the commands must be followed with little or no variation to perform the desired function correctly.

A preferred embodiment of the present invention correctly identifies and performs command and control functions embedded by the user in a stream of dictated text without requiring an exact sequence or predefined format. For instance, to create a table, the user may say:

"Insert a two by three table here," or

"Add a table here," or

"Make a two by three table," or

"Create a table here."

Similar commands composed by the user in a natural language form comfortable at that particular time may be properly recognized and acted upon by the system.

Figs. 1 and 2 illustrate in broad terms a preferred embodiment of the present invention. Figs. 1 and 2 also show various levels of operation of the embodiment. The highest level is the application level **18**, which exists above the line **12** representing the Speech Application Programming Interface (SAPI).

As illustrated in Fig. 1, a preferred embodiment implements a flexible speech recognition dictation system which incorporates both a large vocabulary continuous speech dictation path **13** and one or more limited-vocabulary, application-associated command and control paths **14** which operate simultaneously in parallel on a user provided spoken input. These two paths are here divided by dashed line X-X. The large vocabulary continuous speech dictation path **13** of the embodiment employs a combination of acoustic models and language models to perform multiple search passes on the spoken input and generates scores indicating the degree of match of the input utterance with an identified sequence of the respective models. The limited-vocabulary, application-associated command and control path **14** utilizes acoustic models in combination with context free grammars **141** to generate scores indicating the degree of match of the input utterance with at least one of the recognizable

commands. An arbitration algorithm then selects among the scores and models generated by the recognition networks. Preferably, the scores generated by the respective networks are scaled by a factor or factors empirically trained to minimize incursions by each of the networks on correct results from the other vocabulary. In an embodiment, the input speech may be provided to the system from an audio object, be processed by the speech recognizer, and then provided as output to another computer application, such as a word processing program or accounting spreadsheet, via the speech application programming interface (SAPI) **12**.

In Fig. 1 are also shown successively deeper layers of the embodiment. The large vocabulary dictation context **13** operates across the SAPI **12** in a Voice Dictation Object (VDO) **11** which typically causes the display of dictated text. Just as a user in a keyboard operated word processing program may have multiple documents open at one time, a dictation system user may multiple open VDOs. The natural language process (NLP) grammars and dynamic grammar constituting a part **14** of the embodiment providing command and control contexts may act on an application such as a word processing application. Fig. 1 also shows levels beneath the SAPI including the engine level **15**, wherein the various grammar rules are applied, and the interface level **16**, wherein multiple dictation contexts at the VDO level **18** are mapped to a single dictation context for use at the recognizer level **17**. At the recognizer level **17**, each active context constructs working recognition hypotheses and associated probabilities.

Figure 2 illustrates operation of the system to arbitrate among the hypotheses generated for a given input utterance by each active context. Arbitration consists of selecting the best scoring recognition result hypothesis, identifying the active context which produced that hypothesis and directing the hypothesis to the corresponding application-level path. Figure 2 shows an input utterance **20** being directed to a recognition network for each active context; a dictation network **21**, a select network **22**, and command and control networks

23. The dictation network uses a sophisticated and elaborate processing scheme which requires substantial system computational resources; however, use of intermediate dictation network calculations by the other recognition networks, along with other efficiency measures, allows existing commercial computer systems to perform the simultaneous multiple recognition network processing utilized by preferred embodiments of the present invention.

The recognition hypotheses and their associated probabilities are grouped together with their associated context by the recognition result selector **24** which forms a precedence ordered queue of (hypothesis, context) ordered pairs known as the hypothesis list. The hypothesis list is retrieved by the interface result object **25** which uses the interface map **26** to map the hypothesis list onto the active context models. The arbitrator **27** gets the translated hypothesis list from the interface result object **25,** selects the best scoring hypothesis, and directs it to the appropriate application-level path according to the application map **28.**

In common with many prior art systems, a preferred embodiment of the system of the present invention operates by transforming an input speech signal into a sequence of digitally encoded speech frames as is well known in the art. The input frames are further processed by describing them each with *N* parameters, producing a sequence of *N*-dimensional vectors. A preferred embodiment then utilizes sequence state acoustic models in the form of Hidden Markov Models (HMMs) with continuous observation densities as is known in the art and described, for example, in Rabiner and Juang, *Fundamentals of Speech Recognition*, pp. 350-52, Prentice Hall, 1993, which reference is hereby incorporated herein by reference. The initial processing of the input data also estimates the beginning and end of a word or phrase based on an analysis of energy levels such that some adjustable sensitivity threshold is set to distinguish user speech from background noise.

In a preferred embodiment, the acoustic models represent demi-triphones--models of the acoustic transition from the middle of one phoneme (basic speech sound) to the middle of another phoneme, which are left-right

context sensitive; that is sensitive to the preceding and following sound units. The sequence of states in the models actually is derived from a relatively small pool of approximately 2000 states which model every specific speech sound. The states are represented as mixture models such that the acoustic modeling measures the distance from the input speech to various model parameters. Typically, the model states are a Gaussian family of probability distribution functions in which the conditional probability of the input, given the state, is a weighted sum of Gaussian probabilities. However, a preferred embodiment uses a simplified version in which the family has an assigned weight such that for a given input speech sequence, the distance is measured from the input to the nearest Gaussian.

In some cases, the input frames are relatively far from the nearest Gaussian so that it would be computationally inefficient to perform the relatively complex computation described above. In such cases, a preferred embodiment simply uses the simpler measurement models produced by the process of vector quantization (VQ) in which the *N*-dimensional vector acoustic models are represented by sequences of standard or prototype states. Rather than representing spectral distributions, the state indices identify or correspond to probability distribution functions. The state spectral index essentially serves as a pointer into a table which identifies, for each state index, the set of probabilities that each prototype frame or VQ index will be observed to correspond to that state index. The table is, in effect, a precalculated mapping between all possible frame indices and all state indices. Thus, for comparing a single frame and single state, a distance measurement or a measure of match can be obtained by directly indexing into the tables using the respective indices and combining the values obtained with appropriate weighting. It is thus possible to build a table or array storing a distance metric representing the closeness of match of each standard or prototype input frame with each standard or prototype model state. This matrix is further compressed by methods described

in United States Patent No. 5,546,499, issued August 13, 1996 to Lynch et al., which is hereby incorporated into this application by reference.

Natural variations in speaking rate require that some method be employed for time aligning a sequence of frames representing an unknown speech segment with each sequence of acoustic model states representing a vocabulary word. This process is commonly referred to as time alignment. In word recognition, the sequence of frames which constitute the unknown speech segment taken together with a sequence of states representing a vocabulary model, in effect, define a matrix and the time aligning process involves finding a path across the matrix which produces the best score, e.g., least distance or cost. The distance or cost is typically arrived at by accumulating the cost or distance values associated with each pairing of frame index with state index as described previously with respect to the vector quantization process.

In addition to words which are to be recognized, the vocabulary may also include word initial state and word final state noise models along with models of common intrusive noises, e.g. paper rustling, door closing, or a cough. When an acoustic input is best matched with one of these models, a null output or no output is provided. In addition, a preferred embodiment also uses models of word-to-word transition sounds, also called word "glues."

In continuous speech recognition, states corresponding to phones or other sub-units of speech are typically interconnected in a network and decoded in correspondence with the ongoing utterance. A score is progressively built up as the utterance proceeds. This total score is a function both of the degree of match of the utterance with the decoded path and the length of the utterance. A continuous speech recognition system will typically identify the best scoring model sequence and may also identify a ranked list of possible alternative sequences.

The large vocabulary continuous speech dictation network employs various language models to additionally process the input speech, while in parallel one or more limited-vocabulary, application-associated command and

control networks uses the acoustic model processing together with context free grammars to process and score the same input speech. In a prior art speech recognizer, this would mean implementing two or more independent and complete speech recognizers. However, preferred embodiments of the present invention share the processing results of the acoustic models to reduce the computational load and the total amount of work done by the system. This computational savings allows a single commercially available computer processor to perform the multiple simultaneous path processing needed to realize preferred embodiments of the present invention.

The large vocabulary continuous speech dictation recognition network employs a three pass search utilizing both a tree-structured network, a linear network, and networks for initial and final noises and for the word glues. The first pass uses at every time step, for reasonably good word beginning times, both the linear network and the unigram tree networks. The linear network, however, is used in conjunction with a bigram word history (i.e., the previous word context is used), and therefore, hopefully will not start a large number of words. The tree lexicon, on the other hand, shares the computation of the words (*e.g.*,which share common phone transcription prefixes), and therefore is not that computationally expensive due to shared system resources; furthermore, most state sequences will be pruned after the first few phones. While a transition into a linear network incorporates the appropriate bigram language model score, the unigram scores of the words are factored into the tree representation of the lexicon. Thus, a "smeared" language model score enables earlier pruning of unwanted unigram extensions.

The first search pass sequences the frames into networks with accumulated scores and calculates times of word endings using Hidden Markov Models (HMMs) which are statistical representations of word sequence probabilities. The statistical processing performed by the first pass includes computing for each state the mixture model distances in a four-layered approach employing, in turn, a squeezed matrix, best branch entries, mixture model

approximations, and heuristic stopping of computation. The calculation of the mixture model distances determines the distance of each state from the input frame. The mixture model distance calculations performed by the dictation network are also cached for use by the other recognition networks. When the first pass calculates that a word end has been reached in the input utterance the state scores are updated, and active states of the unigram tree, linear network, word initial and final noises network and the glue network are pruned to delete active states below a threshold value.

The second search pass uses first pass tracking and score array based network construction to produce a word graph of states and probabilities. Starting with the word end scores of the first pass, acoustic and language model distances are calculated for every word that ends that way. This process is described in greater detail in the Lynch patent previously cited. The second pass then uses the mixture model distances (again, also stored to a cache for use by the other recognition networks) to identify sequences of states known as arcs which represent fundamental sound units and compute their acoustic scores, creating a word graph of word hypotheses and associated probabilities. The third pass shares the language model used in the first pass in conjunction with trigram word models to produce a ranking of the most probable hypotheses.

Besides the above described dictation context and its associated network of language model word searches, a preferred embodiment of the present invention implements three other contexts and associated networks:

- select context,
- natural language command and control context, and
- dynamic command and control context.

The select context feature is to the word sequences displayed in the voice dictation object (VDO). When the user wishes to process a portion of the word sequence produced displayed in the VDO, the select context and its associated command network is used to define the word sequence desired much as a

mouse in a keyboard based word processing program may be used to select a portion of text.

A preferred embodiment of the present invention utilizes two independent command context types, one type using a natural language process (NLP) grammar (actually several NLP grammars are used) and one type using a dynamic grammar. When the score from a command and control context is sufficiently small, it means that the given sequence is more probably intended by the user to be a command and control sequence rather than dictated text. Each natural language process grammar utilizes context free grammar rules to parse a sequence of input words and return a score representing the likelihood that a given sequence represents a command and control sequence from the user in natural language format. The dynamic grammar utilizes its own context free grammar and designed to detect and implement the short commands available on a drop-down GUI menu; for example, "Bold on" "Bold off" "Spell check document."

The select and command and control networks utilize acoustic models corresponding to the vocabulary recognized by their respective grammars. They perform recognition searches similar to the first and second pass searches described above with respect to the dictation recognition network, analyzing mixture model distances and producing word graphs of recognition hypotheses and their associated probabilities. By using the cache of mixture model distance calculations produced by the dictation network rather than performing these calculations independently for each recognition network, substantial savings are realized in systems computational resources. Thus, preferred embodiments of the present invention may be implemented in existing commercially available computer systems.

Moreover, preferred embodiments may also realize operating efficiencies by the non-dictation recognition networks "sharing" their respective search beamwidths. As described above, search beamwidth refers to the pruning or deleting of recognition hypotheses which have scores beyond a threshold value

set by the best scoring hypothesis. Preferred embodiments prune the poorly scoring hypotheses in the non-dictation networks according to the best overall hypothesis score by any network in conjunction with common threshold values. Thus, rather than have each network produce its n-best scoring hypotheses, networks in which scores are poor compared to the other networks have all or most of their hypotheses pruned off. In other words, rather than considering the best hypotheses from each network, sharing the pruning beamwidth allows preferred embodiments to consider the best scoring hypotheses from all the networks considered as a whole. Again, this results in significant computational efficiencies being realized.

In an alternative embodiment, even greater computational efficiencies may be realized by utilizing a single recognition network such as that described above with respect to the dictation network. The non-dictation contexts such as select context and command and control context would attempt to parse the output of the single recognition network. Recognition results parsed by a given context would be recognized as commands for that context rather than dictated text. This embodiment requires that a careful compromise be reached between the overall system's recognition of a given word or phrase as dictated text or as commands.

As indicated previously, the different recognition contexts and their associated networks employ different types of models and different scoring mechanisms so that the scores are not directly comparable. Relative scaling of the scores is applied to minimize or avoid intrusions by each recognition network on correct translations from the other recognition network. The scoring produced by the dictation network with its combination of acoustic models, multipass language models, and dictation grammar must be adjusted to be comparable with the scoring produced by the select network and the command and control networks.

An arbitration algorithm for selecting among the competing contexts and hypotheses combines and orders the various scores obtained, and then selects

the top scoring hypothesis and context for output via the SAPI as appropriate to either the visual display of a user application program by means of a voice dictation object (VDO), or to a command and control function within an application speech center which tracks what actions are occurring within the application.

In a preferred embodiment, the large vocabulary continuous recognition dictation context, the select context and some number of command and control contexts have an initial precedence ordering with respect to the others. For instance, in a preferred embodiment, the highest precedence is assigned to the command and control contexts, then to the select context, and lowest is the dictation context. The arbitration algorithm finds the best scoring parseable hypothesis and its context (all dictation and selection results are considered parseable). If the same hypothesis is parseable by more than one context, the arbitration algorithm selects the hypothesis with the highest precedence. Precedence ordering ensures that when an utterance is recognized by both a command and control context and by the dictation context, the utterance will be treated as a command rather than dictation text, since that is the most likely user-desired action.

In a preferred embodiment, the arbitration algorithm initially generates data structures of sets of (hypothesis, score, list of networks) in a score-ordered queue produced at the recognizer level. From the word graph generated by each network, including dictation, the arbitration algorithm gets the score of the best hypothesis and inserts the set of (hypothesis, score, list of networks) in the score-ordered queue. Until the algorithm is done or the queue is empty: the best hypothesis in the queue is found, and all the sets of (hypothesis, score, list of networks) in the queue which have the same hypothesis and score are found. Then, all the hypotheses not yet in the queue having the same hypothesis and score are found by, for each network not in the list of networks associated with the best hypothesis, analyzing the next hypothesis and score, until the score is worse than the current score (or until some maximum number of hypotheses per

context has been reached, *e.g.*, 15). If that is the same hypothesis and score as the current best scoring hypothesis, it is added to the associated list of networks, otherwise, the next hypothesis and score from that network is tested.

At this parsing step, the algorithm determines whether or not to get another (hypothesis, score, list of networks) candidate by checking whether some maximum number of hypotheses has been reached (*e.g.*, 100). There may be more parses than this, since there may be multiple networks per hypothesis, but only a certain maximum number of hypotheses is considered. The hypotheses, in fact, may not be distinct--the same hypothesis may be considered in different networks and with different scores. That is, the same hypothesis may come from the dictation network and from a command and control network at different times, with different scores.

Next, a list of networks associated with this hypothesis is made (often containing only one entry), and this list is ordered by precedence. At the interface level, the list of networks associated with the best scoring hypothesis is translated into a list of contexts associated with the best scoring hypothesis. Then, at the engine level, the first context in the list is found which parses the hypothesis, and if there is such a context it is returned along with its corresponding hypothesis and the algorithm is done. Otherwise, if the queue was empty, the algorithm returns an empty result.

In the normal course, the arbitration algorithm selects a best hypothesis and associated application path to direct it to. If the hypothesis is recognized as dictated text, then that is sent to the VDO. If the hypothesis is recognized as a command to select text or for a command and control grammar, the hypothesis is translated into the appropriate command associated with the recognized hypothesis.

In view of the foregoing it may be seen that several objects of the present invention are achieved and other advantageous results have been attained. As various changes could be made in the above constructions without departing from the scope of the invention, it should be understood that all matter

contained in the above description or shown in the accompanying drawings shall be interpreted as illustrative and not in a limiting sense.

What is claimed is:

1. A method for operating a modeless large vocabulary continuous speech recognition system of the type that represents an input utterance as a sequence of input vectors, the method comprising:

(a) providing a common library of acoustic model states for arrangement in sequences that form acoustic models;

(b) comparing each vector in a sequence of input vectors to a set of model states in the common library to produce a match score for each model state in the set reflecting the likelihood that such state is represented by such vector; and

(c) using, in a plurality of recognition modules operating in parallel, the match scores with the acoustic models to determine at least one recognition result in each of the recognition modules.

2. A method according to claim 1, wherein each acoustic model is composed of a sequence of segment models and each segment model is composed of a sequence of model states.

3. A method according to claim 1, wherein the match score is a probability calculation or a distance measure calculation.

4. A method according to claim 1, wherein each recognition module includes a recognition grammar used with the acoustic models to determine the at least one recognition result.

5. A method according to claim 4, wherein the recognition grammar is a context-free grammar, a natural language grammar, or a dynamic command grammar.

6. A method according to claim 1, further including:

(d) comparing the recognition results of the recognition modules to select at least one system recognition result.

7. A method according to claim 6, wherein the step of comparing uses an arbitration algorithm and a score ordered queue of recognition results and associated recognition modules.

8. A method according to claim 1, wherein the plurality of recognition modules includes a dictation module for producing at least one probable dictation recognition result.

9. A method according to claim 1, wherein the plurality of recognition modules includes a command module for producing at least one probable command recognition result.

10. A method according to claim 1, wherein the plurality of recognition modules includes a select module for recognizing a portion of visually displayed text for processing with a command.

11. A method according to claim 1, wherein the plurality of recognition modules includes a dictation module for producing at least one probable dictation recognition result, a select module for recognizing a portion of visually displayed text for processing with a command, and a command module for producing at least one probable command recognition result.

12. A method for operating a modeless large vocabulary continuous speech recognition system of the type that represents an input utterance as a sequence of input vectors, the method comprising:

(a) providing a common library of acoustic model states for arrangement in sequences that form acoustic models, wherein each acoustic model is composed of a sequence of segment models and each segment model is composed of a sequence of model states;

(b) comparing each vector in a sequence of input vectors to a set of model states in the common library to produce a match score for each model state in the set reflecting the likelihood that such state is represented by such vector;

(c) in a plurality of recognition modules operating in parallel, each having an associated recognition grammar, using the match scores with the acoustic models to determine at least one recognition result in each of the recognition modules, wherein the plurality of recognition modules includes a dictation module for producing at least one probable dictation recognition result, a select module for recognizing a portion of visually displayed text for processing with a command, and a command module for producing at least one probable command recognition result; and

(d) comparing the recognition results of the recognition modules with an arbitration algorithm and a score ordered queue of recognition results and associated recognition modules to select at least one system recognition result.

13. A modeless large vocabulary continuous speech recognition system of the type that represents an input utterance as a sequence of input vectors, the system comprising:

a common library of acoustic model states for arrangement in sequences that form acoustic models;

an input processor that compares each vector in a sequence of input vectors to a set of model states in the common library to produce a match score for each model state in the set reflecting the likelihood that such state is represented by such vector; and

a plurality of recognition modules operating in parallel that use the match scores with the acoustic models to determine at least one recognition result in each of the recognition modules.

14. A system according to claim 13, wherein each acoustic model is composed of a sequence of segment models and each segment model is composed of a sequence of model states.

15. A system according to claim 13, wherein the match score is a probability calculation or a distance measure calculation.

16. A system according to claim 13, wherein each recognition module includes a recognition grammar used with the acoustic models to determine the at least one recognition result.

17. A system according to claim 16, wherein the recognition grammar is a context-free grammar, a natural language grammar, or a dynamic command grammar.

18. A system according to claim 13, further including:

(d) an arbitrator that compares the recognition results of the recognition modules to select at least one system recognition result.

19. A system according to claim 18, wherein the arbitrator includes an arbitration algorithm and a score ordered queue of recognition results and associated recognition modules.

20. A system according to claim 13, wherein the plurality of recognition modules includes a dictation module for producing at least one probable dictation recognition result.

21. A system according to claim 13, wherein the plurality of recognition modules includes a command module for producing at least one probable command recognition result.

22. A system according to claim 13, wherein the plurality of recognition modules includes a select module for recognizing a portion of visually displayed text for processing with a command.

23. A system according to claim 13, wherein the plurality of recognition modules includes a dictation module for producing at least one probable dictation recognition result, a select module for recognizing a portion of visually displayed text for processing with a command, and a command module for producing at least one probable command recognition result.

24. A modeless large vocabulary continuous speech recognition system of the type that represents an input utterance as a sequence of input vectors, the system comprising:

a common library of acoustic model states for arrangement in sequences that form acoustic models, wherein each acoustic model is composed of a sequence of segment models and each segment model is composed of a sequence of model states;

an input processor that compares each vector in a sequence of input vectors to a set of model states in the common library to produce a match score for each model state in the set reflecting the likelihood that such state is represented by such vector;

a plurality of recognition modules and associated recognition grammars, the modules operating in parallel and using the match scores with the acoustic models to determine at least one recognition result in each of the recognition modules, wherein the plurality of recognition modules includes a dictation

module for producing at least one probable dictation recognition result, a select module for recognizing a portion of visually displayed text for processing with a command, and a command module for producing at least one probable command recognition result; and

an arbitrator that uses an arbitration algorithm and a score ordered queue of recognition results together with their associated recognition modules to compare the recognition results of the recognition modules to select at least one system recognition result.

[83931]

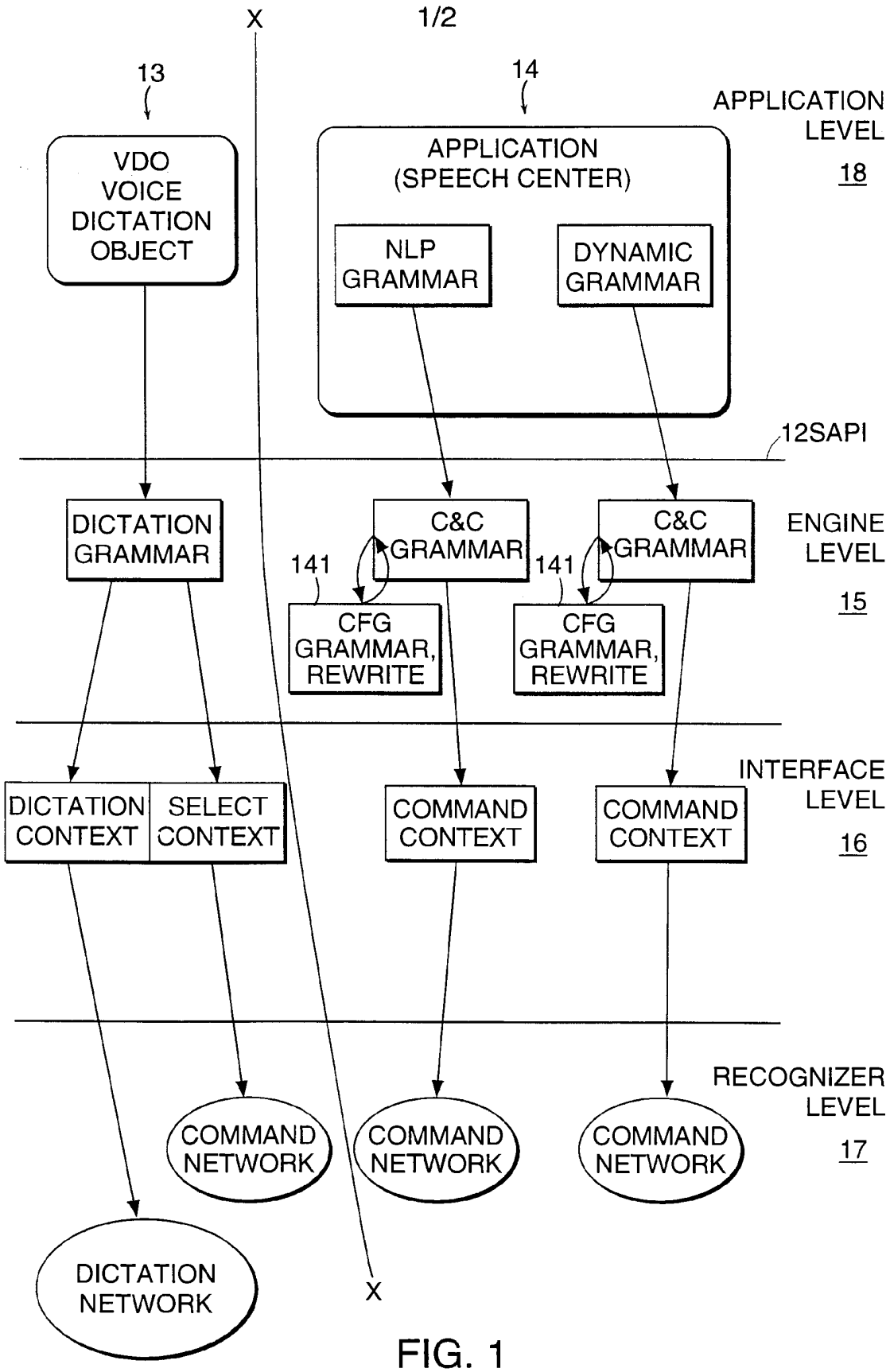

X
1/2
13
14
APPLICATION LEVEL
18
VDO VOICE DICTATION OBJECT
APPLICATION (SPEECH CENTER)
NLP GRAMMAR
DYNAMIC GRAMMAR
12SAPI
DICTATION GRAMMAR
C&C GRAMMAR
C&C GRAMMAR
ENGINE LEVEL
15
141
141
CFG GRAMMAR, REWRITE
CFG GRAMMAR, REWRITE
INTERFACE LEVEL
16
DICTATION CONTEXT
SELECT CONTEXT
COMMAND CONTEXT
COMMAND CONTEXT
RECOGNIZER LEVEL
17
COMMAND NETWORK
COMMAND NETWORK
COMMAND NETWORK
DICTATION NETWORK
X

FIG. 1

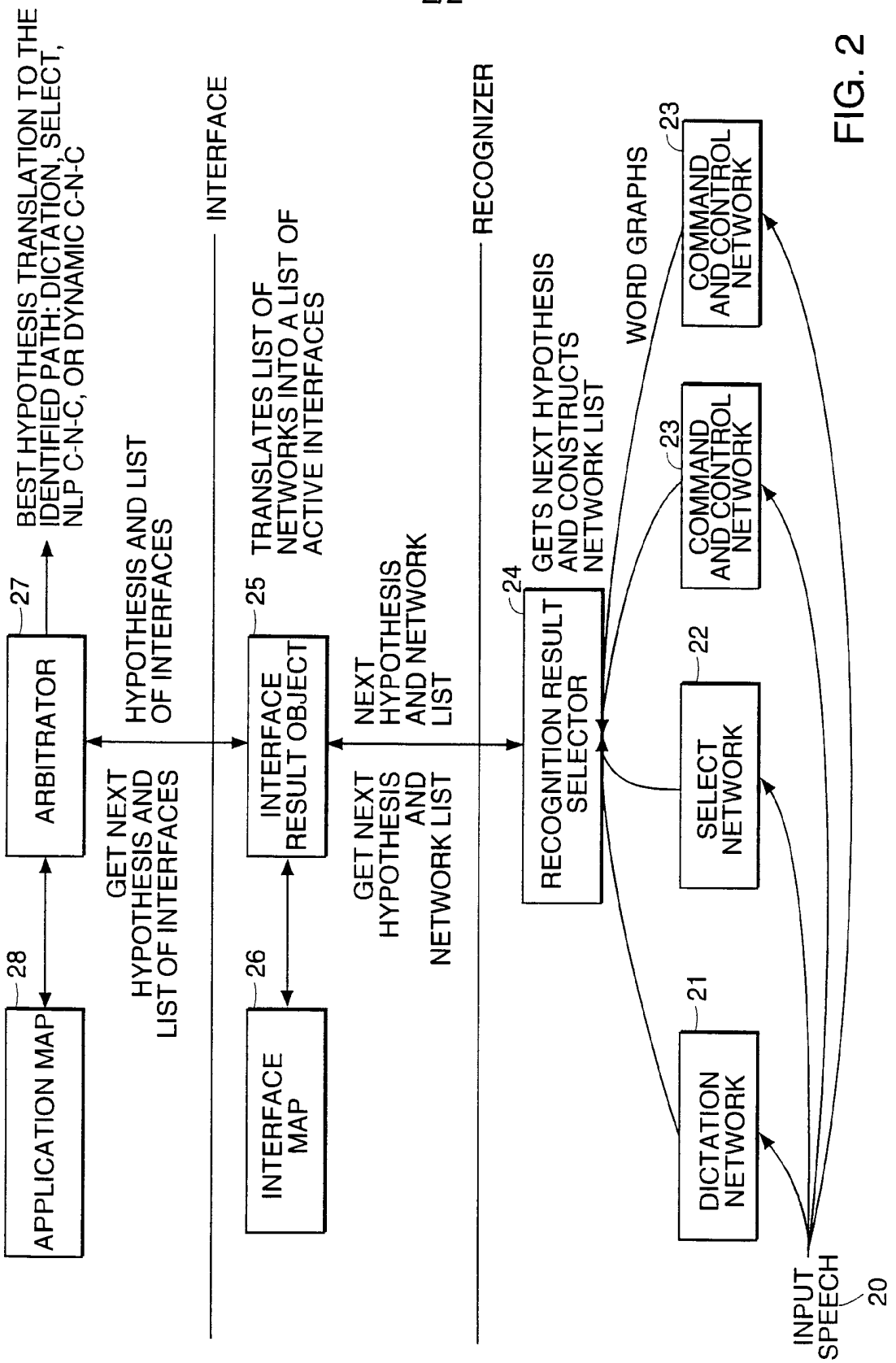

APPLICATION MAP
28
ARBITRATOR
27
BEST HYPOTHESIS TRANSLATION TO THE IDENTIFIED PATH: DICTATION, SELECT, NLP C-N-C, OR DYNAMIC C-N-C
GET NEXT HYPOTHESIS AND LIST OF INTERFACES
HYPOTHESIS AND LIST OF INTERFACES
INTERFACE
INTERFACE MAP
26
INTERFACE RESULT OBJECT
25
TRANSLATES LIST OF NETWORKS INTO A LIST OF ACTIVE INTERFACES
GET NEXT HYPOTHESIS AND NETWORK LIST
NEXT HYPOTHESIS AND NETWORK LIST
RECOGNIZER
RECOGNITION RESULT SELECTOR
24
GETS NEXT HYPOTHESIS AND CONSTRUCTS NETWORK LIST
WORD GRAPHS
DICTATION NETWORK
21
SELECT NETWORK
22
COMMAND AND CONTROL NETWORK
23
COMMAND AND CONTROL NETWORK
23
INPUT SPEECH
20

FIG. 2

INTERNATIONAL SEARCH REPORT

national Application No
PCT/US 99/05090

A. CLASSIFICATION OF SUBJECT MATTER
IPC 6 G10L5/06

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 6 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5 677 991 A (HSU DONG ET AL) 14 October 1997 (1997-10-14) cited in the application figure 2	1,12,13, 24
A	DE 196 35 754 A (SIEMENS AG) 5 March 1998 (1998-03-05) abstract figure 1 page 2, line 50 - line 54	1,12,13, 24
A	WO 96 13829 A (MOTOROLA INC) 9 May 1996 (1996-05-09) abstract page 13, line 11 - line 20 figure 2	1,12,13, 24

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
13 July 1999	23/07/1999

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Krembel, L

1

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Information on patent family members

national Application No
PCT/US 99/05090

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5677991 A	14-10-1997	US 5794196 A	11-08-1998
DE 19635754 A	05-03-1998	WO 9810413 A	12-03-1998
		EP 0925578 A	30-06-1999
WO 9613829 A	09-05-1996	US 5638486 A	10-06-1997
		AU 3274395 A	23-05-1996

Form PCT/ISA/210 (patent family annex) (July 1992)

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : H04L	A2	(11) International Publication Number: WO 00/21232
		(43) International Publication Date: 13 April 2000 (13.04.00)

(21) International Application Number: PCT/US99/23008

(22) International Filing Date: 1 October 1999 (01.10.99)

(30) Priority Data:

60/102,957	2 October 1998 (02.10.98)	US
60/117,595	27 January 1999 (27.01.99)	US

(81) Designated States: CA, CN, IL, IN, JP, KR, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(71) Applicant *(for all designated States except US)*: INTERNATIONAL BUSINESS MACHINES CORPORATION [US/US]; Old Orchard Road, Armonk, NY 10504 (US).

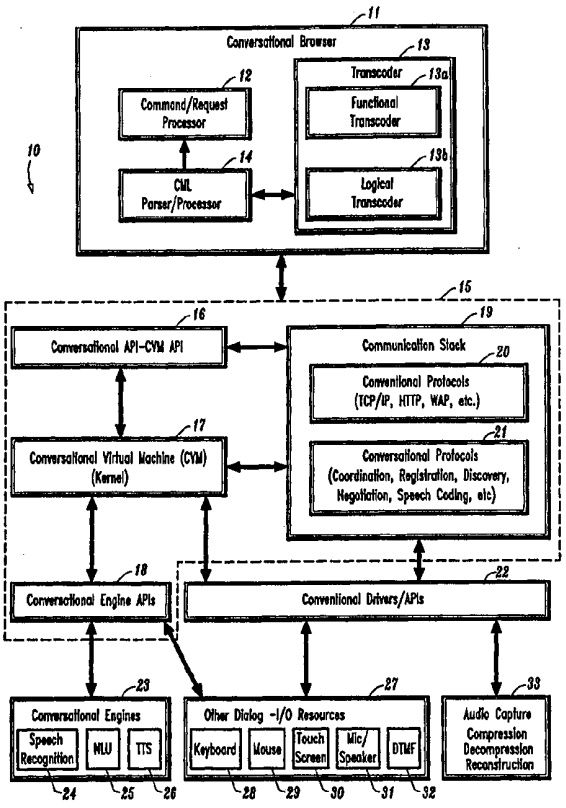
(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: GOPALAKRISHNAN, Ponani [IN/US]; 3073 Radcliff Drive, Yorktown Heights, NY 10598 (US). LUCAS, Bruce, D. [US/US]; 2408 Mill Pond Road, Yorktown Heights, NY 10598 (US). MAES, Stephane, H. [BE/US]; 1 Wintergreen Hill Road, Danbury, CT 06811 (US). NAHAMOO, David [IR/US]; 12 Elmwood Road, White Plains, NY 10605 (US). SEDIVY, Jan [CZ/CZ]; U lesa 11, Praha (CZ).

(74) Agent: OTTERSTEDT, Paul, J.; International Business Machines Corporation, Yorktown IP Law Department, T.J. Watson Research Center, Route 134 and Kitchawan Road, Yorktown Heights, NY 10598 (US).

(54) Title: CONVERSATIONAL BROWSER AND CONVERSATIONAL SYSTEMS

(57) Abstract

A conversational browsing system (10) comprising a conversational browser (11) having a command and control interface (12) for converting speech commands or multi-modal input from I/O resources (27) into navigation request, a processor (14) for parsing and interpreting a CML (conversational markup language) file, the CML file comprising meta-information representing a conversational user interface for presentation to a user. The system (10) comprises conversational engines (23) for decoding input commands for interpretation by the command and control interface and decoding meta-information provided by the CML processor for generating synthesized audio output. The browser (11) accesses the engine (23) via system calls through a system platform (15). The system includes a communication stack (19) for transmitting the navigation request to a content server and receiving a CML file from the content server based on the navigation request. A conversational transcoder (13) transforms presentation material from one modality to a conversational modality. The transcoder (13) includes a functional transcoder (13a) to transform a page of GUI to a page of CUI (conversational user interface) and a logical transcoder (13b) to transform business logic of an application, transaction or site into an acceptable dialog. Conversational transcoding can convert HTML files into CML files that are interpreted by the conversational browser (11).



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

CONVERSATIONAL BROWSER AND CONVERSATIONAL SYSTEMS

This application is based on provisional applications U.S. Serial Number 60/102,957, filed on October 2, 1998, and U.S. Serial No. 60/117,595 filed on January 27, 1999.

BACKGROUND

1. Technical Field:

The present invention relates generally to systems and methods for accessing information and, more particularly, to a conversational browser that provides unification of the access to various information sources to a standard network protocol (such as HTTP) thereby allowing a pure GUI (graphical user interface) modality and pure speech interface modality to be used individually (or in combination) to access the same bank of transaction and information services without the need for modifying the current networking infrastructure.

2. Description of Related Art:

Currently, there is widespread use of IVR (Interactive Voice Response) services for telephony access to information and transactions. An IVR system uses spoken directed dialog and generally operates as follows. A user will dial into an IVR system and then listen to an audio prompts that provide choices for accessing certain menus and particular information. Each choice is either assigned to one number on the phone keypad or associated with a word to be uttered by the user (in voice enabled IVRs) and the user will make a desired selection by pushing the appropriate button or uttering the proper word. Conventional IVR applications are typically written in specialized script languages that are offered by manufacturers in various incarnations and for different HW (hardware) platforms. The development and maintenance of such IVR applications requires qualified staff. Conventional IVR applications use specialized (and expensive) telephony HW, and each IVR applications uses different SW (software) layers for accessing legacy database servers. These layers must be specifically designed for each application.

Furthermore, IVR systems are not designed to handle GUI or other modalities other than DTMF and speech. Although it is possible to mix binary data

and voice on a conventional analog connection, it is not possible to do so with a conventional IVR as the receiver. Therefore, IVR systems typically do not allow data/binary input and voice to be merged. Currently, such service would require a separate system configured for handling binary connections (e.g. a form of modem). In the near future, Voice over IP (VoIP) and wireless communication (e.g., GSM) will allow simultaneous transmission of voice and data. Currently, more than one simultaneous call is needed for simultaneous exchange of binary and voice (as it is explained to be useful later to adequately handle specialized tasks) or it will require a later call or callback for asynchronous transmission of the data. This is typically not convenient. In particular, the data exchange can be more than sending or receiving compressed speech and information related to building a speech UI, it can also be the necessary information to add modalities to the UI (e.g. GUI). Assuming that services will be using multiple lines to offer, for example, a voice in / web out (or voice in / web and voice out) modality where the result of the queries and the presentation material also result into GUI material (e.g. HTML displayed on a GUI browser like Netscape Navigator), the service provider must now add all the IT infrastructure and backend to appropriately networked and synchronize its backends, IVR and web servers. A simple but very difficult task is the coordination between the behavior/evolution of the speech presentation material with respect to the GUI or HTML portion of the presentation.

With the rapidly increasing evolution of mobile and home computing, as well as the prevalence of the Internet, the use of networked PCs, NCs, information kiosks and other consumer devices (as opposed to IVR telephony services) to access information services and transactions has also become widespread. Indeed, the explosion of Internet and Intranet has afforded access to virtually every possible information source, database or transaction accessible through what is generally known as a GUI "Web browser," with the conversion of the data and the transactions being performed via proxies, servers and/or transcoders.

In general, a Web browser is an application program (or client program) that allows a user to view and interact with information on the WWW (World Wide Web or the "Web") (i.e., a client program that utilizes HTTP (Hypertext Transfer Protocol) to make requests of HTTP servers on the Internet). The HTTP servers on the Internet include "Web pages" that are written in standard HTML (Hypertext Markup language). An Internet Web page may be accessed from an HTTP server over a packet-switched network, interpreted by the Web browser, and then presented to the user in graphical form. The textual information presented to the user includes

highlighted hyperlinks to new sources of information. The user can then select a hyperlink by, e.g., clicking on the with mouse, to download a new Web page for presentation by the Web browser. The access to legacy databases over the Internet is enabled by several known standards such as **LiveWire** and JDBC (Java Database Connectivity). Furthermore, Web pages can include executable code such as applets (e.g., java programs) that can be downloaded from a server and executed on the browser or on a JVM (java virtual machine) of the system on top of which the browser is built. Other information can be provided by servlets (e.g., java programs) running on the server and pushing changes in the connected browser. The applets and servlets can include CGI (common gateway interface) functions which allow a Web server and applications to communicate with each other. In addition, other information accessing methods include scripts which are predetermined program languages that are interpreted and executed on the browser. This includes, for example, javascripts and DHTML (Dynamic HTML) languages. Plug-ins are programs outside the browser that can be downloaded by the browser and automatically recognized by the browser to run native on the local device and be executed on arguments that are subsequently provided (via download) by the browser. CGI scripts are server side scripts that implement the business logic and produce as output of them running the next presentation material. Applets and plugins can communicate via RMI (remote method invocation), socket connections, RPC (remote procedure call), etc. In addition, complex transcoding schemes, XML (Extensible Markup Language) extensions and scripting languages are used for specific information or services or to simplify the interaction.

As explained above, the purpose of the Internet Web browser and IVR is to access information. The following example describes a typical scenario in connection with a banking application to demonstrate that the paradigm used for accessing the information via IVR with a telephone and via the Internet using a PC and Web browser is similar. For instance, the typical banking ATM transaction allows a customer to perform money transfers between savings, checking and credit card accounts, check account balances using IVR over the telephone. These transactions can also be performed using a PC with Internet access and a Web browser. In general, using the PC, the customer can obtain information in a form of a text menus. In the case of the telephone, the information is presented via audio menus. The mouse clicks on the PC application are transformed to pushing telephone buttons or spoken commands. More specifically, a typical home banking IVR application begins with a welcome message. Similarly,

the Internet home page of the Bank may display a picture and welcome text and allow the user to choose from a list of services, for example:

a. instant account information;

b. transfer and money payment;

c. fund information;

d. check information;

e. stock quotes; and

f. help.

With the IVR application, the above menu can be played to the user over the telephone, whereby the menu messages are followed by the number or button the user should press to select the desired option:

a. "for instant account information, press one;"

b. "for transfer and money payment, press two;"

c. "for fund information, press three;"

d. "for check information, press four;"

e. "for stock quotes, press five;"

f. "for help, press seven;"

The IVR system may implement speech recognition in lieu of, or in addition to, DTMF keys. Let's assume that user wants to get the credit card related information. To obtain this information via the Internet based application, the user would click on a particular hypertext link in a menu to display the next page. In the telephone application, the user would press the appropriate telephone key to transmit a corresponding DTMF signal. Then, the next menu that is played back may be:

a. "for available credit, press one";

b. "for outstanding balance, press two";

c. "if your account is linked to the checking account, you can pay your credit card balance, press three."

Again, the user can make a desired selection by pressing the appropriate key.

To continue, the user may be prompted to provide identification information. For this purpose, the Internet application may display, for example, a menu with an empty field for the user's account number and another for the users social security number. After the information is

filled in it is posted to the server, processed, the replay is formatted and sent back to the user. Over the telephone the scenario is the same. The IVR system may playback (over the telephone) an audio prompt requesting the user to enter his/her account number (via DTMF or speech), and the information is received from the user by processing the DTMF signaling or recognizing the speech. The user may then be prompted to input his/her SSN and the reply is processed in a similar way. When the processing is complete, the information is sent to a server, wherein the account information is accessed, formatted to audio replay, and then played back to the user over the telephone.

As demonstrated above, IVRs use the same paradigm for information access as Web browsers and fulfill the same functionality. Indeed, beyond their interface and modality differences, IVR systems and Web browsers are currently designed and developed as fundamentally different systems. In the near future, however, banks and large corporations will be moving their publicly accessible information sources to the Internet while keeping the old IVRs. Unfortunately, this would require these institutions to maintain separate systems for the same type of information and transaction services. It would be beneficial for banks and corporations to be able to provide information and services via IVR over the Internet using the existing infrastructure. In view of this, a universal system and method that would allow a user to access information and perform transactions over the Internet using IVR and conventional browsers is desired.

SUMMARY OF THE INVENTION

The present invention is directed to a system and method for unifying the access to applications to a standard protocol, irrespective of the mode of access. In particular, the present invention provides a universal method and system for accessing information and performing transactions utilizing, for example, a standard networking protocol based on TCP/IP (such as HTTP (Hypterext Transfer protocol) or WAP (wireless application protocol) and architecture to access information from, e.g., a HTTP server over the Internet such that a pure GUI (graphical user interface) modality and pure speech interface modality can be used individually (or in combination) to access the same bank of transaction and information services without requiring modification of the current infrastructure of currently available networks.

In one embodiment of the present invention, a conversational browser is provided that translates commands over the telephone to an HTTP protocol. The introduction of the conversational browser allows us to unify Internet and Telephone (IVR) and thereby decrease the cost, enlarge the coverage and flexibility of such applications. In particular, for IVR applications, the conversational browser or (telephony browser) can interpret DTMF signaling and/or spoken commands from a user, generate HTTP requests to access information from the appropriate HTTP server, and then interpret HTML-based information and present it to the user via audio messages. The conversational browser can also decode compressed audio which is received from the HTTP server in the HTTP protocol, and play it reconstructed to the user. Conversely, it can capture the audio and transmit it (compressed or not) to the server for distributed recognition and processing. When the audio is captured locally and shipped to the server, this can be done with a plug-in (native implementation) or for example with a java applet or java program using audio and multimedia API to capture the user's input.

For the new proposed IVR architecture and conversational browser, the content pages are on the same HTTP server that are accessed by conventional modes such as GUI browsers, and use the same information access methods, sharing the legacy database access SW layer, etc. In other words, an IVR is now a special case of a HTTP server with a conversational browser. Similar to the conventional GUI browser and PC, the conversational browser, the information and queries will be sent over the switched packed network using the same protocol (HTTP).

The present invention will allow an application designer to set up the application using one framework, irrespective of the mode of access, whether it is through telephone or a WWW browser. All interactions between the application and the client are standardized to the HTTP protocol, with information presented through html and its extensions, as appropriate. The application on the WWW server has access to the type of client that is accessing the application (telephone, PC browser or other networked consumer device) and the information that is presented to the client can be structured appropriately. The application still needs to only support one standard protocol for client access. In addition, the application and content is presented in a uniformed framework which is easy to design, maintain and modify.

In another aspect of the present invention, a conversational browser interprets conversational mark-up language (CML) which follows the XML specifications. CML allows new experienced application developers to rapidly develop conversational dialogs. In another

aspect, CML may follow other declarative syntax or method. Pursuing further the analogy with HTML and the World Wide Web, CML and conversational browser provide a simple and systematic way to build a conversational user interface around legacy enterprise applications and legacy databases.

CML files/documents can be accessed from HTTP server using standard networking protocols. The CML pages describe the conversational UI to be presented to the user via the conversational browser. Preferably, CML pages are defined by tags which are based on the XML application. The primary elements are <page>, <body><menu>, and <form>. Pages group other CML elements, and serve as the top-level element for a CML document (as required by XML). Bodies specify output to be spoken by the browser. Menus present the user with a list of choices, and associate with each choice a URL identifying a CML element to visit if the user selects that choice. Forms allow the user to provide one or more pieces of information, where the content of each piece of information is described by a grammar. The form element also specifies a URL to visit when the user has completed the form.

In another aspect, conversational mark-up language rules can be added by a content provider to an HTML file (or used in place of HTML) to take full advantage of the conversational browser.

In yet another aspect, a conversational transcoder transforms presentation material form one modality to a conversational modality (typically, speech only and/or speech and GUI). This involves functional transformation to transform one page of GUI to a page of CUI (conversational user interface), as well as logical transcoding to transform business logic of an application, transaction or site into an acceptable dialog. Conversational transcoding can convert HTML files into CML files that are interpreted by the conversational browser. The transcoder may be a proprietary application of the server, browser or content provider.

In another aspect, HTML/GUI based structure skeletons can be used to capture the dialog logic or business logic of a GUI site. This information can be used to map the sit, logic or application. After appropriate organization of the dialog flow, each element can undergo functional transcoding into a speech only content or a multi-modal (synchronized GUI and speech interface) page.

In another aspect, a conversational proxy is provided to modify and/or prepare the content description of the application, logic or site to the capabilities of , e.g., the device,

browser and/or engines, preferences of the user or application, load on the servers, traffic on the network, location of the conversational arguments (data files). For instance, the conversational proxy can directly convert proprietary formats such as screen maps of corporate software.

These and other aspects, features and advantages of the present invention will be described and become apparent from the following detailed description of preferred embodiments, which is to be read in connection with the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a diagram of a conversational browsing system according to a preferred embodiment of the present invention;

Fig. 2 is a block diagram of a system for accessing information implementing a conversational browsing system according to an embodiment of the present invention;

Fig. 3 is a block diagram of a system for accessing information implementing a conversational browsing system according to another embodiment of the present invention;

Fig. 4a is a block diagram illustrating a distributed system for accessing information implementing a conversational browsing system according to an embodiment of the present invention;

Fig. 4b is a block diagram illustrating a distributed system for accessing information implementing a conversational browsing system according to another embodiment of the present invention;

Fig. 5 is a block diagram of a conversational information accessing system using conversational markup language according to an embodiment of the present invention;

Fig. 6 is a general diagram of a distributed conversational system using conversational markup language accordance to an embodiment of the present invention;

Fig. 7 is a diagram of an exemplary distributed conversational system using conversational markup language according to an embodiment of the present invention;

Fig. 8 is a diagram of another exemplary distributed conversational system using conversational markup language according to another embodiment of the present invention;

Fig. 9 is a diagram of a yet another distributed conversational information accessing system using conversational markup language according to an embodiment of the present invention; and

Fig. 10 is a diagram of another exemplary distributed conversational information accessing system using conversational markup language according to an embodiment of the present invention.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

The present invention is directed to a conversational browsing system and CML (conversational markup language) for building a conversational browsing system using a set of interlinked CML pages. This conversational system is conceptually analogous to building conventional GUI browser applications using a set of interlinked pages written using HTML (hypertext markup language). Moreover, just as HTML provides a set of mechanisms for translating GUI actions into application actions such as visiting other pages or communicating with a server, the conversational browser and CML are used for translating spoken inputs into similar application actions. A CML page describes the conversational UI to be interpreted and presented to the user via the conversational browser. Preferably, CML pages are defined by tags which are based on the current XML (extensible markup language) application (as described in detail below).

It is to be understood that the conversational systems and methods described herein may be implemented in various forms of hardware, software, firmware, special purpose processors, or a combination thereof. In particular, the conversational browser is preferably implemented as an application comprising program instructions that are tangibly embodied on a program storage device (e.g., magnetic floppy disk, RAM, ROM, CD ROM and/or Flash memory) and executable by any device or machine comprising suitable architecture such as personal computers and pervasive computing devices such as PDAs and smart phones.

It is to be further understood that, because some of the constituent components of the conversational browser and other system components depicted in the accompanying Figures are preferably implemented in software, the actual connections between such components may differ depending upon the manner in which the present invention is programmed. Given the teachings herein, one of ordinary skill in the related art will be able to contemplate these and similar implementations or configurations of the present invention.

Conversational Browser Architecture

Referring now to Fig. 1, a block diagram illustrates a conversational browser system according to a preferred embodiment of the present invention. In general, a conversational browsing system 10 allows a user to access legacy information services and transactions through voice input (either uniquely or in conjunction with other modalities such as DTMF, keyboard, mouse, etc) using a standard networking protocol such as HTTP. In addition, it is to be understood that the HTTP protocol is a preferred embodiment of the present invention but other similar protocols can be used advantageously. For example, this can be deployed on top of any protocol that such as TCP/IP, WAP (Wireless Application Protocol), GSM, VoIP, etc., or any other protocol that supports IP (and therefore provide TCP/IP or similar features. Even more generally, if TCP/IP is not available we can implement another protocol offering features similar to TCP/IP or at least performing network and transport functions (the present invention is not dependent on the transport and network layer)..

In Fig. 1, a conversational browsing system 10 according to one embodiment of the present invention comprises a conversational browser 11 (conversational application) which executes on top of a CVM (conversational virtual machine) system 15. The conversational browser 11 comprises a transcoder module 11 which, in general, transcodes conventional (legacy) structured document formats such as HTML or DB2 into an intermediate document, or CML (conversational markup language) document in accordance with prespecified transcoding rules (as discussed below). A CML document describes the conversational UI of the legacy information format to be presented to the user. More specifically, a CML document comprises meta- information which is processed by a CML parser/ processor 14 to present, for example, HTML-based information to a user as synthesized audio messages. It is to be understood that various embodiments for a CML document are contemplated for implementation with the present invention. In a preferred embodiment described in detail below, a CML document is defined by tags which are based on XML (extensible markup language). It is to be understood, however, that any declarative method for implementing CML may be employed. XML is currently preferred because of its simplicity, power and current popularity.

The conversational browser 11 further comprises a command/ request processor 12 (a command and control interface) which converts user command (multi- modal) inputs such as speech commands, DTMF signals, and keyboard input into navigation requests such as HTTP

requests. It is to be understood that in a pure speech conversational browser, the only input is speech. However, the conversational browser 11 can be configured for multi-modal input.

When certain conversational functions or services are needed, the conversational browser 11 will make API calls to the CVM 15 requesting such services (as described below). For instance, when interpreting a CML document (via the CML parser/processor 14), the conversational browser 11 may hook to a TTS (text-to-speech syntheses) engine 26 (via the CVM 15) to provide synthesized speech output to the user. In addition, when speech commands or natural language queries (e.g., navigation requests) are input, the conversational browser 11 may hook to a speech recognition engine 24 and NLU (natural language understanding) engine 25 to process such input commands, thereby allowing the command/request processor to generate, e.g., the appropriate HTTP requests. The CVM system 15 is a shell that can run on top of any conventional OS (operating system) or RTOS (real-time operating system). A detailed discussion of the architecture and operation of the CVM system 15 is provided in the patent application IBM Docket No. YO999-111P, filed concurrently herewith, entitled "Conversational Computing Via Conversational Virtual Machine," which is commonly assigned and fully incorporated herein by reference.

Briefly, as shown in Fig. 1, the CVM system 15 comprises a conversational API layer 16 through which the conversational browser 11 can "talk" to a CVM kernel layer 17 to access (via system calls) certain conversational services and behaviors including the conversational engines 23. The CVM kernel 17 is responsible for allocating conversational resources such as engines and arguments (either local and/or distributed) and managing and controlling the dialog and context across applications and devices (locally and/or distributed) on the basis of their registered conversational capabilities and requirements to provide a universal and coordinated CUI (conversational user interface). The CVM system 15 performs conversational services and functions by implementing calls to local conversational engines 23, e.g., a speech recognition engine 24, a NLU (natural language understanding) engine 25 a TTS (text-to-speech) engine 26 (as well as other engines such as an NLG (natural language generation) engine, speaker recognition engine) through a conversational engine API layer 18 (such as SAPI, SRAPI, JSAPI, SVAPI or extensions of such engine APIs). In addition, engine calls can be made to remote speech engines in distributed topologies. In addition, calls to an audio subsystem 33 (providing audio capture, compression, decompression and reconstruction) and any other multi-modal I/O

resources (such as a keyboard 28, Mouse 29, touch screen, microphone/speaker 31 and DTMF 32) are performed via a conventional drivers/API layer 22.

The CVM system 15 further comprises a communication stack 19 for providing network communication via conventional protocols 20 such as TCP/IP, HTTP, WAP, etc. The communication stack 19 further comprises conversational protocols 21(or distributed conversational protocols) which are utilized for distributed applications. As described in the above-incorporated IBM Docket No. YO999-111P, the conversational protocols (or methods) 21 include protocols for (1) discovering network devices and applications that are "conversationally aware" (i.e., that speak conversational protocols); (2) registering conversational capabilities (resources) such as conversational engines and arguments between network devices; (3) negotiating network configurations (such as master/slave, peer-to-peer) based on registered conversational capabilities; (4) exchanging information to coordinate a conversation between network connected devices such as information regarding the state, context and history of a dialog, conversational arguments, applets, ActiveX components, procedural objects, and other executable code; and (5) speech coding protocols to transmit and receive compressed speech (waveforms or features). These conversational protocols 21, as well as their role for providing conversational coordination between networked devices is also disclosed in the patent application IBM Docket No. YO999-113P, filed concurrently herewith, entitled "System and Method For Providing Network Coordinated Conversational Services," which is commonly assigned and incorporated herein by reference.

It is to be understood that the engines 23 and I/O resources 27, conventional drivers/APIs 22 and audio subsystem 33 illustrated in Fig. 1 are components that are part of the underlying device, machine or platform on which the conversational browser 11 and CVM system 15 are executed. It is to be further understood that the conversational browser 11 and CVM system 15 can be provided as separate systems or, alternatively, the conversational browser 11 can be implemented as a stand-alone application carrying its own CVM system 15 (in which case the browser and CVM platform would be the same, i.e., indistinguishable entities). In addition, in the absence of a CVM system 15 as specifically described above, it is to be understood that the conversational browser 11implements functions that are similar to that discussed above for the CVM system 15 (e.g. the conversational browser would make API calls to appropriate engines locally and/or distributed). Indeed, the API, services, features, behaviors, access to engine and

communication mechanisms can all be built directly into, and made part of, thc conversational browser application 11 as part of the features and services provided by the browser (this would be analogous as having a JVM (java virtual machine) provided by the underlying operating system to a Netscape browser or having the browser implement its own JVM). In addition, the conversational browser can take advantage of similar existing services provided by the underlying operating system and/or applications running in parallel of the browser.

It is to be understood that the transcoder 13 may be part of the conversational browser11 as illustrated in Fig. 1. Alternatively, as described below, the transcoder 13 may be located in a network and running, for example, on another machine, conversational proxy server, on a server of the content provider (server-side), or distributed on various machines in the network. In addition, the transcoder may be a proprietary application of the proxy server, browser or content provider. The transcoder should also be able to directly convert other proprietary formats like screen map of corporate software. The transcoder 13 comprises a functional transcoding module 13a for converting , e.g., an HTML document (or other format page) into a CML document. Various rules for implementing this conversion will be explained below. In addition, the transcoder 13 comprises a logical transcoding module 13b which accesses and interprets the logic information behind the application to build a meaningful dialog. This often requires in depth information about the application and /or business logic. It can be conveyed with extra XML or meta information that can define some of these intent. Another method which may be employed for performing logical transcoding is described in described in the patent application IBM Docket No. Y099-114, filed concurrently herewith, entitled "Structure Skeletons For Efficient Voice Navigation Through Generic Hierarchical Objects", which is commonly assigned and fully incorporated herein by reference. This application describes how conversational structure skeletons may be used to encode and transmit logical transcoding information.

Preferably, the transcoder 35 is a generic, general purpose converter, which is capable of transforming any transaction form or HTML page to a CML page. With respect to straightforward dialog like ATM dialog or IVR dialogs, wherein HTML dialog pages with form filling (HTML or CGI) or ATM dialog (and other 3270 type of screen dialogs) have limited information or provide limited user choices, the transcoding may be readily formulated in terms of CML pages. Other complex HTML documents can instead be tagged as CML in the first

instance on the server side (as explained below) without requiring a transcoder for the conversion. Indeed, as explained below, in a preferred embodiment, all the content provided to the conversational browser is directly written in CML without the need for a transcoder or transcoding proxy, as opposed to generic HTML pages, for instance, being transcoded on the fly by the browser or a server- side transcoder. Otherwise, basic composition and design rules can be imposed to simplify the conversion from HTML to CML. For instance, news and e-mail/voice-mail can be converted by filtering the incoming text with conventional TTS filters before tagging it with intonation and TTS markup information à la JSAPI.. In practice, these conversion rules can be either proprietary based on the service provider or the object of a standard. These rules may comprise the following:

1. If the HTML page comprises images, the transcoder will discard the images and generate the necessary meta information for presenting the image name and comment added (if any) in the image tag for audibly presenting such information.

2. If the HTML page comprises scripts and/or applets that have no component which may be presented audibly, the transcoder can generate meta- information for running special warning messages. The warning has several purposes. For instance, the warning message can inform the user that the HTML page contains a script or applet. Indeed in the absence of any speech output provided by the script or applet, the user may never know about the script or applet (as opposed to a GUI browser where an applet animation is immediately visible). In addition, the warning message serves to inform the user that he/she may be missing some information/content contained in the applet that is not capable into being transcoded into meaningful CUI (conversational user interface) features (this holds true for situations where the applet is written in conventional GUI languages like Java, instead of being written with the conversational foundation classes as part of CVM libraries (such as described in the above incorporated IBM Docket No. YO999-111P) to incorporate a CUI component.

In addition, meta-information ca be used to synchronize the GUI (visual) side of a multi- modal browser with the speech side, such as described in provisional application U.S. Serial No. 60/128,081, filed April 7, 1999, entitled "Multi-Modal Shell", which is commonly assigned and incorporated herein by reference. According to that invention, even procedural components (like applets and ActiveX components) can be transcoded into a GUI only or speech only (and variation of this) based on the capabilities of the device/browser. This automatically solves the

problem of the transcoding that was inherently coded by the application developer or the author of the components (foundation classes) the application used to build his procedure.

3. With HTML pages comprising frames, the transcoder can generate meta information for presenting the frames either by their name first or an index number. For instance, the audio playback may be "there are 4 frames on this page: frame 1 presents ....." All the frame links and functions created by HTML and CML can be activated as soon as the entire page is downloaded. This will allow the frames to be used as a menu. Alternatively, a rule can be applied such that only the main frame is presented, while the remaining frames are active but not read.

4. HTML pages can be filtered using TTS filters to produce JSAPI markup (or similar variations).

5. Any hyperlinks in an HTML page can be converted into baseforms, and FSG (finite state grammar). More specifically, hyperlinks and other commands and input that can be followed by the user (by voice commands) may be added to an active list of commands of the browser. This involves adding the vocabulary, the pronunciation rules of these words when needed (i.e the baseforms) and modifying appropriately the grammar or the NLU framework / context to support these new commands.

6. Text may be synthesized with different voices for different fonts. However, the conversational browser can elect not to change the voice.

7. Hyperlinks can be presented with an audio icon (sound, or a different voice, or a background music).

8. Each hyperlink can also be presented as a hidden menu, which may carry over to the following pages, if indicated, for example, with a <!--CML tag-->.

9. Pull down menus and constrained menus can be automatically converted to dialog with appropriate vocabulary and commands.

10. Free fields can activate NLU or dictation capabilities. In general, for general purpose dictation, the speech recognition engine used can be the local default engine of the conversational browser. Alternatively, for distributed systems (as discussed below), the content provider can provide a suggested URL of a speech recognition engine if the browser does not have the necessary capabilities or if the remote speech recognition engine is specifically tuned for the NLU or dictation application. NLU dialogs with NLU form filling can be performed by either passing the appropriate URL of the NLU engine as a CML comment tag in the HTML

page or by directly providing CML pages instead of HTML pages (as discussed in further detail below).

11. Multiple form based screen based transaction GUI dialogs can be directly converted into menus of sequential dialogs if each menu is a stand-alone component (e.g., performing a complete task such as obtaining an address or ordering an item)

The conversational browser 11 may also include a "barge-in" feature that allows the user to utter an attention word or otherwise interrupt the output flow of a TTS presentation before giving a new command or input (as explained in more detail below). This may be implemented with buttons to press when a person speaks, whereby the button is pressed to talk or pressed to start and/or press to stop.

Preferably, the CML documents generated by the transcoder (or the CML documents directly provided by the content provider as discussed below) comprise any combination of the following exemplary features and necessary tags to indicate:

1. Active links which are either spoken to, or hidden from, the user;

2. Links to conversational arguments (e.g., data files of vocabularies, grammars and baseforms) that may be used to decode possible input given by the user for the page;

3. For distributed systems, links to speech recognition URLs, speaker recognition URLs, TTS URLs and/or NLU engine URLs, etc., which may be used to decode user inputs and/or provide location(s) where to send a given page so as to generate the audio output. A local host may be used to impose a local (default) engine. Therefore, if no URL is given, the browser default engine is used. This may be set based on the browser, device, user preference or the service provider (ISP or ASP) preference. This amounts to using any relevant meta-information on the browser or on the server to set the browser default location;

4. URLs to audio files to play back to the user (e.g. IVR prompts);

5. Confirmation request tags that impose user confirmation to some input before following a given link;

6. DTMF code associated to a given link, as well as DTMF grammar for complex key entry;

7. TTS markup (à la JSAPI) describing how text should be played back to a user;

8. Scripting language and procedural code (such as Java, Javascript, ActiveX, etc.) so as to run conversational applications that direct their own dialog. Conversational programming

languages and foundation classes (i.e. CVM libraries) can be used when the browser is built on top of CVM (as discussed above). Otherwise, the code and services can be provided by the browser, the underlying operating system or other applications or services that are available in parallel with the browser.

9. Conversational capability tags, describing the browser minimum requirement for a resource (e.g. a script, an image, etc). For example, non-audio enabled scripts should not be downloaded or run on a speech only device, except if such scripts perform a meaningful task. In addition, images should not be downloaded on a speech only browser (although they are relevant on multimodal and GUI only browsers). When the presentation material is converted according the capabilities of the browser (i.e., the underlying engines), the entity that performs the conversion is a conversational proxy;

10. URLs of plug-ins to use/download for some specific processing (e.g. an encryption plug-in for speaker recognition stage or compression routine and front-end for a specific engine); and/or

11. Meta-information tag(s) for bookmarking and subsequent return to the page (e.g. set of keywords with their corresponding baseforms, grammars, etc.) which, when invoked, will propose the page on the list of the possible link to follow. This is typically a feature related to the categorization / meta-information service as provided by a CVM (as discussed in the above-incorporated IBM Docket No. YO999-111P).

It is to be appreciated that a specific comment **<!--CML proxy url=... -->** tag can be incorporated in an HTML page to specify a custom conversational proxy or transcoding to perform a desired transcoding function (conversational proxy, functional transcoding or logical transcoding). Alternatively, the conversion can be performed by a default transcoder (which is part of a browser, a server or a distributed proxy). CML pages may also be customized by the (proxy) transcoder based on the capabilities of the conversational browser. This functionality can be performed differently than by using the URL tag as a proxy (which implies that the browser connects to a proxy server and sends its HTTP requests through the proxy). If a CVM is present, an underlying

handshake can occur to exchange meta-information about the conversational capabilities of the engines available to CVM for the browser, and then based on that the proxy can automatically

tailor the pages. Also an alternative consists of having in such case the browser forwarded to a different page suitable for the engine capabilities.

In addition, **<!--CML tag...-->** can introduce direct CML tags into an HTML or XML page, which will be directly processed by the conversational browser or conversational proxy. Each of these CML features will be discussed in more detail below.

While CVM is multi-modal, for specific browser/client capabilities (e.g. only a IVR), only the relevant functions are tagged. In the preferred embodiment, the browser typically ignores unknown tag instead of forcing a consistency of the ML. Furthermore, for general purpose systems, the browser should register its capabilities to the proxy or server, using a cookie mechanism or filling a hidden form when connecting for on the first page of that content provider or to the server gateway/portal. The cookie can carry the needed information to identify or authenticate a user. It can also carry all the context information about the current dialog status of the browser. This can also include the rest of the local context, history and/or preference information to be sent or shared with the server upon connection to it. The capability to send dialog information, context and other meta-information to a server via a cookie or other registration protocol enables the server to immediately participate in the conversation. Other registration mechanism may be implemented other than the cookie mechanism.

Conversational mark-up language rules can be added by the content provider to an HTML file or substitute to an HTML file to take full advantage of the conversational browser. This language should present tags to indicate the feature described in the previous section (as described below).

Telephony Applications Using Conversational Browser

Referring now to Fig. 2, a block diagram illustrates a conversational system for providing telephony services according to an embodiment of the present invention. The system 100 comprises an IVR 101 (or call center) which is accessible by a client device 102 over a telephone network 104 (e.g., PSTN (public switched telephone network)). The IVR 101 is connected to one or more content servers 105 over network 104 . In this embodiment, it is assumed that the IVR 101 runs on a machine on the premises of a telephone company (telco). The IVR 101 may be, for example, associated with a service provider of a banking application to provide ATM telephony services. The client device 10 may be, for example, a conventional telephone, a

mobile telephone, or a PC having a telephony card, to establish communication over the telephone network 103 to the IVR 101. The network 104 may be the Internet, an intranet, a LAN, *ad hoc* networks, etc. The content servers 105 may be HTTP servers, for example, which comprise the content pages (HTML, CML, etc) that are associated with the IVR 101 ATM services.

The IVR 101 comprises a conversational browser 106 which is illustrated as a stand alone application running a CVM 106a. It is to be understood that the conversational browser 106 and CVM 106a of Fig. 2 are similar in architecture and functionality as the conversational browser 11 and CVM 15 described above with respect to Fig. 1 (whereby the CVM can be included within the browser or in an underlying operating system. The IVR system100 further comprises an audio system 107 providing audio functions such as audio capture, acoustic front end (acoustic processing) and an audio subsystem (e.g., soundcard). The IVR 101 further comprises a DTMF processor 108 and conversational engines 109 (e.g., speech , NLU, TTS engines, etc.).

When the client 102 is connected to the IVR 101, input commands such as speech commands and DTMF signals are received by the acoustic front end of the audio system 107 and sent to the appropriate conversational engine 109 (e.g., speech engine) and DTMF processor 108, respectively. The processed speech/DTMF commands are then passed as calls to appropriate functions in the conversational browser 106. The conversational browser 106, in turn, generates appropriate requests (e..g, HTTP requests) to obtain desired information from one or more of the content servers 105 over the network 104. The content server 105 retrieves the appropriate information (or completes the specified action) and then transmits results, messages and/or menu options (using TCP/IP or similar protocols discussed above) in either HTML (or other similar formats), CML, compressed audio, or a combination thereof (depending on the manner in which the information is structured on the content server 105). As discussed above, for HTML (and other structure documents), the conversational browser will transcode the information (via a transcoder 13 discussed above) to generate a CML document for processing (via the CML parser/processor 14 discussed above), whereas CML documents from the servers 105 can be directly processed (without transcoding). The conversational browser 106 can hook to necessary conversational engines (via the CVM 106a) to establish a dialog with the user. As discussed below, the conversational engines may be distributed over the network 104.

The conversational browser can also decode compressed audio which transmitted from the content server 105 using a standard network protocol (such as HTTP), and then play back the audio to the user as telephone audio. (thereby allowing, e.g., IP telephony via the conversational browser 106). Conversely, audio input can be captured via the audio capture portion of the audio system 107 and transmitted (via the conversational browser 106) in compressed or non-compressed format to a server over the network 104 for distributed recognition and processing (such as described in the above incorporated patent application IBM Docket No. YO999-113P). As indicted above, the transmission of the compressed speech is performed by the communication stack of the CVM 106a (such as discussed in YO999-111P, as well as YO999-113P). In the absence of a CVM, other communication engines will perform the handshake , communication and coding/decoding. It is to be understood that such communication is not limited to HTTP or VoIP, it can involve any TCP/IP and related protocols. For example, WAP, Bluetooth, Hopping communication protocols may be employed with the present invention. When the audio is captured locally and shipped to a remote server, this can be performed using a plug-in (native implementation) or for example with a java applet or java program using audio and multimedia API to capture the user's input. In addition, the conversational communication(coding) protocols may be used for transmitting compressed speech using, for example, socket connections, RPC, RMI, etc. (such as described in the above incorporated patent applications IBM Docket Nos. YO999-111P and YO999-113P).

It is to be appreciated that the system described in Fig. 2 is only one possible embodiment for providing IVR services. In another embodiment, for instance, the IVR can behave solely as "soundcard" of the system in the sense that it provides the audio capture functionality to capture input speech from a user and then transmit the captured speech to a web server having a conversational browser (similarly it receives audio output from the server and transmits it to the IVR for play back to the user). In this case, the IVR acts as a soundcard or an audio subsystem for the conversational browser. Alternatively, the IVR may bc providing the conversational engines (speech recognition engine, TTS engine, speaker recognition engine etc...) as well as the audio capture and all the system management, call balancing, load balancing functions etc. and only use the conversation browser as a driver of the dialog. Alternatively, also the IVR can only performs the audio capture and the

load balancing/system management while having the conversational engines running on other servers and the conversational browser running also on other servers. Eventually there is a possibility that the IVR or call center is implemented with a farm of networks small workstations (typically NT machines) each with a telephony card that will own the audio capture, the conversational engines and the conversational browser associated to one or a very limited of active ports (i.e active calls).

In the embodiment illustrated in Fig. 2, it is assumed that all the functions of playing back prompts, synthesizing speech, generating dialog, understanding DTMF input or speech input are performed at the level of the IVR 101. It is to be understood that in the absence of a CVM 106a, it is the responsibility of the system integrator or IVR vendor to provide a concrete implementation a system providing abstract layers similar to those provided by the CVM (as discussed above). The actual implementation of the system will determine the actual architecture of the IVR 101, but such architecture would be independent with respect to the functionality and operation of the conversational browser 106.

Advantageously, the conversational browser 106 effectively changes the IVR 101 from what is known in the art as the "all-time" connection to a packet switched network on the service side. In addition, with the universal IVR architecture employing a conversational browser, it is to be appreciated that the content pages are on the same HTTP servers and use the same information access methods, sharing the legacy database access SW layer, etc. as web servers for information access. In other words, an IVR can now be considered a special case of a HTTP server with a conversational browser. Depending on the device, server and content/service provider capabilities, the conversational browser can be located on the client, on the server, shared between the client and the server or distributed across the network.

Referring now to Fig. 3, a block diagram illustrates a conversational system for proving telephony services according to another exemplary embodiment of the present invention. In this example, the conversational browser 106 is located entirely on the client device. All the processing performed by the IVR (as explained above with reference to Fig. 2) is now performed locally on the client device and only the fetching of HTML documents, for example, is performed through HTTP on the content servers 105. The content servers 105 can be HTTP servers having applications executed thereon which are associated with IVR applications. Alternatively, the system of Fig. 3 can simply be used for browsing the WWW.

For instance, as illustrated in Fig. 3, the client device can be a PC 60 (personal computer) having a conversational browser 106 (and CVM 106a) and conventional telephony platform 61 (comprising, for example, a modem, soundboard, speakerphone and voicemail system) for providing telephony services via the PC 60. The PC 60 can be used to establish a dial-up connection via modem (through the conversational browser 106) over the telephone lines 102 to the telco or any other remote access server (RAS) 70 (e.g., Internet Service Provider (ISP)) to access content servers 105 over the network 40 (e.g., Internet, intranet). The PC 60 can also access local/private resources or the Internet through remote connections via IP connection using a dedicated line L. The PC 60 with the browser 106 can be used to provide conversational browsing of, e.g., the WWW. As explained above, the conversational browser 106 performs, e.g., the recognition of voice commands and the conversion of the voice messages from HTTP server coming compressed over Internet. The conversational browser 106 can access any speech enabled local applications 62 (calendar, etc) running on the PC 60 or local client.

The client device may be a mobile phone 80 (e.g., a GSM terminal) having suitable architecture and CPU power to run the conversational browser 106 locally. The mobile telephone 80 can download DSP code (or any other type of CPU programming language) to interpret the Internet protocols for speech (Internet telephone) (or other protocols such as WAP) as well as compression algorithms for speech purposes and any other application or speech processing programs written, for example, in some derivation of Java. It is to be understood that the DSP code may be native and preloaded on the phone (especially if an instantiation of CVM 106a is present on the device). Furthermore, even higher level mobile phone with screen can have higher level of locally running browser with text capabilities or even display capabilities. In addition, the new telephones can be equipped with higher sampling rate A/D converters. This would enable to take advantage of the cepstrum coefficient compression to send the speech more effectively over the communication lines and increase the speech recognition quality (for distributed Speech recognition such as explained in the above-incorporated IBM Docket No. YO999-113P).

It is to be appreciated that the functions of the conversational browser 106 may be partially located on a local client device and partially located on a remote server. For example, the acoustic features can be computed by the client device (e.g., via the front end module 107, Fig. 2)), compressed and then transmitted to the server at low bit rate for processing. In

addition, a portion of the browser processing can be done locally, e.g., playing back regular text and getting simple input, and some processing can be performed on the server side, e.g., obtaining street addresses or complicated input or ascii input to forms. It is to be further appreciated that the browser functions can be distributed across a client device, one or more servers, and a network of resources. For instance, the browser can be located on the server, the client or shared between the two, with conversational subsystems (e.g., speech recognition engine, NLU engine, TTS) being located at other areas on the network. It is to be appreciated that even the browsing functions can be shared between networked devices (servers, computers, PDAs etc). Again, a CVM is a key element to "manage" such cases through communication protocols similar to those described in the incorporated patent applications IBM Docket Nos. YO999-111P and YO999-113P. In addition, for domain specific applications, a content provider can directly provide its own TTS, NLU or speech recognition engine.

Advantageously, the deployment of the present invention can afford the preservation of legacy applications that provide information pages in HTML format, by allowing such information to be presented for pure conversational interaction (besides the capability to display the pages or portions of the pages with multi-modal devices). Each of these concepts will now be illustrated with reference to the illustrative embodiments of Figs. 4a and 4b.

Referring now to Fig. 4a, a block diagram illustrates a distributed system wherein the resources for providing conversational browsing and IVR service are shared across one or more networks. It is to be understood that the dotted circles in Fig. 4a represent, for example, different locations of a given network or individual networks (or subnetworks) on an intranet or the Internet, each comprising corresponding servers, hosts, which interact with one another via standard network protocols (e.g., TCP/IP, HTTP). A PC 302 having a GUI terminal 300 and conversational browser 302a can be used to provide a combination of GUI and conversational browsing (it is to be understood that the device 302 can be any client device (other than a PC) such as a phone which is connected the distributed network via a PSTN , or a bluetooth device connected through bluetooth to a TCP/IP network, a setopbox connected through ISDN or cable or satellite link to the distributed network). Alternatively, if the PC 302 can access a remote server 303 having a conversational browser a via IP connection. Likewise a user can dial into an IVR server/PC 302 via telephone 301 (e.g., regular, cellular, or smart phone) obtain perform conversational browsing or IVR via the local browser 302a or via the remote server 303 and

conversational browser 303a. Alternatively, as explained above, the phone 301, PC/IVR server 302 and remote server 303 can share the functions of the conversational browser depending on their respective architecture and processing power capabilities. In addition, the conversational browser 303a can be shared across one or more servers 304 distributed across a given network or subnetwork. Alternatively, each of the remote server 304 can support one or more specific functions of the browser 304a. Depending on the respective conversational browser function, the PC/IVR server 302 or remote severs 303, 304, each may hook (as necessary) via TCP/IP, HTTP, API (for CVM instantiations using the communication protocols discussed above) to one or more conversational subsystems 309 distributed over the network. In addition, any required conversational arguments 308 can be accessed for use by the conversational browsers 302a, 303a and/or 304a or conversational subsystems 309 over the network.

The exemplary system further includes a conversational proxy server 305 with transcoder 305a, which may be accessed over the network to convert HTML documents and other applications to CML in circumstances where the conversational browsers 302a, 303a, 304a cannot perform such conversion due to inadequate processing power of the device on which the browser runs. In addition, the conversational proxy and transcoder 305a may be associated with a content service provider to perform conversion of HTML documents in instances when the conversational browsers 302a, 303a and/or 304a lack specific proprietary information about the application to perform such conversion (which is known only by the content provider). As shown in Fig. 4a, the HTML documents are generated via web server application of the content provider using a traditional transcoder 306 that interacts with database 307 via SQL or a proprietary API to convert proprietary information into HTML form. It is to be appreciated that the communication between elements 306 and 305/305a can involve the transfer of XML or meta-information to assist the conversational transcoding/proxy task. In addition, structure skeletons can be used as described in the above-incorporated application Y0999-114.

Fig. 4a illustrates a voice portal, which is a CML server where transactions allow the search of HTML or CML pages or web sites, which is similar to today's web portals. The voice portal can retrieve desired sites and pages. These CML pages are directly presented to the conversational browser. Pages retrieved in another presentation language (e.g., HTML) are transcoded before being transmitted to the browser. The main difference between a voice and web portal (i.e., HTML or WML (wireless markup language) based) is that the portal downloads

the requested pages and transcodes them appropriately before sending them to the browser. For client devices (or browsers) having transcoders, the pages can be send to the browser in another format (HTML and \/or HTML-based structure skeleton) and transcoded on the fly by the browser into CML as an intermediate internal language. The voice portal service can be subscription based, paid by the carrier/wireless provider, paid by the referred content provider (on a hit basis), on complete profitable transaction basis, on a percentage of profit from referred business, on a subscription basis to have the service present in the search engine, or the service ranking high as result of the search engine, or for quality CML pages (hand prepared or reviewed) as compared to the automated approach.

Referring now to Fig. 4b, a diagram illustrates an exemplary embodiment of the present invention for providing conversational browsing or IVR services in conjunction with legacy corporate software applications. The system of Fig. 4b is similar to the system described above (Fig. 4a) except that Fig. 4b specifically illustrates conversational browsing of a conventional 3-tier (distributed) corporate application on a given corporate network. The 3-tier application includes a front end terminal/presentation server 310 (Tier 1) which contains the programming to create and present HTML-based documents, a middle tier business logic server 311 or application server (which contains the business applications and serves client requests from workstations in a LAN, for example), and a back-end or third tier 312, 313, which comprises the legacy database and transaction management (data control) applications. As shown, conventional presentation is accomplished via GUI 314, telephone (IVR) 315 and 3270 type screen dialogs (3270 terminal are types of terminals that are typically supported by mainframes and CISC systems, e.g., the traditional greenish terminals that may be found at ATMs which are mostly ascii character only-based screen). Straightforward HTML dialog pages such as form filling (HTML or CGI) or ATM dialog (and other 3270 type of screen dialogs) or other dummy terminals (CT 100, Palm Pilot screen, etc.) can be readily converted to CML via transcoder 305a of proxy server 305, for example, since the business logic of such transactions is typically driven by simple dialog so that limited amount of information or choices is presented at once to the user. The choice or menu can also be conveyed through skeletons as described in the above incorporated IBM Docket No. YO999-114P. In addition, transcoding can be readily extended to other proprietary APIs or screen maps **(e.g.,** maps of the 3270 screens or other screens into an HTML page usually performed by mapping programs called screen scrapers). In particular,

since programs currently exist for transcoding screen maps to HTML pages, and screen maps are limited to simple dialogs driving the business logic, the conversion of the resulting HTML pages is a trivial case of the conversational proxy.

Advantageously, each of the above-described exemplary system architectures advantageously allow universal access via pure speech interface (conversational browser), pure GUI, and a combination of speech and GUI to the same bank of transaction and information services through a same protocol without requiring any redesign of the content currently provided over HTML and other XML/HTTP protocol.

Conversational Applications Employing CML

The following discussion of exemplary embodiments describe network- based conversational applications using CML (conversational markup language) in accordance with the present invention. The term "conversational application" used herein refers to an application that interacts with the user via voice input/output. The term "network-based conversational application" used herein refers to one in which the elements of the conversion that define spoken output and input, i.e., pages of CML, may be obtained over a network or locally. Indeed, just as GUI applications can be built using a set of interlinked pages written using HTML markup language, conversational systems can be built using a set of interlinked pages written in conversational markup language - CML. Moreover, just as HTML provides a set of mechanisms for translating GUI actions into application actions such as visiting other pages or communicating with a server, CML similarly can be implemented to provide mechanisms for translating spoken inputs into application actions. The term application in this context should be understood in the same broad sense in which a set of related HTML pages containing text, links, and forms, together with the code on a server which processes the forms (e.g., servlets or CGI scripts), constitute an application. In addition, similar to the manner in which procedures (such as conversational foundation classes or other dialog components) can also be loaded via Java or ActiveX components, conversational applets and foundation classes can be loaded /downloaded to execute procedural conversational actions (which is a feature of CVM).

Referring now to Fig. 5, a diagram illustrates an exemplary network-based conversational system comprising various classes of conversational applications enabled by CML according to the present invention. The illustrative system comprises a client device 500 having a

conversational browser 501. The client device 500 may be, for example, a desktop PC, a PDA, an automobile computer, a smart telephone, or a "dumb" telephone (the term "dumb" telephone refers to herein as a device that functions as a microphone at the end of a long cord and needs to be connected to a browser at the other end of the phone line). The client device 500 may also comprise one or more speech- enabled local applications 502 running on the client 500 using CML to communicate with the conversational browser 501. For example, the local application may be a car navigation application in which a "Speech Navigation Application" interacts with computer mapping software and a GPS (Global Positioning System) device to provide conversational driving directions. Similarly, other applications and services provided locally by the underlying conventional OS, or other applications can be used to provide services (such as disambiguation, context access, etc.) and dialog modules to the browser when there is no underlying CVM to provide such services.

The client device 500 can access any one of a plurality of server systems S1, S2, and S3 over network 503 (e.g., the Internet) using a standard network protocol (e.g., HTTP, TCP/IP) to transmit CML pages to the client device 500, which are processed by the conversational browser 501 to provide a conversational UI to the user of the device 500. For example, the client device 500 can connect to server S1 to access existing HTML information via a transcoding proxy which, as described above, is capable of interpreting HTML, extracting relevant information, and generating CML information. In addition, the client device 500 can connect to a specialized web server application (S2) such as Lotus Domino server to access Notes data (e.g., Notes e-mail) via a CGI application. In particular, the Domino server can be configured to generating CML and transmit the CML using HTTP. In another example, the client device 500 can connect to a web server application (S3), using a CGI application or Java Servlet to access an legacy database of an enterprise, wherein the web application generates and transmits the information in CML.

This lead to another interesting application example: a CML based universal messaging system. In particular, the same way that conventional and currently available mail and messaging systems can be provided with HTML by ISP, ASP (Application service providers) and web portals, a voice portal or voice web service provider can use CML to offer e-mail access and processing and universal messaging. Universal messaging, means that you can now access via voice different messaging services and program paging, e-mail, voicemail and call processing by

voice. Because the interface is now built in CML through the conversational browser, the user can process his/her messages automatically. Similarly call transfer and processing can be seen as a combination of a CML interface announcing the caller information and possibly call topics and offering a menu of options. The user can select actions (transfer, store, take the call). Each of these actions are processed as procedural call to the call processing unit and the call is accordingly processed (e.g. forwarded or transferred to the user). With a pure speech conversational browser such as shown in Fig. 5, speech and audio are the only output provided to the user. When a user desires, a page is downloaded from the service provider (in CML). The CML is used to describe the conversational UI to be presented to the user. Using the TTS markup information, the conversational browser will read the page to the user and listen for commands from the user to be decoded by the speech recognizer (using possibly clues from the speech recognition portion of the CML page) to follow links, fill forms, or command the browser. For non-embedded applications, the conversational browser preferably utilizes a large vocabulary continuous speech recognition system (e.g., IBM Via Voice) and a "Free text rule-based TTS using JSAPI (java speech) - including dynamic grammar compilation and NL process for example as described in the reference by K.A. Kishore, et al. entitled "Free-flow dialog Management Using Forms," Proc Eurospeech 1999, Budapest Hungary, September 1999 and Davies et al., "The IBM Conversational Telephony System For Financial Applications," Proc. Eurospeech 99, Budapest Hungary, September 1999>

Advantageously, the above capabilities such as server-based CGI scripts and Java Servlets, and local plugins, can provide mechanisms for application-specific code to participate in the conversational application. The interface between the application code and the conversational browser 501 can take the form of URLs for requests and markup data stream (HTML or CML) or sets of attribute value n-uples. Note that instead of URL, other addressing schemes (e.g., sockets and ports) to access local or distributed services may be used.

It is to be appreciated that each of the preceding IVR and conversational browsing applications could alternatively be built with application code with the applications communicating via API and remote calls (i.e., not implementing a conversational browser and CML). In particular, such applications can be written directly to a speech API such as the Java Speech API (JSAPI). However, for distributed applications, the currently available engine APIs (SAPI, JSAPI) have not been designed with network considerations in mind. Therefore,

networked/distributed speech applications directly written on top of the conversational engine APIs will always require re-implementation of the distributed functionalities. It is to be appreciated, however, for applications written on top of the CVM system (IBM Docket No. YO999- 111P), these limitations of the current engine APIs are virtually eliminated through the conversational application APIs that are communicated via the communication stack which hides most of the complexity of the engine APIs through the services and behaviors offered by the CVM shell. Indeed, complex dialog tasks can be efficiently performed though procedural implementations. Such procedures allow simple use of macros for such complex tasks. In addition, procedures are inherently expected to be more efficient than interpreted declarative language such as XML. Depending on the context of use (the importance of the efficiency vs. programming, portability across platforms, etc) , however, both alternative (declarative and procedural) may be used.

CML and the conversational browser, on the other hand, advantageously allow distributed conversational systems to be implemented which take advantage of legacy applications that have already been develop using HTML, for example. Unlike the currently available speech APIs, CML can readily support networked conversational applications. The conversational browser can interpret streams of CML originating from multiple locations thereby providing a user with the same kind of seamless experience provided by a desktop windowing system or by a visual browser. For example, while completing an airline ticket purchase the user can temporarily suspend the transaction and interact with a banking application on a different server to check an account balance. This is related to the concept noted above of having an appropriate mechanism to keep and transmit/restore the context of a dialog with the browser and possibly the server side processing. Furthermore, CML provides a higher-level tool that simplifies the construction of conversational applications (similar to the manner in which HTML simplifies the building of visual applications). A conversational markup language also reduces the amount of expertise in speech required. In addition, as stated above, building distributed applications on heterogenous platforms using an currently available speech APIs is difficult. The large-scale deployment of distributed GUI applications on heterogeneous platforms via the WWW demonstrates the feasibility of a CML data stream approach.

Referring to Fig. 6, a block diagram illustrates a general distributed conversational browsing system using CML in accordance with an embodiment of the present invention. In

general, a distributed conversational browsing system comprises an information server 600, a presentation server 601, a speech server 602, speech models 603 and a conversational browser 604. The information server 600 represents, for example, enterprise databases, newswire feeds and/or WWW servers. The information server 600 provides data in an "information format" which represents application-specific information in a device-independent manner. Examples of this data format include APIs for data access (e.g., Notes DB API, SQL) and application-specific data streams (news wire feeds, information marked up with XML content tags).

The presentation server 601 retrieves information from the information server 600 using an API or a protocol defined by the information server 600. The presentation server 601 then transcodes the received "information format" to a "presentation format" for presentation to a user. The "presentation format" represents information in a device-specific, application-independent manner for presentation to a user which, in accordance with the present invention, is CML (as described in further detail below). Conventional "presentation formats" include, for example, APIs such as the Windows GUI, API and data streams such as HTML, HDML (Handheld-device Markup Language) and WAP (wireless application protocol). The presentation server 601 can be, for example, an HTML server, CGI scripts and Java Servlets, Domino as an HTML server, and the Notes Client, which are configured for transcoding the information received from information server 600 into CML. The CML generated by the presentation server 601 is used to describe the conversational UI to be presented to the user via a conversational browser 604 and provides the mechanisms for translating spoken inputs into application actions.

The speech server 602 comprises the engines which are responsible for speech recognition and parsing and other conversational engines. The speech server 602 addresses the problem unique to speech: the speech recognition process may require large application-specific data sets (i.e., speech models 603) which it is not feasible to transmit to the client. This implies that the speech recognition process must be carried out where the language models reside near the presentation server 601 (in a transmission bandwidth sense). A standard network protocol such as HTTP, XML, VoIP, WAP protocol, as well as the conversational speech coding protocols describe in the above-incorporated YO999- 113P application may be used for transmitting audio from the conversational browser 604 to the speech server 602 and returning

the parsed results, as well as transmitting audio, such as recorded or synthesized speech, from the speech server 602 to the conversational browser 604.

Referring now to Fig. 7, a diagram illustrates an exemplary distributed conversational system using CML in accordance with an embodiment of the present invention. More specifically, the exemplary embodiment illustrates a speech only client device 606 having a conversational browser 604 and microphone/speaker 605. The client device 606 represents, for example, a speech-only PDA (Personal Digital Assistant) or PVA (Personal Vehicle Assistant) client device connected, for example, by a wireless link to a network 607 (e.g., the Internet). The information server 600, presentation server 601 and speech server 602 (with speech models 603) together provide an application (e.g., weather information) to the client device using CML. The dotted line around the presentation server 601, speech server 602 and speech models 603 indicates that these elements are tightly coupled, in that the speech server 602 has speech models specific to the application. In addition, the presentation server provides a GUI to other browsers 608 using HTML. It is to be understood that the conversational browser 604 may access other conversational applications over the network 607.

Referring now to Fig. 8, a diagram illustrates another exemplary distributed conversational system wherein the client device 606 is a speech and GUI client. The client device 606 comprises a display screen 609, a microphone speaker 605 and a conversational + GUI browser 604a for presenting information to a user using speech, GUI and a combination thereof. In this embodiment, the presentation server 601 can provide both HTML and CML documents depending the request by the client 606. In addition, this system can be implemented by embedding CML markup can be embedded in HTML markup and extending a standard HTML browser (using a supplied browser extension mechanism) to run the conversational browser alongside the standard browser to provide a conversational interface alongside the GUI interface. A system and method for integrating and coordinating between A GUI and speech UI is explained in the provisional application U.S. Serial No. 60/128,081, filed April 7, 1999, entitled "Multi- Modal Shell", which is commonly assigned and incorporated herein by reference

Referring now to Fig. 9, another distributed conversational system using distributed telephony is shown. In this exemplary embodiment, a user can dial into a telephony platform 610 of the client 606 (such as a PC) to be connected to a conversational browser 604, which gives the user access to a selection of speech applications via the network 607 (e.g., Internet).

Alternatively, a user can connect to the speech server via dial in or directly via another existing connection (LAN, Bluetooth,, DSL etc...). In this case, the conversational coding algorithm presented earlier can be used to ship the data to the browser. Note that the client "device" is now essentially a telephone handset coupled to the conversational browser 604 running on the telephony platform 610.

Next, in the exemplary embodiment of Fig. 10, the user will dial into a given phone number to access one application. The system is a simplified version of the distributed telephony system of Fig. 9, wherein the conversational browser 604 and application are the same. In this embodiment, however, the conversational browser 604 is not connected to a network but can access only the specific application. The dotted line 611 in this embodiment emphasizes that the presentation server 601, the speech server 602 (and models 603), the conversational browser 604, and the telephony platform 610 are all operated as an integrated service. This is essentially a traditional telephony application, with the conversational browser 604 serving as dialog engine.

Conversational Markup Language: CML

This following describes a preferred embodiment CML which may be employed in the conversational browsing applications described above. In a preferred embodiment, CML is an application of the Extensible Markup Language (XML). XML is an emerging and popular web standard for HTML-style markup languages, defined by W3C, the same body that maintains the HTML standard. XML is essentially a generalization of HTML which borrows a number of design points from HTML. More generically, XML is a declarative universal format for structured documents and data. Advantageously, the use of XML as a basis for CML allows CML to be readily embedded in HTML or combined with HTML to create multi-modal applications (e.g., speech and GUI). It also allows JSML (Java Synthesis Markup Language), another XML application, to be embedded in CML and used as the speech synthesis markup language for CML. Advantageously, the standardization and popularity of XML make it probable that authoring tools for creating CML pages, as well as programming tools for the server code to generating CML, will become standard, therefore providing open architecture for the conversational browsing systems described herein.

XML comprises markup symbols to describe the contents of a page or file. The XML entities comprises pairs of tags of the form:

<**tag** *attr1*="val2" *attr2*="val2"> *arbitrary text*</**tag**>.

The extensibility of XML arises from the fact that the markup symbols of the XML application (and CML in this case) are unlimited and self-defining (i.e., the programmer is free to define his/her own tag names, attribute names and value sets).

In one embodiment of the present invention, CML is written in JAVA and a conversational browser utilizes a "what you hear is what you can say" style of speech input for purposes of local speech recognition and parsing. In other embodiments, the CML is extended for NLU speech input, and distributed client-server speech applications, wherein digitized speech is sent from the client via HTTP, for example, to a speech server and the parsed results are returned using CML (as noted above). Again, besides using HTTP or other conventional protocols, it is to be understood that speech can be transmitted to the speech server via the conversational communication protocols as discussed in the above-incorporated patent applications IBM Docket Nos. YO999-111P and YO999-113P. These embodiments will be discussed in detail below.

In general, CML is preferably defined as a set of tags, wherein the primary CML elements are **<page>**, **<body>**, **<menu>**, and **<form>**. In general, a "page" element groups other CML elements, and serves as the top-level element for a CML document. A <**body**> element specifies output to be spoken by the conversational browser. A <**menu**> element presents the user with a list of choices, and associates with each choice with a target, e.g., a URL, identifying a CML element to visit if the user selects that choice. A <**form**> elements allows the user to provide one or more pieces of information, where the content of each piece of information may be described by a grammar. The **<form>** element may also specify a URL to visit when the user has completed the form. The phrase "visiting an element" (or "visiting a CML element") refers to an action taken by the conversational browser, typically in response to a spoken user input (although it may also be in response to some other kind of user input or some other kind of asynchronous event). Visiting a CML element will cause the conversational browser to produce a spoken output, depending on the type and content of the CML element visited. For example, a <**body**> element is read, the choices of a <**menu**> element are listed, and so on. Visiting an element can also affect the set of spoken response that the browser will

accept from the user, as described for each element in the following sections and as discussed more fully in the section on dialog issues.

The following describes in further detail the tags (elements) and attributes that comprise a CML document or file according to a preferred embodiment.

Page Elements: A <**page**> element is the top-level CML document element that comprises one or more related nested CML pages or units (a unit is a **<body>**, **<menu>**, or **<form>** element) and is preferably structured as follows:

<**page** ttl="seconds"> (bodies, menus, and/or forms units) or nested pages </**page**>

The attribute ttl ("time-to-live") specifies the number of seconds that the CML page may be stored in a cache, wherein a value of 0 prevents caching of the page. A <**page**> element per se does not contribute (no spoken output or input) to the conversation between the conversational browser application and the user, but serves as a container to group other CML elements (e.g., menu, body, and form elements). A **<page>** element may also contain nested pages. The function of the <**page**> element as a (nested) container will be explained later. Visiting a <**page**> element is equivalent to visiting the first unit in the page. The nested page can also include conversational dialog objects or foundation classes. This can be implemented procedurally or with some equivalent CML pages. When a page is reloaded, a cookie or appropriate protocol/service/API is provided to reload the previous context.

Menu Elements: A <**menu**> element serves a function for CML which is analogous to the function that hypertext link and menu serves for HTML in that it presents the user with a set of choices. In general, a <**menu**> element specifies an introductory message and a menu of prompts to be spoken when the entity is visited, and a corresponding set of possible responses, and for each choice a URL, for example, to visit if the user selects that choice:

<**menu**>

introductory message text

<**choice target**=*"URL1"*>*prompt text 1*</**choice**>

<**choice target**=*"URL2"*>*prompt text 2*</**choice**>

</**menu**>

When a <**menu**> element is visited, its title text (introductory text) is spoken followed by the prompt text of any contained <**choice**> elements. A grammar for matching the title text of the

<**menu**> may be activated when the <**menu**> is loaded, and remains active thereafter. When the user utters a word or phrase which matches the title text of a <**menu**> element, that <**menu**> element is visited.

Each <**choice**> has a target URL that the conversational browser visits if the user selects that choice:

<**choice** target="target">*prompt text*</**choice**>

A target may be any valid URL (including relative URLs and URLs specifying HTTP:, file:, FTP:, and local: protocols (see below for a description of the local: protocol) or a URL plus a reference (using the standard URL#reference syntax.). Any other addresses, schemes, and protocols may be employed. The reference (i.e., the part of a URL after the #) is a string that names the unit within the document that has a name attribute whose value is the reference string. A URL without a reference is considered to refer to the first unit within the document. A relative URL is resolved relative to the URL of the containing document.

Furthermore, a "target" may be another address such as a socket address (IP address and port ID). In such case, other IP protocols such as the conversational protocols described in the above-incorporated patent applications IBM Docket No. YO999-111P and YO999-113Pcan be implemented. In situations where a specific non-conventional protocol is used, an additional argument can be used to activate communication such as the following:

<**choice target**="target" **Protocol**="protocolidentifier"> </**choice**>

This provides a mechanism to ship the context and meta-information. It also provides a way to provide system calls to CVM services or, in the absence of CVM, to present calls to equivalent services implemented as applications on top of the underlying OS, or within the browser platform itself. If the browser supports the protocol, the identifier is enough to activate appropriate communication. This is especially applicable for a conversational browser built on top

of CVM (as described above and in IBM Docket No. YO999-111P). In other cases, the protocol identifier can also point to a URL (plug-in or applet) to download so as to initiate the communication. It can also directly point to a local executable on the client platform supporting the conversational browser. Typical examples are "choices" that require distributed recognition (e.g., where the recognition is performed on a remote networked server) (as described in IBM Docket No. YO999-113P). Other examples include choices that download procedural functions

(or local functions) which implement the dialog. Again these conversational procedures are built on top of CVM. It is to be understood that these concepts are valid for all "targets" and not just "choice."

The "prompt text" of a <**choice**> element is spoken when the containing <**menu**> is visited. A grammar for matching the prompt text of a <**choice**> element is activated when the containing <**menu**> is loaded, and remains active thereafter. When the user utters a word or phrase that matches the prompt text of a <**choice**> element, the specified target is visited. For instance, the user may select one of those choices by saying an attention word or phrase ("computer", "go to" or "select"), followed by one or more significant words from the choice. The following example will illustrate the menu element:

<**menu**>

Please choose from the main menu.

<**choice target**="*file:e-mail*">**E-mail.**</**choice**>

<**choice target**="*file:news*">**News.**</**choice**>

<**choice target**="*file:nav*">**Navigation.**</**choice**>

<**choice target**="*file:mcform*">**Food Ordering.**</**choice**>

<**choice target**="*file:weather*">**Weather information.**</**choice**>

<**choice target**="*file:tutorial*">**Tutorial.**</**choice**>

</**menu**>

The main menu may serve as the top-level menu which is heard first when the user initiates a session with a conversational browser. More specifically, when the conversational browser visits this menu it produces the spoken output "Please choose from the main menu" followed by a list of choices: "E-mail. News. Navigation. Food Ordering. Weather Information. Tutorial." Once the conversational browser has loaded this menu, the user may activate (select) any of the choices (during the remainder of the session) by speaking a command. The allowable commands will depend on the input technology being used. In one embodiment which implements a "what you hear is what you can say" approach, the allowable commands may comprise attention phrases (such as "go to" or "select") followed by a subsequence of the words of the prompt text, e.g., "go to e-mail" and "select news."

After the user has uttered a selection, the conversational browser will visit the target address (e.g., URL, socket address) specified by the target attribute associated with the given choice (e.g., the contents of a target URL are fetched and interpreted as a CML entity, and that entity is visited). Note that in the above example menu, all of the target URLs are relative to the URL of the containing page, which in this case is a file: URL that refers to the demo\main file - for example, the target "file:news" is interpreted as a file called "news" in the same directory as the file that contains the main menu.

In summary, the user may speak a phrase that matches the title text of a <**menu**> element to visit the <**menu**> element. The user may speak a phrase that matches the text of a <**choice**> element to visit the target specified by the element. A user's phrase matches the text of a <**menu**> or <**choice**> element if it comprises an attention phrase followed by one or more words from the text in the same sequence that they appear in the text (but not necessarily contiguously). An attention phrase is a phrase such as "go to" that indicates that a command follows.

Body Elements: A <**body**> element specifies some text to be converted to spoken output when the entity is visited:

<**body** name="name" next="target"> *text* </**body**>:

When a <**body**> element is visited, its text is spoken, and then the target specified by the next parameter is visited. The CML <**body**> entity with JSML markup serves a function for speech which is analogous to the function that an HTML body entity with presentation-oriented markup (such as headings, lists, etc.) provides for GUI. For information-retrieval applications (such as e-mail and news), a <**body**> element will typically contain information that the user is seeking. A body element is preferably marked up using Java Synthesis Markup Language (JSML) so as to obtain accurate and efficient text-to-speech synthesis:

<**body**>*Text to be spoken, marked up using* ***<EMP>****JSML****</EMP>***</**body**>

The "next" attribute of the <**body**> element has a value that is a target address (e.g., URL) which specifies another CML element (such as a <**menu**> element or a <**form**> element)

that is to be visited after the body has been spoken to the user. The following example will illustrate the function and operation of the <**body**> element and the use of JSML markup and the "next" attribute.

<**body next**="*#menu1*">
<**JSML**>
Welcome to the IBM ViaVoice<EMP>Speech Browser</EMP>. This tutorial will familiarize you with the spoken commands that you can use to control your speech browser.
</**JSML**>
<**body**>

The body is formatted using JSML to place emphasis on the term "Speech Browser." The "next" attribute in this case is a relative URL (or target address) that instructs the browser to visit the element with the name "menu1" on the current page.

As described above, in one embodiment of CML, the <**body**> and <**menu**> elements provide spoken presentation and navigation of static information using spoken menus. As with HTML, this level of functionality is sufficient for static information that can be organized as a set of interlinked CML pages. A much richer set of applications is possible, however, if the user can supply to the application information taken from large sets, such as search words, dollar amounts, dates, stock names, etc. Such sets of possible inputs are too large to be presented in a menu, so another mechanism for collecting the input at the client is required. Furthermore, a mechanism may be used to compute the application response "on-the-fly" at the presentation server rather than being stored as with menu and body elements. This can be done, for example, via CGI (Common Gate Interface) programs and servlet programs running on the server, or any other backend logic. In cases where a complex logic is hidden behind the transaction or the server application, such server side on-the-fly modification of the response is the only way to proceed (except of course when information about the logic is transmitted using the conversational skeleton methods described in the above incorporated IBM Docket No. Y0999-114 or if the information is transmitted t the client: structure skeleton to disambiguate parts of the dialog).

Collecting such user input may also be accomplished in CML using a **<form>** elements.

Form Elements: A <**form**> element collects one or more pieces of information, or fields, from the user. A <**form**> element is typically used for collecting information, such as names, addresses, telephone numbers, and any other types of information that would be impractical to present as a list of choices in a menu, and has the general form:

<**form action**="URL">
introductory message text
<**field name**="*name 1*" **rule**="*JSGF*">*prompt text 1*</**field**>
<**field name**="*name 2*" **rule**="*JSGF*">*prompt text 2*</**field**>
...
</**form**>

An action is a target URL used in a <**form**>. The values of the <**field**>s of the form are appended to the action as attribute-value pairs, in the same manner as in HTML. A form groups together a set of user inputs that together are sufficient to warrant going back to a presentation server (which may in turn send a request to an information server) to obtain an application response. When a <**form**> element is visited, its title text (introductory text) is spoken, and the user is then prompted one by one for the values of any contained <**field**> elements. A grammar for matching the title text of the <**form**> element is activated when the <**form**> is loaded, and remains active thereafter. The <field> tag has the general form:

<**field** name="name" rule="ruleName" value="value">*prompt text*</**field**>

If a value has not already been supplied (either by the user or by the value parameter of the <field> tag) and if the prompt text is not empty, the user is prompted for the value of a field by speaking the specified prompt text. Acceptable responses for filling a field will depend on the input technology being used. For the "what you hear is what you can say" approach, acceptable responses includes a subsequence of the words of the prompt text followed by a phrase matching a specified grammar (which is preferably in Java Speech Grammar Format (JSGF)) that provides the set of possible values for the field. The portion of the response matching the grammar becomes the value of the field. As discussed below, this can be extended to NLU input. If the user's response fills a field, the user is then prompted for the next unfilled field, and so on.

When all the field values have been specified, the action target is visited using, for example, an HTTP GET method (i.e., the results are transmitted to the presentation server), an a page of CML is returned containing an entity which is then visited. It is to be appreciated that the user may fill in the value of any field either when prompted or at any point as long as the form is still in scope by speaking a portion of the prompt followed by a legal value of the field (i.e., by speaking the prompt text followed by a phrase matching the ruleName specified by the rule parameter, wherein the "ruleName" is a fully qualified JSGF rule name).

The following example will illustrate the function and operation of a form element.

<**form action**="*HTTP://localhost:8080/servlet/McServlet/*">

Please complete the order form.

<**field name**="sandwich"**rule**="*mcgrammar.sandwich*">Sandwich is?</**field**>

<**field name**="drink"**rule**="*mcgrammar.drink*">Drink is?</**field**>

<**form**>

When the above form is visited, the conversational browser outputs the introduction: "Please complete the order form." The user is then prompted "Sandwich is?". The user may then respond "Sandwich is chicken sandwich" or "Sandwich is hamburger" based on the following: The specified JSGF rule, "<*mcgrammar.sandwich*>", is found in JSGF grammar "mcgrammar", which reads

grammar mcgrammar;

public <sandwich> = hamburger chicken sandwich;

public <drink> = coke pepsi;

Note that the user may also fill in this or any field by saying "sandwich is sandwich" or "drink is drink" before being prompted for the value, and for as long as the sandwich form remains active. In this case, the browser selects a value for the field and then prompts the user one by one for the unfilled fields. This is a dialog feature known as "mixed initiative." This means that the user may take the initiative by saying "sandwich is sandwich," or the system may take the initiative by prompting the user for sandwich because, for example, the user has activated a menu choice that leads to a food-ordering form.

CML forms are similar to HTML forms, and in one embodiment, CML uses a mechanism similar to HTML to transmit the field values to the server: when the user completes the form, specifying a value for each field, the browser appends the *name=value* pairs for each

field to the specified form action URL, and the resulting URL is requested from the server via HTTP. Preferably, the form action URL incorporates the name of an application-specific function, such as a CGI script or a Java Servlet, which the server will call on to process the *name=value* pairs and return a response.

As with the case of a graphical web browser wherein the response to a completed form will be a page of HTML to be presented to the user on a display, with the conversational browser, the response will be a page of CML to be presented to the user using speech (by visiting the specified body, menu, or form entity). In either case, the new page will affect the interpretation of subsequent input actions.

Dialog Issues: One consideration in the design of a spoken dialog application is the ability to allow a user to readily determine what he/she can say at any point in the "conversation." Ideally, the application would accept any reasonable spoken input from the user (NLU), which is sometimes difficult to implement. Instead, an application can be designed to accept a limited set of inputs. This design, however, presents the user with the burdensome task of having to learn and remember a "language" comprising an arbitrary subset of his/her natural spoken language.

The CML browser advantageously mitigates this problem by implementing various approaches. One approach is referred to as the "what you hear is what you can say" approach, in which (as explained above) acceptable spoken inputs always echo the spoken prompts that are presented to the user. Each time a CML element is visited, the set of user inputs that the browser will accept is changed. For example, after the user visits a <**menu**> or <**form**>, the browser will accept responses appropriate to that menu or form. Following the "what you hear is what you can say " approach, the inputs accepted are in general echoes of the menu or form prompts, or some abbreviated version of the prompt.

For example, if the user hears: "Choose from the main menu: E-mail. Stock Quotes," the user can say: "Go to [the] main menu"; "Go to stock quotes"; and/or "Go to e-mail." In addition, if the user hears "You have 2 messages: New message 1 from Bill; Smith about Golf tee time. Message 2 from Jane Jones about Project meeting", the user can say: "Go to new messages"; "Go to new message 1"; "Go to message 2"; "Go to message from Bill"; "Go to message about tee time"; and/or "Go to message about project." Moreover, if the user hears: " Stock quotations.

Stock symbol is?", the user can say: "Go to stock quotations"; "Go to quotations"; "Stock symbol is I B M"; "Symbol is I B M" and/or "Stock is I B M."

As described above, the first two examples above may be accomplished in CML using a <**menu**> element that contains some title text ("Choose from the main menu") and a number of <**choice**> elements containing some prompt text ("E-mail", "Stock quotes"). The last example may be accomplished using a <**form**> element that contains some title text ("Stock quotations") and one <**field**> element containing some prompt text ("Stock symbol is?"). Accordingly, the title text of the <**menu**> or <**form**> and the prompt text of the <**choice**> or <**field**> define what the user may say.

Other options for implementing the dialog may be used. For instance, it is possible to pre-load answers to some of the forms. Namely, some scripting elements can be introduced to pre-program via procedures or script based on the answers or the selection from the user (i.e. variables are, for example, the fields of the prompts). Of course, this scripting can use/refer to a variable assigned by other event and procedures local or distributed (e.g. date, result from a query to the presentation server, etc...). It implies catching events, adding logical operations, loops and redirection (e.g. go to) statements. For situations using multiple forms, the capability to catch events is needed. Such events in the context of CVM could catch events also coming from applets and other conversational procedures, or scripts. Scripting can be done in a manner similar to Javascript or ECMAscript. The script may be embedded within the form **<script>...</script>.** This means also that forms, menus, etc., can throw and catch events. This enables appropriate programming capabilities. Scripts can be embedded anywhere within a CML page. Procedures, dialog components, conversational foundation classes and other services can be implemented by **<object>...</object>** tags.

In another aspect, a "Say what you heard" approach allows the user to be in control of the conversation (rather than the browser as described above). More specifically, during a conversational session, every previously visited <**menu**> or <**form**> that has been heard remains active. By way of example, assume the user has heard the main menu choices for "e-mail" and "news." With the "Say what you heard approach," the user can either immediately select the e-mail application by saying "select e-mail" or, at any time subsequent to hearing the main menu, select the news application by saying "go to news" without having to navigate back to the main menu. If the user has forgotten selections that were on the main menu, the use can get back to

the main menu by saying "go to main menu". This principle applies to every menu and form that a user has heard.

Another approach is "Say what you will hear," wherein the menus and forms become active before being visited and remain active throughout the browser session. More specifically, a menu or form becomes active when it is first loaded by the browser, which may be before the menu or form has been visited such as in the case where a page containing several menus and forms is loaded. Even if only one of the menus or forms on that page is visited (depending on the specifics of the URL), generally, all the menus or forms of the page will become active. This is illustrated in the following example:

```
<page>
  <body text="#new2menu">
    New message 2 from Steve about Pay raise.
    Bruce,
    We've decided to give you a 250% pay increase this year.
    Don't spend it all in one place.
    Steve
  </body>

  <menu name="new2menu">
    Choose from the message menu
    <choice target="#new3">Next message.</choice>
    <choice target="#new2forward">Forward message.</choice>
  </menu>
</page>
```

In the above example, assume the user requested to hear message 2. The conversational browser will visit the URL for the page (i.e., it will fetch the page and visit the first unit on the page), which in this case is the body of the message. After the message is finished, the "next" attribute of the **<body>** element causes the conversational browser to visit the "new2menu", which offers the user a couple of choices for disposition of the message. The user does not have to wait, however, to hear the menu before selecting one of the choices. Instead, the user may interrupt the message at any time to say, for example, "go to next message," since the menu

becomes active when the page is first loaded. Advantagcously, this feature is useful in frequently used applications where the user learns to anticipate menus and forms that have not yet been presented. It is to be appreciated that the "say what you will hear" approach may be implemented using the skeletons as described in the above-incorporated IBM Docket No. Y099-114 to preload a dialog object when the conversational browser is built above CVM using Conversational Foundation Classes. Alternatively, the dialog object or dialog component can be implemented in CML.

It is be be appreciated that the "say what you heard" and "say what you will hear" mechanisms contribute to a dialog feature known as "mixed initiative." This means that in some cases the computer takes the initiative in the conversation - for example when the browser prompts the user for a set of menu items - and sometimes the user take the initiative - for example, when the user ignores the prompts and selects a menu item that was heard in the past, or interrupts the computer and selects a menu item that hasn't been presented yet. Thus, the set of spoken responses that are acceptable by the browser is defined by the entire set of CML elements visited and/or loaded by the browser, and not just by the most recent element visited. However, this can sometimes lead to unexpected results because of menus or forms that are no longer relevant being active. Consequently, CML according to the present invention offers a mechanism to limit the scope of activation for menus and forms.

Scope: Scope is a feature that allows a CML application designer to control the duration for which the spoken input associated with a particular <**menu**> or <**form**> is active. The "scope" feature is specified by a scope attribute that may be associated with a <**menu**> or <**form**> element:

<**menu** name="name" scope="scope"> *title text, choices* </**menu**>

<**form** name="name" action="target" scope="scope"> *title text, fields* </**form**>

Exemplary values of the "scope" attribute are the following:

Local: The menu or form is only active if the menu or form itself is the last element that the browser visited.

Page: The menu or form is only active if the last element visited by the browser is on the page (or a subpage of) the page that directly contains the menu or form.

Global: This is a default value, wherein the menu or form is active for the entire duration of the browser session beginning when it is first loaded by the browser (even if the form or menu itself has not yet been visited).

Application specific: This refers to all the page identified with the site or the application (e.g., via meta-information tag).

In many cases, the greatest flexibility for the user is provided if a menu or form becomes active as soon as the browser has first seen the menu or form and remains active thereafter. For example, as explained above, it is advantageous to have main menus or major topic menus active for the entire sessions so that a user can jump directly to the choices provided by the menus without first having to navigate back through a "maze" of menus. In some circumstances, however, this can decrease speech recognition accuracy and produce unexpected results. Consider the following example.

```
<page name="new2">
  <body next="#new2menu">
      New message 2 from Steve about Pay raise.
      Bruce,
      We've decided to give you a 250% pay increase this year.
      Don't spend it all in one place.
      Steve
  </body>

  <menu name="new2menu" scope="page">
      Choose from the message menu.
      <choice target="#new3">Next message.</choice>
      <choice target="#new2forward">Forward message.</choice>
  </menu>
```

The menu associated with each message allows the user to say "go to next message." This selection, however, probably makes little sense after the user has left the e-mail message to do another task, and it could be very surprising if it were activated as the result of a speech recognition error. Therefore, in this example, the message menu is given a scope of "page," meaning that it will only be active as long as the last element visited by the browser was in the

page that contains the message menu. In this case it means that "go to next message" is only active as long as the last browser action involved visiting the body of the e-mail item.

Additional use of the scope and behavior with the scope are provided in the embedded CML (as discussed below) Another scope tag that may be used in conjunction with CML is "Multiple". A discussion on the use of multiple form (not in CML form) is provided in: Kishore, et al, "Free-Flow Dialog Management Using Forms," Proc. Eurospeech 1999, Budapest Hungary, September 1999 and Davies et. al., "The IBM Conversational Telephony System For Financial Applications, Proc. Eurospeech 99, Budapest Hungary, September 1999. Note that according to this, instead of having multiple pages active, multiple forms can be simultaneously activated on a page. These pages can be explicitly nested in the document or they can be referenced by address (to be loaded). Similarly some of these forms can be presented as procedural components.

Typically as described above, under one NL modality, the NL dialog is implemented by activating multiple forms. The forms are filled simultaneously until all the mandatory fields of one form are filled. When filled, the corresponding action or query is executed. When multiple forms are filled, disambiguation dialog is added (e.g. by activating a new form). Multiple scope forms are typically loaded via a page which indicates to load these different forms:

<menu name =mainnl scope =global>

Introduction dialog

<form name = form1 scope = multiple>....</form>

<form name = form2 scope = multiple> ...</form>

<form name = form3 scope = multiple> ...</form>

...

</menu>

Upon any input that implies an action, a new page will be loaded after completion of the action:

<menu name=mainnl scope = global> (we "update" the global menu)

<form name = form1 scope = multiple context = maintain> ...</form>

<form name = form2 scope = multiple context = reset>....</form>

<form name = form3 scope = deactivated context = reset> </form>

<form name = formnew scope = multiple context = reset>....</form>

...

</menu>

Alternatively all these forms can be loaded by :

<form name ... scope ... load = target> </form>

When a load argument is present, the content of the form is to be downloaded from the target and placed between the **<form> </form>** tags. The same concept can be used for any other element of the languagc. Typically, the menu is updated and new forms are added and some of the forms can be deactivated (scope deactivated).

The "context" tag in the above example indicates how to handle past conversation history. In the two examples given above, the form with a context "maintain" will keep the value filled in the form from the previous steps of the dialog (previous input from the user). The "reset" tag refers to resetting the values to default values. Note also that scope can be defined with a duration stamp: scope = 5 s or scope = 1 minute or scope = the name of a scope. The scope is introduced as

<scope name=scope1 status = active> </scope>

When not defined, a scope is non-active. When defined, it can be active or non-active. Elements scoped with that name are activated when the scope is activated and deactivated otherwise.

It is to be understood that the forms used here can be implemented with a procedural object:

<form name=form3 scope = multiple context = reset load = target> </form>

where the target page contains applets, procedures, services etc for example:

<Dialog Object>

<applet ...>

</applet>

</Dialog object>

Again, a dialog object can be an applet (java), a script (interpreted by the browser), a conversational object (Conversational dialog written on top of CVM using conversational foundation classes), plug-in or servlet activation, or any other procedural implementation or service providers.

In the present case, multiple objects can be loaded in parallel. This is typically a case where the browser is implemented on top of CVM, w here the CVM is able to register these

different objects and their context and to determine which dialog is active as a function of the user input.

This last item illustrates that with CML and a browser hooked to appropriate dialog management capabilities, CML can be used to design NL dialogs, independently of the underlying NL technology. Form based, procedural based (decision networks) and grammar based NL dialog managers are fully supported by this approach.

It is also possible to have overloaded tags (i.e., modifying the action associated with a tag). Typically, this can be done via XML or other mechanisms allowing a description of the new meaning of the tag and the scope of the overloading definition.

Browser features: So far we have discussed CML and described how a browser behaves, with respect to spoken output and spoken input, in response to the markup language. The following discussion is directed to features of the conversational browser that are not intrinsic to CML, but rather are characteristic of the browser implementation.

To begin, the conversational browser preferably implements a number of built-in commands (which are similar in spirit to the built-in functions presented on menus and toolbars in a visual browser). Examples of such commands are as follows:

[be] quiet; shut up: These commands will cause the Browser to stop a current spoken output and wait for further instructions from user. This feature is known as "barge-in" and applies to all spoken input so that a user may interrupt the browser at any time. So, for example, the user may interrupt the browser by saying "quiet please," in which case the browser will stop the current output and wait for further instruction. To cause the browser to repeat the interrupted output, the user may then utter, for example, "repeat that please." The user may also interrupt the browser at any point with a spoken command, such as a menu item selection.

say again; repeat that: These commands will cause the Browser to repeat a most recently visited element (menu, form, or body).

go back: This command will cause the Browser to return to a previously visited element in history list.

go forward: This command will cause the Browser to advance to a next visited element in history list. (Only meaningful after some number of "go back" commands).

[[[go]to]the]beginning: These commands will cause the Browser to go to the first element visited (e.g., the "home" page).

The brackets in the above exemplary commands indicate optional words. Any of these commands may be preceded or followed by "please." Other classical commands can be defined for the browser or added by the user.

Inevitably, ambiguities may arise between these various types of spoken input. Ambiguity resolution is handled at the level of units (forms and menus) by maintaining a most-recently-used (MRU) queue of units. More specifically, a spoken input is resolved against a unit if it matches a spoken phrase allowed by the unit (attention phrase followed by title text or choice prompt, field prompt followed by legal field value). The browser attempts to resolve every spoken input against each unit in the MRU queue in order. In one embodiment, the MRU queue may be maintained as follows:

1. When a spoken input is resolved against a unit in the queue, that unit is moved to the head of the queue.

2. Then, if the spoken input causes a target to be visited, all the units in the page containing the target are moved to the head of the queue.

3. Finally, the unit corresponding to the visited target is moved to the head of the queue.

It is to be understood that with a browser built on top of a CVM, the CVM will perform the appropriate ambiguity resolution and dialog with the user if any ambiguity is suspected to arise as discussed in the IBM Docket No. YO999-111P. Essentially, once a CML specification is produced, it will virtually prescribe the manner in which a dialog should be handled. If this needs to be overwritten, or if other behavior(s) need to be introduced, this can be performed by express calls to the underlying CVM, OS, or platform (if available).,

Plugins: The present invention provides a mechanism that provides a plugin- or applet- like capability. A URL specifying a <**choice**> target or a <**form**> action may invoke local plugin code by using thc local: protocol, following one of several exemplary forms:

local:service/function; or

local:service/function?arg1=value1,arg2=value2,...

The first form is useful for <**choice**> targets which have no arguments and for <**form**> actions where the arguments are supplied by the browser from the values of the contained

<**field**>s. The second form is useful for <**choice**> targets where the arguments are supplied in the URL.

In a preferred embodiment of the browser implementation using Java, a local: URL is implemented by looking up the service in a local table that maps it to a class name, instantiating an object of that class if it is not already instantiated, and then interpreting the function as the name of a method to be called. The table of services is preferably located in the file services of the conversational browser, wherein the file itself contains a description of its format. This includes registration of all the available services and other objects (procedural or not)

The plugin code invoked by the local: URL returns a string containing the CML document represented by the URL. Visiting a local: URL causes the returned document to be interpreted in just the same way as visiting any other URL. It is to be appreciated that these features may also be implemented using applets (downloaded or locals), conversational objects and servlets/cgi as well as distributed applications (with for example socket connections and RPC protocols) to act as a local application.

The service may subsequently interact asynchronously with the browser (e.g., to notify the user asynchronously of important events) by starting a thread and causing the browser to asynchronously visit specified URLs (including local: URLs) via an API provided by the browser. The browser API allows local plugin code to cause the browser to visit a specified URL, just as if the user had spoken something that caused the browser to visit the URL. This can be used to provide asynchronous user notifications based on local events. The local plugin will extend the class CML, Service, from which the local plugin inherits a method visit which, when invoked, causes the browser to visit the specified URL with the specified parameters: public class Service {public void visit(URL url, Dictionary parameters) throws IOException;}.

Considerations for Embedded Conversational Browser Applications

Registration mechanism: In situations wherein the functions of an embedded conversational browser may be limited due to inadequate resources (memory, CPU power, etc), a cookie mechanism can be used that allows the browser to transfer to a server a description of its capabilities. For example: Speech_Reco = 1:Speaker_Reco = 0; TTS = 1; Dialog_Manager = 0,NLU = 0; Speech_Reco.Vocabulary_Size=500; Speech_Reco.FSG=1 etc. Upon receiving the cookie, the server side can examine its content and then modify the CML page accordingly. For

example, if the cookie indicates that the vocabulary size is 50, the CML menus can be generated with very limited vocabularies (e.g. selection of a number per item instead of the actual link).

In another embodiment, rather than using a cookie, registration conversational protocols can be used as described in the above incorporated patent applications IBM Docket Nos. YO999-111P and YO999-113P. In such an embodiment, objects are exchanged which describe the characteristics and properties of the engines as well as possibly the requirements of the application. The handshake will determine the conversational responsibilities of the applications. Again, in such case, the preliminary handshake can use HTTP or other protocols such as RPC, RMI, etc. This is especially important for Dialog Objects that can directly dialog with the server to check an possibly adapt their behavior to the capabilities of the local browser. Note also that like in a conventional browsers, these cookies (or procedural protocols) can perform other tasks such as customization of the presentation or service to the user's preference or persistence of these preferences or of the history across sessions. In other words, if it is an option selected by the user, persistence across devices of the context history is guaranteed by having cookies with the past history of a site uploaded upon connection. When performed via a CVM or procedural objects, the information is transmitted via procedural protocols. Now ID cookies are also used to guarantee persistence across devices when stored on the server. When connecting from different devices, the cookies (or procedural protocol) can be stored on the server or the presentation server. Identification of the user can be done using conventional techniques such as userID, caller ID, speaker recognition or speech biometrics. Any other meta-information (persistence, user preference, application preference, and usage history, context, etc) must be transferred (via a cookie or via conversational protocols)

Therefore, if the conversational engine cannot generate the baseforms, the tag can provide a baseform. Various approaches can be utilized: 1) upon determination of the browser capabilities, the browser is sent a different CML page; 2) a transcoder can dynamically modify a CML page sent to the browser based on the registered capabilities; or 3) a CML page can be modified to send the speech and perform complex functions on a networked server (as discussed in IBM Docket No. YO999-113P).

Dynamic deactivation: The conversational browser may use dynamic activities of CML. This is another solution which allows modification of the behavior of the conversational browser for

embedded applications. In particular, instead of completely following the paradigm of "what you hear is what you can say," the conversational browser may only activate locally the markup unless explicitly limited by CML. Furthermore, upon reaching the maximum amount of supported active vocabulary, the speech browser can progressively and hierarchically be deactivated: the oldest command which is not a browser shell command or a global command (as defined by the CML with a Global Tag <**Global**></**Global**> or priority level tags; this is equivalent to a concept of scope tags). A top menu remains active as long as possible commands under a menu item are deactivated first etc. Deactivated commands are cached in series of command caches. When the likelihood (or other confidence measure of the recognized input) obtained for a command are too low or the command is rejected by the user, the utterance is re-decoded against the most recent command cache etc until an acceptable recognition is obtained. Recently, efforts have been initiated to develop appropriate confidence measures for the recognized speech. For instance, In "LVCSR Hub5 Workshop," April 29 - May 1, 1996, MITAGS, MD, organized by NIST and DARPA, different approaches are proposed to attach a confidence level to each word. One method uses decision trees trained on word-dependent features (amount of training utterances, minimum and average triphone occurrences, occurrence in language model training, number of phonemes/lefemes, duration, acoustic score (fast match and detailed match), speech non-speech), sentence-dependent features (signal-to-noise ratio, estimates of speaking rates: number of words or of lefemes or of vowels per seconds, sentence likelihood provided by the language model, trigram occurrence in the language model), word in a context features (trigram occurrence in language model) as well as speaker profile features (accent, dialect, gender, age, speaking rate, identity, audio quality, SNR, etc.). A probability of error is computed on the training data for each of the leaves of the tree. Algorithms for building such trees are described in Breiman et al, "Classification and regression trees," Chapman & Hal,1993. At recognition, all or some of these features are measured during recognition and for each word the decision tree is walked to a leave which provides a confidence level. In the reference by C. Neti, S. Roukos and E. Eide entitled "Word based confidence measures as a guide for stack search in speech recognition," ICASSP97, Munich, Germany, April, 1997, a method is described which relies entirely on scores returned by an IBM stack decoder (using log-likelihood -actually the average incremental log-likelihood, detailed match, fast match). In the LVCSR proceeding, another method to estimate the confidence level is performed using

predictors via linear regression. The predictor used are: the word duration, the language model score, the average acoustic score (best score) per frame and the fraction of the NBEST list with the same word as top choice. The present embodiment offers a combination of these two approaches (confidence level measured via decision trees and via linear predictors) to systematically extract the confidence level in any translation process, not limited to speech recognition.

A "rewind" or "reload" command can be implemented to read back to the user all the commands met on a page. It follows the same principal of dynamic deactivation. More specifically, a entire voice navigation interface is provided by the browser to enable navigation through the previous forms and menus. It can be implemented by a command and control interface such as IBM's ViaVoice VoiceCenter (which performs Command and Control for Windows).

Capability-based frames: Similarly to the frame structure, the CML developers can offer multiple dialog based on different levels of capabilities [cfr<frame> and </frame> tags in HTML].

Forms with minimum requirement: Of course only forms (or menus or other components) which have a limited set of possible entries (limited vocabulary and FSG, selected menu) can be used with the embedded engines. To maintain the capability to have forms, the browser offers at least two other modalities: i) form filling with prescribed grammar (compilation done on the server) and vocabulary list: the other commands, except the <Global> commands are deactivated, ii) applet capturing the audio, computing special features and sending them to a server indicated by a URL, iii) applet to send or receive raw audio to the server. The selection of the form filling method is done by the server or by the browser. This requires that each grammar/vocabulary list to load must contain also a description of the engine requirements. (ii and iii need server/communication capabilities as described in YO999-113P).

A tool will be needed to determine roughly the engine minimum requirements for a given FSG, for the transcoding approach as well as the multiple pages or frame approaches.

MULTI-MODAL BROWSER

With multi-model browsing, HTML and CML can provided on the same page or sent separately and synchronized. Tags to distinguish what is to show on a visual browser (<GUI>) vs. what to show on a multi-modal browser (<M-M>) and speech browser (<speech>): a GUI

browser displays everything while a multi-modal browser can display selectively some items. This is illustrated by the following examples:

<GUI>

.....

</GUI>

<Speech+GUI+M-M>

.....

</Speech+GUI+M-M>

<GUI+M-M>

<img = ...>

</GUI+M-M>

<Speech>

.....

<GUI>

.....

</GUI>

<M-M>

.....

</M-M>

<M-M+Speech>

.....

</M-M+Speech>

In other words, CML and HTML are preserved. Some dynamic HTML features and functions as well as combinations rules are added.

Additional tags to offer Dynamic ML (HTML and CML combined): For example as the TTS reads the text, the item change colors or a local background change color. Change of color of the link selected by voice etc... Selection of recognized text by voice (change color can be selected and modified).

In summary, the introduction of the conversational browser allows Internet and Telephone (IVR) to be unified and thereby decrease the cost, enlarge the coverage and flexibility of such applications. The architecture uses the same access protocols and the information structure (HTML, XML, CML and/or other ML such as WML). The present invention can be applied to many different business solutions. The main advantages are that it is easy to implement, high flexible, platform independent, uses the existing infrastructure, deployed and managed centrally, provides high security, low cost maintenance, and is readily expandable/scalable, all of which are typical for any Internet solutions. Advantageously, the application provider need support only one HW platform which provides ease of maintenance. The content provider can administer various applications from the same WWW server so as to service a large number of different clients (pure speech, pure GUI and a combination of speech/GUI), which provides ease of maintenance. Moreover, the Internet technology enables automatic updating of all servers on the network. The access rights can be also managed centrally from any place accessible over Internet. High level of security can be maintained. The current architecture can be extended to other clients in addition to regular telephones, such as GSM terminals. This is very attractive from the perspective of the investment that is required to maintain a service such as personal banking.

In addition, when only HTTP protocols are used, the traffic on the network between the browser (conversational/GUI) and the WWW server is minimized, the network is used for sending the completed replay in one batch. For example, while entering the account Number and the SSN only the browser/conversational browser are active, the network is idle. The browsers are intelligent enough to catch many errors before the server is queried. Error correction and local disambiguation can for example be performed via CVM services or downloaded applets. The information to the WWW server is sent in both cases using the same protocol the HTTP protocol. The replay information is sent back always using the same protocol and it is the task of the browser to get it to the user in the right form. The suggested architecture enables the designers of thesc applications to use applets, small programs (java or other conversational procedures, especially if the browser is written over a CVM)), which are executed in the browser to preprocess the input and output information. This further increases the flexibility of the application and furthermore decreases the load of the network. For the conversational browser we can easily imagine a custom "Text to Speech" applet will be sent over

and then the required bandwidth will be the same or smaller then the graphical one. In particular, this can be done to synthesize prompts in another language and recognize input in other languages.

Indeed, the present invention provides these advantages to the current telephony applications, typically the IVR applications. The telephone number is perfect for billing purposes. The latency of Internet does not pose any problem to IVR. In any event, latency can be reduced using services like Qos (quality of service) and RSVP (resource reservation protocol). The capability of connecting on the Internet to other servers (IVRs) makes the capabilities much larger. The server can be located anywhere on a computer network with the access to required speech recognition systems being also available anywhere on the network. The packet switching technology with the TCP/IP protocol utilized the net resources better then regular phone connections. The billing may be based on the quality of used channel. The switched packet network is cheaper and transfers less data, therefore, requiring less bandwidth. This will result in providing the services with lower cost. This concept gives anybody chance to set a server and put the information on net.

Although illustrative embodiments have been described herein with reference to the accompanying drawings, it is to be understood that the present system and method is not limited to those precise embodiments, and that various other changes and modifications may be affected therein by one skilled in the art without departing from the scope or spirit of the invention. All such changes and modifications are intended to be included within the scope of the invention as defined by the appended claims.

WHAT IS CLAIMED IS:

1. A conversational browsing system, comprising:

a conversational browser comprising:

a command and control interface for converting an input command into a navigation request, the input command comprising a speech command; and

a CML (conversational markup language) processor for parsing and interpreting a CML file, the CML file comprising meta- information representing a conversational user interface for presentation to a user;

a conversational engine for decoding input commands for interpretation by the command and control interface and decoding meta-information provided by the CML processor for generating synthesized audio output; and

a communication stack for transmitting the navigation request to a content server and receiving a CML file from the content server based on the navigation request.

2. The conversational browsing system of claim 1, where the input command comprise multi-modal input.

3. The conversational browsing system of claim 1, further comprising a (CVM) conversational virtual machine, wherein the conversational browser executes on top of the CVM and makes requests for conversational services to the conversational engines through the CVM.

4. The conversational browsing system of claim 3, wherein the conversational browsing system executes in a client device, and wherein the communication stack comprises conversational protocols for registering conversational capabilities of the client device with the content server.

5. The conversational browsing system of claim 3, wherein one of the conversational browser, CVM and conversational engines are distributed over a network.

6. The conversational browsing system of claim 1, wherein the communication stack implements standard networking protocols for transmitting the navigation request and receiving the CML file.

7. The conversational browsing system of claim 6, wherein the standard networking protocols comprise one of TCP/IP (transmission control protocol/internet protocol), HTTP (hypertext transmission protocol), WAP (wireless application protocol), VoIP (voice over internet protocol) and a combination thereof.

8. The conversational browsing system of claim 1, further comprising a transcoder for converting legacy information formats of a content server to a CML file.

9. The conversational browsing system of claim 8, wherein the transcoder accesses and interprets logic information of an application to generate a dialog with a user through the conversational browser.

10. The conversational browsing system of claim 8, wherein the transcoder is executed on one of the conversational browser, a conversational proxy server, the content server, or distributed among a combination thereof.

11. The conversational browsing system of claim 8, wherein the transcoder generates a custom CML file based on conversational capabilities of the machine on which the conversational browsing system executes.

12. The conversational browsing system of claim 1, wherein the CML file comprises one of (1) a page elements that group other CML elements; (2) a body element which specifies output to be spoken by the conversational browser (3) a menu element comprising introductory text that is spoken to a user and a list of choices, each choice having prompt text and being associated with a target address identifying a CML element if the corresponding choice is selected; (4) a form element for inputting at least on item of information and a target address to sent the at least one item of information; and (5) a combination thereof.

13. The conversational browsing system of claim 12, wherein a target address comprises one of a URL (uniform resource locator) and a socket address.

14. A method for providing conversational browsing, comprising the steps of:

receiving an input command, the input command comprising a speech command;

decoding the input command with at least one of a plurality of conversational engines;

generating a navigation request based on the decoded input command for retrieving a CML (conversational markup language), the CML file comprising meta- information representing a conversational user interface for presentation to a user;

transmitting the navigation request and accessing the requested CML file using a standard networking protocol; and

parsing and interpreting meta information comprising the CML file to provide an audio presentation of the information content of the CML file.

15. The method of claim 14, further comprising the step of:

registering conversational capabilities with an entity from which the CML file is accessed; and

customizing the CML file based on the registered capabilities.

16. The method of claim 14, further comprising the step of:

generating a navigation request for accessing a file comprising a legacy information format; and

converting the legacy information format to a CML file.

17. The method of claim 16, wherein the step of converting is performed by a transcoding proxy associated with a content server from which the file is accessed.

18. The method of claim 14, further comprising the steps of accessing and interpreting logic information of an application associated with the CML file to generate a dialog.

19. A program storage device readable by a machine, tangibly embodying a program of instructions executable by the machine to perform method steps for providing conversational browsing, the method steps comprising:

receiving an input command, the input command comprising a speech command;

decoding the input command with at least one of a plurality of conversational engines;

generating a navigation request based on the decoded input command for retrieving a CML (conversational markup language), the CML file comprising meta- information representing a conversational user interface for presentation to a user;

transmitting the navigation request and accessing the requested CML file using a standard networking protocol; and

parsing and interpreting meta information comprising the CML file to provide an audio presentation of the information content of the CML file.

20. The program storage device of claim 19, further comprising instructions for performing the step of:

registering conversational capabilities with an entity from which the CML file is accessed such that the entity can customize the CML file based on the registered capabilities.

21. The program storage device of claim 19, further comprising instructions for performing the step of:

generating a navigation request for accessing a file comprising a legacy information format; and

converting the legacy information format to a CML file.

22. The program storage device of claim 19, further comprising instructions for performing the step of accessing and interpreting logic information of an application associated with the CML file to generate a dialog.

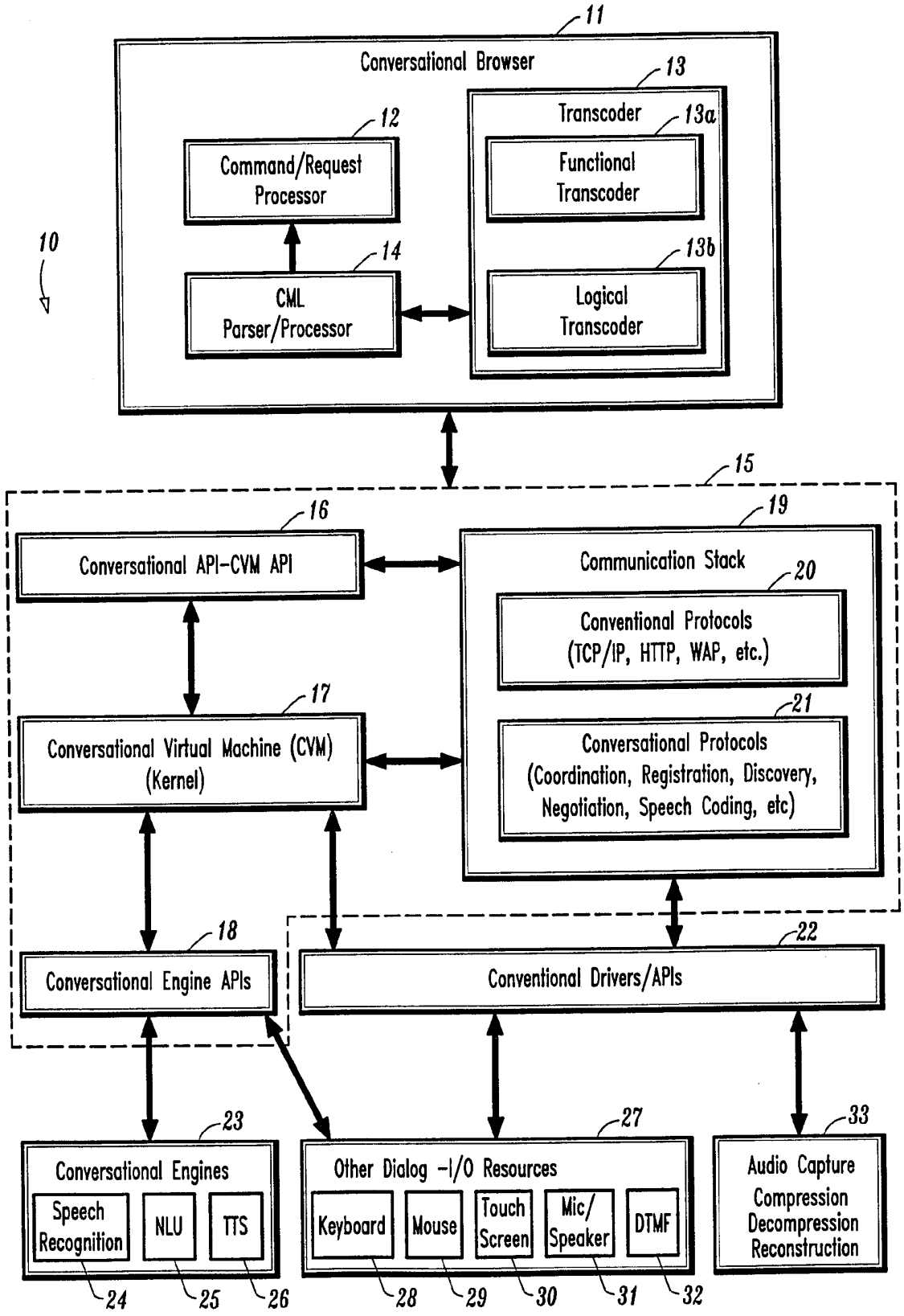


FIG. 1

SUBSTITUTE SHEET (RULE 26)

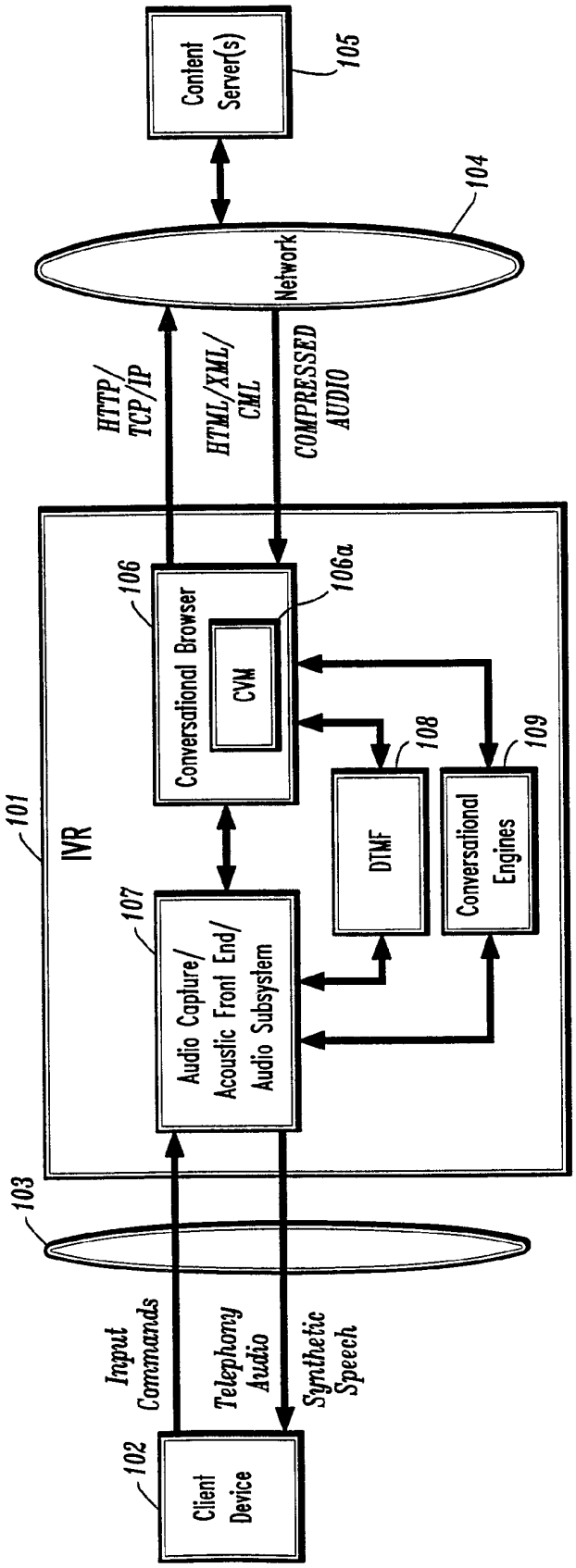

Content Server(s)
105
104
Network
HTTP/ TCP/IP
HTML/XML/ CML
COMPRESSED AUDIO
101
IVR
106
Conversational Browser
CVM
106a
108
DTMF
109
Conversational Engines
107
Audio Capture/ Acoustic Front End/ Audio Subsystem
103
100
Input Commands
Telephony Audio
Synthetic Speech
102
Client Device

FIG. 2

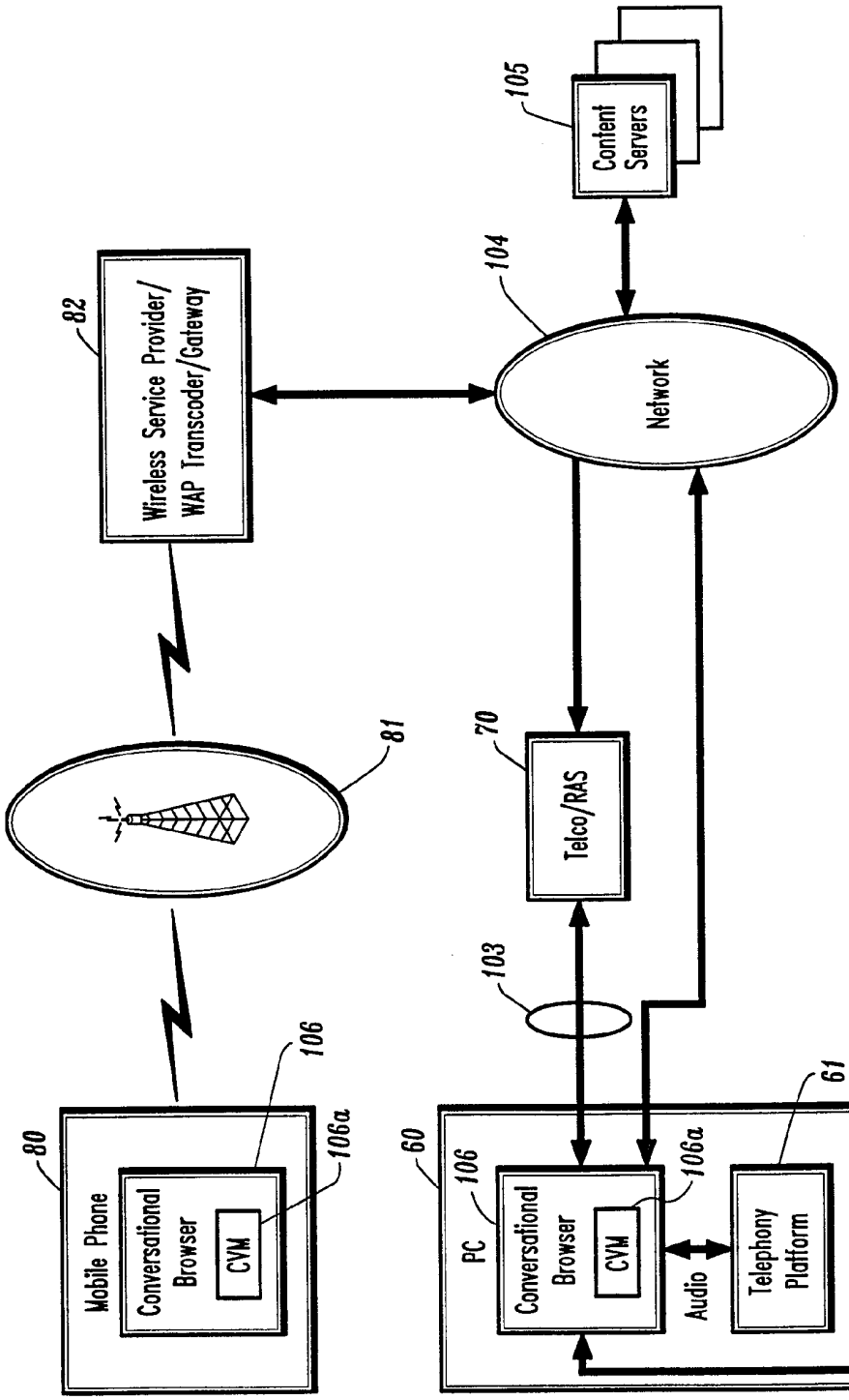

105
Content Servers
104
Network
82
Wireless Service Provider/ WAP Transcoder/Gateway
81
70
Telco/RAS
103
80
Mobile Phone
Conversational Browser
CVM
106
106a
60
PC
Conversational Browser
CVM
Audio
Telephony Platform
Local Application(s)
61
62

FIG. 3

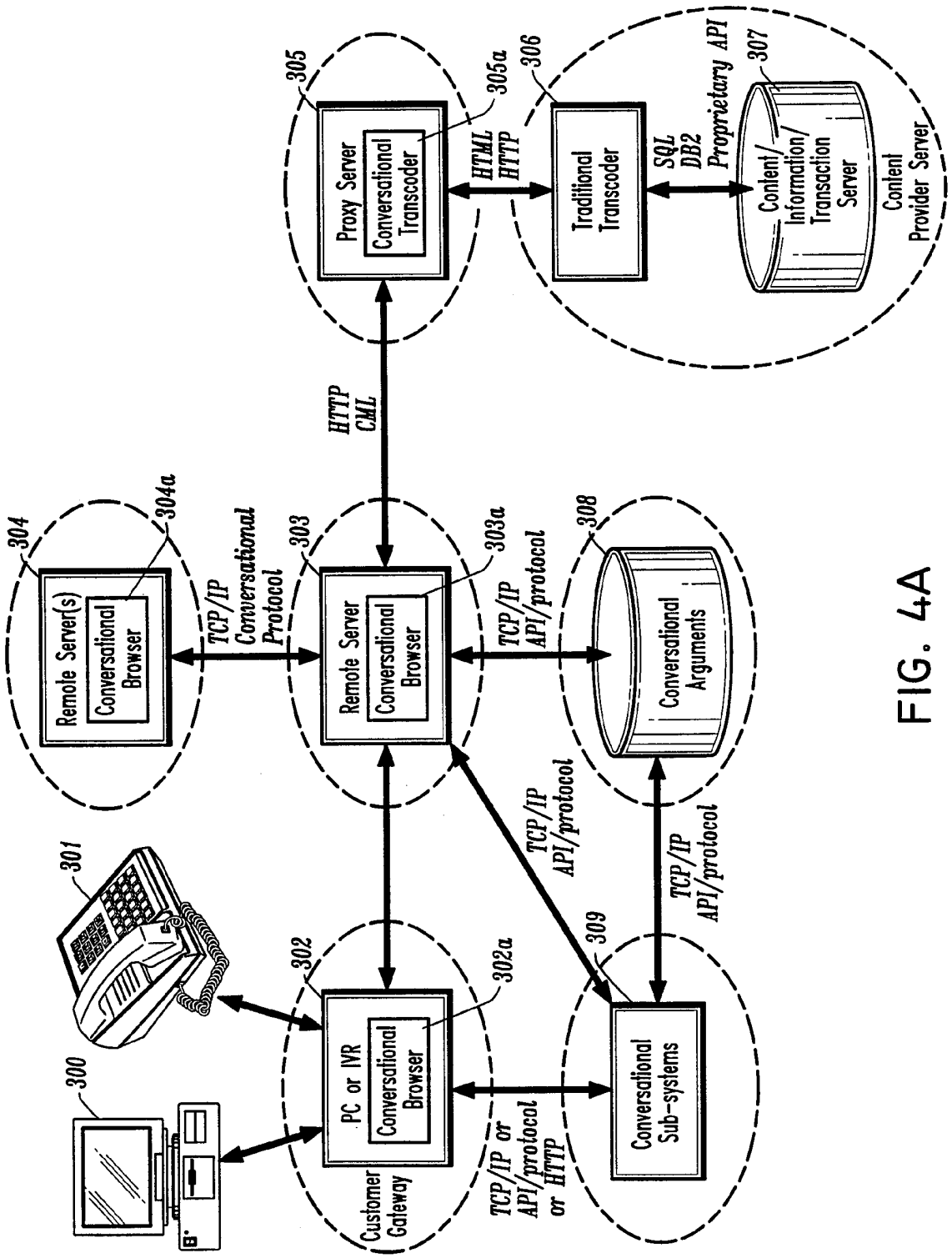

305
Proxy Server
Conversational Transcoder
305a
HTML
HTTP
306
Traditional Transcoder
SQL
DB2
Proprietary API
307
Content/ Information/ Transaction Server
Content Provider Server
HTTP
CML
304
304a
Remote Server(s)
Conversational Browser
TCP/IP
Conversational
Protocol
303
303a
Remote Server
Conversational Browser
TCP/IP
API/protocol
308
Conversational Arguments
TCP/IP
API/protocol
TCP/IP
API/protocol
301
302
302a
PC or IVR
Conversational Browser
Customer Gateway
309
Conversational Sub-systems
TCP/IP or
API/protocol
or HTTP
300

FIG. 4A

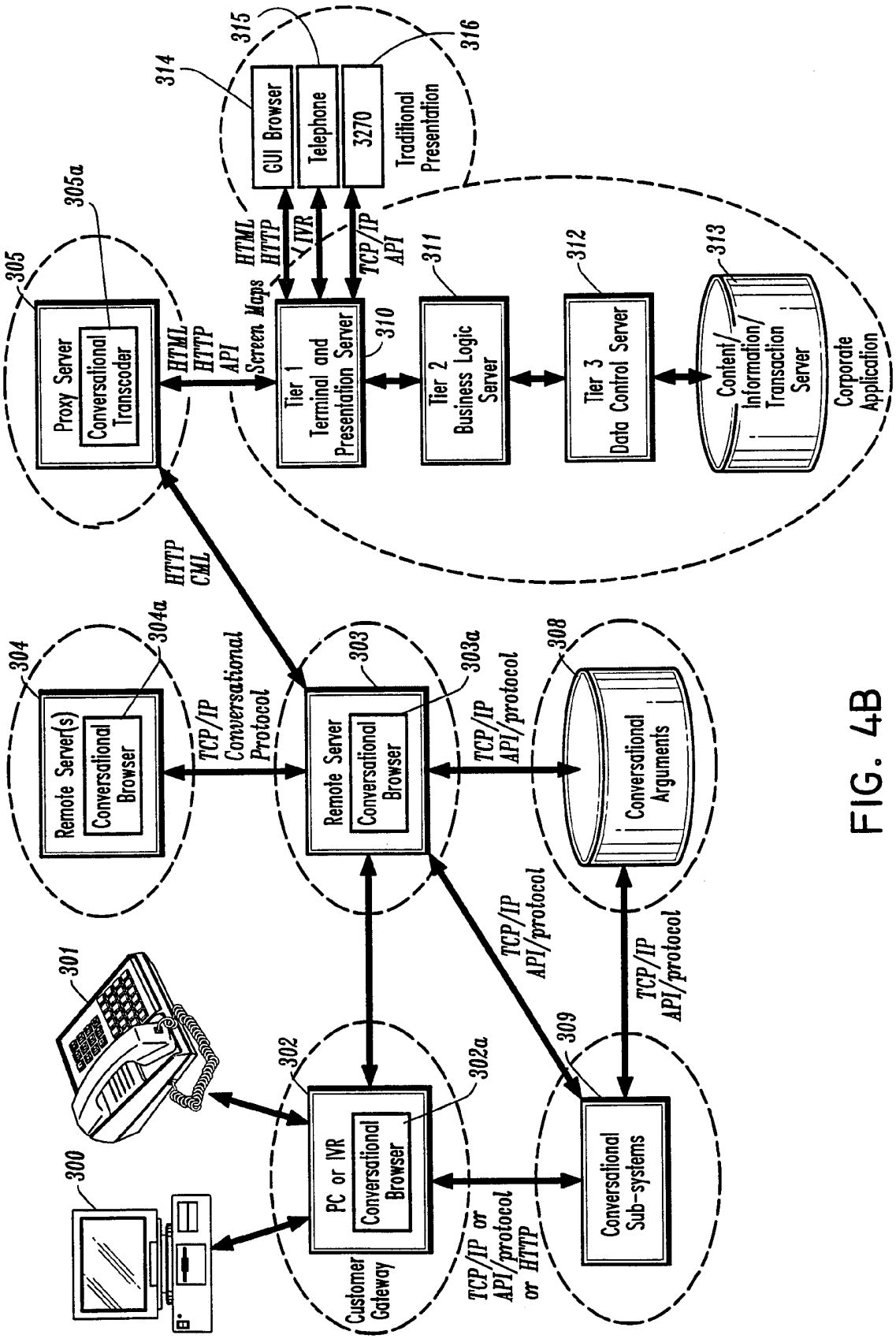

315
316
314
GUI Browser
Telephone
3270
Traditional Presentation
305a
305
Proxy Server
Conversational Transcoder
HTML HTTP API
Screen Maps
HTML HTTP
IVR
TCP/IP API
310
Tier 1 Terminal and Presentation Server
311
Tier 2 Business Logic Server
312
Tier 3 Data Control Server
313
Content/ Information/ Transaction Server
Corporate Application
HTTP CML
304
304a
Remote Server(s)
Conversational Browser
TCP/IP Conversational Protocol
303
303a
Remote Server
Conversational Browser
TCP/IP API/protocol
308
Conversational Arguments
TCP/IP API/protocol
TCP/IP API/protocol
301
302
302a
PC or IVR
Conversational Browser
309
Conversational Sub-systems
300
Customer Gateway
TCP/IP or API/protocol or HTTP

FIG. 4B

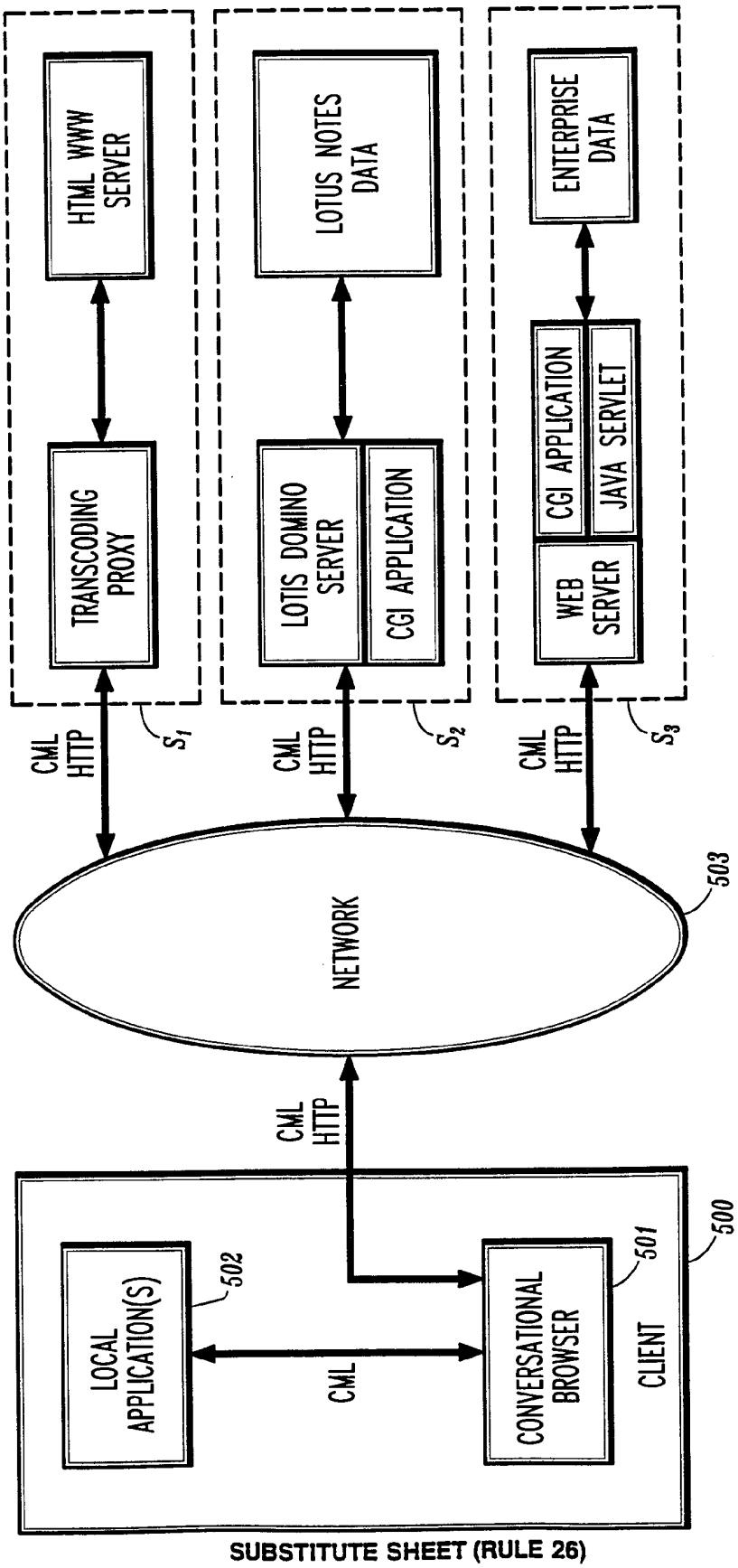

HTML WWW SERVER
TRANSCODING PROXY
LOTUS NOTES DATA
LOTIS DOMINO SERVER
CGI APPLICATION
ENTERPRISE DATA
CGI APPLICATION
JAVA SERVLET
WEB SERVER
CML HTTP
CML HTTP
CML HTTP
S1
S2
S3
NETWORK
503
CML HTTP
LOCAL APPLICATION(S)
502
CML
CONVERSATIONAL BROWSER
501
CLIENT
500

FIG. 5

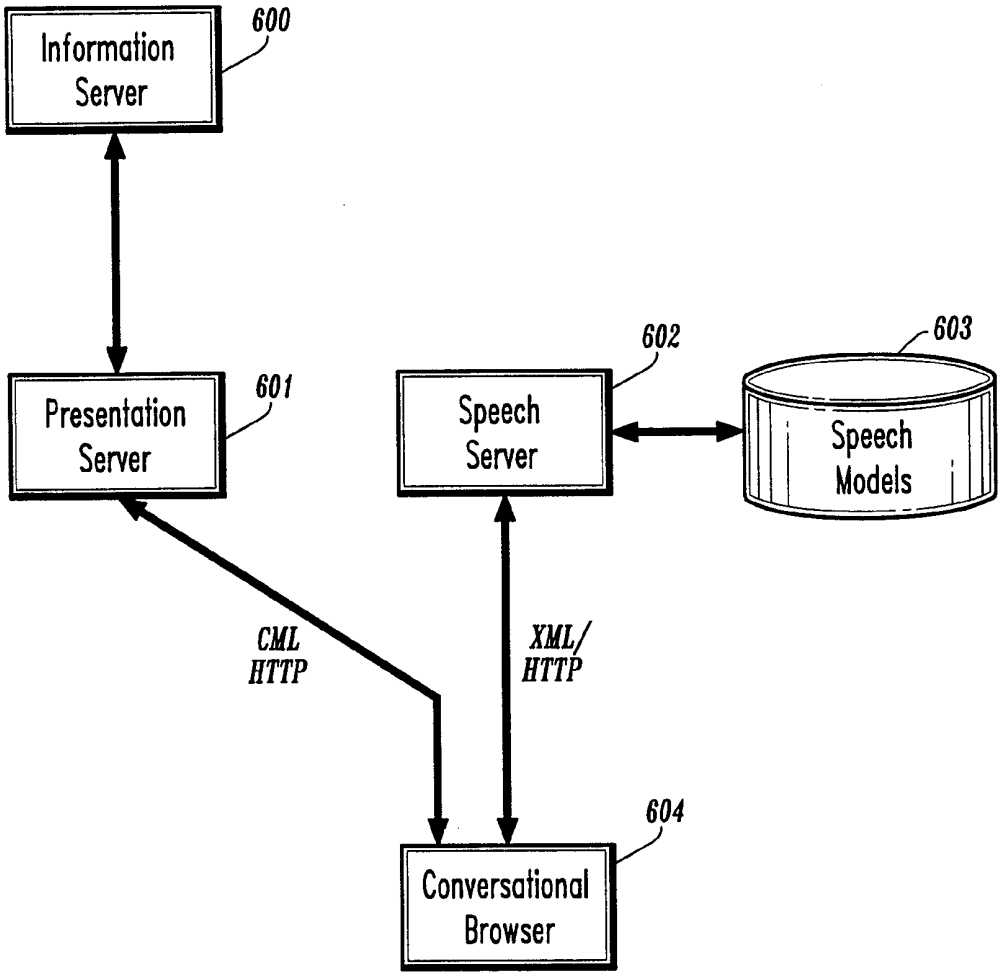

Information Server
600
Presentation Server
601
Speech Server
602
Speech Models
603
CML HTTP
XML/ HTTP
Conversational Browser
604

FIG. 6

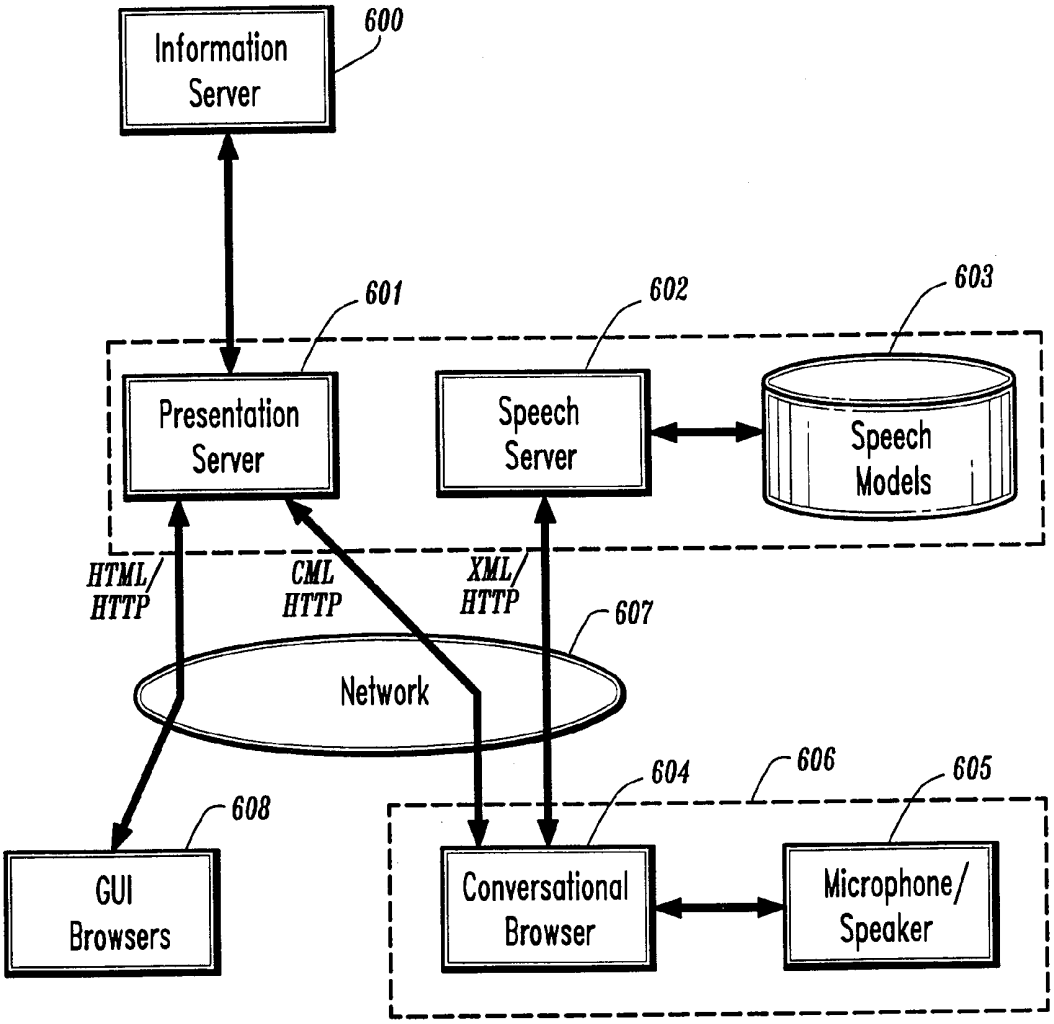

Information Server
600
601
602
603
Presentation Server
Speech Server
Speech Models
HTML/ HTTP
CML HTTP
XML/ HTTP
607
Network
604
606
605
608
GUI Browsers
Conversational Browser
Microphone/ Speaker

FIG. 7

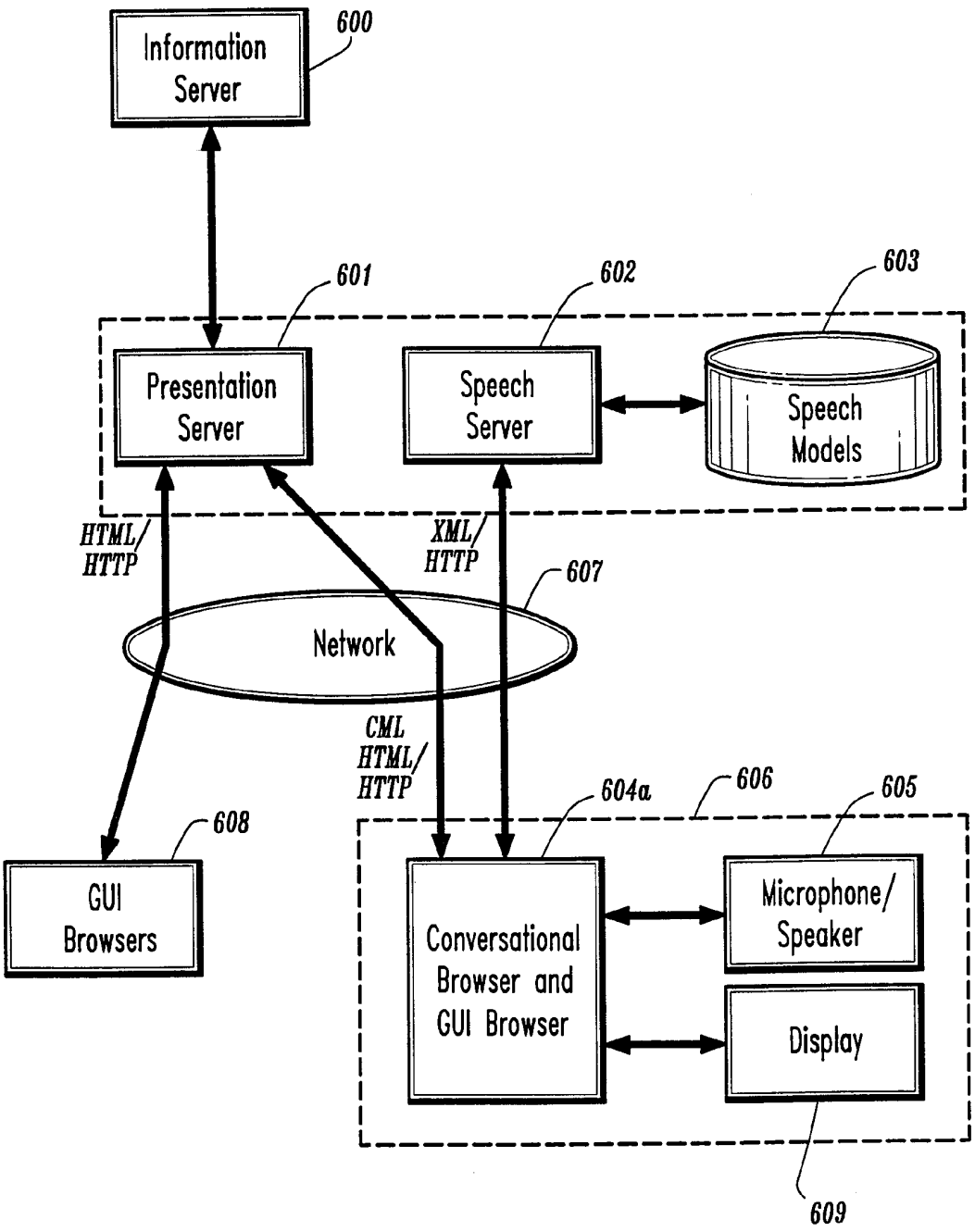

Information Server
600
601
602
603
Presentation Server
Speech Server
Speech Models
HTML/ HTTP
XML/ HTTP
607
Network
CML HTML/ HTTP
604a
606
605
608
GUI Browsers
Conversational Browser and GUI Browser
Microphone/ Speaker
Display
609

FIG. 8

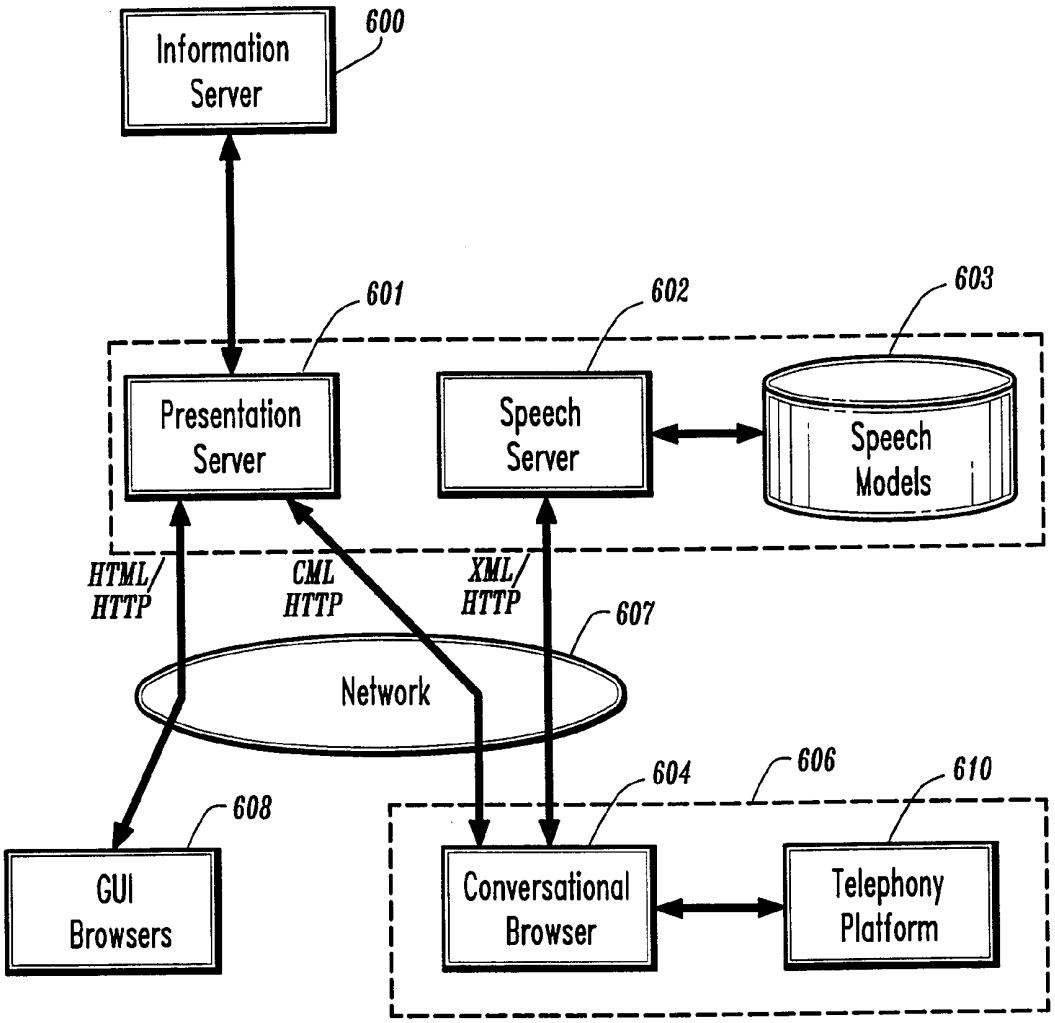

Information Server
600
601
602
603
Presentation Server
Speech Server
Speech Models
HTML/ HTTP
CML HTTP
XML/ HTTP
607
Network
604
606
610
608
GUI Browsers
Conversational Browser
Telephony Platform

FIG. 9

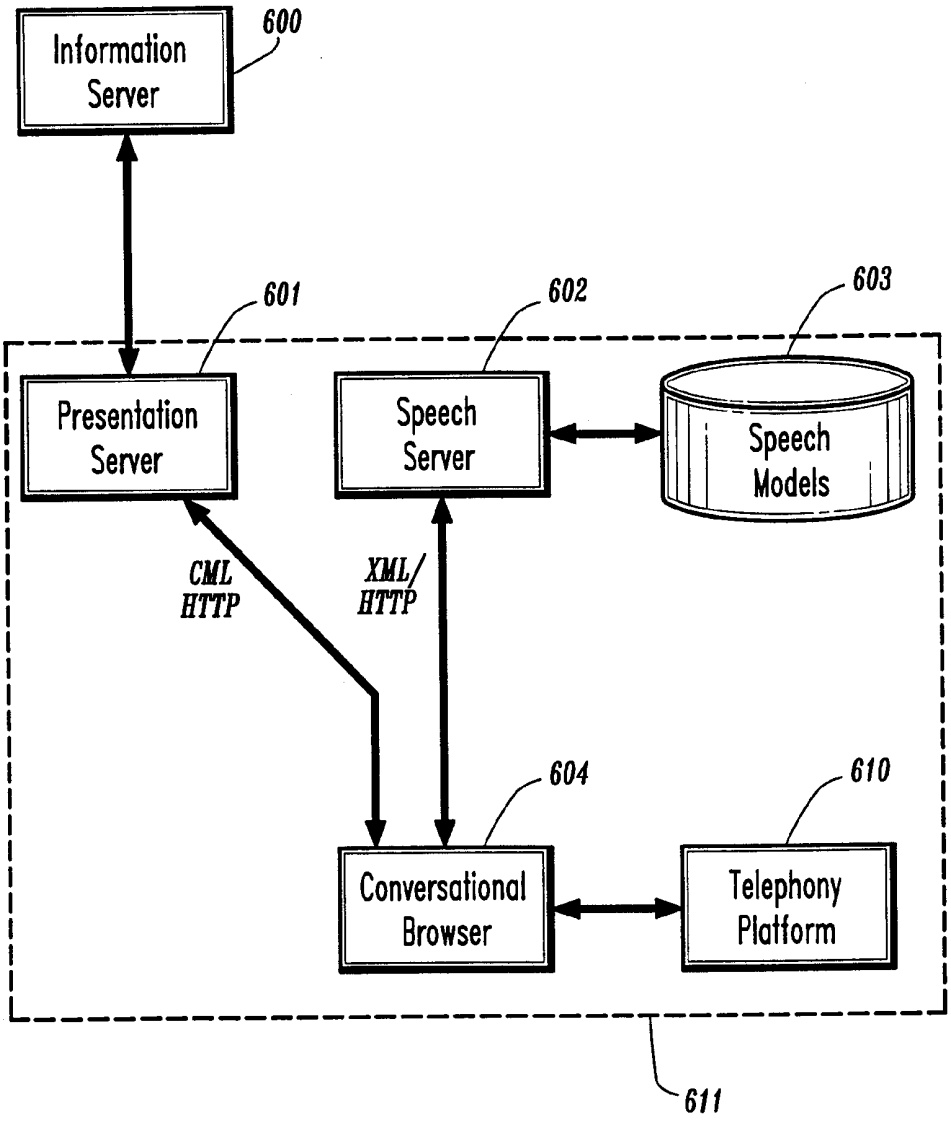

Information Server
600
601
602
603
Presentation Server
Speech Server
Speech Models
CML HTTP
XML/ HTTP
604
610
Conversational Browser
Telephony Platform
611

FIG. 10

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G06F 15/16	A3	(11) International Publication Number: WO 00/21232
		(43) International Publication Date: 13 April 2000 (13.04.00)

(21) International Application Number: PCT/US99/23008

(22) International Filing Date: 1 October 1999 (01.10.99)

(30) Priority Data:

60/102,957	2 October 1998 (02.10.98)	US
60/117,595	27 January 1999 (27.01.99)	US

(71) Applicant *(for all designated States except US)*: INTERNATIONAL BUSINESS MACHINES CORPORATION [US/US]; Old Orchard Road, Armonk, NY 10504 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: GOPALAKRISHNAN, Ponani [IN/US]; 3073 Radcliff Drive, Yorktown Heights, NY 10598 (US). LUCAS, Bruce, D. [US/US]; 2408 Mill Pond Road, Yorktown Heights, NY 10598 (US). MAES, Stephane, H. [BE/US]; 1 Wintergreen Hill Road, Danbury, CT 06811 (US). NAHAMOO, David [IR/US]; 12 Elmwood Road, White Plains, NY 10605 (US). SEDIVY, Jan [CZ/CZ]; U lesa 11, Praha (CZ).

(74) Agent: OTTERSTEDT, Paul, J.; International Business Machines Corporation, Yorktown IP Law Department, T.J. Watson Research Center, Route 134 and Kitchawan Road, Yorktown Heights, NY 10598 (US).

(81) Designated States: CA, CN, IL, IN, JP, KR, US, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

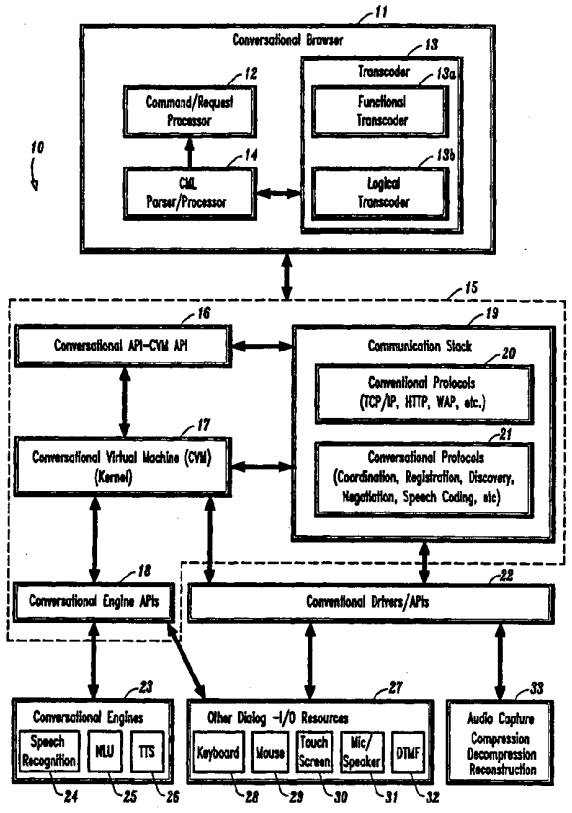
Published
With international search report.

(88) Date of publication of the international search report:
2 November 2000 (02.11.00)

(54) Title: CONVERSATIONAL BROWSER AND CONVERSATIONAL SYSTEMS

(57) Abstract

A conversational browsing system (10) comprising a conversational browser (11) having a command and control interface (12) for converting speech commands or multi-modal input from I/O resources (27) into navigation request. The system (10) comprises conversational engines (23) for decoding input commands for interpretation by the command and control interface and decoding meta-information provided by the CML processor for generating synthesized audio output. The system includes a communication stack (19) for transmitting the navigation request to a content server and receiving a CML file from the content server based on the navigation request. A conversational transcoder (13) transforms presentation material from one modality to a conversational modality. The transcoder (13) includes a functional transcoder (13a) to transform a page of GUI to a page of CUI (conversational user interface) and a logical transcoder (13b) to transform business logic of an application, transaction or site into an acceptable dialog.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/23008

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) : G06F 15/16
US CL : 709/201; 709/219

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 709/201; 709/219

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EAST; WEST; MSDN; DR DOBBS; IEEE; ACM; DIALOG

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y, P	US 5,884,262 A (WISE ET AL) 16 MARCH 1999, col. 5, line 29 - col. 7, line 27.	1-22
Y	US 5,799,063 A (KRANE) 25 AUGUST 1998, all	1-22
Y	ZUE, VICTOR W. et al. "NAVIGATING THE INFORMATION SUPERHIGHWAY USING SPOKEN LANGUAGE INTERFACES" OCTOBER 1995, pages 39-43, especially pg. 41.	1-22
A, E	US 5,987,102 A (ELLIOTT ET AL) 16 NOVEMBER 1999	1, 14, 19
A, P	RASH, WAYNE et al. "HANDS-FREE SURFING", Internetweek, 25 MAY 1998, pg. 1-2, n. 716	1, 14, 19

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 23 MAY 2000	Date of mailing of the international search report 26 JUL 2000
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer ALVIN OBERLEY Joni Hill Telephone No. (703) 305-9716

Form PCT/ISA/210 (second sheet) (July 1998)*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/23008

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	LAU, R., et al. "WEBGALAXY: BEYOND POINT AND CLICK - A CONVERSATIONAL INTERFACE TO A BROWSER", NOVEMBER 1997, V. 29, N. 8-13; abstract	1, 14, 19
A, P	COMPUTING CANADA. "BROWSER TRANSLATES HTML INTO SPEECH, LARGE TYPE", 08 JUNE 1998, V. 24, N. 22; pg. 1.	1, 14, 19

Form PCT/ISA/210 (continuation of second sheet) (July 1998)*

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G10L 15/18

A1

(11) International Publication Number: WO 00/46792

(43) International Publication Date: 10 August 2000 (10.08.00)

(21) International Application Number: PCT/US00/02906

(22) International Filing Date: 4 February 2000 (04.02.00)

(30) Priority Data:

60/118,800	4 February 1999 (04.02.99)	US
09/495,722	1 February 2000 (01.02.00)	US

(71) Applicant *(for all designated States except US)*: SOLILOQUY, INC. [US/US]; 599 Lexington Avenue, New York, NY 10022 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: LUCENTE, Mark [US/US]; Apartment 5C, 204 W. 78th Street, New York, NY 10024 (US). POLISH, Nathaniel [US/US]; 545 West 111th Street, New York, NY 10025 (US).

(74) Agents: WEILD, David, III et al.; Pennie & Edmonds LLP, 1155 Avenue of the Americas, New York, NY 10036 (US).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published

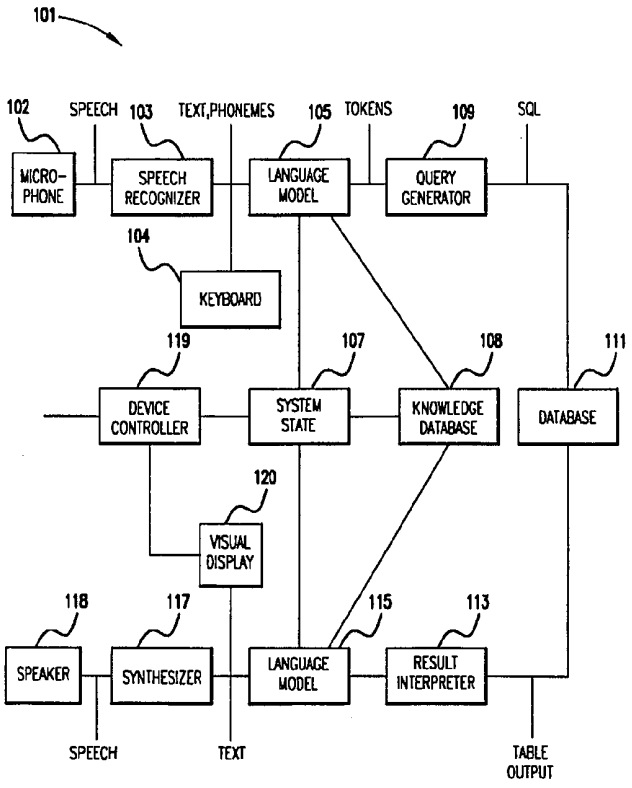
With international search report.

Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: SYSTEM AND METHOD FOR BILATERAL COMMUNICATION BETWEEN A USER AND A SYSTEM

(57) Abstract

A method for a human machine communication system (101) that includes the steps of receiving a first statement in a natural language from a user, generating first information based on the statement and storing context information of at least one of the first statement and the first information. The method also includes the steps of optionally generating a question to be presented to the user in the natural language based on the context information, receiving a second statement in the natural language from the user and generating second information based on the second statement and the context information.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

SYSTEM AND METHOD FOR BILATERAL COMMUNICATION BETWEEN A USER AND A SYSTEM

This application is a continuation-in-part of application serial No. _____ (identified by attorney docket No.10244-003), filed February 1, 2000, which is hereby incorporated by reference in its entirety. Both that application and this application claim priority from provisional patent application serial No. 60/118,800, filed February 4, 1999, which is hereby incorporated by reference in its entirety.

FIELD OF THE INVENTION

The present invention relates to the fields of speech processing, human-machine dialogue, and database retrieval in both client-server and other computer systems.

BACKGROUND OF THE INVENTION

Conventional speech recognition systems are generally unidirectional. Speech recognizers typically receive verbal input from a user and execute commands or produce textual output that could otherwise be effected or created by the user with a keypad, keyboard, or other input device. Speech synthesizers typically simply generate verbal output that otherwise would be available to the user on screens or printouts. Furthermore, conventional speech systems do not fully utilize context information which is an integral part of human-to-human communication.

SUMMARY OF THE INVENTION

Disclosed is a bilateral speech system and method that allows effective communication between a person (a user) and a computer. In a preferred embodiment, the disclosed system comprises a speech recognizer for recognizing spoken input from a user and a language model and query generator for interpreting that input and generating database queries from it. User input may comprise a statement; a question; or an answer to a question posed by the system.

User input is interpreted both on the basis of the input itself and on the basis of the context in which the input was made. This context is developed by storing aspects of a continuing dialogue between the user and the system in memory.

System output may take the form of either information output to the user or actions taken by the system to satisfy user requests. System output is also informed by, and contributes to, the stored context. Information output to the user may comprise a statement; a question; or an answer to a question posed by the user.

In another aspect, the present disclosure includes a method that comprises the steps of receiving a first statement in a natural language from a user, generating first information based on the first statement, and storing context information of at least one of the first statement and the first information.

The method of the present invention further includes the steps of optionally generating a question to be presented to the user in a natural language based on context information, receiving a second statement in the natural language from the user, and generating second information based on the second statement and the context information.

Moreover, the method may include the steps of incorporating content information generated based on the second information into a web page. If desired, the web page may be dynamically generated. The method may also include the steps of generating grammatical data for the first statement, generating one or more parsing tokens based on the grammatical data, and storing the parsing tokens as part of the context information. Further, the parsing tokens can be added, modified, or removed based on the second statement.

The method of the present invention may further include the steps of identifying linguistic structures in the second statement based on the context information. In particular, the method may further include the steps of identifying an antecedent to a pronoun in the second statement or disambiguating a homonym in the second statement.

The method may also include the steps of storing and updating the context information each time a new statement in the natural language is received or providing at least one of a superlative and a comparison of the first information based on the context information. Here, the first information includes a plurality of items that can be compared with each other.

The disclosed system may include a server computer that includes an input device configured to receive a first statement in the natural language from a user. The server computer is configured to generate first information based on the first statement. The server computer also includes a memory bank configured to store context information of at least one of the first statement and the first information. The input device is further configured to receive a second statement in the natural language from the user. In turn, the server computer is further configured to optionally generate a question to be presented to the user in the natural language based on the context information and configured to generate second information based on the second statement and the context information.

The disclosed system may also include a client computer configured to receive a plurality of statements from the user and configured to forward the received plurality of statements to the server computer.

The server computer may include a knowledge database configured to provide language processing information. In particular, the server computer may be

configured to generate a plurality of parsing tokens based on the first statement and the language processing information. The memory bank may further be configured to store the parsing tokens as part of the context information. Furthermore, the server computer may be further configured to add, modify, or remove the stored parsing tokens in the memory bank based on the second statement.

The server computer may be further configured to identify linguistic structures in the second statement based on the context information. In particular, the server computer may be configured to identify an antecedent to a pronoun in the second statement or identify a disambiguating homonym in the second statement.

Moreover, the server computer may be further configured to store and update the context information each time a new statement in the natural language is received by the input device.

The disclosed system may also include an output controller configured to dynamically generate a web page at a client computer based on at least one of the first information and the second information.

The disclosed system may also be used to facilitate a virtual salesman functionality to individuals who visit an e-commerce website.

BRIEF DESCRIPTION OF THE DRAWINGS

Preferred features of the present invention are disclosed in the accompanying drawings, wherein similar reference characters denote similar elements throughout the drawings, and wherein:

FIG. 1A is a block diagram of a preferred embodiment of the disclosed system;

FIG. 1B is a block diagram of various databases of the disclosed system;

FIG. 2 is a diagram illustrating speech recognition in the disclosed system;

FIG. 3 is a diagram illustrating operation of a language model in the disclosed system;

FIG. 4 is a diagram illustrating query generation in the disclosed system;

FIG. 5 is a diagram illustrating results processing in the disclosed system;

FIG. 6 is a diagram illustrating a second language model in the disclosed system;

FIG. 7 is a diagram illustrating speech synthesis in the disclosed system;

FIG. 8 is a diagram illustrating operation of the disclosed system; and

FIG. 9 is a block diagram illustrating a client-server configuration of the disclosed system.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

FIG. 1A illustrates a bilateral communication system 101 adapted to conduct a dialogue with a user and perform tasks that the user requests. As shown in FIG. 1A, system 101 preferably comprises a microphone 102. When a user speaks into microphone 102, the microphone generates an analog speech signal that is transmitted to a speech recognizer 103. Speech recognizer 103 converts the analog speech signal into phonemes in the form of digital text.

The output of speech recognizer 103 is transmitted to a first language model 105 that includes a parser. Language model 105 exchanges grammatical and other data with a system state 107 and a knowledge database 108. Using this data, language model 105 produces parsing tokens that it sends to a database query generator 109.

Database query generator 109 generates queries to a database 111. Database 111 searches data stored therein for the desired information and passes raw information to a result interpreter 113 that interprets the results. A second language model 115 receives the interpreted results from result interpreter 113 and exchanges relevant information with system state 107 and knowledge database 108 to produce a digital text response. Although illustrated in FIG. 1A as distinct components, it should be noted that first and second language models 105, 115 may be implemented as a single language model that performs the functions of first and second language models 105, 115.

The output from language model 115 is transmitted to a synthesizer 117 that converts the response text to analog speech for output to the user via a speaker 118. The components of system 101 described above, with the exception of microphone 102 and speaker 118, may preferably be implemented as software and/or data.

In an alternative preferred embodiment, system 101 may be implemented without microphone 102 and speech recognizer 103. In this embodiment, the user directly types a question or statement on a keyboard 104. In addition, system 101 may be implemented without synthesizer 117 and speaker 118. In that case, output to the user may take the form of text or other information displayed, for example, on a visual display 120.

As discussed above, knowledge database 108 is preferably coupled to first and second language models 105, 115 and also to system state 107. A preferred embodiment of knowledge database 108 is shown in FIG. 1B. As shown in FIG. 1B, knowledge database 108 preferably comprises a general language database 122, a domain specific language database 124, a rules database 126, and a frequently asked question (FAQ) database 128.

General language database 122 preferably stores information relating to conventional natural language understanding such as a dictionary, a list of grammar rules

and/or a synonym knowledge base. The synonym knowledge base is utilized, for example, in correcting misspelled words.

Domain specific language database 124 preferably stores usage specific information. Assume, for example, that system 101 is adapted to respond to queries from a user concerning the time and channel that a particular program will be broadcast. In that event, domain specific language database 124 preferably stores information relating to various television programs including the station that will broadcast them and the time at which they will be broadcast. In a second example, assume that system 101 is adapted to respond to queries from a user concerning computer devices. In that case, domain specific language database 124 preferably stores information relating to computer-related concepts, such as definitions and characteristics of random-access memories (RAMs), read-only memories (ROMs), displays, and central processing units (CPUs).

Rules database 126 preferably stores information directed to general constraints on items stored in the other databases. Thus, if in the first example above one of the programs for which information is stored in knowledge database 108 is Seinfeld, then one illustrative rule might be that the original Seinfeld programs were broadcast on Thursdays. Similarly, if in the second example above one type of computer for which information is stored is laptop computers, then one illustrative rule might be that laptop computers commonly have only one display monitor.

FAQ database 128 stores answers to frequently asked questions. Thus, continuing with the above examples, a frequently asked question in the first example might be: "Where and when is Seinfeld being broadcast this week?", and a frequently asked question in the second example might be "What is RAM?".

In one exemplary embodiment, database 111 may be a product database. For example, it may be a product database for computer devices that includes a list of items such as computers, software applications, computer accessories, memory devices, and display monitors.

System state 107 preferably comprises at least a portion of a memory bank that retains data from first and second language models 105, 115 concerning the context created by both statements made by the user and output from database 111. For instance, when a user asks a first question, system state 107 retains data from language model 105 concerning that question. In a preferred embodiment, the memory bank may be distributed between system state 107 and one or both of language models 105, 115.

After the user receives a response to a first question, the user may wish to ask follow-up questions. Thus, for example, if the user's first question was:

"Where and when is Seinfeld being broadcast this week?",

and the response provided to the user by system 101 was:

"Monday through Friday at 7:30 PM",

then the user might follow that answer with the request:

"Give me a summary of the third and fourth".

In that case, because system state 107 stores the context provided by the user's first question or statement as well as the context provided by the first answer, system 101 is able to respond to this request without the user repeating the facts in the initial question (i.e., that the user is referring to episodes of Seinfeld to be shown this week).

System 101 may then process the follow-up question or request using speech recognizer 103, first language model 105, system state 107, knowledge database 108, query generator 109, database 111, result interpreter 113, second language model 115, and synthesizer 117. System state 107 gives the second question or request a context from which it is understood that the user is referring to the initial inquiry concerning, for example, Seinfeld episodes. The output at speaker 118 will then include the requested summaries.

Continuing with this example, assume that, after receiving the summaries of the third and fourth episodes to be shown that week, the user asks system 101 to:

"Record the third episode".

To respond to that demand, system 101 again processes the input using speech recognizer 103, first language model 105, system state 107, knowledge database 108, query generator 109, database 111, result interpreter 113, second language model 115, and synthesizer 117. Again, system state 107 recollects the context information and permits the system to react to the more specific order from the user without requiring the user to repeat the earlier statements/questions or their factual predicate.

System state 107 may then operate a device controller 119 that takes the database results and the stored context information, and controls a device, such as a video cassette recorder (VCR), web browser, etc., and makes the device perform the action requested by the user. For example, device controller 119 may set a VCR to the time and channel of the episode or episodes to be recorded. At the same time, synthesizer 117 may cause speaker 118 to inform the user:

"Episode 3 set for recording at 11 P.M. Thursday."

Device controller 119 may be adapted to govern any number of devices. For example, if the user wishes to have system 101 browse the web, device controller 119 may control a browser to cause the browser to search for any kind of information and return graphic, text, sounds, pictures, video, multimedia, etc. The result can also be displayed on visual display 120.

Speech recognizer 103 may be any speech recognition engine using computer application programs such as IBM Via Voice™ and Dragon Dictate™ both of which run on standard PC platforms. FIG. 2 illustrates an exemplary embodiment of speech recognizer 103 translating speech into text.

Text from speech recognizer 103 (or that is entered directly by the user), is then processed by first language model 105. In a preferred embodiment, language model 105 may pre-process the text. This pre-processing may include correcting any misspelled words and phrases by matching words and phrases in the text to words and phrases in the various databases of knowledge database 108. The pre-processing may further include expanding any abbreviations in the text.

FIG. 3 illustrates subsequent operation of first language model 105. As shown in Fig. 3, language model 105 first assembles a complete "phrase" in an "Accept Phrase" section 131. Timing cues may provide information as to the end of a phrase. The timing cues can be extracted from stops and pauses when the user speaks to system 101 or when the user types the text using keyboard 104.

In a "Scan Grammar" section 133, language model 105 scans through a set of valid grammars and compares them with each incoming phrase. In a "Pick Grammars" section 135, language model 105 chooses a limited set of grammar candidates from system state 107 to explore based on their weightings. Each of the selected grammars is applied in an "Apply Grammar" section 137. The results of each grammar are evaluated and a single one is chosen based on several possible criteria. The state changes from the final grammar are sent to system state 107. The tokens generated by the final grammar are then sent to query generator 109, as will be described in connection with FIG. 4. Language model 105 can also search for special salient words and phrases in the text.

Turning to FIG. 4, there is shown a preferred embodiment of query generation in the present system. Query generator 109 can retrieve information from domain specific language database 124. Specialized domain processing based on the grammar and the instantiated tokens takes place in a domain specific processing section 141. Processing includes, but is not limited to, the generation of an SQL query from the

tokens in a tokens-to-SQL section 143. The query may then be supplied to an SQL database 145. Examples of suitable database systems are Sybase™ and Oracle™.

If an appropriate answer for the grammar is not found, a fuzzy matching process may be employed to refine the answer, e.g., searching phonetically. The end result is a results table 147 containing the information requested.

As shown in FIG. 5, the contents of results table 147 are sent to result interpreter 113 for additional data mining to produce summarized results. Information summary and consolidation features may also take place.

As shown in FIG. 6, the summarized results arrive at second language model 115. Here the current grammar and conversation history are used to interpret the returned data and generate a formatted response. System state 107 is updated to reflect the new conversation history and updates the grammar weightings for "Pick Grammars" 135 of first language model 105. Updates to system state 107 signal actions for the controlled unit to execute.

As shown in FIG. 7, the text portion of the response is then sent to synthesizer 117. Synthesizer 117 may be a standard speech synthesis engine. Here the text response is converted to speech and played back to the user.

At this point, the user may speak again and the process repeats itself, going from the general to the specific and using system state 107 to remember the user's more general statements and questions and the systems more general statements and answers so as to produce more specific responses.

In this process, speech is received, converted to text, a query is generated, output is created, and another query from the user follows. The result is typically a loop in which the dialogue progresses from general to specific.

Take the following third example. Assume that a user commences a dialogue with system 101 by saying:

"I'm looking for a song with particular lyrics A."

System 101 may then identify the song and output the song's name to the user, in the manner described above. Assume further that the user next says:

"Can I see the lyrics?".

In that event, system 101 may retrieve the lyrics from database 111 because system state 107 has memorized the first question and all the data connected therewith.

As noted, system 101 preserves the context that allows system 101 to respond to the second question. Otherwise, the user would have to say:

"Can I see the lyrics for song X?".

More particularly, because it preserves context, system 101 is able to handle successive increasingly-specific questions. Take the following fourth example. Assume that the user's first statement to system 101 is:

"I'm looking for a Beatles album, the forty-fourth album by the Beatles."

System 101 then provides the name of the forty fourth album, in the manner described above. If the user then says:

"What songs are on that particular album?",

and there are 20 songs on the album, system 101 may retrieve all the song titles from memory and output them to the user because it comprises a memory that retains the place in the conversation.

The memory bank may also store the tokens generated by language model 105 as part of the context. These tokens represent requests and constraints detected by language model 105, both current and previous. These tokens are updated and/or modified by both first and second language models 105, 115 based on the statements and/or questions by the user and the database query results. In a preferred embodiment, these updates may take the form of additions, modifications, or removals.

Take the following fifth example. Assume that at some point in a dialogue with system 101, the user states:

"I am looking for a computer less than $3000".

In that event, an addition occurs, i.e., system 101 adds a price token to the context. If, however, the user subsequently states:

"I want to spend less than $2500",

then a modification occurs, i.e., system 101 modifies the price token to reflect the user's latest statement. Finally, if the user subsequently states that:

"price is not important",

then a removal occurs, i.e., system 101 removes the price token from context.

Since all non-trivial dialogs have a context, the memory bank is preferably configured to continually store the context of the dialog between the user and system 101. The context, built during processing of statements and/or questions entered by the user and the database query results, contains linguistic information, such as a list of possible grammars and weightings thereof. Certain processing is enabled through use of context including the processing of superlatives and comparisons. Take the following sixth example. Assume that a user states at some point during a dialogue with system 101 that:

"I would like something that costs less than that".

In that event, system state 107 may look to the context to see what price range is currently in context. Alternatively, in a seventh example, assume that the user states:

"I would like the the cheapest one".

In that event, the desired item may by identified by checking the list of results posted in the context during a previous output to the user.

The context may also contain information that allows language models 105, 115 to identify or define linguistic structures, such as antecedents for pronouns (e.g., if the user refers to "these" or "it") by utilizing histories of the database query results, histories of statements and/or questions entered by the user, histories of outputs generated, the nature and motivation of such outputs (i.e., answer, prompt, request for clarification), and context at the time of earlier processing by language models 105, 115.

Linguistic information may also be used to disambiguate user statements, such as those containing homonyms (for speech recognition or misspelled words), or references to earlier identified items (e.g., "the second one"). This information may also be used to post expectations in the form of probabilities or weightings to help language models 105, 115 to better process words, phrases, and the occurrence of features or units.

Take the following eighth example. If at some point during a dialogue with a user, system 101 output the question:

"how much would you like to spend",

system 101 may post into context that price is a likely feature to be input next. Then, when the user inputs a response such as:

"2000",

that response may be correctly interpreted as a price parameter. Alternatively, in a ninth example, assume that at some point in a dialogue with the user, system 101 output the question:

"what do you expect to use this computer for?".

In that event, system 101 may post an expectation that an activity will likely be input next. A response from the user of "gaming" may then be properly interpreted grammatically and semantically.

The context may also contain information that is characteristic of a continuing dialogue between the user and system 101. For example, an unfulfilled request, and the number of times such a request has remained unfulfilled, may be utilized to guide language models 105, 115 to generate a helpful statement and/or a helpful question to be presented to the user. Other examples include recommendations made that are close to the desired query, changes in requirements (e.g., "I decided that I don't care about price"), and events when the user failed to respond to a prompt or suggestion. Furthermore, context may contain information gathered about the user, such as preferences. This profiling information can be stored for interactions at a later date.

The context may be updated with each exchange (input and response) of the conversation. The context stores previously detected constraints and requests that are combined with the latest verbal input to create a progressively more refined database query.

Since the context is utilized, output of system 101 may be in the form of the requested information, an indication that the information is not available, a request for clarification, a recommendation, and/or a prompt for more information. Returning to the first example, the reply to the user's initial question concerning Seinfeld episodes may be:

"There are five Seinfeld episodes this week, the first on Sunday at 8:00 PM on channel 110, the second on Wednesday, etc."

rather than simply stating:

"Five episodes are available".

System state 107 then retains data concerning the answer.

In a ninth example, if a request is made for computers that are:

"good for running spreadsheets",

and 3 such computers are found, the system output may be:

"all 3 are good for spreadsheets"

rather than simply:

"here are 3 that you asked for."

In addition, a combination of the user statements and/or questions, the summarized results, and context may be used to control device controller 119.

As noted above, the present invention furnishes database access, in particular access to a relational database accessed via SQL. In a preferred embodiment, system 101 assumes that the user's goal is always to form an SQL query. The language model assisted by the context offered by the working memory forms the basis for the query. In a tenth example, if the user asks for:

"a list of Beatles albums"

then that request is stored in working memory. Therefore if the user subsequently asks for:

"albums from 1969",

the query will be generated based on Beatles albums from 1969.

FIG. 8 is a flow chart depicting operation of system 101. As shown in FIG. 8, in step 181 the user enters a statement or question into a speech receiver such as microphone 102 that may be part of speech recognizer 103. In step 183 speech recognizer 103 converts the speech into phonemes in the form of digital text. In step 185, first language model 105 that includes a parser receives the digital text from speech recognizer

103. "Accept Phrase" section 131 of first language model 105 assembles a complete "phrase." A complete phrase may, for example, be determined by a delay greater than a specific amount of time. Still in step 185, "Pick Grammars" section 135 scans a limited set of grammar candidates to explore in system state 107 based on their weightings.

In step 187, grammars are chosen and applied by "Apply Grammar" section 137. The results of each grammar are evaluated and a single one is chosen based on which is the best fit. In step 189, the grammar is executed and state changes from the final grammar are sent to system state 107 for memorization. In step 191, first language model 105 generates tokens on the basis of the final grammar. In step 193, first language model 105 sends the tokens to query generator 109. In step 195, query generator 109 generates an SQL query from tokens in tokens-to-SQL section 143. In step 197, database 111 is searched to form result table 147 containing the information requested.

In step 199, result interpreter 113 receives table 147 for additional data mining to interpret the results. In step 201, second language model 115 uses the interpreted results and the conversation history from system state 107. Specifically, the current grammar and conversation history are used to interpret the returned data and generate a formatted response. In step 203, system state 107 is updated to reflect the new conversation history and updates the grammar weightings for "Pick Grammars" 135 phase of first language model 105.

If the conversation is complete, in step 205, system state 107 signals control of actions, if any, for the controlled unit to execute. In step 207, synthesizer 117 synthesizes the signals to speech. That is, the text portion of the response is then sent to a standard speech synthesis engine where it is converted to speech and played back to the user. The text to speech synthesizer may be in the form of that available under the trademark Accuvoice.

Step 209 involves return to the speech query step 181 for further specific queries. Here, the user speaks again in step 181, and text is converted in step 183 and the process repeats itself, going from the general to the specific and using system state107 to remember the more general questions so as to produce the more specific answers. In steps189 and 191, the grammar is executed and tokens generated depending upon the conversation stored in system state 107. Also, the conversation history and grammar weightings in system state 107 are updated. This varies depending on the grammar chosen.

FIG. 9 illustrates a preferred client-server environment 301 in which the present invention may be implemented. In particular, client-server environment 301 includes a number of client computers 303 connected to one or more server computers 305 via a network 307. Network 307 may use the Internet communications protocols (TCP/IP) to allow client computers 303 to communicate with server computers 305. Client-server

environment 301 further includes one or more database management systems 309 which may preferably be connected to one of server computers 305. However, database management system 309 may alternatively be located anywhere on network 307 as long as it can be accessed and/or controlled by server computers 305. Database management system 309 is preferably configured to store and manage knowledge database 108 and database 111.

Client computers 303 may be personal computers (PCs), workstations or other computer processing units. Each client computer 303 typically includes one or more processors, memories, and input/output devices. Server computers 305 may be similarly configured. However, in many instances server computers 305 may include one or more high performance computers (e.g., workstations) locally networked together. In fact, network 307 may include hundreds of thousands of individual networks of computers, e.g., the Internet. Although client computers 303 are shown separate from server computers 305, it should be understood that a single computer can perform the client and server roles as well.

In operation, the user at one of client computers 303 desiring to communicate with server computers 305 may use a user interface at client computers 303. In the preferred embodiment, the user interface may be a World-Wide-Web (WWW) browser (or in short a web browser). A web browser may display plain textual information or more complex digitally encoded multimedia content, such as software programs, graphics, audio signals, videos, and etc. Web pages processed by a web browser may be implemented in a variety of computer languages such as using a standard page description language known as Hypertext Markup Language (HTML), Extensible Markup Language (XML) or JAVA™, or any other software language or tool for displaying or otherwise outputting information. For instance, HTML provides basic document formatting and allows the developer to specify "links" to other web pages. Using an HTML-compatible browser (e.g., Netscape Navigator™ or Microsoft Internet Explorer™) at a client computer, the user can make a request to the server (sometimes referred to as a "Web site") identified in the link and, in return, receive a document or other object formatted according to HTML. It should be noted that the request by the user can be in the form of voice statements/questions or text inputs as discussed above.

It should be understood that network 307 can also be another wide or local area network (WAN or LAN) and may employ communications protocols other than TCP/IP.

Referring back to FIG. 1, the various components of system 101 discussed above may reside at a client computer and/or a server computer. For instance, in one exemplary embodiment, microphone 102 and speaker 118 may reside in a client computer, and the remaining components described in FIG. 1 in a server computer. In this exemplary

embodiment, a statement uttered by a user at a client computer is preferably converted to an analog speech signal. The analog speech signal is then forwarded to speech recognizer 103 which can be located either at a client computer or in a server computer. The subsequent processing is performed by the server computer in which the rest of the components reside. After processing the statement as discussed above in connection with FIGs. 1-8, the output generated by the server is then forwarded to the client computer in which speaker 118 resides. It should be noted that the client computer in which microphone 102 resides can be different from the client computer in which speaker 118 resides.

Further, as discussed above, device controller 119 may control a web browser. More specifically, the output generated by the server may include text, lists, pictures, graphics, and other objects that can be output to a user. Therefore, device controller 119 is further configured to dynamically generate, for example, one or more web pages on the client computer by automatically arranging the output into one or more web pages as known in the art.

In another exemplary embodiment, the input can be text input. In this embodiment, speech related components are not provided.

In a preferred embodiment, system 101 may be adapted to provide a virtual salesman functionality in an e-commerce environment. In particular, system 101 may be located at an e-commerce website comprising, for example, a plurality of static, dynamic, and/or active webpages comprising information concerning items or services for sale or rental and additional software to facilitate the sale or rental of the items or services to customers who visit the website. This additional software may, for example, comprise payment software to permit a user to pay for an item or service with, for example, the user's credit card.

In this preferred embodiment, system 101 may be adapted to engage in a sales dialogue. Thus, for example, the system could facilitate the following dialogue:

User input #1: URL of e-commerce site

System response #1: Hi. Welcome to e-commerce.com. What would you like to buy today?

User input #2: I'd like to buy a computer.

System response #2: What kind of computer do you want?

User input #3: I'm not sure.

System response #3: Well, how are you going to use the computer?

User input #3: Well, I travel a fair amount and I like to play computer games.

System response #4: Sounds like you could use a laptop with a good size hard drive.

User input #5: What's a hard drive?

System Response #6: It is one type of memory in a computer. It is typically the biggest memory in the computer and is used to store large programs and chunks of data. Let me show you some laptops that might be good for you.
[output dynamic web page displaying laptops with large hard drives]

User input #7: Those are too expensive!

System response #7: OK, let me see if we've got some that are a little more in your price range. How much were you looking to spend?

User input #8: About $2500.

System response #8: OK, here are some laptops with good size memories in that range.
[output dynamic webpage with reasonably priced laptops with relatively large hard drives]

User input #9: That third one looks good.

System response #9: Would you like to buy it?

User input #10: Yes.

System response #10: How would you like to pay for it?

User input #11: Credit card, please.

System response #11: What kind of card is it?

User input #12: Visa

System response #12: What is the card number?

User input #13: 5555 5555 5555 5555

System response #13: Expiration date?

User input #14: ten-oh-two

System response #14: OK, one second while I confirm the transaction . . . . OK, the transaction is confirmed. How do you want us to deliver it?

User input #15: Can you send it UPS?

System response #15: Sure. What is your address?

User input #16: 555 Walnut Street, Anywhere USA.

System response #16: OK, we will send your computer out today, you should get it tomorrow.

Thus, system 101 may be utilized to provide a virtual salesman functionality to individuals who visit an e-commerce website.

While embodiments of the present invention have been described in detail it will be evident to those skilled in the art that the present invention may be embodied otherwise without departing from its spirit and scope.

CLAIMS

What is claimed is:

1. A human-machine communication method comprising:

receiving a first statement in a natural language from a user;

generating first information based on the first statement;

storing context information of at least one of the first statement and the first information;

optionally generating a question to be presented to the user in the natural language based on the context information;

receiving a second statement in the natural language from the user; and

generating second information based on the second statement and the context information.

2. The method of claim 1 further comprising:

incorporating content information generated based on the second information into a web page.

3. The method of claim 2 further comprising:

dynamically generating the web page based on the content information.

4. The method of claim 1 further comprising:

generating grammatical data for the first statement;

generating one or more parsing tokens based on the grammatical data; and

storing the parsing tokens as part of the context information.

5. The method of claim 4 further comprising:

adding, modifying or removing the stored parsing tokens based on the second statement.

6. The method of claim 1 further comprising:

identifying linguistic structures in the second statement based on the context information.

7. The method of claim 6 further comprising:

identifying an antecedent to a pronoun in the second statement.

8. The method of claim 7 further comprising:

disambiguating homonym in the second statement.

9. The method of claim 1 further comprising:

storing and updating the context information each time a new statement in the natural language is received.

10. The method of claim 1 further comprising:

providing at least one of a superlative and a comparison of the first information based on the context information, wherein the first information includes a plurality of items that can be compared with each other.

11. A human-machine communication system comprising:

a server computer comprising:

an input device configured to receive a first statement in a natural language from a user, wherein the server computer is configured to generate first information based on the first statement;

a memory bank configured to store context information of at least one of the first statement and the first information;

the input device further configured to receive a second statement in the natural language from the user, wherein the server computer is further configured to optionally generate a question to be presented to the user in the natural language based on the context information and configured to generate second information based on the second statement and the context information.

12. The system of claim 11 further comprising:

a client computer configured to receive a plurality of statements from the user and configured to forward the received plurality of statements to the server computer.

13. The system of claim 11 wherein the server computer further comprises:

a knowledge database configured to provide language processing information,

wherein the server computer is further configured to generate a plurality of parsing tokens based on the first statement and the language processing information, and

wherein the memory bank is further configured to store the parsing tokens as part of the context information.

14. The system of claim 13 wherein the server computer is further configured to add, modify or remove the stored parsing tokens in the memory bank based on the second statement.

15. The system of claim 13 wherein the server is further configured to generate the plurality of parsing tokens based on, in part, timing cues of the first statement.

16. The system of claim 11 wherein the server computer is further configured to identify linguistic structures in the second statement based on the context information.

17. The system of claim 16 wherein the server computer is further configured to identify an antecedent to a pronoun in the second statement.

18. The system of claim 16 wherein the server computer is further configured to identify a disambiguating homonym in the second statement.

19. The system of claim 11 wherein the server computer is further configured to store and update the context information each time a new statement in the natural language is received by the input device.

20. The system of claim 11 wherein the server computer is further configured to provide at least one of a superlative and a comparison of the first information based on the context information, wherein the first information includes a plurality of items that can be compared with each other.

21. The system of claim 11 wherein the server computer further comprises:

an output controller configured to dynamically generate a web page at a client computer based on at least one of the first information and the second information.

22. A human-machine communication method, comprising:

receiving a first statement in a natural language from a user;

generating first information based on the first statement;

storing context information in a system state memory that is a function of the first information;

generating a database query that is a function of the first information;

generating database output that is a function of the database query;

generating second information that is a function of the database output and the stored context information;

storing an expectation in the system state memory that is representative of an expected form of a statement from the user;

receiving a second statement in a natural language from a user; and

utilizing the expectation to comprehend the second statement.

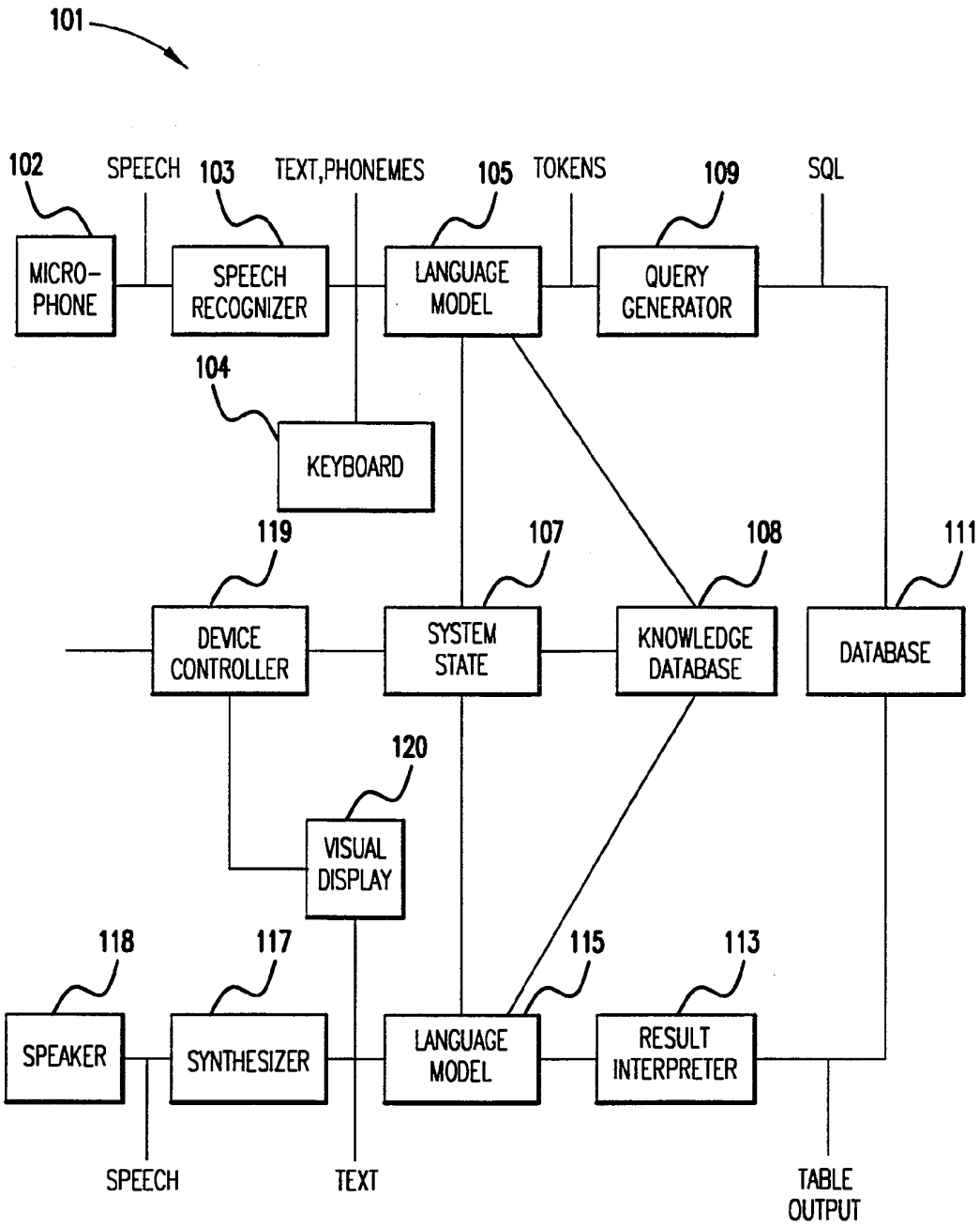

101
102
SPEECH
103
TEXT,PHONEMES
105
TOKENS
109
SQL
MICRO-PHONE
SPEECH RECOGNIZER
LANGUAGE MODEL
QUERY GENERATOR
104
KEYBOARD
119
107
108
111
DEVICE CONTROLLER
SYSTEM STATE
KNOWLEDGE DATABASE
DATABASE
120
VISUAL DISPLAY
118
117
115
113
SPEAKER
SYNTHESIZER
LANGUAGE MODEL
RESULT INTERPRETER
SPEECH
TEXT
TABLE OUTPUT

FIG.1A

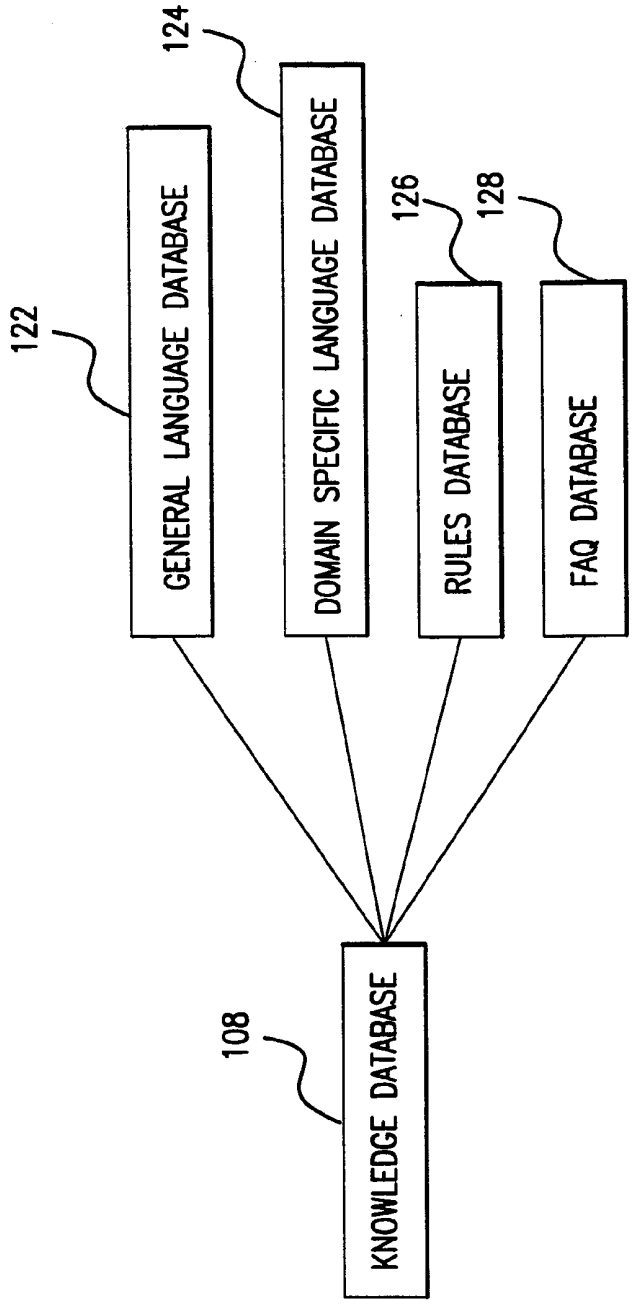

124
GENERAL LANGUAGE DATABASE
DOMAIN SPECIFIC LANGUAGE DATABASE
126
128
122
RULES DATABASE
FAQ DATABASE
108
KNOWLEDGE DATABASE

FIG.1B

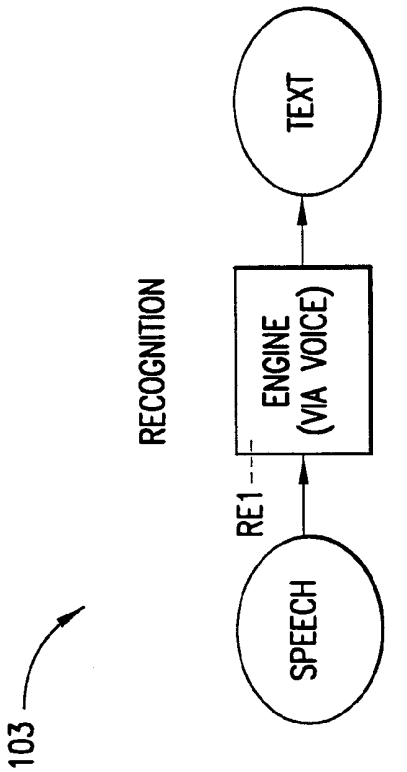


FIG.2

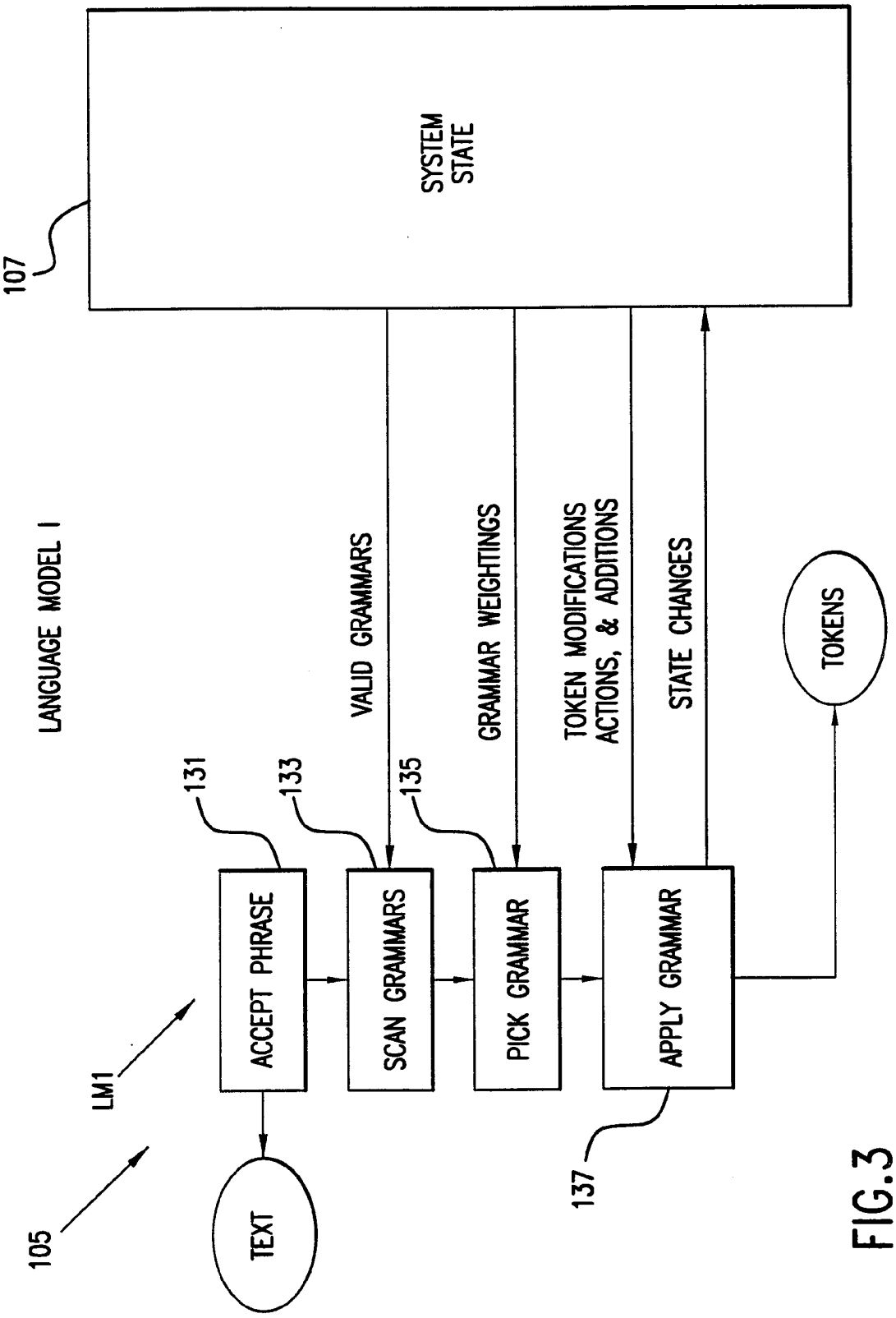

SYSTEM STATE
107
LANGUAGE MODEL I
VALID GRAMMARS
GRAMMAR WEIGHTINGS
TOKEN MODIFICATIONS ACTIONS, & ADDITIONS
STATE CHANGES
TOKENS
131
133
135
ACCEPT PHRASE
SCAN GRAMMARS
PICK GRAMMAR
APPLY GRAMMAR
LM1
105
TEXT
137

FIG.3

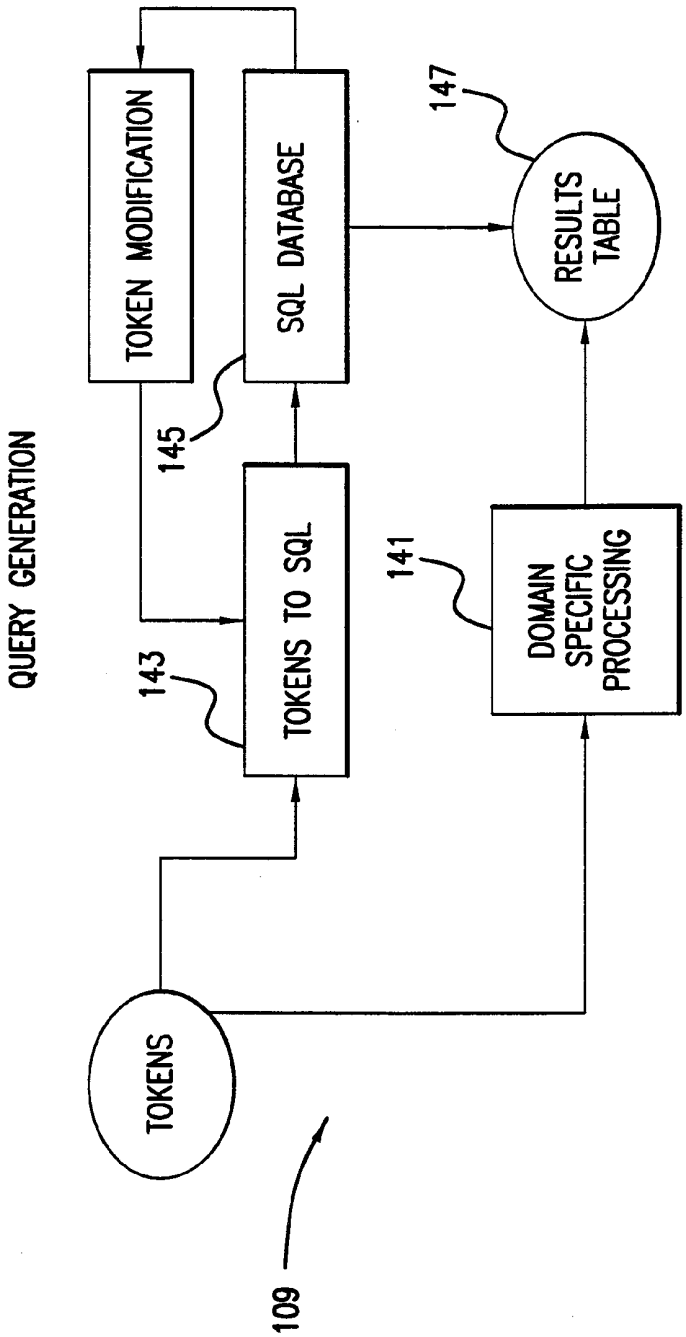

QUERY GENERATION
TOKEN MODIFICATION
SQL DATABASE
RESULTS TABLE
TOKENS TO SQL
DOMAIN SPECIFIC PROCESSING
TOKENS
109
141
143
145
147

FIG.4

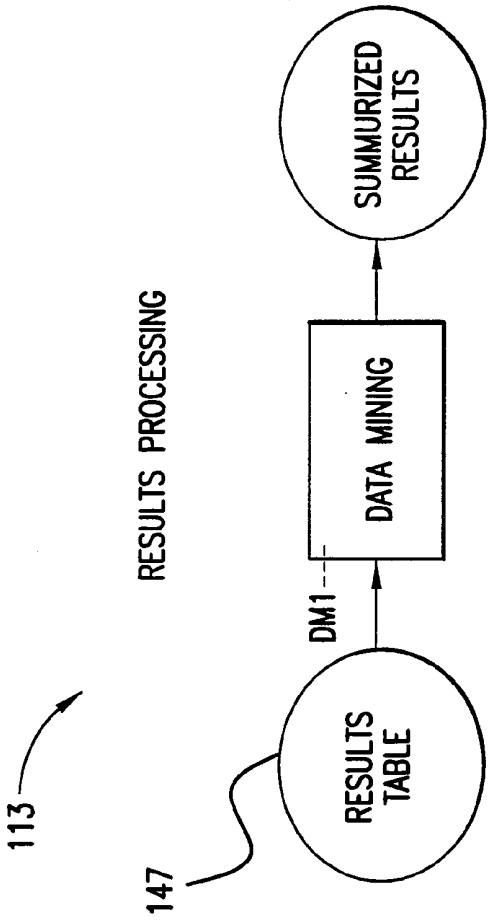

SUMMURIZED RESULTS
DATA MINING
DM1
RESULTS TABLE
RESULTS PROCESSING
113
147

FIG.5

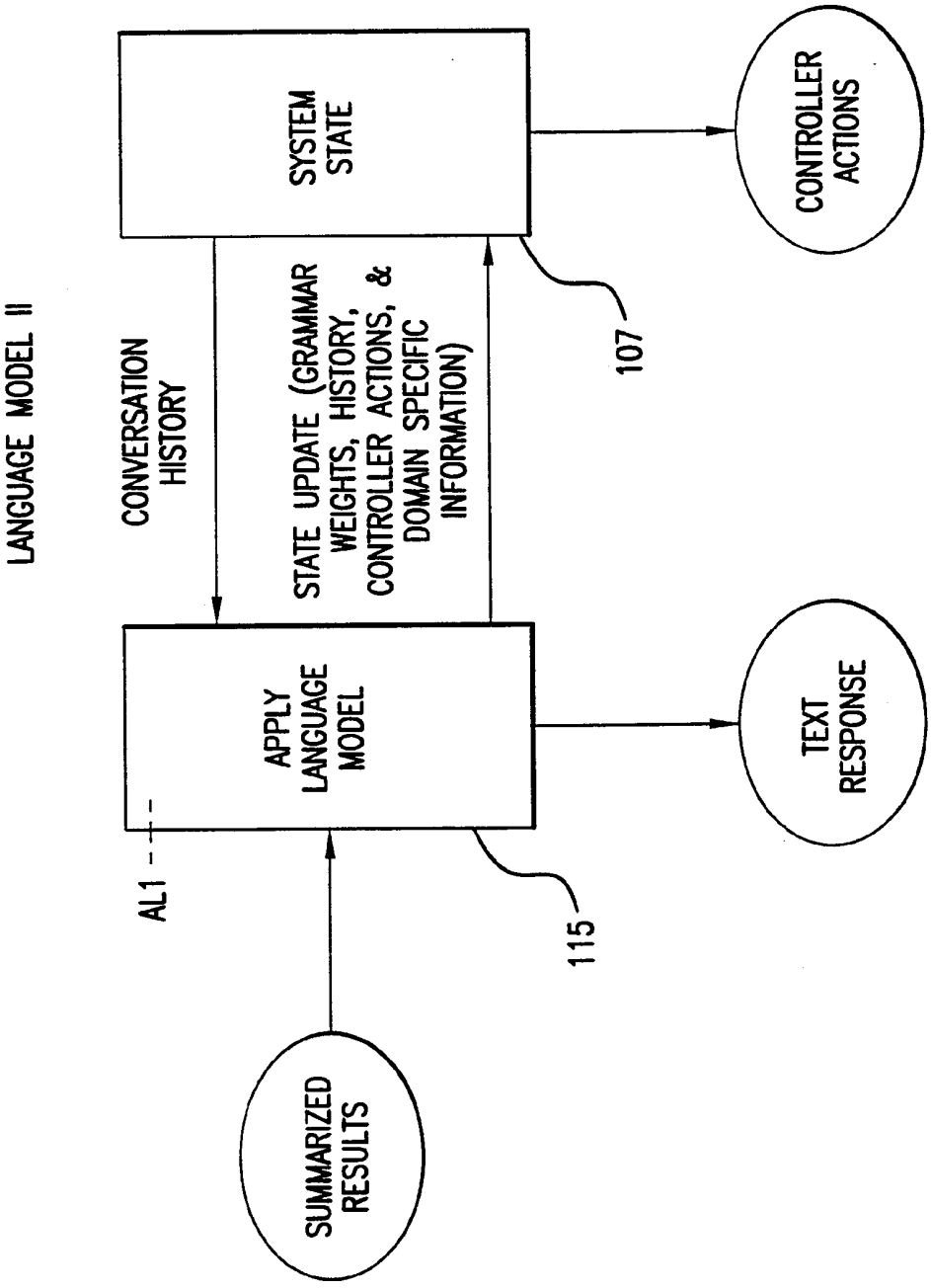

LANGUAGE MODEL II
CONVERSATION HISTORY
STATE UPDATE (GRAMMAR WEIGHTS, HISTORY, CONTROLLER ACTIONS, & DOMAIN SPECIFIC INFORMATION)
SYSTEM STATE
CONTROLLER ACTIONS
107
APPLY LANGUAGE MODEL
TEXT RESPONSE
AL1
115
SUMMARIZED RESULTS

FIG.6

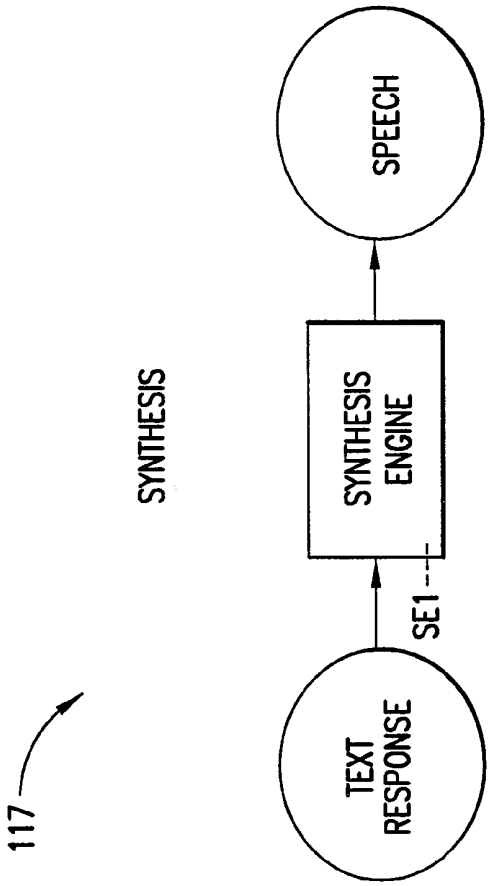


FIG.7

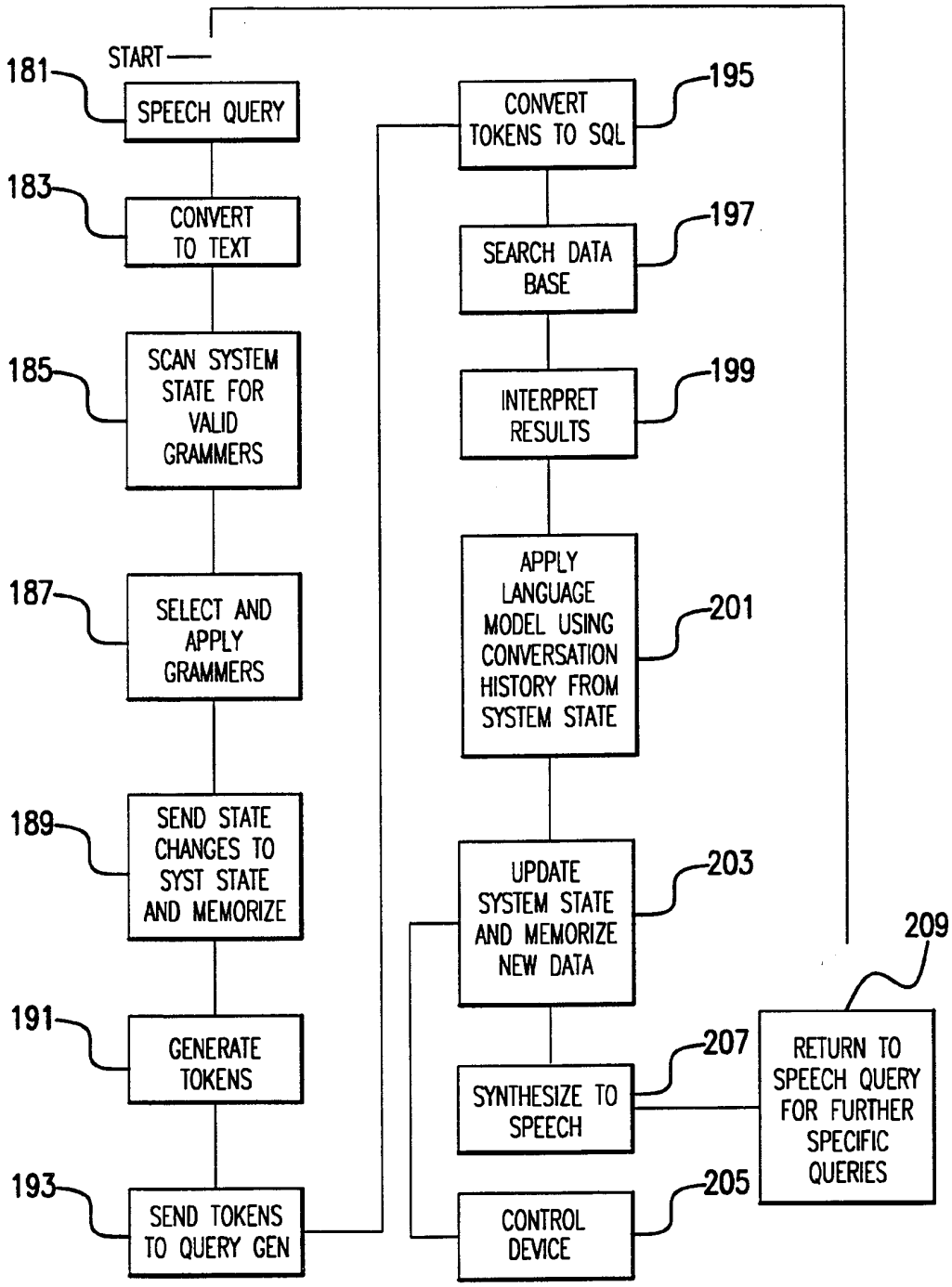

START
181
SPEECH QUERY
183
CONVERT TO TEXT
185
SCAN SYSTEM STATE FOR VALID GRAMMERS
187
SELECT AND APPLY GRAMMERS
189
SEND STATE CHANGES TO SYST STATE AND MEMORIZE
191
GENERATE TOKENS
193
SEND TOKENS TO QUERY GEN
195
CONVERT TOKENS TO SQL
197
SEARCH DATA BASE
199
INTERPRET RESULTS
201
APPLY LANGUAGE MODEL USING CONVERSATION HISTORY FROM SYSTEM STATE
203
UPDATE SYSTEM STATE AND MEMORIZE NEW DATA
205
CONTROL DEVICE
207
SYNTHESIZE TO SPEECH
209
RETURN TO SPEECH QUERY FOR FURTHER SPECIFIC QUERIES

FIG.8

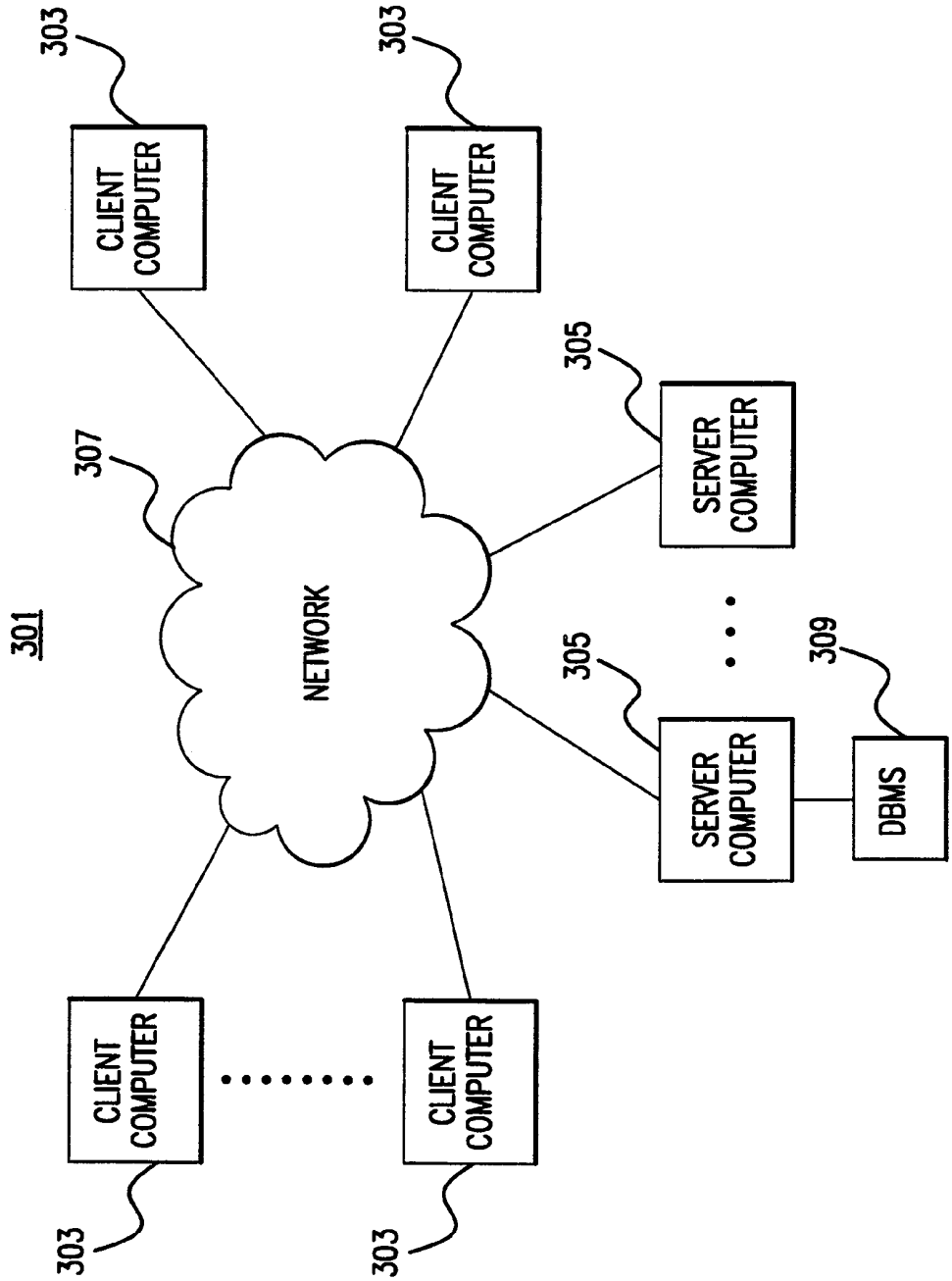

301
NETWORK
307
303
CLIENT COMPUTER
303
CLIENT COMPUTER
305
SERVER COMPUTER
305
SERVER COMPUTER
309
DBMS
303
CLIENT COMPUTER
303
CLIENT COMPUTER

FIG.9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/02906

A. CLASSIFICATION OF SUBJECT MATTER
IPC(7) :G10L 15/18
US CL :704/257
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 704/257, 256, 255, 251, 270, 275

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WEST, USPT, JPAB, EPAB, DWPI, TDBD

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y, P	US 5,983,221 A (CHRISTY) 09 November 1999, see Fig. 2.	1-22
Y	HEMPHILL, CHARLES T. et al. Surfing the Web by Voice. ACM Multimedia 95. Pages 1-11.	1-22
Y, P	US 5,960,399 A (BARCLAY et al.) 28 September 1999, see abstract.	1-22
A	US 5,732,216 A (LOGAN et al.) 24 March 1998, see Fig. 1.	1-22

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"E"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 20 MAY 2000	Date of mailing of the international search report 11 JUL 2000
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer RICHEMOND DORVIL Joni Hill Telephone No. (703) 305-3900

Form PCT/ISA/210 (second sheet) (July 1998)★

(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

(19) Weltorganisation für geistiges Eigentum
Internationales Büro

(43) Internationales Veröffentlichungsdatum
26. August 2004 (26.08.2004)

PCT

(10) Internationale Veröffentlichungsnummer
WO 2004/072954 A1

(51) Internationale Patentklassifikation[7]: G10L 15/18, 15/28

(21) Internationales Aktenzeichen: PCT/EP2004/000351

(22) Internationales Anmeldedatum:
19. Januar 2004 (19.01.2004)

(25) Einreichungssprache: Deutsch

(26) Veröffentlichungssprache: Deutsch

(30) Angaben zur Priorität:
103 06 022.7 13. Februar 2003 (13.02.2003) DE

(71) Anmelder *(für alle Bestimmungsstaaten mit Ausnahme von US)*: **SIEMENS AKTIENGESELLSCHAFT** [DE/DE]; Wittelsbacherplatz 2, 80333 München (DE).

(72) Erfinder; und

(75) Erfinder/Anmelder *(nur für US)*: **BLOCK, Hans-Ulrich** [DE/DE]; Kirchenstr. 42, 81675 München (DE). **SCHACHTL, Stefanie** [DE/DE]; Am Weinberg 31, 82239 Alling (DE).

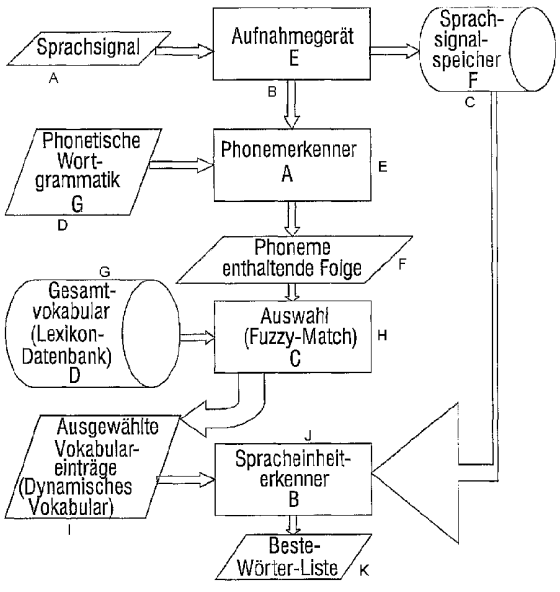
(74) Gemeinsamer Vertreter: **SIEMENS AKTIENGESELLSCHAFT**; Postfach 22 16 34, 80506 München (DE).

(81) Bestimmungsstaaten *(soweit nicht anders angegeben, für jede verfügbare nationale Schutzrechtsart)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NI, NO, NZ, OM, PG,

[Fortsetzung auf der nächsten Seite]

(54) Title: THREE-STAGE WORD RECOGNITION

(54) Bezeichnung: DREISTUFIGE EINZELWORTERKENNUNG



A...VOICE SIGNAL
B...RECORDING DEVICE E
C...VOICE SIGNAL MEMORY
D...PHONETIC GLOSSARY GRAMMAR
E...PHONEME RECOGNITION SYSTEM
F...PHONEME CONTAINING SEQUENCE
G...ENTIRE GLOSSARY (DATA BASE OF DICTIONARIES) D
H...SELECTION (APPROXIMATE CORRESPONDENCE) C
I...SELECTED ENTRY WORDS (DYNAMIC GLOSSARY)
J...VOICE UNIT RECOGNITION SYSTEM B
K...LIST OF BEST WORDS

(57) **Abstract:** In the framework of a three-stage word recognition, a phoneme sequence is associated with a voice unit, entry words most similar to said phoneme sequence are selected in a selection glossary and afterwards, the voice unit is recognised by means of speech recognition which is used by the system as a glossary of the selected entry words which are most similar to said phoneme sequence.

(57) **Zusammenfassung:** Bei einer dreistufigen Spracherkennung wird einer Spracheinheit zunächst eine Phonemfolge zugeordnet, dann werden aus einem Auswahlvokabular diejenigen Vokabulareinträge ausgesucht, die der Phonemfolge am ähnlichsten sind und schliesslich wird die Spracheinheit mit einem Spracheinheiterkenner erkannt, der als Vokabular die ausgewählten, der Phonemfolge ähnlichsten Vokabulareinträge verwendet.

WO 2004/072954 A1

PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) **Bestimmungsstaaten** *(soweit nicht anders angegeben, für jede verfügbare regionale Schutzrechtsart)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), eurasisches (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), europäisches (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Veröffentlicht:

— *mit internationalem Recherchenbericht*

— *vor Ablauf der für Änderungen der Ansprüche geltenden Frist; Veröffentlichung wird wiederholt, falls Änderungen eintreffen*

Zur Erklärung der Zweibuchstaben-Codes und der anderen Abkürzungen wird auf die Erklärungen ("Guidance Notes on Codes and Abbreviations") am Anfang jeder regulären Ausgabe der PCT-Gazette verwiesen.

Beschreibung

Dreistufige Einzelworterkennung

Auf so genannter Embedded Hardware, wie beispielsweise einem Mobiltelefon, einem schnurlosen Telefon, einem PDA, einem Pad oder einem Navigationssystem, die sich derzeit durch Einschränkungen beim Hauptspeicher und geringe Rechenleistung auszeichnet, soll die Spracherkennung von Einzelwörtern aus einem großen Vokabular ermöglicht werden. Unter einem großen Vokabular werden in diesem Zusammenhang Vokabularien von mehr als 60.000 Wörtern verstanden. Es soll zum Beispiel ermöglicht werden, Städtenamen durch ein Sprachkommando in ein Navigationssystem einzugeben oder eine Diktiermöglichkeit für SMS anzubieten.

Darüber hinaus soll auf marktüblicher PC-Hardware die Spracherkennung von Einzelwörtern aus einem sehr großen Vokabular ermöglicht werden. Unter einem sehr großen Vokabular werden in diesem Zusammenhang Vokabularien von mehr als 200.000 Wörtern verstanden. Dadurch soll es beispielsweise ermöglicht werden, Einträge in einem öffentlichen Telefonbuch oder Adressregister durch ein Sprachkommando auszusuchen.

Bislang verwendete Verfahren benötigen für Embedded Hardware zu viel Rechenleistung oder zu viel Speicherplatz. Für marktübliche PC-Hardware stehen zumindest Verfahren für ein großes Vokabular zur Verfügung. Hierfür werden Spracherkenner mit so genanntem "tree search" oder mit so genanntem "FST search" eingesetzt.

Ein interessanter technischer Lösungsweg ist in US 6,256,630 B1 beschrieben. Hier wird ein System vorgeschlagen, das mit einem Ähnlichkeitsmaß arbeitet.

US 5,873,056 beschreibt ebenfalls ein Verfahren zur Spracherkennung.

US 5,638,425 offenbart eine Spracherkennung mit sehr großem Vokabular auf PC-Hardware, bei der ein Worterkenner und ein Phonemerkenner eingesetzt werden und als Suchwort in der Datenbank dasjenige Erkennungsergebnis verwendet wird, das das beste Übereinstimmungsmaß hat.

Davon ausgehend liegt der Erfindung die Aufgabe zugrunde, eine Spracherkennung, insbesondere eine Einzelworterkennung, bei großem oder sehr großem Vokabular mit weniger Speicher und Rechenleistung zu ermöglichen.

Diese Aufgabe wird durch die in den unabhängigen Ansprüchen angegebenen Erfindungen gelöst. Vorteilhafte Ausgestaltungen ergeben sich aus den Unteransprüchen.

In einem Verfahren zur Spracherkennung wird einer Spracheinheit, insbesondere einem Wort, durch einen Phonemerkenner, also Mittel zum Erkennen von Phonemen, eine Phoneme enthaltende Folge zugeordnet. Dies geschieht, indem der Phonemerkenner in der Spracheinheit vorhandene Phoneme erkennt. Dann wird die so generierte, Phoneme enthaltende Folge mit Vokabulareinträgen eines Auswahlvokabulars verglichen, das groß oder sehr groß ist und entsprechend sehr viele Vokabulareinträge aufweist. Aus den sehr vielen Vokabulareinträgen wird eine Anzahl von Vokabulareinträgen ausgewählt, die der Phoneme enthaltenden Folge ähnlich sind. Vorzugsweise sind dies die der Phoneme enthaltenden Folge ähnlichsten Vokabulareinträge und ihre Anzahl liegt erheblich unter der Anzahl der Vokabulareinträge des Auswahlvokabulars. Die so ausgewählten, der Phoneme enthaltenden Folge ähnlichen Vokabulareinträge werden dann als Erkennungsvokabular eines Spracheinheiterkenners, also Mitteln zum Erkennen von Spracheinheiten, geladen und die Spracheinheit wird vom Spracheinheiterkenner aus diesem Erkennungsvokabular erkannt.

Daraus ergibt sich der Vorteil, dass sowohl bei der Erkennung durch den Phonemerkenner als auch bei der Erkennung durch den Spracheinheiterkenner die Suche in einem relativ kleinen Vokabular ausgeführt werden kann, was eine erhebliche Einsparung an Speicher- und Rechenressourcen bedeutet. Auf dem Auswahlvokabular wird dagegen nur eine Ähnlichkeitssuche ausgeführt, die wesentlich weniger rechenintensiv ist.

Übliche Phonemerkenner liefern leider häufig eine relativ schlechte Erkennungsleistung. Deshalb ist es sinnvoll, die Erkennungsleistung des in der ersten Stufe verwendeten Phonemerkenners durch bestimmte Maßnahmen zu verbessern.

Dazu können als Erkennungsvokabular des Phonemerkenners nicht nur die in der deutschen Sprache beispielsweise 36 Phoneme verwendet werden, sondern auch noch andere Einheiten, die als Pseudophoneme in das Vokabular des Phonemekenners eingefügt werden. Da größere Einheiten in der Regel besser erkannt werden, können als Pseudophoneme größere Einheiten, die für sich noch keine Wörter bilden, verwendet werden, insbesondere Konsonantencluster, Kombinationen aus einem Vokal und einem Konsonanten und/oder die am häufigsten vorkommenden Silben der Sprache der zu erkennenden Spracheinheit. Der Phonemerkenner ist also ein Erkenner, der insbesondere, aber nicht nur Phoneme erkennt.

Die Anzahl der Vokabulareinträge des Phonemerkenners sollte allerdings nicht so groß werden, dass die Ressourcen des embedded oder PC-Spracherkenners zu stark belastet werden, wodurch die Vorteile der Erfindung verloren gingen. In der Praxis hat es sich als vorteilhaft herausgestellt, wenn die Anzahl der Vokabulareinträge des Phonemerkenners etwa 0,01% bis 10% der Vokabulareinträge des Auswahlvokabulars beträgt, insbesondere etwa 0,1% bis 1%. Je nach embedded Lösung sollte die Anzahl der Vokabulareinträge des Phonemerkenners also etwa 300 oder auch etwa 2000 betragen.

Bei einer relativ großen Anzahl der Vokabulareinträge des Phonemerkenners, also so etwa 2000, können als Pseudophoneme neben Kombinationen aus einem Vokal und einem Konsonanten auch Kombinationen aus mehreren Vokalen und Konsonanten berücksichtigt werden.

Eine andere Möglichkeit, die Leistung des Phonemerkenners zu verbessern, besteht darin, nur bestimmte Abfolgen der Phoneme oder der Phoneme und größeren Einheiten in der Folge von Phonemen zuzulassen. Das sind insbesondere die Abfolgen, die phonotaktisch korrekt sind. Dies lässt sich durch die Verwendung eines Sprachmodells und/oder durch eine die phonetischen Spracheinheitbildungsregeln der jeweiligen Sprache beschreibende Spracheinheitgrammatik bei der Erkennung erzielen.

Auch bei der zweiten Stufe, der Auswahl der der Phoneme enthaltenden Folge ähnlichen Vokabulareinträge aus dem Auswahlvokabular lassen sich durch bestimmte Maßnahmen erhebliche Effizienzsteigerungen erzielen.

Da als Spracheinheiten in der Regel Einzelworte erkannt werden sollen, führt eine falsch erkannte größere Einheit, insbesondere eine falsch erkannte Silbe, zu einem relativ großen Fehler beim Bestimmen der Ähnlichkeit und damit bei der Auswahl, da sie allein schon einen großen Anteil des Wortes ausmacht. Oftmals enthält diese falsch erkannte größere Einheit aber zumindest eines oder mehrere richtig erkannte Phoneme. Dies lässt sich ausnutzen, indem die größeren Einheiten vor der Auswahl von der Phoneme enthaltenden Folge ähnlichen Vokabulareinträgen wieder in Phoneme zerlegt werden, so dass die Phoneme enthaltende Folge als Phonemfolge ausschließlich oder nahezu ausschließlich Phoneme enthält. Dadurch liefern die in der falsch erkannten größeren Einheit enthaltenen richtigen Phoneme einen korrekten Beitrag bei der Bestimmung der Ähnlichkeit.

Bei der Auswahl der der Phoneme enthaltenden Folge ähnlichen Vokabulareinträgen kann grundsätzlich ein beliebiges Ähnlichkeitsmaß verwendet werden, um die Ähnlichkeit zwischen der Phoneme enthaltenden Folge und dem jeweils zu vergleichenden Vokabulareintrag zu bestimmen. Die Vokabulareinträge des Auswahlvokabulars werden dann entsprechend der Größe ihres Ähnlichkeitsmaßes geordnet und die der Phoneme enthaltenden Folge ähnlichsten Vokabulareinträge werden ausgewählt.

Allerdings würde bei einem großen oder sehr großen Auswahlvokabular eine Bestimmung der Ähnlichkeitsmaße sämtlicher Vokabulareinträge einen unvertretbaren Zeitaufwand bedeuten. Deshalb ist es vorteilhaft, wenn die Auswahl der der Phoneme enthaltenden Folge ähnlichen Vokabulareinträgen mit Hilfe eines Fuzzy-Matches und/oder mit Hilfe einer Index-Suche erfolgt. Dazu kann die Phoneme enthaltende Folge in die Menge ihrer Phonembigramme zerlegt werden.

Für die Größe des Erkennungsvokabulars des Spracheinheiterkenners ergeben sich in etwa die gleichen Anforderungen wie für die Größe des Erkennungsvokabulars des Phonemerkenners. Entsprechend sind das Vokabular des Spracheinheiterkenners und das Vokabular des Phonemerkenners vorzugsweise in etwa gleich umfangreich. Da die ausgewählten, der Phoneme enthalten Folge ähnlichen Vokabulareinträge als Erkennungsvokabular des Spracheinheiterkenners dienen oder zumindest den wesentlichen Teil davon bilden, heißt das, dass die Anzahl der ausgewählten, der Phoneme enthaltenden Folge ähnlichen Vokabulareinträgen etwa 0,01% bis 10% der Vokabulareinträge des Auswahlvokabulars beträgt, insbesondere etwa 0,1% bis 1%. Daraus ergibt sich bei heutigen Lösungen eine Anzahl von etwa 300 oder auch eine Anzahl von etwa 2000.

Für sprecherunabhängige Spracherkennung kann als Spracheinheiterkenner vorzugsweise ein HMM-Erkenner

eingesetzt werden. Alternativ oder ergänzend ist aber auch der Einsatz eines DTW-Erkenners möglich. Der Spracheinheiterkenner arbeitet insbesondere mit einer Viterbisuche.

In den meisten Einsatzfällen wird die Spracheinheit ein Wort sein, insbesondere ein Name etwa in Form eines Städtenamens, Straßennamens und/oder Personennamens. Der Spracheinheiterkenner ist dann dementsprechend in der Regel ein Worterkenner.

Eine Vorrichtung, die eingerichtet ist und Mittel aufweist, um das vorstehende Verfahren auszuführen, lässt sich beispielsweise durch entsprechendes Einrichten und Programmieren einer Datenverarbeitungsanlage realisieren. Dadurch können insbesondere der Phonemerkenner, die Auswahlmittel und der Spracheinheiterkenner zur Verfügung gestellt werden. Da die Spracheinheit sowohl für den Erkennungsvorgang des Phonemerkenners als auch für den Erkennungsvorgang des Spracheinheiterkenners benötigt wird, enthält die Vorrichtung vorzugsweise Mittel, durch die die Spracheinheit sowohl dem Phonemerkenner als auch später dem Spracheinheiterkenner zuführbar ist. Diese Mittel können beispielsweise einen Sprachsignalspeicher aufweisen.

Andere vorteilhafte Ausgestaltungen der Vorrichtung ergeben sich analog zu den vorteilhaften Ausgestaltungen des Verfahrens.

Ein Programmprodukt für eine Datenverarbeitungsanlage, das Codeabschnitte enthält, mit denen eines der geschilderten Verfahren auf der Datenverarbeitungsanlage ausgeführt werden kann, lässt sich durch geeignete Implementierung des Verfahrens in einer Programmiersprache und Übersetzung in von der Datenverarbeitungsanlage ausführbaren Code ausführen. Die Codeabschnitte werden dazu gespeichert. Dabei wird unter einem Programmprodukt das Programm als handelbares Produkt

verstanden. Es kann in beliebiger Form vorliegen, so zum Beispiel auf Papier, einem computerlesbaren Datenträger oder über ein Netz verteilt.

Weitere wesentliche Vorteile und Merkmale der Erfindung ergeben sich aus der Beschreibung eines Ausführungsbeispiels anhand der Zeichnung. Dabei zeigt

Figur 1 ein Ablaufdiagramm eines Verfahrens zur Spracherkennung;

Figur 2 eine Wortgrammatik.

Eine Vorrichtung zur Spracherkennung in Form eines Spracherkennungssystems für große Wortschätze enthält beispielsweise folgende Komponenten:

- Ein erstes Spracherkennungssystem in Form eines Phonemerkenners A, das auf der Zielplattform Einheiten aus einem festgelegten Vokabular von mindestens 300 Wörtern kontinuierlich gesprochen erkennen kann, was eine sehr geringe Anforderung ist, und die Angabe eines Sprachmodells in Form eines endlichen Automaten erlaubt;
- ein zweites Spracherkennungssystem in Form eines Spracheinheiterkenners B, das auf der Zielplattform Einzelwörter aus einem dynamisch ladbaren Vokabular von mindestens 300 Wörtern erkennen kann, was ebenfalls eine sehr geringe Anforderung ist;
- Mittel zur Auswahl von der Phoneme enthaltenden Folge ähnlichen Vokabulareinträgen aus einem Auswahlvokabular in Form eines Fuzzy-Match-Index-Systems C, wie es etwa aus der Technologie der Translation Memories bekannt ist und dessen Einsatz beispielsweise in US 5,724,593 beschrieben ist;
- eine Datenbank D, in der das zu erkennende Vokabular als Auswahlvokabular in seiner graphemischen und phonetischen Form abgelegt ist;

- ein Aufnahmegerät E, mit dem das vom Benutzer geäußerte Sprachsignal aufgenommen wird;
- einen Sprachsignalspeicher F, in dem das mit dem Aufnahmegerät E aufgenommene Sprachsignal gespeichert wird;
- optional eine Spracheinheitgrammatik in Form einer Grammatik G, die die phonetischen Regeln der Sprache beschreibt.

In der ersten Stufe eines Verfahrens zur Spracherkennung wird eine Phonemerkennung durchgeführt. Dazu wird eine Äußerung eines Benutzers mit dem Aufnahmegerät E aufgenommen und in dem Sprachsignalspeicher F gespeichert. Mittels des ersten Spracherkennungssystems in Form eines Phonemerkenners A wird dem Sprachsignal eine Phoneme enthaltende Folge zugeordnet.

Dabei ist die Auswahl des Erkennungsvokabulars des Phonemerkenners besonders wichtig. Bisherige Verfahren für die Phonemerkennung definieren als Einheiten des Phonemerkenners die Phoneme der Sprache, also beispielsweise für das Deutsche a, a:, ä, ä:, e, ..., die ohne Sprachmodell erkannt werden. Systeme, die so aufgebaut sind, sind sehr störanfällig und liefern im Allgemeinen sehr schlechte Erkennungsergebnisse, die nur sehr bedingt zu gebrauchen sind. Insbesondere variieren die erkannten Phonemfolgen bei Eingabe desselben Worts stark von Sprecher zu Sprecher.

Die Erkennung größerer Zusammenhänge wie beispielsweise Wörter funktioniert besser, da sich die Lautqualität von Konsonanten auch noch auf den Nachbarlauten widerspiegelt.

Deshalb wird der Phonemerkenner vorzugsweise derart gestaltet, dass möglichst größere Einheiten als Phoneme erkannt werden, aber dennoch das Vokabular so klein gehalten wird, dass der Phonemerkenner auf Embedded Hardware auflauffähig ist. Die Größe des Erkennungsvokabulars eines

Spracherkennungssystems determiniert nämlich Speicher- und CPU-Bedarf.

Als Einheiten des Erkennungsvokabulars des Phonemerkenners werden daher gewählt: die Phoneme (Laute) der Sprache (wie in herkömmlichen Phonemerkennern, im Deutschen 36) + alle im Vokabular vorkommenden Konsonantencluster (im Deutschen beispielsweise lchst, mpft etc., in Summe ca. 180) + die 120 häufigsten Silben. Das gibt insgesamt ca. 300 Vokabulareinträge, was der üblichen Leistungsfähigkeit von Embedded Hardware entspricht. Bei leistungsfähigerer Hardware können auch etwa 2000 Vokabulareinträge verwendet werden, wobei hier Kombinationen aus einem oder mehreren Vokalen mit einem oder mehreren Konsonanten und weitere häufige Silben hinzutreten.

Die Abfolge dieser Einheiten des Erkennervokabulars (Vokabularelemente) in einem Wort wird entweder durch ein statisches, am Vokabular trainiertes, bi- oder trigramm-Sprachmodell oder durch eine die phonetische Wortbildungsregeln der jeweiligen Sprache beschreibende Grammatik G eingeschränkt.

Eine solche Grammatik G, die in Form und Mächtigkeit eines endlichen Automaten spezifiziert werden kann, beschreibt beispielsweise ein Wort als eine Folge von einer oder mehreren Silben und diese Silben dann wiederum jeweils als Folge von optionalem Anlaut, Vokal und optionalem Auslaut. Sie legt fest, aus welchen Konsonanten und Konsonantenclustern die Anlaute und Auslaute der betrachteten Sprache bestehen können und welche Gruppen von Auslauten auf welche Vokale folgen können.

Weiterhin beschreibt die Grammatik gegebenenfalls auch konsonantische Silben wie dl in einem Eigenname Brandl. Eine typische Grammatik ist in Figur 2 dargestellt.

Das erste Spracherkennungssystem in Form des Phonemerkenners A gibt nun eine Folge von solchen Einheiten des Vokabulars als Ergebnis zurück. Für das Wort Mönchengladbach ist das beispielsweise die Folge m-ö-nch-en-gl-a-t-b-a-x oder auch - bei schlechterer Erkennung - b-öh-mch-nen-k-au-b-a-k.

Es hat sich als ganz besonders vorteilhaft erwiesen, wenn für die später folgende Auswahl des Erkennervokabulars des zweiten Spracherkennungssystems in Form des Spracheinheiterkenners B diese erkannten Folgen von Phonemen und größeren Einheiten nun wieder in Phonemfolgen zerlegt werden. Für das oben genannte Beispiel ergeben sich die Phonemfolgen m-ö-n-ch-e-n-g-l-a-t-b-a-x bzw. b-öh-m-ch-n-e-n-k-au-b-a-k, mit denen weitergearbeitet wird.

In der zweiten Stufe des Verfahrens zur Spracherkennung erfolgt die Auswahl der der Phoneme enthaltenden Folge ähnlichen Vokabulareinträge aus einem Auswahlvokabular mit Hilfe eines Fuzzy-Matches. Mit dem Fuzz-Match-Index-System C werden dazu in der Datenbank D diejenigen 300 Einträge gesucht, die der erkannten Phonemfolge im Sinne eines Fuzzy-Matches am ähnlichsten sind.

Um diese Fuzzy-Suche zu ermöglichen, wird nach folgendem Verfahren ein Index aufgebaut. Die n Einträge der Datenbank werden jeweils mit einer eindeutigen Indexnummer aus {1...n} versehen. Zur phonetischen Repräsentation inklusive Anfangs- und Endmarkierer (#) eines jeden Eintrags i wird die Menge der in ihr enthaltenen Bigramme gebildet. Für Mönchengladbach und die Phonemfolge m-ö-n-ch-e-n-g-l-a-t-b-a-x werden dementsprechend die Phonembigramme {#-m,m-ö,ö-n,n-ch,ch-e,e-n,n-g,g-l,l-a,a-t,t-b,b-a,a-x,x-#} gebildet.

Für jedes Bigramm b aus dieser Menge wird ein Indexeintrag b-{i} gespeichert. Kommt Bigramm b noch in einem weiteren Eintrag j vor, wird der Indexeintrag entsprechend zu b {i,j}

erweitert. Dies wird für alle Einträge und alle Bigramme fortgeführt.

Das Ergebnis ist eine spärlich besiedelte Matrix von Bigrammen (types) auf die jeweiligen Indexnummern der Einträge, in denen das Bigramm vorkommt.

Der Fuzzy-Match wird nun derart realisiert, dass die erkannte Phonemfolge, beispielsweise b-öh-m-ch-n-e-n-k-au-b-a-k in die Menge ihrer Bigramme BM zerlegt wird {#-b,b-öh,öh-m,m-ch,ch-n,n-e,e-n,n-k,k-au,au-b,b-a,a-k,k-#}.

Zu jedem Bigramm wird der zugeordnete Vektor der Indexnummern aus dem Speicher ausgelesen und es wird eine unscharfe Vereinigungsmenge VM derjenigen Indexnummern gebildet, auf die durch Bigramme in BM verwiesen wird. Jedes Element e der Vereinigungsmenge VM ist dabei mit einer Häufigkeitsangabe versehen, die angibt, wie viele Bigramme aus BM auf e verweisen.

Die Häufigkeitsangabe im gegebenen Beispiel für den Eintrag i (Mönchengladbach) ist beispielsweise 2, da nur e-n und b-a auf i verweisen.

Die errechnete Häufigkeit h der Indexnummer wird abschließend in ein über die Länge des Musters lm und die Länge der phonetischen Form des Eintrags le normiertes Gewichtsmaß g umgewandelt. Dies geschieht nach der Formel
$g = -\log(h/lm)*0,5 + -\log(h/le)*0,5$ oder nach einer anderen geeigneten Formel.

Die Elemente der unscharfen Vereinigungsmenge VM werden nun nach ihrem Gewicht aufsteigend sortiert.

Für die dritte Stufe, die Spracheinheiterkennung durch den Spracheinheiterkenner, werden beispielsweise die den ersten 300 Elementen der sortierten Menge VM zugeordneten Einträge

als Vokabular in das zweite Spracherkennungssystem in Form des Spracheinheiterkenners B geladen. Mit dem Spracheinheiterkenner, der insbesondere als Worterkenner realisiert ist, wird nun noch einmal das im Sprachspeicher F gespeicherte Sprachsignal analysiert. Typischerweise sollte das Spracherkennungssystem die m am besten erkannten Wörter ausgeben.

Die genannte Vokabulargröße von 300 für das erste Spracherkennungssystem in Form des Phonemerkenners A ist eine arbiträre Größe. Generell gilt, dass das Erkennungsergebnis um so besser ist, je mehr größere Einheiten im Lexikon enthalten sind. Wenn die Rechenleistung der Hardware es gestattet, kann es beispielsweise vorteilhaft sein, statt der genannten 300 Einträge ein Vokabular zu verwenden, das aus den Silben einer Sprache oder allen in der Datenbank vorkommenden Silben besteht. Im Deutschen ist hier mit etwa 7000 Einträgen zu rechnen. Die Formulierung der phonetischen Wortgrammatik kann sich dann auf die triviale Regel Wort = Silbe* (ein Wort besteht aus einer Folge von einer oder mehreren Silben) beschränken.

Das Fuzzy-Match-System kann gegebenenfalls verbessert werden, wenn bekanntermaßen leicht verwechselbare Laute wie m und n sowohl in der Datenbank als auch in der erkannten Phonemfolge auf einen Laut, beispielsweise n, normiert werden. Im obigen Beispiel hätte der Eintrag für Möchengladbach bei einer m-n-Normierung zwei Punkte mehr erhalten. Erste empirische Tests haben allerdings bisher gezeigt, dass der Einsatz solcher Normierungen auch stark zur Auswahl falscher Einträge beiträgt.

Weiterhin kann das Fuzzy-Match-System gegebenenfalls dadurch verbessert werden, dass die Häufigkeitsberechnung so gestaltet wird, dass Bigramme, die bekanntermaßen schlecht erkennbare Laute, wie beispielsweise die stimmlosen Plosive

p, t, k enthalten, nur zu einem geringeren Teil als 1,0 zur Häufigkeit der entsprechenden Indexnummern beitragen.

Das Fuzzy-Match-System kann dadurch verbessert werden, dass in das normierte Gewichtsmaß neben der Häufigkeit der Bigramme auch die absolute Differenz der Anzahl der Vokale in Datenbankeintrag ve und Suchmuster vs sowie die absolute Differenz der Länge von Datenbankeintrag le und Suchmuster lm eingehen. Dies kann beispielsweise durch Berechnung eines Gewichtsmaßes g nach folgender Formel geschehen:

$$g = -\log(h/lm)*f_1 + -\log(h/le)*f_2 + -\log(1/(abs(vm-ve)+1))*f_3 + -\log(1/(abs(lm-le)+1))*f_4$$

wobei die Koeffizienten f_n sinnvoll einzusetzen sind.

Auch die genannte Vokabulargröße von 300 für das zweite Spracherkennungssystem in Form des Spracheinheiterkenners B ist eine arbiträre Größe. Wenn die Rechenleistung der Hardware es gestattet, ist es sicherlich vorteilhaft, statt der genannten 300 Einträge ein größeres dynamisches Vokabular der Größe v zu verwenden. Beim Fuzzy-Match werden dann entsprechend die ersten v Elemente aus der sortierten Menge VM zur weiteren Betrachtung gewählt.

Durch die Erfindung wird eine Einzelworterkennung aus großen Wortschätzen bei geringen Hardwareanforderungen durch Aufteilung in drei einfachere Operationen realisiert.

Die Spracherkennung aus Vokabularen von bis zu 300 Wörtern auf einem PDA oder einer Autoelektronik (Navigationssystem) ist bereits heute Stand der Technik und machbar.

Der Speicherplatz- und CPU-Bedarf für den Fuzzy-Match ist nachweisbar gering. Zum Beleg dienen die folgenden Beispiele.

Speicherbedarf: ein Vokabular aus 65.000 Wörtern enthält bei einer grob geschätzten durchschnittlichen Wortlänge von 10 Phonemen/Wort 650.000 Phonemvorkommen (Tokens), denn die durchschnittliche Wortlänge im Deutschen ist ca. 8, plus Wortanfang- und -endemarkierung ergibt 10. Der Wert ist auch an 65.000 Städtenamen nachgewiesen. Da die Anzahl der Bigramm-Vorkommen trivialerweise gleich der Anzahl der Phonemvorkommen ist, ergibt sich eine Menge von ca. 650.000 Bigramm-Vorkommen. Die Anzahl der unterschiedlichen Bigramm-Typen ist natürlich geringer. An empirischen Daten gemessen ergibt sich eine Menge von ca. 1.200 Bigramm-Typen. Trivialerweise ist die Anzahl der Einträge in der spärlich besiedelten Indexmatrix gleich der Anzahl der Bigramm-Vorkommen, also ca. 650.000. Da für die Kodierung von 65.000 (>64k) Indexnummern 3 Byte pro Indexnummer benötigt werden, beträgt der Speicherbedarf für die spärlich besiedelte Indexmatrix 1.950.000 Byte bzw. ca. 2 MB. Hinzu kommt der Speicherbedarf für die graphemische Repräsentation der Einträge (ca. 700 KB) plus jeweils ein Byte für die Speicherung der Länge der Phonemkette und die Anzahl der Vokale (zusammen 130.000 Byte) sowie ein Arbeitsspeicher für die Bildung der unscharfen Vereinigungsmenge von maximal 65.000 Elementen. Der Speicherbedarf für jedes Element liegt bei 7 Byte: 3 Byte für die Indexnummer und 4 Byte für die Realzahl-Darstellung des Gewichts. Für den Arbeitsspeicher sind somit 455.000 Byte zu reservieren. Der Speicherbedarf für den Fuzzy-Match beträgt also für ein Vokabular von 65.000 Einträgen insgesamt maximal 3,2 MB.

CPU-Bedarf: Der CPU-Bedarf hängt von der Bildung der unscharfen Vereinigungsmenge ab. Bei 650.000 Bigramm-Vorkommen und 1200 unterschiedlichen Bigramm-Typen, ergeben sich im Durchschnitt ca. 540 Indexnummern pro Bigramm. Die Anzahl der notwendigen Additionen zur Berechnung der Häufigkeiten beträgt also im Schnitt n*540 bei einer erkannten Phonemkette der Länge n. Hinzu kommt die Berechnung

des Gewichts und die entsprechende Sortierung des Arbeitsspeichers.

Patentansprüche

1. Verfahren zur Spracherkennung,
- bei dem einer Spracheinheit durch einen Phonemerkenner eine Phoneme enthaltende Folge zugeordnet wird,
- bei dem aus einem Auswahlvokabular der Phoneme enthaltenden Folge ähnliche Vokabulareinträge ausgewählt werden,
- bei dem die Spracheinheit durch einen Spracheinheiterkenner erkannt wird, wobei das Erkennungsvokabular des Spracheinheiterkenners die ausgewählten, der Phoneme enthaltenden Folge ähnlichen Vokabulareinträge enthält.

2. Verfahren nach Anspruch 1,
bei dem als Erkennungsvokabular des Phonemerkenners neben Phonemen auch größere Einheiten als Phoneme verwendet und der Spracheinheit zugeordnet werden, insbesondere Konsonantencluster, Kombinationen aus einem Vokal und einem Konsonanten und/oder die am häufigsten vorkommenden Silben.

3. Verfahren nach Anspruch 1 oder Anspruch 2,
bei dem die Anzahl der Vokabulareinträge des Phonemerkenners weniger als 10% der Vokabulareinträge des Auswahlvokabulars beträgt, insbesondere weniger als 1%.

4. Verfahren nach einem der vorhergehenden Ansprüche,
bei dem die Abfolge in der Folge von Phonemen durch ein Sprachmodell und/oder durch eine die phonetischen Spracheinheitsbildungsregeln beschreibende Spracheinheitsgrammatik eingeschränkt wird.

5. Verfahren nach Anspruch 2,
bei dem die größeren Einheiten vor der Auswahl der der Phoneme enthaltenden Folge ähnlichen Vokabulareinträge in Phoneme zerlegt werden.

6. Verfahren nach einem der vorhergehenden Ansprüche, bei dem die Auswahl der der Phoneme enthaltenden Folge ähnlichen Vokabulareinträge mit Hilfe eines Fuzzy-Matches erfolgt.

7. Verfahren nach einem der vorhergehenden Ansprüche, bei dem die Auswahl der der Phoneme enthaltenden Folge ähnlichen Vokabulareinträge mit Hilfe einer Indexsuche erfolgt.

8. Verfahren nach einem der vorhergehenden Ansprüche, bei dem zur Auswahl der der Phoneme enthaltenden Folge ähnlichen Vokabulareinträge der Phoneme enthaltenden Folge entsprechende Phonembigramme gebildet werden.

9. Verfahren nach einem der vorhergehenden Ansprüche, bei dem die Anzahl der ausgewählten, der Phoneme enthaltenden Folge ähnlichen Vokabulareinträge weniger als 10% der Vokabulareinträge des Auswahlvokabulars beträgt, insbesondere weniger als 1%.

10. Verfahren nach einem der vorhergehenden Ansprüche, bei dem der Spracheinheiterkenner einen HMM-Erkenner aufweist.

11. Verfahren nach einem der vorhergehenden Ansprüche, bei dem die Spracheinheit ein Wort ist.

12. Vorrichtung zur Spracherkennung, die eingerichtet ist und Mittel aufweist, um ein Verfahren nach einem der Ansprüche 1 bis 11 auszuführen.

13. Programmprodukt, das, wenn es auf eine Datenverarbeitungsanlage geladen und darauf ausgeführt wird, ein Verfahren nach den Ansprüchen 1 bis 11 oder eine Vorrichtung nach Anspruch 12 in Kraft setzt.

FIG 1

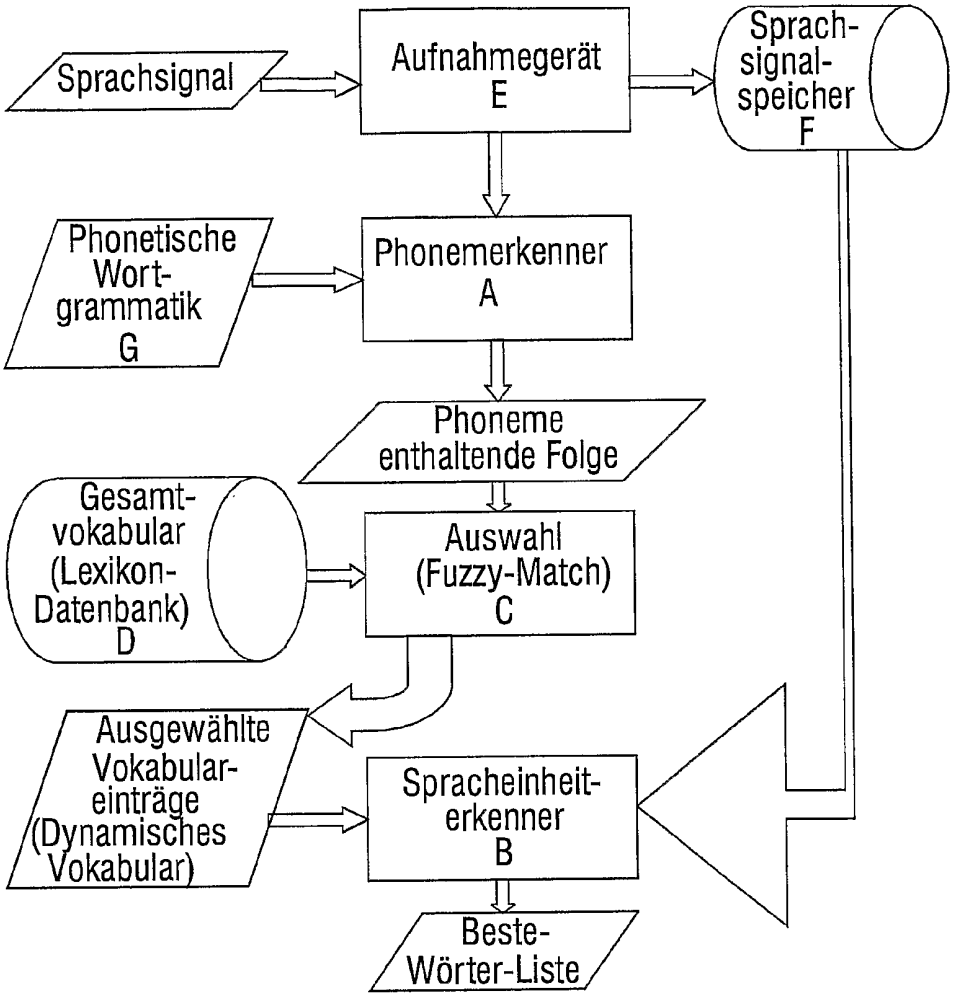

Sprachsignal
Aufnahmegerät
E
Sprach-
signal-
speicher
F
Phonetische
Wort-
grammatik
G
Phonemerkenner
A
Phoneme
enthaltende Folge
Gesamt-
vokabular
(Lexikon-
Datenbank)
D
Auswahl
(Fuzzy-Match)
C
Ausgewählte
Vokabular-
einträge
(Dynamisches
Vokabular)
Spracheinheit-
erkenner
B
Beste-
Wörter-Liste

FIG 2

```
wort(W) --> silbe(t(W,B)), silben(t(B,[])).
wort(W) --> silbe(t(W,[])).

silben(t(A,B)) --> silbe_2(t(A,C)), silben(t(C,B)).
silben(t(A,B)) --> silbe_2(t(A,B)).

silbe(t(A,B)) --> anlaut(t(A,C)), silbe_2(t(C,B)).
silbe(t(A,B)) --> silbe_2(t(A,B)).

silbe_2(t(A,B)) --> ( vokal_1(t(A,D)),
                      ( auslaut_1(t(D,B))
                      ; auslaut_2(t(D,B))
                      )
                    ; vokal_2(t(A,D)), auslaut_2(t(D,B))
                    ).
silbe_2(t(A,B)) --> vokal(t(A,B)).

vokal(T) --> vokal_1(T).
vokal(T) --> vokal_2(T).

vokal_1(t([a|A],A)) --> [a].
vokal_1(t([ae|A],A)) --> [ae].
...
vokal_1(t([uh|A],A)) --> [uh].

vokal_2(t([ai|A],A)) --> [ai].
...
vokal_2(t([au|A],A)) --> [au].

anlaut(t([b|A],A)) --> [b].
anlaut(t([d|A],A)) --> [d].
...
anlaut(t([z|A],A)) --> [z].

anlaut(t([b,l|A],A)) --> [bl].
anlaut(t([b,r|A],A)) --> [br].
...
anlaut(t([t,s|A],A)) --> [ts].

anlaut(t([sh,p,l|A],A)) --> [shpl].
anlaut(t([sh,p,r|A],A)) --> [shpr].
anlaut(t([sh,t,r|A],A)) --> [shtr].

auslaut_1(t([l,t,s|A],A)) --> [lts].
...
auslaut_1(t([x|A],A)) --> [x].
...
auslaut_2(t([r,l,t|A],A)) --> [rlt].
auslaut_2(t([r,m,t|A],A)) --> [rmt].
...
auslaut_2(t([r,n|A],A)) --> [rn].
auslaut_2(t([r,t,s|A],A)) --> [rts].
auslaut_2(t([n g|A],A)) --> [ng].
```

INTERNATIONAL SEARCH REPORT

International Application No
PCT/EP2004/000351

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G10L15/18 G10L15/28

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, INSPEC, WPI Data, PAJ

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category °	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2003/004722 A1 (BUTZBERGER JOHN W ET AL) 2 January 2003 (2003-01-02) paragraph '0027! - paragraph '0034!	1-13
A	EP 1 162 602 A (SONY INT EUROP GMBH) 12 December 2001 (2001-12-12) abstract; figure 1	1-13
A	US 5 638 425 A (TRESSLER ROBERT C ET AL) 10 June 1997 (1997-06-10) cited in the application abstract; figure 4	1-13
P,X	DE 102 07 895 A (HARMAN BECKER AUTOMOTIVE SYSTE) 18 September 2003 (2003-09-18) column 1, paragraph 8 - column 4, paragraph 12; figure 2	1-3,5,7, 10-13

☐ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

° Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
10 June 2004	17/06/2004

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31–70) 340–2040, Tx. 31 651 epo nl, Fax: (+31–70) 340–3016	Burchett, S

Form PCT/ISA/210 (second sheet) (January 2004)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No
PCT/EP2004/000351

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2003004722 A1	02-01-2003	NONE	
EP 1162602 A	12-12-2001	EP 1162602 A1	12-12-2001
		JP 2002006878 A	11-01-2002
		US 2002013706 A1	31-01-2002
US 5638425 A	10-06-1997	AU 5803394 A	04-07-1994
		WO 9414270 A1	23-06-1994
DE 10207895 A	18-09-2003	DE 10207895 A1	18-09-2003
		US 2004034527 A1	19-02-2004

Form PCT/ISA/210 (patent family annex) (January 2004)

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
PCT/EP2004/000351

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES
IPK 7 G10L15/18 G10L15/28

Nach der Internationalen Patentklassifikation (IPK) oder nach der nationalen Klassifikation und der IPK

B. RECHERCHIERTE GEBIETE

Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)
IPK 7 G10L

Recherchierte aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)
EPO-Internal, INSPEC, WPI Data, PAJ

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie°	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
A	US 2003/004722 A1 (BUTZBERGER JOHN W ET AL) 2. Januar 2003 (2003-01-02) Absatz '0027! - Absatz '0034!	1-13
A	EP 1 162 602 A (SONY INT EUROP GMBH) 12. Dezember 2001 (2001-12-12) Zusammenfassung; Abbildung 1	1-13
A	US 5 638 425 A (TRESSLER ROBERT C ET AL) 10. Juni 1997 (1997-06-10) in der Anmeldung erwähnt Zusammenfassung; Abbildung 4	1-13
P,X	DE 102 07 895 A (HARMAN BECKER AUTOMOTIVE SYSTE) 18. September 2003 (2003-09-18) Spalte 1, Absatz 8 - Spalte 4, Absatz 12; Abbildung 2	1-3,5,7, 10-13

☐ Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen

☒ Siehe Anhang Patentfamilie

° Besondere Kategorien von angegebenen Veröffentlichungen :
"A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist
"E" älteres Dokument, das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist
"L" Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt)
"O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht
"P" Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist
"T" Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist
"X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden
"Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren anderen Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist
"&" Veröffentlichung, die Mitglied derselben Patentfamilie ist

Datum des Abschlusses der internationalen Recherche	Absendedatum des internationalen Recherchenberichts
10. Juni 2004	17/06/2004

Name und Postanschrift der Internationalen Recherchenbehörde	Bevollmächtigter Bediensteter
Europäisches Patentamt, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Burchett, S

Formblatt PCT/ISA/210 (Blatt 2) (Januar 2004)

INTERNATIONALER RECHERCHENBERICHT

Angaben zu Veröffentlichungen, die zur selben Patentfamilie gehören

Internationales Aktenzeichen
PCT/EP2004/000351

Im Recherchenbericht angeführtes Patentdokument	Datum der Veröffentlichung	Mitglied(er) der Patentfamilie	Datum der Veröffentlichung
US 2003004722 A1	02-01-2003	KEINE	
EP 1162602 A	12-12-2001	EP 1162602 A1	12-12-2001
		JP 2002006878 A	11-01-2002
		US 2002013706 A1	31-01-2002
US 5638425 A	10-06-1997	AU 5803394 A	04-07-1994
		WO 9414270 A1	23-06-1994
DE 10207895 A	18-09-2003	DE 10207895 A1	18-09-2003
		US 2004034527 A1	19-02-2004

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 February 2007 (15.02.2007)

PCT

(10) International Publication Number
WO 2007/019318 A2

(51) International Patent Classification:
G06F 17/27 (2006.01)

(21) International Application Number:
PCT/US2006/030458

(22) International Filing Date: 4 August 2006 (04.08.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/197,504 5 August 2005 (05.08.2005) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [—/US]; 10505 NE 38th Place, Building 9, Kirkland, WA 98033 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **DI CRISTO, Philippe** [FR/US]; 177 107th Avenue, N.E., #1007, Bellevue, WA 98004 (US). **KE, Min** [US/US]; 3221 201st Place, S.E., Bothell, WA 98012 (US). **KENNEWICK, Robert, A.** [US/US]; 2204 S. Eddy Street, Seattle, WA 98108 (US).

(74) Agents: **GATTO, James, G.** et al.; PILLSBURY WINTHROP SHAW PITTMAN LLP, P.O. Box 10500, Mclean, VA 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 2007/019318 A2

(54) Title: SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE

(57) Abstract: Systems and methods are provided for receiving speech and non-speech communications of natural language questions and/or commands, transcribing the speech and non-speech communications to textual messages, and executing the questions and/or commands. The invention applies context, prior information, domain knowledge, and user specific profile data to achieve a natural environment for one or more users presenting questions or commands across multiple domains. The systems and methods creates, stores and uses extensive personal profile information for each user, thereby improving the reliability of determining the context of the speech and nou-speech communications and presenting the expected results for a particular question or command.

SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE

BACKGROUND OF THE INVENTION

Cross-Reference to Related Applications

[01] This application claims the benefit of U.S. Patent Application Serial No. 11/197,504 filed August 5, 2005, which is herein incorporated by reference in its entirety.

Field of the Invention

[02] The invention relates to retrieval of information or processing of commands through a speech interface and/or a combination of a speech interface and a non-speech interface. More specifically, the invention provides a fully integrated environment that allows users to submit natural language questions and commands via the speech interface and the non-speech interface. Information may be obtained from a wide range of disciplines, making local and network inquiries to obtain the information and presenting results in a natural manner, even in cases where the question asked or the responses received are incomplete, ambiguous or subjective. The invention may further allow users to control devices and systems either locally or remotely.

Background of the Related Art

[03] A machine's ability to communicate with humans in a natural manner remains a difficult problem. Cognitive research on human interaction shows that verbal communication, such as a person asking a question or giving a command, typically relies heavily on context and domain knowledge of the target person. By contrast, machine-based queries (a query may be a question, a command, a request and/or other types of communications) may be highly structured and may not be inherently natural to the human user. Thus, verbal communications and machine processing of queries taken from the verbal communications may be fundamentally incompatible. Yet the ability to allow a person to make natural language speech-based queries remains a desirable goal.

[04] Speech recognition has steadily improved in accuracy and today is successfully used in a wide range of applications. Natural language processing has been applied to the parsing of speech queries. Yet, current systems do not reliably provide a complete environment for users to submit verbal and/or textual communications through natural language queries that are processed to provide natural responses. There remains a number of significant barriers to creation of a complete speech-based and/or non-speech-based natural language query and response environment.

SUMMARY OF THE INVENTION

[05] According to an aspect of the invention, one object of the invention is to overcome these and other drawbacks of prior systems.

[06] According to one aspect of the invention, users may provide the system with spoken utterances and/or textual communications in a natural language format, including imperfect information such as, incomplete thoughts, incomplete sentences, incomplete phrases, slang terminology, repeated words, word variations, synonyms, or other imperfect information. In order for machines (e.g., computer devices) to properly respond to questions and/or commands that are presented in the natural language format, the questions and/or commands may be parsed and interpreted to formulate machine processable queries and algorithms.

[07] In another aspect of the invention, systems are provided for overcoming the deficiencies of prior systems through the application of a complete speech-based information query, retrieval, presentation and command environment or a combination of speech-based and non-speech-based information query, retrieval, presentation and command environment. This environment makes maximum use of context, prior information, domain knowledge, and user specific profile data to achieve a natural environment for one or more users submitting queries or commands to multiple domains. Through this integrated approach, a complete speech-based natural language query and response environment may be created. Alternatively, a combination of speech-based and non-speech-based natural language query and response environment may be created. Further, at each step in the process, accommodation may be made for full or partial failure and graceful recovery. The robustness to partial failure is achieved through the use of probabilistic and fuzzy reasoning at several stages of the process. This robustness to partial failure promotes the feeling of a natural response to questions and commands.

[08] According to another aspect of the invention, the system may include a speech unit interface device that receives utterances, including spoken natural language queries, commands and/or other utterances from a user, and a computer device or system that receives input from the speech unit and processes the input (e.g., retrieves information responsive to the query, takes action consistent with the command and performs other functions as detailed herein). The system may further generate a natural language response and/or generate a natural language message, such as an alert message, without any prompting from a user. According to another embodiment of the invention, the non-speech interface may be provided in addition to, or in place of, the speech unit interface. For example, the non-speech interface

may communicate non-speech information, such as textual communications, graphical or tabular information, or other non-speech information.

[09] According to one embodiment of the invention, infrastructure may be provided to maintain context information during multimodal interactions, such as speech and/or non-speech interactions. According to one exemplary embodiment of the invention, context information may be maintained in a multimodal environment by providing communication channels between multimodal devices, or mobile device, and the system. The communication channels allow the system to receive multimodal input such as text-based commands and questions and/or voice commands and questions. According to another embodiment of the invention, the multimodal input may include a string of text, such as keywords, that are received as commands or questions. According to yet another embodiment of the invention, the system may synchronize the context between the multimodal devices and the speech-based units. In order to send a response to the corresponding device, the system may track the source and send the response to the corresponding speech interface or the non-speech interface.

[10] According to an alternative embodiment of the invention, context information may be maintained using a context manager that may be centrally positioned to receive input from multiple sources and to provide output to multiple sources. According to one embodiment, the devices that communicate with the context manager may register through a registration module and may subscribe to one or more events. According to another embodiment of the invention, the context manager may receive input in Context XML form, for example. The other registered devices may be informed of context changes through a context tracking module to enable synchronizing of context across the registered modules. According to one embodiment of the invention, registered modules may be added or removed from the system. The registered modules may include dynamic link libraries (DLLs) that are specific to multimodal devices.

[11] According to yet another alternative embodiment of the invention, context information may be determined from a command or request that is presented in a textual format and/or a command or request that is presented as an utterance and processed using a multi-pass automatic speech recognition module that transcribes the utterance to a text message. The command or request may be compared against a context description grammar to identify a match. Any active grammars in the context description grammar may be scored against the command or request and a best match may be sent to a response generator module. Agents may be associated with corresponding response generator modules and may

retrieve the requested information for generation of a response. The agents may update a context stack to enable follow-up requests.

[12] According to another aspect of the invention, the speech unit and/or multimodal device may be incorporated into the computer device or system, or may be separate structures. If separate structures are provided, the speech unit and/or multimodal devices may be connected to the computer device via a wired or wireless connection. If a wireless connection is provided, a base unit may be connected to the computer, internally or externally, to communicate with the speech unit and/or multimodal device.

[13] According to another aspect of the invention, the computer devices or systems may comprise stand alone or networked PCs, personal digital assistants (PDAs), cellular telephones, or other computer devices or systems. For convenience, these and other computer alternatives are referred to as computers. One aspect of the invention comprises software that may be installed onto the computer, where the software may include one or more of the following modules: a non-speech information receiving module; a speech recognition module that captures user utterances; a parser that parses the utterance; a text to speech engine module that converts the text to speech; a network interface that enables the computer to interface with one or more networks; a non-speech interface module and an event manager for managing events. Preferably, the event manager is in communication with a context description grammar, a user profile module that enables user profiles to be created, modified and accessed, a personality module that enables various personalities to be created and used, an agent module, an update manager, a cognitive model that provides statistical abstracts of user interaction patterns with the system, one or more databases, and other components.

[14] According to another aspect of the invention domain specific behavior and information may be organized into data managers. Data managers are autonomous executables that receive, process and respond to user questions, queries and commands. The data managers provide complete, convenient and re-distributable packages or modules of functionality, typically for a specific domain of application. Data managers may be complete packages of executable code, scripts, links to information, and other forms of communication data that provide a specific package of functionality, usually in a specific domain. In other words, data managers include components for extending the functionality to a new domain. Further, data managers and their associated data may be updated remotely over a network as new behavior is added or new information becomes available. Data managers may use system resources and the services of other, typically more specialized, data managers. Data

managers may be distributed and redistributed in a number of ways including on removable storage media, transfer over networks or attached to emails and other messages. An update manager may be used to add new data managers to the system or update existing data managers.

[15] According to another aspect of the invention, license management capabilities allowing the sale of data managers by third parties to one or more users on a one time or subscription basis may be provided. In addition, users with particular expertise may create data managers, update existing data managers by adding new behaviors and information, and making these data managers for other users as agents.

[16] In order to enhance the natural query and response environment, the system may format results to increase understandability to users. Formatting and presentation of results may be based on the context of the questions, the contents of the response being presented, the history of the interaction with the user, the user's preferences and interests and the nature of the domain. By contrast, rigid, highly formatted, or structured presentation of results may be deemed unnatural by many users.

[17] According to another embodiment of the invention, the system may simulate some aspects of a human "personality". In some cases, the presentation of the response and the terms that are used to provide the response may be randomized to avoid the appearance of rigidly formatted or mechanical responses. The use of other simulated personality characteristics is also desirable. For example, a response that may be upsetting to the user may be presented in a sympathetic manner. Furthermore, results of requests may be long text strings, lists, tables or other lengthy sets of data. Natural presentation of this type of information presents particular challenges because simply reading the long response is generally not preferred. Instead, the system may parse important sections from the response and may initially provide only reports. Determining what parts of a long response are presented may be based on the context of the questions, the contents of the response being presented, the history of the interaction with the user, the user's preferences and interests and the nature of the domain. At the same time, the system may give the user interactive control over what information to present and how much information to present, to stop the response all together, or to take other actions.

[18] According to another aspect of the invention, the system may process and respond to questions, requests and/or commands. Keywords or context may be used to determine whether the received utterance and/or textual message includes a request or command. For example, utterances may include aspects of questions, requests and/or commands. For

example, a user may utter "record my favorite TV program". A request is processed to determine the name, the channel, and time for the users favorite TV program. A command must be executed to set a video recorder to capture this program.

[19] For utterances including questions and/or requests, the system may perform multiple steps that may include one or more of:

- capturing the user's questions and/or requests through speech recognition components that operate in a variety of real-world environments;
- parsing and interpreting the question and/or request;
- determining the domain of expertise and context, invoking the proper resources, including agents;
- formulating one or more requests to one or more local and/or network data sources or sending appropriate commands to local or remote devices or the system itself;
- performing presentation formatting, variable substitutions and transformations to modify the requests to a form that yields desired results from the available sources;
- executing the multiple requests or commands in an asynchronous manner and dealing gracefully with failures;
- extracting or scraping the desired information from the one or more results, which may be returned in any one of a number of different formats;
- evaluating and interpreting the results, including processing of errors to provide one or more results judged to be "best," even if the results are ambiguous, incomplete, or conflicting;
- performing formatting, variable substitutions and transformations to modify the results to a form most easily understood by the user; and
- presenting the compound result, through a text to speech engine or a multimodal interface, to the user in a useful and expected manner.

[20] The above steps may be performed with knowledge of the domain of expertise, the context for the question or command, domain specific information, the history of the user's interactions, user preferences, available information sources or commands, and responses obtained from the sources.

[21] Probabilistic or fuzzy set decision and matching methods may be applied to deal with inconsistent, ambiguous, conflicting and incomplete information or responses. In addition,

asynchronous queries may be used to provide rapid and graceful failure of requests or commands that allow the system to robustly return results quickly and in a manner that seems natural to the user.

[22] Many everyday questions are inherently subjective and result in answers that are a matter of opinion or consensus, as much as fact. Such questions are often ad hoc in their nature, as well. The system, according to another aspect of the invention, may use adaptive, probabilistic, and fuzzy set decision and matching methods to identify the subjective nature of the question and to evaluate a range of possible answers, wherein one or more answers may be selected that most accurately represent the type of result desired by the user.

[23] The context and expected results from a particular question may be highly dependent on the individual asking the question. Therefore, the system may create, store and use personal profile information for each user. Information in the profile may be added and updated automatically as the user uses the system or may be manually added or updated by the user or by others. Domain specific agents may collect, store and use specific profile information, as may be required for optimal operations. Users may create commands for regularly used reports, automatically generated alerts, and other requests and for the formatting and presentation of results. The system may use profile data in interpreting questions, formulating requests, interpreting request results and presenting answers to the user. Examples of information in a user profile includes, history of questions asked, session histories, formatting and presentation preferences, special word spelling, terms of interest, special data sources of interest, age, sex, education, location or address, place of business, type of business, investments, hobbies, sports interests, news interests and other profile data.

[24] According to one aspect of the invention, the system may attempt to provide rapid responses in order to provide a natural question and response environment. The rapid responses may be provided without obtaining additional information. The system may determine agent composition, context and/or domain for a user's question or command, for example, by using a real-time scoring system or other technique. Based on this determination, the system may trigger one or more agents to respond to the user's question or command. The agents may make one or more requests and rapidly return a formatted response. Thus, users may receive direct responses to a set of questions, each with a different response or context. In some cases, the available information, including the request results, may not adequately answer the questions presented. In such situations, the user may be asked one or more follow-up questions to resolve the ambiguity. Additional requests may then be made before an adequate response is provided. In these cases, the system may use context

information, user profile information and/or domain specific information to minimize the interaction with the user required to deliver a response.

[25] If the confidence level of the domain or context score is not high enough to ensure a reliable response, the system may request that the user verify that the question or command is correctly understood. In general, the question may be phrased to indicate the context of the question including all criteria or parameters. If the user confirms that the question is correct, the system may proceed to produce a response. Otherwise, either the user can rephrase the original question, perhaps adding additional information to remove ambiguity, or the system may ask one or more questions to attempt to resolve the ambiguity or other actions may taken.

[26] According to one aspect of the invention, the system may accept any natural language question or command and, as a result, may be subject to ambiguous requests. To assist users in formulating concise questions and commands, the system may support a voice query language. The language may help users clearly specify the keywords or contexts of the question or command along with the parameters or criteria. The system may provide built in training capabilities to help the user learn the best methods to formulate their questions and commands.

[27] In order to make the responses to user's questions and commands seem more natural, the system may employ one or more dynamically invokable personalities and/or emotional models. Personalities and emotional models have specific characteristics that simulate the behavioral characteristics of real humans. Examples of these characteristics include, sympathy, irritation, and helpfulness and associated emotions. The personality also randomizes aspects of responses, just as a real human would do. This behavior includes randomization of terms used and the order of presentation of information. Characteristics of the personality and/or emotions are invoked using probabilistic or fuzzy set decision and matching methods and using criteria, including the context for the question, the history of the user's interaction, user preferences, information sources available, responses obtained from the sources.

[28] According to another aspect of the invention, special procedures may be employed to present information in the form of long text strings, tables, lists or other long response sets. Simply presenting a long set of information in an ordered manner may not be considered natural or what most users have in mind. The system may use probabilistic or fuzzy set matching methods to extract relevant information and present these subsets first. Further the

system may provide commands allowing users to skip through the list, find keywords or key information in the list or stop processing the list altogether.

[29] In addition to the ability to process user questions and commands, the invention may provide local or remote control functions for the system or for other devices. Users may initiate commands locally or remotely. Several techniques for executing remote operations are possible, for example, the use of a telephone or other audio connection. Other remote command techniques may be used. The processing of these commands may be performed in a nearly identical manner to a question. One difference being that the result of the command is generally an action rather than a response. Examples of control operations supported by the invention include, finding and playing music, film, games or other entertainment media, control of the appliances or other electronic or electro-mechanical devices, security systems control, audio, text and video conferencing (including Voice over IP conferencing), operating integrated messaging environments and other control operations.

[30] By using an optional analog or digital telephone interface or a data network interface, the invention can provide remote control capability of the computer or and auxiliary devices connected to the computer. For example, if the computer is connected via an interface to one or more auxiliary devices, a remote user can use this capability to remotely control or program the operation of the auxiliary device. For example, this enables recording of a television program, checking the status of a home security system, checking voice mail messages, hearing and responding to email messages, and much more. For each user command utterance, the system may execute a number of steps possibly including:

- capturing the user's command through speech recognition that operates in a variety of real-world environments;
- parsing and interpreting the command;
- determining the domain for the command and context, invoking the proper resources, including agents;
- formulating device specific commands for the system or external devices;
- routing commands to the system, or external devices, including external devices connected to data networks;
- receiving and processing results of commands, including errors; and
- optionally, providing responses to users indicating the success or failure of the command, and possibly including state information.

[31] The system may be used in conjunction with a wide range of platform environments. These environments may include, as a peripheral to a PC, a portable or wearable computer or other computer, embedded in a PC or other computer, on a personal digital assistant, such as a PDA with wireless networking capability, in a specialized device such as a bar code scanner or payment terminal, on a wireless telephone, or other platforms. If the invention is used with a mobile or portable device that has position location capability, the location data can be used by the invention as part of the context for user questions. A user may use the system on one or more devices. In this case, history and profile information for the user may be synchronized between the multiple devices on a periodic basis or other basis.

[32] According to another aspect of the invention, the system may be deployed in a network of devices that use a common base of agents, data, information, user profiles and histories. Each user can then interact with, and receive, the same services and applications at any location equipped with the required device on the network. For example, multiple devices on which the invention is deployed, and connected to a network, can be placed at different locations throughout a home, place of business or other location. In such a case, the system may use the location of the particular device addressed by the user as part of the context for the questions asked.

[33] According to one embodiment of the invention, the system may support multiple users that access the system at different times. According to another embodiment of the invention, the system may support multiple users that access the system during a same session in an interleaved or overlapping manner. The system may recognize the multiple users by name, voice, or other characteristic and may invoke a correct profile for each user. If multiple users are addressing the system in overlapping or interleaved sessions, the system may identify the multiple users and may invoke one or more corresponding profiles. For applications that require security safeguards, the multiple users may be verified using voiceprint matching, password or pass-phrase matching, or other security safeguards.

[34] When multiple users are engaged in interleaved sessions, the system may gracefully resolve conflicts using a probabilistic or fuzzy set decision method for each user. This process may simulate the manner in which a human would address multiple questions from various sources. For example, the system may answer short questions first in time while answering longer questions later in time. Alternatively, the system may answer questions in the order that they are received, among other configurations.

[35] Given the desire for domain specific behavior, user specific behavior and domain specific information, the system, according to another aspect of the invention, may allow both

users and content providers to extend the system capabilities, add data to local data sources, and add references to network data sources. To allow coverage of the widest possible range of topics, the system may allow third party content developers to develop, distribute and sell specialized or domain specific system programs and information. Content may be created though creation of new data managers, scripting existing data managers, adding new data to agents or databases and adding or modifying links to information sources. Distribution of this information is sensitive to the users interests and use history and to their willingness to pay or not.

[36] According to another aspect of the invention, the system may allow users to disseminate their knowledge without programming. The system may include mechanisms that allow users to post and distribute data managers and information in their particular areas of expertise, to improve system capability. Further, users can extend the system and configure it to their own preferences, add information to their profile to define new questions or queries, extend and modify existing questions and queries, add new data sources, update data sources, set preferences and specify presentation parameters for results.

[37] Since the invention may operate in many environments, including environments with background noise, point noise sources and people holding conversations, filtering of speech input may be performed. The invention may use, for example, either one-dimensional or two-dimensional array microphones to receive human speech. The array microphones may be fixed or employ dynamic beam forming techniques. The array pattern may be adjusted to maximize gain in the direction of the user and to null point noise sources. Speech received at the microphones may then be processed with analog or digital filters to optimize the bandwidth, cancel echoes, and notch-out narrow band noise sources. Following filtering, the system may use variable rate sampling to maximize the fidelity of the encoded speech, while minimizing required bandwidth. This procedure may be particularly useful in cases where the encoded speech is transmitted over a wireless network or link.

[38] Some example applications for the invention includes but are not limited to the following:

- white pages and yellow pages lookups to find, email addresses, telephone numbers, street addresses and other information for businesses and individuals;
- personal address book, calendars and reminders for each user;

- automatic telephone dialing, reading and sending emails and pages by voice and other communications control functions;
- map, location and direction applications;
- movie or other entertainment locator, review information and ticket purchasing;
- television, radio or other home entertainment schedule, review information and device control from a local or remote user;
- weather information for the local area or other locations;
- stock and other investment information including, prices, company reports, profiles, company information, business news stories, company reports, analysis, price alerts, news alerts, portfolio reports, portfolio plans, etc.;
- flight or other scheduled transportation information and ticketing;
- reservations for hotels, rental cars and other travel services;
- local, national and international news information including headlines of interest by subject or location, story summaries, full stories, audio and video retrieval and play for stories.;
- sports scores, news stories, schedules, alerts, statistics, back ground and history information, etc.;
- ability to subscribe interactively to multimedia information channels, including sports, news, business, different types of music and entertainment, applying user specific preferences for extracting and presenting information;
- rights management for information or content used or published;
- horoscopes, daily jokes and comics, crossword puzzle retrieval and display and related entertainment or diversions;
- recipes, meal planning, nutrition information and planning, shopping lists and other home organization related activities;
- as an interface to auctions and online shopping, and where the system can manage payment or an electronic wallet;
- management of network communications and conferencing, including telecommunications, email, instant messaging, Voice over IP communications and conferencing, local and wide area video and audio conferencing, pages and alerts;

- location, selection, management of play lists and play control of interactive entertainment from local or network sources including, video on demand, digital audio, such as MP3 format material, interactive games, web radio and video broadcasts;
- organization and calendar management for families, businesses and other groups of users including the management of, meetings, appointments, and events; and
- interactive educational programs using local and network material, with lesson material level set based on user's profile, and including, interactive multimedia lessons, religious instruction, calculator, dictionary and spelling, language training, foreign language translation and encyclopedias and other reference material.

[39] It will be appreciated that the foregoing statements of the features of the invention are not intended as exhaustive or limiting, the proper scope thereof being appreciated by reference to this entire disclosure and reasonably apparent variations and extensions thereof.

BRIEF DESCRIPTION OF THE DRAWINGS

[40] The invention will be described by reference to the preferred and alternative embodiments thereof in conjunction with the drawings in which:

[41] FIG. 1 is an overall diagrammatic view according to one embodiment of the invention;

[42] FIG. 2 is a schematic block diagram showing the agent architecture according to one embodiment of the invention;

[43] FIG. 3 is a high level process for receiving natural language speech-based queries and/or commands and generating a response according to one embodiment of the invention;

[44] FIG. 4A is a process for receiving natural language speech-based queries and/or commands and generating a response according to one embodiment of the invention;

[45] FIG. 4B is a process for receiving natural language speech-based commands in order to locally or remotely control functions of devices according to one embodiment of the invention;

[46] FIG. 5 is a process for correctly interpreting a user's utterance according to one embodiment of the invention; and

[47] FIG. 6 is a process for determining the proper domain agents to invoke and the properly formatted queries and/or commands that is to be submitted to the agents according to one embodiment of the invention.

[48] FIG. 7 illustrates a diagram for centrally maintaining context, according to one embodiment of the invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[49] Barriers to natural human request-response interaction between humans and machines include that natural language queries may be incomplete in their definition. Furthermore, context of previous questions, knowledge of the domain, and/or the user's history of interests and preferences may be needed to properly interpret questions and/or commands. Thus, natural language questions may not be easily transformed to machine processable form. Compounding this problem, natural language questions may be ambiguous or subjective. In these cases, forming a machine processable query and returning a natural language response may be difficult.

[50] In order for machines (e.g., computer devices) to properly respond to questions or commands that are submitted in a natural language form, machine processable queries and algorithms may be formulated after the natural form questions or commands have been parsed and interpreted. Algorithms describe how the machines should gather data to respond to the questions or commands. Depending on the nature of the questions or commands, there may not be a simple set of queries and algorithms that will return an adequate response. Several queries and algorithms may need to be initiated and even these queries and algorithms may need to be chained or concatenated to achieve a complete response. Further, no single available source may contain the entire set of results needed to generate a complete response. Thus, multiple queries and/or algorithms, perhaps with several parts, may be generated to access multiple data sources that are located both locally or remotely. Not all of the data sources, queries and/or algorithms may return useful results or any results at all.

[51] Useful results that are returned are often embedded in other information and may need to be extracted from the other information. For example, a few key words or numbers may need to be "scraped" from a larger amount of other information in a text string, table, list, page, video stream or other information. At the same time, extraneous information including graphics or pictures may be removed to process the response. In any case, the multiple results must be evaluated and combined to form the best possible response, even in cases where some queries do not return useful results or fail to produce results entirely. In cases where the question is determined to be ambiguous or the result is inherently subjective,

determining the results to present in the response is a complex process. Finally, to maintain a natural interaction, responses need to be returned rapidly to the user. Managing and evaluating complex and uncertain queries, while maintaining real-time performance, is a significant challenge.

[52] The following detailed description refers to the accompanying drawings, and describes exemplary embodiments of the invention. Other embodiments are possible and modifications may be made to the exemplary embodiments without departing from the spirit, functionality and scope of the invention. Therefore, the following detailed descriptions are not meant to limit the invention.

[53] The invention provides a complete speech-based information query, retrieval, processing and presentation environment or a combination of speech-based and non-speech-based information query, retrieval, processing and presentation environment. In addition, the invention may be useful for controlling the system itself and/or external devices. This integrated environment makes maximum use of context, prior information and domain and user specific profile data to achieve a natural environment for one or more users submitting requests or commands in multiple domains. Through this integrated approach, a complete speech-based natural language query, algorithm and response environment or a combination of speech-based and non-speech-based may be created. FIG. 1 shows one exemplary schematic diagram view of a system 90 according to an embodiment of the invention.

[54] The system 90 may include a main unit 98, a speech unit 128, and multi-modal device 155. Alternatively, the system 98 may comprise completely separate systems for main unit 98, speech unit 128 and multi-modal device 155. The event manager 100 may mediate interactions between other components of the main unit 98. The event manager 100 provides a multi-threaded environment allowing the system 98 to operate on multiple commands or questions from multiple user sessions, without conflict, and in an efficient manner, maintaining real-time response capabilities.

[55] Agents 106 may include a collection of grammars, criteria handlers, and algorithms that are accessed to respond to a set of requests and/or commands. Agents 106 further may contain packages of both generic and domain specific behavior for the system 98. Agents 106 may use nonvolatile storage for data, parameters, history information, and locally stored content provided in the system databases 102 or other local sources. One or more user profiles 110 may be provided that includes user specific data, parameters, and session and history information for determining the behavior of agents 106. One or more personality modules 108 may be provided in a data determining system that includes personality

characteristics for agents 106. The update manager 104 manages the automatic and manual loading and updating of agents 106 and their associated data from the Internet 146 or other network through the network interface 116.

[56] According to one embodiment of the invention, the speech-based interface for the system 90 may include one or more speech units 128. Speech units 128 may include one or more microphones, for example array microphone 134, to receive the utterances from the user. The speech received at the microphone 134 may be processed by filter 132 and passed to the speech coder 138 for encoding and compression. In one embodiment, a transceiver module 130 transmits the coded speech to the main unit 98. Coded speech received from the main unit 98 is detected by the transceiver 130, then decoded and decompressed by the speech coder 138 and annunciated by the speaker 136.

[57] According to one embodiment of the invention, the non-speech-based interface for the system 90 may include one or more multi-modal devices 155 that may include mobile devices, stand alone or networked computers, personal digital assistances (PDAs), portable computer devices, or other multi-modal devices.

[58] The speech units 128, multi-modal devices 155 and the main unit 98 may communicate over a communication link. The communication link may include a wired or wireless link. According to one embodiment, the communication link may comprise an RF link. The transceiver 130 on the speech unit may communicate coded speech data bi-directionally over the communication link with the transceiver 126 on the main unit 98. According to another embodiment, RF link may use any standard local area wireless data protocols including the IEEE 802.11, Bluetooth or other standards. Alternatively, an infrared data link conforming to any suitable standard such as IrDA or other infrared standards can be used. In an alternative embodiment, wires connect the speech unit 128 and the main unit 98, eliminating the need for one speech coder 138. Other wired or wireless analog or digital transmission techniques can be used.

[59] According to one embodiment of the invention, coded speech received at the transceiver 126 on the main unit 98 may be passed to the speech coder 122 for decoding and decompression. The decoded speech may be processed by the speech recognition engine 120 using the context description grammar module 112, among other information. Any recognized information may be processed by the parser 118, which transforms information into complete algorithms and questions using data supplied by knowledge agents. Knowledge agents may be selected from the plurality of agents 106 using a grammar stack, wherein the knowledge agents provide information for generating a response to the question

or command. The knowledge agents may then process the commands or questions by creating queries that are submitted to local databases 102 or submitted through the network interface 116 to external data sources over the Internet 146 or external other networks. Algorithms typically result in actions taken by the system 90 itself (i.e., pause or stop), or to a remote device or data source (i.e., download data or program, or control a remote device), through the network interface to the Internet or other data interface.

[60] According to one embodiment of the invention, knowledge agents may return results of questions as responses to users. The responses may be created using the results of information queries, the system personality 108, the user preferences, other data in the user profile 110, and/or other information. Agents 106 may present these results using the speech unit 128. The agents 106 may create a response string, which is sent to the text to speech engine 124. The text to speech engine 124 generates the utterances, which may be encoded and compressed by the speech coder 122. Once coded, the utterances are transmitted from the main unit 98 by the transceiver 126 to the transceiver 130 on the speech unit 128. The utterance is then decoded and decompressed by the speech coder 138 and output by the speaker 136. Alternatively, agents 106 may present the results using multi-modal devices 155.

[61] The non-speech interface 114 may be part of, or separate from, the multi-modal devices 155 and may be used as a substitute or may complement the speech interface. For example, non-speech interface 114 may be used to present and to interact with non-speech (e.g., graphical or tabular) information in a manner more easily digested by the user. According to one embodiment of the invention, multimodal support may be provided to maintain the context during both voice interaction and interaction through the non-speech interface 114. In one exemplary embodiment, a communication channel may be opened between multimodal devices 155 and the main user interface system 90 to allow multimodal devices 155 to input text commands and questions. The multimodal devices 155 may receive and/or provide one or more types of communications including speech, text, digital audio files, and other communications. According to another embodiment of the invention, multimodal devices 155 may send a string of text or keywords for a command or question. The main interface system 90 may synchronize the context between multimodal device 155 and the speech units 128. In order to send a response to the corresponding device, the main user interface system 90 may track where the input came from so that the response may be sent to a TTS or multi-modal device 155.

[62] According to an alternative embodiment of the invention illustrated in FIG. 7, a context manager 702 may be centrally maintained to allow input from and output to multiple mobile devices 36. Each of the mobile devices may communicate with the context manager 702 to register via registration module 712, wherein registration may indicate events that the mobile devices 36 may subscribe to. The context manager 702 may receive input in Context XML form, for example. The other registered mobile devices 36 may be informed of a context changes through context tracking module 714, thereby synchronizing the context across the registered mobile devices 36. According to one embodiment of the invention, registered mobile devices 36 may be added or removed. The registered mobile devices 36 may be dynamic link libraries (DLLs) that may be specific to the mobile devices 36.

[63] According to yet another alternative embodiment of the invention, context information may be determined from a command or request that is presented as a text message and/or a command or request that is presented as a verbal utterance and processed using a multi-pass automatic speech recognition module that transcribes the verbal utterance to a text message. The multi-pass automatic speech recognition module may use a dictation grammar or a large vocabulary grammar, among other resources, to transcribe the verbal utterance into a text message. On platforms where a dictation grammar is not available, the multi-pass automatic speech recognition may use a virtual dictation grammar that uses decoy words for out-of-vocabulary words. Decoy words may include utility words, nonsense words, isolated syllables, isolated distinct sounds and other decoy words.

[64] According to one embodiment of the invention, the text message may be searched for a particular character, group of characters, word, group of words, and other text combinations. The text combination may be compared against entries in a context description grammar that is associated with each agent 106. If a match is identified between an active grammar in the context description grammar and the command and/or request, then the match may be scored. The agents 106 may be ranked based on the determined score. In generating an aggregate response from the one or more responses received from the agents 106, the ordering of the responses from the individual agents may be determined based on the rank of agents 106. The aggregate response may be generated by a response generator module. Agents 106 may update a context stack, that includes an ordered list of command contexts, to enable follow-up requests.

[65] According to another embodiment of the invention, if a match is not found, or only a partial match is found, between the text message and active grammars, then a knowledge-enhanced speech recognition system may be used to semantically broaden the search. The

knowledge-enhanced speech recognition system may be used to determine the intent of the request and/or to correct false recognitions. The knowledge-enhanced speech recognition may access a set of expected contexts that are stored in a context stack to determine a most likely context. The knowledge-enhanced speech recognition may use context specific matchers that are able to identify context such as time, location, numbers, dates, categories (e.g., music, movies, television, addresses, etc.) and other context. The matching may be performed by comparing a character, group of characters, a word, group of words, and other text combinations. Alternatively, or in addition to text based matching, the matching may be performed using phonetic matching, among other techniques. The results of any match may be used to generate a command and/or request that is communicated to agents 106 for additional processing. According to one embodiment of the invention, non-speech interface 114 may show system, state and history information in a more concise manner than is possible through the speech interface. Non-speech interface 114 may be accessed to create or extend capabilities of agents 106. These operations may include scripting of agents, adding data to the agent or databases 102 used by the agent, adding links to information sources, among other operations.

[66] According to another embodiment of the invention, system 90 may include different types of agents 106. For example, generic and domain specific behavior and information may be organized into domain agents. A system agent, on the other hand, may provide default functionality and basic services. The domain agents provide complete, convenient and re-distributable packages or modules for each application area. In other words, the domain agents include data that is needed to extend or modify the functionality of the system 90 in a current or new domain. Further, domain agents and their associated data can be updated remotely over a network as new behavior is added or new information becomes available. Domain agents may use the services of other, typically more specialized data managers and the system agent 150. Agents are distributed and redistributed in a number of ways including on removable storage media, transfer over networks or attached to emails and other messages. The invention may provide license management capability allowing the sale of data managers by third parties to one or more users on a one time or subscription basis. In addition, users with particular expertise may create data managers, update existing data managers by adding new behaviors and information and making these data managers to other users as agents. A block diagram of an agent architecture according to an embodiment of the invention is shown in FIG. 2.

[67] Agents 106 may receive and return events to the event manager 100. Both system agents 150 and domain agents 156 may receive questions and commands from the parser 118. Based on keywords in the questions and commands and the structures of the questions and commands, the parser may invoke selected agents. Agents use the nonvolatile storage for data, parameters, history information and local content provided in the system databases 102.

[68] According to one embodiment of the invention, when the system starts-up or boots-up, the agent manager 154 may load and initialize the system agent 150 and the one or more domain agents 156. Agent manager 154 includes knowledge of agents 106 and maps agents 106 to the agent library 158. At shutdown, the agent manager may unload the agents 106. The agent manager 154 also performs license management functions for the domain agents 156 and content in the databases 102.

[69] The system agent 150 manages the criteria handlers 152 that handle specific parameters or values (criteria) used to determine context for questions and commands. According to one embodiment of the invention, criteria handlers 152 include parsing routines that are specialized to recognize particular parts of speech, such as times, locations, movie titles, and other parts of speech. The criteria handlers 152 may identify matching phrases and extract semantic attributes from the phrases. Both the system agent 150 and the domain agents 156 may use the criteria handlers 152. The various domain agents 156 may use the services of the system agent 150 and of other, typically more specialized, domain agents 156. The system agent 150 and the domain agents 156 may use the services of the agent library 158, which contains utilities for commonly used functions. According to one embodiment of the invention, the agent library may be a dynamic link library that implements one or more agents. The agent library may include utilities for text and string handling, network communications, database lookup and management, fuzzy and probabilistic evaluation, text to speech formats, and other utilities.

[70] Domain agents 156 may be data-driven, scripted or created with compiled code. A base of generic agent may be used as the starting point for data-driven or scripted agents. Agents created with compiled code are typically built into dynamically linkable or loadable libraries. Developers of agents can add new functionality to the agent library 158 as required. Details of agent distribution and update, and agent creation or modification are discussed in sections below.

[71] According to another embodiment of the invention, capabilities are provided to distribute and update system agents 150, domain agents 156, agent library 158 components, databases 102, and context description grammar 112 over wireless or wired networks 136,

including dial-up networks using the update manager 104. The network interface 116 may provide connections to one or more networks. The update manager 104 may also manage the downloading and installation of core system updates. The agent manager 154 may perform license management functions for the domain agents and the databases. The update manager 104 and agent manager 154 may perform these functions for all agents and database content including, agents and content available to all users or agents and/or content available to selected users. Examples of agent and database components added or updated on a periodic basis include:

- agents for new domains;
- additional domain knowledge for agents;
- new keywords for a domain, which can include names of politicians, athletes, entertainers, names of new movies or songs, etc. who have achieved recent prominence;
- links to a preferred set of information sources for the domains covered including links for, entertainment, news, sports, weather, etc.;
- updates to domain information based on, for example, changes to tax laws, company mergers, changing political boundaries;
- updates to content, including dictionaries, encyclopedias and almanacs; and
- other content and database components.

[72] When a user requires or selects a new domain agent 156 or database element 102, the update manager 104 may connect to their source on the network 146 though the network interface 116, download and install the agent or data. To save system resources and to comply with any license conditions, the update manger 104 may uninstall agents that are no longer in use. In one embodiment of the invention, the update manager 104 may periodically query one or more sources of the licensed agents and database components to locate and download updates to agent executables, scripts or data as they become available. Alternatively, the agent sources may initiate the downloading of agent updates of the registered or licensed agents to the update manager as they become available.

[73] The agent manager 154 may provide license management clients that are capable of executing most any license terms and conditions. When a particular agent 106 and/or database element 102 is selected based on a submitted command, the agent manager 154 verifies that the use of the agent or data element is within the allowed terms and conditions, and if so, invokes the agent or allows access to the data element. License management

schemes that can be implemented through the agent manager 154 include outright purchase, subscription for updates, one time or limited time use. Use of shared agents and data elements (such as those down-loaded from web sites maintained by groups of domain experts) may also be managed by the agent manager 154.

[74] If questions or commands do not match an agent that is currently loaded on the system, the agent manager 154 may search the network 146 through the network interface 116 to find a source for a suitable agent. Once located, the agent can be loaded under the control of the update manager 104, within the terms and conditions of the license agreement, as enforced by the agent manager.

[75] New commands, keywords, information, or information sources can be added to any domain agent 156 by changing agent data or scripting. These configuration capabilities may allow users and content developers to extend and modify the behavior of existing domain agents 156 or to create new domain agents 156 from a generic agent without the need to create new compiled code. Thus, the modification of the domain agents 156 may range from minor data-driven updates by even the most casual users, such as specifying the spelling of words, to development of complex behavior using the scripting language as would typically be done by a domain expert. The user can create and manage modifications to domain agents 156 through speech interface commands or using non-speech interface 114. User-specific modifications of domain agents 156 are stored in conjunction with the user's profile 110 and accessed by the domain agent 156 at run-time.

[76] The data used to configure data driven agents 156 are structured in a manner to facilitate efficient evaluation and to help developers with organization. These data are used not only by the agents 156, but also by the speech recognition engine 120, the text to speech engine 124, and the parser 118. Examples of some major categories of data include:

1. Content packages include questions or commands. Each command or question or group of commands or questions includes contexts used for creation of one or more requests. The domain agent 156 passes a context description grammar expression to the parser 118 for evaluation of a context or question. An initial or default context is typically supplied for each command or question. The command or question may include a grammar for the management and evaluation of the context stack.

2. Page lists or pointers to other local or network content sources. For each page or content source there is a pointer (e.g. URL, URI, or other pointer) to the page or source. Each page has specific scraping information used to extract the data of

interest. The scraping information includes matching patterns, HTML or other format parsing information.

3. A response list, determining the response of the domain agent 156 to a particular command or question given the context, the user profile and the information retrieved. Responses can include diagnostic error messages or requests for more information if the question or command cannot yet be resolved from the known information. Responses can be based on or dependent on thresholds or probabilistic or fuzzy weights for the variables.

4. Substitution lists containing variable substitutions and transformations, often applied by the agents 150, 156 in the formatting of queries and results. For example, a stock domain specific agent 156 would use a substitution list of company trading symbols, company names and commonly used abbreviations. Substitutions and transformations can be performed on commands and questions to create precise queries, which can be applied against one or more information sources or to results for creating more meaningful output to the user. Substitution lists also include information for optimally dealing with structured information, such as HTTP formatted page parsing and evaluation.

5. Personalities used for responses. Personalities are constructed by combining multiple traits in a weighted manner. Weights can be specified for each agent's domain area to create one or more specific personalities. Examples of personality traits include sarcasm, humor, irritation, and sympathy, and other traits.

6. Public and user specific parameters for sources, substitutions, transformations, variables or criteria. The public parameter lists are part of the agent package 156. The user specific parameters are contained in the user profile 110.

[77] Commands and questions are interpreted, queries formulated, responses created and results presented based on the users personal or user profile 110 values. Personal profiles may include information specific to the individual, their interests, their special use of terminology, the history of their interactions with the system, and domains of interest. The personal profile data may be used by the agents 106, the speech recognition engine 120, the text to speech engine 124, and the parser 118. Preferences can include, special (modified) commands, past behavior or history, questions, information sources, formats, reports, and alerts. User profile data can be manually entered by the user and/or can be learned by the system 90 based on user behavior. User profile values may include:

- spelling preferences;
- date of birth for user, family and friends;
- income level;
- gender;
- occupation;
- location information such as, home address, neighborhood, and business address;
- car type;
- telecommunications and other service providers and services;
- financial and investment information;
- synonyms (i.e., a nick name for someone);
- special spelling;
- keywords;
- transformation or substitution variables;
- domains of interest; and
- other values.

[78] End users may use the data driven agent 156 extension and modification facilities and values stored in user profiles 110 to create special reports, packages of queries, alerts and output formats. A single alert or report can be configured to use multiple data sources and other variable (i.e., time, location) value to determine when alerts should be sent. For example, an alert can be generated by sampling a stock price every 15 minutes and sending an alert if the price drops below some value. To create a report, the user first specifies a set of commands or questions. Next, the user creates or selects a format for the report. Finally the user may name the report. A report can have variable parameters. For example, users may create a company stock report, and execute the report by stating its name and the company name, which gives the user selected information and in a specified format for that company. In another example, users can create a "morning" report, which presents selected multimedia information from different sources (news, sports, traffic, weather) in the order and formats desired. Alerts and reports can be created using only voice commands and responses, commands and responses through non-speech interface 114, or a combination of the two. To create a report, alert, or other specialized behavior, the user performs a number of steps including:

- specify the command to run a report or alert;

- specify the question or questions, including keywords, used for a query;
- set the criteria for running the report such as on command or when a particular condition is met;
- define preferred information sources;
- define preferences for order of result evaluation by source, value, etc.;
- specify the presentation medium for a report or alert, such as an email, the text to speech engine, a message to a pager, or a text and graphics display; and
- specify the preferred format for the report, such as information to be presented, order of information to be presented, preferred abbreviations or other variable substitutions.

[79] Filtering and noise elimination may be important in facilitating the various functionalities of the system 90. Recognition and parsing of the user's speech is implemented with good signal to noise ratio at the input to the speech recognition engine 120. To provide acceptable results, a set of acoustic models, an array microphone 134, a filter 132, or other components, may be employed. If a good signal to noise ration cannot be achieved, a noise identification algorithm may be used and the appropriate acoustic model, for example, one that has been trained in conditions similar to the identified noise, may be selected. According to an embodiment of the invention, the microphone array, filters and speech coder 138 are physically separated from the main unit 98 into a speech unit 128, and connected using a wireless link. Since bandwidth on a wireless connection is at a premium, the speech coder dynamically adapts the digitization rate and compression of the captured speech.

[80] The speech unit 128 may use an array of microphones 134 to provide better directional signal capture and noise elimination than can be achieved with a single microphone. The microphone array can be one-dimensional (a linear array) or two-dimensional (a circle, square, triangle or other suitable shape). The beam pattern of the array can be fixed or made adaptive though use of analog or digital phase shifting circuitry. The pattern of the active array is steered to point in the direction of the one or more users speaking. At the same time, nulls can be added to the pattern to notch out point or limited area noise sources. The use of the array microphone also helps reduce the cross talk between output from the text to speech engine 124 through the speaker 136 and detection of the user's speech.

[81] The microphone can be complemented with an analog or digital (i.e., Voice over IP) speech interface. This interface allows a remote user to connect to the system and interact with the system in the same manner possible if they were physically present.

[82] The speech unit 128, according to an embodiment of the invention, may use an analog or digital filter 132 between the array microphone 134 and the speech coder 138. The pass band of the filter is set to optimize the signal to noise ratio at the input to the speech recognition engine 120. In some embodiments, the filter is adaptive, using band shaping combined with notch filtering to reject narrow-band noise. In one embodiment of the invention, the system 90 may employ adaptive echo cancellation in the filter. The echo cancellation helps prevent cross talk between output from the text to speech engine and detection of the user's speech as well as suppression of environmentally caused echoes. Calculations are conducted for comparing the background noise to the signal received from the users speech are used to optimize the band-shaping parameters of the adaptive filter. Those skilled in the art will be familiar with multiple techniques used to construct suitable digital and analog filters.

[83] The speech received by the array microphone 134 and passed through the filter 132 may be sent to the speech digitizer or coder 138. The speech coder 138 may use adaptive lossy audio compression to optimize bandwidth requirements for the transmission of the coded speech to the speech recognition engine 120 over a wireless link. The lossy coding may be optimized to preserve only the components of the speech signal required for optimal recognition. Further, the lossy compression algorithms that may be used may be designed to prevent even momentary gaps in the signal stream, which can cause severe errors in the speech recognition engine. The digitized speech may be buffered in the coder and the coder may adapt the output data rate to optimize the use of the available bandwidth. The use of the adaptive speech coder is particularly advantageous when a band-limited wireless link is used between the coder and the speech recognition engine.

[84] In an alternative embodiment, the array microphone can be replaced by a set of physically distributed microphones or a headset worn by the user. The distributed microphones can be placed in different parts of a room or in different rooms. The distributed microphones can create a three-dimensional array to improve signal to noise ration. The headset may use a wireless or wired connection.

[85] While the invention is intended to be able to accept most any natural language question or command, ambiguity may be a problem. To assist users formulate concise questions and commands, the system 90 may support a voice query language. The language

is structured to allow a variety of queries with minimal ambiguity. Thus, the voice query language helps users clearly specify the keywords or contexts of the question or command along with the parameters or criteria. The language provides a grammar to clearly specify the keyword used to determine the context and a set of one or criteria or parameters. A user asking a question or stating a command in the voice query language is nearly always guaranteed to receive a response.

[86] The voice query language may be sensitive to the contents of the context stack, wherein a context defines a set of questions that can be activated or deactivated during a conversation. According to one embodiment, each agent may designate one context to be the root context that defines base algorithms that the agent implements. Thus, following-on questions can be asked using an abbreviated grammar, since key words and criteria can be inherited from the context stack. For example, the user can simply asked about another keyword if the criteria of the question remain constant.

[87] The system 90, may provide built in training capabilities to help the user learn the best methods to formulate their questions and commands. The interactive training allows users to audibly or visibly see the machine interpretation of their queries and provides suggestions on how to better structure a query. Using the interactive training users can quickly become comfortable with the voice query language and at the same time learn how to optimize the amount of information required with each step of a dialog.

[88] The output of the speech coder 122 may be fed to the speech recognition engine 120. The speech recognition engine 120 recognizes words and phrases, using information in the context description grammar 112, and passes these to the parser 118 for interpretation. The speech recognition engine 120 may determine the user's identity by voice and name for each utterance. Recognized words and phrases may be tagged with this identity in all further processing. Thus, as multiple users engage in overlapping sessions, the tags added by the speech recognition engine 120 to each utterance allows other components of the system 90 to tie that utterance to the correct user and dialog. The user recognition capability may further be used as a security measure for applications, such as auctions or online shopping, where this is required. Voice characteristics of each user may be contained in the user profile 110.

[89] According to one embodiment of the invention, users may start a dialog with the system 90 when they first address it. This can be done by speaking a generic word ("computer") or addressing a specific name ("Fred"), which may be generally tied to a system personality 108. Once the user starts the dialog, it may be recognized by the speech

recognition engine 120, using unique characteristics of the user's speech. At the end of a dialog or to interrupt a dialog, the user may utter a dismissal word ("good bye").

[90] According to another embodiment of the invention, the system 90 may employ a speech recognition engine 124 that gains improved word recognition accuracy using data from context description grammar 112, user profiles 110, and the agents 106, among other components. At the same time, the fuzzy set possibilities or prior probabilities for the words in the context description grammar may be dynamically updated to maximize the probability of correct recognition at each stage of the dialog. The probabilities or possibilities may be dynamically updated based on a number of criteria including the application domain, the questions or commands, contexts, the user profile and preferences, user dialog history, the recognizer dictionary and phrase tables, and word spellings, among other criteria.

[91] For uncommon words or new vocabulary words, a user may be given the option to spell the words. The spelling may be done by saying the names or the letters or using a phonetic alphabet. The phonetic alphabet can be a default one or one of the user's choosing.

[92] Alternatively, when a user submits a word that is not recognized at all or is not correctly recognized by the speech recognition engine 120 then the user may be asked to spell the word. The speech recognition engine 120 determines this condition based on confidence level for the scoring process. The word may be looked up in a dictionary and the pronunciation for the word is added to either the dictionary, the agent 106, or the user's profile 110. The word pronunciation can then be associated with the domain, the question, the context and the user. Through this process, the speech recognition engine learns with time and improves in accuracy. To assist users in spelling words, an individualized phonetic alphabet can be used. Each user can modify the standard phonetic alphabets with words, which they can remember more easily.

[93] Once the words and phrases have been recognized by the speech recognition engine 120, the tokens and user identification is passed to the parser 118. The parser 118 examines the tokens for the questions or commands, context and criteria. The parser 118 determines a context for an utterance by applying prior probabilities or fuzzy possibilities to keyword matching, user profile 110, and dialog history. The context of a question or command determines the domain and thereby, the domain agent 156, if any, to be evoked. For example, a question with the keywords "temperature" implies a context value of weather for the question. The parser dynamically receives keyword and associated prior probability or fuzzy possibility updates from the system agent 150 or an already active domain agent 156. Based

on these probabilities or possibilities the possible contexts are scored and the top one or few are used for further processing.

[94] The parser 118 uses a scoring system to determine the most likely context or domain for a user's question and/or command. The score is determined from weighing a number of factors including, the user profile 110, the domain agent's data content and previous context. Based on this scoring, the system 90 invokes the correct agent. If the confidence level of the score is not high enough to ensure a reliable response, the system 90 may ask the user to verify whether the question and/or command is correctly understood.

[95] In general, the question that is asked by the system 90 may be phrased to indicate the context of the question including all criteria or parameters. For example, the question can be in the form of: "Did I understand that you want such-and-such" If the user confirms that the question is correct, the system proceeds to produce a response. Otherwise, the user can rephrase the original question, perhaps adding additional information to remove ambiguity, or the system can ask one or more questions to attempt to resolve the ambiguity.

[96] Once the context for the question or command has been determined, the parser 118 can invoke the correct agent 156, 150. To formulate a question or command in the regular grammar used by agents, the parser 118 may determine required and optional values for the criteria or parameters. These criteria may have been explicitly supplied by the user or may need to be inferred. The parser 118 may make use of the criteria handlers 152 supplied by the system agent 150. The criteria handlers 152 may provide context sensitive procedures for extracting the criteria or parameters from the user's questions or commands. Some criteria are determined by executing algorithms in the agent, while others may be determined by applying probabilistic of fuzzy reasoning to tables of possible values. Prior probabilities or fuzzy possibilities and associated values are received from a number of sources including the history of the dialog, the user profile 110, and the agent. Based on user responses, the prior probabilities or fuzzy possibilities are updated as the system learns the desired behavior. For a weather context, examples of criteria include, location, date and time. Other criteria can include command criteria (i.e., yes/no, on/off, pause, stop), and spelling. Special criteria handlers are available from the system agent for processing lists, tables, barge-in commands, long strings of text and system commands.

[97] The criteria handlers 152 operate iteratively or recursively on the criteria extracted to eliminate ambiguity. This processing helps reduce the ambiguity in the user's question or command. For example, if the user has a place name (or other proper noun) in their utterance, the parser 118 can use services of the domain agent 156 to look up tables in the

databases 102 for place names or can attempt to determine which word is the proper noun from the syntax of the utterance. In another example, the user asks "what about flight one hundred and twenty too?". The parser and domain agent use flight information in the database and network information along with context to determine the most plausible interpretation among; flight 100 and flight 20 also, flight 100 and flight 22, flight 122, and the like.

[98] Once the context and the criteria are determined, the parser 118 may form the question or command in a standard format or hierarchical data structure used for processing by the agents 150, 156. The parser 118 may fill in all required and some optional tokens for the grammar of the context. Often the tokens must be transformed to values and forms acceptable to the agents. The parser obtains the required transformations from the agents, dialog history or user profile 110. Examples of transformations or substitutions performed by the parser on tokens include:

- substituting a stock symbol for a company name or abbreviation;
- substituting a numerical value for a word or words;
- adding a zip code to an address; and
- changing a place or other name to a commonly used standard abbreviation.

[99] The agents 150, 156 may receive a command or question once the parser 118 has placed it in the required standard format. Based on the context, the parser 118 evokes the correct agent to process the question or command.

[100] Commands can be directed to the system 90 or to an external entity. System commands are generally directed to the system agent 150. Commands for external entities are generally processed by a domain agent 156, which includes the command context and behavior for the external entity.

[101] Specific questions are generally directed to one of the domain agents 156. Based on the question or context and the parameters or criteria, the domain agent creates one or more queries to one or more local or external information sources. Questions can be objective or subjective in nature. Results for objective questions can often be obtained by structured queries to one or more local or network information sources. Even for objective questions, the system 90 may need to apply probabilistic or fuzzy set analysis to deal with cases of conflicting information or incomplete information. Information to answer subjective questions is generally obtained by one or more ad-hoc queries to local or network data

sources, followed by probabilistic or fuzzy set evaluation of the one results to determine a best answer.

[102] Once the domain agent 156 has formulated the one or more queries, they may be sent to local and/or network information sources. The queries are performed in an asynchronous manner to account for the fact that sources respond at different speeds or may fail to respond at all. Duplicate queries are sent to different information sources to ensure that at least one source responds with a useful result in a timely manner. Further, if multiple results are received in a timely manner, they can be scored by the system to determine which data is most reliable or appropriate. Examples of data sources accommodated include, HTTP data sources, sources with meta-data in various formats including XML, entertainment audio, video and game files including MP3, databases using query languages and structured responses such as SQL, and other data sources.

[103] The local information sources can be stored in one or more system databases 102 or can be on any local data storage such as a set of CDs or DVDs in a player or other local data storage. Network information sources can be connected to the Internet 136 or other network and accessed through a series of plug-ins or adaptors, known a pluggable sources, in the network interface 116. The pluggable sources in the network interface 116 may be capable of executing the protocols and interpreting the data formats for the data sources of interest. The pluggable sources may provide information, scrapping data and procedures for each source to the domain agents 156. If a new type of data source is to be used, a new plug-in or adaptor can be added to the network interface 116.

[104] The domain agent 156 evaluates the results of the one or more queries as they arrive. The domain agent 156 scores the relevance of the results based on results already received, the context, the criteria, the history of the dialog, the user profile 110 and domain specific information using probabilistic or fuzzy scoring techniques. Part of the dialog history is maintained in a context stack. The weight of each context for the scoring is based on the relevance of one context to another and the age of the contexts. Other scoring variables can be associated through the context stack. Contexts can also be exclusive, so that previous contexts have no weight in the scoring.

[105] Based on the continuous scoring processes, the domain agent 156 may determine if a single best answer can be extracted. For most questions, the desired result has a set of tokens that must be found to formulate an answer. Once a value has been found for each of these tokens, the results are ready for presentation to the user. For example, for a question on weather, the tokens can include the date, day of week, predicted high temperature, predicted

low temperature, chance of precipitation, expected cloud cover, expected type of precipitation and other tokens. Results processed in this manner include error messages. For subjective questions, this determination is made by determining a most likely answer or answers, extracted by matching of the results received. If no satisfactory answer can be inferred from the results of the query, the agent can do one of the following:

1. Ask the user for more information, typically through the speech interface, and based on the results obtained formulate new queries. This approach is applied when an irresolvable ambiguity arises in the formulation of a response.
2. Formulate new queries based on the results received from the first set of queries. This approach is typically applied in cases where the responses received do not contain all the required information. Information sources to query can inferred from the results already obtained (i.e., links in an HTML document) or from other sources. Using this approach one or more sets of queries and responses can be chained without the need for action by the user.
3. Wait for additional queries to return results.

[106] In any case, the domain agent 156 may continue to make requests and evaluate results until a satisfactory response is constructed. In doing so, the agent can start several overlapping query paths or threads of inquiry, typically mediated by the event manager 100. This technique, combined with the use of asynchronous queries from multiple data sources, provides the real-time response performance required for a natural interaction with the user.

[107] The domain agent 156 may apply conditional scraping operations to each query response as it is received. The conditional scraping actions depend on the context, the criteria, user profile 110, and domain agent coding and data. For each token to be extracted, a scraping criteria 152 may be created using the services of the system agent 150. The scraping criteria use format specific scraping methods including, tables, lists, text, and other scrapping methods. One or more scraping criteria can be applied to a page or results set. Once additional results are received, the domain agent 156 can create new scraping criteria to apply to results already acquired. The conditional scarping process removes extraneous information, such as graphics, which need not be further processed or stored, improving system performance.

[108] The domain agent 156 may strongly influence how the results are presented. For instance, once the domain agent 156 has created a satisfactory response to a question, the agent 156 may format that response for presentation. Typically, the domain agent 156 formats the response into the markup format used by the text to speech engine 124. The

domain agent 156 may also format the result presentation using available format templates and based on the context, the criteria, and the user profile 110. Variable substitutions and transformations may be performed by the agent 156 to produce a response best understood and most natural to the user. Further, the order of presentation of tokens and the exact terminology used to create a more natural response to the user may be varied by the agent 156. The domain agent 156 may also select the presentation personality 108 to be used.

[109] The domain agent 156 may select the presentation template, determine order of presentation for tokens and determine variable substitutions and transformations using probabilistic or fuzzy set decision methods. The template used to form the presentation can be from the domain agent itself or from the user profile 110. The user profile 110 can completely specify the presentation format or can be used to select and then modify an existing presentation format. Selection and formatting of presentation templates can also depend on the presentation personality modules 108. At the same time, the characteristics of the personality used for the response are dynamically determined using probabilities or fuzzy possibilities derived from the context, the criteria, the domain agent itself and the user profile 110.

[110] The domain agent 156 may apply a number of transformations to the tokens before presentation to the user. These variable substitutions and transformations are derived from a number of sources including, domain information carried by the agent, the context, the token values, the criteria, the personality module 108 to be used, and the user profile 110. Examples of variable substitutions and transformations include:

- substitution of words for numbers;
- substitution of names for acronyms or symbols (i.e., trading symbols);
- use of formatting information derived from the information sources (i.e., HTML tags);
- nature of the response including, text, long text, list, table;
- possible missing information or errors;
- units for measurement (i.e., English or metric); and
- preferred terminology from the user profile 110 or presentation personality 108.

[111] The system 90, according to another embodiment of the invention, may provide special purpose presentation capabilities for long text strings, tables, lists and other large results sets. Domain agents 156 may use special formatting templates for such results. The

system agent 150 may provide special criteria handlers 152 for presentation and user commands for large results sets. The presentation templates used by the domain agents 156 for large results sets typically include methods for summarizing the results and then allowing the user to query the result in more detail. For example, initially only short summaries such as headlines or key numbers are presented. The user can then query the results set further. The criteria handlers 152 provide users with the capability to browse large results sets. Commands provided by the criteria handlers 152 for large results sets include, stop, pause, skip, rewind, start, and forward.

[112] Some information, in formats such as video, pictures and graphics, may be best presented in a displayed format. The domain agents 156 may apply suitable presentation templates in these cases and present the information through the graphical user interface 114. The system agent 150 provides special criteria handlers 152 for presentation and user commands for display presentation and control.

[113] FIG. 3 illustrates a high level process 300 for receiving natural language speech-based queries and/or commands and generating a response according to an embodiment of the invention. The process 300 comprises of individual steps, several of which may be repetitive in order to accommodate partial failures. The process 300 may begin when a user's utterance is received and processed at 302. The user utterance may be a query and/or a command and may be of a natural language speech-based form. After receiving the natural language query and/or command, the query and/or command may be processed so that the data is in a more machine understandable format such as in a digital format. Once the utterance has been converted into an understandable format, the utterance may be processed so that at least an approximate meaning of the utterance can be determined at 304. This may be accomplished by, for example, the use of stored dictionaries, phrases, user profiles, domain agent data, and other data. Based on the approximate meaning of the utterance, one or more correctly formatted queries and/or commands may be generated.

[114] A user created query and/or command may require that multiple requests and/or commands may be needed in order to generate the desired response or action. Further, depending upon the context in which the query and/or command has been submitted (e.g., who is the user, what is the subject the query and/or command directed to, at what time or location is the query and/or command being directed to, and other parameters used in defining the query and/or command), a specific format for the query and/or command may be required. For instance, each of the domain agents 156 may require that requests and/or commands be formatted in a specific manner. Thus at 306, one or more properly formatted

queries and/or commands may be generated. Operations 304 and 306 may be repeated in order to obtain the correct interpretation of the user utterance and result in the desired response and/or action. Once, the properly formatted queries and/or commands are generated, the appropriate action may be taken at 308. This may require that the properly formatted queries and/or commands be sent to specific domain agent[s], information source[s], device[s], or other appropriate destination[s] that can fulfill the requirements of the query and/or command. Once the appropriate action[s] have been executed, this event may be recorded to, for example, the user's profile, database and/or one or more agents at 310. Such data may be useful for future user inquires and commands. After the action[s] have been performed, a response, if need be, may be generated and forwarded to the user and/or third parties at 312.

[115] In the case of a query for retrieving data, the response would contain the requested information. In the case of a command, the response may be a confirmation that a specific action[s] has been executed. The response may be in the form of a natural language format. The response may also be formatted to reflect a particular personality or tone to the response in order to make the response more "human." The response may relayed to the user and/or third parties as an audio message and/or a visual message displayed on a user interface.

[116] FIG. 4A is a process 400 for receiving natural language speech-based queries and/or commands and generating a response using the system 90 according to another embodiment of the invention. The process 400 may begin when a user's utterance (i.e., user query and/or command) is captured through speech recognition operating in a variety of real-world environments at 402. Once the utterance is captured, it is parsed and interpreted to determine the query and/or command that is contained in the utterance at 404. Next, the query and/or command are reviewed to determine the domain of expertise required and the context of the query, invoking the proper resources including, for example, agents at 406. The original query and/or command submitted by the user may require that multiple queries and/or commands be generated.

[117] For instance, suppose a user is interested in retrieving the value of her stock portfolio. The user may utter "please get the value of my stock portfolio." The system 90 may review this request together with stored data such as the user's profile and determine keywords such as "get the value" and "my stock portfolio." The system 90 may then generate queries to determine the stocks in the user's portfolio, the number of shares and the source for current pricing information. These queries may then be sent to one or more agent domains, such as a

domain which may access a database containing the user's profile and a domain which accesses stock pricing sources to determine the answers to these questions.

[118] Queries may be sent to these domain agents or sources in order to obtain the desired data. Thus, at operation 408, one or more queries may be formulated and sent to one or more local and/or network data sources and/or appropriate commands may be sent to local or remote devices or the system itself. The queries are then sent to the designated agent[s]. The agents may then, in turn, generate their own queries and/or commands to be sent to, for example, local or remote information sources to retrieve needed data. The agent generated queries and/or commands may be formatted according to the requirements of the target sources and variable substitutions and transformations are performed to modify the queries to a form most likely to yield desired results from the available sources at 410. Once the queries are formatted correctly, they may be executed in an asynchronous manner and dealing gracefully with failures at 412. As a result of the execution of the queries, results may be returned by the domain agents and/or sources. The system 90 may then extract or scrape the desired information from the one or more results, which may be returned in any one of a number of different formats at 414. That is, the results sought by a user may be the summary or the results of further processing of information obtained from several sources for example.

[119] Next, the results may be evaluated and interpreted including processing of errors, and gathering and combining them into a single best result judged to be "best" even if the results are ambiguous, incomplete, or conflicting at 416. Once the best results are determined, any required formatting is performed. At operation 418, variable substitutions and transformations may be used to modify the results. Finally, at operation 420, the compound results may be presented to the user in a useful and expected manner through the text to speech engine 124. The process 400 may be performed while accounting for the domain of expertise required, the context in which the question or command is presented, the domain specific information available, the history of the user's interaction, the user preferences, the information sources or commands that are available, and responses obtained from the sources.

[120] At each stage of the process 400, probabilistic or fuzzy set decision and matching methods may be applied to deal with inconsistent, ambiguous, conflicting and incomplete information or responses. In addition, the use of asynchronous queries that may result in rapid and graceful failure of some queries or commands may allow the system 90 to robustly return results quickly and in a manner that seems natural to the user.

[121] FIG. 4B illustrates a process 450 for receiving natural language speech-based commands in order to locally or remotely control functions of the system 90, or for other devices, according to another embodiment of the invention. The process 450 may begin when a user's utterance (i.e., user query and/or command) is captured through speech recognition operating in a variety of real-world environments at 452. Once the utterance is captured, it is parsed and interpreted to determine the command that is contained in the utterance at operation 454.

[122] Next, at operation 456, the command is reviewed to determine the domain for the command and context, invoking the proper resources including agents. The original command submitted by the user will often require that multiple queries and/or commands be generated. For instance, suppose a user is interested in recording his favorite television program. The user may utter "please record my favorite TV program." The system 90 may review this request together with stored data such as the user's profile and determine keywords such as "record" and "my favorite TV program." The system 90 may then generate queries to determine the name, the channel, and time for the user's favorite TV program. These queries may then be sent to an agent domain, which may access a database containing the user's profile and determine the answers to these questions. A command may then be sent to a video recorder ordering the video recorder to record the selected TV program. Thus, at operation 458, one or more queries may be formulated that are to be sent to one or more local and/or network data sources. Appropriate commands may be sent to local or remote devices or the system itself. Next, at operation 460, the generated commands may be routed to the appropriate system[s] and/or external devices. Once the commands have been executed, results including errors may be received and processed at 462. The results of the executed commands may be optionally presented to the user at operation 464.

[123] Some of the operations depicting in FIG. 4A may be performed using multiple steps that may be repetitive. For instance, in operation 404, the user's utterance is parsed and interpreted in order to determine the meaning of the utterance. System 90 may make an initial interpretation of the utterance based the data stored in, for example, the context description grammar module 112, user profiles, 110, agents 106 and the databases 104. A real-time scoring system or other techniques may be applied to the interpretation results in order to generate a domain or a context score. If the confidence level of the domain or context score is not high enough to ensure a reliable response, the system 90 can request that the user verify the question or command is correctly understood. In general, the question may be phrased to indicate the context of the question including all criteria or parameters. If

the user confirms that the question is correct, the system 90 may proceed to produce a response. Otherwise, either the user can rephrase the original question, perhaps adding additional information to remove ambiguity, or the system may ask one or more questions to attempt to resolve the ambiguity or other actions may taken.

[124] FIG. 5 illustrates a process 500 for correctly interpreting a user's utterance according to one embodiment of the invention. The process 500 generally represents operations 402 and 404 of process 400. Initially the user's utterance is inputted into the system 90 at operation 502. The input is then interpreted using, for example, various stored data such as user profiles, agent data, dictionary and phrases, and other relevant data. The interpretation is then scored using, for instance, a weighted scoring system as described previously at 506. Once a score is obtained, the confidence level of the interpretation may be analyzed to determine if it is satisfactory at operation 508. In other words, a determination is made whether the score given to the interpretation exceeds a certain value. If the confidence level is determined to be unsatisfactory, then a request may be submitted to the user requesting that the user verify the interpretation at 510 and 512. If the user is unsatisfied with the interpretation, he/she may be asked to rephrase the utterance and/or provide additional information at 514. Once the user provides the rephrased utterance and/or additional information, the process 500 returns to the beginning at operation 502. If, on the other hand, the user is satisfied with the interpretation, then the process 400 for receiving natural language speech-based queries and/or commands, and generating a response, may continue at operation 516. Once the meaning of the utterance has been determined, the appropriate domain agent[s] and the query and/or commands may be properly formatted for the agent[s].

[125] FIG. 6 illustrates a process 600 for determining the proper domain agent[s] to invoke and the proper format for queries and/or commands that are submitted to the agents 106, as generally depicted in steps 406-408 of FIG. 4A. In order to formulate a question or command in the regular grammar used by agents 106, a determination may be made as to the required and optional values for the criteria or parameters of the queries and/or commands. These criteria may have been explicitly supplied by the user or may need to be inferred. Thus, the properly interpreted query and/or command that is generated by, for example, operations 404 and 454 of FIGS. 4A and 4B, is parsed at operation 602.

[126] The content of the properly interpreted query and/or command is then analyzed to determine the criteria and/or parameters of the properly interpreted query and/or command at 604. A context sensitive procedure for extracting the criteria or parameters from the properly interpreted query and/or command may be used. Some criteria are determined by executing

algorithms in the agent, while others may be determined by applying probabilistic of fuzzy reasoning to tables of possible values. Prior probabilities or fuzzy possibilities and associated values are received from a number of sources including the history of the dialog, the user profile 110, and the agent.

[127] Based on user responses, the prior probabilities or fuzzy possibilities are updated as the system 90 learns the desired behavior. For a weather context, examples of criteria include, location, date and time. Other criteria can include command criteria (i.e., yes/no, on/off, pause, stop), and spelling. The process for determining criteria may be iterative or recursive in order to eliminate ambiguity in the user's question or command. For example, if the user has a place name (or other proper noun) in their utterance, tables in the databases 102 may be reviewed for place names or an attempt may be made to determine which word is the proper noun from the syntax of the utterance. In another example, the user asks "what about flight one hundred and twenty too?" In such a situation, flight information in the database and network information along with context may be used to determine the most plausible interpretation among; flight 100 and flight 20 also, flight 100 and flight 22, flight 122, and the like. Once the parameters and criteria for the query and/or command have been established, the proper agents 106 may be selected at 606.

[128] After selecting the agents 106, at operation 608, a properly formatted query and/or command that will be submitted to the agents 106 may be determined. The query and/or command may be in a standard format or may be a hierarchical data structure used for processing by the agent 106. In order to submit the properly formatted query and/or command to the agent 106, all of the required, and some optional tokens, for the grammar of the context may be filled in. Often the tokens must be transformed to values and forms acceptable to the agents. The required transformations may be obtained from the agents, dialog history or user profile 110. Examples of transformations or substitutions that may be performed were provided above. Once the properly formatted queries and/or commands are generated, the process (e.g., process 400) may continue at 610.

[129] While specific embodiments of the invention have been described above, it will be appreciated that the invention may be practiced otherwise than as described. For example, the invention may take the form of a computer program containing one or more sequences of machine-readable instructions describing a method as disclosed above, or a data storage medium (e.g. semiconductor memory, magnetic or optical disk) having such a computer program stored therein.

[130] The descriptions above are intended to be illustrative, not limiting. Thus, it will be apparent to one skilled in the art that modifications may be made to the invention as described without departing from the scope of the claims set out below.

We claim:

1. A system for processing speech and non-speech communications, comprising:

a terminal device that receives the speech and the non-speech communications;

a transcription module that transcribes the speech and the non-speech communications to create a speech-based textual message and a non-speech-based textual message;

a merging module that merges the speech-based textual message and the non-speech-based textual message to generate a query;

a search module that searches the query for text combinations;

a comparison module that compares the text combinations to entries in a context description grammar;

a plurality of domain agents that are associated with the context description grammar;

a scoring module that provides relevance scores based on results from the comparison module;

a domain agent selector that selects domain agents based on results from the scoring module; and

a response generating module that communicates with the selected domain agents to obtain content that is gathered by the selected domain agents and that generates a response from the content, wherein the content is arranged in a selected order based on results from the scoring module.

2. The system according to claim 1, wherein the response generating module generates an aggregate response that includes the content gathered by the selected domain agents.

3. The system according to claim 1, wherein the terminal device includes (i) a personal digital assistant, (ii) a cellular telephone, (iii) a portable computer, (iv) a desktop computer, or any combination of (i) to (iv).

4. The system according to claim 1, wherein the terminal device receives follow-up speech and non-speech communications and wherein the transcription module transcribes the follow-up speech and non-speech communications to create a follow-up speech-based textual message and a follow-up non-speech-based textual message.

5. The system according to claim 4, wherein the merging module merges the follow-up speech-based textual message and the follow-up non-speech-based textual message to generate a follow-up query.

6. The system according to claim 1, further comprising a personality module that facilitates formatting the response.

7. The system according to claim 1, further comprising a context stack that includes one or more contexts that are selected based on the query.

8. The system according to claim 7, wherein the scoring module determines the one or more contexts based on at least applying prior probabilities or fuzzy possibilities to (i) keyword matching, (ii) user profiles, (iii) a dialog history, or any combination of (i) to (iii).

9. The system according to claim 1, wherein at least one of the domain agents creates and directs a request to at least one of a local information source and a network information source.

10. The system according to claim 1, wherein at least one of the domain agents creates and directs a command to a remote or local device.

11. The system according to claim 5, wherein at least one of the domain agents evaluates multiple queries from multiple sources.

12. The system according to claim 5, wherein the follow-up query is associated with a same context as the query.

13. A method of processing speech and non-speech communications, comprising:

receiving the speech and non-speech communications;

transcribing the speech and non-speech communications to create a speech-based textual message and a non-speech-based textual message;

merging the speech-based textual message and the non-speech-based textual message to generate a query;

searching the query for text combinations;

comparing the text combinations to entries in a context description grammar;

accessing a plurality of domain agents that are associated with the context description grammar;

generating a relevance score based on results from comparing the text combinations to entries in the context description grammar;

selecting one or more domain agents based on results from the relevance score;

obtaining content that is gathered by the selected domain agents; and

generating a response from the content, wherein the content is arranged in a selected order based on results from the relevance score.

14. The method according to claim 13, further comprising generating an aggregate response that includes the content that is gathered by the selected domain agents.

15. The method according to claim 13, further comprising:

receiving a follow-up speech and non-speech communications;

transcribing the follow-up speech and non-speech communications to create a follow-up speech-based textual message and a follow-up non-speech-based textual message; and

merging the follow-up speech-based textual message and the follow-up non-speech-based textual message to generate a follow-up query.

16. The method according to claim 13, further comprising a personality module that communicates the response to a user.

17. The method according to claim 13, further comprising generating a context stack that includes one or more contexts that are selected based on the query.

18. The method according to claim 17, wherein the one or more contexts are generated based on applying prior probabilities or fuzzy possibilities to (i) keyword matching, (ii) user profiles, (iii) a dialog history, or any combination of (i) to (iii).

19. A multimodal system for processing speech and non-speech communications, comprising:

a terminal device that receives one or more types of input;

a transcription module that transcribes the one or more types of input into one or more textual messages;

a merging module that merges the one or more textual messages to generate a query;

a search module that searches the query for text combinations;

a comparison module that compares the text combinations to entries in a context description grammar;

a plurality of domain agents that are associated with the context description grammar;

a scoring module that provides relevance scores based on results from the comparison module;

a domain agent selector that selects domain agents based on results from the scoring module; and

a response generating module that communicates with the selected domain agents to obtain content that is gathered by the selected domain agents and that generates a response from the content, wherein the content is arranged in a selected order based on results from the scoring module;

wherein the terminal device delivers the response using one or more types of output.

20. The multimodal system according to claim 19, wherein the one or more types of input includes (i) speech, (ii) text, (iii) digital audio files, or any combination of (i) to (iii).

21. The multimodal system according to claim 19, wherein the one or more types of output includes (i) speech, (ii) text, (iii) digital audio files, or any combination of (i) to (iii).

22. The multimodal system according to claim 19, wherein the response generating module generates an aggregate response that includes the content gathered by the selected domain agents and wherein the terminal device delivers the aggregate response using the one or more types of output.

23. The multimodal system according to claim 19, wherein the terminal device includes (i) a personal digital assistant, (ii) a cellular telephone, (iii) a portable computer, (iv) a desktop computer, or any combination of (i) to (iv).

24. The multimodal system according to claim 19, wherein the terminal device receives one or more types of follow-up input, wherein the transcription module transcribes the one or more types of follow-up input to create one or more follow-up textual messages, and wherein the one or more types of follow-up input are different from the one or more types of input.

25. The multimodal system according to claim 24, wherein the terminal device that receives the one or more types of follow-up input is different from the terminal device that receives the one or more types of input.

26. The multimodal system according to claim 24, wherein the merging module merges the one or more follow-up textual messages to generate a follow-up query.

27. The multimodal system according to claim 19, further comprising a personality module that facilitates formatting the response.

28. The multimodal system according to claim 19, further comprising a context stack that includes one or more contexts that are selected based on the query.

29. The multimodal system according to claim 28, wherein the scoring module determines the one or more contexts based on at least applying prior probabilities or fuzzy possibilities to (i) keyword matching, (ii) user profiles, (iii) a dialog history, or any combination of (i) to (iii).

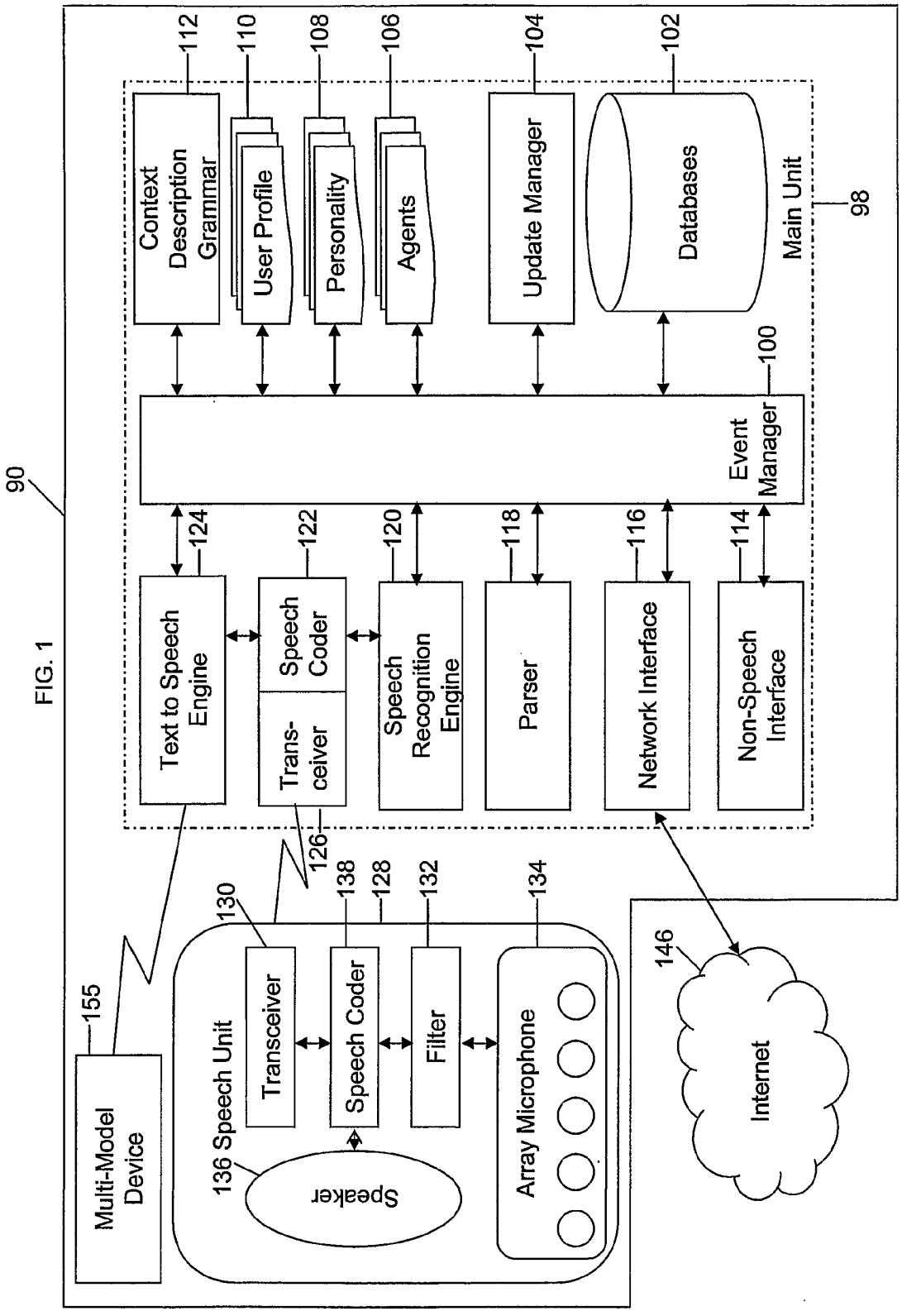

90
Main Unit
98
Context Description Grammar
112
User Profile
110
Personality
108
Agents
106
Update Manager
104
Databases
102
Event Manager
100
Text to Speech Engine
124
Speech Coder
Trans-ceiver
122
126
Speech Recognition Engine
120
Parser
118
Network Interface
116
Non-Speech Interface
114
Multi-Model Device
155
136 Speech Unit
130
Transceiver
138
Speech Coder
128
Filter
132
Array Microphone
134
Speaker
Internet
146

FIG. 1

FIG. 2

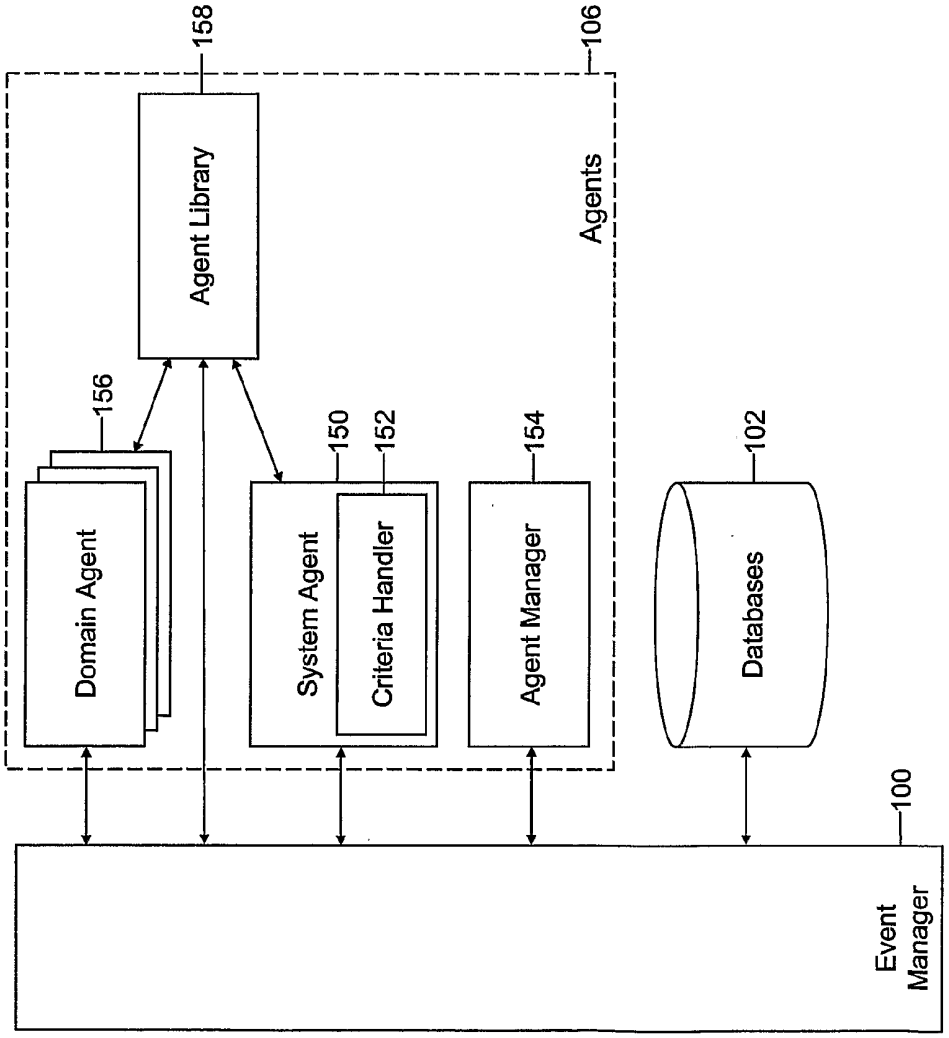

158
106
Agent Library
Agents
156
150
152
154
102
Domain Agent
System Agent
Criteria Handler
Agent Manager
Databases
100
Event Manager

FIG. 3

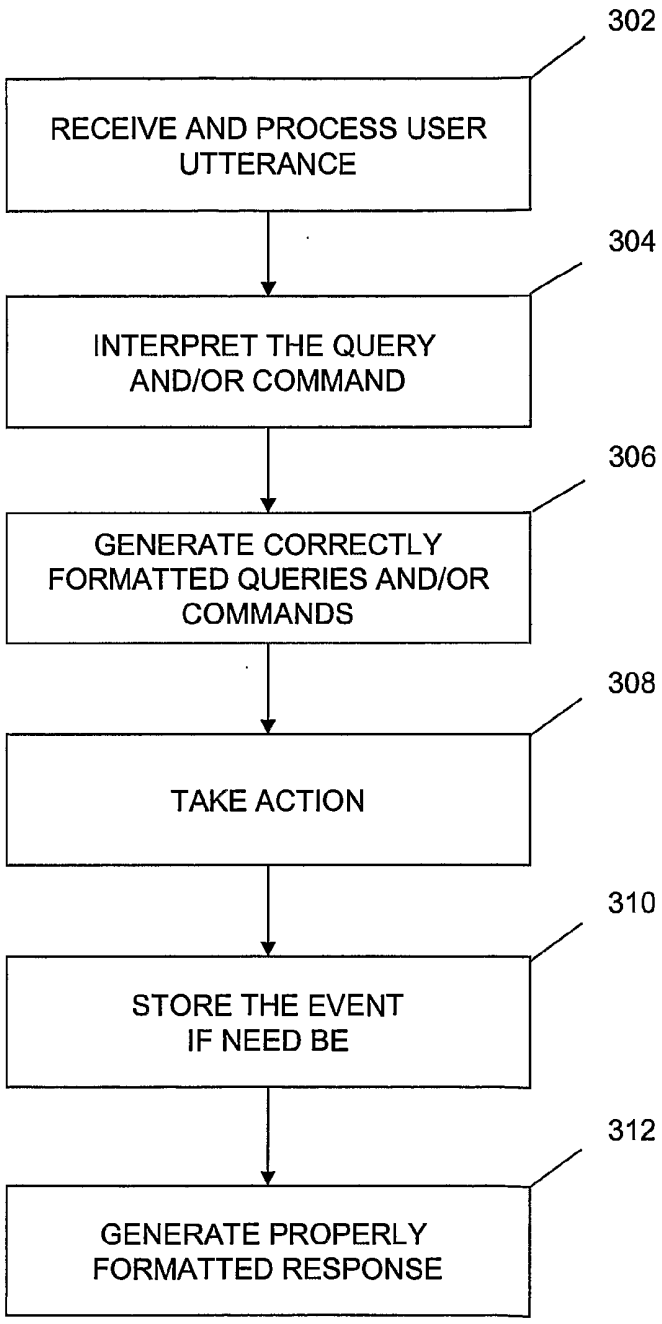
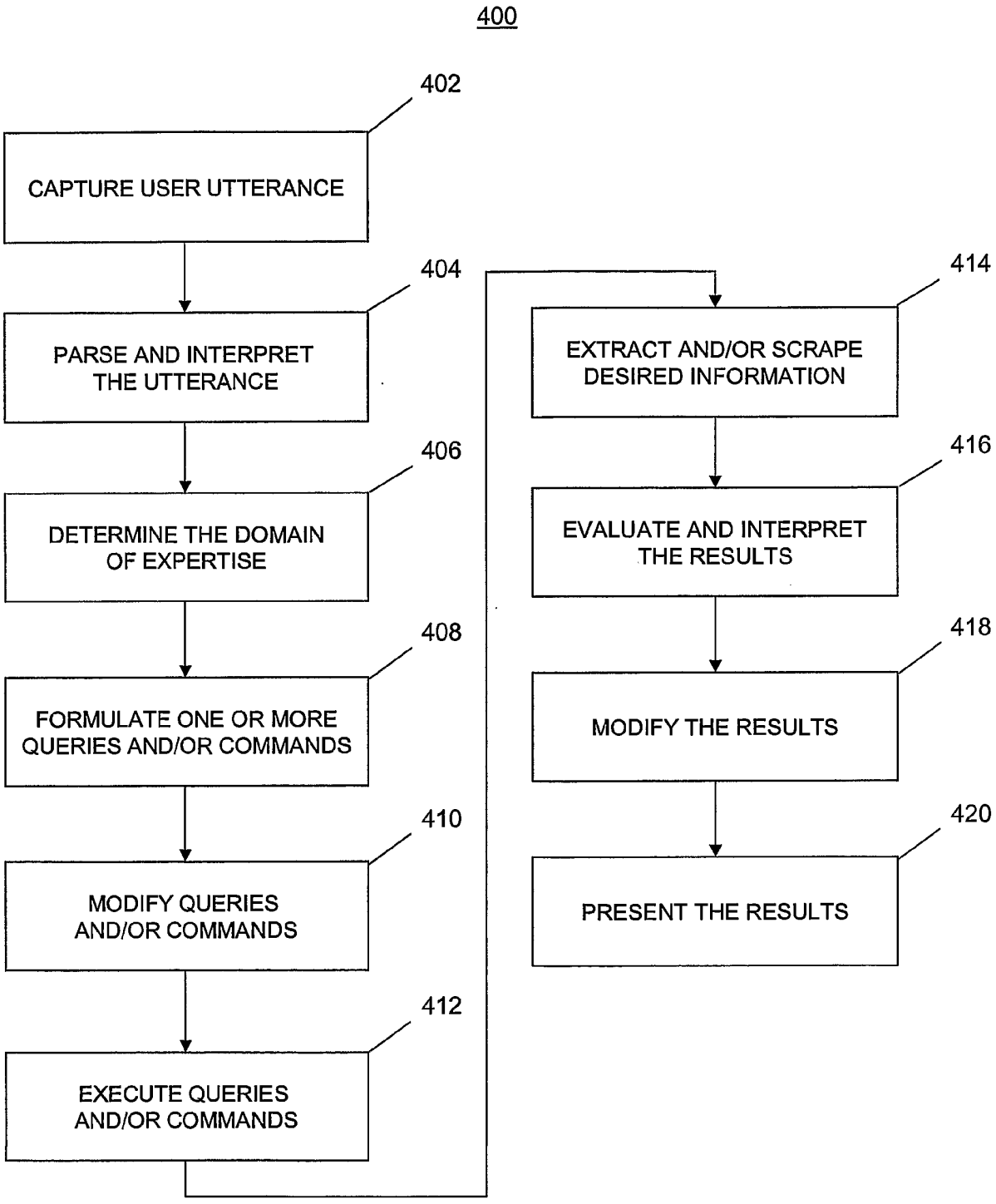
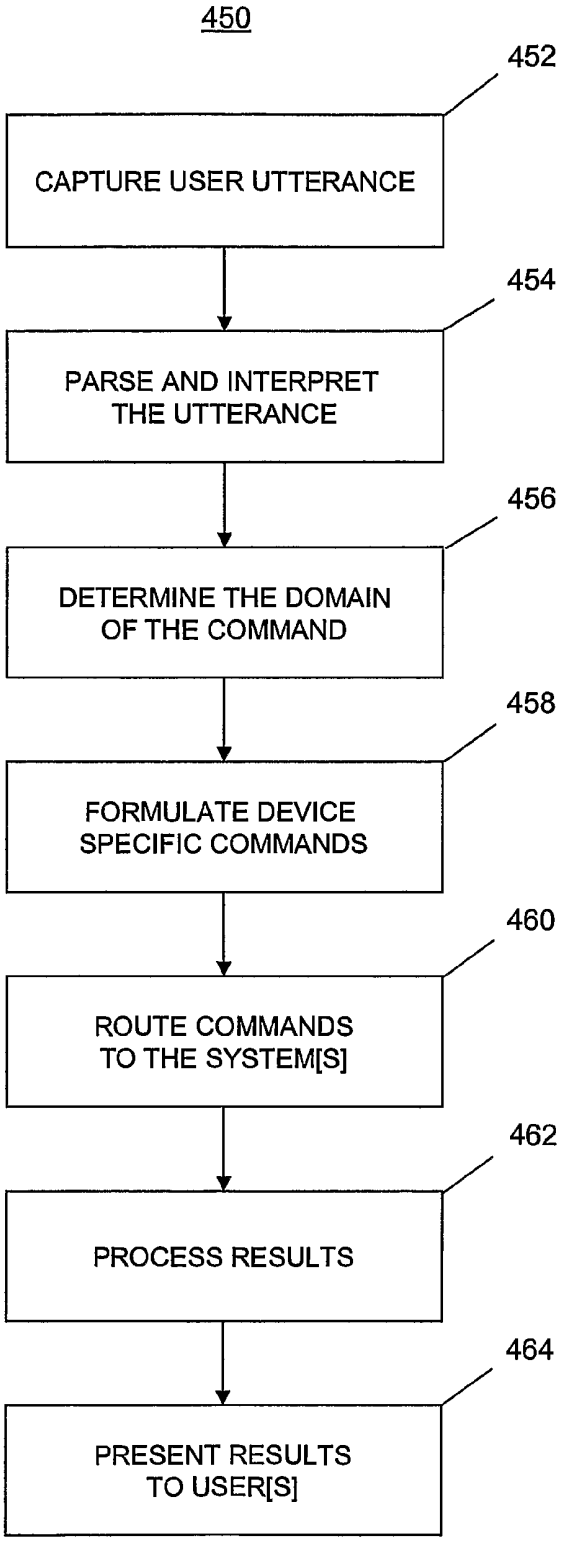

300
302
RECEIVE AND PROCESS USER UTTERANCE
304
INTERPRET THE QUERY AND/OR COMMAND
306
GENERATE CORRECTLY FORMATTED QUERIES AND/OR COMMANDS
308
TAKE ACTION
310
STORE THE EVENT IF NEED BE
312
GENERATE PROPERLY FORMATTED RESPONSE

FIG. 4A


400
402
CAPTURE USER UTTERANCE
404
PARSE AND INTERPRET THE UTTERANCE
406
DETERMINE THE DOMAIN OF EXPERTISE
408
FORMULATE ONE OR MORE QUERIES AND/OR COMMANDS
410
MODIFY QUERIES AND/OR COMMANDS
412
EXECUTE QUERIES AND/OR COMMANDS
414
EXTRACT AND/OR SCRAPE DESIRED INFORMATION
416
EVALUATE AND INTERPRET THE RESULTS
418
MODIFY THE RESULTS
420
PRESENT THE RESULTS

5/8

FIG. 4B


450
452
CAPTURE USER UTTERANCE
454
PARSE AND INTERPRET THE UTTERANCE
456
DETERMINE THE DOMAIN OF THE COMMAND
458
FORMULATE DEVICE SPECIFIC COMMANDS
460
ROUTE COMMANDS TO THE SYSTEM[S]
462
PROCESS RESULTS
464
PRESENT RESULTS TO USER[S]

6/8

FIG. 5

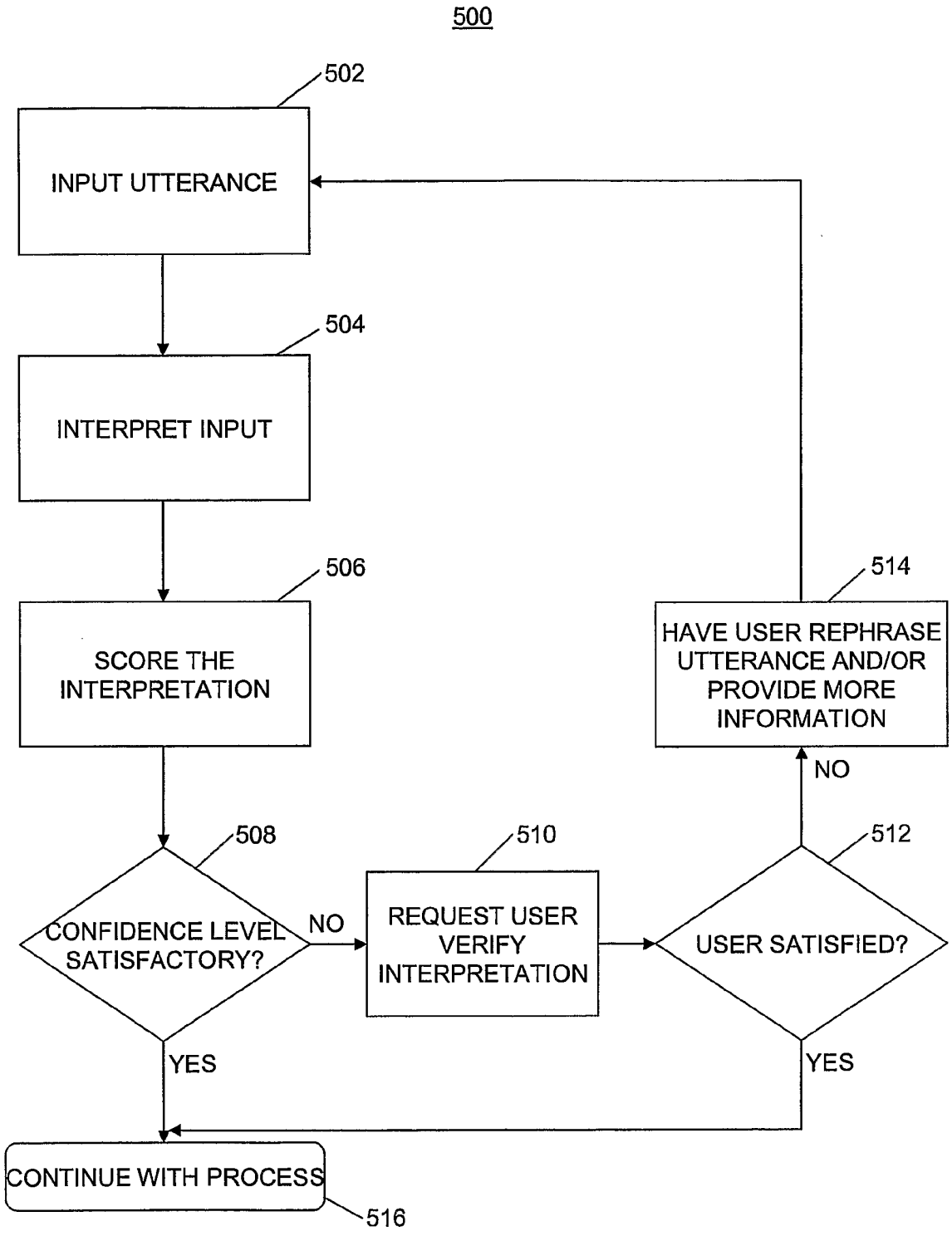

500
502
INPUT UTTERANCE
504
INTERPRET INPUT
506
SCORE THE INTERPRETATION
508
CONFIDENCE LEVEL SATISFACTORY?
NO
YES
510
REQUEST USER VERIFY INTERPRETATION
512
USER SATISFIED?
NO
YES
514
HAVE USER REPHRASE UTTERANCE AND/OR PROVIDE MORE INFORMATION
CONTINUE WITH PROCESS
516

FIG. 6

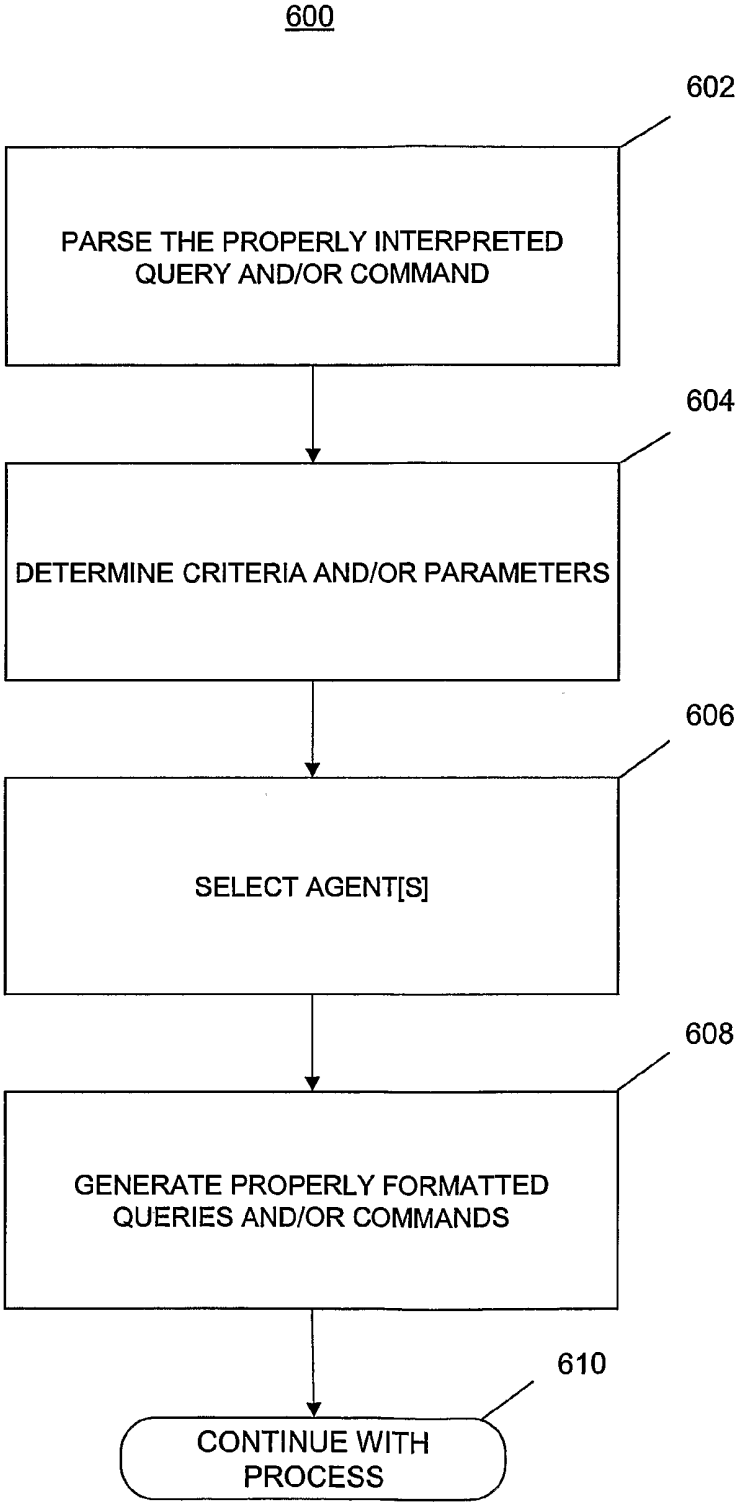
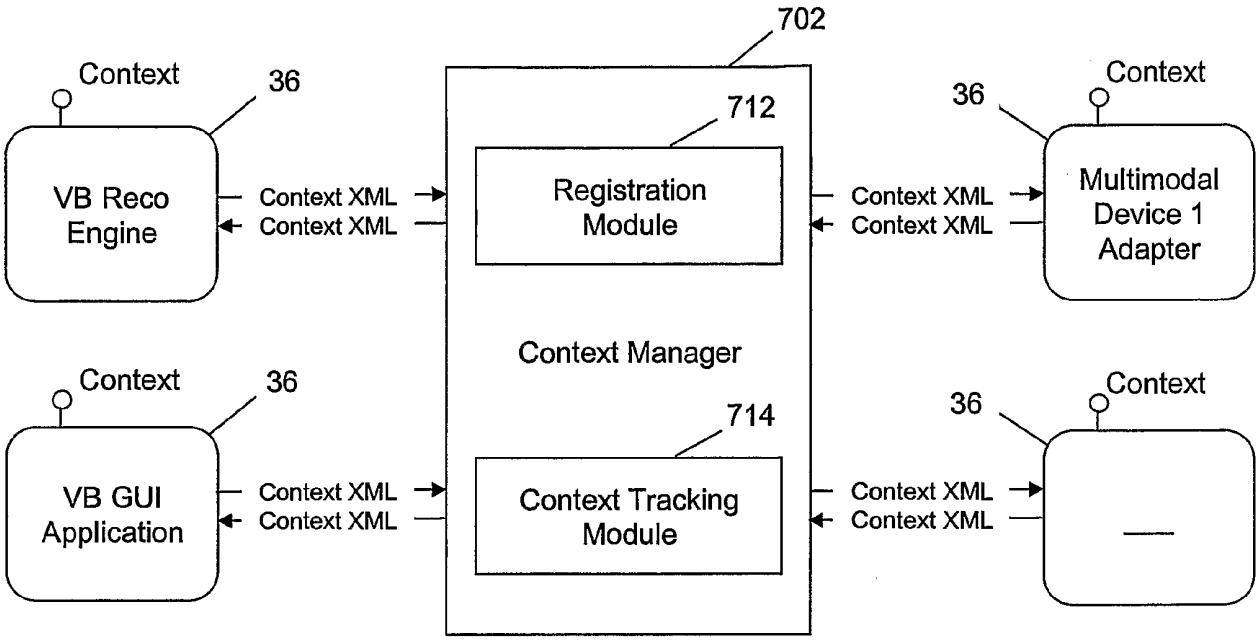

600
602
PARSE THE PROPERLY INTERPRETED QUERY AND/OR COMMAND
604
DETERMINE CRITERIA AND/OR PARAMETERS
606
SELECT AGENT[S]
608
GENERATE PROPERLY FORMATTED QUERIES AND/OR COMMANDS
610
CONTINUE WITH PROCESS

FIG. 7


700
702
Context
36
VB Reco Engine
Context XML
Context XML
712
Registration Module
Context XML
Context XML
36
Context
Multimodal Device 1 Adapter
Context Manager
Context
36
VB GUI Application
Context XML
Context XML
714
Context Tracking Module
Context XML
Context XML
36
Context

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 February 2007 (15.02.2007)

PCT

(10) International Publication Number
WO 2007/019318 A3

(51) International Patent Classification:
G10L 21/00 (2006.01) *G06F 17/30* (2006.01)
G06F 3/16 (2006.01)

(21) International Application Number:
PCT/US2006/030458

(22) International Filing Date: 4 August 2006 (04.08.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/197,504 5 August 2005 (05.08.2005) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [—/US]; 10505 NE 38th Place, Building 9, Kirkland, WA 98033 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **DI CRISTO, Philippe** [FR/US]; 177 107th Avenue, N.E., #1007, Bellevue, WA 98004 (US). **KE, Min** [US/US]; 3221 201st Place, S.E., Bothell, WA 98012 (US). **KENNEWICK, Robert, A.** [US/US]; 2204 S. Eddy Street, Seattle, WA 98108 (US).

(74) Agents: **GATTO, James, G.** et al.; PILLSBURY WINTHROP SHAW PITTMAN LLP, P.O. Box 10500, Mclean, VA 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

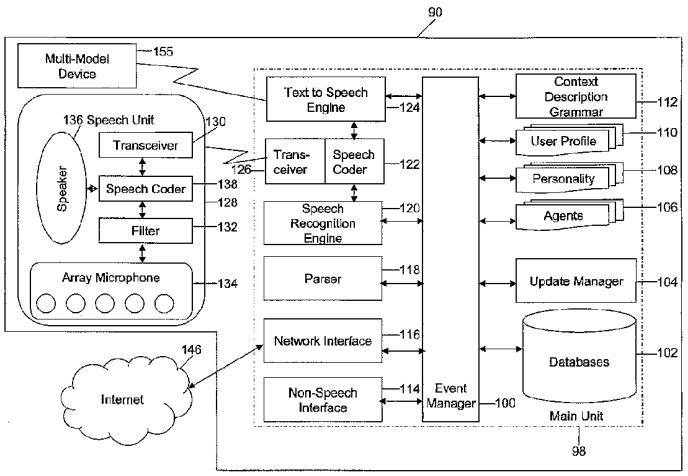
(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

(88) Date of publication of the international search report:
29 November 2007

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE



(57) Abstract: Systems and methods are provided for receiving speech and non-speech communications of natural language questions and/or commands, transcribing the speech and non-speech communications to textual messages, and executing the questions and/or commands. The invention applies context, prior information, domain knowledge, and user specific profile data to achieve a natural environment for one or more users presenting questions or commands across multiple domains. The systems and methods creates, stores and uses extensive personal profile information for each user, thereby improving the reliability of determining the context of the speech and non-speech communications and presenting the expected results for a particular question or command.

WO 2007/019318 A3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US06/30458

A. CLASSIFICATION OF SUBJECT MATTER
IPC: **G10L 21/00**(2006.01);**G06F 3/16**(2006.01);**G06F 17/30**(2006.01)

USPC: 704/270.1;715/727;707/3
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
U.S. : 704/270.1;715/727;707/3

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US 7,137,126 B1 (COFFMAN et al) 14 November 2006 (14.11.2006), column 28 to column 32.	1-8 and 12-29 --- 9-11
Y	US 2004/0044516 A1 (KENNEWICK et al) 4 March 2004 (04.03.2004), paragraphs 161 to 166.	

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
17 August 2007 (17.08.2007)

Date of mailing of the international search report
25 SEP 2007

Name and mailing address of the ISA/US
Mail Stop PCT, Attn: ISA/US
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450
Facsimile No. (571) 273-3201

Authorized officer
David Hudspeth
Telephone No. (571) 272-2600

Form PCT/ISA/210 (second sheet) (April 2005)

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 February 2007 (22.02.2007)

PCT

(10) International Publication Number
WO 2007/021587 A2

(51) International Patent Classification:
G06F 17/27 (2006.01)

(21) International Application Number:
PCT/US2006/030459

(22) International Filing Date: 4 August 2006 (04.08.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/200,164 10 August 2005 (10.08.2005) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [—/US]; 10505 NE 38TH PLACE, Building 9, Kirkland, WA 98033 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **DI CRISTO, Philippe** [FR/US]; 177 107th Avenue, N.E., #1007, Bellevue, WA 98004 (US). **WEIDER, Chris** [US/US]; 2818 Grand Avenue, #B-508, Everett, WA 98201 (US). **KENNEWICK, Robert, A.** [US/US]; 2204 S. Eddy Street, Seattle, WA 98108 (US).

(74) Agents: **GATTO, James, G.** et al.; PILLSBURY WINTHROP SHAW PITTMAN LLP, P.O. Box 10500, McLean, VA 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 2007/021587 A2

(54) Title: SYSTEMS AND METHODS OF SUPPORTING ADAPTIVE MISRECOGNITION IN CONVERSATIONAL SPEECH

(57) Abstract: A system and method are provided for receiving speech and/or non-speech communications of natural language questions and/or commands and executing the questions and/or commands. The invention provides a conversational human-machine interface that includes a conversational speech analyzer, a general cognitive model, an environmental model, and a personalized cognitive model to determine context, domain knowledge, and invoke prior information to interpret a spoken utterance or a received non-spoken message. The system and method creates, stores and uses extensive personal profile information for each user, thereby improving the reliability of determining the context of the speech or non-speech communication and presenting the expected results for a particular question or command.

SYSTEM AND METHOD OF SUPPORTING ADAPTIVE MISRECOGNITION IN CONVERSATIONAL SPEECH

BACKGROUND OF THE INVENTION

Cross-Reference to Related Applications

[01] This application claims the benefit of U.S. Patent Application Serial No. 11/200,164 filed August 10, 2005, which is herein incorporated by reference in its entirety.

Field of the Invention

[02] The invention relates to providing conversational human-machine interfaces for controlling devices through a speech interface and/or a combination of a speech interface and a non-speech interface. More specifically, the invention allows users to submit natural language commands via the speech interface and/or the non-speech interface. The invention further allows users to control devices and systems either locally or remotely.

Background of the Related Art

[03] A machine's ability to communicate with humans in a natural manner remains a difficult problem. Cognitive research on human interaction shows that verbal communication, such as a person asking a question or giving a command, typically relies heavily on context and domain knowledge of the target person. By contrast, machine-based queries (a query may be a question, a command, a request and/or other types of communications) may be highly structured and may not be inherently natural to the human user. Thus, verbal communications and machine processing of requests that are extracted from the verbal communications may be fundamentally incompatible. Yet the ability to allow a person to make natural language speech-based request remains a desirable goal.

[04] Speech recognition has steadily improved in accuracy and today is successfully used in a wide range of applications. Natural language processing has been applied to the parsing of speech queries. Yet, current systems do not reliably provide a complete environment for users to submit speech and/or non-speech communications through natural language queries that are processed to provide natural responses. There remains a number of significant barriers to creation of a complete speech-based and/or non-speech-based natural language query and response environment.

SUMMARY OF THE INVENTION

[05] According to an aspect of the invention, one object of the invention is to overcome these and other drawbacks of prior systems.

[06] According to one aspect of the invention, users may provide the system with spoken utterances and/or other communications in a natural language format, including imperfect

information such as, incomplete thoughts, incomplete sentences, incomplete phrases, slang terminology, repeated words, word variations, synonyms, or other imperfect information. In order for machines (e.g., computer devices) to properly respond to questions and/or commands that are presented in the natural language format, the questions and/or commands may be parsed and interpreted to formulate machine processable queries and algorithms.

[07] In another aspect of the invention, systems are provided for overcoming the deficiencies of prior systems through the application of a complete speech-based command generation, information retrieval, presentation and command environment or a combination of speech-based and non-speech-based command generation, information retrieval, presentation and command environment. This environment makes maximum use of context, prior information, domain knowledge, and user specific profile data to achieve a natural environment for one or more users to submit queries or commands to multiple domains. Through this integrated approach, a complete speech-based natural language query and response environment may be created. Alternatively, a combination of speech-based and non-speech-based natural language query and response environment may be created. Further, at each step in the process, accommodation may be made for full or partial failure and graceful recovery. The robustness to partial failure is achieved through the use of probabilistic and fuzzy reasoning at several stages of the process. This robustness to partial failure promotes the feeling of a natural response to questions and commands.

[08] According to another aspect of the invention, the system may include a speech unit interface device that receives utterances, including spoken natural language queries, commands and/or other utterances from a user, and a computer device or system that receives input from the speech unit and processes the input (e.g., retrieves information responsive to the query, takes action consistent with the command and performs other functions as detailed herein). The system may further generate a natural language response and/or generate a natural language message, such as an alert message, without any prompting from a user. According to another embodiment of the invention, the non-speech interface may be provided in addition to, or in place of, the speech unit interface. For example, the non-speech interface may communicate non-speech information, such as textual communications, graphical or tabular information, or other non-speech information.

[09] According to one embodiment of the invention, multi-modal input may be received from users at a conversational speech analyzer that is coupled to several components, including a general cognitive model, an environmental model, a personalized cognitive model, and an adaptive misrecognition analysis engine, or other components. According to

one embodiment of the invention, the conversational speech analyzer may include one or more speech recognition engines that transcribe utterances to text.

[10] According to one embodiment of the invention, the environmental model may be accessed to determine user location, user activity, track user actions, and/or other environmental information to invoke context, domain knowledge, preferences, and/or other cognitive qualities to enhance the interpretation of questions and/or commands. The conversational speech analyzer may also access the general cognitive model and/or the personalized cognitive model to further refine context, domain knowledge, preferences, and/or other cognitive qualities to enhance the interpretation of questions and/or commands. Based on information received from the general cognitive model, environmental model and/or the personalized cognitive model, the system may enhance responses to commands and questions by including a prediction of user behavior.

[11] An adaptive misrecognition analysis engine may be provided to analyze and store speech and non-speech data, including transcribed utterances, that are identified as being unrecognized or incorrectly recognized by the conversational speech analyzer. Upon a determination that data is unrecognized, the system may generate an unrecognized event. According to one embodiment of the invention, the system may implement one or more techniques to determine that data is incorrectly recognized. When the system misrecognizes a request, the user typically provides immediate feedback, such as overriding the command in a time shorter than the expected time execution time of the command, issuing a stop command, or taking other action. The user's actions may be analyzed in a statistical model to determine a frequency occurrence of misrecognitions for particular commands, with the results being used to update the corresponding personalized cognitive model.

[12] According to another embodiment of the invention, the conversational speech analyzer may access the personalized cognitive model to proactively select a next best (or nth best) match for the received data. A match may be confirmed by user action that include the user not immediately canceling the command. The misrecognitions may also be analyzed to potentially determine personalized tuning parameters for the speech recognition components of the system. For example, the system may tune the speech recognition engine over time to increase recognition by analyzing how the speech recognition engine misrecognizes the utterance.

[13] According to one embodiment of the invention, infrastructure may be provided to maintain context information during multimodal interactions, such as speech and/or non-speech interactions. According to one exemplary embodiment of the invention, context

information may be maintained in a multimodal environment by providing communication channels between multimodal devices, or mobile devices, and the system. The communication channels allow the system to receive multimodal input such as text-based commands and questions and/or voice-based commands and questions. According to another embodiment of the invention, the multimodal input may include a string of text, such as keywords, that are received as commands or questions. According to yet another embodiment of the invention, the system may synchronize the context between the multimodal devices and the speech-based units. In order to send a response to the corresponding device, the system may track the source and send the response to the corresponding speech interface or the non-speech interface.

[14] According to an alternative embodiment of the invention, context information may be maintained using a context manager that may be centrally positioned to receive input from multiple sources and to provide output to multiple sources. According to one embodiment, the devices that communicate with the context manager may register through a registration module and may subscribe to one or more events. According to another embodiment of the invention, the context manager may receive input in Context XML form, for example. The other registered devices may be informed of context changes through a context tracking module to enable synchronizing of context across the registered modules. According to one embodiment of the invention, registered modules may be added or removed from the system. The registered modules may include dynamic link libraries (DLLs) that are specific to multimodal devices.

[15] According to yet another alternative embodiment of the invention, context information may be determined from a command or request that is presented in a textual format and/or a command or request that is presented as a verbal utterance and processed using a multi-pass automatic speech recognition module that transcribes the verbal utterance to a textual format. The multi-pass automatic speech recognition module may use a dictation grammar or a large vocabulary grammar, among other resources, to transcribe the verbal utterance into a text message. On platforms where a dictation grammar is not available, the multi-pass automatic speech recognition may use a virtual dictation grammar that uses decoy words for out-of-vocabulary words. Decoy words may include utility words, nonsense words, isolated syllables, isolated distinct sounds and other decoy words.

[16] According to one embodiment of the invention, the text message may be searched for a particular character, group of characters, word, group of words, and other text combinations. The text combination may be compared against entries in a context

description grammar that is associated with each agent. If a match is identified between an active grammar in the context description grammar and the command and/or request, then the match may be scored. The agents may be ranked based on the determined score. In generating an aggregate response from the one or more responses received from the agents, the ordering of the responses from the individual agents may be determined based on the rank of agents. The aggregate response may be generated by a response generator module. Agents may update a context stack, that includes an ordered list of command contexts, to enable follow-up requests.

[17] According to another aspect of the invention, the speech unit and/or multimodal device may be incorporated into the computer device or system, or may be separate structures. If separate structures are provided, the speech unit and/or multimodal devices may be connected to the computer device via a wired or wireless connection. If a wireless connection is provided, a base unit may be connected to the computer, internally or externally, to communicate with the speech unit and/or multimodal device.

[18] According to another aspect of the invention, the computer devices or systems may comprise stand alone or networked PCs, personal digital assistants (PDAs), cellular telephones, or other computer devices or systems. For convenience, these and other computer alternatives are referred to as computers. One aspect of the invention comprises software that may be installed onto the computer, where the software may include one or more of the following modules: a non-speech information receiving module; a speech recognition module that captures user utterances; a parser that parses the utterance; a text to speech engine module that converts the text to speech; a network interface that enables the computer to interface with one or more networks; a non-speech interface module and an event manager for managing events. Preferably, the event manager is in communication with a context description grammar, a user profile module that enables user profiles to be created, modified and accessed, a personality module that enables various personalities to be created and used, an agent module, an update manager, a cognitive model that provides statistical abstracts of user interaction patterns with the system, one or more databases, and other components.

[19] According to another aspect of the invention domain specific behavior and information may be organized into data managers. Data managers are autonomous executables that receive, process and respond to user questions, queries and commands. The data managers provide complete, convenient and re-distributable packages or modules of functionality, typically for a specific domain of application. Data managers may be complete

packages of executable code, scripts, links to information, and other forms of communication data that provide a specific package of functionality, usually in a specific domain. In other words, data managers include components for extending the functionality to a new domain. Further, data managers and their associated data may be updated remotely over a network as new behavior is added or new information becomes available. Data managers may use system resources and the services of other, typically more specialized, data managers. Data managers may be distributed and redistributed in a number of ways including on removable storage media, transfer over networks or attached to emails and other messages. An update manager may be used to add new data managers to the system or update existing data managers.

[20] According to another aspect of the invention, license management capabilities allowing the sale of data managers by third parties to one or more users on a one time or subscription basis may be provided. In addition, users with particular expertise may create data managers, update existing data managers by adding new behaviors and information, and making these data managers for other users as agents.

[21] In order to enhance the natural query and response environment, the system may format results to increase understandability to users. Formatting and presentation of results may be based on the context of the questions, the contents of the response being presented, the history of the interaction with the user, the user's preferences and interests and the nature of the domain. By contrast, rigid, highly formatted, or structured presentation of results may be deemed unnatural by many users.

[22] According to another embodiment of the invention, the system may simulate some aspects of a human "personality". In some cases, the presentation of the response and the terms that are used to provide the response may be randomized to avoid the appearance of rigidly formatted or mechanical responses. The use of other simulated personality characteristics is also desirable. For example, a response that may be upsetting to the user may be presented in a sympathetic manner. Furthermore, results of requests may be long text strings, lists, tables or other lengthy sets of data. Natural presentation of this type of information presents particular challenges because simply reading the long response is generally not preferred. Instead, the system may parse important sections from the response and may initially provide only reports. Determining what parts of a long response are presented may be based on the context of the questions, the contents of the response being presented, the history of the interaction with the user, the user's preferences and interests and the nature of the domain. At the same time, the system may give the user interactive control

over what information to present and how much information to present, to stop the response all together, or to take other actions.

[23] According to another aspect of the invention, the system may process and respond to questions, requests and/or commands. Keywords or context may be used to determine whether the received utterance and/or textual message includes a request or command. For example, utterances may include aspects of questions, requests and/or commands. For example, a user may utter "record my favorite TV program". A request is processed to determine the name, the channel, and time for the users favorite TV program. A command must be executed to set a video recorder to capture this program.

[24] For utterances including questions and/or requests, the system may perform multiple steps that may include one or more of:

- capturing the user's questions and/or requests through speech recognition components that operate in a variety of real-world environments;
- parsing and interpreting the question and/or request;
- determining the domain of expertise and context, invoking the proper resources, including agents;
- formulating one or more requests to one or more local and/or network data sources or sending appropriate commands to local or remote devices or the system itself;
- performing presentation formatting, variable substitutions and transformations to modify the requests to a form that yields desired results from the available sources;
- executing the multiple requests or commands in an asynchronous manner and dealing gracefully with failures;
- extracting or scraping the desired information from the one or more results, which may be returned in any one of a number of different formats;
- evaluating and interpreting the results, including processing of errors to provide one or more results judged to be "best," even if the results are ambiguous, incomplete, or conflicting;
- performing formatting, variable substitutions and transformations to modify the results to a form most easily understood by the user; and
- presenting the compound result, through a text to speech engine or a multimodal interface, to the user in a useful and expected manner.

[25] The above steps may be performed with knowledge of the domain of expertise, the context for the question or command, domain specific information, the history of the user's interactions, user preferences, available information sources or commands, and responses obtained from the sources.

[26] Probabilistic or fuzzy set decision and matching methods may be applied to deal with inconsistent, ambiguous, conflicting and incomplete information or responses. In addition, asynchronous queries may be used to provide rapid and graceful failure of requests or commands that allow the system to robustly return results quickly and in a manner that seems natural to the user.

[27] Many everyday questions are inherently subjective and result in answers that are a matter of opinion or consensus, as much as fact. Such questions are often ad hoc in their nature, as well. The system, according to another aspect of the invention, may use adaptive, probabilistic, and fuzzy set decision and matching methods to identify the subjective nature of the question and to evaluate a range of possible answers, wherein one or more answers may be selected that most accurately represent the type of result desired by the user.

[28] The context and expected results from a particular question may be highly dependent on the individual asking the question. Therefore, the system may create, store and use personal profile information for each user. Information in the profile may be added and updated automatically as the user uses the system or may be manually added or updated by the user or by others. Domain specific agents may collect, store and use specific profile information, as may be required for optimal operations. Users may create commands for regularly used reports, automatically generated alerts, and other requests and for the formatting and presentation of results. The system may use profile data in interpreting questions, formulating requests, interpreting request results and presenting answers to the user. Examples of information in a user profile includes, history of questions asked, session histories, formatting and presentation preferences, special word spelling, terms of interest, special data sources of interest, age, sex, education, location or address, place of business, type of business, investments, hobbies, sports interests, news interests and other profile data.

[29] According to one aspect of the invention, the system may attempt to provide rapid responses in order to provide a natural question and response environment. The rapid responses may be provided without obtaining additional information. The system may determine agent composition, context and/or domain for a user's question or command, for example, by using a real-time scoring system or other technique. Based on this determination, the system may trigger one or more agents to respond to the user's question or

command. The agents may make one or more requests and rapidly return a formatted response. Thus, users may receive direct responses to a set of questions, each with a different response or context. In some cases, the available information, including the request results, may not adequately answer the questions presented. In such situations, the user may be asked one or more follow-up questions to resolve the ambiguity. Additional requests may then be made before an adequate response is provided. In these cases, the system may use context information, user profile information and/or domain specific information to minimize the interaction with the user required to deliver a response.

[30] If the confidence level of the domain or context score is not high enough to ensure a reliable response, the system may request that the user verify that the question or command is correctly understood. In general, the question may be phrased to indicate the context of the question including all criteria or parameters. If the user confirms that the question is correct, the system may proceed to produce a response. Otherwise, either the user can rephrase the original question, perhaps adding additional information to remove ambiguity, or the system may ask one or more questions to attempt to resolve the ambiguity or other actions may taken.

[31] According to one aspect of the invention, the system may accept any natural language question or command and, as a result, may be subject to ambiguous requests. To assist users in formulating concise questions and commands, the system may support a voice query language. The language may help users clearly specify the keywords or contexts of the question or command along with the parameters or criteria. The system may provide built in training capabilities to help the user learn the best methods to formulate their questions and commands.

[32] In order to make the responses to user's questions and commands seem more natural, the system may employ one or more dynamically invokable personalities and/or emotional models. Personalities and emotional models have specific characteristics that simulate the behavioral characteristics of real humans. Examples of these characteristics include, sympathy, irritation, and helpfulness and associated emotions. The personality also randomizes aspects of responses, just as a real human would do. This behavior includes randomization of terms used and the order of presentation of information. Characteristics of the personality and/or emotions are invoked using probabilistic or fuzzy set decision and matching methods and using criteria, including the context for the question, the history of the user's interaction, user preferences, information sources available, responses obtained from the sources.

[33] According to another aspect of the invention, special procedures may be employed to present information in the form of long text strings, tables, lists or other long response sets. Simply presenting a long set of information in an ordered manner may not be considered natural or what most users have in mind. The system may use probabilistic or fuzzy set matching methods to extract relevant information and present these subsets first. Further the system may provide commands allowing users to skip through the list, find keywords or key information in the list or stop processing the list altogether.

[34] In addition to the ability to process user questions and commands, the invention may provide local or remote control functions for the system or for other devices. Users may initiate commands locally or remotely. Several techniques for executing remote operations are possible, for example, the use of a telephone or other audio connection. Other remote command techniques may be used. The processing of these commands may be performed in a nearly identical manner to a question. One difference being that the result of the command is generally an action rather than a response. Examples of control operations supported by the invention include, finding and playing music, film, games or other entertainment media, control of the appliances or other electronic or electro-mechanical devices, security systems control, audio, text and video conferencing (including Voice over IP conferencing), operating integrated messaging environments and other control operations.

[35] By using an optional analog or digital telephone interface or a data network interface, the invention can provide remote control capability of the computer or and auxiliary devices connected to the computer. For example, if the computer is connected via an interface to one or more auxiliary devices, a remote user can use this capability to remotely control or program the operation of the auxiliary device. For example, this enables recording of a television program, checking the status of a home security system, checking voice mail messages, hearing and responding to email messages, and much more. For each user command utterance, the system may execute a number of steps possibly including:

- capturing the user's command through speech recognition that operates in a variety of real-world environments;
- parsing and interpreting the command;
- determining the domain for the command and context, invoking the proper resources, including agents;
- formulating device specific commands for the system or external devices;

- routing commands to the system, or external devices, including external devices connected to data networks;
- receiving and processing results of commands, including errors; and
- optionally, providing responses to users indicating the success or failure of the command, and possibly including state information.

[36] The system may be used in conjunction with a wide range of platform environments. These environments may include, as a peripheral to a PC, a portable or wearable computer or other computer, embedded in a PC or other computer, on a personal digital assistant, such as a PDA with wireless networking capability, in a specialized device such as a bar code scanner or payment terminal, on a wireless telephone, or other platforms. If the invention is used with a mobile or portable device that has position location capability, the location data can be used by the invention as part of the context for user questions. A user may use the system on one or more devices. In this case, history and profile information for the user may be synchronized between the multiple devices on a periodic basis or other basis.

[37] According to another aspect of the invention, the system may be deployed in a network of devices that use a common base of agents, data, information, user profiles and histories. Each user can then interact with, and receive, the same services and applications at any location equipped with the required device on the network. For example, multiple devices on which the invention is deployed, and connected to a network, can be placed at different locations throughout a home, place of business or other location. In such a case, the system may use the location of the particular device addressed by the user as part of the context for the questions asked.

[38] According to one embodiment of the invention, the system may support multiple users that access the system at different times. According to another embodiment of the invention, the system may support multiple users that access the system during a same session in an interleaved or overlapping manner. The system may recognize the multiple users by name, voice, or other characteristic and may invoke a correct profile for each user. If multiple users are addressing the system in overlapping or interleaved sessions, the system may identify the multiple users and may invoke one or more corresponding profiles. For applications that require security safeguards, the multiple users may be verified using voiceprint matching, password or pass-phrase matching, or other security safeguards.

[39] When multiple users are engaged in interleaved sessions, the system may gracefully resolve conflicts using a probabilistic or fuzzy set decision method for each user. This

process may simulate the manner in which a human would address multiple questions from various sources. For example, the system may answer short questions first in time while answering longer questions later in time. Alternatively, the system may answer questions in the order that they are received, among other configurations.

[40] Given the desire for domain specific behavior, user specific behavior and domain specific information, the system, according to another aspect of the invention, may allow both users and content providers to extend the system capabilities, add data to local data sources, and add references to network data sources. To allow coverage of the widest possible range of topics, the system may allow third party content developers to develop, distribute and sell specialized or domain specific system programs and information. Content may be created though creation of new data managers, scripting existing data managers, adding new data to agents or databases and adding or modifying links to information sources. Distribution of this information is sensitive to the users interests and use history and to their willingness to pay or not.

[41] According to another aspect of the invention, the system may allow users to disseminate their knowledge without programming. The system may include mechanisms that allow users to post and distribute data managers and information in their particular areas of expertise, to improve system capability. Further, users can extend the system and configure it to their own preferences, add information to their profile to define new questions or queries, extend and modify existing questions and queries, add new data sources, update data sources, set preferences and specify presentation parameters for results.

[42] Since the invention may operate in many environments, including environments with background noise, point noise sources and people holding conversations, filtering of speech input may be performed. The invention may use, for example, either one-dimensional or two-dimensional array microphones to receive human speech. The array microphones may be fixed or employ dynamic beam forming techniques. The array pattern may be adjusted to maximize gain in the direction of the user and to null point noise sources. Speech received at the microphones may then be processed with analog or digital filters to optimize the bandwidth, cancel echoes, and notch-out narrow band noise sources. Following filtering, the system may use variable rate sampling to maximize the fidelity of the encoded speech, while minimizing required bandwidth. This procedure may be particularly useful in cases where the encoded speech is transmitted over a wireless network or link.

[43] Some example applications for the invention includes but are not limited to the following:

- white pages and yellow pages lookups to find, email addresses, telephone numbers, street addresses and other information for businesses and individuals;
- personal address book, calendars and reminders for each user;
- automatic telephone dialing, reading and sending emails and pages by voice and other communications control functions;
- map, location and direction applications;
- movie or other entertainment locator, review information and ticket purchasing;
- television, radio or other home entertainment schedule, review information and device control from a local or remote user;
- weather information for the local area or other locations;
- stock and other investment information including, prices, company reports, profiles, company information, business news stories, company reports, analysis, price alerts, news alerts, portfolio reports, portfolio plans, etc.;
- flight or other scheduled transportation information and ticketing;
- reservations for hotels, rental cars and other travel services;
- local, national and international news information including headlines of interest by subject or location, story summaries, full stories, audio and video retrieval and play for stories.;
- sports scores, news stories, schedules, alerts, statistics, back ground and history information, etc.;
- ability to subscribe interactively to multimedia information channels, including sports, news, business, different types of music and entertainment, applying user specific preferences for extracting and presenting information;
- rights management for information or content used or published;
- horoscopes, daily jokes and comics, crossword puzzle retrieval and display and related entertainment or diversions;
- recipes, meal planning, nutrition information and planning, shopping lists and other home organization related activities;
- as an interface to auctions and online shopping, and where the system can manage payment or an electronic wallet;

- management of network communications and conferencing, including telecommunications, email, instant messaging, Voice over IP communications and conferencing, local and wide area video and audio conferencing, pages and alerts;
- location, selection, management of play lists and play control of interactive entertainment from local or network sources including, video on demand, digital audio, such as MP3 format material, interactive games, web radio and video broadcasts;
- organization and calendar management for families, businesses and other groups of users including the management of, meetings, appointments, and events; and
- interactive educational programs using local and network material, with lesson material level set based on user's profile, and including, interactive multimedia lessons, religious instruction, calculator, dictionary and spelling, language training, foreign language translation and encyclopedias and other reference material.

[44] It will be appreciated that the foregoing statements of the features of the invention are not intended as exhaustive or limiting, the proper scope thereof being appreciated by reference to this entire disclosure and reasonably apparent variations and extensions thereof.

BRIEF DESCRIPTION OF THE DRAWINGS

[45] The invention will be described by reference to the preferred and alternative embodiments thereof in conjunction with the drawings in which:

[46] FIGURE 1 is an overall diagrammatic view according to one embodiment of the invention;

[47] FIGURE 2 is a schematic block diagram showing the agent architecture according to one embodiment of the invention;

[48] FIGURE 3 is a high level process for receiving natural language speech-based queries and/or commands and generating a response according to one embodiment of the invention;

[49] FIGURE 4A is a process for receiving natural language speech-based queries and/or commands and generating a response according to one embodiment of the invention;

[50] FIGURE 4B is a process for receiving natural language speech-based commands in order to locally or remotely control functions of devices according to one embodiment of the invention;

[51] FIGURE 5 is a process for correctly interpreting a user's utterance according to one embodiment of the invention; and

[52] FIGURE 6 is a process for determining the proper domain agents to invoke and the properly formatted queries and/or commands that is to be submitted to the agents according to one embodiment of the invention.

[53] FIGURE 7 illustrates a diagram for centrally maintaining context, according to one embodiment of the invention.

[54] FIGURE 8 illustrates a system diagram for an enhanced conversational voice user interface according to one embodiment of the invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[55] Barriers to natural human request-response interaction between humans and machines include that natural language queries may be incomplete in their definition. Furthermore, context of previous questions, knowledge of the domain, and/or the user's history of interests and preferences may be needed to properly interpret questions and/or commands. Thus, natural language questions may not be easily transformed to machine processable form. Compounding this problem, natural language questions may be ambiguous or subjective. In these cases, forming a machine processable query and returning a natural language response may be difficult.

[56] In order for machines (e.g., computer devices) to properly respond to questions or commands that are submitted in a natural language form, machine processable queries and algorithms may be formulated after the natural form questions or commands have been parsed and interpreted. Algorithms describe how the machines should gather data to respond to the questions or commands. Depending on the nature of the questions or commands, there may not be a simple set of requests and/or algorithms that will return an adequate response. Several requests and algorithms may need to be initiated and even these requests and algorithms may need to be chained or concatenated to achieve a complete response. Further, no single available source may contain the entire set of results needed to generate a complete response. Thus, multiple requests and/or algorithms, perhaps with several parts, may be generated to access multiple data sources that are located both locally or remotely. Not all of the data sources, requests and/or algorithms may return useful results or any results at all.

[57] Useful results that are returned are often embedded in other information and may need to be extracted from the other information. For example, a few key words or numbers may need to be "scraped" from a larger amount of other information in a text string, table, list, page, video stream or other information. At the same time, extraneous information including

graphics or pictures may be removed to process the response. In any case, the multiple results must be evaluated and combined to form the best possible response, even in cases where some requests do not return useful results or fail to produce results entirely. In cases where the question is determined to be ambiguous or the result is inherently subjective, determining the results to present in the response is a complex process. Finally, to maintain a natural interaction, responses should be returned to the user rapidly. Managing and evaluating complex and uncertain requests, while maintaining real-time performance, is a significant challenge.

[58] The following detailed description refers to the accompanying drawings and describes exemplary embodiments of the invention. Other embodiments are possible and modifications may be made to the exemplary embodiments without departing from the spirit, functionality and scope of the invention. Therefore, the following detailed descriptions are not meant to limit the invention.

[59] The invention provides a complete speech-based information query, processing and presentation environment or a combination of speech-based and non-speech-based information query, processing and presentation environment. In addition, the invention may be useful for controlling the system itself and/or external devices. This integrated environment makes maximum use of context, prior information and domain and user specific profile data to achieve a natural environment for one or more users submitting requests or commands in multiple domains. Through this integrated approach, a complete speech-based natural language query, algorithm and response environment or a combination of speech-based and non-speech-based command, algorithm and response environment may be created. FIG. 1 shows one exemplary schematic diagram view of a system 90 according to an embodiment of the invention.

[60] The system 90 may include a main unit 98, speech units 128, and multi-modal devices 155. Alternatively, the system 98 may include completely separate systems for main unit 98, speech units 128 and multi-modal devices 155. The event manager 100 may mediate interactions between other components of main unit 98. The event manager 100 provides a multi-threaded environment allowing the system 98 to operate on multiple commands or questions from multiple user sessions, without conflict, and in an efficient manner to maintain real-time response capabilities.

[61] FIG. 8 illustrates one exemplary embodiment of an enhanced conversational user interface 800 that receives multi-modal input from users 802 and communicates with a conversational speech analyzer 804. The conversational speech analyzer 804 may be coupled

to several components, including a general cognitive model 806, an environmental model 808, a personalized cognitive model 810, and an adaptive misrecognition analysis engine 812, or other components. According to one embodiment of the invention, the conversational speech analyzer 804 may include one or more speech recognition engines that transcribe utterances to textual messages. The transcription may be performed using one or more conversational grammars, a context description grammar 112, or other transcription methods. According to one embodiment of the invention, data that is entered in a text format may be merge with data that is transcribed to a textual format from the utterance. Conversational speech analyzer 804 also may include a semantic knowledge-based model that analyzes the text message and detects command components.

[62] According to one embodiment of the invention, personalized cognitive model 810 is a model derived from a user's interaction pattern with the system and may be used to predict what actions the user may take next in time, thus assisting with speech recognition and/or question or command recognition. Personalized cognitive model 810 may track actions performed by the user. When the system is attempting to predict user behavior, the personalized cognitive model may be consulted first. The system may have multiple personalized cognitive models, wherein one or more may correspond to each user.

[63] According to another embodiment of the invention, general cognitive model 806 is a statistical abstract that corresponds to interaction patterns with the system for multiple users. Data stored within general cognitive model 806 may be used to predict a user's next action, thus assisting with speech recognition and/or question or command recognition. The general cognitive model 806 may also track what actions a particular user has performed and may be used when the user interacts with the system in a way that is not handled in the personalized cognitive model.

[64] According to one embodiment of the invention, the environmental model 808 may include information associated with the user's environment and surroundings. The information may include the type of environment that a user is in (e.g., quiet or noisy); details of a microphone and/or speaker system; the user's current global position and movement, as may be determined by GPS; current system status, such as what song/movie is playing, is the system in the midst of retrieving something, or other system status; details on all voice-enabled devices in the immediate vicinity, such as a presence of a voice-enabled TV, stereo, and DVD player in the same room; a user's credit card information, such as numbers and current balances, wherein the user may ask a mobile telephone to download and pay for a video and the system may respond that there are insufficient funds; or other information. The

information may be accessed to invoke a context, domain knowledge, preferences, and/or other cognitive quality that enhances the interpretation of questions and/or commands.

[65] The conversational speech analyzer 804 may also access the general cognitive model 806 and/or the personalized cognitive model 810 to further refine context, domain knowledge, preferences, and/or other cognitive qualities to enhance the interpretation of questions and/or commands. Based on information received from general cognitive model 806, environmental model 808 and/or the personalized cognitive model 810, the system may enhance responses to commands and questions by including a prediction of user behavior.

[66] Adaptive Misrecognition Analysis Engine 812 may analyze and store textual messages, including transcribed utterances that are identified as being unrecognized or incorrectly recognized by conversational speech analyzer 804. Upon a determination that text is unrecognized, the system may generate an unrecognized event. For example, an unrecognized event may result from not finding a match to text and/or transcribed utterance.

[67] According to one embodiment of the invention, the system may implement one or more techniques to determine that textual messages are incorrectly recognized. For example, the user may command the system to play a specific song title and the system may misrecognize the requested song title and provide a song having a different title, the system may misrecognize an incorrect song title and provide a song having a different than requested title, the system may misrecognize a correct song title and provide a song having an invalid song title, among other variations. When the system misrecognizes a request, the user typically provides immediate feedback such as, overriding the command in a time shorter than the expected time execution time of the command, repeating the original request, issuing a stop command, or taking other action, wherein the action may be presented verbally, non-verbally by pushing a selected button on a cell phone or remote control, or both, among other configurations. According to one embodiment of the invention, the system may detect the user action and may prompt the user to re-phrase the request to enable the system to modify words in the query. The user's actions may be analyzed in a statistical model to determine a frequency occurrence of misrecognitions for particular commands, with the results being used to update the corresponding personalized cognitive model 810.

[68] According to another embodiment of the invention, the conversational speech analyzer 804 may access the personalized cognitive model 810 to proactively select a next best (or nth best) match for the received text. A match may be confirmed by user action that include the user not immediately canceling the command. The misrecognitions may also be analyzed to potentially determine personalized tuning parameters for the speech recognition

components of the system. For example, the system may, over time, tune the speech recognition engine to increase recognition by analyzing how the speech recognition engine misrecognizes the verbal utterance.

[69] Agents 106 may include a collection of grammars, criteria handlers, and algorithms that are accessed to respond to a set of requests and/or commands. Agents 106 further may contain packages of both generic and domain specific behavior for the system 98. Agents 106 may use nonvolatile storage for data, parameters, history information, and locally stored content provided in the system databases 102 or other local sources. One or more user profiles 110 may be provided that includes user specific data, parameters, and session and history information for determining the behavior of agents 106. One or more personality modules 108 may be provided in a data determining system that includes personality characteristics for agents 106. The update manager 104 manages the automatic and manual loading and updating of agents 106 and their associated data from the Internet 146 or other network through the network interface 116.

[70] According to one embodiment of the invention, the speech-based interface for the system 90 may include one or more speech units 128. Speech units 128 may include one or more microphones, for example array microphone 134, to receive the utterances from the user. The speech received at the microphone 134 may be processed by filter 132 and passed to the speech coder 138 for encoding and compression. In one embodiment, a transceiver module 130 may transmit the coded speech to the main unit 98. Coded speech received from the main unit 98 is detected by the transceiver 130, then decoded and decompressed by the speech coder 138 and annunciated by the speaker 136.

[71] According to one embodiment of the invention, the non-speech-based interface for the system 90 may include one or more multi-modal devices 155 that may include mobile devices, stand alone or networked computers, personal digital assistances (PDAs), portable computer devices, or other multi-modal devices.

[72] The speech units 128, multi-modal devices 155 and the main unit 98 may communicate over a communication link. The communication link may include a wired or wireless link. According to one embodiment, the communication link may comprise an RF link. The transceiver 130 on the speech unit may communicate coded speech data bi-directionally over the communication link with the transceiver 126 on the main unit 98. According to another embodiment, RF link may use any standard local area wireless data protocols including the IEEE 802.11, Bluetooth or other standards. Alternatively, an infrared data link conforming to any suitable standard such as IrDA or other infrared standards can be

used. In an alternative embodiment, wires may connect the speech unit 128 and the main unit 98, eliminating the need for one speech coder 138. Other wired or wireless analog or digital transmission techniques can be used.

[73] According to one embodiment of the invention, coded speech received at the transceiver 126 on the main unit 98 may be passed to the speech coder 122 for decoding and decompression. The decoded speech may be processed by the speech recognition engine 120 using the context description grammar module 112, among other information. Any recognized information may be processed by the parser 118, which transforms information into complete algorithms and questions using data supplied by knowledge agents. Knowledge agents may be selected from the plurality of agents 106 using a grammar stack, wherein the knowledge agents provide information for generating a response to the question or command. The knowledge agents may then process the commands or questions by creating requests that are submitted to local databases 102 or submitted through the network interface 116 to external data sources over the Internet 146 or external other networks. Algorithms typically result in actions taken by the system 90 itself (i.e., pause or stop), or to a remote device or data source (i.e., download data or program, or control a remote device), through the network interface to the Internet or other data interface.

[74] According to one embodiment of the invention, knowledge agents may return results of questions as responses to users. The responses may be created using the results of information requests, the system personality 108, the user preferences, other data in the user profile 110, and/or other information. Agents 106 may present these results using the speech unit 128. The agents 106 may create a response string, which is sent to the text to speech engine 124. The text to speech engine 124 may generate the utterances, which may be encoded and compressed by the speech coder 122. Once coded, the utterances are transmitted from the main unit 98 by the transceiver 126 to the transceiver 130 on the speech unit 128. The utterance is then decoded and decompressed by the speech coder 138 and output by the speaker 136. Alternatively, agents 106 may present the results using multi-modal devices 155.

[75] The non-speech interface 114 may be integral with, or separate from, the multi-modal devices 155 and may be used as a substitute or may complement the speech interface. For example, non-speech interface 114 may be used to present and to interact with non-speech (e.g., graphical or tabular) information in a manner more easily understood by the user. According to one embodiment of the invention, multimodal support may be provided to maintain the context during both voice interaction and interaction through the non-speech

interface 114. In one exemplary embodiment, a communication channel may be opened between multimodal devices 155 and the main user interface system 90 to allow multimodal devices 155 to input text commands and questions. The multimodal devices 155 may received and/or provide one or more types of communications, including speech, text, digital audio files, dual-tone-multi-frequency (DTMF) or other communication types. According to another embodiment of the invention, multimodal devices 155 may send a string of text or keywords for a command or question. The main interface system 90 may synchronize the context between multimodal device 155 and the speech units 128. In order to send a response to the corresponding device, the main user interface system 90 may track where the input came from so that the response may be sent to a TTS or multi-modal device 155.

[76] According to an alternative embodiment of the invention illustrated in FIG. 7, a context manager 702 may be centrally maintained to allow input from and output to multiple mobile devices 36. Each of the mobile devices may communicate with the context manager 702 to register via registration module 712, wherein registration may indicate events that the mobile devices 36 may subscribe to. The context manager 702 may receive input in Context XML form, for example. The other registered mobile devices 36 may be informed of context changes through context tracking module 714, thereby synchronizing the context across the registered mobile devices 36. According to one embodiment of the invention, registered terminal devices may be added or removed. The registered terminal devices may be dynamic link libraries (DLLs) that may be specific to a multimodal device.

[77] According to yet another alternative embodiment of the invention, context information may be determined from a command or request that is presented as a text message and/or a command or request that is presented as a verbal utterance and processed using a multi-pass automatic speech recognition module that transcribes the verbal utterance to a text message. The multi-pass automatic speech recognition module may use a dictation grammar or a large vocabulary grammar, among other resources, to transcribe the verbal utterance into a text message. On platforms where a dictation grammar is not available, the multi-pass automatic speech recognition may use a virtual dictation grammar that uses decoy words for out-of-vocabulary words. Decoy words may include utility words, nonsense words, isolated syllables, isolated distinct sounds and other decoy words.

[78] According to one embodiment of the invention, the text message may be searched for a particular character, group of characters, word, group of words, and other text combinations. The text combination may be compared against entries in a context description grammar that is associated with each agent 106. If a match is identified between

an active grammar in the context description grammar and the command and/or request, then the match may be scored. The agents 106 may be ranked based on the determined score. In generating an aggregate response from the one or more responses received from the agents 106, the ordering of the responses from the individual agents may be determined based on the rank of agents 106. The aggregate response may be generated by a response generator module. Agents 106 may update a context stack, that includes an ordered list of command contexts, to enable follow-up requests.

[79] According to another embodiment of the invention, if a match is not found, or only a partial match is found, between the text message and active grammars, then a knowledge-enhanced speech recognition system may be used to semantically broaden the search. The knowledge-enhanced speech recognition system may be used to determine the intent of the request and/or to correct false recognitions. The knowledge-enhanced speech recognition may access a set of expected contexts that are stored in a context stack to determine a most likely context. The knowledge-enhanced speech recognition may use context specific matchers that are able to identify context such as time, location, numbers, dates, categories (e.g., music, movies, television, addresses, etc.) and other context. The matching may be performed by comparing a character, group of characters, a word, group of words, and other text combinations. Alternatively, or in addition to text based matching, the matching may be performed using phonetic matching, among other techniques. The results of any match may be used to generate a command and/or request that is communicated to agents 106 for additional processing. According to one embodiment of the invention, non-speech interface 114 may show system, state and history information in a more concise manner than is possible through the speech interface. Non-speech interface 114 may be accessed to create or extend capabilities of agents 106. These operations may include scripting of agents, adding data to the agent or databases 102 used by the agent, adding links to information sources, among other operations.

[80] According to another embodiment of the invention, system 90 may include different types of agents 106. For example, generic and domain specific behavior and information may be organized into domain agents. A system agent, on the other hand, may provide default functionality and basic services. The domain agents provide complete, convenient and re-distributable packages or modules for each application area. In other words, the domain agents include data that is needed to extend or modify the functionality of the system 90 in a current or new domain. Further, domain agents and their associated data can be updated remotely over a network as new behavior is added or new information becomes

available. Domain agents may use the services of other, typically more specialized, data managers and the system agent 150. Agents are distributed and redistributed in a number of ways including on removable storage media, transfer over networks or attached to emails and other messages. The invention may provide license management capability allowing the sale of data managers by third parties to one or more users on a one time or subscription basis. In addition, users with particular expertise may create data managers, update existing data managers by adding new behaviors and information and making these data managers to other users as agents. A block diagram of an agent architecture according to an embodiment of the invention is shown in FIG. 2.

[81] Agents 106 may receive and return events to the event manager 100. Both system agents 150 and domain agents 156 may receive questions and commands from the parser 118. Based on keywords in the questions and commands and the structures of the questions and commands, the parser may invoke selected agents. Agents use the nonvolatile storage for data, parameters, history information and local content provided in the system databases 102.

[82] According to one embodiment of the invention, when the system starts-up or boots-up, the agent manager 154 may load and initialize the system agent 150 and the one or more domain agents 156. Agent manager 154 includes knowledge of agents 106 and maps agents 106 to the agent library 158. At shutdown, the agent manager may unload the agents 106. The agent manager 154 also performs license management functions for the domain agents 156 and content in the databases 102.

[83] The system agent 150 manages the criteria handlers 152 that handle specific parameters or values (criteria) used to determine context for questions and commands. According to one embodiment of the invention, criteria handlers 152 include parsing routines that are specialized to recognize particular parts of speech, such as times, locations, movie titles, and other parts of speech. The criteria handlers 152 may identify matching phrases and extract semantic attributes from the phrases. Both the system agent 150 and the domain agents 156 may use the criteria handlers 152. The various domain agents 156 may use the services of the system agent 150 and of other, typically more specialized, domain agents 156. The system agent 150 and the domain agents 156 may use the services of the agent library 158, which contains utilities for commonly used functions. According to one embodiment of the invention, the agent library may be a dynamic link library that implements one or more agents. The agent library may include utilities for text and string handling, network communications, database lookup and management, fuzzy and probabilistic evaluation, text to speech formats, and other utilities.

[84] Domain agents 156 may be data-driven, scripted or created with compiled code. A base of generic agent may be used as the starting point for data-driven or scripted agents. Agents created with compiled code are typically built into dynamically linkable or loadable libraries. Developers of agents can add new functionality to the agent library 158 as required. Details of agent distribution and update, and agent creation or modification are discussed in sections below.

[85] According to another embodiment of the invention, capabilities are provided to distribute and update system agents 150, domain agents 156, agent library 158 components, databases 102, and context description grammar 112 over wireless or wired networks 136, including dial-up networks using the update manager 104. The network interface 116 may provide connections to one or more networks. The update manager 104 may also manage the downloading and installation of core system updates. The agent manager 154 may perform license management functions for the domain agents and the databases. The update manager 104 and agent manager 154 may perform these functions for all agents and database content including, agents and content available to all users or agents and/or content available to selected users. Examples of agent and database components added or updated on a periodic basis include:

- agents for new domains;
- additional domain knowledge for agents;
- new keywords for a domain, which can include names of politicians, athletes, entertainers, names of new movies or songs, etc. who have achieved recent prominence;
- links to a preferred set of information sources for the domains covered including links for, entertainment, news, sports, weather, etc.;
- updates to domain information based on, for example, changes to tax laws, company mergers, changing political boundaries;
- updates to content, including dictionaries, encyclopedias and almanacs; and
- other content and database components.

[86] When a user requires or selects a new domain agent 156 or database element 102, the update manager 104 may connect to their source on the network 146 though the network interface 116, download and install the agent or data. To save system resources and to comply with any license conditions, the update manger 104 may uninstall agents that are no longer in use. In one embodiment of the invention, the update manager 104 may periodically

queries one or more sources of the licensed agents and database components to locate and download updates to agent executables, scripts or data as they become available. Alternatively, the agent sources may initiate the downloading of agent updates of the registered or licensed agents to the update manager as they become available.

[87] The agent manager 154 may provide license management clients that are capable of executing most any license terms and conditions. When a particular agent 106 and/or database element 102 is selected based on a submitted command, the agent manager 154 verifies that the use of the agent or data element is within the allowed terms and conditions, and if so, invokes the agent or allows access to the data element. License management schemes that can be implemented through the agent manager 154 include outright purchase, subscription for updates, one time or limited time use. Use of shared agents and data elements (such as those down-loaded from web sites maintained by groups of domain experts) may also be managed by the agent manager 154.

[88] If questions or commands do not match an agent that is currently loaded on the system, the agent manager 154 may search the network 146 through the network interface 116 to find a source for a suitable agent. Once located, the agent can be loaded under the control of the update manager 104, within the terms and conditions of the license agreement, as enforced by the agent manager.

[89] New commands, keywords, information, or information sources can be added to any domain agent 156 by changing agent data or scripting. These configuration capabilities may allow users and content developers to extend and modify the behavior of existing domain agents 156 or to create new domain agents 156 from a generic agent without the need to create new compiled code. Thus, the modification of the domain agents 156 may range from minor data-driven updates by even the most casual users, such as specifying the spelling of words, to development of complex behavior using the scripting language as would typically be done by a domain expert. The user can create and manage modifications to domain agents 156 through speech interface commands or using non-speech interface 114. User-specific modifications of domain agents 156 are stored in conjunction with the user's profile 110 and accessed by the domain agent 156 at run-time.

[90] The data used to configure data driven agents 156 are structured in a manner to facilitate efficient evaluation and to help developers with organization. These data are used not only by the agents 156, but also by the speech recognition engine 120, the text to speech engine 124, and the parser 118. Examples of some major categories of data include:

1. Content packages include questions or commands. Each command or question or group of commands or questions includes contexts used for creation of one or more requests. The domain agent 156 passes a context description grammar expression to the parser 118 for evaluation of a context or question. An initial or default context is typically supplied for each command or question. The command or question may include a grammar for the management and evaluation of the context stack.

2. Page lists or pointers to other local or network content sources. For each page or content source there is a pointer (e.g. URL, URI, or other pointer) to the page or source. Each page has specific scraping information used to extract the data of interest. The scraping information includes matching patterns, HTML or other format parsing information.

3. A response list, determining the response of the domain agent 156 to a particular command or question given the context, the user profile and the information retrieved. Responses can include diagnostic error messages or requests for more information if the question or command cannot yet be resolved from the known information. Responses can be based on or dependent on thresholds or probabilistic or fuzzy weights for the variables.

4. Substitution lists containing variable substitutions and transformations, often applied by the agents 150, 156 in the formatting of queries and results. For example, a stock domain specific agent 156 would use a substitution list of company trading symbols, company names and commonly used abbreviations. Substitutions and transformations can be performed on commands and questions to create precise queries, which can be applied against one or more information sources or to results for creating more meaningful output to the user. Substitution lists also include information for optimally dealing with structured information, such as HTTP formatted page parsing and evaluation.

5. Personalities used for responses. Personalities are constructed by combining multiple traits in a weighted manner. Weights can be specified for each agent's domain area to create one or more specific personalities. Examples of personality traits include sarcasm, humor, irritation, and sympathy, and other traits.

6. Public and user specific parameters for sources, substitutions, transformations, variables or criteria. The public parameter lists are part of the agent package 156. The user specific parameters are contained in the user profile 110.

[91] Commands and questions are interpreted, queries formulated, responses created and results presented based on the users personal or user profile 110 values. Personal profiles may include information specific to the individual, their interests, their special use of terminology, the history of their interactions with the system, and domains of interest. The personal profile data may be used by the agents 106, the speech recognition engine 120, the text to speech engine 124, and the parser 118. Preferences can include, special (modified) commands, past behavior or history, questions, information sources, formats, reports, and alerts. User profile data can be manually entered by the user and/or can be learned by the system 90 based on user behavior. User profile values may include:

- spelling preferences;
- date of birth for user, family and friends;
- income level;
- gender;
- occupation;
- location information such as, home address, neighborhood, and business address;
- car type;
- telecommunications and other service providers and services;
- financial and investment information;
- synonyms (i.e., a nick name for someone);
- special spelling;
- keywords;
- transformation or substitution variables;
- domains of interest; and
- other values.

[92] End users may use the data driven agent 156 extension and modification facilities and values stored in user profiles 110 to create special reports, packages of queries, alerts and output formats. A single alert or report can be configured to use multiple data sources and other variable (i.e., time, location) value to determine when alerts should be sent. For example, an alert can be generated by sampling a stock price every 15 minutes and sending an alert if the price drops below some value. To create a report, the user first specifies a set of commands or questions. Next, the user creates or selects a format for the report. Finally the user may name the report. A report can have variable parameters. For example, users

may create a company stock report, and execute the report by stating its name and the company name, which gives the user selected information and in a specified format for that company. In another example, users can create a "morning" report, which presents selected multimedia information from different sources (news, sports, traffic, weather) in the order and formats desired. Alerts and reports can be created using only voice commands and responses, commands and responses through non-speech interface 114, or a combination of the two. To create a report, alert, or other specialized behavior, the user performs a number of steps including:

- specify the command to run a report or alert;
- specify the question or questions, including keywords, used for a query;
- set the criteria for running the report such as on command or when a particular condition is met;
- define preferred information sources;
- define preferences for order of result evaluation by source, value, etc.;
- specify the presentation medium for a report or alert, such as an email, the text to speech engine, a message to a pager, or a text and graphics display; and
- specify the preferred format for the report, such as information to be presented, order of information to be presented, preferred abbreviations or other variable substitutions.

[93] Filtering and noise elimination may be important in facilitating the various functionalities of system 90. Recognition and parsing of the user's speech may be implemented with good signal to noise ratio at the input to the speech recognition engine 120. To provide acceptable results, a set of acoustic models, an array microphone 134 a filter 132, or other components, may be employed. If a good signal to noise ratio cannot be achieved, a noise identification algorithm may be used and the appropriate acoustic model, for example, one that has been trained in conditions similar to the identified noise, may be selected. According to an embodiment of the invention, the microphone array, filters and speech coder 138 are physically separated from the main unit 98 into a speech unit 128, and connected using a wireless link. Since bandwidth on a wireless connection is at a premium, the speech coder dynamically adapts the digitization rate and compression of the captured speech.

[94] The speech unit 128 may use an array of microphones 134 to provide better directional signal capture and noise elimination than can be achieved with a single microphone. The microphone array can be one-dimensional (a linear array) or two-

dimensional (a circle, square, triangle or other suitable shape). The beam pattern of the array can be fixed or made adaptive though use of analog or digital phase shifting circuitry. The pattern of the active array is steered to point in the direction of the one or more users speaking. At the same time, nulls can be added to the pattern to notch out point or limited area noise sources. The use of the array microphone also helps reduce the cross talk between output from the text to speech engine 124 through the speaker 136 and detection of the user's speech.

[95] The microphone can be complemented with an analog or digital (i.e., Voice over IP) speech interface. This interface allows a remote user to connect to the system and interact with the system in the same manner possible if they were physically present.

[96] The speech unit 128, according to an embodiment of the invention, may use an analog or digital filter 132 between the array microphone 134 and the speech coder 138. The pass band of the filter is set to optimize the signal to noise ratio at the input to the speech recognition engine 120. In some embodiments, the filter is adaptive, using band shaping combined with notch filtering to reject narrow-band noise. In one embodiment of the invention, the system 90 may employ adaptive echo cancellation in the filter. The echo cancellation helps prevent cross talk between output from the text to speech engine and detection of the user's speech as well as suppression of environmentally caused echoes. Calculations are conducted for comparing the background noise to the signal received from the users speech are used to optimize the band-shaping parameters of the adaptive filter. Those skilled in the art will be familiar with multiple techniques used to construct suitable digital and analog filters.

[97] The speech received by the array microphone 134 and passed through the filter 132 may be sent to the speech digitizer or coder 138. The speech coder 138 may use adaptive lossy audio compression to optimize bandwidth requirements for the transmission of the coded speech to the speech recognition engine 120 over a wireless link. The lossy coding may be optimized to preserve only the components of the speech signal required for optimal recognition. Further, the lossy compression algorithms that may be used may be designed to prevent even momentary gaps in the signal stream, which can cause severe errors in the speech recognition engine. The digitized speech may be buffered in the coder and the coder may adapt the output data rate to optimize the use of the available bandwidth. The use of the adaptive speech coder is particularly advantageous when a band-limited wireless link is used between the coder and the speech recognition engine.

[98] In an alternative embodiment, the array microphone can be replaced by a set of physically distributed microphones or a headset worn by the user. The distributed microphones can be placed in different parts of a room or in different rooms. The distributed microphones can create a three-dimensional array to improve signal to noise ration. The headset may use a wireless or wired connection.

[99] While the invention is intended to be able to accept most any natural language question or command, ambiguity may be a problem. To assist users formulate concise questions and commands, the system 90 may support a voice query language. The language is structured to allow a variety of queries with minimal ambiguity. Thus, the voice query language helps users clearly specify the keywords or contexts of the question or command along with the parameters or criteria. The language provides a grammar to clearly specify the keyword used to determine the context and a set of one or criteria or parameters. A user asking a question or stating a command in the voice query language is nearly always guaranteed to receive a response.

[100] The voice query language may be sensitive to the contents of the context stack, wherein a context defines a set of questions that can be activated or deactivated during a conversation. According to one embodiment, each agent may designate one context to be the root context that defines base algorithms that the agent implements. Thus, following-on questions can be asked using an abbreviated grammar, since key words and criteria can be inherited from the context stack. For example, the user can simply asked about another keyword if the criteria of the question remain constant.

[101] The system 90, may provide built in training capabilities to help the user learn the best methods to formulate their questions and commands. The interactive training allows users to audibly or visibly see the machine interpretation of their queries and provides suggestions on how to better structure a query. Using the interactive training users can quickly become comfortable with the voice query language and at the same time learn how to optimize the amount of information required with each step of a dialog.

[102] The output of the speech coder 122 may be fed to the speech recognition engine 120. The speech recognition engine 120 recognizes words and phrases, using information in the context description grammar 112, and passes these to the parser 118 for interpretation. The speech recognition engine 120 may determine the user's identity by voice and name for each utterance. Recognized words and phrases may be tagged with this identity in all further processing. Thus, as multiple users engage in overlapping sessions, the tags added by the speech recognition engine 120 to each utterance allows other components of the system 90 to

tie that utterance to the correct user and dialog. The user recognition capability may further be used as a security measure for applications, such as auctions or online shopping, where this is required. Voice characteristics of each user may be contained in the user profile 110.

[103] According to one embodiment of the invention, users may start a dialog with the system 90 when they first address it. This can be done by speaking a generic word ("computer") or addressing a specific name ("Fred"), which may be generally tied to a system personality 108. Once the user starts the dialog, it may be recognized by the speech recognition engine 120, using unique characteristics of the user's speech. At the end of a dialog or to interrupt a dialog, the user may utter a dismissal word ("good bye").

[104] According to another embodiment of the invention, the system 90 may employ a speech recognition engine 124 that gains improved word recognition accuracy using data from context description grammar 112, user profiles 110, and the agents 106, among other components. At the same time, the fuzzy set possibilities or prior probabilities for the words in the context description grammar may be dynamically updated to maximize the probability of correct recognition at each stage of the dialog. The probabilities or possibilities may be dynamically updated based on a number of criteria including the application domain, the questions or commands, contexts, the user profile and preferences, user dialog history, the recognizer dictionary and phrase tables, and word spellings, among other criteria.

[105] For uncommon words or new vocabulary words, a user may be given the option to spell the words. The spelling may be done by saying the names or the letters or using a phonetic alphabet. The phonetic alphabet can be a default one or one of the user's choosing.

[106] Alternatively, when a user submits a word that is not recognized at all or is not correctly recognized by the speech recognition engine 120 then the user may be asked to spell the word. The speech recognition engine 120 determines this condition based on confidence level for the scoring process. The word may be looked up in a dictionary and the pronunciation for the word is added to either the dictionary, the agent 106, or the user's profile 110. The word pronunciation can then be associated with the domain, the question, the context and the user. Through this process, the speech recognition engine learns with time and improves in accuracy. To assist users in spelling words, an individualized phonetic alphabet can be used. Each user can modify the standard phonetic alphabets with words, which they can remember more easily.

[107] Once the words and/or phrases have been recognized by the speech recognition engine 120, the tokens and user identification is passed to the parser 118. The parser 118 examines the tokens for the questions or commands, context and criteria. The parser 118

determines a context for an utterance by applying prior probabilities or fuzzy possibilities to keyword matching, user profile 110, and dialog history. The context of a question or command determines the domain and thereby, the domain agent 156, if any, to be evoked. For example, a question with the keywords "temperature" implies a context value of weather for the question. The parser dynamically receives keyword and associated prior probability or fuzzy possibility updates from the system agent 150 or an already active domain agent 156. Based on these probabilities or possibilities the possible contexts are scored and the top one or few are used for further processing.

[108] The parser 118 uses a scoring system to determine the most likely context or domain for a user's question and/or command. The score is determined from weighing a number of factors including, the user profile 110, the domain agent's data content and previous context. Based on this scoring, the system 90 invokes the correct agent. If the confidence level of the score is not high enough to ensure a reliable response, the system 90 may ask the user to verify whether the question and/or command is correctly understood.

[109] In general, the question that is asked by the system 90 may be phrased to indicate the context of the question including all criteria or parameters. For example, the question can be in the form of: "Did I understand that you want such-and-such" If the user confirms that the question is correct the system proceeds to produce a response. Otherwise, the user can rephrase the original question, perhaps adding additional information to remove ambiguity, or the system can ask one or more questions to attempt to resolve the ambiguity.

[110] Once the context for the question or command has been determined, the parser 118 can invoke the correct agent 156, 150. To formulate a question or command in the regular grammar used by agents, the parser 118 may determine required and optional values for the criteria or parameters. These criteria may have been explicitly supplied by the user or may need to be inferred. The parser 118 may make use of the criteria handlers 152 supplied by the system agent 150. The criteria handlers 152 may provide context sensitive procedures for extracting the criteria or parameters from the user's questions or commands. Some criteria are determined by executing algorithms in the agent, while others may be determined by applying probabilistic of fuzzy reasoning to tables of possible values. Prior probabilities or fuzzy possibilities and associated values are received from a number of sources including the history of the dialog, the user profile 110, and the agent. Based on user responses, the prior probabilities or fuzzy possibilities are updated as the system learns the desired behavior. For a weather context, examples of criteria include, location, date and time. Other criteria can include command criteria (i.e., yes/no, on/off, pause, stop), and spelling. Special criteria

handlers are available from the system agent for processing lists, tables, barge-in commands, long strings of text and system commands.

[111] The criteria handlers 152 operate iteratively or recursively on the criteria extracted to eliminate ambiguity. This processing helps reduce the ambiguity in the user's question or command. For example, if the user has a place name (or other proper noun) in their utterance, the parser 118 can use services of the domain agent 156 to look up tables in the databases 102 for place names or can attempt to determine which word is the proper noun from the syntax of the utterance. In another example, the user asks "what about flight one hundred and twenty too?". The parser and domain agent use flight information in the database and network information along with context to determine the most plausible interpretation among; flight 100 and flight 20 also, flight 100 and flight 22, flight 122, and the like.

[112] Once the context and the criteria are determined, the parser 118 may form the question or command in a standard format or hierarchical data structure used for processing by the agents 150, 156. The parser 118 may fill in all required and some optional tokens for the grammar of the context. Often the tokens must be transformed to values and forms acceptable to the agents. The parser obtains the required transformations from the agents, dialog history or user profile 110. Examples of transformations or substitutions performed by the parser on tokens include:

- substituting a stock symbol for a company name or abbreviation;
- substituting a numerical value for a word or words;
- adding a zip code to an address; and
- changing a place or other name to a commonly used standard abbreviation.

[113] The agents 150, 156 may receive a command or question once the parser 118 has placed it in the required standard format. Based on the context, the parser 118 evokes the correct agent to process the question or command.

[114] Commands can be directed to the system 90 or to an external entity. System commands are generally directed to the system agent 150. Commands for external entities are generally processed by a domain agent 156, which includes the command context and behavior for the external entity.

[115] Specific questions are generally directed to one of the domain agents 156. Based on the question or context and the parameters or criteria, the domain agent creates one or more queries to one or more local or external information sources. Questions can be objective or

subjective in nature. Results for objective questions can often be obtained by structured queries to one or more local or network information sources. Even for objective questions, the system 90 may need to apply probabilistic or fuzzy set analysis to deal with cases of conflicting information or incomplete information. Information to answer subjective questions is generally obtained by one or more ad-hoc queries to local or network data sources, followed by probabilistic or fuzzy set evaluation of the one results to determine a best answer.

[116] Once the domain agent 156 has formulated the one or more queries, they may be sent to local and/or network information sources. The queries are performed in an asynchronous manner to account for the fact that sources respond at different speeds or may fail to respond at all. Duplicate queries are sent to different information sources to ensure that at least one source responds with a useful result in a timely manner. Further, if multiple results are received in a timely manner, they can be scored by the system to determine which data is most reliable or appropriate. Examples of data sources accommodated include, HTTP data sources, sources with meta-data in various formats including XML, entertainment audio, video and game files including MP3, databases using query languages and structured responses such as SQL, and other data sources.

[117] The local information sources can be stored in one or more system databases 102 or can be on any local data storage such as a set of CDs or DVDs in a player or other local data storage. Network information sources can be connected to the Internet 136 or other network and accessed through a series of plug-ins or adaptors, known a pluggable sources, in the network interface 116. The pluggable sources in the network interface 116 may be capable of executing the protocols and interpreting the data formats for the data sources of interest. The pluggable sources may provide information, scrapping data and procedures for each source to the domain agents 156. If a new type of data source is to be used, a new plug-in or adaptor can be added to the network interface 116.

[118] The domain agent 156 evaluates the results of the one or more queries as they arrive. The domain agent 156 scores the relevance of the results based on results already received, the context, the criteria, the history of the dialog, the user profile 110 and domain specific information using probabilistic or fuzzy scoring techniques. Part of the dialog history is maintained in a context stack. The weight of each context for the scoring is based on the relevance of one context to another and the age of the contexts. Other scoring variables can be associated through the context stack. Contexts can also be exclusive, so that previous contexts have no weight in the scoring.

[119] Based on the continuous scoring processes, the domain agent 156 may determine if a single best answer can be extracted. For most questions, the desired result has a set of tokens that must be found to formulate an answer. Once a value has been found for each of these tokens, the results are ready for presentation to the user. For example, for a question on weather, the tokens can include the date, day of week, predicted high temperature, predicted low temperature, chance of precipitation, expected cloud cover, expected type of precipitation and other tokens. Results processed in this manner include error messages. For subjective questions, this determination is made by determining a most likely answer or answers, extracted by matching of the results received. If no satisfactory answer can be inferred from the results of the query, the agent can do one of the following:

1. Ask the user for more information, typically through the speech interface, and based on the results obtained formulate new queries. This approach is applied when an irresolvable ambiguity arises in the formulation of a response.
2. Formulate new queries based on the results received from the first set of queries. This approach is typically applied in cases where the responses received do not contain all the required information. Information sources to query can inferred from the results already obtained (i.e., links in an HTML document) or from other sources. Using this approach one or more sets of queries and responses can be chained without the need for action by the user.
3. Wait for additional queries to return results.

[120] In any case, the domain agent 156 may continue to make requests and evaluate results until a satisfactory response is constructed. In doing so, the agent can start several overlapping query paths or threads of inquiry, typically mediated by the event manager 100. This technique, combined with the use of asynchronous queries from multiple data sources, provides the real-time response performance required for a natural interaction with the user.

[121] The domain agent 156 may apply conditional scraping operations to each query response as it is received. The conditional scraping actions depend on the context, the criteria, user profile 110, and domain agent coding and data. For each token to be extracted, a scraping criteria 152 may be created using the services of the system agent 150. The scraping criteria use format specific scraping methods including, tables, lists, text, and other scrapping methods. One or more scraping criteria can be applied to a page or results set. Once additional results are received, the domain agent 156 can create new scraping criteria to apply to results already acquired. The conditional scarping process removes extraneous

information, such as graphics, which need not be further processed or stored, improving system performance.

[122] The domain agent 156 may strongly influence how the results are presented. For instance, once the domain agent 156 has created a satisfactory response to a question, the agent 156 may format that response for presentation. Typically, the domain agent 156 formats the response into the markup format used by the text to speech engine 124. The domain agent 156 may also format the result presentation using available format templates and based on the context, the criteria, and the user profile 110. Variable substitutions and transformations may be performed by the agent 156 to produce a response best understood and most natural to the user. Further, the order of presentation of tokens and the exact terminology used to create a more natural response to the user may be varied by the agent 156. The domain agent 156 may also select the presentation personality 108 to be used.

[123] The domain agent 156 may select the presentation template, determine order of presentation for tokens and determine variable substitutions and transformations using probabilistic or fuzzy set decision methods. The template used to form the presentation can be from the domain agent itself or from the user profile 110. The user profile 110 can completely specify the presentation format or can be used to select and then modify an existing presentation format. Selection and formatting of presentation templates can also depend on the presentation personality modules 108. At the same time, the characteristics of the personality used for the response are dynamically determined using probabilities or fuzzy possibilities derived from the context, the criteria, the domain agent itself and the user profile 110.

[124] The domain agent 156 may apply a number of transformations to the tokens before presentation to the user. These variable substitutions and transformations are derived from a number of sources including, domain information carried by the agent, the context, the token values, the criteria, the personality module 108 to be used, and the user profile 110. Examples of variable substitutions and transformations include:

- substitution of words for numbers;
- substitution of names for acronyms or symbols (i.e., trading symbols);
- use of formatting information derived from the information sources (i.e., HTML tags);
- nature of the response including, text, long text, list, table;
- possible missing information or errors;

- units for measurement (i.e., English or metric); and
- preferred terminology from the user profile 110 or presentation personality 108.

[125] The system 90, according to another embodiment of the invention, may provide special purpose presentation capabilities for long text strings, tables, lists and other large results sets. Domain agents 156 may use special formatting templates for such results. The system agent 150 may provide special criteria handlers 152 for presentation and user commands for large results sets. The presentation templates used by the domain agents 156 for large results sets typically include methods for summarizing the results and then allowing the user to query the result in more detail. For example, initially only short summaries such as headlines or key numbers are presented. The user can then query the results set further. The criteria handlers 152 provide users with the capability to browse large results sets. Commands provided by the criteria handlers 152 for large results sets include, stop, pause, skip, rewind, start, and forward.

[126] Some information, in formats such as video, pictures and graphics, may be best presented in a displayed format. The domain agents 156 may apply suitable presentation templates in these cases and present the information through the graphical user interface 114. The system agent 150 provides special criteria handlers 152 for presentation and user commands for display presentation and control.

[127] FIG. 3 illustrates a high level process 300 for receiving natural language speech-based queries and/or commands and generating a response according to an embodiment of the invention. The process 300 comprises of individual steps, several of which may be repetitive in order to accommodate partial failures. The process 300 may begin when a user's utterance is received and processed at 302. The user utterance may be a query and/or a command and may be of a natural language speech-based form. After receiving the natural language query and/or command, the query and/or command may be processed so that the data is in a more machine understandable format such as in a digital format. Once the utterance has been converted into an understandable format, the utterance may be processed so that at least an approximate meaning of the utterance can be determined at 304. This may be accomplished by, for example, the use of stored dictionaries, phrases, user profiles, domain agent data, and other data. Based on the approximate meaning of the utterance, one or more correctly formatted queries and/or commands may be generated.

[128] A user created query and/or command may require that multiple requests and/or commands may be needed in order to generate the desired response or action. Further, depending upon the context in which the query and/or command has been submitted (e.g., who is the user, what is the subject the query and/or command directed to, at what time or location is the query and/or command being directed to, and other parameters used in defining the query and/or command), a specific format for the query and/or command may be required. For instance, each of the domain agents 156 may require that requests and/or commands be formatted in a specific manner. Thus at 306, one or more properly formatted queries and/or commands may be generated. Operations 304 and 306 may be repeated in order to obtain the correct interpretation of the user utterance and result in the desired response and/or action. Once, the properly formatted queries and/or commands are generated, the appropriate action may be taken at 308. This may require that the properly formatted queries and/or commands be sent to specific domain agent[s], information source[s], device[s], or other appropriate destination[s] that can fulfill the requirements of the query and/or command. Once the appropriate action[s] have been executed, this event may be recorded to, for example, the user's profile, database and/or one or more agents at 310. Such data may be useful for future user inquires and commands. After the action[s] have been performed, a response, if need be, may be generated and forwarded to the user and/or third parties at 312.

[129] In the case of a query for retrieving data, the response would contain the requested information. In the case of a command, the response may be a confirmation that a specific action[s] has been executed. The response may be in the form of a natural language format. The response may also be formatted to reflect a particular personality or tone to the response in order to make the response more "human." The response may relayed to the user and/or third parties as an audio message and/or a visual message displayed on a user interface.

[130] FIG. 4A illustrates an exemplary process 400 for receiving natural language speech-based requests and/or commands and generating a response using the system 90 according to another embodiment of the invention. The process 400 may begin when a user's utterance (i.e., user query and/or command) is captured using speech recognition that operates in a variety of real-world environments at 402. Once the utterance is captured, it is parsed and interpreted to determine the query and/or command that is contained in the utterance at 404. According to one embodiment of the invention, the utterance may be transcribed to text and analyzed using the general cognitive model 806, the environmental model 808, and/or the personalized cognitive model 810. Next, the request and/or command are reviewed to

determine the domain of expertise required and the context of the query, invoking the proper resources including, for example, agents at 406. The original query and/or command submitted by the user may require that multiple queries and/or commands be generated.

[131] For instance, suppose a user is interested in retrieving the value of her stock portfolio. The user may utter "please get the value of my stock portfolio." The system 90 may review this request together with stored data such as the user's profile, environmental model 808, and/or personalized cognitive model 810, to determine keywords such as "get the value" and "my stock portfolio." The system 90 may then generate requests to determine the stocks in the user's portfolio, the number of shares and the source for current pricing information. These queries may then be sent to one or more agent domains, such as a domain which may access a database containing the user's profile and a domain which accesses stock pricing sources to determine the answers to these questions.

[132] Requests may be sent to these domain agents or sources in order to obtain the desired data. Thus, at operation 408, one or more requests may be formulated and sent to one or more local and/or network data sources and/or appropriate commands may be sent to local or remote devices or the system itself. The requests are then sent to the designated agent[s]. The agents may then, in turn, generate their own requests and/or commands to be sent to, for example, local or remote information sources to retrieve needed data. The agent generated requests and/or commands may be formatted according to the requirements of the target sources and variable substitutions and transformations are performed to modify the requests to a form most likely to yield desired results from the available sources at 410. Once the requests are formatted correctly, they may be executed in an asynchronous manner and dealing gracefully with failures at 412. As a result of the execution of the requests, results may be returned by the domain agents and/or sources. The system 90 may then extract or scrape the desired information from the one or more results, which may be returned in any one of a number of different formats at 414. That is, the results sought by a user may be the summary or the results of further processing of information obtained from several sources for example.

[133] Next, the results may be evaluated and interpreted including processing of errors, and gathering and combining them into a single best result judged to be "best" even if the results are ambiguous, incomplete, or conflicting at 416. Once the best results are determined, any required formatting is performed. At operation 418, variable substitutions and transformations may be used to modify the results. Finally, at operation 420, the compound results may be presented to the user in a useful and expected manner through the text to

speech engine 124. The process 400 may be performed while accounting for the domain of expertise required, the context in which the question or command is presented, the domain specific information available, the history of the user's interaction, the user preferences, the information sources or commands that are available, and responses obtained from the sources.

[134] At each stage of process 400, probabilistic or fuzzy set decision and matching methods may be applied to deal with inconsistent, ambiguous, conflicting and incomplete information or responses. In addition, the use of asynchronous queries that may result in rapid and graceful failure of some queries or commands may allow the system 90 to robustly return results quickly and in a manner that seems natural to the user.

[135] FIG. 4B illustrates an exemplary process 450 for receiving natural language speech-based commands in order to locally or remotely control functions of the system 90, or for other devices, according to another embodiment of the invention. The process 450 may begin when a user's utterance (i.e., user query and/or command) is captured using speech recognition systems that operate in a variety of real-world environments at 452. Once the utterance is captured, it is parsed and interpreted to determine the command that is contained in the utterance at operation 454.

[136] Next, at operation 456, the command is reviewed to determine the domain for the command and context, invoking the proper resources including agents. The original command submitted by the user will often require that multiple queries and/or commands be generated. For instance, suppose a user is interested in recording his favorite television program. The user may utter "please record my favorite TV program." The system 90 may review this request together with stored data such as the user's profile and determine keywords such as "record" and "my favorite TV program." The system 90 may then generate queries to determine the name, the channel, and time for the user's favorite TV program. These queries may then be sent to an agent domain, which may access a database containing the user's profile and determine the answers to these questions. A command may then be sent to a video recorder ordering the video recorder to record the selected TV program. Thus, at operation 458, one or more queries may be formulated that are to be sent to one or more local and/or network data sources. Appropriate commands may be sent to local or remote devices or the system itself. Next, at operation 460, the generated commands may be routed to the appropriate system[s] and/or external devices. Once the commands have been executed, results including errors may be received and processed at 462. The results of the executed commands may be optionally presented to the user at operation 464.

[137] Some of the operations depicting in FIG. 4A may be performed using multiple steps that may be repetitive. For instance, in operation 404, the user's utterance is parsed and interpreted in order to determine the meaning of the utterance. System 90 may make an initial interpretation of the utterance based the data stored in, for example, the context description grammar module 112, user profiles, 110, agents 106 and the databases 104. A real-time scoring system or other techniques may be applied to the interpretation results in order to generate a domain or a context score. If the confidence level of the domain or context score is not high enough to ensure a reliable response, the system 90 can request that the user verify the question or command is correctly understood. In general, the question may be phrased to indicate the context of the question including all criteria or parameters. If the user confirms that the question is correct, the system 90 may proceed to produce a response. Otherwise, either the user can rephrase the original question, perhaps adding additional information to remove ambiguity, or the system may ask one or more questions to attempt to resolve the ambiguity or other actions may taken.

[138] FIG. 5 illustrates a process 500 for correctly interpreting a user's utterance according to one embodiment of the invention. The process 500 generally represents operations 402 and 404 of process 400. Initially the user's utterance is inputted into the system 90 at operation 502. The input is then interpreted using, for example, various stored data such as user profiles, agent data, dictionary and phrases, and other relevant data. The interpretation is then scored using, for instance, a weighted scoring system as described previously at 506. Once a score is obtained, the confidence level of the interpretation may be analyzed to determine if it is satisfactory at operation 508. In other words, a determination is made whether the score given to the interpretation exceeds a certain value. If the confidence level is determined to be unsatisfactory, then a request may be submitted to the user requesting that the user verify the interpretation at 510 and 512. If the user is unsatisfied with the interpretation, he/she may be asked to rephrase the utterance and/or provide additional information at 514. Once the user provides the rephrased utterance and/or additional information, the process 500 returns to the beginning at operation 502. If, on the other hand, the user is satisfied with the interpretation, then the process 400 for receiving natural language speech-based requests and/or commands, and generating a response, may continue at operation 516. Once the meaning of the utterance has been determined, the appropriate domain agent[s] and the query and/or commands may be properly formatted for the agent[s].

[139] FIG. 6 illustrates a process 600 for determining the proper domain agent[s] to invoke and the proper format for queries and/or commands that are submitted to the agents 106, as

generally depicted in steps 406-408 of FIG. 4A. In order to formulate a question or command in the regular grammar used by agents 106, a determination may be made as to the required and optional values for the criteria or parameters of the queries and/or commands. These criteria may have been explicitly supplied by the user or may need to be inferred. Thus, the properly interpreted query and/or command that is generated by, for example, operations 404 and 454 of FIGS. 4A and 4B, is parsed at operation 602.

[140] The content of the properly interpreted request and/or command is then analyzed to determine the criteria and/or parameters of the properly interpreted request and/or command at 604. A context sensitive procedure for extracting the criteria or parameters from the properly interpreted request and/or command may be used. Some criteria are determined by executing algorithms in the agent, while others may be determined by applying probabilistic of fuzzy reasoning to tables of possible values. Prior probabilities or fuzzy possibilities and associated values are received from a number of sources including the history of the dialog, the user profile 110, and the agent.

[141] Based on user responses, the prior probabilities or fuzzy possibilities are updated as the system 90 learns the desired behavior. For a weather context, examples of criteria include, location, date and time. Other criteria can include command criteria (i.e., yes/no, on/off, pause, stop), and spelling. The process for determining criteria may be iterative or recursive in order to eliminate ambiguity in the user's question or command. For example, if the user has a place name (or other proper noun) in their utterance, tables in the databases 102 may be reviewed for place names or an attempt may be made to determine which word is the proper noun from the syntax of the utterance. In another example, the user asks "what about flight one hundred and twenty too?" In such a situation, flight information in the database and network information along with context may be used to determine the most plausible interpretation among; flight 100 and flight 20 also, flight 100 and flight 22, flight 122, and the like. Once the parameters and criteria for the query and/or command have been established, the proper agents 106 may be selected at 606.

[142] After selecting the agents 106, at operation 608, a properly formatted query and/or command that will be submitted to the agents 106 may be determined. The query and/or command may be in a standard format or may be a hierarchical data structure used for processing by the agent 106. In order to submit the properly formatted query and/or command to the agent 106, all of the required, and some optional tokens, for the grammar of the context may be filled in. Often the tokens must be transformed to values and forms acceptable to the agents. The required transformations may be obtained from the agents,

dialog history or user profile 110. Examples of transformations or substitutions that may be performed were provided above. Once the properly formatted queries and/or commands are generated, the process (e.g., process 400) may continue at 610.

[143] While specific embodiments of the invention have been described above, it will be appreciated that the invention may be practiced otherwise than as described. For example, the invention may take the form of a computer program containing one or more sequences of machine-readable instructions describing a method as disclosed above, or a data storage medium (e.g. semiconductor memory, magnetic or optical disk) having such a computer program stored therein.

[144] The descriptions above are intended to be illustrative, not limiting. Thus, it will be apparent to one skilled in the art that modifications may be made to the invention as described without departing from the scope of the claims set out below.

We claim:

1. A system for processing speech and non-speech communications, comprising:

a terminal device that receives the speech and the non-speech communications;

a transcription module that transcribes the speech and the non-speech communications to create a speech-based textual message and a non-speech-based textual message;

a merging module that merges the speech-based textual message and the non-speech-based textual message to generate a query;

an analyzer that accesses information associated with the query to enhance query recognition; and

an adaptive misrecognition analysis engine that monitors actions corresponding to the query and classifies the query as recognized, not recognized or incorrectly recognized.

2. The system according to claim 1, wherein the analyzer includes (i) a personalized cognitive model, (ii) a general cognitive model, (iii) an environmental model, or any combination of (i) to (iii).

3. The system according to claim 2, wherein the personalized cognitive model is derived from an interaction pattern between a user and the system.

4. The system according to claim 3, wherein the personalized cognitive model accesses a user profile and predicts user actions.

5. The system according to claim 3, wherein the personalized cognitive model tracks actions performed by a plurality users.

6. The system according to claim 2, wherein the generalized cognitive model is a statistical abstract that corresponds to interactions patterns between a plurality of users and the system.

7. The system according to claim 6, wherein the generalized cognitive model predicts actions to be performed by the plurality of users.

8. The system according to claim 6, wherein the generalized cognitive model tracks actions performed by the plurality of users.

9. The system according to claim 2, wherein the environmental model includes information associated with user environmental conditions or surroundings.

10. The system according to claim 9, wherein user environmental conditions include (i) a user global position, (ii) user movement information, (iii) noisy conditions, (iv) quiet conditions, (v) vicinity to voice-enabled devices or any combination of (i) to (v).

11. The system according to claim 9, wherein the environmental model provides (i) context, (ii) domain knowledge, (iii) preferences or any combination of (i) to (iii) in order to enhance an interpretation of the query.

12. The system according to claim 1, further comprising:

a search module that searches the query for text combinations;

a comparison module that compares the text combinations to entries in a context description grammar;

a plurality of domain agents that are associated with the context description grammar;

a scoring module that provides relevance scores based on results from the comparison module;

a domain agent selector that selects domain agents based on results from the scoring module; and

a response generating module that communicates with the selected domain agents to obtain content that is gathered by the selected domain agents and that generates a response from the content, wherein the content is arranged in a selected order based on results from the scoring module.

13. The system according to claim 12, wherein the response generating module generates an aggregate response that includes the content gathered by the selected domain agents.

14. The system according to claim 1, wherein the terminal device includes (i) a personal digital assistant, (ii) a cellular telephone, (iii) a portable computer, (iv) a desktop computer or any combination of (i) to (iv).

15. The system according to claim 1, wherein the terminal device receives follow-up speech and non-speech communications and wherein the transcription module transcribes the follow-up speech and non-speech communications to create a follow-up speech-based textual message and a follow-up non-speech-based textual message.

16. The system according to claim 15, wherein the merging module merges the follow-up speech-based textual message and the follow-up non-speech-based textual message to generate a follow-up query.

17. The system according to claim 15, wherein the adaptive misrecognition analysis engine monitors actions performed by users including issuing the follow-up speech and non-speech communications that are presented proximate in time to the speech and non-speech communications.

18. The system according to claim 1, further comprising a tuning module that determines how the analyzer incorrectly recognizes or does not recognize the query and tunes parameters to recognize the query in a subsequent speech or non-speech communication.

19. The system according to claim 12, further comprising a personality module that facilitates formatting the response.

20. The system according to claim 12, further comprising a context stack that includes one or more contexts that are selected based on the query.

21. The system according to claim 12, wherein the scoring module determines the one or more contexts based on at least applying prior probabilities or fuzzy possibilities to (i) keyword matching, (ii) user profiles, (iii) a dialog history, or any combination of (i) to (iii).

22. The system according to claim 12, wherein at least one of the domain agents creates and directs a request to at least one of a local information source and a network information source.

23. The system according to claim 12, wherein at least one of the domain agents creates and directs a command to a remote or local device.

24. The system according to claim 16, wherein at least one of the domain agents evaluates multiple queries from multiple sources.

25. The system according to claim 16, wherein the follow-up query is associated with a same context as the query.

26. A method of processing speech and non-speech communications, comprising:

receiving the speech and non-speech communications;

transcribing the speech and non-speech communications to create a speech-based textual message and a non-speech-based textual message;

merging the speech-based textual message and the non-speech-based textual message to generate a query;

accessing information associated with the query to enhance query recognition;

monitoring actions corresponding to the query; and

classifying the query as recognized, unrecognized or incorrectly recognized.

27. The method according to claim 26, wherein accessing information associated with the query includes accessing (i) a personalized cognitive model, (ii) a general cognitive model, (iii) an environmental model, or any combination of (i) to (iii).

28. The method according to claim 26, wherein the personalized cognitive model is derived from an interaction pattern between a user and the system.

29. The method according to claim 28, wherein the personalized cognitive model accesses a user profile and predicts user actions.

30. The method according to claim 28, wherein the personalized cognitive model tracks user actions.

31. The method according to claim 27, wherein the generalized cognitive model is a statistical abstract that corresponds to interactions patterns between a plurality of users and the system.

32. The method according to claim 31, wherein the generalized cognitive model predicts actions performed by the plurality of users.

33. The method according to claim 31, wherein the generalized cognitive model tracks actions performed by the plurality of users.

34. The method according to claim 27, wherein the environmental model includes information associated with user environmental conditions or surroundings.

35. The method according to claim 34, wherein user environmental conditions include (i) a user global position, (ii) user movement information, (iii) noisy conditions, (iv) quiet conditions, (v) vicinity to voice-enabled devices or any combination of (i) to (v).

36. The method according to claim 34, wherein the environmental model provides (i) context, (ii) domain knowledge, (iii) preferences or any combination of (i) to (iii) in order to enhance an interpretation of the query.

37. The method according to claim 26, further comprising:

searching the query for text combinations;

comparing the text combinations to entries in a context description grammar;

accessing a plurality of domain agents that are associated with the context description grammar;

generating a relevance score based on results from comparing the text combinations to entries in the context description grammar;

selecting one or more domain agents based on results from the relevance score;

obtaining content that is gathered by the selected domain agents; and

generating a response from the content, wherein the content is arranged in a selected order based on results from the relevance score.

38. The method according to claim 37, further comprising generating an aggregate response that includes the content that is gathered by the selected domain agents.

39. The method according to claim 26, further comprising:

receiving a follow-up speech and non-speech communications;

transcribing the follow-up speech and non-speech communications to create a follow-up speech-based textual message and a follow-up non-speech-based textual message; and

merging the follow-up speech-based textual message and the follow-up non-speech-based textual message to generate a follow-up query.

40. The method according to claim 39, wherein monitoring actions includes monitoring the follow-up speech and non-speech communications that are presented proximate in time to the speech and non-speech communications.

41. The method according to claim 40, further comprising determining how the query is incorrectly recognized or not recognized and tuning parameters to recognize the query in a subsequent speech or non-speech communication.

42. The method according to claim 37, further comprising a personality module that communicates the response to a user.

43. The method according to claim 37, further comprising generating a context stack that includes one or more contexts that are selected based on the query.

44. The method according to claim 43, wherein the one or more contexts are generated based on applying prior probabilities or fuzzy possibilities to (i) keyword matching, (ii) user profiles, (iii) a dialog history, or any combination of (i) to (iii).

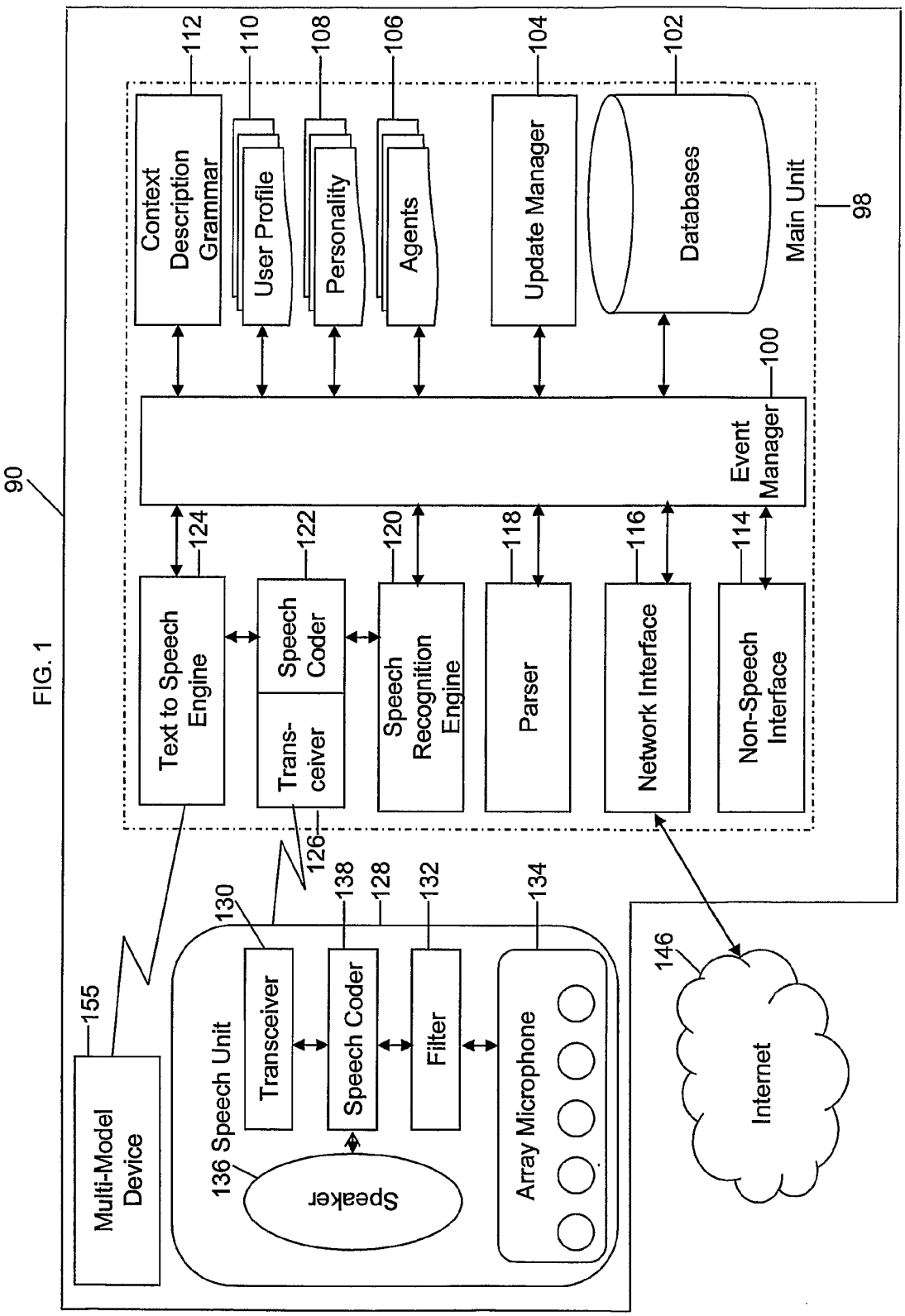

90
Main Unit
98
Context Description Grammar
User Profile
Personality
Agents
Update Manager
Databases
Event Manager
100
Text to Speech Engine
Speech Coder
Trans-ceiver
Speech Recognition Engine
Parser
Network Interface
Non-Speech Interface
102
104
106
108
110
112
114
116
118
120
122
124
126
128
130
132
134
136 Speech Unit
Transceiver
Speech Coder
Filter
Array Microphone
Speaker
138
146
Internet
155
Multi-Model Device

FIG. 1

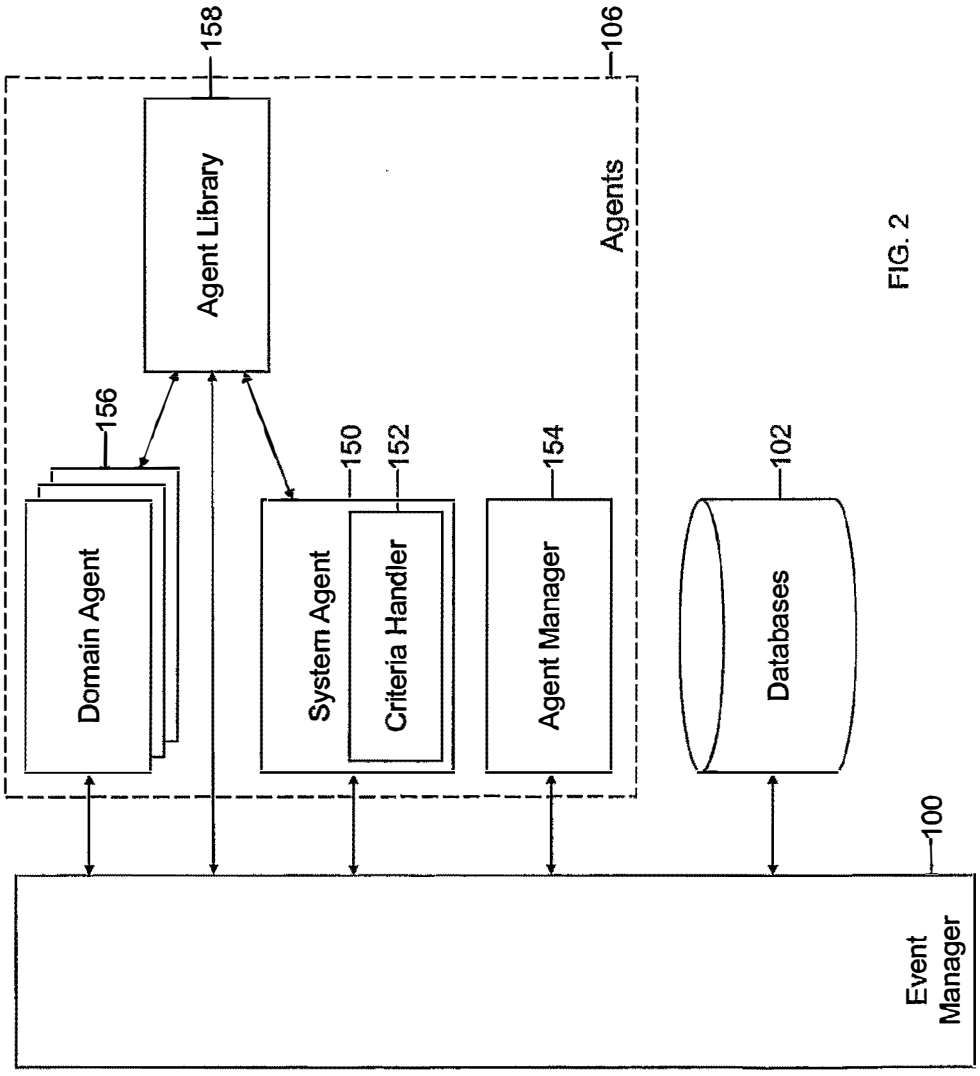

158
106
Agent Library
Agents
156
150
152
154
102
Domain Agent
System Agent
Criteria Handler
Agent Manager
Databases
100
Event Manager

FIG. 2

FIG. 3

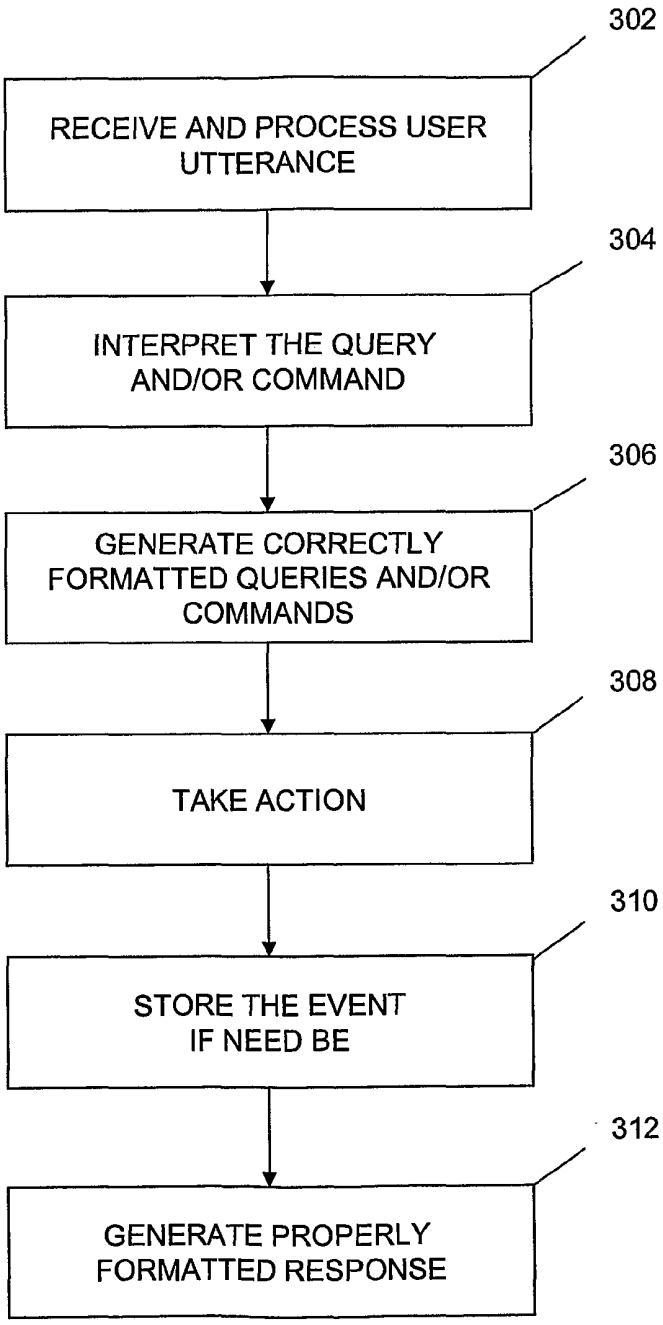

300
302
RECEIVE AND PROCESS USER UTTERANCE
304
INTERPRET THE QUERY AND/OR COMMAND
306
GENERATE CORRECTLY FORMATTED QUERIES AND/OR COMMANDS
308
TAKE ACTION
310
STORE THE EVENT IF NEED BE
312
GENERATE PROPERLY FORMATTED RESPONSE

FIG. 4A

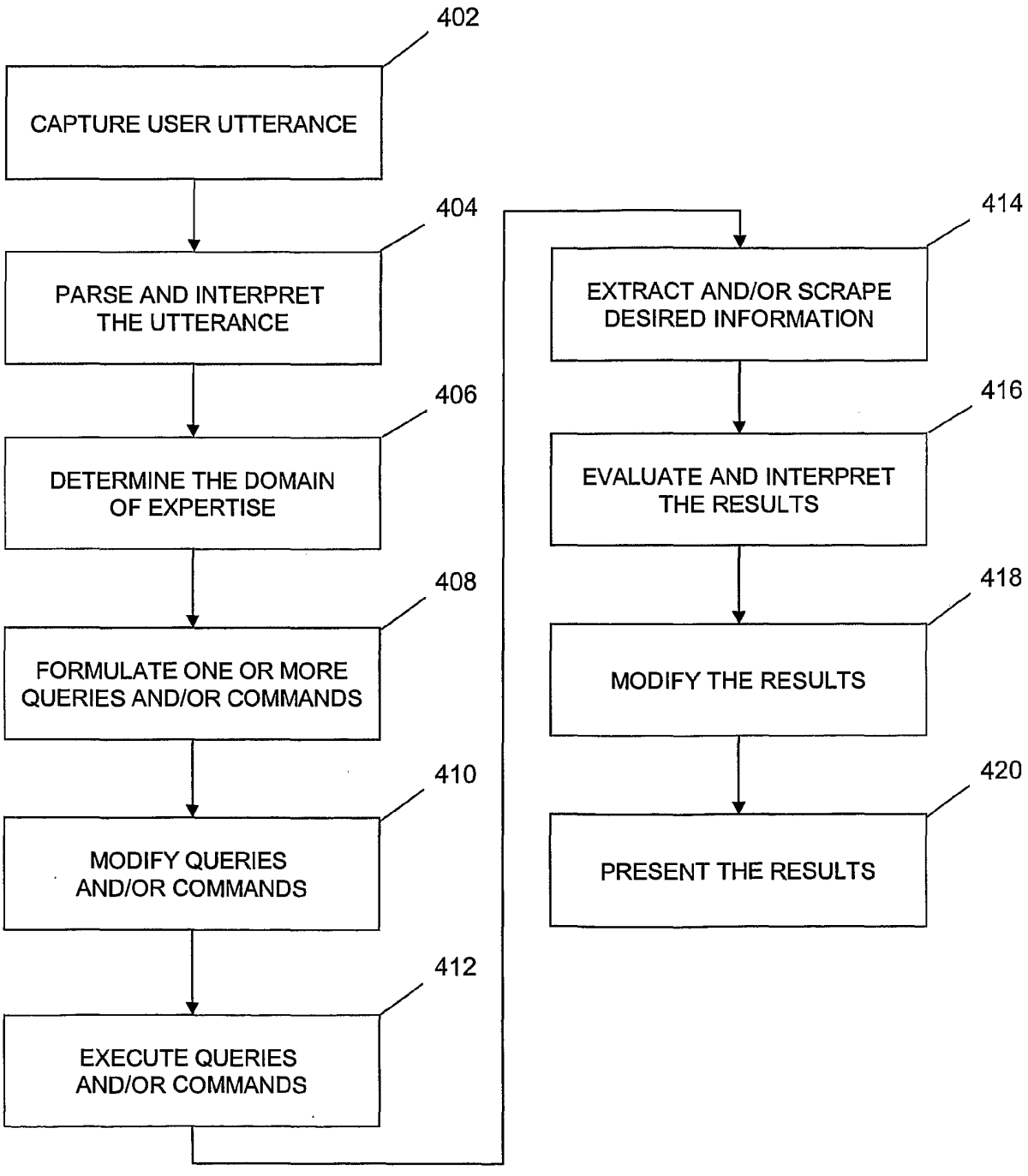

400
402
CAPTURE USER UTTERANCE
404
PARSE AND INTERPRET THE UTTERANCE
406
DETERMINE THE DOMAIN OF EXPERTISE
408
FORMULATE ONE OR MORE QUERIES AND/OR COMMANDS
410
MODIFY QUERIES AND/OR COMMANDS
412
EXECUTE QUERIES AND/OR COMMANDS
414
EXTRACT AND/OR SCRAPE DESIRED INFORMATION
416
EVALUATE AND INTERPRET THE RESULTS
418
MODIFY THE RESULTS
420
PRESENT THE RESULTS

FIG. 4B

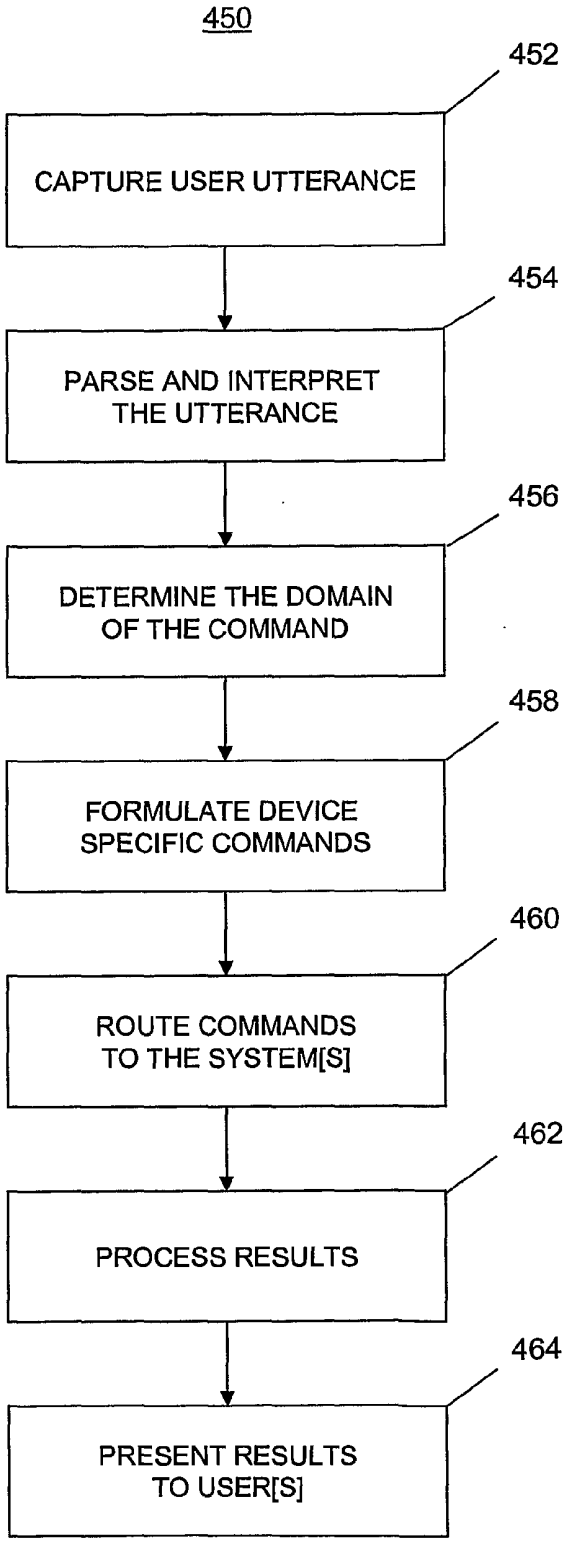

450
452
CAPTURE USER UTTERANCE
454
PARSE AND INTERPRET THE UTTERANCE
456
DETERMINE THE DOMAIN OF THE COMMAND
458
FORMULATE DEVICE SPECIFIC COMMANDS
460
ROUTE COMMANDS TO THE SYSTEM[S]
462
PROCESS RESULTS
464
PRESENT RESULTS TO USER[S]

FIG. 5

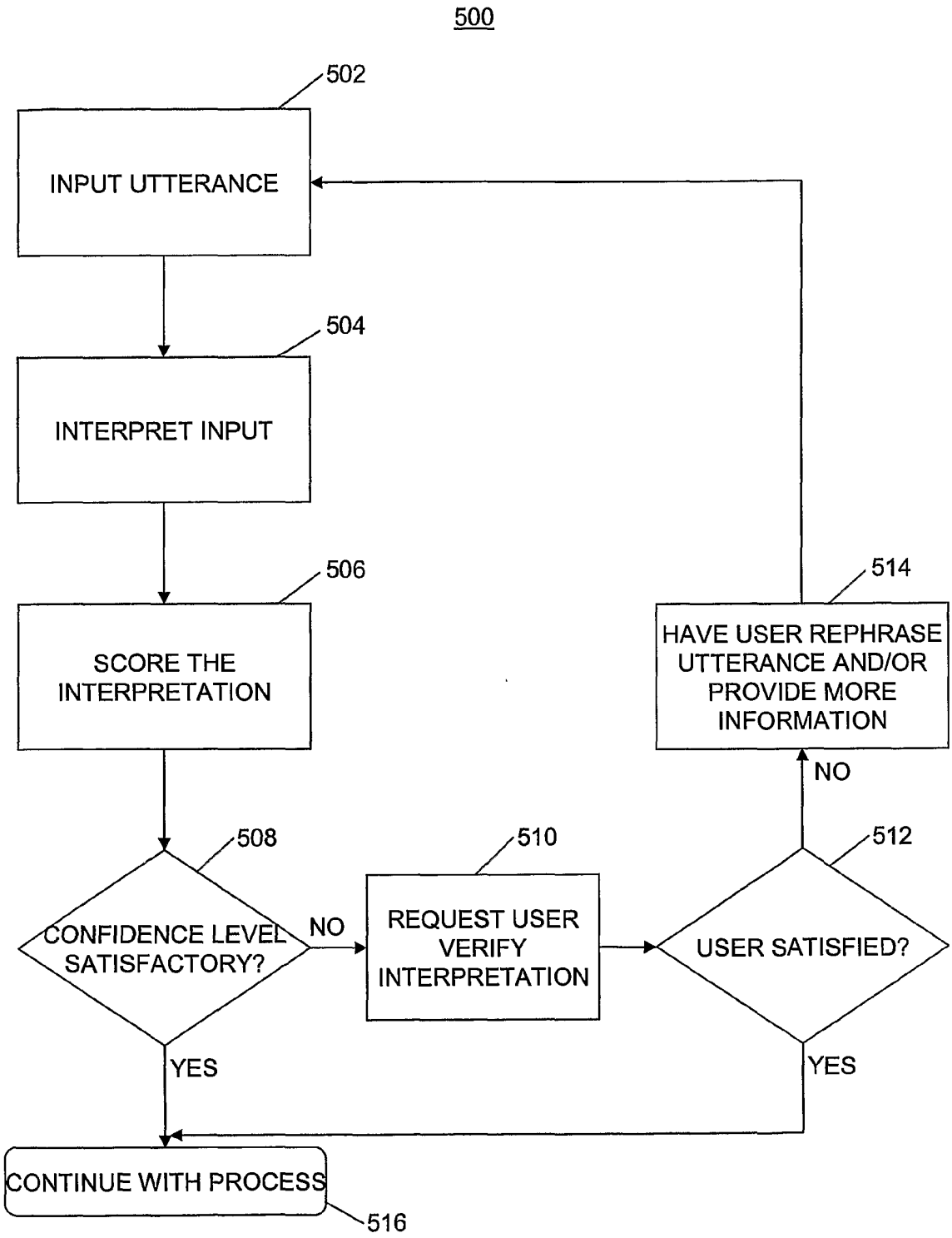

500
502
INPUT UTTERANCE
504
INTERPRET INPUT
506
SCORE THE INTERPRETATION
508
CONFIDENCE LEVEL SATISFACTORY?
NO
YES
510
REQUEST USER VERIFY INTERPRETATION
512
USER SATISFIED?
NO
YES
514
HAVE USER REPHRASE UTTERANCE AND/OR PROVIDE MORE INFORMATION
CONTINUE WITH PROCESS
516

FIG. 6

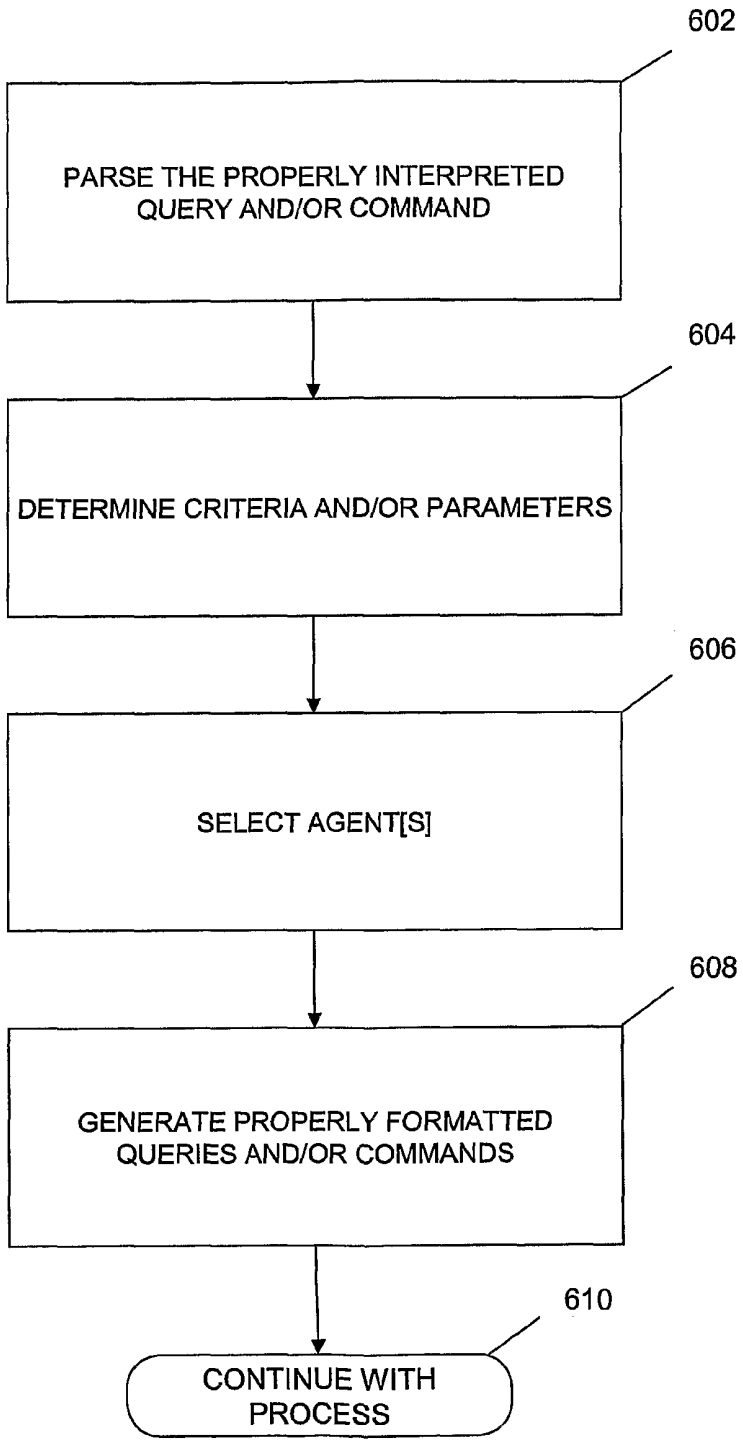
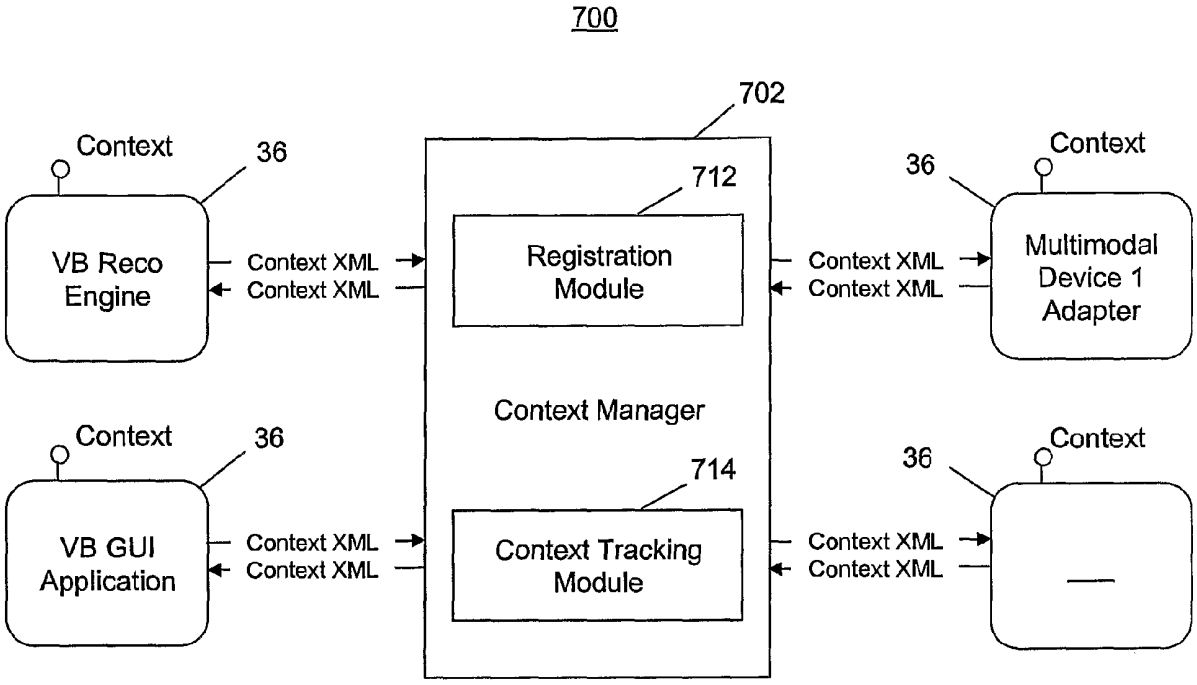

600
602
PARSE THE PROPERLY INTERPRETED QUERY AND/OR COMMAND
604
DETERMINE CRITERIA AND/OR PARAMETERS
606
SELECT AGENT[S]
608
GENERATE PROPERLY FORMATTED QUERIES AND/OR COMMANDS
610
CONTINUE WITH PROCESS

FIG. 7


700
702
Context
36
VB Reco Engine
Context XML
Context XML
712
Registration Module
Context XML
Context XML
36
Context
Multimodal Device 1 Adapter
Context Manager
Context
36
VB GUI Application
Context XML
Context XML
714
Context Tracking Module
Context XML
Context XML
36
Context

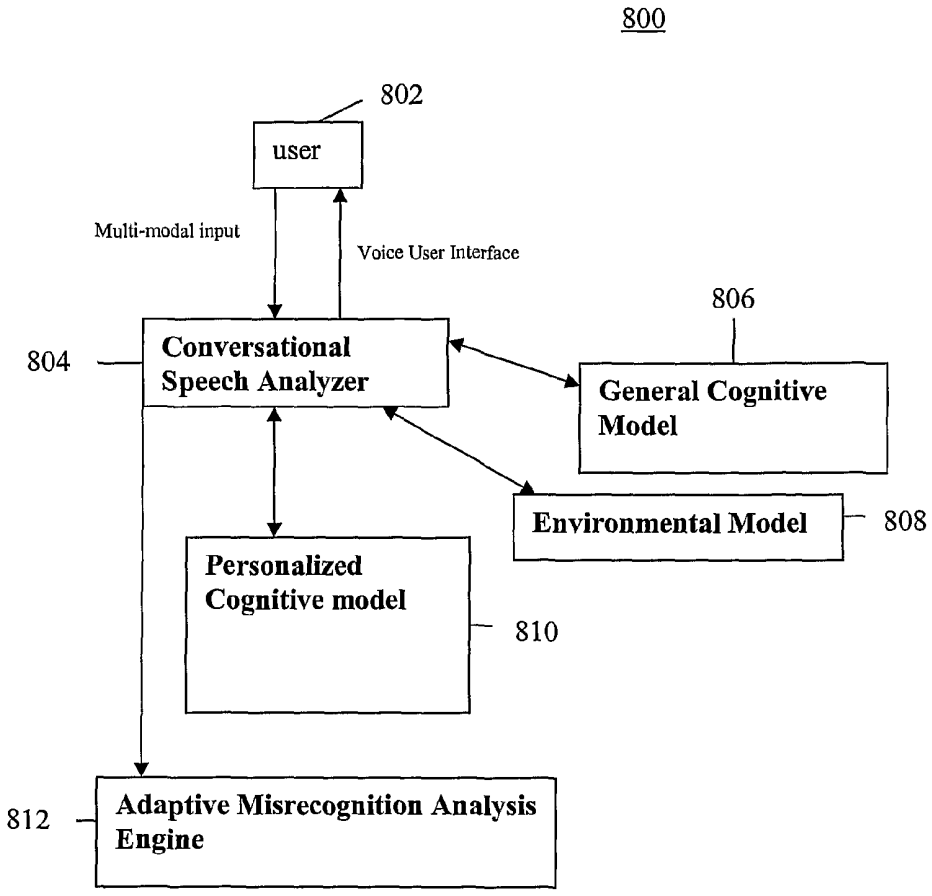

800
802
user
Multi-modal input
Voice User Interface
806
804
Conversational Speech Analyzer
General Cognitive Model
Environmental Model
808
Personalized Cognitive model
810
812
Adaptive Misrecognition Analysis Engine

FIGURE 8

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 February 2007 (22.02.2007)

PCT

(10) International Publication Number
WO 2007/021587 A3

(51) International Patent Classification:
G06F 17/27 (2006.01)

(21) International Application Number:
PCT/US2006/030459

(22) International Filing Date: 4 August 2006 (04.08.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/200,164 10 August 2005 (10.08.2005) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [—/US]; 10505 NE 38TH PLACE, Building 9, Kirkland, WA 98033 (US).

(72) Inventors; aud

(75) Inventors/Applicants *(for US only)*: **DI CRISTO, Philippe** [FR/US]; 177 107th Avenue, N.E., #1007, Bellevue, WA 98004 (US). **WEIDER, Chris** [US/US]; 2818 Grand Avenue, #B-508, Everett, WA 98201 (US). **KENNEWICK, Robert, A.** [US/US]; 2204 S. Eddy Street, Seattle, WA 98108 (US).

(74) Agents: **GATTO, James, G.** et al.; PILLSBURY WINTHROP SHAW PITTMAN LLP, P.O. Box 10500, McLean, VA 22102 (US).

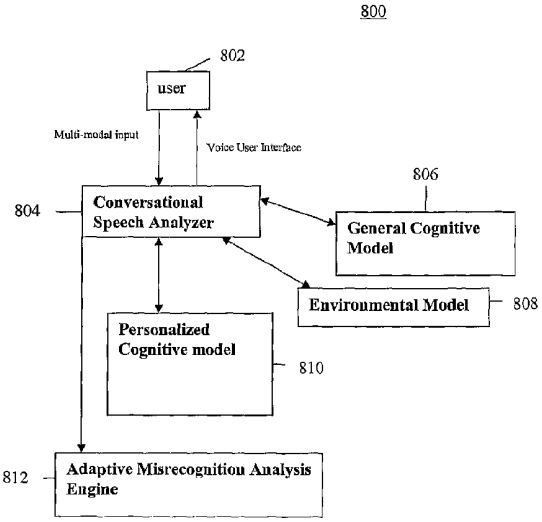
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

[Continued on next page]

(54) Title: SYSTEMS AND METHODS OF SUPPORTING ADAPTIVE MISRECOGNITION IN CONVERSATIONAL SPEECH



(57) Abstract: A system and method are provided for receiving speech and/or non-speech communications of natural language questions and/or commands and executing the questions and/or commands. The invention provides a conversational human-machine interface that includes a conversational speech analyzer, a general cognitive model, an environmental model, and a personalized cognitive model to determine context, domain knowledge, and invoke prior information to interpret a spoken utterance or a received non-spoken message. The system and method creates, stores and uses extensive personal profile information for each user, thereby improving the reliability of determining the context of the speech or non-speech communication and presenting the expected results for a particular question or command.

WO 2007/021587 A3

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

(88) Date of publication of the international search report:
27 September 2007

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 06/30459

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8): G06F 17/27 (2007.01)
USPC: 704/9
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC: 704/9

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 704/200; 704/251

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PubWEST(USPT, PGPB, EPAB, JPAB), GOOGLE SCHOLAR - terms: natural language recognition, non-speech, user profile or preference or personality

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2004/0193420 A1 (KENNEWICK et al.) 30 September 2004 (30.09.2004), entire document.	1-44
A	US 2002/0143535 A1 (KIST et al.) 03 October 2002 (03.10.2002), entire document.	1-44

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
29 April 2007 (29.04.2007)	27 JUL 2007
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (April 2005)

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 March 2007 (08.03.2007)

PCT

(10) International Publication Number
WO 2007/027546 A2

(51) International Patent Classification:
G10L 21/00 (2006.01)

(21) International Application Number:
PCT/US2006/033378

(22) International Filing Date: 29 August 2006 (29.08.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/212,693 29 August 2005 (29.08.2005) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [US/US]; 10505 NE 38th Place, Building 9, Kirkland, WA 98033 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **WEIDER, Chris** [US/US]; 2818 Grand Ave. #B-508, Everett, WA 98201 (US). **KENNEWICK, Rich** [US/US]; 14222 227th Ave., NE, Woodinville, WA 98077 (US). **KENNEWICK, Mike** [US/US]; 13407 37th Place, NE, Bellevue, WA 98005 (US). **DI CRISTO, Phillippe** [FR/US]; 177 107th Ave. #1007, NE, Bellevue, WA 98004 (US). **KENNEWICK, Robert, A.** [US/US]; 2204 S. Eddy St., Seattle, WA 98108 (US).

(74) Agents: **GATTO, James, G.** et al.; Pillsbury Winthrop Shaw Pittman LLP, P.O. Box 10500, McLean, VA 22102 (US).

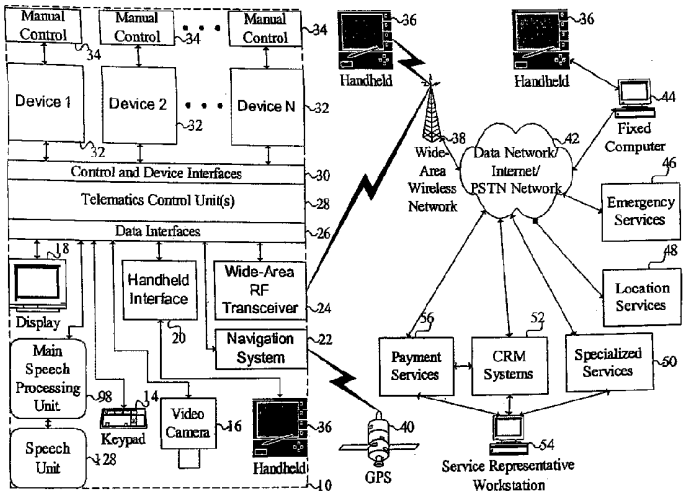
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

[Continued on next page]

(54) Title: MOBILE SYSTEMS AND METHODS OF SUPPORTING NATURAL LANGUAGE HUMAN-MACHINE INTERACTIONS



(57) Abstract: A mobile system is provided that includes speech-based and non-speech-based interfaces for telematics applications. The mobile system identifies and uses context, prior information, domain knowledge, and user specific profile data to achieve a natural environment for users that submit requests and/or commands in multiple domains. The invention creates, stores and uses extensive personal profile information for each user, thereby improving the reliability of determining the context and presenting the expected results for a particular question or command. The invention may organize domain specific behavior and information into agents, that are distributable or updateable over a wide area network.

WO 2007/027546 A2

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

MOBILE SYSTEMS AND METHODS OF SUPPORTING NATURAL LANGUAGE HUMAN-MACHINE INTERACTIONS

FIELD OF THE INVENTION

[001] The invention provides mobile devices having a speech interface and/or a combination of a speech interface and a non-speech interface to enable natural language human-machine interactions. More specifically, the invention enables mobile users to submit natural language speech and/or non-speech questions or commands in a wide range of domains. The mobile device is configured to present responses in a natural manner.

BACKGROUND OF THE INVENTION

[002] Telematic systems are systems that bring human-computer interfaces to mobile environments. Conventional computer interfaces use some combination of keyboards, keypads, point and click techniques and touch screen displays. These conventional interface techniques are generally not suitable for a mobile environments, due at least in part to the speed of interaction and the inherent danger and distraction. Therefore, speech interfaces are being adopted in many telematic applications.

[003] However, creating a natural language speech interface that is suitable for use in the mobile environment has proved difficult. A general-purpose telematics system should accommodate commands and requests from a wide range of domains and from many users with diverse preferences and needs. Further, multiple mobile users may want to use such systems, often simultaneously. Finally, most mobile environments are relatively noisy, making speech recognition inherently difficult.

[004] Retrieval of both local and network hosted online information and processing of commands in a natural manner remains a difficult problem in any environment, especially a mobile environment. Cognitive research on human interaction shows that verbal communication, such as a person asking a question or giving a command, typically relies heavily on context and domain knowledge of the target person. By contrast, machine-based requests (a request may be a question, a command, and/or other types of communications) may be highly structured and may not be inherently natural to the human user. Thus, verbal communications and machine processing of requests that are extracted from the verbal communications may be fundamentally

incompatible. Yet the ability to allow a person to make natural language speech-based requests remains a desirable goal.

[005] Research has been performed on multiple fields of natural language processing and speech recognition. Speech recognition has steadily improved in accuracy and today is successfully used in a wide range of applications. Natural language processing has previously been applied to the parsing of speech queries. Yet, a limited number of systems have been developed that provide a complete environment for users to make natural language speech requests and/or commands and receive natural sounding responses in a mobile environment. There remain a number of significant barriers to creation of a complete natural language verbal and/or textual-based query and response environment.

[006] The fact that most natural language requests and commands are incomplete in their definition is a significant barrier to natural language query-response interaction. Further, some questions can only be interpreted in the context of previous questions, knowledge of the domain, or the user's history of interests and preferences. Thus, some natural language questions and commands may not be easily transformed to machine processable form. Compounding this problem, many natural language questions may be ambiguous or subjective. In these cases, the formation of a machine processable query and returning of a natural language response is difficult at best.

[007] Even once a question is asked, parsed and interpreted, machine processable requests and commands must be formulated. Depending on the nature of the question, there may not be a simple set of requests that return an adequate response. Several requests may need to be initiated and even these requests may need to be chained or concatenated to achieve a complete result. Further, no single available source may include the entire set of results required. Thus, multiple requests, perhaps with several parts, may need to be made to multiple data sources, which can be located both locally or remotely. Not all of these sources and requests may return useful results or any results at all.

[008] In a mobile environment, the use of wireless communications may further reduce the chances that requests will be complete or that successful results will be returned. Useful results that are returned are often embedded in other information and may need to be extracted therefrom. For example, a few key words or numbers often need to be "scraped" from a larger amount of other information in a text string, table,

list, page, or other information. At the same time, other extraneous information such as graphics or pictures may need to be removed to process the response in speech. In any case, the multiple results should be evaluated and combined to form the best possible answer, even in the case where some requests do not return useful results or fail entirely. In cases where the question is ambiguous or the result inherently subjective, determining the best result to present is a complex process. Finally, to maintain a natural interaction, responses should be returned rapidly to the user. Managing and evaluating complex and uncertain requests, while maintaining real-time performance, is a significant challenge.

[009] These and other drawbacks exist in existing systems.

SUMMARY OF THE INVENTION

[010] The invention overcomes these and other drawbacks of prior telematic systems.

[011] According to one aspect of the invention, speech-based and non-speech-based systems are provided that act on commands and retrieve information. The invention uses context, prior information, domain knowledge, and user specific profile data to achieve a natural environment for users that submit requests and/or commands in multiple domains. At each step in the process, accommodation may be made for full or partial failure and graceful recovery. The robustness to partial failure may be achieved through the use of probabilistic and fuzzy reasoning at several stages of the process. This robustness to partial failure promotes the feeling of a natural response to questions and commands.

[012] According to another aspect of the invention, the interactive natural language system (herein "the system") may be incorporated into mobile devices or may be connected to the mobile device via a wired or wireless connection. The mobile device may interface with computers or other electronic control systems through wired or wireless links. The mobile device may also operate independently of a mobile structure and may be used to remotely control devices through a wireless local area connection, a wide area wireless connection or through other communication links.

[013] According to one aspect of the invention, software may be installed onto the mobile device that includes an input module that captures the user input; a parser that parses the input, a text to speech engine module for converting text to speech; a network interface for enabling the device to interface with one or more networks; a

non-speech interface module, an event manager for managing events and/or other modules. In some embodiments, the event manager may be in communication with a context description grammar, a user profile module that enables user profiles to be created, modified and/or accessed, a personality module that enables various personalities to be created and/or used, an agent module, an update manager and one or more databases. It will be understood that this software may be distributed in any way between a mobile device, a computer attached to a mobile structure, a desktop computer or a server without altering the function, features, scope, or intent of the invention.

[014] According to one aspect of the invention, the system may include a speech unit interface device that receives spoken natural language requests, commands and/or other utterances from users, and a computer device or system that receives input from the speech unit and processes the input and responds to the user with a natural language speech response.

[015] According to another aspect of the invention, the system may be interfaced by wired or wireless connections to one or more other systems. The other systems may themselves be distributed between electronic controls or computers that are attached to a mobile structure or are located external to the mobile structure. The other systems may include, electronic control systems, entertainment devices, navigation equipment, measurement equipment or sensors, or other systems. External systems may also be provided with features that include payment systems, emergency assistance networks, remote ordering systems, automated or attended customer service functions, or other features.

[016] According to another aspect of the invention, the system may be deployed in a network of devices that share a common base of agents, data, information, user profiles, histories or other components. Each user may interact with, and receive, the same services and applications at any location equipped with the mobile device on the network. For example, multiple mobile devices may be placed at different locations throughout a home, place of business, vehicle or other location. In such a case, the system can use the location of the particular device addressed by the user as part of the context for the questions asked.

[017] According to one embodiment of the invention, processing may be performed at the mobile devices. The commands may be processed on-board to enable the

mobile devices to control themselves and/or to control other mobile devices, fixed computers, mobile telephones, and other devices. Additionally, mobile devices may track context.

[018] According to one embodiment of the invention, infrastructure may be provided to maintain context information during multimodal interactions, such as speech and/or non-speech interactions. According to one exemplary embodiment of the invention, context information may be maintained in a multimodal environment by providing communication channels between mobile device, or multimodal devices, and the system. The communication channels allow the system to receive multimodal input such as text-based commands and questions and/or voice-based commands and questions. According to another embodiment of the invention, the multimodal input may include a string of text, such as keywords, that are received as commands or questions. According to yet another embodiment of the invention, the system may synchronize the context between the mobile devices and the speech-based units. In order to send a response to the corresponding mobile device, the system may track the source and send the response to the corresponding speech interface or the non-speech interface.

[019] According to an alternative embodiment of the invention, context information may be maintained using a context manager that may be centrally positioned to receive input from multiple mobile devices and to provide output to multiple mobile devices. According to one embodiment, the mobile devices that communicate with the context manager may register through a registration module and may subscribe to one or more events. According to another embodiment of the invention, the context manager may receive input in Context XML form, for example. The other registered mobile devices may be informed of context changes through a context tracking module to enable synchronizing of context across the registered modules. According to one embodiment of the invention, registered modules may be added or removed from the system. The registered modules may include dynamic link libraries (DLLs), or other information sources, that are specific to multimodal devices.

[020] According to yet another alternative embodiment of the invention, context information may be determined from a command or request that is presented in a textual format and/or a command or request that is presented as an utterance and processed using a multi-pass automatic speech recognition module that transcribes the

utterance to a text message. The command or request may be compared against a context description grammar to identify a match. Any active grammars in the context description grammar may be scored against the command or request and a best match may be sent to a response generator module. Agents may be associated with corresponding response generator modules and may retrieve the requested information for generation of a response. The agents may update a context stack to enable follow-up requests.

[021] According to another embodiment of the invention, mobile devices may be configured to allow verbal annotations of objects stored thereon. Mobile devices may transcribe the verbal annotation to text and store the textual annotation with the object. Alternatively, mobile devices may be configured to enable users to manually input textual descriptions that are stored along with the objects. According to one embodiment of the invention, the textual annotations and/or textual descriptions may be classified and searched. In an alternative embodiment, mobile devices may classify and search the verbal annotations rather than textual annotations. However, classifying and searching verbal annotations may be considerably more difficult than classifying and searching textual annotations and/or textual descriptions.

[022] According to one embodiment of the invention, the textual annotations and textual descriptions may be communicated using a short message service on the mobile telephone or other device. Short message service is a text message service that enables sending and receiving of short textual messages. The textual messages may be stored at data centers for forwarding to intended recipients. Other configurations may be used.

[023] According to another aspect of the invention, domain specific behavior and information may be organized into data managers. Data managers are autonomous executables that receive, process and respond to user questions, queries and commands. The data managers provide complete, convenient and re-distributable packages or modules of functionality that are typically directed to a specific domain of applications. Data managers may be complete packages of executable code, scripts, links to information, and other forms of communication data that provide a specific package of functionality, usually in a specific domain. In other words, data managers may include components for extending the functionality to a new domain. Further, data managers and their associated data may be updated remotely over a

network as new behavior is added or new information becomes available. Data managers may use system resources and the services of other, typically more specialized, data managers. Data managers may be distributed and redistributed in a number of ways including on removable storage media, transfer over networks or attached to emails and other messages. An update manager may be used to add new data managers to the system or update existing data managers.

[024] In order to enhance the natural query and response environment, the system may format results to increase understandability to users. Formatting and presentation of results may be based on the context of the questions, the contents of the response being presented, the history of the interaction with the user, the user's preferences and interests and the nature of the domain. By contrast, rigid, highly formatted, or structured presentation of results may be deemed unnatural by many users.

[025] According to another embodiment of the invention, the system may simulate some aspects of a human "personality". In some cases, the presentation of the response and the terms that are used to provide the response may be randomized to avoid the appearance of rigidly formatted or mechanical responses. The use of other simulated personality characteristics is also desirable. For example, a response that may be upsetting to the user may be presented in a sympathetic manner. Furthermore, results of requests may be long text strings, lists, tables or other lengthy sets of data. Natural presentation of this type of information presents particular challenges because simply reading the long response is generally not preferred. Instead, the system may parse important sections from the response and may initially provide only reports. Determining what parts of a long response are presented may be based on the context of the questions, the contents of the response being presented, the history of the interaction with the user, the user's preferences and interests and the nature of the domain. At the same time, the system may give the user interactive control over what information to present and how much information to present, to stop the response all together, or to take other actions.

[026] The invention may be applied as a user interface to telematics systems in a wide variety of environments. These environments may include, but are not limited to, the following: 1) personal automobiles, rented automobiles, or fleet automobiles; 2) motorcycles, scooters, and other two wheeled or open-air vehicles; 3)commercial long-haul and short haul trucks; 4) delivery service vehicles; 5) fleet service vehicles;

6) industrial vehicles; 7) agricultural and construction machinery; 8) water-borne vehicles; 9) aircraft; and, 10) specialized military, law enforcement and emergency vehicles.

[027] According to another aspect of the invention, the system may process and respond to questions, requests and/or commands. Keywords or context may be used to determine whether the received utterance and/or textual message includes a request or command. For example, utterances may include aspects of questions, requests and/or commands. For example, a user may utter "tune in my favorite radio station." A request is processed to determine the name, the channel, and time for the users favorite radio station. If the programming on that station is of a type the user generally does not listen to, the system can suggest using an alterative option, such as listening to a CD more likely to please the user. A command must be executed to set a tune the radio station.

[028] The invention can be used for generalized local or network information query, retrieval and presentation in a mobile environment. For each user utterance, including a question or query or set of questions or queries, the system may perform multiple steps possibly including: 1) capturing the user's question or query through speech recognition operating in a variety of real-world environments; 2) parsing and interpreting the question or query; 3) determining the domain of expertise required and the context to invoke the proper resources, including agents; 4) formulating one or more queries to one or more local and/or network data sources or sending appropriate commands to local or remote devices or the system itself; 5) performing required formatting, variable substitutions and transformations to modify the queries to a form most likely to yield desired results from the available sources; 6) executing the multiple queries or commands in an asynchronous manner and dealing gracefully with failures; 7) extracting or scraping the desired information from the one or more results, which may be returned in any one of a number of different formats; 8) evaluating and interpreting the results, including processing of errors, gathered and combine them into a single best result judged to be "best" even if the results are ambiguous, incomplete, or conflicting; 9) performing required formatting, variable substitutions and transformations to modify the results to a form most easily understood by the user; 10) presenting the compound result, through a text to speech engine or multimodal interface, to the user in a useful and/or expected manner; 11)

optionally, providing a response to users indicating the success or failure of the command, and possibly including state information; or other steps.

[029] The above steps may be performed with knowledge of the domain of expertise, the context for the question or command, domain specific information, the history of the user's interactions, user preferences, available information sources or commands, and responses obtained from the sources.

[030] Probabilistic or fuzzy set decision and matching methods may be applied to deal with inconsistent, ambiguous, conflicting and incomplete information or responses. In addition, asynchronous queries may be used to provide rapid and graceful failure of requests or commands that allow the system to robustly return results quickly and in a manner that seems natural to the user.

[031] Many everyday questions are inherently subjective and result in answers that are a matter of opinion or consensus, as much as fact. Such questions are often ad hoc in their nature, as well. The system, according to another aspect of the invention, may use adaptive, probabilistic, and fuzzy set decision and matching methods to identify the subjective nature of the question and to evaluate a range of possible answers, wherein one or more answers may be selected that most accurately represent the type of result desired by the user.

[032] The context and expected results from a particular question may be highly dependent on the individual asking the question. Therefore, the system may create, store and use personal profile information for each user. Information in the profile may be added and updated automatically as the user uses the system or may be manually added or updated by the user or by others. Domain specific agents may collect, store and use specific profile information, as may be required for optimal operations. Users may create commands for regularly used reports, automatically generated alerts, and other requests and for the formatting and presentation of results. The system may use profile data in interpreting questions, formulating requests, interpreting request results and presenting answers to the user. Examples of information in a user profile includes, history of questions asked, session histories, formatting and presentation preferences, special word spelling, terms of interest, special data sources of interest, age, sex, education, location or address, place of business, type of business, investments, hobbies, sports interests, news interests and other profile data.

[033] According to one aspect of the invention, the system may attempt to provide rapid responses in order to provide a natural question and response environment. The rapid responses may be provided without obtaining additional information. The system may determine agent composition, context and/or domain for a user's question or command, for example, by using a real-time scoring system or other technique. Based on this determination, the system may trigger one or more agents to respond to the user's question or command. The agents may make one or more requests and rapidly return a formatted response. Thus, users may receive direct responses to a set of questions, each with a different response or context. In some cases, the available information, including the request results, may not adequately answer the questions presented. In such situations, the user may be asked one or more follow-up questions to resolve the ambiguity. Additional requests may then be made before an adequate response is provided. In these cases, the system may use context information, user profile information and/or domain specific information to minimize the interaction with the user required to deliver a response.

[034] If the confidence level of the domain or context score is not high enough to ensure a reliable response, the system may request that the user verify that the question or command is correctly understood. In general, the question may be phrased to indicate the context of the question including all criteria or parameters. If the user confirms that the question is correct, the system may proceed to produce a response. Otherwise, either the user can rephrase the original question, perhaps adding additional information to remove ambiguity, or the system may ask one or more questions to attempt to resolve the ambiguity or other actions may taken.

[035] According to one aspect of the invention, the system may accept any natural language question or command and, as a result, may be subject to ambiguous requests. To assist users in formulating concise questions and commands, the system may support a voice query language. The language may help users clearly specify the keywords or contexts of the question or command along with the parameters or criteria. The system may provide built in training capabilities to help the user learn the best methods to formulate their questions and commands.

[036] In order to make the responses to user's questions and commands seem more natural, the system may employ one or more dynamically invokeable personalities and/or emotional models. Personalities and emotional models have specific

characteristics that simulate the behavioral characteristics of real humans. Examples of these characteristics include, sympathy, irritation, and helpfulness and associated emotions. The personality also randomizes aspects of responses, just as a real human would do. This behavior includes randomization of terms used and the order of presentation of information. Characteristics of the personality and/or emotions are invoked using probabilistic or fuzzy set decision and matching methods and using criteria, including the context for the question, the history of the user's interaction, user preferences, information sources available, responses obtained from the sources.

[037] According to another aspect of the invention, special procedures may be employed to present information in the form of long text strings, tables, lists or other long response sets. Simply presenting a long set of information in an ordered manner may not be considered natural or what most users have in mind. The system may use probabilistic or fuzzy set matching methods to extract relevant information and present these subsets first. Further the system may provide commands allowing users to skip through the list, find keywords or key information in the list or stop processing the list altogether.

[038] According to one embodiment of the invention, the system may support multiple users that access the system at different times. According to another embodiment of the invention, the system may support multiple users that access the system during a same session, in an interleaved or overlapping manner. The system may recognize the multiple users by name, voice, or other characteristic and may invoke a correct profile for each user. If multiple users are addressing the system in overlapping or interleaved sessions, the system may identify the multiple users and may invoke one or more corresponding profiles. For applications that require security safeguards, the multiple users may be verified using voiceprint matching, password or pass-phrase matching, or other security safeguards.

[039] When multiple users are engaged in interleaved sessions, the system may gracefully resolve conflicts using a probabilistic or fuzzy set decision method for each user. This process may simulate the manner in which a human would address multiple questions from various sources. For example, the system may answer short questions first in time while answering longer questions later in time. Alternatively, the system may answer questions in the order that they are received, among other configurations.

[040] Since the invention may operate in many environments, including mobile environments with background noise, point noise sources and people holding conversations, filtering of speech input may be advantageous. The invention may use, for example, either one-dimensional or two-dimensional array microphones (or other devices) to receive human speech. The array microphones may be fixed or employ dynamic beam forming techniques. The array pattern may be adjusted to maximize gain in the direction of the user and to null point noise sources. Alternatively, microphones may be placed at particular locations within the mobile environment near where occupants are likely to use the system. These microphones can be single microphones, directional microphones or an array of microphones. Speech received at the microphones may then be processed with analog or digital filters to optimize the bandwidth, cancel echoes, notch-out narrow band noise sources, or perform other functions. Following filtering, the system may use variable rate sampling to maximize the fidelity of the encoded speech, while minimizing required bandwidth. This procedure may be particularly useful in cases where the encoded speech is transmitted over a wireless network or link.

[041] The invention can be applied to a wide range of telematics applications. General applications areas can include, but are not limited to remote or local vehicle control, information query, retrieval and presentation from local or network sources, safety applications, and security applications.

[042] The system may provide local or remote control functions for the system, including devices that are located on the mobile structure or off the mobile structure. Users may initiate commands locally or remotely. Typically, remote operation may be conducted through an IP connection, a telephone connection, or other connections. The user may address spoken commands to a mobile device or desktop unit, which may send the commands to controllers on the vehicle over wireless links. Other remote command techniques may be used. The system may process commands in a nearly identical manner to a request. One difference being that the result of the command is generally an action rather than a response. In many cases, the system may give the user a cue or response to indicate that the command has been successfully executed or has failed. In cases of failure, an interactive session may be started to allow the user to resolve the difficulty or formulate a command more likely to succeed.

[043] The invention provides users, including mobile structure operators, with the capability to control most any mobile system functions using interactive speech and non-speech commands and/or requests. Generally, controls of a critical nature, or having safety implications, may employ fail-safe checks that, before execution, verify that a command will not create a hazardous condition. Manual overrides may also be provided as an extra precaution. The invention may provide built in help and user guidance for the devices under control. This guidance may include step-by-step training for operators that are learning to use the features of the mobile structure. The system can provide extensive interactive guidance when commands cannot be executed or when commands fail. This advice may include suggestions to reformulate the command to succeed, suggestions to work around a failure, suggestions for alternative commands that may achieve a similar function, or other suggestions. Examples of control functions that can be performed from local or remote locations by the invention include:

1. Control of mobile structure multimedia entertainment electronics, such as radio, CD player, or video player. This control can be based on user specified play lists and may be sensitive to user profile information including, preferences history, or other information. The invention may include the ability to control multiple or individual multimedia entertainment stations.
2. Control of communications devices such as cell phones, voice mail systems, fax systems, text or instant messaging systems, call and message forwarding systems, email systems and other communication devices. This control includes controlling features such as, address books, phone books, call forwarding, conference calling, and voice mail, among others.
3. Local or remote control of mobile structure systems. Most any device on the mobile structure may be controlled including, door locks, window controls, interior temperature controls, shifting of the transmission, turn signals, lights, safety equipment, engine ignition, cruse control, fuel tank switches, seat adjustments, specialized equipment such as winches, lifting systems or loading systems, and other mobile structure systems.
4. Systems that are located external to the mobile structure typically may be controlled through wireless links including, garage door openers, gate

controllers, mobile structure entry security passes, automated toll collection systems, and mobile structure weighing systems and other external systems.

5. Mobile structure power management and systems control. The invention may provide the mobile structure operator with information on limitations and on tips for better power management or fuel utilization or other systems control.

6. Diagnostic information management. The invention may provide diagnostic information announcements and warnings for the mobile structure operator. These announcements and warnings may be interactive allowing the operator to request additional information or to suggest various courses of action. The invention may mediate a solution to the problem, including providing access to a scheduling service, summoning help or providing instructions for remedial action until a permanent solution can be achieved. The system can ask the operator to authorize the ordering of needed parts and may provide cost estimates. The system may receive data for these announcements and warnings from a wide range of sources including sensors and vehicle control computers. Sensors may include fuel level sensors, coolant temperature sensors, oil temperature sensors, axel temperature sensors, tire pressure sensors, and other sensors.

7. System status inquiry. Mobile structure operators may use the interactive natural language interface of the invention to query and receive reports on the status of systems that are located on the mobile structure including, fuel level, interior temperature, outside temperature, engine or other mobile structure system status. The operator can further query the system to receive more information or determine a course of action if problem is detected.

8. Vehicle service history. The invention can provide the vehicle operator or other personnel with interactive access to the vehicle service history. The invention can provide announcements or warnings as the time for service draws near. The user can interact with the system to schedule the required service, order required parts, receive cost estimates, or update the service history. Users may customize the nature of this interaction to suit their desires or policies.

9. Diagnostic and service history. The invention may provide diagnostic and service history information to service personnel. This information may include vehicle fault codes and other information for devices that are under control of or measured by the system. Alternatively, the invention may receive information on the state and history of vehicle operation from other control computers. The invention may provide interactive service information and history. The service history may be queried and presented through the speech interface or the non-speech interface. The system may prompt service personnel for more information if the record is deemed incomplete. In other cases, the invention may prompt service personnel for information on their actions if a change is detected in system status, such as replacement of a part.

[044] The invention can provide users or operators of a mobile structure with specialized safety functions through the interactive speech interface and/or non-speech interface. The invention may use a dynamically evocable personality that is capable of creating announcements that are appropriate for the severity of the situation. The announcements and personalities may be under user control and configuration. Some examples of these safety applications can include:

1. The invention can provide automated detection and reporting of accident situations through the wireless communications link. Information on an accident situation can be gathered from airbag control systems or other sensors. Once an accident situation has been detected the invention may use the interactive speech interface and/or non-speech interface to determine the nature of the accident and/or the condition of victims. This information, along with location information and other relevant information, may be reported over the wireless link. Alternatively, the invention may establish voice channel communications or other communications between occupants of the vehicle and emergency personnel.

2. The invention can be used to store and retrieve medical information on vehicle occupants. Following an accident, emergency personnel may query the system for this information. Alternatively, the system may annunciate a warning to emergency personnel if a person has a special medical condition. The system maintains the security of medical information through a number of techniques, including not annunciating medical information unless an accident

is detected, or not annunciating medical information unless that person or another authorized person gives permission.

3. Occupants of the vehicle can summon help in the event of a crime using the speech interface and/or non-speech interface. Typical crimes may include robberies and hijackings. The invention allows vehicle occupants to set panic or emergency words or phrases that indicate to the system a crime is occurring.

4. The invention can provide the vehicle operator with safety announcements if an unsafe or potentially unsafe situation is detected. The operator can use the interactive speech interface and/or non-speech interface to obtain more information on the situation or dismiss the alert. The operator can annunciate or otherwise provide commands to remedy or mitigate the situation during this dialog. Conditions that can be announced include, following another vehicle too closely, too great a speed for the road or conditions, obstruction on roadway, a fire in some part of the vehicle, high cargo pressure or temperature, leaks, and other information.

5. The interactive speech interface and non-speech interface of the invention can provide the operator with real-time assistance. This assistance may include, aid in parking or backing a vehicle, aid with complex maneuvers, aid with optimal operation of the vehicle, and other operations. The operator can ask the system for advice or assistance with a planned maneuver or operation. Alternatively, the invention can proactively offer assistance if certain situations are detected.

6. The interactive speech interface and non-speech interface may be used to improve vehicle security. Voiceprints or voice authentication can be used to gain access to the vehicle or start the vehicle. Alternatively or in addition, a password or pass-phrase can be used. In another alternative speech security can be used as a supplement to other vehicle security techniques.

7. The invention can provide measurement of operator fatigue and alert the operator or remote personnel if unacceptable levels of fatigue are detected. The interactive speech interface and/or non-speech interface can be used to query the operator to detect fatigue. Alternatively, or in addition, other measurements of operator fatigue can be used. If a fatigue situation is

detected the invention may initiate a dialog with the operator to determine the extent of the problem and if required, ask the operator to cease operation.

[045] The invention can offer vehicle operators and occupants a variety of services that are useful while in the vehicle and/or while arriving at a destination. Further, users can employ the interactive natural language interface to customize these services to suit each individual. Some examples of services that can be supported by the natural-language interactive speech interface of the invention, include:

1. Providing vehicle operators with interactive directions to a destination or waypoint, wherein the user can specify a desired destination and any preferred waypoints. A destination can be specified in any manner including providing the name of a place, an address, a name of a person, a name of a business, or other type of information. As the trip progresses the system may provide the operator with continued directions and warnings if a mistake has been made. The operator can query the system for additional information or request less information. Generally, the system is interfaced with one or more navigation sensors and local or remote map databases. The invention may provide operators or passengers with alerts of upcoming points or interest, exits or stops, hazards, or other conditions. The users can query the system for more specific information. Alternatively, the invention can provide operators and occupants of the vehicle with an interactive guided tour. The system's information query, retrieval and presentation capability can be employed by users to receive additional information or points or items of interest during the tour and may take into account stored personal profile information for a user.
2. The invention can provide the operator of a vehicle with interactive dynamic routing information. The routing can be updated based on traffic conditions, weather conditions, facilities availability, and information provided by the operator. Generally, the system is interfaced with one or more navigation sensors, local or remote map databases, and sources of traffic, weather, and facilities use data.
3. The direction, routing and communications capabilities of the invention can be combined in an interactive system that helps one or more operators rendezvous at a predetermined destination or any other convenient midpoint. The operators may use the interactive natural language interface to

communicate with the system to arrange the rendezvous, to receive directions as they travel to the rendezvous point and/or to communicate with the other operators.

4. The navigational capabilities of the invention may be used to place limits on where a vehicle is allowed to go and/or for how long. The system may employ the interactive natural language speech interface and/or non-speech interface to inform the operator when the vehicle is approaching or has exceeded a limit. The operator can query the system to determine the best course of action to return to limits or prevent exceeding them. Alternatively, the system can query the operator to determine why they are exceeding the limits or to mediate a negotiation to extend the limits if this is required by circumstances. This capability is useful in several situations including, keeping a delivery or passenger vehicle on a regular route, setting and enforcing use limits on teenagers, preventing an operator from using the vehicle in an unauthorized manner, or other situations.

5. The interactive natural language interface of the invention can be used to provide Customer Relationship Management (CRM) services to vehicle operators and passengers. The user can interact with the services offered via data networks, video signals, or audio. The interaction can be performed with automated services or a live Customer Service Representative. Interactions with the customer service representatives can be through any combination of possible techniques, such as, live audio, live video, electronic messaging or email, instant messaging, and other techniques. These services can be offered by a number of entities including, vehicle manufacturers, vehicle dealers, vehicle service organizations, automobile or travel clubs, wireless carriers, travel service organizations, or other organizations. The services offered can be personalized to the occupants of the vehicle using a variety of information including, user profile information, history, location, paths traveled, time of day, day of week, or other information. In addition, the system can offer customized services based on information about the vehicle, including paths traveled, distance, service history, type of equipment on vehicle. These services can be accessed while a person is an occupant in a vehicle, while the person is not an occupant but are using a wireless or wired network equipped

mobile device or while using a wired or wireless network desktop system, among other situations. Examples of these services include:

a. Location based marketing programs wherein occupants of the vehicle receive promotional offers from merchants along a route of travel. Occupants can query the system for offers and promotions for goods, services, or other information, along the travel route. The system may apply other available information to form a response, including, the user profile, history, location, or other information. The system can provide optimized interactive routing assistance to the vehicle operator. Alternatively, the system can provide interactive offers and promotions for goods and services along the route, or in advance of a particular trip. Promotions can be offered for goods and services can include but are not limited to, travel services, groceries, prepared foods, vehicle service, fuel, entertainment, or other goods and services.

b. Remote ordering and payment for goods and services. The system can interactively present the menu or product catalog using the list and table presentation capabilities of the invention. The system may facilitate remote ordering by using location information, customer preferences, customer order history, etc. The system can manage a secure payment wallet for the users. Voiceprints, spoken passwords, and non-speech security methods (i.e. PIN pad, etc.) can be combined to create the appropriate level of security.

c. Travel services for occupants of the vehicle. These services can include, directories of travel and entertainment services, or reservations for entertainment restaurants, hotels and other accommodation. The system may present directories, lists and menus using its interactive list and table presentation capabilities. The travel service capability can be used in conjunction with the remote ordering and payment capabilities and the dynamic interactive routing capability.

d. Answer specialized travel related questions in areas such as vehicle registration, taxes, safety laws, required inspections, weight

limits, insurance coverage requirements, insurance policy provisions, or other areas.

6. The invention can provide an operator or other occupant of a vehicle with an interactive location sensitive shopping list or a location and time sensitive task reminder list using the natural language interface. Users can create the list while in the vehicle, while on foot using a mobile device, or at a fixed location using a handheld or desktop device. A user may grant permission to other users to add tasks or shopping items to their lists. Once in the vehicle, the system may provide occupants with routing assistance to optimize travel time and may provide reminders of items to be purchased and tasks to be completed as the vehicle comes in close proximity to a particular location, type of merchant or other service provider, or when a set time has been reached, among other criteria.

7. Automatic interactive dispatch and reporting for fleet vehicles having vehicle operators or other vehicle occupants that use the speech interface and/or non-speech interface to interact with these services. These services can include, dynamic optimal routing, inventory of parts and other materials, ordering of required parts and materials, work orders, receipt generation, payments, or other services.

8. Sales force automation, sales reporting, contact database management, calendar management, and call routing. The system may employ its interactive list and table presentation capabilities to supply catalog and pricing information or other information. These services can use local or network data. Additional services can include, memos, reminders, activity lists, or other information.

[046] Vehicle operators and other occupants can use the interactive natural language interactive interface of the invention to perform many types of information query, retrieval and presentation operations. Using the natural language interactive interface, users can modify the parameters of queries or specify the presentation formats for results. Data used to create a response can be from any combination of local and remote data sources. User specific data can be synchronized between systems fixed to one or more vehicles, mobile structures and desktop systems. Some examples of

information query, retrieval and presentation applications for the invention include, but are not limited to the following:

1. White pages and yellow pages lookups to find, email addresses, telephone numbers, street addresses and other information for businesses and individuals. These services can be used in conjunction with other services, including remote ordering and payment, offers and promotions, mapping, and driving directions;

2. Management and access to personal address book, calendars and reminders for users;

3. Automatic telephone dialing, reading and sending emails, pages, instant messaging by voice, text or video and other communications control functions;

4. Selection, schedules and play list management for television, satellite broadcast, radio or other entertainment schedule. The available information can include reviews and other information on programming. The system may provide device control for users;

5. Weather information for the local area or other locations;

6. Stock and other investment information including, prices, company reports, profiles, company information, business news stories, company reports, analysis, price alerts, news alerts, portfolio reports, portfolio plans, or other information;

7. Local, national and international news information including headlines of interest by subject or location, story summaries, full stories, audio and video retrieval and play for stories;

8. Sports scores, news stories, schedules, alerts, statistics, background and history information, or other information;

9. The ability to subscribe interactively to multimedia information channels, including sports, news, business, different types of music and entertainment, applying user specific preferences for extracting and presenting information;

10. Rights management for information or content that is used or published;

11. Horoscopes, daily jokes and comics, crossword puzzle retrieval and display and related entertainment or diversions;

12. Interactive educational programs using local and network material, with lesson material level that is set based on user's profile, location of the vehicle, planned route of the vehicle, planned activities during the trip and including, interactive multimedia lessons, religious instruction, calculator, dictionary and spelling, geographic information, instruction for specialized tasks planned during the trip, language training, foreign language translation, presentation of technical manuals, and encyclopedias and other reference material.

[047] It will be appreciated that the foregoing statements of the features of the invention are not intended as exhaustive or limiting, the proper scope thereof being appreciated by reference to this entire disclosure and reasonably apparent variations and extensions thereof.

BRIEF DESCRIPTION OF THE DRAWINGS

[048] The invention will be described by reference to the preferred and alternative embodiments thereof, in conjunction with the drawings in which:

[049] Figure 1 is an overall block diagram of the system according to a first embodiment of the invention;

[050] Figure 2 is an overall block diagram of the system according to a second embodiment of the invention;

[051] Figure 3 is an overall block diagram of a mobile device according to an embodiment of the invention;

[052] Figure 4 is an overall block diagram of a fixed computer according to an embodiment of the invention;

[053] Figure 5 is an overall diagrammatic view of the interactive natural language processing system according to an embodiment of the invention;

[054] Figure 6 is a schematic block diagram showing the agent architecture according to an embodiment of the invention;

[055] Figure 7 illustrates a diagram for centrally maintaining context, according to one embodiment of the invention; and

[056] Figure 8 illustrates a system diagram for an enhanced conversational voice user interface according to one embodiment of the invention.

DETAILED DESCRIPTION

[057] The following detailed description refers to the accompanying drawings, and describes exemplary embodiments of the invention. Other embodiments are contemplated and modifications may be made to the exemplary embodiments without departing from the spirit, functionality and scope of the invention. Therefore, the following detailed descriptions are not meant to limit the invention.

[058] According to one embodiment of the invention, a telematic natural language speech interface and non-speech interface are provided for use in mobile environments and telematic applications. The system, or portions thereof, may be used in vehicles, while on foot or at a fixed location such as an office or home, or at other locations. An overall block diagram of one embodiment of the invention is illustrated in Figure 1.

[059] A speech unit 128 and/or a keypad 14 may be integrally coupled to a mobile structure 10 or may be part of mobile devices 36, fixed home or office computer systems 44, or other devices. Mobile devices 36 may include mobile telephones, personal digital assistants, digital radios, compact disk players, navigation systems, or other mobile devices. The mobile devices 36 may be configured to integrate with set-top boxes, alarm clocks, radios, or other electronic components. The speech unit 128 and/or keypad 14 may be interfaced to a Telematics Control Unit (TCU) 28 through one or more data interfaces 26. According to some embodiments, the main speech-processing unit 98 may be embedded in one or more TCU 28. In some embodiments, the components of the speech unit 128 may also be distributed between one or more TCUs.

[060] A speech-processing unit may be built into mobile devices 36 and may be coupled with the data interfaces 26 though a wireless, or wired, handheld interface 20. Other user interface peripherals may be connected to the TCU through the data interfaces and may include displays 18, such as touch screen displays for displaying text, graphics and video; keypads 14 for receiving textual data input; video cameras 16 for receiving multimedia communications or conferences; a pointing device or stylus, or other user interface peripherals. Other devices may be connected to the TCU through the data interfaces including wide-area RF transceivers 24, navigation

system components 22, or other devices. The navigation system may include several components such as, Global Positioning System (GPS) receivers or other radiolocation system receivers, gyroscopes or other inertial measurement equipment, distance measurement sensors, such as odometers, or other components. Radiolocation equipment may receive coded signals from one or more satellites or terrestrial sources 40. The one or more location service servers 48 may assist the navigation system. Other systems that can connect to the TCU through the data interfaces may include automotive control computers, digital control interfaces for devices such as media players or other electronic systems, measurement sensors, or other specialized electronic equipment.

[061] The control and device interfaces 30 may connect the TCU 28 to various devices 32 within the mobile structure 10. The control and device interfaces 30 may be used to execute local or remote commands from users of the natural language interface. In some cases, the control and device interfaces 30 may include specialized hardware for interacting with different types of devices. The hardware interfaces may include analog or digital signal interfaces for device control, along with analog or digital interfaces for measurements that may control devices 32. These interfaces may also include specialized software that encapsulates or abstracts specific behavior of the devices 32. The interface software may include one or more drivers that are specific to the hardware interface and to one or more agents. The domain agents may include the specialized software behavior and data required to control a particular device or class of devices. New or updated behavior may be added to the system by updating data managers that are associated with specific devices or class of devices. The devices 32 may include manual controls or manual overrides 34. For safety reasons, the control and device interfaces 30 may incorporate fail-safe systems that, for example, may verify operating limits before changing settings, ensuring that commands do not conflict with settings from manual controls, and will not, in some combination with other commands or device settings, create an unsafe situation. The software behavior and data that ensures safe operations may be included within the domain agent that is specific to the device or class of devices. Examples of devices and system that can be controlled through the control and device interfaces 30 include, power management systems, measurement sensors, door locks, window controls, interior temperature controls, shifting of the transmission, turn signals,

lights, safety equipment, engine ignition, cruise control, fuel tank switches, seat adjustments, specialize equipment such as winches, lifting systems or loading system, and other systems.

[062] The wide-area RF transceiver 24 may communicate with one or more wide-area wireless networks 38, which may be connected to data networks 42, including the Internet, the Public Switched Telephone Network (PSTN) 42 or other data networks. The wide-area wireless networks can be of any suitable terrestrial or satellite based type. Mobile devices 36 may communicate with one or more local or wide-area wireless networks. Home or office systems 44, equipped with wired or wireless network interfaces, may communicate through the data networks or PSTN.

[063] According to one embodiment of the invention, data and agents may be stored and synchronized in mobile structures 10, mobile devices 36 and/or fixed systems 44 having one or more main speech-processing units 98. The synchronization between these different systems can occur on the wide area wireless network 38, the data network 42, through the handheld interface 20, or other local data connections. The synchronization may be performed automatically when any two or more of the computers are connected to these networks. Alternatively, the synchronization may be performed on demand under user control. The synchronization process attempts to determine which version of a data element or an agent are the newest or most up-to-date and propagates that element. Thus, synchronization is an incremental change processes. In some cases, a complete replacement of a database, a portion of the database or of one or more agents may be performed rather than performing a series of incremental updates.

[064] The wide-area wireless networks 38, the data networks 42 or PSTN, may connect mobile structures 10, mobile devices 36 and fixed computers 44 to one or more servers that provide one of more services. According to one embodiment of the invention, an interactive natural language user interface may be provided that supports a transfer of data or transmission of speech, text, video and other formats. For data centric applications, a standardized data transfer format may be used including, for example, Hypertext Markup Language over Hypertext Transfer Protocol (HTTP), Extensible Markup Language (XML), and other data formats or schemas, over HTTP or other transfer protocol, Electronic Data Interchange formats over a variety of transport protocols, etc. It will be understood that the exact configuration of the

servers may be determined on a case-by-case basis with consideration being given to the exact combinations of services being offered, the service providers providing the services, the contractual relationships between the service provider, and other factors. The invention supports any suitable configuration. In each case these servers may themselves be distributed over one or more public or private networks. Some examples of servers that may be used to deliver these services, are given below:

1. One or more payment service providers 56 that supply payment capabilities to users. These payment services may include electronic wallet capabilities for one or more payment accounts and may include payment security information, payment account information, transaction histories, account balance information, and other information. Exemplary payment types include, stored value accounts, promotional accounts, credit accounts, telecommunications billing accounts, debit accounts, or other payment types that are accessed using online or offline methods. Payments can be computed in any manner including payment for a specific good or service, subscription payment, metered payment, or other payments. The payment services can be distributed in a number of ways. Examples of computers and servers used to store and process payment transaction information include, smart cards, main speech processing units 128, mobile devices 36, TCUs 28, fixed personal computers 44, payment gateways, payment servers 56, or other systems.
2. One or more Customer Relationship Management (CRM) systems 52 may supply any number of consumer and business customer services as discussed above. The CRM system can supply automated services or services that are partly or completely manual. For manual services one or more customer service representatives may use one or more service representative workstations 54. The CRM system and the service representative workstation may be connected to one or more data networks 42 or the PSTN. Any other servers may also have connections to one or more service representative workstations, which may be in common or independent of each other.
3. One or more specialized service servers 50 may support specialized consumer and business services.
4. One or more location service servers 48 may supply location information and location based services. The location data is used as a data

input to the location services that may then be distributed in any suitable manner including, on main speech processing units 128, mobile devices 36, TCUs 28, fixed personal computers 44, other servers (i.e. 46, 50, 52, 56) or the one or more location services servers 48.

5. One or more emergency services servers 46 may supply both public and private emergency services to the users.

[065] Main speech processing unit 98, speech unit 128 and keyboard 14 may be distributed in mobile devices 36 in a number of ways. For example, these units may be attached to the mobile devices 36 as independent components or as a single integrated component. In another embodiment of the invention, some or all of the main speech processing unit 98 and speech unit 128 may be embedded in one or more of the TCUs 28, mobile devices 36, fixed computer systems 44, or other devices.

[066] Figure 2 illustrates an alternative block diagram of a second embodiment of the invention. In this embodiment, the main speech processing unit 98 and speech unit 128 are external to the TCU 28. These components may be housed in one or more packages or may be included in a single integrated package.

[067] In all other respects, the second embodiment resembles the first embodiment illustrated in Figure 1. It will be understood that the exact distribution and packaging of the main speech processing unit 98 and speech unit 128 may be determined by the details of the deployment situation and should not change the functionality, capabilities or spirit of the invention in any way.

[068] Figure 3 illustrates a block diagram of one embodiment of the mobile devices 36 that include main speech processing unit 98 and speech unit 128 embedded in the mobile device 36, mobile telephone or other mobile device. The main speech processing unit may interface to one or more processing units 70 located in the mobile device 36. The processing units 70 may include one or more central processing units, one or more data and address busses, data interfaces, volatile memory, or other components. The processing units 70 may use one of more types of nonvolatile memory 80 for software and data storage. Suitable types of nonvolatile memory 80 may include flash memory and hard disk drives. In some embodiments, the main speech processing unit 98 may be integrated with the one or more processing units 70.

[069] According to one embodiment, users may interact with the mobile device 36 through the speech unit 128, the keypad 74 or keyboard, a display 72 that displays

text, graphics, video, or other peripheral. In some embodiments the display may be a touch screen type. Alternatively, a pointing device (not shown) may be used, among other devices.

[070] The mobile device 36 may connect to one or more wired or wireless wide-area or local-area networks through one or more interfaces. A wide-area network transceiver 78 may connect to the wide-area wireless network 38 or the data network 42 using a wireless or wired connection, including an IP connection, a dial-up PSTN network connection or other connections. The local-area network transceiver 76 may connect to wired or wireless local area networks. These networks may include the handheld interface 20 or connections to fixed computer systems 44. In both mobile device environments and fixed computer environments, communications between humans and machines may not provide accurate results at least because natural language commands may be incomplete in their definition. The occurrence of inaccurate results may be reduced by leveraging context of previous utterances, knowledge of the domain and/or the user's history of interests and preferences or other factors.

[071] According to one embodiment of the invention, processing may be performed at the mobile devices 36. Alternatively, the processing may be performed on the server side. In yet another embodiment, the processing may occur on both the mobile devices 36 and the server side. The commands may be processed on-board the mobile devices 36 to enable the mobile devices 36 to control themselves and/or to control other mobile devices 36, fixed computers 44, mobile telephones, or other devices. Additionally, mobile devices 36 may track context. According to another embodiment of the invention illustrated in FIG. 7, a context manager 702 may be centrally maintained to allow input to and output from multiple mobile devices 36. Each of the mobile devices may communicate with the context manager 702 to register via registration module 712, wherein registration may indicate events that the mobile devices 36 may subscribe to. The context manager 702 may receive input in Context XML form, for example. The other registered mobile devices 36 may be informed of context changes through context tracking module 714, thereby synchronizing the context across the registered mobile devices 36. According to one embodiment of the invention, registered mobile devices 36 may be added or removed.

The registered mobile devices 36 may be dynamic link libraries (DLLs), object code, or other data that may be specific to the mobile device 36.

[072] According to yet another alternative embodiment of the invention, context information may be determined from a command or request that is presented as a text message and/or a command or request that is presented as a verbal utterance and processed using a multi-pass automatic speech recognition module that transcribes the verbal utterance to a text message. The multi-pass automatic speech recognition module may use a dictation grammar or a large vocabulary grammar, among other resources, to transcribe the verbal utterance into a text message. On platforms where a dictation grammar is not available, the multi-pass automatic speech recognition may use a virtual dictation grammar that uses decoy words for out-of-vocabulary words. Decoy words may include utility words, nonsense words, isolated syllables, isolated distinct sounds and other decoy words.

[073] According to one embodiment of the invention, the text message may be searched for a particular character, group of characters, word, group of words, and other text combinations. The text combination may be compared against entries in a context description grammar that is associated with each agent 106. If a match is identified between an active grammar in the context description grammar and the command and/or request, then the match may be scored. The agents 106 may be ranked based on the determined score. In generating an aggregate response from the one or more responses received from the agents 106, the ordering of the responses from the individual agents may be determined based on the rank of agents 106. The aggregate response may be generated by a response generator module. Agents 106 may update a context stack, that includes an ordered list of command contexts, to enable follow-up requests.

[074] According to another embodiment of the invention, if a match is not found, or only a partial match is found, between the text message and active grammars, then a knowledge-enhanced speech recognition system may be used to semantically broaden the search. The knowledge-enhanced speech recognition system may be used to determine the intent of the request and/or to correct false recognitions. The knowledge-enhanced speech recognition may access a set of expected contexts that are stored in a context stack to determine a most likely context. The knowledge-enhanced speech recognition may use context specific matchers that are able to

identify context such as time, location, numbers, dates, categories (e.g., music, movies, television, addresses, etc.) and other context. The matching may be performed by comparing a character, group of characters, a word, group of words, and other text combinations. Alternatively, or in addition to text based matching, the matching may be performed using phonetic matching, among other techniques. The results of any match may be used to generate a command and/or request that is communicated to agents 106 for additional processing. According to one embodiment of the invention, non-speech interface 114 may show system, state and history information in a more concise manner than is possible through the speech interface. Non-speech interface 114 may be accessed to create or extend capabilities of agents 106. These operations may include scripting of agents, adding data to the agent or databases 102 used by the agent, adding links to information sources, among other operations.

[075] According to another embodiment of the invention, mobile devices 36 may be configured to allow speech annotations of objects stored thereon. The objects may include photographs, calendar entries, email messages, instant messages, phonebook entries, voice mail entries, digital movies or other objects. Mobile devices 36 may transcribe the speech annotations to textual annotations and store the textual annotations with the object. Alternatively, mobile devices 36 may be configured to enable users to input non-speech annotations, such as textual descriptions, that are stored along with the objects.

[076] According to one embodiment of the invention, the annotated objects may be stored on a server side, a client side, a combination of server side and client side, or according to other configurations. The invention further contemplates collaboratively exchanging and sharing the annotated objects among distributed workgroups that may include centralized servers having shared workspaces for providing common object storage and retrieval facilities. The shared workspaces may be implemented on the centralized servers and may be accessed from different platforms using the mobile devices 36. The system may include a peer-to-peer system for accessing the annotated objects.

[077] According to one embodiment of the invention, the non-speech annotations may be classified and searched. In an alternative embodiment, the speech annotations may be classified and searched. However, classifying and searching speech

annotations may be considerably more difficult than classifying and searching non-speech annotations.

[078] According to another embodiment of the invention, the textual annotations and/or textual descriptions that are associated with the objects may be stored as metadata, thereby enabling searching for the objects using the metadata. The metadata may include GPS information, environmental information, geographic information, or other information. For example, proximity to famous landmarks may be determined using GPS information, environmental information, geographic information, or other information, and this information may be integrated into the metadata associated with the objects. According to one embodiment of the invention, GPS coordinates may be stored in the metadata associated with the objects and users may search for selected objects based on the GPS coordinates. A user may provide a speech command such as "show me all the photos of Greece." In this case, the system would limit the type of object to photographs and would determine the GPS coordinates of Greece. The system would then search the metadata for objects that correspond to photographs and that also satisfy GPS coordinates for Greece. According to another embodiment of the invention, including GPS coordinates in the metadata of objects enables post-processing of the objects based on the GPS coordinates. For example, the objects may be subjected to an initial sort based on generalized GPS coordinates that are stored in the metadata and may be subjected to additional sorting based on more particular criteria for the GPS coordinates. Thus, a user may initially search for object metadata that corresponds to a location near a famous landmark and may use image matching to label the objects with the searchable metadata (i.e., textual descriptions). For example, using the GPS coordinates stored in the metadata, users may first determine that photographs were taken at the Jefferson Memorial and may use this information to label the photographs with searchable metadata (i.e., textual descriptions) including "photo of the Jefferson Memorial." One of ordinary skill in the art will readily appreciate that metadata may include various types of information and may be searched using the various types of information.

[079] According to one embodiment of the invention, the textual annotations may be communicated using a short message service on the mobile telephone or other device. Short message service is a text message service that enables sending and receiving of

short textual messages. The textual messages may be stored at data centers for forwarding to intended recipients. Other configurations may be used.

[080] According to another embodiment of the invention, the mobile devices 36 may support multi-modal communications that enable displaying of non-speech search results on a graphical interface and receipt of speech commands to provide a follow-up search, among other configurations. For example, the user may be presented with textual search results corresponding to a name of a famous person and the user may provide a speech command to find a biography of the famous person. The system may maintain the context of the textual search results to find the biography associated with the famous person. By contrast, known systems may perform a follow-up search on the term "biography" and may present a dictionary definition of the term "biography."

[081] Figure 4 illustrates an embodiment that uses a fixed computer 44. In some embodiments, the main speech processing unit 98 and speech unit 128 may be embedded into the fixed computer 44. The main speech processing unit may interface with one or more processing units 84 that are associated with the fixed computer. The processing units may include one or more central processing units, one or more data and address busses, data interfaces, volatile memory, or other components. The processing unit may use one of more types of nonvolatile memory 94 for software and data storage. Suitable types of nonvolatile memory include, for example, flash memory and hard disk drives. In some embodiments, the main speech processing unit 98 may be integrated with the one or more processing units 84. According to one embodiment of the invention, processing may be performed at the fixed computers 44. The commands may be processed on-board to enable the fixed computers 44 to control themselves and/or to control other fixed computers 44, mobile devices 36, mobile telephones, or other devices. Additionally, fixed computers 44 may track context as described above.

[082] According to another embodiment of the invention, fixed computers 44 may be configured to allow verbal annotations of objects stored thereon. Fixed computers 44 may transcribe the verbal annotation to text and store the textual annotation with the object. Alternatively, fixed computers 44 may be configured to enable users to manually input textual descriptions that are stored along with the objects. According to one embodiment of the invention, the textual annotations and/or textual

descriptions may be classified and searched. In an alternative embodiment, fixed computers 44 may classify and search the verbal annotations rather than textual annotations. However, classifying and searching verbal annotations may be considerably more difficult than classifying and searching textual annotations and/or textual descriptions.

[083] In another embodiment of the invention, users may interact with fixed computers 44 using speech unit 128, the keyboard 88 or keypad, a display 86 for displaying text, graphics, video, or other peripherals. According to some embodiments of the invention, the display may be a touch screen type. Alternatively, a pointing device (not shown) may be used, along with other devices. Fixed computers 44 may be coupled to one or more wired or wireless wide-area or local-area networks through one or more interfaces. A wide-area transceiver 92 may connect to the wide-area wireless network 38 or the data network 42, using a wireless or wired connection, including an IP network, a dial-up PSTN network connection, or other connections. The local-area network transceiver 90 may connect to wired or wireless local area networks. These networks may include connections to mobile devices 36.

[084] In order for devices to properly respond to requests and/or commands that are submitted in a natural language form, machine processable requests and/or algorithms may be formulated after the natural form questions or commands have been parsed and interpreted. Algorithms describe how the machines should gather data to respond to the questions or commands. Depending on the nature of the requests or commands, there may not be a simple set of requests and/or algorithms that will return an adequate response. Several requests and algorithms may need to be initiated and even these requests and algorithms may need to be chained or concatenated to achieve a complete response. Further, no single available source may contain the entire set of results needed to generate a complete response. Thus, multiple requests and/or algorithms, perhaps with several parts, may be generated to access multiple data sources that are located both locally or remotely. Not all of the data sources, requests and/or algorithms may return useful results or any results at all. Useful results that are returned are often embedded in other information and may need to be extracted from the other information. For example, a few key words or numbers may need to be "scraped" from a larger amount of other information in a text string, table, list, page,

video stream or other information. At the same time, extraneous information including graphics or pictures may be removed to process the response. In any case, the multiple results must be evaluated and combined to form the best possible response, even in cases where some requests do not return useful results or fail to produce results entirely. In cases where the command is determined to be ambiguous or the result is inherently subjective, determining the results to present in the response is a complex process. Finally, to maintain a natural interaction, responses should be returned to the user rapidly. Managing and evaluating complex and uncertain requests, while maintaining real-time performance, is a significant challenge.

[085] The invention provides a complete speech-based command generation, information query, retrieval, processing and presentation environment or a combination of speech-based and non-speech-based command generation, information query, retrieval, processing and presentation environment for telematic applications. In addition, the invention may be useful for controlling the system itself and/or external devices. This integrated environment makes maximum use of context, prior information and domain and user specific profile data to achieve a natural environment for one or more users submitting requests or commands in multiple domains. Through this integrated approach, a complete speech-based natural language command, algorithm and response environment or a combination of speech-based and non-speech-based command, algorithm and response environment may be created.

[086] The telematic natural language interface may be deployed as part of, or a peripheral to a TCU or other mobile devices 36, as part of a mobile device interfaced to vehicle computers and other mobile system through wired, wireless, optical, or other types of connections or fixed computers interfaced to the vehicle computers or other systems through a combination of wired, wireless, optical and/or other types of connections. Alternatively, the components of the interactive natural language telematic interface may be distributed in any suitable manner between these multiple computing platforms.

[087] Figure 5 illustrates one exemplary schematic diagram view of a system 90 according to an embodiment of the invention. The system 90 may include a main unit 98, speech units 128, and multi-modal devices 155. In an alternative embodiment, the system 98 may include completely separate systems for main unit 98, speech units

128 and multi-modal devices 155. The event manager 100 may mediate interactions between components of main unit 98. The event manager 100 provides a multi-threaded environment allowing the system 98 to operate on multiple commands or questions from multiple user sessions, without conflict, and in an efficient manner to maintain real-time response capabilities.

[088] FIG. 8 illustrates one exemplary embodiment of an enhanced conversational user interface 800 that receives multi-modal input from users 802 and communicates with a conversational speech analyzer 804. The conversational speech analyzer 804 may be coupled to several components, including a general cognitive model 806, an environmental model 808, a personalized cognitive model 810, and an adaptive misrecognition analysis engine 812 or other components. According to one embodiment of the invention, the conversational speech analyzer 804 may include one or more speech recognition engines that transcribe utterances to textual messages. The transcription may be performed using one or more conversational grammars, a context description grammar 112, or other transcription methods. According to one embodiment of the invention, data that is entered in a text format may be merged with data that is transcribed to a textual format from the utterance. Conversational speech analyzer 804 also may include a semantic knowledge-based model that analyzes the textual message and detects command components.

[089] According to one embodiment of the invention, personalized cognitive model 810 is a model derived from a user's interaction pattern with the system and may be used to predict what actions the user may take next in time, thus assisting with speech recognition and/or question or command recognition. Personalized cognitive model 810 may track actions performed by the user. When the system is attempting to predict user behavior, the personalized cognitive model may be consulted first. The system may have multiple personalized cognitive models, wherein one may correspond to each user.

[090] According to another embodiment of the invention, general cognitive model 806 is a statistical abstract that corresponds to interaction patterns with the system for multiple users. Data stored within general cognitive model 806 may be used to predict a user's next action, thus assisting with speech recognition and/or question or command recognition. The general cognitive model 806 may also track what actions

a particular user has performed and may be used when the user interacts with the system in a way that is not handled in the personalized cognitive model.

[091] According to one embodiment of the invention, the environmental model 808 may include information associated with the user's environment and surroundings. The information may include the type of environment that a user is in (e.g., quiet or noisy); details of a microphone and/or speaker system; the user's current global position and movement, as may be determined by GPS; current system status, such as what song/movie is playing, is the system in the midst of retrieving something, or other system status; details on all voice-enabled devices in the immediate vicinity, such as a presence of a voice-enabled TV, stereo, and DVD player in the same room; a user's credit card information, such as numbers and current balances, wherein the user may ask a mobile telephone to download and pay for a video and the system may respond that there are insufficient funds; or other information. The information may be accessed to invoke a context, domain knowledge, preferences, and/or other cognitive qualities that enhance the interpretation of questions and/or commands.

[092] The conversational speech analyzer 804 may also access the general cognitive model 806 and/or the personalized cognitive model 810 to further refine a context, domain knowledge, preferences, and/or other cognitive qualities to enhance the interpretation of questions and/or commands. Based on information received from general cognitive model 806, environmental model 808 and/or the personalized cognitive model 810, the system may enhance responses to commands and questions by including a prediction of user behavior.

[093] Adaptive Misrecognition Analysis Engine 812 may analyze and store textual messages, including transcribed utterances, that are identified as being unrecognized or incorrectly recognized by conversational speech analyzer 804. Upon a determination that text is unrecognized, the system may generate an unrecognized event. For example, an unrecognized event may result from not finding a match to text and/or the transcribed utterance.

[094] According to one embodiment of the invention, the system may implement one or more techniques to determine that textual messages are incorrectly recognized. For example, the user may command the system to play a specific song title and the system may misrecognize the requested song title and provide a song having a different title, the system may misrecognize an incorrect song title and

provide a song having a different than requested title, the system may misrecognize a correct song title and provide a song having an invalid song title, among other variations. When the system misrecognizes a request, the user typically provides immediate feedback such as, overriding the command in a time shorter than the expected execution time of the command, repeating the original request, issuing a stop command, or taking other action, wherein the action may be presented verbally, non-verbally by pushing a selected button on a cell phone or remote control, or both, among other configurations. According to one embodiment of the invention, the system may detect the user action and may prompt the user to re-phrase the request to enable the system to modify words in the query. The user's actions may be analyzed in a statistical model to determine a frequency occurrence of misrecognitions for particular commands, with the results being used to update the corresponding personalized cognitive model 810.

[095] According to another embodiment of the invention, the conversational speech analyzer 804 may access the personalized cognitive model 810 to proactively select a next best (or nth best) match for the received text. A match may be confirmed by user action that include the user not immediately canceling the command or taking other action. The misrecognitions may also be analyzed to potentially determine personalized tuning parameters for the speech recognition components of the system. For example, the system may, over time, tune the speech recognition engine to increase recognition by analyzing how the speech recognition engine misrecognizes the utterance.

[096] The event manager 100 may mediate interactions between other components of the invention. The event manager can provide a multi-threaded environment allowing the system to operate on multiple commands or questions from multiple user sessions without conflict and in an efficient manner, maintaining real-time response capabilities.

[097] Agents 106 may include a collection of grammars, criteria handlers, and algorithms that are accessible to respond to a set of requests and/or commands. Agents 106 may further contain packages of both generic and domain specific behavior for the system 98. Agents 106 may use nonvolatile storage for data, parameters, history information, and locally stored content provided in the system databases 102 or other local sources. One or more user profiles 110 may be provided

that include user specific data, parameters, and session and history information for determining the behavior of agents 106. One or more personality modules 108 may be provided in a data determining system that include personality characteristics for agents. The update manager 104 manages the automatic and manual loading and updating of agents 106 and their associated data from the Internet 146 or other network through the network interface 116.

[098] According to one embodiment of the invention, the speech-based interface for the system 90 may include one or more speech units 128. Speech units 128 may include one or more microphones, for example array microphone 134, to receive the utterances from the user. The speech received at the microphone 134 may be processed by filter 132 and passed to the speech coder 138 for encoding and compression. In one embodiment, a transceiver module 130 may transmit the coded speech to the main unit 98. Coded speech received from the main unit 98 is detected by the transceiver 130, then decoded and decompressed by the speech coder 138 and annunciated by the speaker 136.

[099] According to one embodiment of the invention, the non-speech-based interface for the system 90 may include one or more multi-modal devices 155 that may include mobile devices, stand alone or networked computers, personal digital assistances (PDAs), portable computer devices, or other multi-modal devices.

[0100] The speech units 128, multi-modal devices 155 and the main unit 98 may communicate over a communication link. The communication link may include a wired or wireless link. According to one embodiment, the communication link may comprise an RF link. The transceiver 130 on the speech unit may communicate coded speech data bi-directionally over the communication link with the transceiver 126 on the main unit 98. According to another embodiment, the RF link may use any standard local area wireless data protocols including the IEEE 802.11, Bluetooth or other standards. Alternatively, an infrared data link conforming to any suitable standard such as IrDA or other infrared standards can be used. In an alternative embodiment, wires may connect the speech unit 128 and the main unit 98, eliminating the need for one speech coder 138. Other wired or wireless analog or digital transmission techniques can be used.

[0101] According to one embodiment of the invention, coded speech received at the transceiver 126 on the main unit 98 is passed to the speech coder 122 for decoding

and decompression. The decoded speech may be processed by the speech recognition engine 120 using the context description grammar module 112, among other information. Any recognized information may be processed by the parser 118, which transforms information into complete algorithms and questions using data supplied by knowledge agents. Knowledge agents may be selected from the plurality of agents 106 using a grammar stack, wherein the knowledge agents provide information for generating a response to the question or command. The knowledge agents may then process the commands or questions by creating requests that are submitted to local databases 102 or submitted through the network interface 116 to external data sources over the Internet 146 or external other networks. Algorithms typically result in actions taken by the system 90 itself (i.e., pause or stop), or to a remote device or data source (i.e., download data or program, or control a remote device), through the network interface to the Internet or other data interface.

[0102] According to one embodiment of the invention, knowledge agents may return results of questions as responses to users. The responses may be created using the results of information requests, the system personality 108, the user preferences, other data in the user profile 110, and/or other information. Agents 106 may present these results using the speech unit 128. The agents 106 may create a response string, which is sent to the text to speech engine 124. The text to speech engine 124 may generate the utterances, which may be encoded and compressed by the speech coder 122. Once coded, the utterances are transmitted from the main unit 98 by the transceiver 126 to the transceiver 130 on the speech unit 128. The utterance is then decoded and decompressed by the speech coder 138 and output by the speaker 136. Alternatively, agents 106 may present the results using multi-modal devices 155.

[0103] The non-speech interface 114 may be part of, or separate from, the multi-modal devices 155 and may be used as a substitute or complement to the speech interface. For example, non-speech interface 114 may be used to present and interact with non-speech (e.g., graphical or tabular) information in a manner more easily understood by the user. According to one embodiment of the invention, multimodal support may be provided to maintain the context during both voice interaction and interaction through the non-speech interface 114. In one exemplary embodiment, a communication channel may be opened between multimodal devices 155 and the main user interface system 90 to allow multimodal devices 155 to input text

commands and questions. According to another embodiment of the invention, multimodal devices 155 may send a string of text or keywords for a command or question. The main interface system 90 may synchronize the context between multimodal device 155 and the speech units 128. In order to send a response to the corresponding device, the main user interface system 90 may track where the input came from so that the response may be sent to a TTS or multi-modal device 155.

[0104] According to one embodiment of the invention, non-speech interface 114 may show system, state and history information in a more concise manner than is possible through the speech interface. Non-speech interface 114 may be accessed to create or extend capabilities of agents 106. These operations may include scripting of agents, adding data to the agent or databases 102 used by the agent, adding links to information sources, among other operations.

[0105] According to another embodiment of the invention, system 90 may include different types of agents 106. In some embodiments of the invention, generic and domain specific behavior and information may be organized into domain agents 156. The system agent, on the other hand, may provide default functionality and basic services. The domain specific agents may provide complete, convenient and re-distributable packages or modules for each application area. In other words, the domain agents may include data that is needed to extend or modify the functionality of the system 90 in a current or new domain. Further, domain agents and their associated data can be updated remotely over a network as new behavior is added or new information becomes available. Domain agents may access a plurality of sources that may provide various services. Domain agents may use the services of other, typically more specialized, data managers and system agent. Agents are distributed and redistributed in a number of ways including on removable storage media, transfer over networks or attached to emails and other messages. The invention may provide license management capability allowing the sale of data managers by third parties to one or more users on a one time or subscription basis. In addition, users with particular expertise can create data managers, update existing data managers by adding new behaviors and information and making these data managers available to other users. A block diagram of the agent architecture according to an embodiment of the invention is shown in Figure 6.

[0106] Agents 106 may receive and return events to the event manager 100. Both system agents 150 and domain agents 156 may receive questions and commands from the parser 118. Based on keywords in the questions and commands and the structures of the questions and commands, the parser may invoke selected agents. Agents use the nonvolatile storage for data, parameters, history information and local content provided in the system databases 102.

[0107] According to one embodiment of the invention, when the system starts-up or boots-up, the agent manager 154 may load and initialize the system agent 150 and the one or more domain agents 156. Agent manager 154 includes knowledge of agents 106 and maps agents 106 to the agent library 158. At shutdown, the agent manager may unload the agents 106. The agent manager 154 also performs license management functions for the domain agents 156 and content in the databases 102.

[0108] The system agent 150 manages the criteria handlers 152 that handle specific parameters or values (criteria) used to determine context for questions and commands. According to one embodiment of the invention, criteria handlers 152 include parsing routines that are specialized to recognize particular parts of speech, such as times, locations, movie titles, and other parts of speech. The criteria handlers 152 may identify matching phrases and extract semantic attributes from the phrases. Both the system agent 150 and the domain agents 156 may use the criteria handlers 152. The various domain agents 156 may use the services of the system agent 150 and of other, typically more specialized, domain agents 156. The system agent 150 and the domain agents 156 may use the services of the agent library 158, which contains utilities for commonly used functions. According to one embodiment of the invention, the agent library may be a dynamic link library that implements one or more agents. The agent library may include utilities for text and string handling, network communications, database lookup and management, fuzzy and probabilistic evaluation, text to speech formats, and other utilities.

[0109] Domain agents 156 may be data-driven, scripted or created with compiled code. A base of generic agent may be used as the starting point for data-driven or scripted agents. Agents created with compiled code are typically built into dynamically linkable or loadable libraries. Developers of agents can add new functionality to the agent library 158 as required. Details of agent distribution and update, and agent creation or modification are discussed in sections below.

[0110] According to another embodiment of the invention, capabilities are provided to distribute and update system agents 150, domain agents 156, agent library 158 components, databases 102, and context description grammar 112 over wireless or wired networks 136, including IP networks and dial-up networks using the update manager 104. The network interface 116 may provide connections to one or more networks. The update manager 104 may also manage the downloading and installation of core system updates. The agent manager 154 may perform license management functions for the domain agents and the databases. The update manager 104 and agent manager 154 may perform these functions for all agents and database content including, agents and content available to all users or agents and/or content available to selected users. Examples of agent and database components added or updated on a periodic basis include:

- agents for new domains;
- additional domain knowledge for agents;
- new keywords for a domain, which can include names of politicians, athletes, entertainers, names of new movies or songs, etc. who have achieved recent prominence;
- links to a preferred set of information sources for the domains covered including links for, entertainment, news, sports, weather, etc.;
- updates to domain information based on, for example, changes to tax laws, company mergers, changing political boundaries;
- updates to content, including dictionaries, encyclopedias and almanacs; and
- other content and database components.

[0111] When a user requires or selects a new domain agent 156 or database element 102, the update manager 104 may connect to their source on the network 146 though the network interface 116, download and install the agent or data. To save system resources and to comply with any license conditions, the update manger 104 may uninstall agents that are no longer in use. In one embodiment of the invention, the update manager 104 may periodically query one or more sources of the licensed agents and database components to locate and download updates to agent executables, scripts or data as they become available. Alternatively, the agent sources may initiate

the downloading of agent updates of the registered or licensed agents to the update manager as they become available.

[0112] The agent manager 154 may provide license management clients that are capable of executing most any license terms and conditions. When a particular agent 106 and/or database element 102 is selected based on a submitted command, the agent manager 154 verifies that the use of the agent or data element is within the allowed terms and conditions, and if so, invokes the agent or allows access to the data element. License management schemes that can be implemented through the agent manager 154 include outright purchase, subscription for updates, one time or limited time use. Use of shared agents and data elements (such as those down-loaded from web sites maintained by groups of domain experts) may also be managed by the agent manager 154.

[0113] If questions or commands do not match an agent that is currently loaded on the system, the agent manager 154 may search the network 146 through the network interface 116 to find a source for a suitable agent. This process may be triggered, for example, when a query is made in a domain for which an agent is not available, when a new device is added to a mobile structure, or when the behavior of a mobile device is updated. Once located, the agent can be loaded under the control of the update manager 104, within the terms and conditions of the license agreement, as enforced by the agent manager.

[0114] New commands, keywords, information, or information sources can be added to any domain agent 156 by changing agent data or scripting. These configuration capabilities may allow users and content developers to extend and modify the behavior of existing domain agents 156 or to create new domain agents 156 from a generic agent without the need to create new compiled code. Thus, the modification of the domain agents 156 may range from minor data-driven updates by even the most casual users, such as specifying the spelling of words, to development of complex behavior using the scripting language as would typically be done by a domain expert. The user can create and manage modifications to domain agents 156 through speech interface commands or using non-speech interface 114. User-specific modifications of domain agents 156 are stored in conjunction with the user's profile 110 and accessed by the domain agent 156 at run-time.

[0115] The data used to configure data driven agents 156 may be structured in a manner to facilitate efficient evaluation and to help developers with organization. This data is used not only by the agent, but also in the speech recognition engine 120, the text to speech engine 124, and the parser 118. Examples of some major categories of data include:

1. Content packages may include questions or commands. Each command or question or group of commands or questions may include contexts used for creation of one or more requests. The domain agent 156 may pass a context description grammar expression to the parser 118 for evaluation of a context or question. An initial or default context is typically supplied for each command or question. The command or question includes a grammar for the management and evaluation of the context stack.
2. Parameters and other operating data on devices that are under control of the natural language speech interface. The domain agent 156 may use this data and parameters to determine how to execute a command, how to formulate the command string for the parser 118, determine if the command is feasible, and determine if the command can be executed within safety and operating limits.
3. Page lists or pointers to other local or network content sources. For each page or content source, there may be a pointer (e.g. URL, URI, or other pointer) to the page or source. Each page may have specific scraping information used to extract the data of interest. The scraping information may include, for example, matching patterns, HTML or other format parsing information.
4. A response list, determining the response of the domain agent 156 to a particular command or question given the context, the user profile and the information retrieved. Responses may include diagnostic error messages or requests for more information if the question or command cannot yet be resolved from the known information. Responses can be based on or be dependent on thresholds or probabilistic or fuzzy weights for the variables.
5. Substitution lists that include variable substitutions and transformations, often applied by the agents 150, 156 in the formatting of requests and results. For example, a stock domain specific agent 156 would

use a substitution list of company trading symbols, company names and commonly used abbreviations. Substitutions and transformations can be performed on commands and questions to create precise queries, which can be applied against one or more information sources or to results for creating more meaningful output to the user. Substitution lists also include information for optimally dealing with structured information, such as HTML formatted page parsing and evaluation.

6. Personalities used for responses. Personalities can be constructed by combining multiple traits in a weighted manner. Weights can be specified for each agent's domain area to create one or more specific personalities. Examples of personality traits include, sarcasm, humor, irritation, and sympathy, and other traits.

7. Public and user specific parameters for sources, substitutions, transformations, variables or criteria. The public parameter lists are part of the agent package 156. The user specific parameters are included in the user profile 110.

[0116] Commands and questions are interpreted, requests formulated, responses created and results presented can be based on the user's personal or user profile 110 values. Personal profiles may include information specific to the individual, their interests, their special use of terminology, the history of their interactions with the system, domains of interest, or other factors. The personal profile data can be used by the agents 106, the speech recognition engine 120, the text to speech engine 124, and the parser 118. Preferences can include, special (modified) commands, past behavior or history, questions, information sources, formats, reports, alerts or other preferences. User profile data can be manually entered by the user and/or can be learned by the system based on user behavior. User profile values can include: 1) spelling preferences; 2) date of birth for user, family and friends; 3) income level; 4) gender; 5) occupation; 6)location information such as, home address, neighborhood, and business address, paths traveled, locations visited; 7) vehicle type or types; 8) vehicle operator certifications, permits or special certificates; 9) history of commands and queries; 10) telecommunications and other service providers and services; 11) financial and investment information; 12) synonyms (i.e., a nick name for someone,

different terms for the same item); 13) special spelling; 14) keywords; 15) transformation or substitution variables; 16) domains of interest; or, 17) other values.

[0117] End users may use the data driven agent 156 extension and modification facilities and values stored in user profiles 110 to create special reports, packages of queries, alerts and output formats. A single alert or report can be configured to use multiple data sources and other variables (i.e., time, location, measured value) value to determine when alerts should be sent. For example, an alert can be generated by sampling a stock price every 15 minutes and sending an alert if the price drops below some value. In another example, an alert may be generated when a particular condition or combination of conditions occur for the vehicles. Alerts and reports can be directed to a local or remote output.

[0118] To create a report, the user may first specify a set of commands or questions. Next, the user can create or select a format for the report. Finally the user may name the report. A report can have variable parameters. For example, a user may create a company stock report and execute the report by stating its name and the company name, which gives the user selected information in a specified format for that company. In another example, a user can create a "morning" report, which presents selected multimedia information from different sources (news, sports, traffic, weather) in the order and formats desired. In yet another example, the user can create a report on the status of one or more vehicle systems. Alerts and reports can be created using only voice commands and responses, commands and responses through the graphical user interface 114, or a combination of the two. Reports can be run locally or remotely with respect to the vehicle. To create a report, alert, or other specialized behavior, the user performs a number of steps including: 1) specifying the command to run a report or alert; 2) specifying the question or questions, including keywords, used for a query; 3) setting the criteria for running the report such as on command or when a particular condition is met; 4) defining preferred information sources; 5)defining preferences for order of result evaluation by source, value, and other parameters; 6) specifying the presentation medium for a report or alert, such as an email, the text to speech engine, a message to a pager, or a text and graphics display; and, 7) specifying the preferred format for the report, such as information to be presented, order of information to be presented, preferred abbreviations or other variable substitutions.

[0119] Filtering and noise elimination may be important in facilitating the various functionalities of system 90 and may improve operation in noisy mobile environments. Recognition and parsing of the user's speech is implemented with good signal to noise ratio at the input to the speech recognition engine 120. To provide acceptable results, a set of acoustic models, an array microphone 134, a filter 132, or other components, may be employed. If a good signal to noise ratio cannot be achieved, a noise identification algorithm may be used and the appropriate acoustic model, for example, one that has been trained in conditions similar to the identified noise, may be selected. According to one embodiment of the invention, the microphone array, filters and speech coder 138 are physically separated from the main unit 98 into a speech unit 128, and connected using a wireless link. Since bandwidth on a wireless connection is at a premium, the speech coder dynamically adapts the digitization rate and compression of the captured speech.

[0120] Some embodiments of the invention may use one or more arrays of microphones 134 to provide better directional signal capture and noise elimination than can be achieved with a single microphone. The microphone array can be one-dimensional (a linear array) or two-dimensional (a circle, square, triangle or other suitable shape). The beam pattern of the array can be fixed or made adaptive though use of analog or digital phase shifting circuitry. The pattern of the active array is steered to point in the direction of the one or more users speaking. At the same time, nulls can be added to the pattern to notch out point or limited area noise sources. The use of the array microphone also helps reduce the cross talk between output from the text to speech engine 124 through the speaker 136 or from another user talking and detection of the user's speech.

[0121] The invention may use an analog or digital filter 132 between the array microphone or conventional microphone 134 and the speech coder 138. The pass band of the filter can be set to optimize the signal to noise ratio at the input to the speech recognition engine 120. In some embodiments, the filter is adaptive, using band shaping combined with notch filtering to reject narrow-band noise. One embodiment employs adaptive echo cancellation in the filter. The echo cancellation helps prevent cross talk between output from the text to speech engine and detection of the user's speech as well as suppression of environmentally caused echoes.

Algorithms comparing the background noise to the signal received from the users speech may be used to optimize the band-shaping parameters of the adaptive filter.

[0122] The speech received by the array microphone 134 and passed through the filter 132 may be sent to the speech digitizer or coder 138. The speech coder may use adaptive lossy audio compression to optimize bandwidth requirements for the transmission of the coded speech to the speech recognition engine 120 over a wireless link. The lossy coding is optimized to preserve only the components of the speech signal required for optimal recognition. Further, the lossy compression algorithms that may be used are designed to prevent even momentary gaps in the signal stream, which can cause errors in the speech recognition engine. The digitized speech may be buffered in the coder and the coder may adapt the output data rate to optimize the use of the available bandwidth. The use of the adaptive speech coder is particularly advantageous when a band-limited wireless link is used between the coder and the speech recognition engine.

[0123] The microphone can be complemented with an analog or digital (i.e., Voice over IP) speech interface. This interface allows a remote user to connect to the system and interact with it in the same manner possible if they were physically present.

[0124] In an alternative embodiment, the array microphone can be replaced by a set of physically distributed microphones or headsets worn by the users. The distributed microphones can be placed in different parts of a vehicle, different parts of a room or in different rooms of a building. The distributed microphones can create a three-dimensional array to improve signal to noise ratio. The headset can use a wireless or wired connection.

[0125] While the invention is intended to be able to accept most any natural language question or command, ambiguity may be a problem. To assist users formulate concise questions and commands, the system can support a voice query language. The language may be structured to allow a variety of queries and commands with minimal ambiguity. Thus, the voice query language helps users clearly specify the keywords or contexts of the question or command along with the parameters or criteria. The language can provide a grammar to clearly specify the keyword used to determine the context and present a set of one or criteria or parameters. A user asking a question or

stating a command in the voice query language may nearly always be guaranteed to receive a response.

[0126] The voice query language may be sensitive to the contents of the context stack, wherein a context defines a set of questions that can be activated or deactivated during a conversation. According to one embodiment, each agent may designate one context to be the root context that defines base algorithms that the agent implements. Thus, follow-on questions can be asked using an abbreviated grammar, since key words and criteria can be inherited from the context stack. For example, the user can simply ask about another keyword if the criteria of the question remain constant.

[0127] The system 90, may provide built in training capabilities to help users learn the best methods to formulate their questions and commands. The interactive training allows users to audibly or visibly see the machine interpretation of their queries and provides suggestions on how to better structure a query. Using the interactive training, users can quickly become comfortable with the voice query language and, at the same time, learn how to optimize the amount of information required with each step of a dialog.

[0128] The output of the speech coder 122 may be fed to the speech recognition engine 120. The speech recognition engine 120 recognizes words and phrases, using information in the context description grammar 112 and passes these to the parser 118 for interpretation. The speech recognition engine 120 may determine the user's identity by voice and name for each utterance. Recognized words and phrases may be tagged with this identity in all further processing. Thus, as multiple users engage in overlapping sessions, the tags added by the speech recognition engine 120 to each utterance allows other components of the system 90 to tie that utterance to the correct user and dialog. The user recognition capability may further be used as a security measure for applications, such as auctions or online shopping, where this is required. Voice characteristics of each user may be contained in the user profile 110.

[0129] According to one embodiment of the invention, users may start a dialog with the system 90 when they first address it. This can be done by speaking a generic word ("computer") or addressing a specific name ("Fred"), which may be generally tied to a system personality 108. Once the user starts the dialog, it may be recognized by the speech recognition engine 120, using unique characteristics of the user's speech. At

the end of a dialog or to interrupt a dialog, the user may utter a dismissal word ("good bye").

[0130] According to another embodiment of the invention, the system 90 may employ a speech recognition engine 124 that gains improved word recognition accuracy using data from context description grammar 112, user profiles 110, and the agents 106, among other components. At the same time, the fuzzy set possibilities or prior probabilities for the words in the context description grammar may be dynamically updated to maximize the probability of correct recognition at each stage of the dialog. The probabilities or possibilities may be dynamically updated based on a number of criteria including the application domain, the questions or commands, contexts, the user profile and preferences, user dialog history, the recognizer dictionary and phrase tables, and word spellings, among other criteria.

[0131] For uncommon words or new vocabulary words, the user may be given the option to spell the words. The spelling can be done by saying the names or the letters or using a phonetic alphabet. The phonetic alphabet can be a default one or one of the user's choosing.

[0132] Alternatively, when a user submits a word that is not recognized at all or is not correctly recognized by the speech recognition engine 120, then the user may be asked to spell the word. The speech recognition engine 120 determines this condition based on confidence level for the scoring process. The word may be looked up in a dictionary and the pronunciation for the word is added to either the dictionary, the agent 106, or the user's profile 110. The word pronunciation can then be associated with the domain, the question, the context and the user. Through this process, the speech recognition engine learns with time and improves in accuracy. To assist users in spelling words, an individualized phonetic alphabet can be used. Each user can modify the standard phonetic alphabets with words, which they can remember more easily.

[0133] Once the words and phrases have been recognized by the speech recognition engine 120, the tokens and user identification may be passed to the parser 118. The parser examines the tokens for the questions or commands, context and criteria. The parser may determine a context for an utterance by applying prior probabilities or fuzzy possibilities to keyword matching, user profile 110, dialog history, and context stack contents. The context of a question or command may determine the domain

and, thereby, the domain agent 156, if any, to be invoked. For example, a question with the keywords "temperature" implies a context value of weather for the question. Within a different dialog, the keyword "temperature" can imply a context for a measurement. The parser dynamically receives keyword and associated prior probability or fuzzy possibility updates from the system agent 150 or an already active domain agent 156. Based on these probabilities or possibilities the possible contexts are scored and the top one or few are used for further processing.

[0134] The parser 118 uses a scoring system to determine the most likely context or domain for a user's question and/or command. The score is determined from weighing a number of factors including, the user profile 110, the domain agent's data content and previous context. Based on this scoring, the system 90 invokes the correct agent. If the confidence level of the score is not high enough to ensure a reliable response, the system 90 may ask the user to verify whether the question and/or command is correctly understood.

[0135] In general, the question that is asked by the system 90 may be phrased to indicate the context of the question including all criteria or parameters. For example, the question can be in the form of: "Did I understand that you want such-and-such" If the user confirms that the question is correct the system proceeds to produce a response. Otherwise, the user can rephrase the original question, perhaps adding additional information to remove ambiguity, or the system can ask one or more questions to attempt to resolve the ambiguity.

[0136] Once the context for the question or command has been determined, the parser 118 can invoke the correct agent 156, 150. To formulate a question or command in the regular grammar used by agents, the parser will preferably determine required and optional values for the criteria or parameters. These criteria may have been explicitly supplied by the user or may need to be inferred. The parser may make use of the criteria handlers 152 supplied by the system agent. The criteria handlers can provide context sensitive procedures for extracting the criteria or parameters from the user's question or command. Some criteria may be determined by executing algorithms in the agent, while others may be determined by applying probabilistic or fuzzy reasoning to tables of possible values. Prior probabilities or fuzzy possibilities and associated values may be received from a number of sources including, for example, the history of the dialog, the user profile 110, and the agent. Based on user

responses, the prior probabilities or fuzzy possibilities may be updated as the system learns the desired behavior. For a weather context, examples of criteria include, location, date and time. Other criteria can include command criteria (i.e., yes/no, on/off, pause, stop), and spelling. Special criteria handlers are available from the system agent for processing lists, tables, barge-in commands, long strings of text and system commands.

[0137] The criteria handlers 152 can operate iteratively or recursively on the criteria extracted to eliminate ambiguity. This processing may help reduce the ambiguity in the user's question or command. For example, if the user has a place name (or other proper noun) in their utterance the parser 118 can use services of the domain agent 156 to look up tables in the databases 102 for place names or can attempt to determine which word is the proper noun from the syntax of the utterance. In another example, the user asks, "what about flight one hundred and twenty too?" The parser and domain agent use flight information in the database and network information along with context to determine the most plausible interpretation among; flight 100 and flight 20 also, flight 100 and flight 22, flight 122, etc.

[0138] Once the context and the criteria are determined, the parser 118 may form the question or command in a standard format or hierarchical data structure used for processing by the agents 150, 156. The parser 118 may fill in all required and some optional tokens for the grammar of the context. Often the tokens must be transformed to values and forms acceptable to the agents. The parser obtains the required transformations from the agents, dialog history or user profile 110. Examples of transformations or substitutions performed by the parser on tokens include: 1) substituting a stock symbol for a company name or abbreviation; 2) substituting a numerical value for a word or words; 3) adding a zip code to an address; and, 4) changing a place or other name to a commonly used standard abbreviation.

[0139] The agents 150, 156 may receive a command or question once the parser 118 has placed it in the required standard format. Based on the context, the parser can evoke the correct agent to process the question or command.

[0140] Commands can be directed to the system or to an external entity. System commands are generally directed to the system agent 150. Commands for external entities are generally processed by a domain agent 156, which includes the command context and behavior for the external entity.

[0141] Specific questions may be generally directed to one of the domain agents 156. The real-time selection of the correct agent allows the invention to dynamically switch contexts. Based on the question, command or context and the parameters or criteria, the domain agent may create one or more queries to one or more local or external information sources. Questions can be objective or subjective in nature. Results for objective questions can often be obtained by structured queries to one or more local or network information sources. Even for objective questions, the system may need to apply probabilistic or fuzzy set analysis to deal with cases of conflicting information or incomplete information. Information to answer subjective questions is generally obtained by one or more ad-hoc queries to local or network data sources, followed by probabilistic or fuzzy set evaluation of the one results to determine a best answer.

[0142] Once the domain agent 156 has formulated the one or more queries, they may be sent to local and/or network information sources. The queries may be performed in an asynchronous manner to account for the fact that sources respond at different speeds or may fail to respond at all. Duplicate queries can be sent to different information sources to ensure that at least one source responds with a useful result in a timely manner. Further, if multiple results are received in a timely manner, they can be scored by the system to determine which data is most reliable or appropriate. Examples of data sources accommodated include, HTTP data sources, sources with meta-data in various formats including XML, measurement data from sensors using various formats, device 32 setting parameters, entertainment audio, video and game files including MP3, databases using query languages and structured responses such as SQL, and other data sources.

[0143] The local information sources can be stored in one or more system databases 102 or can be on any local data storage such as a set of CDs or DVDs in a player or other local data storage. In other cases, local information can be obtained from vehicle system settings or measurement devices. Network information sources can be connected to, the control and device interfaces 30, the data interfaces 26, the Internet 42 or other network and accessed through a series of plug-ins or adaptors, known a pluggable sources, in the network interface 116. The pluggable sources are capable of executing the protocols and interpreting the data formats for the data sources of interest. The pluggable source provides information scrapping forms and procedures

for each source to the domain agents 156. If a new type of data source is to be used a new plug-in or adaptor can be added to the appropriate interface.

[0144] The domain agent 156 can evaluate the results of the one or more queries as they arrive. The domain agent may score the relevance of the results based on results already received, the context, the criteria, the history of the dialog, the user profile 110 and domain specific information using probabilistic or fuzzy scoring techniques. Part of the dialog history is maintained in a context stack. The weight of each context for the scoring may be based on the relevance of one context to another and the age of the contexts. Other scoring variables can be associated through the context stack. Contexts can also be exclusive, so that previous contexts have no weight in the scoring.

[0145] Based on the on-going scoring processes, the domain agent 156 may determine if a single best answer can be extracted. For most questions, the desired result may include a set of tokens that may be found to formulate an answer. Once a value has been found for each of these tokens, the results are ready for presentation to the user. For example, for a question on weather, the tokens can include the date, day of week, predicted high temperature, predicted low temperature, chance of precipitation, expected cloud cover, expected type of precipitation and other tokens. Results processed in this manner may include error messages. For subjective questions, this determination is made by determining a most likely answer or answers, extracted by matching of the results received. If no satisfactory answer can be inferred from the results of the query, the agent can do one of the following:

1. Ask the user for more information, typically through the speech interface, and based on the results obtained formulate new queries. This approach is applied when an irresolvable ambiguity arises in the formulation of a response.

2. Formulate new queries based on the results received from the first set of queries. This approach is typically applied in cases where the responses received do not include all the required information. Information sources to queries can be inferred from the results already obtained (i.e., links in an HTML document or measurements or settings from other devices 32) or from other sources. Using this approach one or more sets of queries and responses can be chained without the need for action by the user.

3. Wait for additional queries to return results.

[0146] In any case, the domain agent 156 may continue to make queries and evaluate results until a satisfactory response is constructed. In doing so, the agent can start several overlapping query paths or threads of inquiry, typically mediated by the event manager 100. This technique, combined with the use of asynchronous queries from multiple data sources provides the real-time response performance required for a natural interaction with the user.

[0147] The domain agent 156 may apply conditional scraping operations to each query response as it is received. The conditional scraping actions may depend on the context, the criteria, user profile 110, and domain agent coding and data. For each token to be extracted a scraping criteria 152 can be created using the services of the system agent 150. The scraping criteria may use format specific scraping methods including, tables, lists, text, and other methods. One or more scraping criteria can be applied to a page or results set. Once additional results are received, the domain agent can create new scraping criteria to apply to results already acquired. The conditional scarping process removes extraneous information, such as graphics, which need not be further processed or stored, improving system performance.

[0148] Specific commands are generally directed to one of the domain agents 156. The real-time selection of the correct agent allows the invention to dynamically switch contexts. Command oriented domain agents 156 evaluate the command and the state of vehicle systems, system capabilities, and measurements to determine if the command can be executed at all or if doing so will exceed operating or safety limits. If the command is ambiguous or cannot be executed for some other reason, the system may ask the user for more information or may suggest what the problem is and a likely approach to the solution. The domain agent may format the command for the specific device 32 and control and device interface 30. This formatting may involve variable substitution, inference of missing values and other formatting. Variable substitution and inference depends on the command context, the user profile 110, command history, state of vehicle systems and measured values, and other factors. A complex command can result in more atomic commands being sent to multiple devices, perhaps in a sequence. The sequence and nature of subsequent commands may depend on the previous commands, results of pervious commands, device settings and other measurements. As a command is executed, measurements are made

and results collected to determine if the execution was correct and the desired state or states were reached.

[0149] Once the domain agent 156 has created a satisfactory response to a question, or to a command, the agent may format that response for presentation. Typically, the domain agent can format the response into the markup format used by the text to speech engine 124. The domain agent may format the result presentation using available format templates and based on the context, the criteria, and the user profile 110. The domain agent may perform variable substitutions and transformations to produce a response best understood and most natural to the user. The domain agent may vary the order of presentation of tokens and the exact terminology used to create a more natural response to the user. The domain agent may also select the presentation personality 108 to be used.

[0150] For both command and query responses, the domain agent 156 may select the presentation template, determine order of presentation for tokens and determine variable substitutions and transformations using probabilistic or fuzzy set decision methods. The template used to form the presentation can be from the domain agent itself or from the user profile 110. The user profile can completely specify the presentation format or can be used to select and then modify an existing presentation format. Selection and formatting of presentation template can also depend on the presentation personality 108. At the same time, the characteristics of the personality used for the response are dynamically determined using probabilities or fuzzy possibilities derived from the context, the criteria, the domain agent itself and the user profile 110.

[0151] The domain agent 156 may apply a number of transformations to the tokens before presentation to the user. These variable substitutions and transformations may be derived from a number of sources including, domain information carried by the agent, the context, the token values, the criteria, the personality 108 to be used, and the user profile 110. Examples of variable substitutions and transformations include: 1) substitution of words for numbers; 2) substitution of names for acronyms or symbols (i.e., trading symbols); 3) use of formatting information derived from the information sources (i.e., HTML tags); 4) nature of the response including, text, long text, list, table; 5) possible missing information or errors; 6) units for measurement

(i.e., English or metric); and, 7) preferred terminology from the user profile or presentation personality 108.

[0152] The invention may provide special purpose presentation capabilities for long text strings, tables, lists and other large results sets. Domain agents 156 may use special formatting templates for such results. The system agent 150 can provide special criteria handlers 152 for presentation and user commands for large results sets. The presentation templates used by the domain agents for large results sets typically include methods for summarizing the results and then allowing the user to query the result in more detail. For example, initially only short summaries, such as headlines or key numbers, are presented. The user can then query the results set further. The criteria handlers provide users with the capability to browse large results sets. Commands provided by the criteria handlers for large results sets include, stop, pause, skip, rewind, start, and forward.

[0153] Some information, in formats such as video, pictures and graphics, may be best presented in a displayed format. The domain agents 156 apply suitable presentation templates in these cases and present the information through the non-speech interface 114. The system agent 150 provides special criteria handlers 152 for presentation and user commands for display presentation and control.

[0154] Although particular embodiments of the invention have been shown and described, it will be understood that it is not intended to limit the invention to the embodiments that are disclosed and it will be obvious to those skilled in the art that various changes and modifications may be made without departing from the spirit and scope of the invention. Thus, the invention is intended to cover alternatives, modifications, and equivalents, which may be included within the spirit and scope of the invention as defined by the claims.

We claim:

1. A mobile device for processing speech and non-speech communications, comprising:

an identification module that identifies a user that communicates with the mobile device;

a matching module that matches the identified user to one or more profiles;

a transcription module that transcribes the speech and the non-speech communications to create a speech-based textual message and a non-speech-based textual message, wherein the transcription module accesses the one or more profiles;

a merging module that merges the speech-based textual message and the non-speech-based textual message to generate a query;

a search module that searches the query for text combinations;

a comparison module that compares the text combinations to entries in a context description grammar;

a plurality of domain agents that are associated with the context description grammar;

a scoring module that provides relevance scores based on results from the comparison module;

a domain agent selector that selects domain agents based on results from the scoring module; and

a response generating module that communicates with the selected domain agents to obtain content that is gathered by the selected domain agents and that generates a response from the content, wherein the content is arranged in a selected order based on results from the scoring module.

2. The mobile device according to claim 1, wherein the response generating module generates an aggregate response that includes the content gathered by the selected domain agents.

3. The mobile device according to claim 1, wherein the identification module supports multiple users that access the mobile device during a same session.

4. The mobile device according to claim 3, wherein the session may include an overlapping session or an interleaved session.

5. The mobile device according to claim 3, wherein the identification module verifies an identify of the multiple users based on voiceprint matching or password matching.

6. The mobile device according to claim 4, wherein, during the interleaved session, the mobile device processes one or more queries in an order of receipt.

7. The mobile device according to claim 1, wherein, during the interleaved session, the mobile device processes one or more queries based on an order determined by a length of the queries.

8. The mobile device according to claim 1, wherein the mobile device includes (i) a personal digital assistant, (ii) a cellular telephone, (iii) a portable computer, (iv) a desktop computer, or any combination of (i) to (iv).

9. The mobile device according to claim 1, wherein the mobile device receives follow-up speech and non-speech communications and wherein the transcription module transcribes the follow-up speech and non-speech communications to create a follow-up speech-based textual message and a follow-up non-speech-based textual message.

10. The mobile device according to claim 9, wherein the merging module merges the follow-up speech-based textual message and the follow-up non-speech-based textual message to generate a follow-up query.

11. The mobile device according to claim 1, further comprising a personality module that facilitates formatting the response.

12. The mobile device according to claim 1, wherein the mobile device is configured to communicate with a context stack that includes one or more contexts that are selected based on the query.

13. The mobile device according to claim 12, wherein the scoring module determines the one or more contexts based on at least applying prior probabilities or fuzzy possibilities to (i) keyword matching, (ii) user profiles, (iii) a dialog history, or any combination of (i) to (iii).

14. The mobile device according to claim 1, wherein at least one of the domain agents creates and directs a request to a local information source or a network information source.

15. The mobile device according to claim 1, wherein at least one of the domain agents creates and directs a command to a remote or local device.

16. The mobile device according to claim 10, wherein the follow-up query is associated with a same context as the query.

17. The mobile device according to claim 1, wherein the mobile device supports Internet browsing and wherein the generated query is used to conduct the Internet browsing.

18. The mobile device according to claim 1, wherein the mobile device supports multimodal communications.

19. The mobile device, comprising:

an object that is accessible by the mobile device;

an input device that receives speech and non-speech annotations corresponding to the object;

a transcription module that transcribes the speech and the non-speech annotations to create a speech-based textual message and a non-speech-based textual message; and

an association module that associates the object with the speech-based textual message and the non-speech-based textual message.

20. The mobile device according to claim 19, further comprising:

a classification module that classifies the speech-based textual message and the non-speech-based textual message; and

a searching module that searches the speech-based textual message and the non-speech-based textual message based on a search request.

21. The mobile device according to claim 19, wherein the association module associates the object with the speech and non-speech annotations.

22. The mobile device according to claim 21, further comprising:

a classification module that classifies the speech and non-speech annotations; and

a searching module that searches the speech and non-speech annotations based on a search request.

23. The mobile device according to claim 19, wherein the object includes digital media.

24. The mobile device according to claim 23, wherein the object includes a digital photograph, a calendar entry, an email message, an instant message, a phonebook entry, a voicemail entry or a digital movie.

25. The mobile device according to claim 19, wherein the association module associates the object to the speech and non-speech annotations, the speech-based textual message, and the non-speech-based textual message.

26. The mobile device according to claim 25, wherein the object is retrievable using (i) a voice command, (i) a text command, or a combination of (i) and (ii).

27. The mobile device according to claim 19, wherein the object is stored (i) in the mobile device, (ii) at a server, or a combination of (i) and (ii).

28. The mobile device according to claim 19, wherein the object includes metadata.

29. The mobile device according to claim 28, wherein the metadata includes global positioning system information, environmental information, or geographic information.

30. The mobile device according to claim 28, further comprising a metadata searching module that enables searching of the metadata.

31. A system, comprising:

a plurality of mobile devices that support multimodal communications, wherein two or more mobile devices are associated with a user; and

a context module that is coupled to the plurality of mobile devices, wherein a same context is maintained at the two or more mobile devices that are associated with the user.

32. The system according to claim 31, wherein the plurality of mobile devices comprise:

an identification module that identifies a user that communicates with the mobile device;

a matching module that matches the identified user to one or more profiles;

a transcription module that transcribes the speech and the non-speech communications to create a speech-based textual message and a non-speech-based textual message, wherein the transcription module accesses the one or more profiles;

a merging module that merges the speech-based textual message and the non-speech-based textual message to generate a query;

a search module that searches the query for text combinations;

a comparison module that compares the text combinations to entries in a context description grammar;

a plurality of domain agents that are associated with the context description grammar;

a scoring module that provides relevance scores based on results from the comparison module;

a domain agent selector that selects domain agents based on results from the scoring module; and

a response generating module that communicates with the selected domain agents to obtain content that is gathered by the selected domain agents and that generates a response from the content, wherein the content is arranged in a selected order based on results from the scoring module.

33. The system according to claim 31, wherein the context module is configured to broadcast the same information to two or more mobile devices and to receive a response from one of the mobile devices.

34. The system according to claim 31, wherein the context module broadcasts a first type of information and receives a second type of information.

35. The system according to claim 34, wherein first type of information and the second type of information include speech information or non-speech information.

36. The system according to claim 35, wherein the first type of information includes non-speech information that is presented on a graphical user interface and the second type of information includes speech information that is applied to search a network.

37. The system according to claim 36, wherein the first type of information and the second type of information correspond to the same context.

38. A method of processing speech and non-speech communications in a multimodal environment, comprising:

receiving the speech and non-speech communications;

identifying a user that provides the speech and non-speech communications;

matching the identified user to one or more profiles;

transcribing the speech and non-speech communications to create a speech-based textual message and a non-speech-based textual message, wherein the transcribing includes accessing the one or more profiles;

merging the speech-based textual message and the non-speech-based textual message to generate a query;

searching the query for text combinations;

comparing the text combinations to entries in a context description grammar;

accessing a plurality of domain agents that are associated with the context description grammar;

generating a relevance score based on results from comparing the text combinations to entries in the context description grammar;

selecting one or more domain agents based on results from the relevance score;

obtaining content that is gathered by the selected domain agents; and

generating a response from the content, wherein the content is arranged in a selected order based on results from the relevance score.

39. The method according to claim 38, further comprising generating an aggregate response that includes the content that is gathered by the selected domain agents.

40. The method according to claim 38, further comprising:

receiving a follow-up speech and non-speech communications;

transcribing the follow-up speech and non-speech communications to create a follow-up speech-based textual message and a follow-up non-speech-based textual message; and

merging the follow-up speech-based textual message and the follow-up non-speech-based textual message to generate a follow-up query.

41. The method according to claim 38, further comprising a personality module that communicates the response to a user.

42. The method according to claim 38, wherein the system supports multiple users during a same session.

43. The method according to claim 42, wherein the session may include an overlapping session or an interleaved session.

44. The method according to claim 42, wherein the multiple users are identified based on voiceprint matching or password matching.

45. The method according to claim 43, further comprising processing one or more queries in an order of receipt.

46. The method according to claim 43, further comprising processing one or more queries based on an order determined by a length of the queries.

47. The method according to claim 38, further comprising generating a context stack that includes one or more contexts that are selected based on the query.

48. The method according to claim 47, wherein the one or more contexts are generated based on applying prior probabilities or fuzzy possibilities to (i) keyword matching, (ii) user profiles, (iii) a dialog history, or any combination of (i) to (iii).

49. The method according to claim 38, further comprising conducting Internet browsing with the generated query.

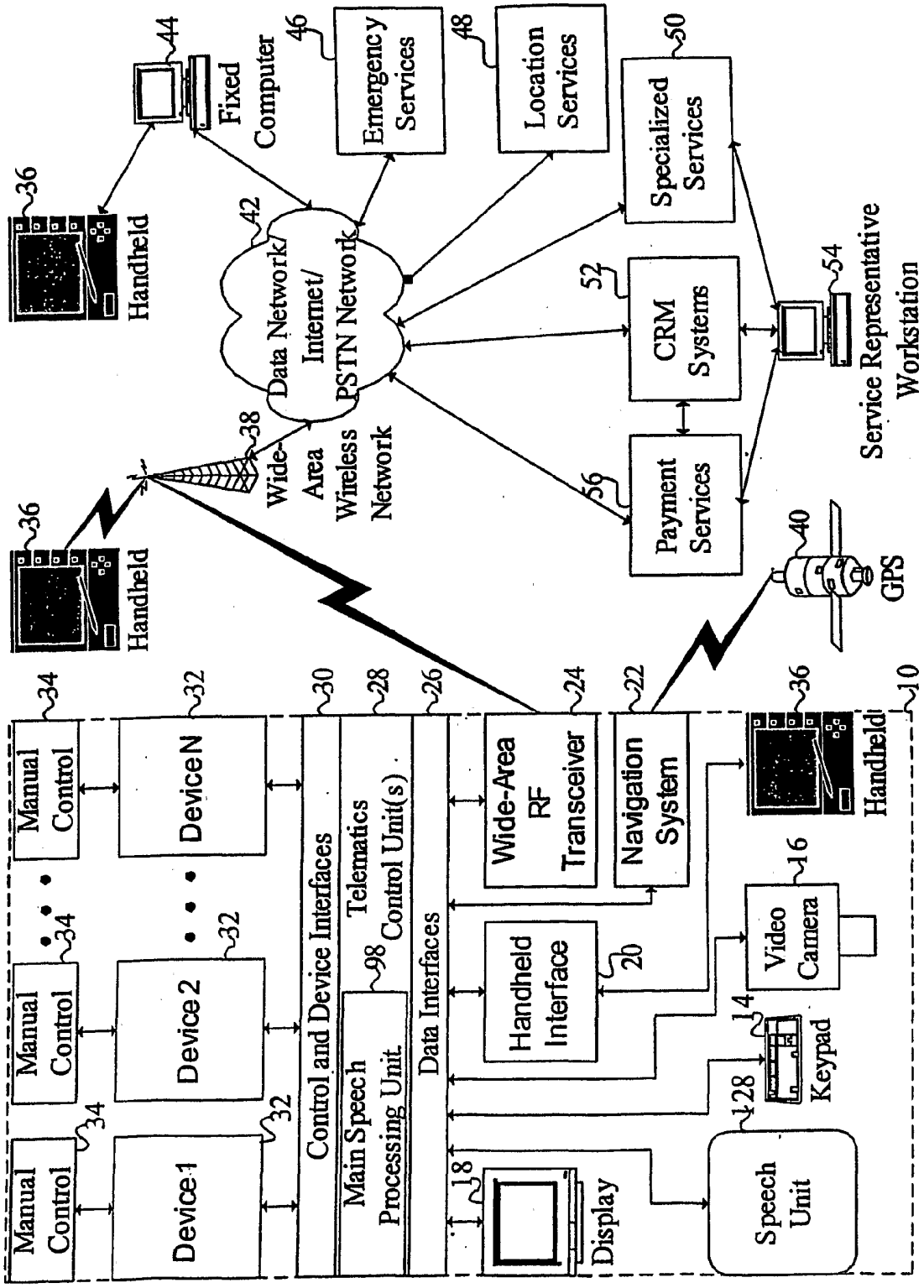

Manual Control
Device 1
Device 2
DeviceN
Control and Device Interfaces
Main Speech Processing Unit
Telematics Control Unit(s)
Data Interfaces
Wide-Area RF Transceiver
Navigation System
Handheld Interface
Display
Speech Unit
Keypad
Video Camera
Handheld
GPS
Payment Services
CRM Systems
Service Representative Workstation
Specialized Services
Location Services
Emergency Services
Fixed Computer
Data Network/ Internet/ PSTN Network
Wide-Area Wireless Network
Handheld
10
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
52
54
56
98
128

Figure 1

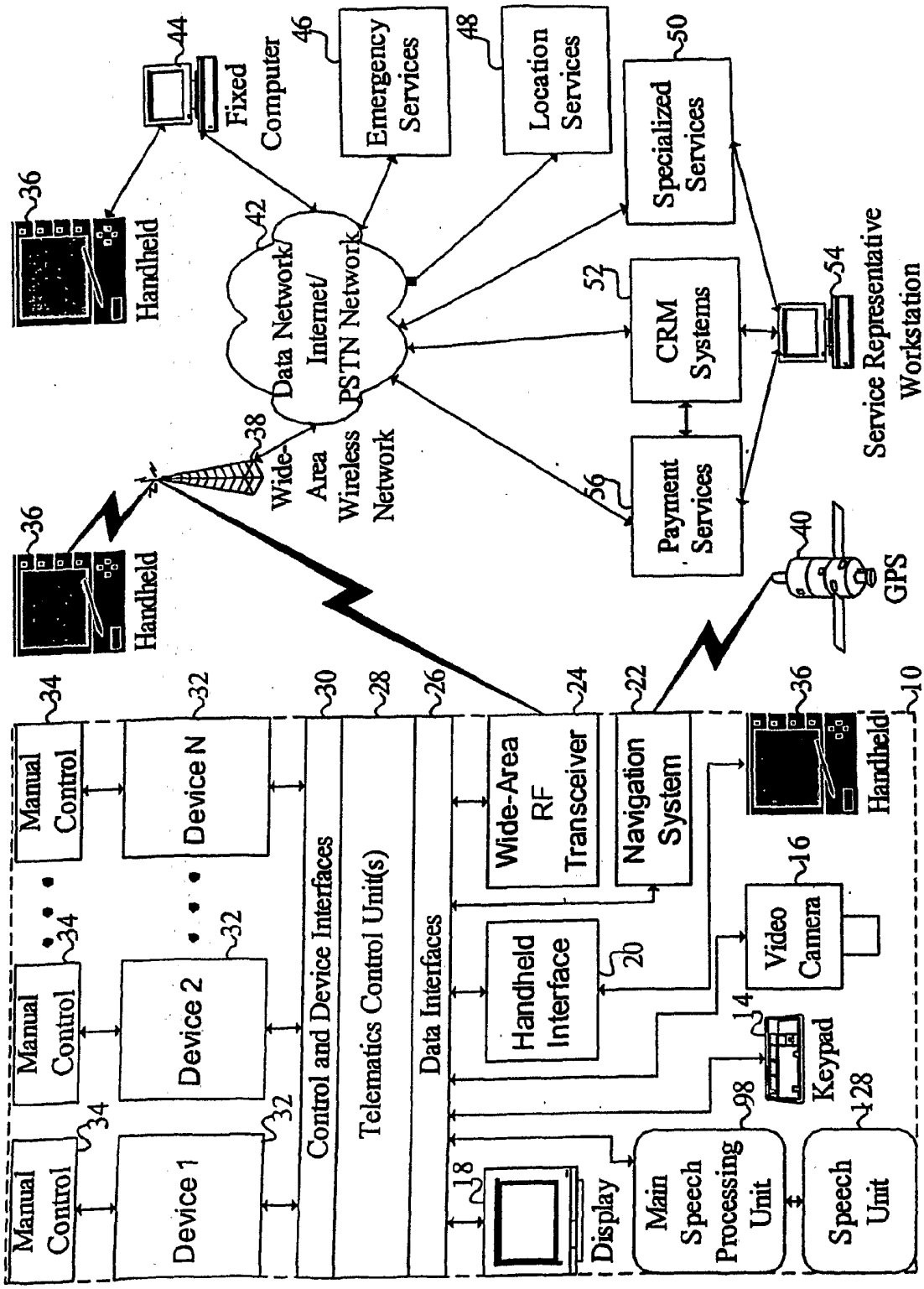

Manual Control
Device 1
Manual Control
Device 2
Manual Control
Device N
Control and Device Interfaces
Telematics Control Unit(s)
Data Interfaces
Wide-Area RF Transceiver
Navigation System
Handheld Interface
Video Camera
Keypad
Display
Main Speech Processing Unit
Speech Unit
Handheld
Handheld
Handheld
Wide-Area Wireless Network
Data Network/ Internet/ PSTN Network
Fixed Computer
Emergency Services
Location Services
Specialized Services
CRM Systems
Payment Services
Service Representative Workstation
GPS

Figure 2

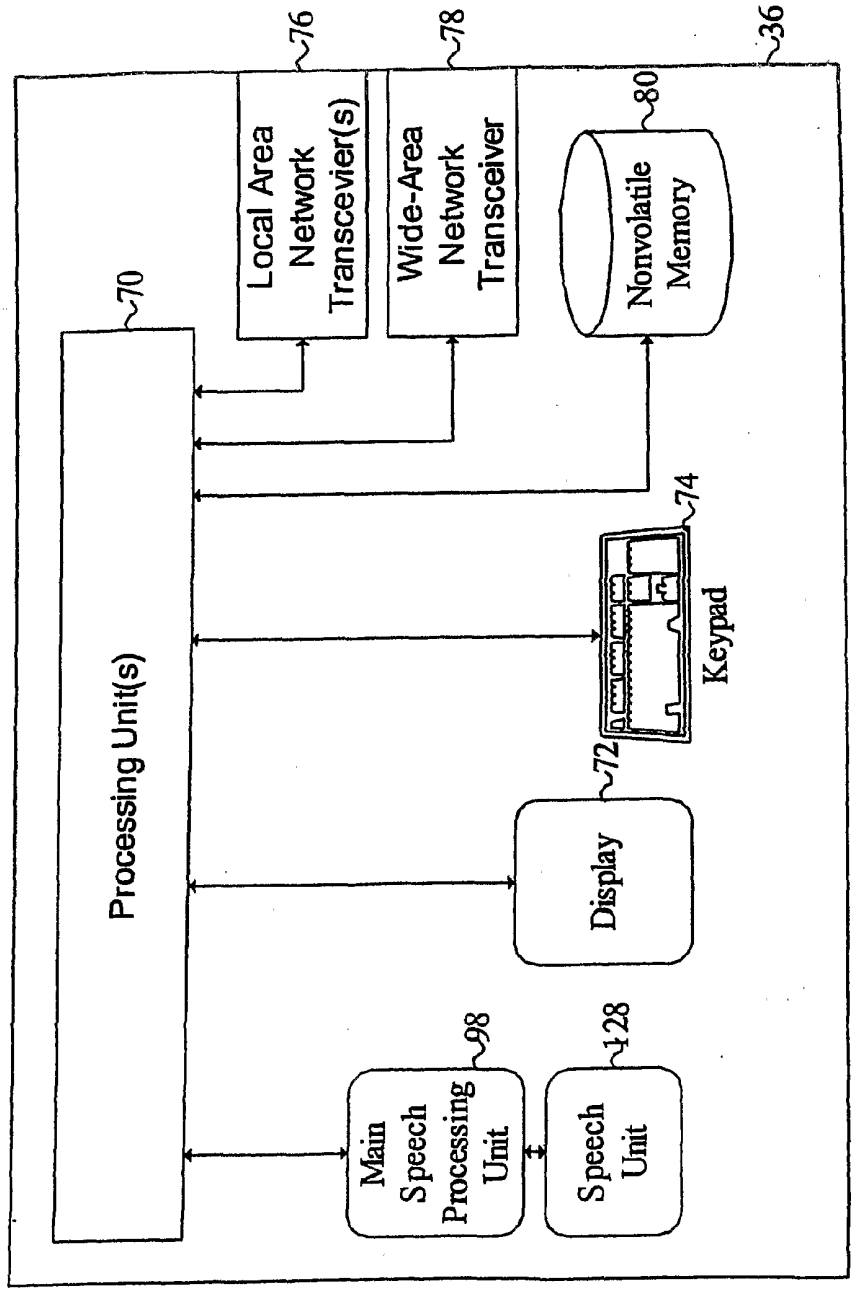

Processing Unit(s)
70
Local Area Network Transcevier(s)
76
Wide-Area Network Transceiver
78
Nonvolatile Memory
80
36
74
Keypad
72
Display
Main Speech Processing Unit
98
Speech Unit
128

Figure 3.

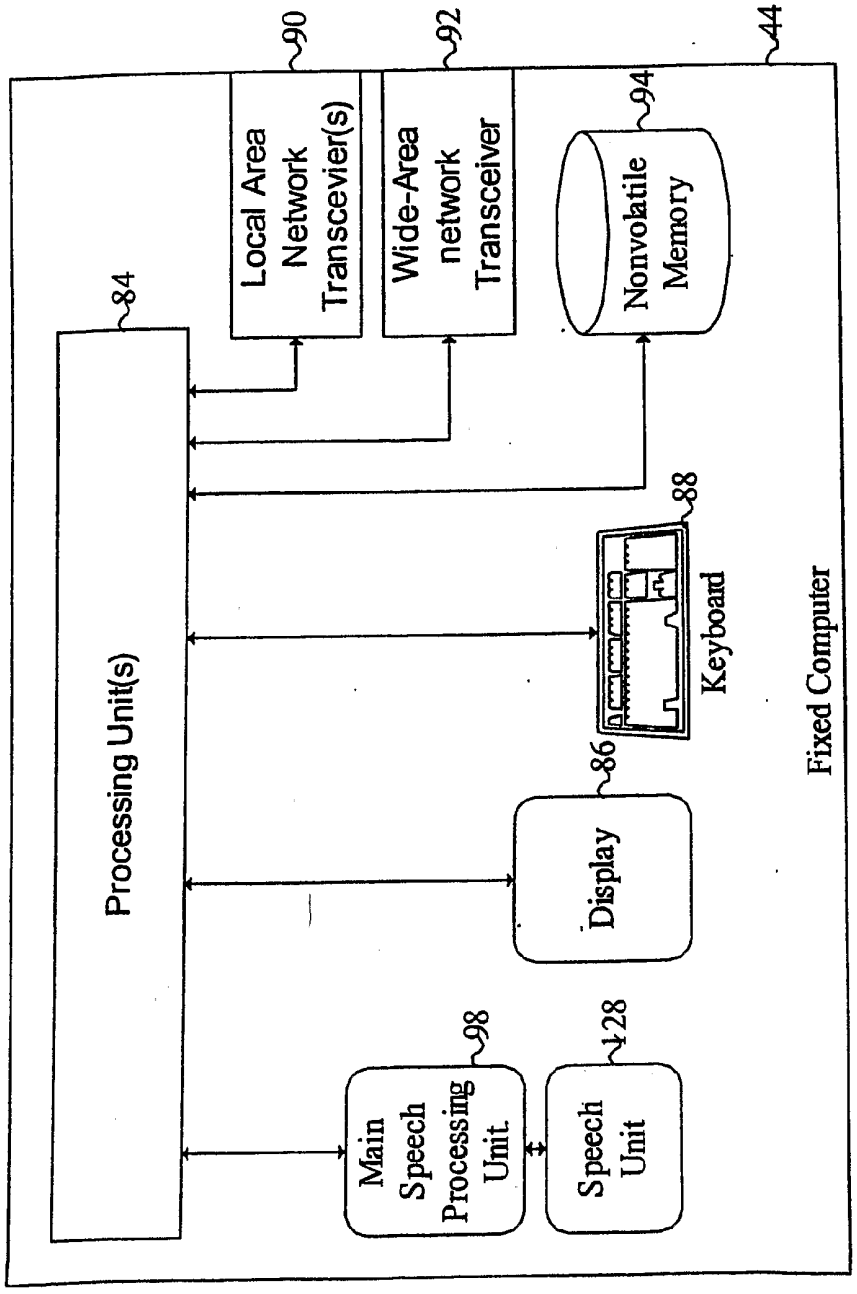

Fixed Computer
44
Local Area Network Transcevier(s)
90
Wide-Area network Transceiver
92
Nonvolatile Memory
94
Processing Unit(s)
84
Keyboard
88
Display
86
Main Speech Processing Unit
98
Speech Unit
128

Figure 4

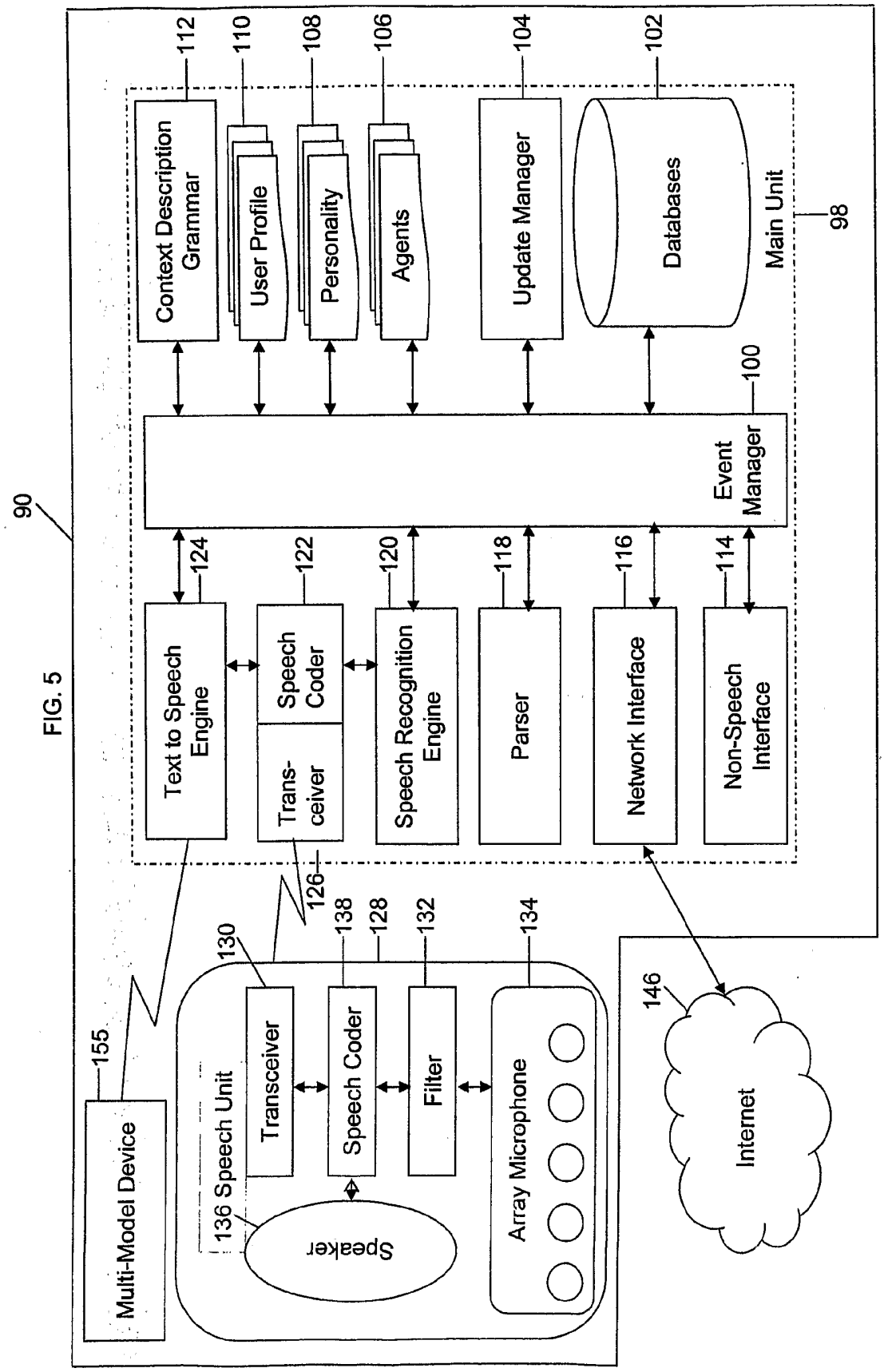

90
Main Unit
98
Context Description Grammar
112
User Profile
110
Personality
108
Agents
106
Update Manager
104
Databases
102
Event Manager
100
Text to Speech Engine
124
Speech Coder
Trans-ceiver
122
Speech Recognition Engine
120
Parser
118
Network Interface
116
Non-Speech Interface
114
126
Multi-Model Device
155
136 Speech Unit
130
Transceiver
138
Speech Coder
128
Filter
132
Array Microphone
134
Speaker
Internet
146

FIG. 5

FIG. 6

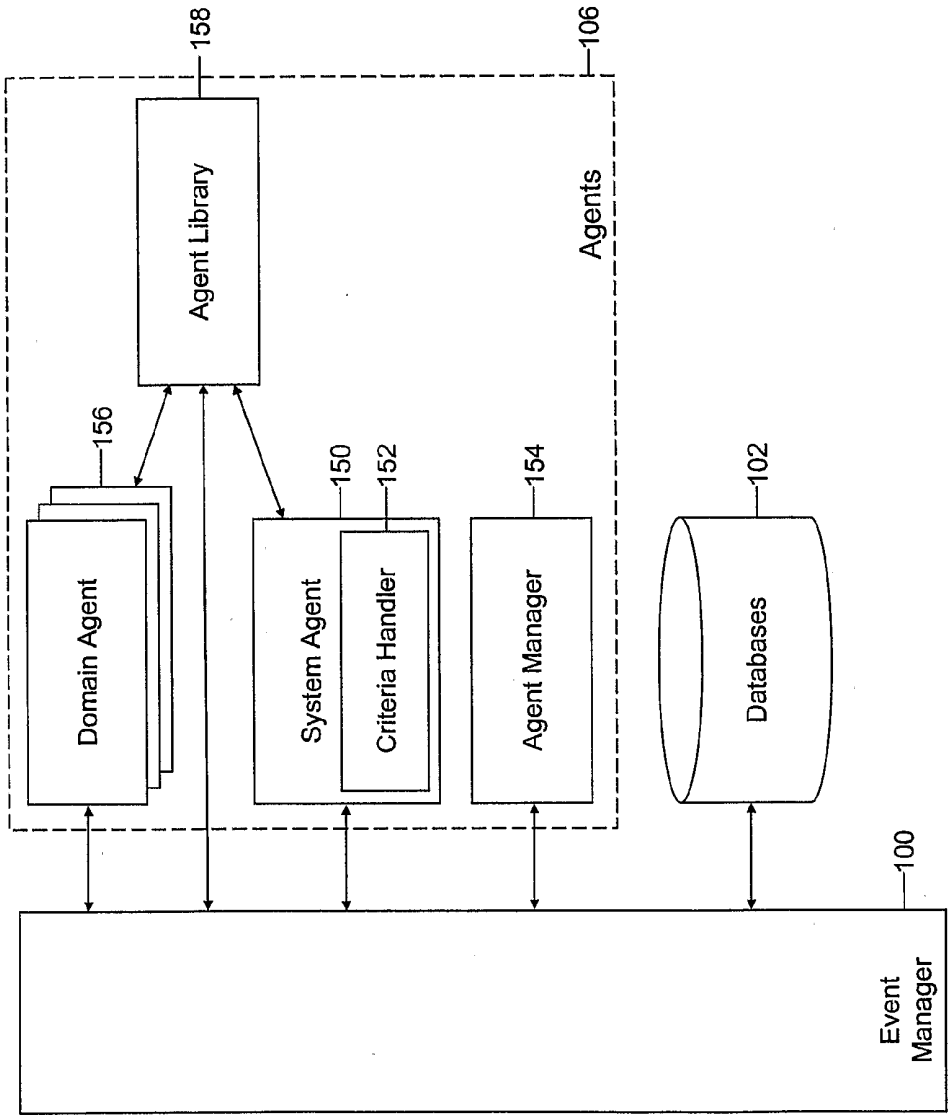
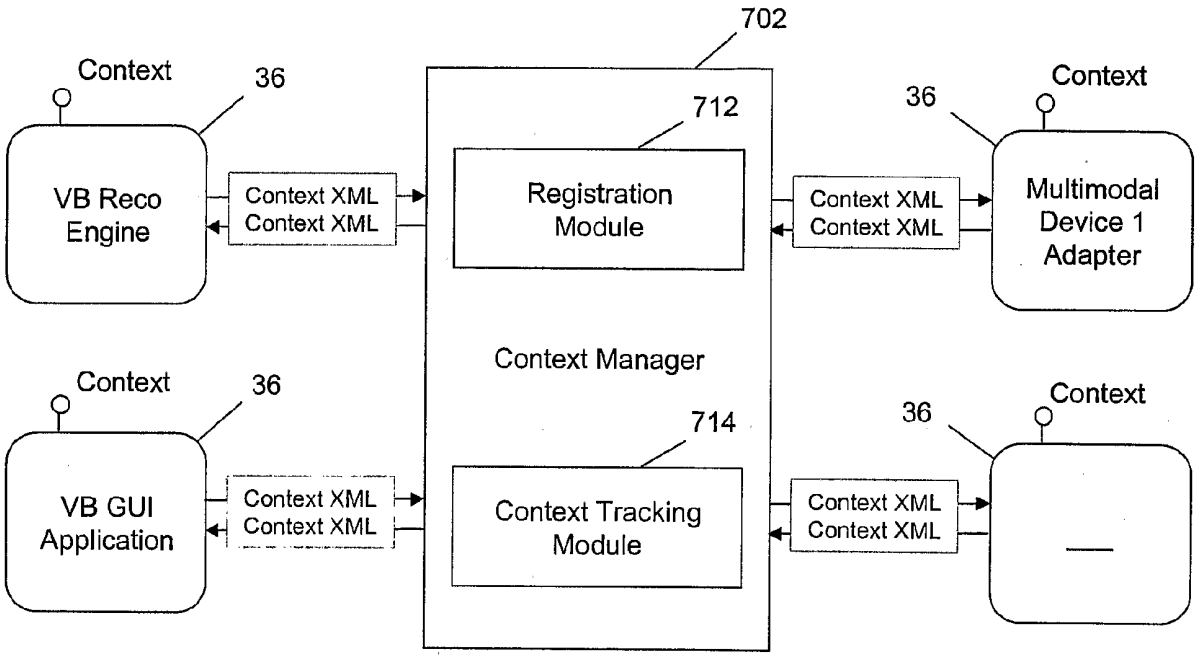

158
106
Agent Library
Agents
156
150
152
154
102
Domain Agent
System Agent
Criteria Handler
Agent Manager
Databases
100
Event Manager

FIG. 7


700
702
Context
36
VB Reco Engine
Context XML
Context XML
712
Registration Module
Context XML
Context XML
36
Context
Multimodal Device 1 Adapter
Context Manager
Context
36
VB GUI Application
Context XML
Context XML
714
Context Tracking Module
Context XML
Context XML
36
Context

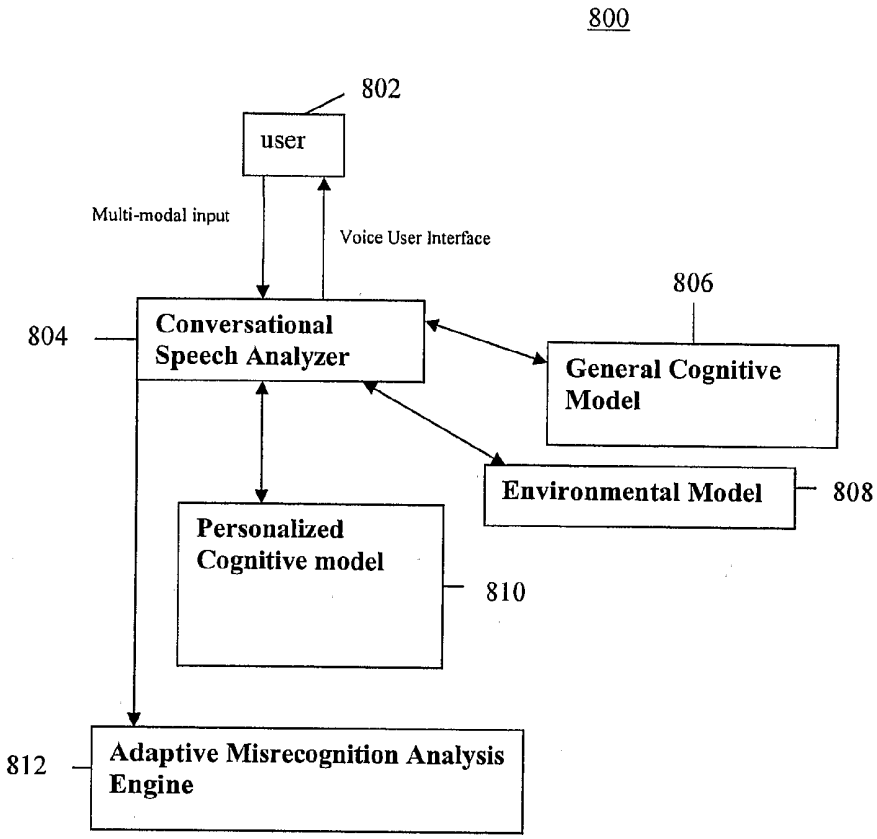

800
802
user
Multi-modal input
Voice User Interface
806
804
Conversational Speech Analyzer
General Cognitive Model
Environmental Model
808
Personalized Cognitive model
810
812
Adaptive Misrecognition Analysis Engine

FIGURE 8

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 March 2007 (08.03.2007)

PCT

(10) International Publication Number
WO 2007/027989 A2

(51) International Patent Classification:
G10L 15/18 (2006.01)

(21) International Application Number:
PCT/US2006/034184

(22) International Filing Date: 31 August 2006 (31.08.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/712,412 31 August 2005 (31.08.2005) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [US/US]; 10505 NE 38th Place, Building 9, Kirkland, WA 98033 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **DI CRISTO, Philippe** [FR/US]; 177 107th Avenue, N.E., #1007, Bellevue, WA 98004 (US). **KE, Min** [US/US]; 3221 201st Place, S.E., Bothell, WA 98012 (US). **KENNEWICK, Robert, A.** [US/US]; 2204 S. Eddy St., Seattle, WA 98108 (US). **TJALVE, Michael** [DK/US]; 14904 SE 47th Court, Bellevue, WA 98006 (US).

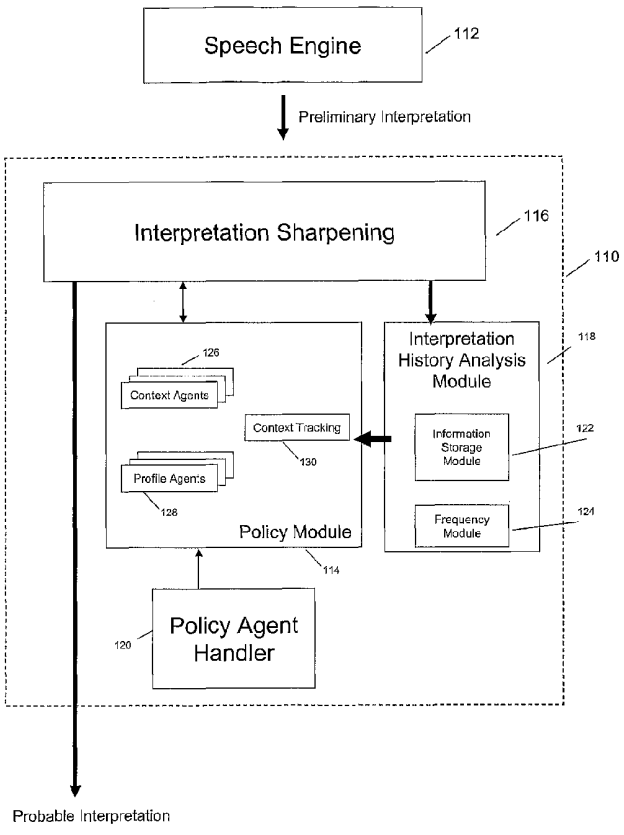
(74) Agents: **GATTO, James, G.** et al.; PILLSBURY WINTHROP SHAW PITTMAN LLP, P.O. Box 10500, Mclean, VA 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI,

[Continued on next page]

(54) Title: DYNAMIC SPEECH SHARPENING



(57) Abstract: An enhanced system for speech interpretation is provided. The system may include receiving a user verbalization and generating one or more preliminary interpretations of the verbalization by identifying one or more phonemes in the verbalization. An acoustic grammar may be used to map the phonemes to syllables or words, and the acoustic grammar may include one or more linking elements to reduce a search space associated with the grammar. The preliminary interpretations may be subject to various post-processing techniques to sharpen accuracy of the preliminary interpretation. A heuristic model may assign weights to various parameters based on a context, a user profile, or other domain knowledge. A probable interpretation may be identified based on a confidence score for each of a set of candidate interpretations generated by the heuristic model. The model may be augmented or updated based on various information associated with the interpretation of the verbalization.

WO 2007/027989 A2

FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

DYNAMIC SPEECH SHARPENING

CROSS REFERENCE TO RELATED APPLICATIONS

[001] This application claims the benefit of provisional U.S. Patent Application Serial No. 60/712,412, entitled "Dynamic Speech Sharpening," filed August 31, 2005, and which is hereby incorporated by reference in its entirety.

[002] This application is related to the following co-pending applications, which are hereby incorporated by reference in their entirety: "Systems and Methods for Responding to Natural Language Speech Utterance," Ser. No. 10/452,147, filed June 3, 2003, and published as US Patent Application Publication No. US 2004/0044516; "Mobile Systems and Methods for Responding to Natural Language Speech Utterance," Ser. No. 10/618,633, filed July 15, 2003, and published as US Patent Application Publication No. US 2004/0193420; "Systems and Methods for Responding to Natural Language Speech Utterance," Ser. No. 11/197,504, filed August 5, 2005; "System and Method of Supporting Adaptive Misrecognition in Conversational Speech," Ser. No. 11/200,164, filed August 10, 2005; and "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," Ser. No. 11/212,693, filed August 29, 2005.

FIELD OF THE INVENTION

[003] The invention is related generally to automated speech interpretation, and in particular, to enhancing the accuracy and performance of speech interpretation engines.

BACKGROUND OF THE INVENTION

[004] The field of automated speech interpretation is in increasingly higher demand. One use of automated speech interpretation is to provide voice requests to electronic devices. This may enable a user to simply speak to an electronic device rather than manually inputting requests, or other information, through pressing buttons, uploading information, or by other request input methods. Controlling various electronic devices through speech may enable the user to use the electronic devices more efficiently.

[005] However, existing technology in the field of automated speech interpretation, such as standard speech engines, automatic speech recognition (ASR), and other systems for interpreting speech, are unable to process a speech signal in an efficient manner, often constructing large grammars that include a large number of items,

nodes, and transitions, which is a concern particularly for large-list recognition for embedded applications. If the grammar for an embedded application grows too much, it may not fit within the constrained space of an embedded application. With limited CPU power, response time and performance is easily affected due to the significant time needed to compile and load the grammar. Response time is further degraded because the speech engine has to parse through a large number of transition states to come up with a recognition result. Even when the speech engine is able recognize a word, the results are often unreliable because large grammars introduce greater risk of confusion between items as the size of the grammar increases. Existing techniques focus on reducing the size of a grammar tree by removing command variants or criteria items, but this approach strips functionality from the application.

[006] In addition to the performance problems associated with speech recognition engines that employ large word grammars, existing speech processing engines are unable to interpret natural human speech with a suitable accuracy to sufficiently control some electronic devices. In particular, speech interpretation engines still have substantial problems with accuracy and interpreting words that are not defined in a predetermined vocabulary or grammar context. Poor quality microphones, extraneous noises, unclear or grammatically incorrect speech by the user, or an accent of the user may also cause shortcomings in accuracy, such as when a particular sound cannot be mapped to a word in the grammar.

[007] In light of these and other problems, there is a need for enhanced automated speech interpretation that may interpret natural human speech with an augmented accuracy.

SUMMARY OF THE INVENTION

[008] According to one aspect of the invention, a system for enhancing automated speech interpretation is provided. The system may include a set of techniques for use in a speech-to-text engine to enhance accuracy and performance, for example, by reducing the search space of the speech engine. The problems with large-list recognition for embedded applications may also be improved by using phonetic dictation, which may recognize a phoneme string by disregarding the notion of words. The system may also use one or more post-processing techniques to sharpen an output of a preliminary speech interpretation made by a speech engine. The system may be modeled at least partially after one or more speech pattern

recognition techniques used by humans, such as interpreting speech using words, word sequences, word combinations, word positions, context, phonetic similarities between two or more words, parts of speech, or other techniques.

[009] In one implementation of the invention, the system may receive a verbalization made by a user, where a speech engine may receive the verbalization. The speech engine may output information relating to a plurality of preliminary interpretations of the verbalization, where the plurality of preliminary interpretations represent a set of best guesses at the user verbalization. According to one aspect of the invention, the performance of the speech engine may be improved by using phoneme recognition. Phoneme recognition may disregard the notion of words, instead interpreting a verbalization as a series of phonemes, which may provide out-of-vocabulary (OOV) capabilities, such as when a user misspeaks or an electronic capture devices drops part of a speech signal, or for large-list applications, such as city and street names or song titles, for example. Phoneme recognition may be based on any suitable acoustic grammar that maps a speech signal into a phonemic representation. For example, the English language may be broken down into a detailed grammar of the phonotactic rules of the English language. Portions of a word may be represented by a syllable, which may be further broken down into core components of an onset, a nucleus, and a coda, which may be further broken down into sub-categories. Various different acoustic grammars may be formed as trees with various branches representing many different syllables forming a speech signal.

[010] According to another aspect of the invention, the performance of the speech engine and the phonemic recognition may be improved by pruning the search space used by the speech engine using a common phonetic marker. In one implementation, the acoustic grammar may be represented entirely by a loop of phonemes. In another implementation, the speech engine may reduce the search space by reducing the number of transitions in a grammar tree, thereby speeding up the process of compiling, loading, and executing the speech engine. For example, the phoneme loop may include a linking element between transitions. This may reduce the number of grammar transitions, such that grammar paths merge after a first transition and diverge after the linking element. In one implementation of the invention, a common acoustic element that is part of a speech signal may be used as the linking element. In one implementation of the invention, the acoustic element may be one that is very likely to be triggered even if it is unpronounced. For

example, a *schwa* in the English language may be used as the linking element because *schwa* represents an unstressed, central vowel that is likely to be spoken even if unintended. Those skilled in the art will appreciate that acoustic models for different languages may use other frequently elided phonemes as linking elements to reduce the search space used by the speech engine.

[011] The speech engine may generate a plurality of preliminary interpretations representing a set of best guesses at the user verbalization. The preliminary interpretations may be stored in a matrix, array, or another form, and may be provided to an interpretation sharpening module to determine a probable interpretation of a verbalization made by a user by applying heuristic policies against the preliminary interpretation to identify dominant words and/or phrases. According to various aspects of the invention, the interpretation sharpening module may include a policy module that may manage and/or provide one or more policies that enable the sharpening module to generate a plurality of probable interpretations of the verbalization made by the user. For example, according to one aspect of the invention, the plurality of preliminary interpretations may be applied against one or more policies to generate a set of hypotheses as to a candidate recognition. Each hypothesis may be reanalyzed to generate an interpretation score that may relate to a likelihood of the probable interpretation being a correct interpretation of the verbalization, and the preliminary interpretation corresponding to the highest (or lowest) interpretation score may then be designated as a probable interpretation of the verbalization. The designated probable interpretation may be stored and used for augmenting the policies to improve accuracy.

[012] According to one aspect of the invention, the policy module may include one or more agents that represent domains of knowledge. The agents may compete using a weighted model to revise a preliminary interpretation by determining context and intent. Relevant substitution of suspect words and phrases may be based on phonetic similarities or domain appropriateness. A domain agent may include one or more domain parameters for determining a probable interpretation from a preliminary interpretation. For example, domain parameters may include a policy vocabulary, a word position in the verbalization, a word combination, a sentence structure, or other parameters. A domain agent may include a parameter weighting scheme that may weight individual parameters according to one or more weighting factors, such as, a frequency of use, a difficulty to understand, or other factors.

[013] According to one aspect of the invention, the domain agents may revise a preliminary interpretation into a probable interpretation using phonetic fuzzy matching (PFM). In one implementation of the invention, the speech engine may output a phoneme stream that is applied against a model of phoneme feature similarities, drawn from domain agents, to identify a closest phonetic match using a multi-pass method. Domain agents may be loaded and prioritized into an M-Tree, which accounts for the possibility of the speech engine dropping or adding phonemes. An M-Tree may be an index structure that resolves similarity queries between phonemes using a closest-distance metric based on relative weightings of phoneme misrecognition, phoneme addition, and phoneme deletion. The M-Tree may be updated using an adaptive misrecognition model. For example, information about a verbalization and its components, as well as a probability that the probable interpretation was correct, may be stored and used for adapting the policy module for the user.

[014] In one implementation of the invention, the domain agents in the policy module may include one or more profile agents that may manage and/or provide one or more profile policies for revising a preliminary interpretation of a phoneme stream. For example, a profile agent may correspond to a user and may include one or more profile parameters tailored to the user. The profile agent may be used as a base policy to interpret any verbalizations made by the user. In other implementations, a profile agent may correspond to a particular language, a regional accent, or other profiles for interpreting a user verbalization. The profile agents may be augmented to enable the system to provide more accurate interpretations of verbalizations made by the user. The augmentation may include a user augmentation, such as providing additional vocabulary (e.g., names in an address book), one or more personalized pronunciations or other pronunciation information, or other user provided augmentations. The augmentation may also include a non-user provided augmentation, such as updates generated by a third party (e.g., a commercial administration and/or maintenance entity), or other non-user provided augmentations. The augmentation may be automated, such as adjusting a profile parameter-weighting scheme through an adaptive misrecognition model, as discussed above.

[015] In another implementation of the invention, the domain agents in the policy module may include one or more context agents that may manage and/or provide

one or more context policies for revising a preliminary interpretation of a phoneme stream. For example, a context agent may correspond to a context, such as song titles, city and street names, movie titles, finance, or other contexts. A context agent may include one or more context parameters that may be tailored to a verbalization context. The context policy may enhance an ability of the system related to interpreting verbalizations made by the user in the verbalization context corresponding to the context agent. The context agents may be augmented to enable the system to provide more accurate interpretations of verbalizations made in a verbalization context corresponding to the context agent. The augmentation may include a user provided augmentation, a non-user provided augmentation, an automated augmentation, or other augmentations. The augmentation may be automated, such as adjusting a profile parameter-weighting scheme through an adaptive misrecognition model, as discussed above.

[016] According to various implementations of the invention, the policy module may determine which profile agents and/or which context agents to use through a set of heuristics provided in a context-tracking module. In one implementation, the context-tracking module may use phonetic fuzzy matching to track a series of verbalizations by the user to identify a verbalization context. The context-tracking module may utilize one or more M-Trees to track the series of verbalizations and determine a closest phonetic match. The context-tracking module may track one or more past verbalization contexts for the series of verbalizations, one or more current verbalization contexts for the series of verbalizations, and/or make predictions regarding one or more future verbalization contexts for the series of verbalizations. The policy module may utilize information about the verbalization context of the series of verbalizations generated by the context tracking module to manage and/or provide one or more profile and/or context agents.

[017] According to one aspect of the invention, the system may include an interpretation history analysis module that may enable the system to augment one or more domain agents based on an analysis of past interpretations related to previously interpreted verbalizations. The augmentations enabled by the interpretation history analysis module may include a user augmentation, a third-party augmentation, an automated augmentation, or other augmentations. The interpretation history analysis module may include an information storage module that may store interpretation information related to past verbalizations, such as one

or more preliminary interpretations associated with a past verbalization, one or more interpretation scores associated with a past verbalization, one or more probable interpretations associated with a past verbalization, whether or not a past verbalization was interpreted correctly, or other information. A frequency module may be included in the interpretation history module, and the frequency module may use some or all of the information stored in the information storage module to generate one or more frequencies related to one or more past verbalizations. For example, the frequency module may calculate a frequency of word usage, word combinations, phonetic homonyms, interpretation errors for a particular verbalization, or other frequencies.

[018] The Information generated and/or stored by the interpretation history analysis module may be used to augment the profile and/or context agents in order to enhance the accuracy of subsequent interpretations. In some implementations, an adaptive misrecognition model may use one or more generated frequencies to augment one or more agents. For example, one or more parameters and/or weighting schemes of an agent or phonetic model may be augmented based on a frequency generated by the interpretation history analysis module. Other augmentations using information stored and/or generated by the interpretation history analysis module may be made, and the system may include a policy agent handler that may augment, update, remove, and/or provide one or more domain agents to the system. A domain agent may comprise a profile or context agent, and the policy agent handler may be controlled, directly or indirectly by a third party (e.g. a commercial entity, etc.). The policy agent handler may augment, update, remove, and/or provide domain agents to the system as part of a commercial agreement, such as a licensing agreement, a subscription agreement, a maintenance agreement, or other agreements.

[019] Other objects and advantages of the invention will be apparent to those skilled in the art based on the following detailed description and accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[020] Fig. 1 illustrates an exemplary system for enhancing automated speech interpretation according to one implementation of the invention.

[021] Fig. 2 illustrates an exemplary grammar tree for enhancing the performance of a speech engine according to one implementation of the invention.

[022] Fig. 3 illustrates an exemplary flow chart of a method for enhancing automated speech interpretation according to one implementation of the invention.

DETAILED DESCRIPTION

[023] Fig. 1 illustrates an exemplary system for enhancing automated speech interpretation according to one implementation of the invention. A speech-to-text processing engine 112 may receive a user verbalization, and speech engine 112 may generate one or more preliminary interpretations of the user verbalization. The preliminary interpretations may represent a set of best guesses as to the user verbalization arranged in any predetermined form or data structure, such as an array, a matrix, or other forms. In one implementation of the invention, speech engine 112 may generate the preliminary interpretations by performing phonetic dictation to recognize a stream of phonemes, instead of a stream of words. Phonemic recognition provides several benefits, particularly in the embedded space, such as offering out-of-vocabulary (OOV) capabilities, improving processing performance by reducing the size of a grammar, and eliminating the need to train Statistic Language Models (SLMs). Those skilled in the art will recognize other advantages of phonemic recognition.

[024] Speech engine 112 may apply the phoneme stream against one or more acoustic grammars that reliably map a speech signal to a phonemic representation in order to generate the plurality of preliminary interpretations. Characteristics of a speech signal may be mapped to a phonemic representation to construct a suitable acoustic grammar, and various acoustic grammars may be included in speech engine 112 to generate one or more preliminary interpretations according to the various acoustic grammars. For example, the English language may be mapped into a detailed acoustic grammar representing the phonotactic rules of English, where words may be divided into syllables, which may further be divided into core components of an onset, a nucleus, and a coda, which may be further broken down into one or more sub-categories.

[025] Once the phonotactic rules of a speech signal have been identified, a detailed acoustic grammar tree may be constructed that accounts for the nuances of the speech signal. The acoustic grammar may include a loop of phonemes, or the phoneme loop may include a linking element to reduce a size of a search space associated with the grammar. Using the English language as an example, the grammar tree may include various branches representing English language

syllables. The speech engine may traverse one or more grammar trees to generate one or more preliminary interpretations of a phoneme stream as a series of syllables that map to a word or phrase. By using phonemic recognition rather than word recognition, the size of the grammar can be reduced, which reduces the amount of time required to compile, load, and execute speech interpretation. Moreover, because the grammar maintains a high level of phonotactic constraints and therefore a large number of syllables, speech engine 112 may be very precise in generating phonemic representations of human verbalizations.

[026] An acoustic grammar used by speech engine 112 may be further optimized to reduce compile time, load time, and execution time by reducing the size of a search space associated with the acoustic grammar. Referring now to Fig. 2, a traditional grammar tree 120 is compared to an exemplary grammar tree according to one aspect of the invention to demonstrate the performance enhancements of speech engine 112. In traditional speech processing engines, nodes in a grammar tree 210 tend to represent words, or large-list applications may be supported provided through a grammar tree 210 where the nodes represent items in the large-list. This requires the speech engine to parse through a large number of transition states to come up with a recognition result, which degrades response time. An example of this is seen in the following grammar structure:

[027] "<street name> <city name>" → e.g., "NE 24th Street Bellevue"

[028] In the above example, a large list of street names is followed by a large list of city names. Assuming three elements in the list of street names, and three elements in the list of city names, this results in twenty-one transitions, which may be represented by traditional grammar tree 210. Every end-node of the first list is followed by all entries in the second list, potentially leading to very large grammars because most real-world large-list applications are likely to include much more than three list items. For example, a city may have hundreds or thousands of street names, and there may be hundreds or thousands of city names. Moreover, every element in the second segment of traditional grammar tree 210 is repeated, once for each first segment, which introduces redundancy.

[029] According to an aspect of the invention, the problems with traditional grammar trees may be resolved by using phonemic acoustic grammars instead of large-lists. The grammar may further be improved by including linking elements to reduce the number of transition states in the grammar. Thus, a grammar tree with a linking

element 220 will merge after a first segment and then spread out again at a second segment, where the segments may represent a phoneme in an acoustic grammar, as discussed above. For example, assume a two-syllable word in an acoustic grammar consisting of three phonemes, which is able to reduce the number of transitions from twenty-one in a traditional grammar tree 210 to twelve in a grammar tree with a linking element 220. Two syllables and three phonemes are chosen to show the reduction in search space in a grammar tree with a linking element 220 as opposed to a corresponding traditional grammar tree 210, although a real-world acoustic grammar modeled after a language is likely to have a maximum of roughly fifty phonemes. Moreover, the search space may be further reduced by restricting available transitions based on phonotactic constraints for an acoustic model.

[030] Using the approach described in Fig. 2, adding a linking element to an acoustic grammar may reduce both grammar size and response time. Part of a speech signal may be mapped to the linking element in order to maintain the phonotactic rules of the acoustic grammar. The linking element may be an acoustic element that is likely to be triggered even if unpronounced. For example, a *schwa* represents an unstressed, central vowel in the English language (e.g., the first and last sound in the word "arena" is *schwa*). The phoneme *schwa* is an ideal linking element because of how it is represented in a frequency spectrum. That is, *schwa* is a brief sound and when a person opens their mouth to speak, there is a strong likelihood of passing through the frequencies of *schwa* even if unintended. Those skilled in the art will recognize that this approach may be extended to acoustic models of speech signals for other languages by using frequently elided phonemes as linking elements to reduce the search space of an acoustic grammar.

[031] Referring again to Fig. 1, speech engine 112 may receive a user verbalization and process the verbalization into a plurality of preliminary interpretations using the techniques described above. That is, the verbalization may be interpreted as a series of phonemes, and the series of phonemes may be mapped to one or more preliminary interpretations by traversing one or more acoustic grammars that are modeled after grammar 220 of Fig. 2. The plurality of preliminary interpretations may take the form of words, parts of words, phrases, utterances, or a combination thereof, and the plurality of preliminary interpretations may be arranged as a matrix, an array, or in another form. The plurality of preliminary interpretations are then

passed to a speech sharpening engine 110 for deducing a most probable interpretation of the verbalization.

[032] According to various aspects of the invention, speech sharpening engine 110 may include an interpretation sharpening module 116, a policy module 114, an interpretation history analysis module 118, and a policy agent handler 120. The plurality of preliminary interpretations may be received by interpretation sharpening module 116, which forwards the preliminary interpretations to policy module 114 for further processing. Policy module 114 may include one or more context agents 126, one or more profile agents 128, and a context tracking module 130 that collectively revise the plurality of preliminary interpretations into a set of hypotheses that represent candidate recognitions of the verbalization. Policy module 114 may assign each hypothesis an interpretation score, and interpretation sharpening module 116 may designate the hypothesis with the highest (or lowest) interpretation score as a probable interpretation.

[033] According to one aspect of the invention, policy module 114 may include one or more context agents 126. Context agents 126 may represent domains of knowledge corresponding to a given context, such as song titles, city and street names, finance, movies, or other contexts. Context agents 126 may use context objects and associated dynamic languages to represent a corresponding context. Policy module 114 may also include one or more profile agents 128. Profile agents 128 may represent domains of knowledge corresponding to a given profile, such as a specific user, language, accent, or other profiles. Profile agents 128 may use profile objects and dynamic languages to represent a corresponding profile. Dynamic languages for context agents 126 or profile agents 128 may specify vocabularies, word combinations, phrases, sentence structures, criteria, and priority weightings for any given context or profile, respectively. The priority weightings may weight individual parameters according to one or more weighting factors, such as assigning a weight according to a frequency of use, a difficulty to understand, or other factors. Policy module 114 may also include a context-tracking module 130. Context tracking module 130 may track a verbalization context of a consecutive series of verbalizations. Context tracking module 130 may utilize one or more conversation trees to track the series of verbalizations. Context tracking sub-module 214 may track one or more past or current verbalization contexts of the series of verbalizations, and/or may make predictions regarding one or more future

verbalization contexts of the series of verbalizations. Policy module 114 may utilize information about the verbalization context, generated by context tracking module 130, to generate one or more sharpened interpretations and corresponding interpretation scores.

[034] In some implementations, policy module 114 may use context tracking module 130 to apply objects from one or more context agents 126 and/or profile agents 128 to the preliminary interpretations provided by speech engine 112. The various agents may compete with each other using a set of heuristics in a phonetic fuzzy matcher, where an intent or context of the user may be identified based on the set of heuristics about how a request may be phrased in a given domain. A closest phonetic match may be identified for suspect words and/or phrases among the plurality of preliminary interpretations.

[035] The phonetic fuzzy matcher may include an M-Tree that is populated with context objects, profile objects, and/or dynamic language data from one or more of context agents 126 and/or profile agents 128. M-Trees are known to those skilled in the art. The M-Tree may assign relative priority weights to the context objects, profile objects, and/or dynamic language data in order to account for the possibility of misrecognized phonemes, extraneous phonemes, or erroneously deleted phonemes. A closest distance metric associated with the M-Tree may be used given the relative weightings of phoneme misrecognition, phoneme addition, and phoneme deletion for various contexts and/or profiles.

[036] According to one aspect of the invention, one or more passes may be taken over the plurality of preliminary interpretations to identify dominant words and/or phrases among the plurality of preliminary interpretations. Using the M-Tree weighted model, one or more candidate interpretations may be made based on relevant substitution of suspect words and/or phrases based on phonetic similarities and/or domain appropriateness. For example, if a set of dominant words appear to be a movie name, a candidate interpretation will substitute the relevant words and/or phrases to generate a candidate interpretation about movies. After a set of candidate interpretations have been generated, the candidate interpretations are analyzed using the M-Tree weighted model. With the relevant domains constrained by the candidate interpretations, a confidence or interpretation score may be assigned to each candidate interpretation, with the interpretation score representing a likelihood that a particular candidate interpretation is a correct interpretation of the

verbalization. The candidate interpretations may then be returned to interpretation sharpening module 116, and interpretation sharpening module 116 may select a candidate interpretation with a highest (or lowest) interpretation score as a probable interpretation of the verbalization.

[037] According to various implementations of the invention, speech sharpening engine 110 may include an interpretation history analysis module 118. Interpretation history analysis module 118 may include an information storage module 122 a frequency module 124. Information storage module 122 may store information related to verbalizations, including components of verbalizations, preliminary interpretations, dominant words and/or phrases, candidate interpretations, probable interpretations, and/or interpretation scores associated with verbalizations, as well as whether or not a verbalization was interpreted correctly, or other information. Interpretation history analysis module 118 may also include a frequency module 124. Frequency module 124 may use some or all of the information stored in information storage module 122 to generate one or more frequencies related to one or more past verbalizations. For example, frequency module 124 may calculate a word usage frequency, a word combination frequency, a frequency related to a set of verbalizations that are phonetically similar but have distinct meanings, an interpretation error frequency for a particular verbalization, or other frequencies.

[038] Information stored and/or generated by interpretation history analysis module 118 may be used to augment speech sharpening engine 110. In some implementations, the information may be used to adjust various weights used in phonetic models, such as context agents 126 or profile agents 128, as well as adapting the relative weights in the M-Tree in context tracking module 130 to enhance accuracy for subsequent verbalizations. In another implementation, the stored information may be sent to a third-party or commercial entity for analyzing the data and developing new domain agents or further improving the accuracy of speech sharpening engine 110. For example, one or more parameters and/or weighting schemes of an agent may be augmented based on a frequency generated by interpretation history analysis module 118. Other augmentations related to information stored on and/or generated by interpretation history analysis module 118 may be made. Speech sharpening engine 110 may also include a policy agent handler 120 that may augment, update, remove, and/or provide one or more domain agents to policy module 114. A domain agent may include one or more new,

modified, or updated context agents 126 and/or profile agents 128. Policy agent handler 120 may also augment or update the M-Tree in context tracking module 130 to adjustments in priority weighting schemes or phonetic models. Policy agent handler 120 may be controlled, directly or indirectly, by a third party, such as a commercial entity, and domain agents may be augmented, updated, removed, and/or provided by policy agent handler 120 as part of a commercial agreement, licensing agreement, subscription agreement, maintenance agreement, or other agreement.

[039] Referring to Fig. 3, a flow chart demonstrating an exemplary method for enhancing the performance and accuracy of speech interpretation is provided. The method may begin by receiving a user verbalization at an operation 312. The received user verbalization may be electronically captured at operation 312, such as by a microphone or other electronic audio capture device. The electronically captured verbalization may be provided to a speech interpretation engine, such as speech engine 112 in Fig. 1.

[040] The speech interpretation may then generate one or more preliminary interpretations of the received verbalization at an operation 314. According to one implementation of the invention, the plurality of preliminary interpretations may be generated using phonetic dictation, grammar trees with linking elements, or any combination thereof to improve performance and enhance accuracy. Phonetic dictation and reducing a search space of a grammar tree by including linking elements is discussed in greater detail above. The preliminary interpretations may be arranged in any predetermined form, such as an array, a matrix, or other forms.

[041] In an operation 320, the preliminary interpretations may be provided to a speech sharpening engine. The speech sharpening engine may take one or more passes over the plurality of preliminary interpretations to identify dominant words and/or phrases in operation 320. This information may then be used to generate one or more candidate interpretations. The candidate interpretations may be based on various domain agents, such as context agents and/or profile agents, which may be organized as a weighted domain model, such as an M-Tree. For example, if a set of dominant words sound like a movie name, apply policies operation 320 may generate a candidate interpretation that substitutes relevant words and/or phrases based on a domain agent populated with movie titles. Additional passes may be made over the candidate interpretations, which may be constrained by domain

information associated with the candidate interpretations, to thereby generate a confidence score or interpretation score for each candidate interpretation. The interpretation score may represent a likelihood that a particular candidate interpretation is a correct interpretation of the verbalization received in operation 312. The operation of apply policies 320 is described in greater detail above in reference to Fig. 1.

[042] The candidate interpretations and corresponding interpretation scores may then be analyzed to determine a probable interpretation in an operation 322. In one implementation of the invention, a candidate interpretation with a highest (or lowest) score may be designated as a probable interpretation. The probable interpretation may then be output in an operation 324, such as for use in a voice-activated vehicular navigation system, a voice-controlled server or desktop computer, or other electronic device that can be controlled using voice commands.

[043] Information relating to the verbalization and the interpretations of the verbalization may be provided in a store interpretation operation 325. Store interpretation operation 324 may store interpretation information related to verbalizations, such as components of verbalizations, preliminary interpretations, dominant words and/or phrases, candidate interpretations, probable interpretations, and/or interpretation scores associated with verbalizations, as well as whether or not a verbalization was interpreted correctly, or other information. In some implementations of the invention, some or all of the interpretation information stored at store interpretation operation 324 may be used to determine one or more frequencies at a determine frequencies operation 326. The frequencies calculated at determine frequencies operation 326 may include one or more frequencies related to past verbalizations, such as, a word usage frequency, a word combination frequency, a frequency related to a set of verbalizations that are phonetically similar but have distinct meanings, an interpretation error frequency for a particular verbalization, or other frequencies. Determine frequencies operation 326 may be performed by interpretation history analysis module 118.

[044] In various implementations, a decision may be made whether to augment a speech sharpening engine in an augmentation decision operation 328. The decision concerning system augmentation may be based at least in part on information generated at determine frequencies block 326, such as one or more frequencies, or other information. If it is decided that no augmentation is needed, no further action is

taken until another verbalization is captured, and the method ends. In some instances, decision operation 328 may determine that augmentation may be made and control passes to an augment system operation 330. Augment system operation 330 may include making an augmentation to a speech sharpening engine. For example, one or more domain agents may be augmented to reflect probabilities of an interpretation being a correct interpretation of a verbalization, to update a user profile, or other augmentation. Dynamic languages associated with context agents and/or profile agents may be augmented, or parameters weights may be augmented to enhance accuracy when interpreting subsequent verbalizations. For example, an adaptive misrecognition technique may adjust the various weights in a phonetic model or update similarity weights for regional accents, or other augmentations may be made. In parallel to augment system operation 330, new agent policies may be received in an operation 332. For example, a third party or commercial entity may redesign or modify various domain agents, new domain agents may be developed and installed as plug-ins, domain agents that are unreliable may be removed, or other augmentations or modifications may be made. Thus, the method continually refines the domain agents and the weighting of various parameters in order to refine the accuracy of the speech sharpening engine for subsequent verbalizations.

[045] The above disclosure has been described in terms of specific exemplary aspects, implementations, and embodiments of the invention. However, those skilled in the art will recognize various changes and modifications that may be made without departing from the scope and spirit of the invention. For example, references throughout the specification to "one implementation," "one aspect," "an implementation," or "an aspect" may indicate that a particular feature, structure, or characteristic is included in at least one implementation. However, the particular features, structures, or characteristics may be combined in any suitable manner in one or more implementations. Therefore, the specification and drawings are to be regarded as exemplary only, and the scope of the invention is to be determined solely by the appended claims.

CLAIMS

What is claimed is:

1. A computer-implemented method for interpreting speech, the method comprising:

receiving a user verbalization;

parsing the verbalization into a plurality of phonemes; and

generating at least one interpretation of the verbalization based on the plurality of phonemes.

2. The method of claim 1, wherein generating the at least one preliminary interpretation includes traversing an acoustic grammar based on the plurality of phonemes.

3. The method of claim 2, wherein the acoustic grammar is a phonemic representation of a speech signal.

4. The method of claim 3, wherein the acoustic grammar includes a linking element between segments, the linking element being an acoustic element in the speech signal.

5. The method of claim 4, wherein the acoustic element is a phoneme representing an unstressed vowel.

6. The method of claim 1, further comprising:

revising the at least one interpretation, wherein the revising includes generating a plurality of candidate interpretations of the verbalization;

assigning an interpretation score to each candidate interpretation; and

selecting a candidate interpretation with a highest interpretation score as a probable interpretation of the verbalization.

7. The method of claim 6, wherein revising the at least one interpretation includes:

applying the at least one interpretation against at least one domain agent;

identifying suspect words and/or phrases in the at least one interpretation;

identifying a closest phonetic match for the suspect words and/or phrases based on a relative weighting of parameters in at least one domain associated with the at least one domain agent; and

substituting the closest phonetic match for the suspect words and/or phrases to generate a candidate interpretation.

8. The method of claim 7, wherein the closest phonetic match is stored in an M-Tree, and wherein identifying the closest phonetic match includes using a closest-distance metric associated with the M-Tree.

9. The method of claim 7, wherein the parameters include at least one of a context of the verbalization, a verbalization criteria, a history of user verbalizations, a user profile, or domain specific information.

10. The method of claim 7, further comprising adjusting the relative weighting of parameters in response to the selection of a probable interpretation of the verbalization.

11. A method for sharpening a preliminary interpretation of a user verbalization, the preliminary interpretation generated by a speech interpretation engine, the method comprising:

revising the preliminary interpretation, wherein the revising includes generating a plurality of candidate interpretations of the verbalization;

assigning an interpretation score to each candidate interpretation; and

selecting a candidate interpretation with a highest interpretation score as a probable interpretation of the verbalization.

12. The method of claim 11, wherein revising the preliminary interpretation includes:

applying the preliminary interpretation against at least one domain agent;

identifying suspect words and/or phrases in the preliminary interpretation;

identifying a closest phonetic match for the suspect words and/or phrases based on a relative weighting of parameters in at least one domain associated with the at least one domain agent; and

substituting the closest phonetic match for the suspect words and/or phrases to generate a candidate interpretation.

13. The method of claim 12, wherein the closest phonetic match is stored in an M-Tree, and wherein identifying the closest phonetic match includes using a closest-distance metric associated with the M-Tree.

14. The method of claim 12, wherein the parameters include at least one of a context of the verbalization, a verbalization criteria, a history of user verbalizations, a user profile, or domain specific information.

15. The method of claim 12, further comprising adjusting the relative weighting of parameters in response to the selection of a probable interpretation of the verbalization.

16. A system for interpreting speech, the system comprising:

a device that receives a user verbalization and generates an electronic signal corresponding to the user verbalization; and

a speech interpretation engine that receives the electronic signal corresponding to the user verbalization, wherein the speech interpretation engine parses the signal into a plurality of phonemes and generates at least one interpretation of the verbalization based on the plurality of phonemes.

17. The system of claim 16, wherein generating the at least one preliminary interpretation includes traversing an acoustic grammar based on the plurality of phonemes.

18. The system of claim 17, wherein the acoustic grammar is a phonemic representation of a speech signal.

19. The system of claim 18, wherein the acoustic grammar includes a linking element between segments, the linking element being an acoustic element in the speech signal.

20. The system of claim 19, wherein the acoustic element is a phoneme representing an unstressed vowel.

21. The system of claim 16, further comprising a speech sharpening engine configured to:

revise the at least one interpretation by generating a plurality of candidate interpretations of the verbalization;

assign an interpretation score to each candidate interpretation; and

select a candidate interpretation with a highest interpretation score as a probable interpretation of the verbalization.

22. The system of claim 21, wherein the speech sharpening engine is configured to revise the at least one interpretation by:

applying the at least one interpretation against at least one domain agent;

identifying suspect words and/or phrases in the at least one interpretation;

identifying a closest phonetic match for the suspect words and/or phrases based on a relative weighting of parameters in at least one domain associated with the at least one domain agent; and

substituting the closest phonetic match for the suspect words and/or phrases to generate a candidate interpretation.

23. The system of claim 22, wherein the closest phonetic match is stored in an M-Tree, and wherein identifying the closest phonetic match includes using a closest-distance metric associated with the M-Tree.

24. The system of claim 22, wherein the parameters include at least one of a context of the verbalization, a verbalization criteria, a history of user verbalizations, a user profile, or domain specific information.

25. The system of claim 22, wherein the speech sharpening engine is further configured to adjust the relative weighting of parameters in response to the selection of a probable interpretation of the verbalization.

26. A system for sharpening a preliminary interpretation of a user verbalization received by an electronic capture device, the preliminary interpretation generated by a speech interpretation engine, the system comprising a speech sharpening engine containing at least one processor configured to:

revise the preliminary interpretation by generating a plurality of candidate interpretations of the verbalization;

assign an interpretation score to each candidate interpretation; and

select a candidate interpretation with a highest interpretation score as a probable interpretation of the verbalization.

27. The system of claim 26, wherein the speech sharpening engine is further configured to revise the preliminary interpretation by:

applying the preliminary interpretation against at least one domain agent;

identifying suspect words and/or phrases in the preliminary interpretation;

identifying a closest phonetic match for the suspect words and/or phrases based on a relative weighting of parameters in at least one domain associated with the at least one domain agent; and

substituting the closest phonetic match for the suspect words and/or phrases to generate a candidate interpretation.

28. The system of claim 27, wherein the closest phonetic match is stored in an M-Tree, and wherein identifying the closest phonetic match includes using a closest-distance metric associated with the M-Tree.

29. The system of claim 27, wherein the parameters include at least one of a context of the verbalization, a verbalization criteria, a history of user verbalizations, a user profile, or domain specific information.

30. The system of claim 27, wherein the speech sharpening engine is further configured to adjust the relative weighting of parameters in response to the selection of a probable interpretation of the verbalization.

31. A computer-readable medium containing computer-executable instructions for interpreting speech, the computer-executable instructions configured to:

receive a user verbalization;

parse the verbalization into a plurality of phonemes; and

generate at least one interpretation of the verbalization based on the plurality of phonemes.

32. The computer-readable medium of claim 31, wherein the computer-executable instructions are configured to generate the at least one preliminary interpretation by traversing an acoustic grammar based on the plurality of phonemes.

33. The computer-readable medium of claim 32, wherein the acoustic grammar is a phonemic representation of a speech signal.

34. The computer-readable medium of claim 33, wherein the acoustic grammar includes a linking element between segments, the linking element being an acoustic element in the speech signal.

35. The computer-readable medium of claim 34, wherein the acoustic element is a phoneme representing an unstressed vowel.

36. The computer-readable medium of claim 31, wherein the computer-executable instructions are further configured to:

revise the at least one interpretation by generating a plurality of candidate interpretations of the verbalization;

assign an interpretation score to each candidate interpretation; and

select a candidate interpretation with a highest interpretation score as a probable interpretation of the verbalization.

37. The computer-readable medium of claim 36, wherein the computer-executable instructions are configured to revise the at least one interpretation by:

applying the at least one interpretation against at least one domain agent;

identifying suspect words and/or phrases in the at least one interpretation;

identifying a closest phonetic match for the suspect words and/or phrases based on a relative weighting of parameters in at least one domain associated with the at least one domain agent; and

substituting the closest phonetic match for the suspect words and/or phrases to generate a candidate interpretation.

38. The computer-readable medium of claim 37, wherein the closest phonetic match is stored in an M-Tree, and wherein identifying the closest phonetic match includes using a closest-distance metric associated with the M-Tree.

39. The computer-readable medium of claim 37, wherein the parameters include at least one of a context of the verbalization, a verbalization criteria, a history of user verbalizations, a user profile, or domain specific information.

40. The computer-readable medium of claim 37, wherein the computer-executable instructions are further configured to adjust the relative weighting of parameters in response to the selection of a probable interpretation of the verbalization.

41. A computer-readable medium containing computer-executable instructions for sharpening a preliminary interpretation of a user verbalization, the preliminary interpretation generated by a speech interpretation engine, the computer-executable instructions configured to:

revise the preliminary interpretation by generating a plurality of candidate interpretations of the verbalization;

assign an interpretation score to each candidate interpretation; and

select a candidate interpretation with a highest interpretation score as a probable interpretation of the verbalization.

42. The computer-readable medium of claim 41, wherein the computer-executable instructions are configured to revise the preliminary interpretation by:

applying the preliminary interpretation against at least one domain agent;

identifying suspect words and/or phrases in the preliminary interpretation;

identifying a closest phonetic match for the suspect words and/or phrases based on a relative weighting of parameters in at least one domain associated with the at least one domain agent; and

substituting the closest phonetic match for the suspect words and/or phrases to generate a candidate interpretation.

43. The computer-readable medium of claim 42, wherein the closest phonetic match is stored in an M-Tree, and wherein identifying the closest phonetic match includes using a closest-distance metric associated with the M-Tree.

44. The computer-readable medium of claim 42, wherein the parameters include at least one of a context of the verbalization, a verbalization criteria, a history of user verbalizations, a user profile, or domain specific information.

45. The computer-readable medium of claim 42, wherein the computer-executable instructions are further configured to adjust the relative weighting of parameters in response to the selection of a probable interpretation of the verbalization.

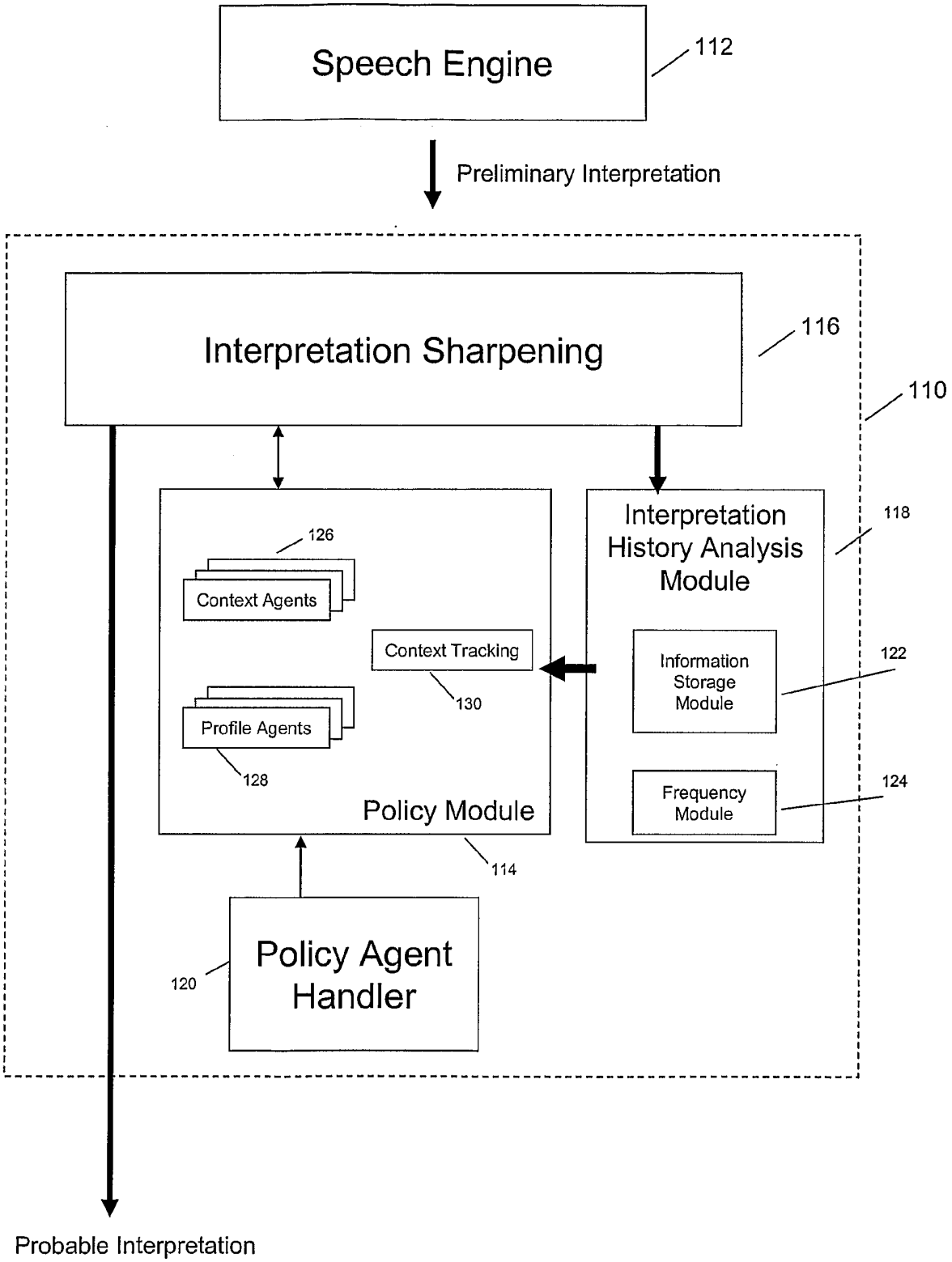
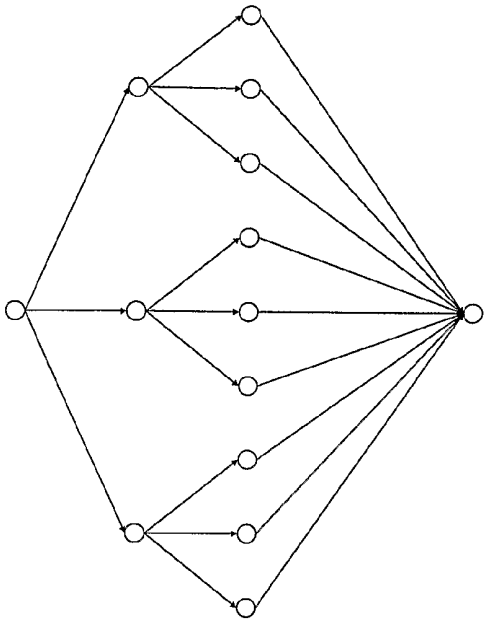

Speech Engine
112
Preliminary Interpretation
Interpretation Sharpening
116
110
Interpretation History Analysis Module
118
126
Context Agents
Context Tracking
130
Information Storage Module
122
Profile Agents
128
Frequency Module
124
Policy Module
114
Policy Agent Handler
120
Probable Interpretation

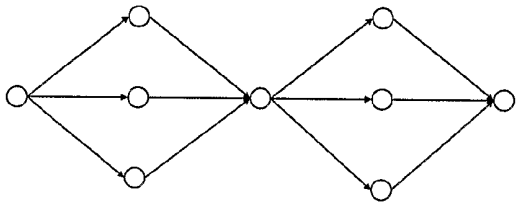
Fig. 1

Traditional Grammar Tree



210

Grammar Tree with Linking Element



220

Figure 2

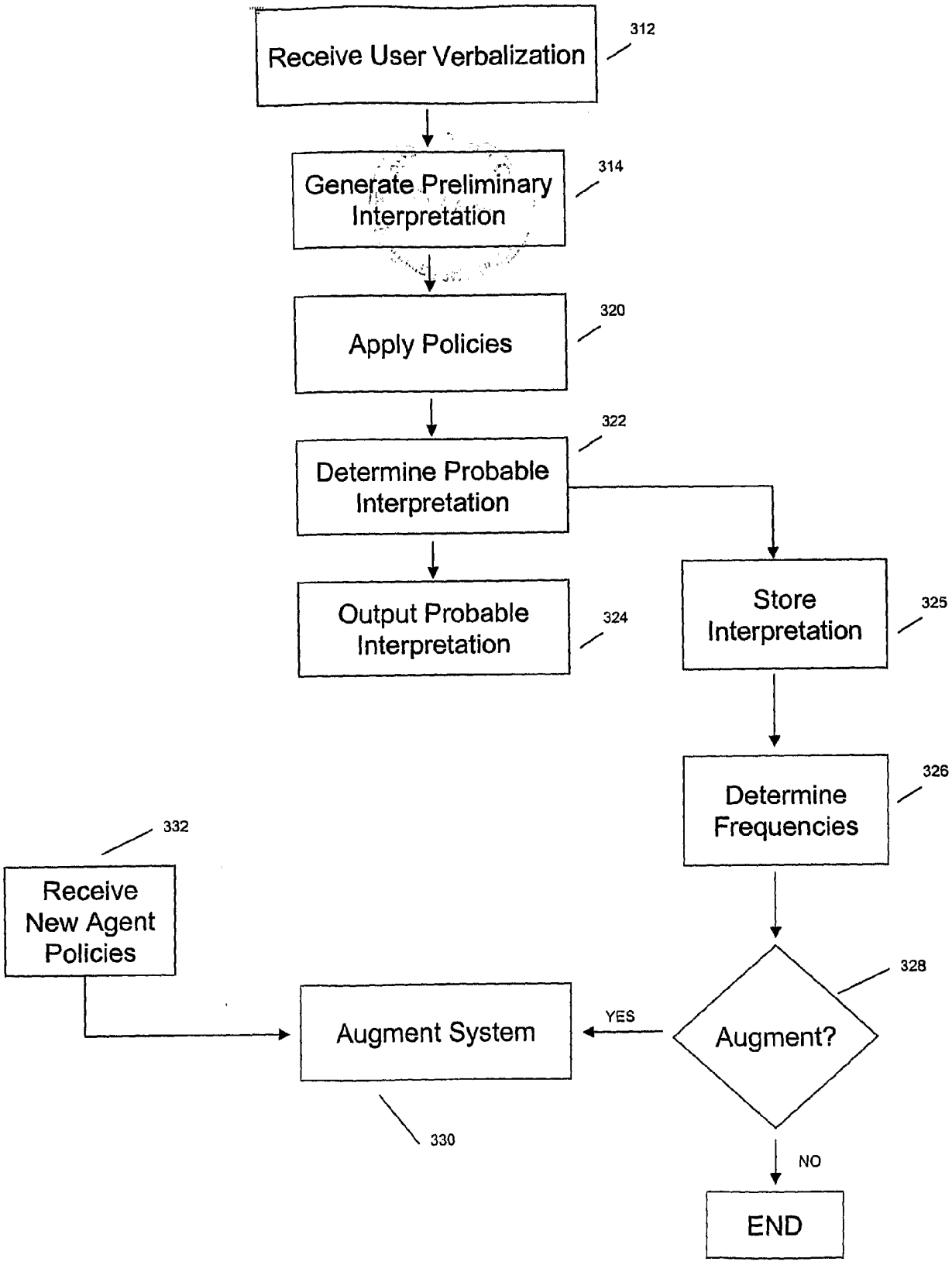

Receive User Verbalization
312
Generate Preliminary Interpretation
314
Apply Policies
320
Determine Probable Interpretation
322
Output Probable Interpretation
324
Store Interpretation
325
Determine Frequencies
326
Receive New Agent Policies
332
Augment System
330
YES
Augment?
328
NO
END

Fig. 3

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
14 August 2008 (14.08.2008)

PCT

(10) International Publication Number
WO 2008/098039 A1

(51) International Patent Classification:
G06Q 30/00 (2006.01)

(21) International Application Number:
PCT/US2008/053157

(22) International Filing Date: 6 February 2008 (06.02.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/671,526 6 February 2007 (06.02.2007) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [US/US]; 11980 NE 24th Street, Bellevue, WA 98005 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **FREEMAN, Tom** [US/US]; 6735 83rd Avenue, S.E., Mercer Island, WA 98040 (US). **KENNEWICK, Mike** [US/US]; 13407 Ne 37th Place, Bellevue, WA 98005 (US).

(74) Agents: **GATTO, James, G.** et al.; Pillsbury Winthrop Shaw Pittman LLP, P.O. Box 10500, McLean, VA 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

(54) **Title:** SYSTEM AND METHOD FOR SELECTING AND PRESENTING ADVERTISEMENTS BASED ON NATURAL LANGUAGE PROCESSING OF VOICE-BASED INPUT

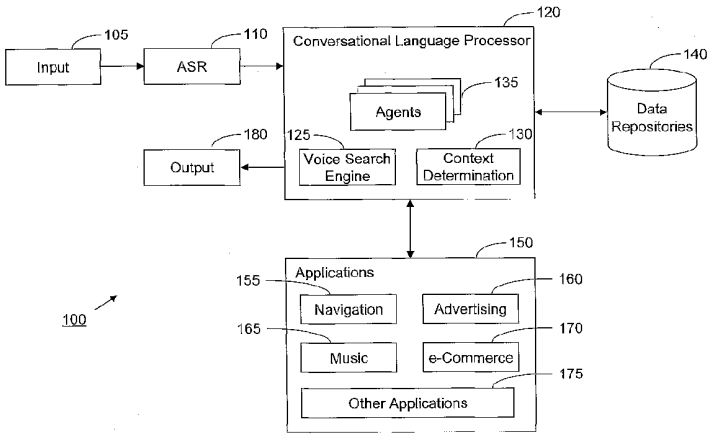


Fig. 1

(57) **Abstract:** A system and method for selecting and presenting advertisements based on natural language processing of voice-based inputs is provided. A user utterance may be received at an input device, and a conversational, natural language processor may identify a request from the utterance. At least one advertisement may be selected and presented to the user based on the identified request. The advertisement may be presented as a natural language response, thereby creating a conversational feel to the presentation of advertisements. The request and the user's subsequent interaction with the advertisement may be tracked to build user statistical profiles, thus enhancing subsequent selection and presentation of advertisements.

WO 2008/098039 A1

SYSTEM AND METHOD FOR SELECTING AND PRESENTING ADVERTISEMENTS BASED ON NATURAL LANGUAGE PROCESSING OF VOICE-BASED INPUT

FIELD OF THE INVENTION

[001] The present invention relates to selecting and presenting advertisements based on natural language processing of voice-based input.

BACKGROUND OF THE INVENTION

[002] As technology advances, consumer electronics devices tend to play larger roles due to increased functionality and mobility. For example, mobile phones, navigation devices, embedded devices, and other such devices provide a wealth of functionality beyond core applications. However, increased functionality adds difficulty to the learning curve associated with using electronic devices, and increased mobility intensifies the demand for simple mechanisms to interact with devices on the go. For example, existing systems tend to have complex human to machine interfaces, which may inhibit mass-market adoption for various technologies. For example, when a user wishes to perform a relatively simple task on a mobile phone, such as purchasing a ring tone, the user often is forced to navigate through a series of menus and press a series of buttons. In some instances, this may result in the transaction not necessarily occurring, as the user may prefer to avoid the hassles altogether. As such, there is ever-growing demand for ways to exploit technology in intuitive ways.

[003] Voice recognition software may enable a user to exploit applications and features of a device that may otherwise be unfamiliar, unknown, or difficult to use. However, many existing voice user interfaces (when they actually work) still require significant learning on the part of the user. For example, users often cannot directly issue a request for a system to retrieve information or perform an action without having to memorize specific syntaxes, words, phrases, concepts, semantic indicators, or other keywords/qualifiers. Similarly, when users are uncertain of particular needs, many existing systems do not engage the user in a productive, cooperative dialogue to resolve requests and advance a conversation. Instead, many existing speech interfaces force users to use a fixed set commands or

keywords to communicate requests in ways that systems can understand. Using existing voice user interfaces, there is virtually no option for dialogue between the user and the system to satisfy mutual goals.

[004] The lack of adequate voice user interfaces results in missed opportunities for providing valuable and relevant information to users. Not only does this potentially leave user requests unresolved, in certain instances, providers of goods and services may lose out on potential business. In an increasingly global marketplace, where marketers are continually looking for new and effective ways to reach consumers, the problems with existing voice user interfaces leaves a large segment of consumer demand unfulfilled. Furthermore, existing techniques for marketing, advertising, or otherwise calling consumers to action fail to effectively utilize voice-based information, which is one of the most natural, intuitive methods of human interaction.

[005] Existing systems suffer from these and other problems.

SUMMARY OF THE INVENTION

[006] According to various aspects of the invention, a system and method for selecting and presenting advertisements based on natural language processing of voice-based inputs is provided. A natural language voice-based input may be received by a voice user interface. The voice-based input may include a user utterance, and a request may be identified from the utterance. Appropriate action may be taken to service the request, while one or more advertisements may be selected and presented to the user. Advertisements may be selected based on various criteria, including content of the input (e.g., concepts, semantic indicators, etc.), an activity related to the input (e.g., a relation to a request, a requested application, etc.), user profiles (e.g., demographics, preferences, location, etc.), or in other ways. A user may subsequently interact with the advertisement (e.g., via a voice-based input), and action may be taken in response to the interaction. Furthermore, the interaction may be tracked to build statistical profiles of user behavior based on affinities or clusters among advertisements, user profiles, contexts, topics, semantic indicators, concepts, or other criteria.

[007] According to various aspects of the invention, advertisers may create advertisements, which may be stored in an advertisement repository. For example,

advertisements may include sponsored messages, calls to action, purchase opportunities, trial downloads, or any other marketing communication, as would be apparent to those skilled in the art. Advertisers may specify various parameters to associate with the advertisements, such as various contexts or topic concepts (e.g., semantic indicators for a "music" concept may include words such as "music," "tunes," "songs," etc.), target demographics (e.g., a preferred audience), marketing criteria or prices for insertion (e.g., dynamic or static pricing based on various marketing criteria), or other information, as would be apparent. The advertisement repository may be associated with a server, where in response to a voice-based input from a user (e.g., at a voice-enabled device), a communications link may be established with the server. Information may be extracted from the voice-based input (e.g., words in the input, applications requested by the input, etc.), and the extracted information may be correlated with user profiles, advertisement parameters, or other information to determine which advertisements to select in relation to the voice-based input. The server may subsequently communicate the selected advertisements to the user, and the server may track the user's subsequent interaction with the selected advertisements.

[008] Other objects and advantages of the invention will be apparent based on the following drawings and detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

[009] Fig. 1 illustrates a block diagram of an exemplary system for implementing a voice user interface according to various aspects of the invention.

[010] Fig. 2 illustrates a block diagram of an exemplary advertising system according to various aspects of the invention.

[011] Fig. 3 illustrates a flow diagram of an exemplary method for selecting and presenting advertisements based on voice-based inputs according to various aspects of the invention.

DETAILED DESCRIPTION

[012] Referring to Fig. 1, an exemplary system 100 for implementing a voice user interface is illustrated according to various aspects of the invention. System 100 may enable users to perform various tasks on a voice-enabled device. For example, users may control navigation

devices, media devices, personal computers, personal digital assistants, or any other device supporting voice-based inputs. System 100 may enable users to request voice-enabled devices to retrieve information or perform various tasks, among other things, using natural language voice-based inputs. For example, system 100 may interpret natural language voice-based inputs and generate responses using, among other things, techniques described in co-pending U.S. Patent Application Ser. No. 10/452,147, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed June 3, 2003; and U.S. Patent Application Ser. No. 10,618,633, entitled "Mobile Systems and Methods for Responding to Natural Language Speech Utterance," filed June 15, 2003, both of which are hereby incorporated by reference in their entirety.

[013] According to various aspects of the invention, system 100 may receive a user input, including at least a voice-based user utterance, at an input device 105. Input device 105 may include any suitable device, or combination of devices, for receiving a voice-based input (e.g., a microphone). In various implementations, input device 105 may include a multi-modal input, such as a touch-screen interface, keypad, or other input. The received utterance may be processed by an Automatic Speech Recognizer 110. Automatic Speech Recognizer 110 may generate one or more preliminary interpretations of the utterance using various techniques. For example, Automatic Speech Recognizer 110 may interpret the utterance using techniques of phonetic dictation to recognize a stream of phonemes. Further, Automatic Speech Recognizer 110 may perform post-processing to enhance the preliminary interpretations. For example, Automatic Speech Recognizer 110 may vary interpretations of an utterance, or components of an utterance, from one context to another. Other techniques for enhancing an interpretation of a user utterance may be used, such as those described in co-pending U.S. Patent Application Ser. No. 11/513,269, entitled "Dynamic Speech Sharpening," filed August 31, which is hereby incorporated by reference in its entirety.

[014] The one or more preliminary interpretations may be provided to a conversational language processor 120. Conversational language processor 120 may include a voice search engine 125, a context determination module 130, and one or more agents 135, among other things, to enable cooperative, conversational interaction between the user and system 100. Conversational language processor 120 may be communicatively coupled to one or more data repositories 140 and one or more applications 150. Conversational language processor

120 may generate a domain-specific conversational response, which may be returned to the user as an output 180. Output 180 may include a multi-modal output (e.g., by simultaneously returning a voice-based response and displaying information on a display device).

[015] System 100 may further include an interaction with one or more applications 150 to service one or more requests in the utterance. For example, the utterance may include one or more requests for performing an action, retrieving information, or various combinations thereof. Output 180 may include a conversational response to advance a conversation to service requests by invoking one or more applications 150, as appropriate. For example, applications 150 may include a navigation application 155, an advertising application 160, a music application, an electronic commerce application 170, and/or other applications 175. Furthermore, Automatic Speech Recognizer 110, conversational language processor 120, data repositories 140, and/or applications 150 may reside locally (e.g., on a user device), remotely (e.g., on a server), and/or hybrid local/remote processing models may be used (e.g., lightweight applications may be processed locally while computationally intensive applications may be processed remotely).

[016] Conversational language processor 120 may build long-term and/or short-term shared knowledge in one or more knowledge source. For example, shared knowledge sources may include information about previous utterances, requests, and other user interactions to inform generating an appropriate response to a current utterance. The shared knowledge may include public/non-private (i.e., environmental) knowledge, as well as personal/private (i.e., historical) knowledge. For example, conversational language processor 120 may use context determination module 130 to establish a context for a current utterance by having domain agents 135 competitively generate a context-based interpretation of the utterance (e.g., by scoring possible interpretations and selecting a highest scoring interpretation). As such, agents 135 may model various domains (e.g., navigation, music, a specific user, global users, advertising, e-commerce, etc.), and conversational language processor 120 may interpret and/or respond to a voice-based input accordingly. For example, context-based interpretations and responses to a voice-based input may be generated using techniques described in co-pending U.S. Patent Application Ser. No. 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed August 5, 2005; and U.S. Patent Application Ser. No. 11/212,693,

entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," filed August 29, 2005, both of which are hereby incorporated by reference in their entirety.

[017] Furthermore, conversational language processor 120 may support adaptive misrecognition to reinterpret a current utterance and/or one or more previous utterances. For example, information contained in a current utterance may indicate that interpretations for one or more previous utterances were incorrect, and therefore, the previous utterances may be reinterpreted to improve subsequent interpretations. Accordingly, conversational language processor 120 may use the techniques described herein, along with various other techniques, to interpret and respond to conversational, natural language utterances. Conversational language processor 120 may use various other techniques as will be apparent, such as those described in co-pending U.S. Patent Application Ser. No. 11/200,164, entitled "System and Method of Supporting Adaptive Misrecognition in Conversational Speech," filed August 10, 2005; and U.S. Patent Application Ser. No. 11/580,926, entitled "System and Method for a Cooperative Conversational Voice User Interface," filed October 16, 2006, both of which are hereby incorporated by reference in their entirety.

[018] Referring to Fig. 2, an exemplary advertising system 200 is illustrated according to various aspects of the invention. System 200 may include a server 230 for receiving one or more advertisements from an advertiser 220, wherein the advertisements may be stored in a data repository 260 associated with server 230. For example, advertisements may include sponsored messages or marketing communications, calls to action, purchase opportunities, trial downloads, coupons, or any other suitable marketing, advertising, campaign, or other information, as would be apparent to those skilled in the art. A voice-enabled device 210 may receive a voice-based input and establish communications with advertising server 230. Subsequently, advertising server 230 may select one or more advertisements from among the advertisements stored in data repository 260, and the selected advertisements may be provided to the voice-enabled device for presentation to a user.

[019] Advertiser 220 may access advertising server 230 via an advertiser interface 245. Advertisers 220 may upload targeted advertisements to server 230 via advertiser interface 245, and server 230 may store the advertisements in data repository 260. The advertisements may include graphically-based advertisements that include banners, images,

audio, video, or any suitable combination thereof. Furthermore, the advertisements may include interactive or embedded information, such as links, metadata, or computer-executable instructions, or any suitable combination thereof. Advertisers may specify criteria for a campaign or targeting information for an advertisement (e.g., a start date, an end date, budget information, geo-targeting information, conceptual or contextual information, or any other suitable criteria), which may be used to facilitate selecting an advertisement in relation to a particular voice-based input.

[020] In addition to providing interface 245 for advertisers, server 230 may include a content/action identification module 235, a user profile module 240, an advertisement selection module 250, and a tracking module 255. Users may submit voice-based requests to voice-enabled device 210, and voice-enabled device 210 may communicate information about the voice-based input to server 230. Server 230 may invoke advertisement selection module 250 to extract relevant information from about the voice-based input, where advertisement selection module 250 may select one or more advertisements relevant to the voice-based input based on information extracted using content/action identification module 235 and/or user profile module 240.

[021] For example, content/action identification module 235 may identify content of the voice-based input (e.g., words in the input), requested information (e.g., search results, a web page, music, video, graphics, or other information), requested actions (e.g., calculating a navigation route, placing a telephone call, playing a song, etc.), a category or topic related to the input (e.g., music, business, stocks, sports, navigation, movies, etc.), or other criteria to use in selecting an advertisement. Further, user profile module 240 may identify characteristics of a specific user (e.g., demographics, personal preferences, location-based information, etc.), global user profiles (e.g., demographic profiles, click-through rates, etc.), or other criteria to use in selecting an advertisement. Moreover, advertisement selection module 250 may account for where a request originates from. For example, advertisements may be selected based on a default user location (e.g., identified from a user profile), current geolocation information (e.g., identified from a navigation device), whether an affiliate or partner of server 230 initiated the request, or other criteria.

[022] For instance, a user may request airline reservations via voice-enabled device 210, and content/action identification module 235 may identify specific words used in the request, a category related to the request (e.g., travel, airlines, hotels, etc.), or other

information. Furthermore, user profile module 240 may identify relevant characteristics of the user (e.g., user-specific demographics, location information, preferred airlines or hotels, etc.), as well as global user characteristics (e.g., most popular airlines). In various implementations, advertisements may be selected by assigning a score to each advertisement (e.g., based on click-through rates, relevance metrics, target audiences, etc.). As such, advertisement selection module 250 may correlate the information about the request to select advertisements stored in data repository 260, and server 230 may communicate the selected advertisements to voice-enabled device 210. Furthermore, selected advertisements may be presented according to a predetermined ordering or ranking (e.g., based on a ranking of relevance to an advertisement).

[023] In various implementations, advertisement selection module 250 may retrieve a predetermined number of advertisements for any given request. Furthermore, the selected advertisements may depend upon a presentation format. For example, advertisements may be selected based on an amount of available space on a display of voice-enabled device 210 and/or a size/shape of the selected advertisements. In another example, voice-based advertisements may be selected and presented to the user audibly (e.g., a "hands-free" advertisement may be preferred when voice-enabled device 210 is a telematics device).

[024] Furthermore, the user's subsequent interaction with an advertisement may be tracked using tracking module 255. For example, tracking module 255 may determine whether a conversion or click-through occurs for each advertisement presented to users. Further, tracking module 255 may maintain accounting and/or billing information associated with advertisers 220. For example, advertisers 220 may specify a maximum insertion cost, a cost-per-click-through, an average insertion cost, or other criteria specifying a budget constraint for an advertisement. As such, tracking module 255 may track which advertisements are selected and/or presented, which advertisements result in a conversion or click-through, whether a click-through or conversion results in a transaction or sale, associations between advertisements and users, requests, concepts, semantic indicators, and/or other criteria. For example, tracking user interaction with advertisements may be used to build user-specific and/or global statistical profiles that map or cluster advertisements to topics, semantic indicators, contexts, concepts, etc. based on user behavior, demographics, targeting constraints, content of advertisements, content of requests, actions associated with requests, or other statistically relevant information.

Accordingly, the tracking information may be used to bill or invoice advertisers 220, as well as to improve subsequent performance and relevance of advertisements selected using advertisement selection module 250. Other techniques and features of selecting and presenting advertisements based on voice-based inputs may suitably be employed, as would be apparent.

[025] Referring to Fig. 3, an exemplary method for selecting and presenting advertisements based on a voice-based input is illustrated according to various aspects of the invention. The method may begin in an operation 305, where a voice-based input, including at least a user utterance, may be received at a voice user interface. The voice user interface may include any suitable mechanism for receiving the utterance (e.g., a microphone), and may interface with any suitable voice-enabled device, as would be apparent, including personal navigation devices, personal digital assistants, media devices, telematics devices, personal computers, mobile phones, or others.

[026] Subsequently, one or more requests included in the voice-based input may be identified in an operation 310. For example, the requests may include requests to retrieve information, perform tasks, explore or gather information, or otherwise interact with a system or device. For example, a voice-based input to a navigation device may include a request to calculate a route or retrieve location-based information. In another example, a voice-based input to a mobile phone may include a request to place a telephone call, purchase a ringtone, or record a voice-memo. Furthermore, in various implementations, voice-based inputs may include multiple requests, multi-modal requests, cross-device requests, cross-application requests, or other types of requests. For example, an utterance received in operation 305 may be: "Get me a route to Chang's Restaurant, and call them so I can make a reservation." The utterance may thus include multiple requests, including cross-device requests (e.g., calculate a route using a navigation device, and make a telephone call using a mobile phone), as well as cross-application requests (e.g., search for an address and/or phone number using a voice search engine, and calculate a route using a navigation application).

[027] The requests may be part of a conversational interaction between a user and a system or device, whereby an interpretation of requests in a current utterance may be based upon previous utterances in a current conversation, utterances in previous conversations, context-based information, local and/or global user profiles, or other

information. For example, a previous request may be reinterpreted based on information included in subsequent requests, a current request may be interpreted based on information included in previous requests, etc. Furthermore, the conversational interaction may take various forms, including query-based conversations, didactic conversations, exploratory conversations, or other types of conversations. For example, the conversational language processor may identify a type of conversation, and information may be extracted from the utterance accordingly to identify the one or more requests in operation 310. Moreover, the conversational language processor may determine whether any of the requests are incomplete or ambiguous, and action may be taken accordingly (e.g., a system response may prompt a user to clarify an incomplete and/or ambiguous request). The conversational language processor may therefore use various techniques to identify a conversation type, interpret utterances, identify requests, or perform other tasks, such as those described in the aforementioned co-pending patent applications, which are hereby incorporated by reference in their entirety.

[028] Upon identifying the one or more requests, action may be taken based on the identified requests in an operation 315, while one or more advertisements may be selected in an operation 320 (described in greater detail below). For example, one or more context-appropriate applications may be invoked to service the requests in operation 315 (e.g., a voice search engine, a navigation application, an electronic commerce application, or other application may be invoked depending upon the request). Furthermore, in operation 320, information may be communicated to an advertising server to select one or more advertisements related to the request. Thus, as shown in Fig. 3, taking action in operation 315 and selecting advertisements in operation 320 may be related operations (e.g., advertisements may be selected to help in interpreting incomplete and/or ambiguous requests).

[029] Upon taking action in operation 315 (e.g., to service the request) and selecting one or more advertisements in operation 320 (e.g., in relation to the request), an output may be presented to the user in operation 325. The output may indicate a result of the action associated with operation 315. For example, the output may include requested information, an indication of whether a requested task was successfully completed, whether additional information is needed to service the request (e.g., including a prompt for the information), or other information relating to an action based on the request. Furthermore,

the output may include advertisements, as selected in operation 320. For example, the output may include text-based, graphic-based, video-based, audio-based, or other types of advertisements, as would be apparent to those skilled in the art. Further, the output may include other types of advertisements, including calls to action (e.g., a location-based coupon or purchase opportunity, trial downloads, or other actionable advertising or marketing).

[030] Advertisements may be selected in relation to a request based on various criteria. For example, an advertisement may be selected based on words or other content of the request, relevant words or content related to the words or content of the request, etc. In another example, the advertisement may be selected based on requested tasks/information (e.g., a request for movie showtimes may result in an advertisement being selected for a particular theater). In yet another example, the advertisement may be selected based on a topic or category associated with the requested tasks/information (e.g., a request to purchase airline tickets may result in an advertisement being selected for a hotel in a destination associated with a reserved flight). In still other examples, the advertisement may be selected based on location information, (e.g., advertisements may be selected based on a proximity to a user geolocation identified using a navigation device), user-specific and/or global user profiles (e.g., advertisements may be selected based on user-specific and/or global preferences, advertiser campaign criteria, etc.).

[031] Content of a voice-based input may be determined based on various criteria, including contextual or conceptual information (e.g., semantic indicators, qualifiers, or other information). For example, a given concept may include various semantically equivalent indicators having an identical meaning. Thus, for instance, a voice-based input may be "Play some tunes!" or "Play some music!" or other variants thereof, each of which may be interpreted as relating to a specific idea (or concept) of "Music." Thus, concept or content information in a request may be used to select an advertisement. For example, a user may request to calculate a route in Seattle, Washington (e.g., "How do I get to the Space Needle?"). Based on a context of the requested task (e.g., "Navigation," "Seattle," etc.), a voice search engine may retrieve an address of the Space Needle and a navigation application may calculate the route. Furthermore, user profile information may indicate that the user is visiting Seattle from out-of-town (e.g., the profile may indicate that the user's home is Sacramento), and therefore, an advertisement for popular points-of-interest

in Seattle may be selected. In another example, the user may request information about a sporting event (e.g., "Get me the kickoff time for the Eagles game on Sunday"). Based on a context of the requested information (e.g., "Search," "Sports," "Philadelphia," etc.), the requested information may be retrieved, while an advertisement for Eagles apparel or memorabilia may be selected.

[032] In various instances, concepts, semantic indicators, qualifiers, or other information included in, or inferred from, a request may indicate an exploratory nature for the request. In other words, the exploratory request may identify a goal for a conversation, instead of a particular task to perform or information to retrieve. As such, in various implementations, an advertisement may be selected in operation 320 in an effort to advance the conversation towards the goal. For example, an exploratory request may include a request for a navigation route (e.g., "I feel like going to a museum, find me something interesting"). Based on a context of the requested task (e.g., "Navigation," "Points of Interest," etc.), the goal of the conversation may be identified, and the request may be serviced in operation 315 (e.g., a voice search engine may locate nearby points of interest based on user preferred topics). Further, the advertising application may select an appropriate advertisement in operation 320, where the advertisement may be selected in an attempt to advance the conversation towards the goal. For example, statistical profiles (e.g., user profiles, global profiles, topic-based profiles, etc.) may reflect an affinity between an advertisement for a particular museum and other users sharing similar demographics or other characteristics with the requesting user. Thus, in addition to retrieving information about museums in operation 315, an advertisement for a museum likely to be of interest to the user may be selected in operation 320.

[033] In various instances, a request may include incomplete, ambiguous, unrecognized, or otherwise insufficient semantic indicators, context, qualifiers, or other information needed to identify the request. In other words, the request may include inadequate information to identify or infer a task to perform, information to retrieve, or a goal for a conversation. Thus, as much information as possible may be extracted and/or inferred from the request based on shared knowledge such as context, user or global profile information, previous utterances, previous conversations, etc. As such, servicing the request may include generating a response and/or communicating with an advertising application to advance a conversation toward a serviceable request. For example, servicing the request in operation

315 and selecting an advertisement in operation 320 may include generating a response and/or selecting an advertisement to frame a subsequent user input, thereby advancing the conversation.

[034] For example, the request may include incomplete, ambiguous, or unrecognized information (e.g., "Do you know [*mumbled words*] Seattle?"). A context of the requested task may be identified (e.g., "Seattle"), yet the identified context may be insufficient to adequately take action to service the request. Additional information may be inferred based on previous utterances in the conversation, profile information, or other information. However, when the additional information fails to provide adequate information to infer a reasonable hypothesis, servicing the request in operation 315 may include generating a response to frame a subsequent user input and advance the conversation (e.g., information about various topics may be retrieved based on a user's preferred topics). Further, the advertising application may select an advertisement in operation 320 to advance the conversation (e.g., advertisements may be selected based on user and/or global profiles reflecting an affinity between certain advertisements associated with Seattle and user preferences, profiles, etc.). Thus, by selecting an advertisement, indicating dissatisfaction with an advertisement, or otherwise interacting with an advertisement, the interaction may be used to build context and shared knowledge for a subsequent course of the conversation. For example, a user may select an advertisement, and an interpretation of a subsequent voice-based input (e.g., "Call them," "What's the price range?" etc.) may be interpreted with shared knowledge of the advertisement that the voice-based input relates to. Thus, advertisements may be used in a way that enables advertisers to market to consumers, while also improving the consumers' interaction with a device. Other advantages will be apparent to those skilled in the art.

[035] It will be apparent that operation 320 may use various techniques to select advertisements based on voice-based inputs and/or requests included therein. For example, an advertiser may specify a target audience, marketing criteria, campaign strategies, budget constraints, concepts, semantic indicators, related topics, categories, and/or any other suitable information to associate with an advertisement. For instance, advertisers may pay a premium to prioritize an advertisement in relation to similar advertisements (e.g., advertisements associated with competitors). In another example, various statistical profiles may define affinities between advertisements, topics, users, etc.

(e.g., based on click-through or conversion rates, or other tracking information, as described in more detail below). Thus, advertisements may be selected in operation 320 using various techniques, including content of the request, an activity/action associated with the request, user profiles, user preferences, statistical metrics, advertiser-specified criteria, to advance a conversation, to resolve ambiguous requests, or in various other ways, as will be apparent.

[036] The output presented to the user in operation 325 may be provided to the user in various ways. For example, in various implementations, the output may include a voice-based or otherwise audible response. In another example, when an associated device includes a display mechanism, the output may be displayed on the display device. It will be apparent that many combinations or variants thereof may be used, such as augmenting a voice-based response with information on a display device. For example, a user may request information about restaurants, and an advertisement may be selected based on a user preference indicating a favorite type of restaurant (e.g., a Chinese restaurant may be selected based on a user profile indicating a preference for Chinese). Therefore, in one example, the output presented in operation 325 may display information about various restaurants matching the requested information, while a voice-based advertisement for the Chinese restaurant may be played to the user (e.g., via a speaker or other suitable mechanism for playing voice back to the user). Many other variations will be apparent (e.g., a graphical advertisement may be displayed on a display device, while a corresponding or different voice-based advertisement may be played audibly).

[037] Subsequent interaction between the user and the presented advertisements may be monitored in a decisional operation 330. For instance, when the user elects to interact with the advertisement, action may be taken based on the interaction in an operation 335. The interaction may take various forms, including additional voice-based inputs or other suitable mechanisms for interacting with advertisements (e.g., clicking on an advertisement displayed on a personal digital assistant using an associated stylus). For example, a user may initially request information from a voice-enabled media device (e.g., a satellite radio player) about a song currently playing (e.g., "What is this song?"). In addition to outputting the requested information about the song (e.g., "This song is Double Barrel by Dave and Ansel Collins."), a selected advertisement may enable the user to purchase a ringtone for a mobile phone that corresponds to the song. In this example, the interaction may include a request to purchase the ringtone (e.g., "Yeah, I'll buy that"), and action taken in operation

335 may include completing a transaction for the ringtone and/or downloading the ringtone to the mobile phone. Furthermore, additional advertisements may be selected in an operation 340 based on the interaction, using similar techniques as described in connection with operation 320 (e.g., advertisements for additional ringtones, similar musicians, etc. may be selected). Processing may subsequently return to operation 325 to present output resulting from the interaction.

[038] User advertisement interaction may be tracked in an operation 345. For example, operation 345 may track historical data about users, conversations, topics, contexts, or other criteria to associate information with the selected advertisement. The tracking information may therefore be used to build statistical profiles defining affinities, click-through or conversion rates, or other information about various advertisements, topics, or other criteria on a user-specific and/or a global-user level. Thus, clusters or mappings may be created between advertisements, topics, concepts, demographics, or other criteria based on user behavior with the advertisements (e.g., whether a user interacts with the advertisement in operation 330).

[039] For instance, certain advertisements may experience high click-through rates in relation to a first context and/or topic, but low click-through rates in relation to a second context and/or topic, and therefore, when requests relate to the first context and/or topic, the advertisement may be more likely to be selected in subsequent operations 320/340. In another example, global statistical profiles may indicate that an advertisement experiences more click-throughs by users of a particular demographic, and therefore, the advertisement may be more likely to be selected for users falling within the demographic. Many different techniques for tracking and building statistical profiles will be apparent.

[040] Implementations of the invention may be made in hardware, firmware, software, or any combination thereof. The invention may also be implemented as instructions stored on a machine-readable medium, which may be read and executed by one or more processors. A machine-readable medium may include any mechanism for storing or transmitting information in a form readable by a machine (e.g., a computing device). For example, a machine-readable storage medium may include read only memory, random access memory, magnetic disk storage media, optical storage media, flash memory devices, and others, and a machine-readable transmission media may include forms of propagated signals, such as carrier waves, infrared signals, digital signals, and others. Further, firmware, software,

routines, or instructions may be described in the above disclosure in terms of specific exemplary aspects and implementations of the invention, and performing certain actions. However, it will be apparent that such descriptions are merely for convenience and that such actions in fact result from computing devices, processors, controllers, or other devices executing the firmware, software, routines, or instructions.

[041] Aspects and implementations may be described as including a particular feature, structure, or characteristic, but every aspect or implementation may not necessarily include the particular feature, structure, or characteristic. Further, when a particular feature, structure, or characteristic is described in connection with an aspect or implementation, it will be apparent to effect such feature, structure, or characteristic in connection with other aspects or implementations whether or not explicitly described. Thus, various changes and modifications may be made, without departing from the scope and spirit of the invention. The specification and drawings are to be regarded as exemplary only, and the scope of the invention is to be determined solely by the appended claims.

CLAIMS

What is claimed is:

1. A method for selecting and presenting advertisements in response to voice-based inputs, the method comprising:

receiving a voice-based input, the received input including a request;

retrieving information from at least one knowledge source based on the received request;

selecting an advertisement based on the received request and the retrieved information; and

presenting the selected advertisement to a user.

2. The method of claim 1, wherein the retrieved information is related to one or more of content of the voice-based input, an action associated with the request, a location of the user, a user-specific profile, or a global user profile.

3. The method of claim 1, wherein selecting the advertisement includes identifying an advertisement related to resolving the request when the received input includes insufficient information to resolve the request.

4. The method of claim 1, further comprising tracking an interaction between the user and the presented advertisement.

5. The method of claim 4, wherein the tracking is used to build statistical profiles of one or more users.

6. The method of claim 5, wherein the statistical profiles identify affinities among one or more advertisements, concepts, categories, and/or user profiles.

7. The method of claim 4, wherein the tracking is used to build a context for interpreting subsequent voice-based inputs.

8. The method of claim 4, further comprising taking an action in response to the interaction.

9. The method of claim 8, wherein the action includes performing a task and/or retrieving information based on the interaction.

10. The method of claim 1, wherein the at least one knowledge source includes public information generic to the user, and private information specific to the user.

11. A system for selecting and presenting advertisements in response to voice-based inputs, the system comprising:

at least one input device that receives a voice-based input, the received input including a request; and

at least one processor operable to:

retrieve information from at least one knowledge source based on the received request;

select an advertisement based on the received request and the retrieved information; and

present the selected advertisement to a user.

12. The system of claim 11, wherein the retrieved information is related to one or more of content of the voice-based input, an action associated with the request, a location of the user, a user-specific profile, or a global user profile.

13. The system of claim 11, wherein selecting the advertisement includes identifying an advertisement related to resolving the request when the received input includes insufficient information to resolve the request.

14. The system of claim 11, further comprising tracking an interaction between the user and the presented advertisement.

15. The system of claim 14, wherein the tracking is used to build statistical profiles of one or more users.

16. The system of claim 15, wherein the statistical profiles identify affinities among one or more advertisements, concepts, categories, and/or user profiles.

17. The system of claim 14, wherein the tracking is used to build a context for interpreting subsequent voice-based inputs.

18. The system of claim 14, further comprising taking an action in response to the interaction.

19. The system of claim 18, wherein the action includes performing a task and/or retrieving information based on the interaction.

20. The system of claim 11, wherein the at least one knowledge source includes public information generic to the user, and private information specific to the user.

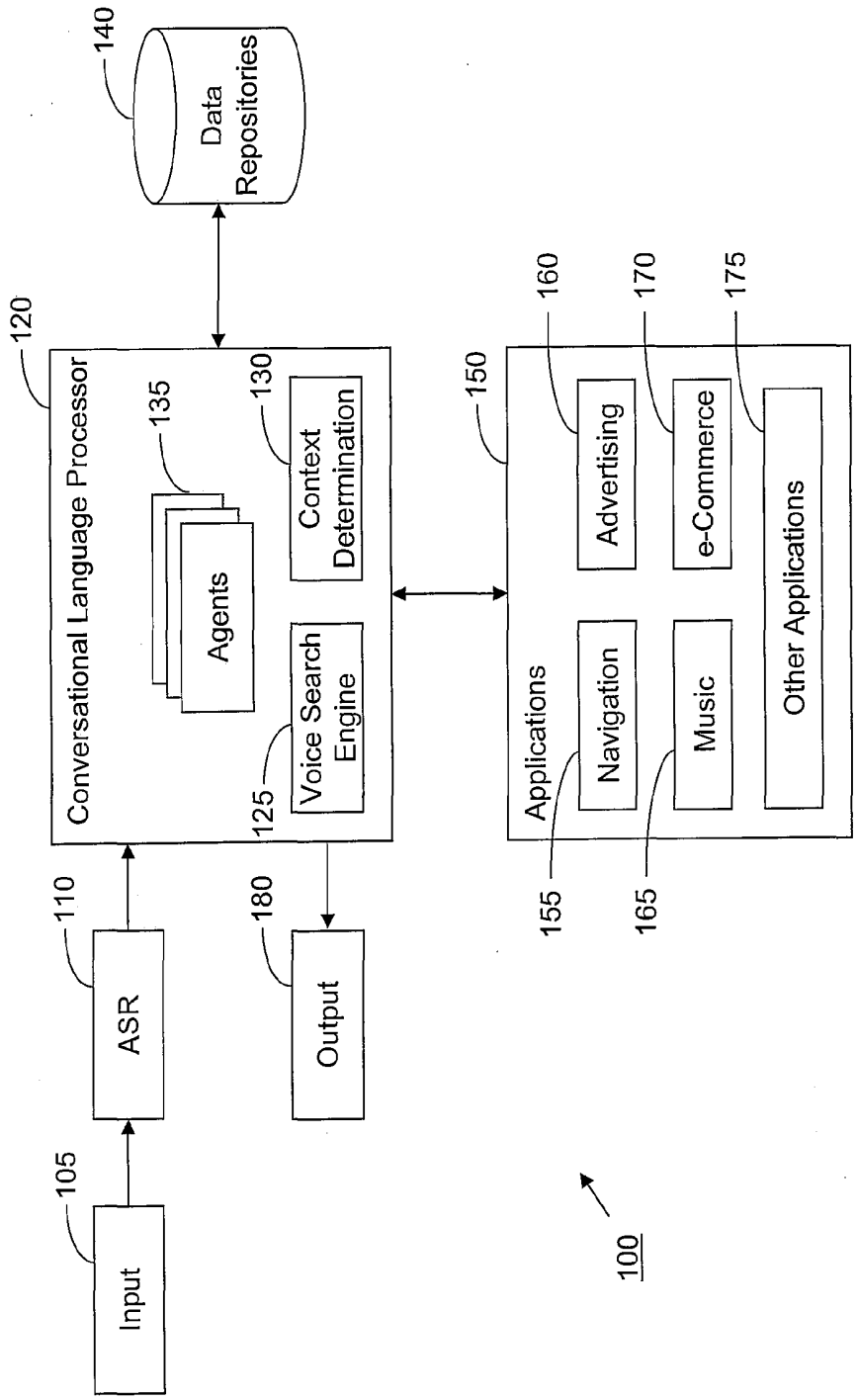

140
Data Repositories
120
Conversational Language Processor
135
Agents
130
Context Determination
125
Voice Search Engine
150
Applications
160
Advertising
170
e-Commerce
175
Other Applications
Navigation
Music
155
165
110
ASR
180
Output
105
Input
100

Fig. 1

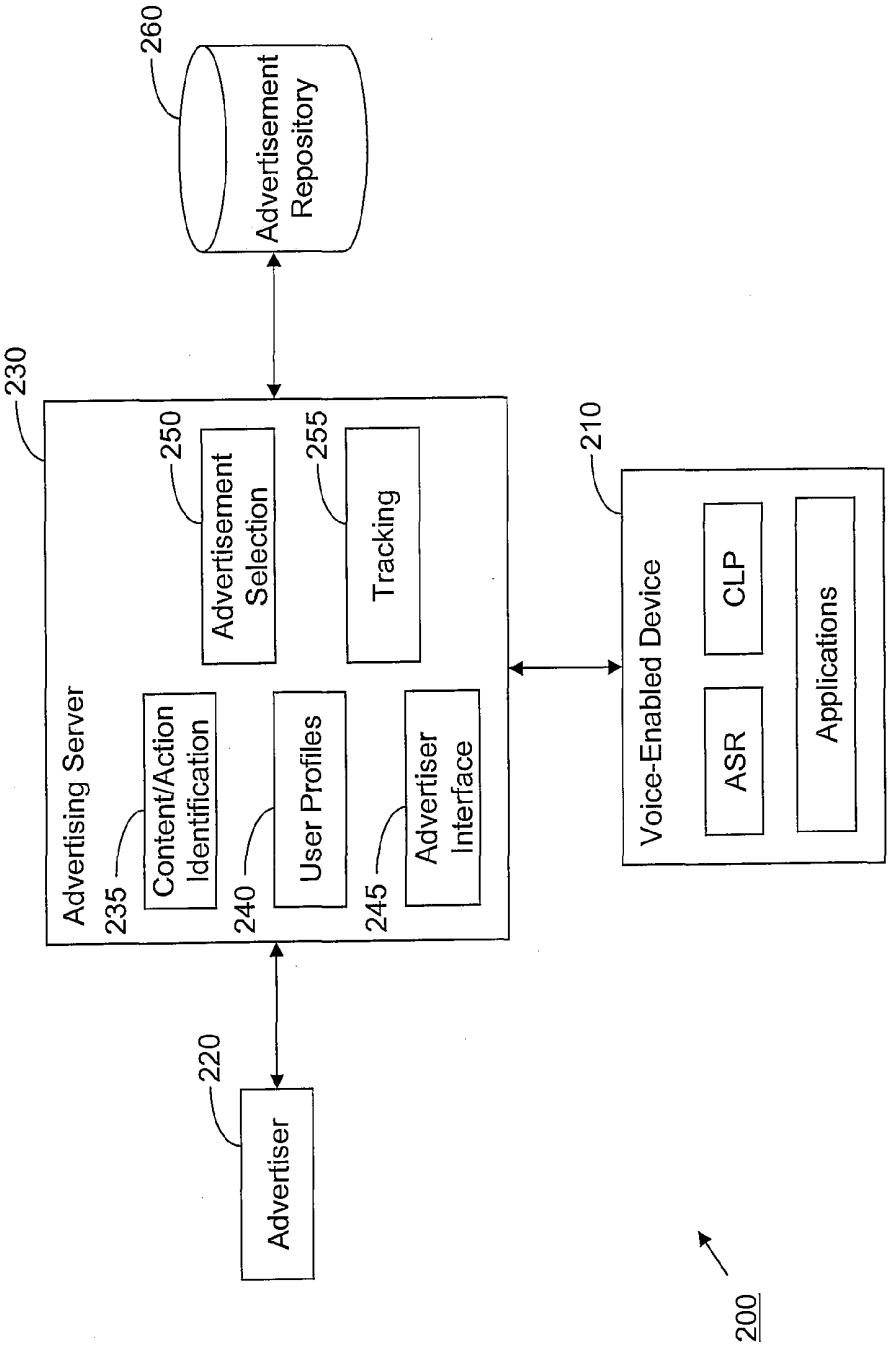

260
Advertisement Repository
230
Advertising Server
250
Advertisement Selection
255
Tracking
235
Content/Action Identification
240
User Profiles
245
Advertiser Interface
210
Voice-Enabled Device
ASR
CLP
Applications
220
Advertiser
200

Fig. 2

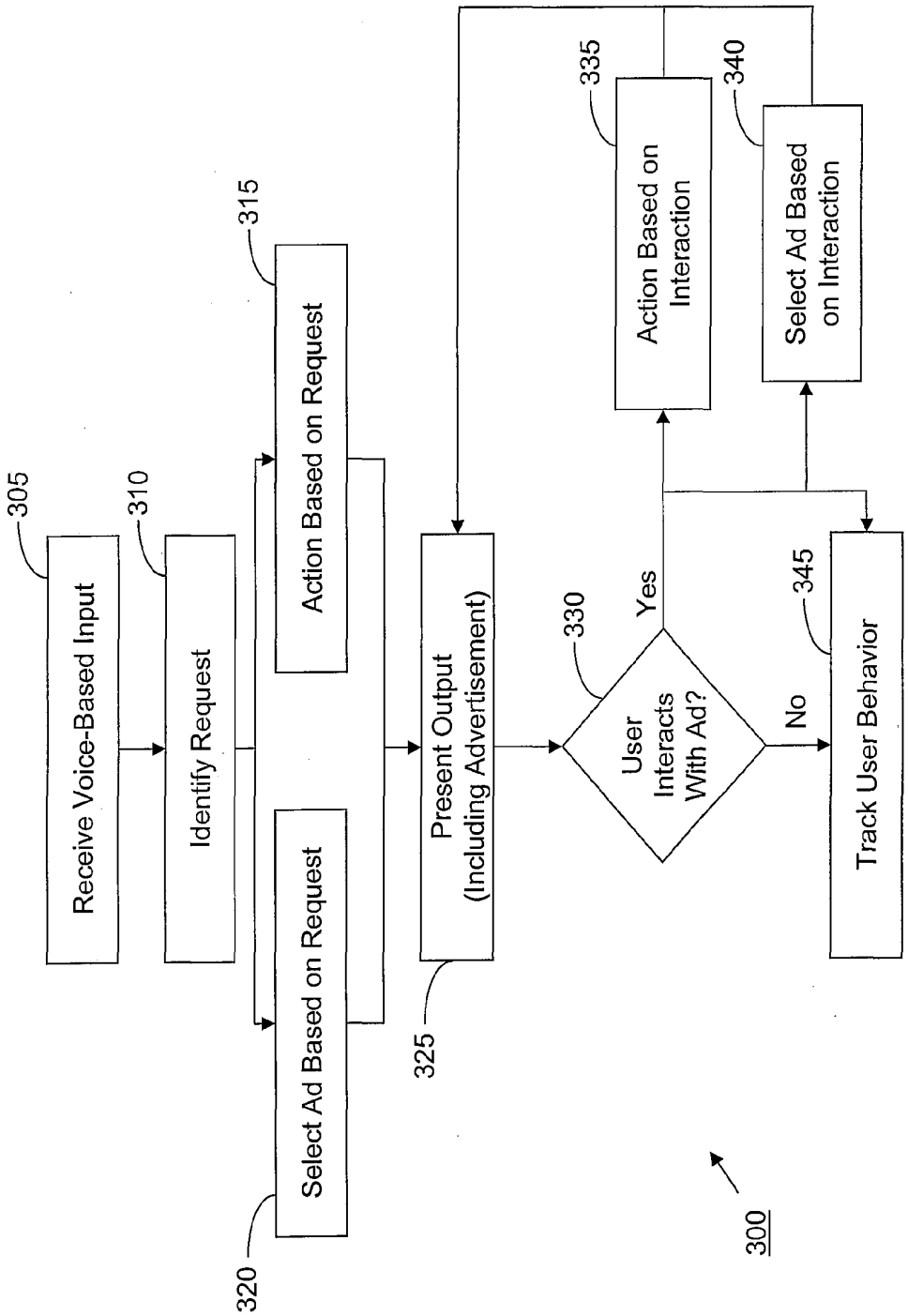

Receive Voice-Based Input
305
Identify Request
310
Action Based on Request
315
Select Ad Based on Request
320
Present Output (Including Advertisement)
325
User Interacts With Ad?
330
Yes
No
Action Based on Interaction
335
Select Ad Based on Interaction
340
Track User Behavior
345
300

Fig. 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 08/53157

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06Q 30/00
USPC - 705/14
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8) - G06Q 30/00 (2008.04)
USPC - 705/14

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
705/1, 7, 8; 725/37, 38, 42

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PubWEST(USPT,PGPB,EPAB,JPAB); Google Scholar; Google
Search Terms Used: method to provide/select/display advertisement, advertisement, voice input, knowledge source, database, track, profile, user profile, advertisement database/source

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,696,965 A (DEDRICK) 09 December 1997 (09.12.1997) entire document, especially col 4, ln 60-66, col 5, ln 10-20, col 6, ln 12-30 and col 8, ln 24 - col 9, ln 10.	1-20

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 April 2008 (24.04.2008)	22 MAY 2008
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (April 2007)

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
2 October 2008 (02.10.2008)

PCT

(10) International Publication Number
WO 2008/118195 A2

(51) International Patent Classification:
G06F 17/28 (2006.01)

(21) International Application Number:
PCT/US2007/081481

(22) International Filing Date: 16 October 2007 (16.10.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/580,926 16 October 2006 (16.10.2006) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [US/US]; 11980 Ne 24th Street, Bellevue, WA 98005 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **BALDWIN, Larry** [US/US]; 25498 S.e. 274th Place, Maple Valley, WA 98038 (US). **FREEMAN, Tom** [US/US]; 6735 83rd Avenue S.e., Mercer Island, WA 98040 (US). **TJALVE, Michael** [DK/US]; 14904 S.e. 47th Court, Bellevue, WA 98006 (US). **EBERSOLD, Blane** [US/US]; 4721 Admiral Way, S.w., Seattle, WA 98116 (US). **WEIDER, Chris** [US/US]; 2618 Grand Avenue, #b-508, Everett, WA 98201 (US).

(74) Agents: **BARUFKA, Jack, S.** et al.; Pillsbury Winthrop Shaw Pittman Llp, P.O. Box 10500, Mclean, VA 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

(54) **Title:** SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

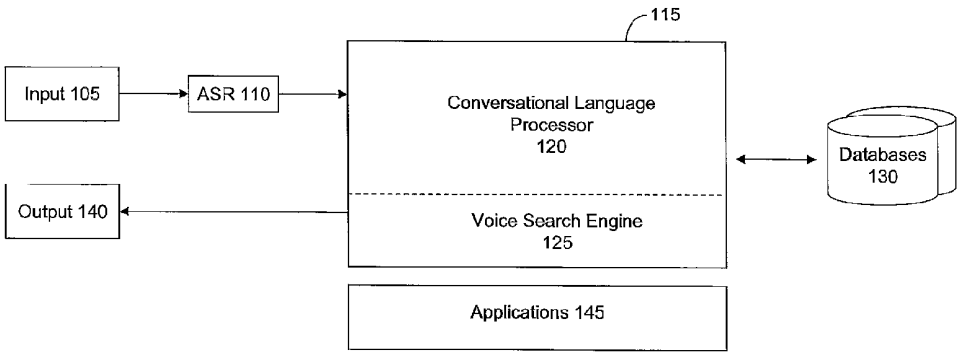


FIG. 1

(57) **Abstract:** A cooperative conversational voice user interface is provided. The cooperative conversational voice user interface may build upon short-term and long-term shared knowledge to generate one or more explicit and/or implicit hypotheses about an intent of a user utterance. The hypotheses may be ranked based on varying degrees of certainty, and an adaptive response may be generated for the user. Responses may be worded based on the degrees of certainty and to frame an appropriate domain for a subsequent utterance. In one implementation, misrecognitions may be tolerated, and conversational course may be corrected based on subsequent utterances and/or responses.

WO 2008/118195 A2

SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

FIELD OF THE INVENTION

[001] The invention relates to a cooperative conversational model for a human to machine voice user interface.

BACKGROUND OF THE INVENTION

[002] Advances in technology, particularly within the convergence space, have resulted in an increase in demand for voice recognition software that can exploit technology in ways that are intuitive to humans. While communication between human beings is most often "cooperative," in that information and/or context is shared to advance mutual conversational goals, existing Human-to-Machine interfaces fail to provide the same level of intuitive interaction. For example, each human participant in a conversation can contribute to an exchange for the benefit of the exchange. This is done through shared assumptions and expectations regarding various aspects of the conversation, such as the topic, participant knowledge about the topic, expectations of the other participant's knowledge about the topic, appropriate word usage for the topic and/or participants, conversational development based on previous utterances, the participants' tone or inflection, the quality and quantity of contribution expected from each participant, and many other factors. Participating in conversations that continually build and draw upon shared information is a natural and intuitive way for humans to converse.

[003] In contrast, complex Human-to-Machine interfaces do not allow users to exploit technology in an intuitive way, which inhibits mass-market adoption for various technologies. Incorporating a speech interface helps to alleviate this burden by making interaction easier and faster, but existing speech interfaces (when they actually work) still require significant learning on the part of the user. That is, existing speech interfaces are unable to bridge the gap between archaic Human-to-Machine interfaces and conversational speech that would make interaction with systems feel normal. Users should be able to directly request what they want from a system in a normal, conversational fashion, without having to memorize exact words or phrases. Alternatively, when users are uncertain of particular needs, they should be able to engage the system in a productive, cooperative dialogue to resolve their

requests. Instead, existing speech interfaces force users to dumb down their requests to match simple sets of instructions in simple languages in order to communicate requests in ways that systems can understand. Using existing speech interfaces, there is virtually no option for dialogue between the user and the system to satisfy mutual goals.

[004] Therefore, existing systems lack a conversational speech model that can provide users with the ability to interact with systems in ways that are inherently intuitive to human beings. Existing systems suffer from these and other problems.

SUMMARY OF THE INVENTION

[005] According to various embodiments and aspects of the invention, a cooperative conversational voice user interface may understand free form human utterances, freeing users from being restricted to a fixed set of commands and/or requests. Rather, users can engage in cooperative conversations with a machine to complete a request or series of requests using a natural, intuitive, free form manner of expression.

[006] According to an aspect of the invention, an exemplary system architecture for implementing a cooperative conversational voice user interface is provided. The system may receive an input, which may include a human utterance received by an input device, where the utterance may include one or more requests. As used herein, an "utterance" may be words, syllables, phonemes, or any other audible sound made by a human being. As used herein, a "request" may be a command, directive, or other instruction for a device, computer, or other machine to retrieve information, perform a task, or take some other action. In one implementation, the input may be a multi-modal input, where at least part of the multi-modal input is an utterance. The utterance component of the input may be processed by an Automatic Speech Recognizer to generate one or more preliminary interpretations of the utterance. The one or more preliminary interpretations may then be provided to a conversational speech engine for further processing, where the conversational speech engine may communicate with one or more databases to generate an adaptive conversational response, which may be returned to the user as an output. In one implementation, the output may be a multi-modal output. For example, the utterance may include a request to perform an action, and the output may include a conversational response reporting success or failure, as well as an execution of the action.

[007] According to another aspect of the invention, an exemplary conversational

speech engine may generate an adaptive conversational response to a request or series of requests. The conversational speech engine may include a free form voice search module that may understand an utterance made using typical, day-to-day language (i.e., in free form), and may account for variations in how humans normally speak, the vocabulary they use, and the conditions in which they speak. To account for intangible variables of human speech, the free form search module may include models of casual human speech. For example, in one implementation, the free form search module may understand specialized jargon and/or slang, tolerate variations in word order, and tolerate verbalized pauses or stuttered speech. For example, formalized English requests, where a verb precedes a noun, may be treated in an equivalent manner to requests where the noun precedes the verb. In another implementation, compound requests and/or compound tasks with multiple variables may be identified in a single utterance. By identifying all relevant information for completing one or more tasks from a single utterance, advantages may be provided over existing voice user interfaces, such as Command and Control systems that use verbal menus to restrict information that a person can provide at a given point. In another implementation, inferring intended requests from incomplete or ambiguous requests may provide a conversational feel. By modeling what contextual signifiers, qualifiers, or other information may be required to perform a task in an identified context, an adaptive response may be generated, such as prompting a user for missing contextual signifiers, qualifiers, or other information. In one implementation, the response may ask for missing information in a way that most restricts possible interpretations, and the response may be framed to establish a domain for a subsequent user utterance. In another implementation, common alternatives for nouns and verbs may be recognized to reflect variations in usage patterns according to various criteria. Thus, variations in expression may be supported because word order is unimportant or unanticipated, and nouns and/or verbs may be represented in different ways to give simplistic, yet representative, examples. In another implementation, requests may be inferred from contradictory or otherwise inaccurate information, such as when an utterance includes starts and stops, restarts, stutters, run-on sentences, or other imperfect speech. For example, a user may sometimes change their mind, and thus alter the request in mid-utterance, and the imperfect speech feature may nonetheless be able to infer a request based on models of human speech. For example, various models may indicate that a last criterion is most likely to be correct, or intonation, emphasis, stress, use of the word "not," or other

models may indicate which criterion is most likely to be correct.

[008] According to another aspect of the invention, the conversational speech engine may include a noise tolerance module that may discard words or noise which has no meaning in a given context to reduce a likelihood of confusion. Moreover, the noise tolerance module may filter out environmental and non-human noise to further reduce a likelihood of confusion. In one implementation, the noise tolerance module may cooperate with other modules and features to filter out words that do not fit into an identified context. For example, the noise tolerance module may filter other human conversations and/or utterances within a range of one or more microphones. For example, a single device may include multiple microphones, or multiple devices may each include one or more microphones, and the noise tolerance module may collate inputs and cooperatively filter out sound by comparing a speech signal from the various microphones. The noise tolerance module may also filter out non-human environmental noise within range of the microphones, out-of-vocabulary words caused by speaker ambiguity or malapropisms, or other noise that may be unrelated to a target request. Performance benchmarks for the noise tolerance module may be defined by noise models based on human criteria. For example, if a driver of a car is 92% likely to be understood by a passenger when traveling at 65 miles-per-hour with windows cracked, then performance benchmarks for the noise tolerance module may have a similar performance under such conditions.

[009] According to another aspect of the invention, the conversational speech engine may include a context determination process that determines one or more contexts for a request to establish meaning within a conversation. The one or more contexts may be determined by having one or more context domain agents compete to determine a most appropriate domain for a given utterance. Once a given domain "wins" the competition, the winning domain may be responsible for establishing or inferring further contexts and updating short-term and long-term shared knowledge. If there is a deadlock between context domain agents, an adaptive conversational response may prompt the user to assist in disambiguating between the deadlocked agents. Moreover, the context determination process may infer intended operations and/or context based on previous utterances and/or requests, whereas existing systems consider each utterance independently, potentially making the same errors over and over again. For example, if a given interpretation turns

out to be incorrect, the incorrect interpretation may be removed as a potential interpretation from one or more Automatic Speech Recognizer grammars and/or from possible interpretations determined by the conversational speech engine, thereby assuring that a mistake will not be repeated for an identical utterance.

[010] The context determination process may provide advantages over existing voice user interfaces by continually updating one or more models of an existing context and establishing context as a by-product of a conversation, which cannot be established a priori. Rather, the context determination process may track conversation topics and attempt to fit a current utterance into recent contexts, including switching between contexts as tasks are completed, partially completed, requested, etc. The context determination process may identify one or more context domains for an utterance by defining a collection of related functions that may be useful for users in various context domains. Moreover, each context domain may have relevant vocabularies and thought collections to model word groupings, which when evaluated together, may disambiguate one context domain from another. Thus, eliminating out-of-context words and noise words when searching for relevant combinations may enhance accuracy of inferences. This provides advantages over existing systems that attempt to assign meaning to every component of an utterance (i.e., including out-of-context words and noise words), which results in nearly infinite possible combinations and greater likelihood of confusion. The context determination process may also be self-aware, assigning degrees of certainty to one or more generated hypotheses, where a hypothesis may be developed to account for variations in environmental conditions, speaker ambiguity, accents, or other factors. By identifying a context, capabilities within the context, vocabularies within the context, what tasks are done most often historically in the context, what task was just completed, etc., the context determination process may establish intent from rather meager phonetic clues. Moreover, just as in human-to-human conversation, users may switch contexts at any time without confusion, enabling various context domains to be rapidly selected, without menu-driven dead ends, when an utterance is unambiguous.

[011] According to another aspect of the invention, an exemplary cooperative conversational model may build upon free form voice search, noise tolerance, and context determination to implement a conversational Human-to-Machine interface that reflects human interaction and normal conversational behavior. That is, the cooperative

conversational model enables humans and machines to participant in a conversation with an accepted purpose or direction, with each participant contributing to the conversation for the benefit of the conversation. By taking advantage of human presumptions about utterances that humans rely upon, both as speakers and listeners, a Human-to-Machine interface may be analogous to everyday human-to-human conversation. In one implementation, the exemplary cooperative conversation model may take incoming data (shared knowledge) to inform a decision (intelligent hypothesis building), and then may refine the decision and generate a response (adaptive response building).

[012] According to another aspect of the invention, shared knowledge may include both short-term and long-term knowledge. Short-term knowledge may accumulate during a single conversation, where input received during a single conversation may be retained. The shared knowledge may include cross-modality awareness, where in addition to accumulating input relating to user utterances, requests, locations, etc., the shared knowledge may accumulate a current user interface state relating to other modal inputs to further build shared knowledge models. The shared knowledge may be used to build one or more intelligent hypotheses using current and relevant information, build long-term shared knowledge by identifying information with long-term significance, and generate adaptive responses with relevant state and word usage information. Moreover, because cooperative conversations model human conversations, short-term session data may be expired after a psychologically appropriate amount of time, thereby humanizing system behavior, reducing a likelihood of contextual confusion based on stale data, while also adding relevant information from an expired session context to long-term knowledge models. Long-term shared knowledge may generally be user-centric, rather than session-based, where inputs may be accumulated over time to build user, environmental, cognitive, historical, or other long-term knowledge models. Long-term and short-term shared knowledge may be used simultaneously anytime a user engages in a cooperative conversation. Long-term shared knowledge may include explicit and/or implicit user preferences, a history of recent contexts, requests, tasks, etc., user-specific jargon related to vocabularies and/or capabilities of a context, most often used word choices, or other information. The long-term shared knowledge may be used to build one or more intelligent hypotheses using current and relevant information, generate adaptive responses with appropriate word choices when unavailable via short-term shared knowledge, refine long-term shared knowledge models, identify a frequency of

specific tasks, identify tasks a user frequently has difficulty with, or provide other information and/or analysis to generate more accurate conversational responses. Shared knowledge may also be used to adapt a level of unprompted support (e.g., for novices versus experienced users, users who are frequently misrecognized, etc.) Thus, shared knowledge may enable a user and a voice user interface to share assumptions and expectations such as topic knowledge, conversation history, word usage, jargon, tone, or other assumptions and/or expectations that facilitate a cooperative conversation between human users and a system.

[013] According to another aspect of the invention, a conversation type may be identified for any given utterance. Categorizing and developing conceptual models for various types of exchanges may consistently align user expectations and domain capabilities. One or more intelligent hypotheses may be generated as to a conversation type by considering conversational goals, participant roles, and/or an allocation of information among the participants. Based on the conversational goals, participant roles, and allocation of information, the intelligent hypotheses may consider various factors to classify a conversation (or utterance) into general types of conversations that can interact with one another to form many more variations and permutations of conversation types (e.g., a conversation type may change dynamically as information is reallocated from one participant to another, or as conversational goals change based on the reallocation of information).

[014] According to another aspect of the invention, the intelligent hypotheses may include one or more hypotheses of a user's intent in an utterance. In addition, the intelligent hypotheses may use short-term and/or long-term shared knowledge to proactively build and evaluate interaction with a user as a conversation progresses or over time. The hypotheses may model human-to-human interaction to include a varying degree of certainty for each hypothesis. That is, just as humans rely on knowledge shared by participants to examine how much and what kind of information was available, the intelligent hypotheses may leverage the identified conversation type and shared knowledge to generate a degree of certainty for each hypothesis.

[015] According to another aspect of the invention, syntactically, grammatically, and contextually sensitive "intelligent responses" may be generated from the intelligent hypotheses that can be used to generate a conversational experience for a user, while

also guiding the user to reply in a manner favorable for recognition. The intelligent responses may create a conversational feel by adapting to a user's manner of speaking, framing responses appropriately, and having natural variation and/or personality (e.g., by varying tone, pace, timing, inflection, word use, jargon, and other variables in a verbal or audible response).

[016] According to another aspect of the invention, the intelligent responses may adapt to a user's manner of speaking by using contextual signifiers and grammatical rules to generate one or more sentences that may cooperate with the user. By taking advantage of shared knowledge about how a user utters a request, the responses may be modeled using similar techniques used to recognize requests. The intelligent responses may rate possible responses statistically and/or randomize responses, which creates an opportunity to build an exchange with natural variation and conversational feel. This provides advantages over existing voice user interfaces where input and output is incongruous, as the input is "conversational" and the output is "computerese."

[017] According to another aspect of the invention, the intelligent responses may frame responses to influence a user reply utterance for easy recognition. For example, the responses may be modeled to illicit utterances from the user that may be more likely to result in a completed request. Thus, the responses may conform to a cooperative nature of human dialog and a natural human tendency to "parrot" what was just heard as part of a next utterance. Moreover, knowledge of current context may enhance responses to generate more meaningful conversational responses. Framing the responses may also deal with misrecognitions according to human models. For example, humans frequently remember a number of recent utterances, especially when one or more previous utterances were misrecognized or unrecognized. Another participant in the conversation may limit correction to a part of the utterance that was misrecognized or unrecognized, or over subsequent utterances and/or other interactions, clues may be provided to indicate the initial interpretation was incorrect. Thus, by storing and analyzing multiple utterances, utterances from earlier in a conversation may be corrected as the conversation progresses.

[018] According to another aspect of the invention, the intelligent responses may include multi-modal, or cross-modal, responses to a user. In one implementation, responses may be aware of and control one or more devices and/or interfaces, and users may respond by using whichever input method, or combination of input methods, is most

convenient.

[019] According to another aspect of the invention, the intelligent responses may correct a course of a conversation without interrupting conversational flow. That is, even though the intelligent responses may be reasonably "sure," the intelligent responses may nonetheless sometimes be incorrect. While existing voice user interfaces tend to fail on average conversational missteps, normal human interactions may expect missteps and deal with them appropriately. Thus, responses after misrecognitions may be modeled after clarifications, rather than errors, and words may chosen in subsequent responses to move conversation forward and establish an appropriate domain to be explored with the user.

[020] Other objects and advantages of the invention will be apparent to those skilled in the art based on the following drawings and detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

[021] Fig. 1 is an exemplary block diagram of a system architecture according to one aspect of the invention.

[022] Fig. 2 is an exemplary block diagram of a conversational speech engine according to one aspect of the invention.

[023] Fig. 3 is an exemplary block diagram of a cooperative conversational model according to one aspect of the invention.

DETAILED DESCRIPTION

[024] Referring to Fig. 1, an exemplary system architecture for implementing a cooperative conversational voice user interface is illustrated according to one aspect of the invention. The system may receive an input 105 from a user, where in one implementation, input 105 may be an utterance received by an input device (e.g., a microphone), where the utterance may include one or more requests. Input 105 may also be a multi-modal input, where at least part of the multi-modal input is an utterance. For example, the input device may include a combination of a microphone and a touch-screen device, and input 105 may include an utterance that includes a request relating to a portion of a display on the touch-screen device that the user is touching. For instance, the touch-screen device may be a navigation device, and input 105 may include an utterance

of "Give me directions to here," where the user may be requesting directions to a desired destination on the display of the navigation device.

[025] The utterance component of input 105 may be processed by an Automatic Speech Recognizer 110 to generate one or more preliminary interpretations of the utterance. Automatic Speech Recognizer 110 may process the utterance using any suitable technique known in the art. For example, in one implementation, Automatic Speech Recognizer 110 may interpret the utterance using techniques of phonetic dictation to recognize a phoneme stream, as described in copending U.S. Patent Application Ser. No. 11/513,269, entitled "Dynamic Speech Sharpening," which is hereby incorporated by reference in its entirety. The one or more preliminary interpretations generated by Automatic Speech Recognizer 110 may then be provided to a conversational speech engine 115 for further processing. Conversational speech engine 115 may include a conversational language processor 120 and/or a voice search engine 125, described in greater detail in Fig. 2 below. Conversational speech engine 115 may communicate with one or more databases 130 to generate an adaptive conversational response, which may be returned to the user as an output 140. In one implementation, output 140 may be a multi-modal output and/or an interaction with one or more applications 145 to complete the request. For example, output 140 may include a combination of an audible response and a display of a route on a navigation device. For example, the utterance may include a request to perform an action, and output 140 may include a conversational response reporting success or failure, as well as an execution of the action. In addition, in various implementations, Automatic Speech Recognizer 110, conversational speech engine 115, and/or databases 130 may reside locally (e.g., on a user device), remotely (e.g., on a server), or a hybrid model of local and remote processing may be used (e.g., lightweight applications may be processed locally while computationally intensive applications may be processed remotely).

[026] Referring to Fig. 2, an exemplary block diagram is provided illustrating a conversational speech engine 215 according to one aspect of the invention. Conversational speech engine 215 may include a conversational language processor 220 that generates an adaptive conversational response to a request or series of requests using a free form voice search module 245, a noise tolerance module 250, and/or a context determination process 255. According to one aspect of the invention, modules 245-255 may communicate with a voice search engine 225 that includes one or more context

domain agents 230 and/or one or more vocabularies 235 to aid in interpreting utterances and generating responses, as described in "Enhancing the VUE™ (Voce-User-Experience) Through Conversational Speech," by Tom Freeman and Larry Baldwin, which is herein incorporated by reference in its entirety. Conversational speech engine 215 may generate an adaptive conversational response to one or more requests, where the requests may depend on unspoken assumptions, incomplete information, context established by previous utterances, user profiles, historical profiles, environmental profiles, or other information. Moreover, conversational speech engine 215 may track which requests have been completed, which requests are being processed, and/or which requests cannot be processed due to incomplete or inaccurate information, and the response may be generated accordingly.

[027] According to one aspect of the invention, free form voice search module 245 may understand an utterance made using typical, day-to-day language (i.e., in free form), and may account for variations in how humans normally speak, the vocabulary they use, and the conditions in which they speak. Because variables such as stress, distraction, and serendipity are always different and infinitely varied, free form search module 245 may be designed with a goal of understanding that no human will come to the same Human-to-Machine interface situation in the same way twice. Thus, free form search module 245 may implement one or more features that model casual human speech. In various implementations, free form search module 245 may include, among other things, a free form utterance feature, a one-step access feature, an inferencing intended operations feature, an alternative expression feature, and/or an imperfect speech feature.

[028] The free form utterance feature may understand specialized jargon and/or slang, tolerate variations in word order (e.g., whether a subject of a request comes before or after a verb may be irrelevant), and tolerate verbalized pauses (e.g., "um," "ah," "eh," and other utterances without meaning). For example, the free form utterance feature may treat formalized English verb-before-noun requests in an equivalent manner to free form requests where a noun may precede a verb. For example, user utterances of "Change it to the Squizz" and "You know, um, that Squizz channel, ah, switch it there" may be treated equivalently (where Squizz is a channel on XM Satellite Radio). In either case, the free form utterance feature is able to identify "Squizz" as a subject of the utterance and "Change it" or "switch it" as a verb or request for the utterance (e.g., by

cooperating with context determination process 255, or other features, and identifying a relevant context domain agent 230 and/or vocabulary 235 to interpret the utterance).

[029] The one-step access feature may understand utterances that include compound requests with multiple variables. For example, a user utterance may be "What is the forecast for Boston this weekend?" The one-step access feature may identify "weather" as a context (e.g., by cooperating with context determination process 255, or other features, and identifying "forecast" as a synonym of "weather"), and search for a city equal to "Boston" and a time equal to "weekend." By identifying all relevant information for completing a task from a single utterance, the one-step access feature may overcome drawbacks of existing voice user interfaces, such as Command and Control systems that use verbal menus to restrict information that a person can provide at a given point (e.g., a Command and Control system for a phone directory service may say: "State please," . . . "City please," . . . "What listing," etc.). Moreover, some utterances may include compound requests, and the one-step access feature may decompose the compound requests into sub-tasks. For example, a user utterance of "I need to be at a meeting tomorrow in San Francisco at 8:00 am" may be decomposed into a set of sub-tasks such as (1) checking availability and reserving a flight on an evening before the meeting, (2) checking availability and reserving a hotel, (3) checking availability and reserving a car, etc., where users may further designate preferences for various tasks (e.g., first check availability on an airline for which the user is a frequent flyer). Depending on a level of shared knowledge about a user's preferences and/or historical patterns, the one-step access feature may infer additional tasks from a request. For example, in the above example, the one-step access feature may also check a weather forecast, and if the weather is "nice" (as defined by the user preferences and/or as inferred from historical patterns), the one-step access feature may schedule a tee-time at a preferred golf course in San Francisco.

[030] The inferencing intended operations feature may identify an intended request from incomplete or ambiguous requests. For example, when a user utters "Route <indecipherable> Chicago <indecipherable> here," where the user intended to say "Route calculation to Chicago from here," the inferencing intended operations feature may model what is required to calculate a route (an origination point and a destination point). Because the utterance includes the origination point and the destination point, a request to calculate a route from the user's present location to Chicago may be inferred.

Similarly, when the inferencing intended operations feature does not have sufficient information to infer a complete request, an adaptive conversational response may be generated to prompt the user for missing information. For example, when an utterance includes a request for a stock quote but not a company name (e.g., "Get me the stock price for <indecipherable>"), the response may be "What company's stock quote do you want?" The user may then provide an utterance including the company name, and the request may be completed. In one implementation, the response may ask for missing information in a way that most restricts possible interpretations (e.g., in a request for a task that requires both a city and a state, the state may be asked for first because there are fewer states than cities). Moreover, the inferencing intended operations feature may model compound tasks and/or requests by maintaining context and identifying relevant and/or missing information at both a composite and sub-task level.

[031] The alternative expression feature may recognize common alternatives for nouns and verbs to reflect variations in usage patterns according to various criteria. For example, users may vary expression based on age, socio-economics, ethnicity, user whims, or other factors. Thus, the alternative expression feature may support variations in expression where word order is unimportant or unanticipated. Alternatives in expression based on various criteria or demographics may be loaded into context domain agents 230 and/or vocabularies 235, and the alternative expression feature may update context domain agents 230 and/or vocabularies 235 based on inferred or newly discovered variations. In one implementation, conversational speech engine 215 may include a subscription interface to update changes to context domain agents 230 and/or vocabularies 235 (e.g., a repository may aggregate various user utterances and deploy updates system wide). In operation, the alternative expression feature may allow nouns and/or verbs to be represented in different ways to give simplistic, yet representative, examples. For example, a user interested in a weather forecast for Washington, D.C. may provide any of the following utterances, each of which are interpreted equivalently: "What's the weather like in DC," "Is it raining inside the Beltway," Gimme the forecast for the capital," etc. Similarly, utterances of "Go to my home," "Go home," "Show route to home," and "I would like to know my way home" may all be interpreted equivalently, where a user profile may include the user's home address and a navigation route to the home address may be calculated.

[032] The imperfect speech feature may be able to infer requests from contradictory or

otherwise inaccurate information, such as when an utterance includes starts and stops, restarts, stutters, run-on sentences, or other imperfect speech. For example, a user may sometimes change their mind, and thus alter the request in mid-utterance, and the imperfect speech feature may nonetheless be able to infer a request based on models of human speech. For example, for an utterance of "Well, I wanna . . . Mexi . . . no, steak restaurant please, I'm hungry," existing voice user interfaces make no assumptions regarding models of human speech and would be unable to infer whether the user wanted a Mexican or steak restaurant. The imperfect speech feature overcomes these drawbacks by using various models of human understanding that may indicate that a last criterion is most likely to be correct, or intonation, emphasis, stress, use of the word "not," or other models may indicate which criterion is most likely to be correct. Thus, in the above example, the imperfect speech feature may infer that the user wants a steak restaurant.

[033] According to one aspect of the invention, noise tolerance module 250 may be closely related to the imperfect speech feature, and may operate to discard words or noise that has no meaning in a given context so as not to create confusion. Moreover, noise tolerance module 250 may filter out environmental and non-human noise to further reduce a likelihood of confusion. In one implementation, noise tolerance module 250 may cooperate with other modules and features to filter out words that do not fit into a context. For example, one or more contexts may be identified, and words that have no meaning with respect to system capabilities, random human utterances without meaning and other noise may be filtered out. Thus, noise tolerance module 250 may model real-world conditions to identify meaningful requests. For example, noise tolerance module 250 may filter other human conversations and/or utterances within a range of one or more microphones, For example, a single device may include multiple microphones, or multiple devices may each include one or more microphones, and the noise tolerance module may collate inputs and cooperatively filter out sound by comparing a speech signal from the various microphones. Noise tolerance module 250 may also filter out non-human environmental noise within the range of the microphones, out-of-vocabulary words, which could be a result of speaker ambiguity or malapropisms, or other noise that may be unrelated to a target request. Noise models in noise tolerance module 250 may define performance benchmarks based on human criteria. For example, if a driver of a car, traveling at 65 miles-per-hour, with windows cracked is 92% likely to be understood by a passenger, then noise tolerance module 250 may have a similar performance under

those conditions.

[034] According to one aspect of the invention, conversational speech engine 215 may include a context determination process 255 that determines one or more contexts for a request to establish meaning within a conversation. The one or more contexts may be determined by having one or more context domain agents compete to determine a most appropriate domain for a given utterance, as described in copending U.S. Patent Application Ser. No. 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," and copending U.S. Patent Application Ser. No. 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," both of which are hereby incorporated by reference in their entirety. Once a given context domain agent "wins" the competition, the winning agent may be responsible for establishing or inferring further contexts and updating short-term and long-term shared knowledge. If there is a deadlock between context domain agents, an adaptive conversational response may prompt the user to assist in disambiguating between the deadlocked agents. For example, a user utterance of "What about traffic?" may have a distinct meaning in various contexts. That is, "traffic" may have a first meaning when the user is querying a system's media player (i.e., "traffic" would be a Rock and Roll band led by singer/songwriter Steve Winwood), a second meaning when the user is querying a search interface regarding Michael Douglas films (i.e., "traffic" would be a film directed by Steven Soderbergh), a third meaning when the user is querying a navigation device for directions to an airport (i.e., "traffic" would be related to conditions on roads along a route to the airport).

[035] Moreover, context determination process 255 may infer intended operations and/or context based on previous utterances and/or requests, whereas existing systems consider each utterance independently, potentially making the same errors over and over again. For example, if a given interpretation turns out to be incorrect, the incorrect interpretation may be removed as a potential interpretation from one or more Automatic Speech Recognizer grammars and/or from possible subsequent interpretations determined by context determination process 255, thereby assuring that a mistake will not be repeated for an identical utterance.

[036] Context determination process 255 may overcome drawbacks of existing systems by continually updating one or more models of an existing context, where establishing

context may be a by-product of a conversation, which cannot be established a priori. Context determination process 255 may establish a first context domain, change to a second context domain, change back to the first context domain, and so on, as tasks are completed, partially completed, requested, etc, and a context stack may track conversation topics and attempt to fit a current utterance into a most-recent context, next-most-recent topic, etc., traversing the context stack until a most likely intent can be established. For example, a user may utter "What's the traffic report," and context determination process 255 may establish Traffic as a context, and return an output including a traffic report, which does not happen to mention traffic on Interstate-5. The user may then utter "What about I-5?" and context determination process 255 may know that the current context is Traffic, a traffic report including information about Interstate-5 may be searched for, and the traffic report indicating that Interstate-5 is crowded may be returned as an output. The user may then utter "Is there a faster way?" and context determination module 255 may know that the current context is still Traffic, and may search for routes to a specified destination with light traffic and avoiding Interstate-5. Moreover, context determination process 255 may build context based on user profiles, environmental profiles, historical profiles, or other information to further refine the context. For example, the profiles may indicate that Interstate-5 is a typical route taken Monday through Friday.

[037] The profiles may be particularly meaningful when attempting to disambiguate between contexts where a word has different meanings in different contexts. For example, a user may utter "What's the weather in Seattle?" and context determination process 255 may establish Weather as a context, as well as establishing Seattle as an environmental context. The user may then utter "and Portland?" and context determination process 255 may return a weather report for Portland, Oregon based on the Weather and an environmental proximity between Portland, Oregon and Seattle, Washington. The user may then ask "What time does the game start?"" and a search for sports events with teams from Seattle and/or Portland may occur, with results presented conversationally according to methods described in greater detail below in Fig. 3. Correlatively, had user originally uttered "What's the weather in Portsmouth, New Hampshire," in the second utterance, context determination process 255 may instead retrieve a weather report for Portland, Maine based on an environmental proximity to New Hampshire. Moreover, when environmental profiles, contextual shared knowledge, and/or other short-term and/or long-term shared knowledge does not provide enough information to

disambiguate between possibilities, responses may prompt the user with a request for further information (e.g., "Did you mean Portland, Maine, or Portland, Oregon?").

[038] Context determination process 255 may cooperate with context domain agents 230, where each context domain agent 230 may define a collection of related functions that may be useful for users. Moreover, each context domain agent 230 may include a relevant vocabulary 235 and thought collections that model word groupings, which when evaluated together, may disambiguate one context domain from another (e.g., a Music context domain agent 230 may include a vocabulary 235 for songs, artists, albums, etc., whereas a Stock context domain agent 230 may include a vocabulary 235 for company names, ticker symbols, financial metrics, etc.). Thus, accuracy in identifying meaning may be enhanced by eliminating out-of-context words and noise words when searching for relevant combinations. In contrast, existing systems attempt to assign meaning to every component of an utterance (e.g., including out-of-context words and noise words), which results in nearly infinite possible combinations and greater likelihood of confusion. Moreover, context domain agents 230 may include metadata for each criteria to further assist in interpreting utterances, inferring intent, completing incomplete requests, etc. (e.g., a Space Needle vocabulary word may include metadata for Seattle, landmark, tourism, Sky City restaurant, etc.). Given a disambiguated criterion, context determination process 255 may thus be able to automatically determine other information needed to complete a request, discard importance of word order, and perform other enhancements for conversational speech.

[039] Context domain agents 230 may also be self-aware, assigning degrees of certainty to one or more generated hypotheses, where a hypothesis may be developed to account for variations in environmental conditions, speaker ambiguity, accents, or other factors. Conceptually, context domain agents 230 may be designed to model utterances as a hard-of-hearing person would at a noisy party. By identifying a context, capabilities within the context, vocabularies within the context, what tasks are done most often historically in the context, what task was just completed, etc., a context domain agent 230 may establish intent from rather meager phonetic clues. Moreover, the context stack may be one of a plurality of components for establishing context, and thus not a constraint upon the user. All context domains may be accessible, allowing the user to switch contexts at any time without confusion. Thus, just as in human- to-human conversation, context domains

may be rapidly selected, without menu-driven dead ends, when an utterance is unambiguous. For example, a user may utter, "Please call Rich Kennewick on his cell phone," and a system response of "Do you wish me to call Rich Kennewick on his cell?" may be generated. The user may decide at that point to call Rich Kennewick later, and instead, listen to some music. Thus, the user may then utter, "No, play the Louis Armstrong version of Body and Soul from my iPod," and a system response of "Playing Body and Soul by Louis Armstrong" may be generated as Body and Soul is played through a media player. In this example, the later utterance has no contextual connection to the first utterance, yet because request criteria in the utterances are unambiguous, contexts can be switched easily without relying on the context stack.

[040] Referring to Fig. 3, an exemplary cooperative conversational model 300 is illustrated according to an aspect of the invention. Cooperative conversational model 300 may build upon free form voice search 245, noise tolerance 250, and context determination 255 to implement a conversational Human-to-Machine interface that reflects how humans interact with each other and their normal behavior in conversation. Simply put, cooperative conversational model 300 enables humans and machines to participate in a conversation with an accepted purpose or direction, with each participant contributing to the conversation for the benefit of the conversation. That is, cooperative conversational model 300 incorporates technology and process-flow that takes advantage of human presumptions about utterances that humans rely upon, both as speakers and listeners, thereby creating a Human-to-Machine interface that is analogous to everyday human-to-human conversation. In one implementation, a cooperative conversation may take incoming data (shared knowledge) 305 to inform a decision (intelligent hypothesis building) 310, and then may refine the decision and generate a response (adaptive response building) 315.

[041] According to one aspect of the invention, shared knowledge 305 includes both short-term and long-term knowledge about incoming data. Short-term knowledge may accumulate during a single conversation, while long-term knowledge may accumulate over time to build user profiles, environmental profiles, historical profiles, cognitive profiles, etc.

[042] Input received during a single conversation may be retained in a Session Input Accumulator. The Session Input Accumulator may include cross-modality

awareness, where in addition to accumulating input relating to user utterances, requests, locations, etc., the Session Input Accumulator may accumulate a current user interface state relating to other modal inputs to further build shared knowledge models and more accurate adaptive responses (e.g., when a user utters a request relating to a portion of a touch-screen device, as described above). For example, the Session Input Accumulator may accumulate inputs including recognition text for each utterance, a recorded speech file for each utterance, a list-item selection history, a graphical user interface manipulation history, or other input data. Thus, the Session Input Accumulator may populate Intelligent Hypothesis Builder 310 with current and relevant information, build long-term shared knowledge by identifying information with long-term significance, provide Adaptive Response Builder 315 with relevant state and word usage information, retain recent contexts for use with Intelligent Hypothesis Builder 310, and/or retain utterances for reprocessing during multi-pass evaluations. Moreover, because cooperative conversations 300 model human conversations, short-term session data may be expired after a psychologically appropriate amount of time, thereby humanizing system behavior. For example, a human is unlikely to recall a context of a conversation from two years ago, but because the context would be identifiable by a machine, session context is expired after a predetermined amount of time to reduce a likelihood of contextual confusion based on stale data. However, relevant information from an expired session context may nonetheless be added to user, historical, environmental, cognitive, or other long-term knowledge models.

[043] Long-term shared knowledge may generally be user-centric, rather than session-based. That is, inputs may be accumulated over time to build user, environmental, cognitive, historical, or other long-term knowledge models. Long-term and short-term shared knowledge (collectively, shared knowledge 305) may be used simultaneously anytime a user engages in a cooperative conversation 300. Long-term shared knowledge may include explicit and/or implicit user preferences, a history of most recently used agents, contexts, requests, tasks, etc., user-specific jargon related to vocabularies and/or capabilities of an agent and/or context, most often used word choices, or other information. The long-term shared knowledge may be used to populate Intelligent Hypothesis Builder 310 with current and relevant information, provide Adaptive Response Builder 315 with appropriate word choices when the appropriate word choices are unavailable via the Session Input Accumulator, refine long-term shared

knowledge models, identify a frequency of specific tasks, identify tasks a user frequently has difficulty with, or provide other information and/or analysis to generate more accurate conversational responses.

[044] As described above, shared knowledge 305 may be used to populate Intelligent Hypothesis Builder 310, such that a user and a voice user interface may share assumptions and expectations such as topic knowledge, conversation history, word usage, jargon, tone (e.g., formal, humorous, terse, etc.), or other assumptions and/or expectations that facilitate interaction at a Human-to-Machine interface.

[045] According to an aspect of the invention, one component of a successful cooperative conversation may be identifying a type of conversation from an utterance. By categorizing and developing conceptual models for various types of exchanges, user expectations and domain capabilities may be consistently aligned. Intelligent Hypothesis Builder 310 may generate a hypothesis as to a conversation type by considering conversational goals, participant roles, and/or an allocation of information among the participants. Conversational goals may broadly include: (1) getting a discrete piece of information or performing a discrete task, (2) gathering related pieces of information to make a decision, and/or (3) disseminating or gathering large amounts of information to build expertise. Participant roles may broadly include: (1) a leader that controls a conversation, (2) a supporter that follows the leader and provides input as requested, and/or (3) a consumer that uses information. Information may be held by one or more of the participants at the outset of a conversation, where a participant may hold most (or all) of the information, little (or none) of the information, or the information may be allocated roughly equally amongst the participants. Based on the conversational goals, participant roles, and allocation of information, Intelligent Hypothesis Builder 310 may consider various factors to classify a conversation (or utterance) into general types of conversations that can interact with one another to form many more variations and permutations of conversation types (e.g., a conversation type may change dynamically as information is reallocated from one participant to another, or as conversational goals change based on the reallocation of information).

[046] For example, in one implementation, a query conversation may include a conversational goal of getting a discrete piece of information or performing a particular task, where a leader of the query conversation may have a specific goal in

mind and may lead the conversation toward achieving the goal. The other participant may hold the information and may support the leader by providing the information. In a didactic conversation, a leader of the conversation may control information desired by a supporter of the conversation. The supporter's role may be limited to regulating an overall progression of the conversation and interjecting queries for clarification. In an exploratory conversation, both participants share leader and supporter roles, and the conversation may have no specific goal, or the goal may be improvised as the conversation progresses. Based on this model, Intelligent Hypothesis Builder 310 may broadly categorize a conversation (or utterance) according to the following diagram:

QUERY		
	Participant A	*Participant B*
	User	Voice User Interface
GOAL	Get information/action	Provide information/action
ROLE	Leader/Consumer	Supporter/Dispenser
INFORMATION ALLOCATION	Less	More

DIDACTIC		
	Participant A	*Participant B*
	User	Voice User Interface
GOAL	Get information	Provide information
ROLE	Follower/Consumer	Leader/Dispenser
INFORMATION ALLOCATION	Less	More

EXPLORATORY		
	Participant A	*Participant B*
	User	Voice User Interface
GOAL	Gather/share information	Gather/share information
ROLE	Follower/Consumer and Leader/Dispenser	Follower/Consumer and Leader/Dispenser
INFORMATION ALLOCATION	Equal or alternating	Equal or alternating

[047] Intelligent Hypothesis Builder 310 may use an identified conversation type to assist in generating a set of hypotheses as to a user's intent in an utterance. In addition, Intelligent Hypothesis Builder 310 may use short-term shared knowledge from the Session Input Accumulator to proactively build and evaluate interaction with a user as a conversation progresses, as well as long-term shared knowledge to proactively build and evaluate interaction with the user over time. Intelligent Hypothesis Builder 310 may thus adaptively arrive at a set of n-best hypotheses about user intent, and the n-best hypotheses may be provided to an Adaptive Response Builder 315. In addition, Intelligent Hypothesis Builder 310 may model human-to-human interaction by calculating a degree of certainty for each of the hypotheses. That is, just as humans rely on knowledge shared by participants to examine how much and what kind of information was available, Intelligent Hypothesis Builder 310 may leverage the identified conversation type and short-term

and long-term shared knowledge to generate a degree of certainty for each hypothesis.

[048] According to another aspect of the invention, Intelligent Hypothesis Builder 310 may generate one or more explicit hypotheses of a user's intent when an utterance contains all information (including qualifiers) needed to complete a request or task. Each hypothesis may have a corresponding degree of certainty, which may be used to determine a level of unprompted support to provide in a response. For example, a response may include a confirmation to ensure the utterance was not misunderstood. or the response may adaptively prompt a user to provide missing information.

[049] According to another aspect of the invention, Intelligent Hypothesis Builder 310 may use short-term knowledge to generate one or more implicit hypotheses of a user's intent when an utterance may be missing required qualifiers or other information needed to complete a request or task. Each hypothesis may have a corresponding degree of certainty. For instance, when a conversation begins, short-term knowledge stored in the Session Input Accumulator may be empty, and as the conversation progresses, the Session Input Accumulator may build a history of the conversation. Intelligent Hypothesis Builder 310 may use data in the Session Input Accumulator to supplement or infer additional information about a current utterance. For example, Intelligent Hypothesis Builder 310 may evaluate a degree of certainty based on a number of previous requests relevant to the current utterance. In another example, when the current utterance contains insufficient information to complete a request or task, data in the Session Input Accumulator may be used to infer missing information so that a hypothesis can be generated. In still another example, Intelligent Hypothesis Builder 310 may identify syntax and/or grammar to be used by Adaptive Response Builder 315 to formulate personalized and conversational response. In yet another example, when the current utterance contains a threshold amount of information needed to complete a request or task, data in the Session Input Accumulator may be relied upon to tune a degree of certainty.

[050] According to another aspect of the invention, Intelligent Hypothesis Builder 310 may use long-term shared knowledge to generate one or more implicit hypotheses of a user's intent when an utterance is missing qualifiers or other information needed to complete a request or task. Each hypothesis may have a corresponding degree of certainty. Using long-term knowledge may be substantially similar to using short-term shared knowledge, except that information may be unconstrained by a current session,

and an input mechanism may include information from additional sources other than conversational sessions. For example, Intelligent Hypothesis Builder 310 may use information from long-term shared knowledge at any time, even when a new conversation is initiated, whereas short-term shared knowledge may be limited to an existing conversation (where no short-term shared knowledge would be available when a new conversation is initiated). Long-term shared knowledge may come from several sources, including user preferences or a plug-in data source (e.g., a subscription interface to a remote database), expertise of a user (e.g., based on a frequency of errors, types of tasks requested, etc., the user may be identified as a novice, intermediate, experienced, or other type of user), agent-specific information and/or language that may also apply to other agents (e.g., by decoupling information from an agent to incorporate the information into other agents), frequently used topics passed in from the Session Input Accumulator, frequently used verbs, nouns, or other parts of speech, and/or other syntax information passed in from the Session Input Accumulator, or other sources of long-term shared knowledge may be used.

[051] According to another aspect of the invention, knowledge-enabled utterances, as generated by Intelligent Hypothesis Builder 310, may include one or more explicit (supplied by a user), and one or more implicit (supplied by Intelligent Hypothesis Builder 310) contextual signifiers, qualifiers, criteria, and other information that can be used to identify and evaluate relevant tasks. At that point, Intelligent Hypothesis Builder 310 may provide an input to Adaptive Response Builder 315. The input received by Adaptive Response Builder 315 may include at least a ranked list of hypotheses, including explicit and/or implicit hypotheses, each of which may have a corresponding degree of certainty. A hypothesis may be assigned one of four degrees of certainty: (1) "sure," where contextual signifiers and qualifiers relate to one task, context and qualifiers relate to one task, and an ASR confidence level exceeds a predetermined threshold; (2) "pretty sure," where contextual signifiers and qualifiers relate to more than one task (select top-ranked task) and criteria relates to one request, and/or the ASR confidence level is below the predetermined threshold; (3) "not sure," where additional contextual signifiers or qualifiers are needed to indicate or rank a task; and (4) "no hypothesis," where little or no information can be deciphered. Each degree of certainty may further be classified as explicit or implicit, which may be used to adjust a response. The input received by Adaptive Response Builder 310 may also include a context, user syntax and/or

grammar, context domain agent specific information and/or preferences (e.g., a travel context domain agent may know a user frequently requests information about France, which may be shared with a movie context domain agent so that responses may occasionally include French movies).

[052] According to another aspect of the invention, Adaptive Response Builder 315 may build syntactically, grammatically, and contextually sensitive "intelligent responses" that can be used with one or more agents to generate a conversational experience for a user, while also guiding the user to reply in a manner favorable for recognition. In one implementation, the intelligent responses may include a verbal or audible reply played through an output device (e.g., a speaker), and/or an action performed by a device, computer, or machine (e.g., downloading a web page, showing a list, executing an application, etc.). In one implementation, an appropriate response may not require conversational adaptation, and default replies and/or randomly selected response sets for a given task may be used.

[053] According to another aspect of the invention, Adaptive Response Builder 310 may draw on information maintained by Intelligence Hypothesis Builder 310 to generate responses that may be sensitive to context, task recognition of a current utterance, what a user already knows about a topic, what an application already knows about the topic, shared knowledge regarding user preferences and/or related topics, appropriate contextual word usage (e.g., jargon), words uttered by the user in recent utterances, conversational development and/or course correction, conversational tone, type of conversation, natural variation in wording of responses, or other information. As a result, Adaptive Response Builder 315 may generate intelligent responses that create conversational feel, adapt to information that accumulates over a duration of a conversation, maintain cross-modal awareness, and keep the conversation on course.

[054] According to another aspect of the invention, Adaptive Response Builder 315 may create a conversational feel by adapting to a user's manner of speaking, framing responses appropriately, and having natural variation and/or personality (e.g., by varying tone, pace, timing, inflection, word use, jargon, and other variables in a verbal or audible response). Adapting to a user's manner of speaking may include using contextual signifiers and grammatical rules to generate one or more sentences for use as response sets that may cooperate with the user. By taking advantage of short-term (from the Session Input

Accumulator) and long-term (from one or more profiles) shared knowledge about how a user utters a request, the responses may be modeled using techniques used to recognize requests. Adaptive Response Builder 315 may rate possible responses statistically and/or randomize responses, which creates an opportunity to build an exchange with natural variation and conversational feel. This may be a significant advantage over existing voice user interfaces with incongruous input and output, where the input is "conversational" and the output is "computerese." The following examples may demonstrate how a response may adapt to a user's input word choices and manner of speaking:

User	**Do you know** [*mumbled words*] Seattle [*more mumbled words*]?
Voice User Interface	Did you want Seattle sports scores, weather, traffic, or news?

User	**Find me** [*mumbled words*] Seattle [*more mumbled words*]?
Voice User Interface	I *found* Seattle, did you want sports scores, weather, traffic, or news?

User	**Get me** [*mumbled words*] Seattle [*more mumbled words*]?
Voice User Interface	I've *got* Seattle, did you want me to *get* sports scores, weather, traffic, or news?

[055] According to another aspect of the invention, Adaptive Response Builder 315 may frame responses to influence a user to reply with an utterance that may be easily recognized. For example, a user may utter, "Get me the news" and a voice user interface response may be "Which of these categories? Top news stories, international news, political news, or sports news?" The response may be likely to illicit utterances from the user, such as "Top news stories" or "International news," which are more likely to result in a completed request. Thus, the responses may conform to a cooperative nature of human dialog, and a natural human tendency to "parrot" what was just heard as part of a next utterance. Moreover, knowledge of current context may enhance responses to generate more meaningful conversational responses, such as in the following exchange:

User	What's the weather like in Dallas?

Voice User Interface	In Dallas, it's sunny and 90 degrees.
User	What theaters are showing the movie "The Fantastic Four" there?
Voice User Interface	10 theaters in Dallas are showing "The Fantastic Four." Do you want show times for a particular theater?

[056] Framing the responses may also deal with misrecognitions according to human models. For example, humans frequently remember a number of recent utterances, especially when one or more previous utterances were misrecognized or unrecognized. Another participant in the conversation may limit correction to a part of the utterance that was misrecognized or unrecognized, or over subsequent utterances and/or other interactions, clues may be provided to indicate the initial interpretation was incorrect. Thus, by storing and analyzing multiple utterances, utterances from earlier in a conversation may be corrected as the conversation progresses.

[057] According to another aspect of the invention, Adaptive Response Builder 315 may generate multi-modal, or cross-modal, responses to a user. In one implementation, responses may be aware of and control one or more devices and/or interfaces, and users may respond by using whichever input method, or combination of input methods, is most convenient. For example, a response asking the user to direct an utterance with a "Yes" or "No" in a multi-modal environment may also display alternatives visually.

[058] According to another aspect of the invention, Adaptive Response Builder 315 may correct a course of a conversation without interrupting conversational flow. Adaptive Response Builder 315 may generate intelligent responses based on the ranked list of hypotheses and corresponding degrees of certainty, which may be used to correct a course of a conversation without interrupting conversational flow. That is, even though the intelligent responses may be reasonably "sure," the intelligent responses may nonetheless sometimes be incorrect. While existing voice user interfaces tend to fail on average conversational missteps, normal human interactions may expect missteps and deal with them appropriately. Thus, responses after a misrecognition may be modeled after clarifications, rather than errors, and words may chosen in subsequent responses that move conversation forward and establish an appropriate domain to be explored with the user. For example, course correction may result in the following exchange:

User	Can you **get** [*mumbled words here*] Mariners [*more mumbled words*]?
Voice User Interface	*I've got* the score for the Mariners game? 4-2 Mariners.
User	No, the start time for tomorrow's game.
Voice User Interface	Oh, 7:05.

[059] The above disclosure has been described in terms of specific exemplary aspects, implementations, and embodiments of the invention. However, those skilled in the art will recognize various changes and modifications that may be made without departing from the scope and spirit of the invention. Therefore, the specification and drawings are to be regarded as exemplary only, and the scope of the invention is to be determined solely by the appended claims.

CLAIMS

What is claimed is:

1. A method for providing a voice user interface that generates cooperative conversational responses to user utterances, the responses being based on shared knowledge about a current conversation and/or one or more past conversations, the method comprising:

receiving an input, the input including at least one utterance;

updating the shared knowledge to include knowledge about the at least one utterance;

determining an intended meaning for the at least one utterance based on the updated shared knowledge; and

generating a response based on the determined intended meaning.

2. The method of claim 1, wherein updating the shared knowledge includes populating a short-term context stack with information about the at least one utterance.

3. The method of claim 2, further comprising expiring the information about the at least one utterance in the short-term context stack after a predetermined amount of time.

4. The method of claim 1, wherein updating the shared knowledge includes updating one or more long-term profiles based on information about the at least one utterance.

5. The method of claim 1, wherein determining the intended meaning includes identifying a conversation type for the at least one utterance, the conversation type based on one or more of a goal for the utterance, one or more participant roles for the utterance, and/or an allocation of information for the utterance.

6. The method of claim 1, wherein determining the intended meaning includes generating at least one hypothesis of an intended meaning for the at least one utterance, the at least one hypothesis including a degree of certainty.

7. The method of claim 6, wherein generating the at least one hypothesis includes identifying contextual signifiers and/or qualifiers in the at least one

utterance.

8. The method of claim 7, further comprising inferring missing contextual signifiers and/or qualifiers for the at least one utterance based on the shared knowledge.

9. The method of claim 8, wherein the degree of certainty for the at least one hypothesis is based on one or more of the identified contextual signifiers and/or qualifiers, the inferred contextual signifiers and/or qualifiers, and/or contextual signifiers and/or qualifiers necessary for generating a response.

10. The method of claim 1, wherein generating the response includes adapting contextual signifiers and/or grammar in the response to reflect contextual signifiers and/or grammar in the at least one utterance.

11. The method of claim 1, wherein generating the response includes framing the response to influence a subsequent utterance.

12. The method of claim 1, further comprising:

learning that the determined intended meaning is incorrect;

further updating the shared knowledge to include knowledge that the determined intended meaning is incorrect; and

determining a second intended meaning for the at least one utterance based on the further updated shared knowledge, wherein the incorrect determined intended meaning is excluded as a possible second intended meaning.

13. The method of claim 1, wherein the input is a multi-modal input.

14. The method of claim 1, wherein the response is a multi-modal response.

15. A computer readable medium containing computer-executable instructions for providing a voice user interface that generates cooperative conversational responses to user utterances, the responses being based on shared knowledge about a current conversation and/or one or more past conversations, the computer-executable

instructions operable when executed to:

receive an input, the input including at least one utterance;

update the shared knowledge to include knowledge about the at least one utterance;

determine an intended meaning for the at least one utterance based on the updated shared knowledge; and

generate a response based on the determined intended meaning.

16. The computer readable medium of claim 15, wherein the computer-executable instructions are operable to update the shared knowledge by populating a short-term context stack with information about the at least one utterance.

17. The computer readable medium of claim 16, wherein the computer-executable instructions are operable to expire the information about the at least one utterance in the short-term context stack after a predetermined amount of time.

18. The computer readable medium of claim 15, wherein the computer-executable instructions are operable to update the shared knowledge by updating one or more long-term profiles based on information about the at least one utterance.

19. The computer readable medium of claim 15, wherein the computer-executable instructions are operable to determine the intended meaning by identifying a conversation type for the at least one utterance, the conversation type based on one or more of a goal for the utterance, one or more participant roles for the utterance, and/or an allocation of information for the utterance.

20. The computer readable medium of claim 15, wherein the computer-executable instructions are operable to determine the intended meaning by generating at least one hypothesis of an intended meaning for the at least one utterance, the at least one hypothesis including a degree of certainty.

21. The computer-readable medium of claim 20, wherein the computer-executable instructions are operable to generate the at least one hypothesis by identifying contextual signifiers and/or qualifiers in the at least one utterance.

22. The computer readable medium of claim 21, wherein the computer-executable instructions are operable to infer missing contextual signifiers and/or qualifiers for the at least one utterance based on the shared knowledge.

23. The computer readable medium of claim 22, wherein the degree of certainty for the at least one hypothesis is based on one or more of the identified contextual signifiers and/or qualifiers, the inferred contextual signifiers and/or qualifiers, and/or contextual signifiers and/or qualifiers necessary for generating a response.

24. The computer readable medium of claim 15, wherein the computer-executable instructions are operable to generate the response by adapting contextual signifiers and/or grammar in the response to reflect contextual signifiers and/or grammar in the at least one utterance.

25. The computer readable medium of claim 15, wherein the computer-executable instructions are operable to generate the response by framing the response to influence a subsequent utterance.

26. The computer readable medium of claim 15, wherein the computer-executable instructions are further operable when executed to:

learn that the determined intended meaning is incorrect;

further update the shared knowledge to include knowledge that the determined intended meaning is incorrect; and

determine a second intended meaning for the at least one utterance based on the further updated shared knowledge, wherein the incorrect determined intended meaning is excluded as a possible second intended meaning.

27. The computer readable medium of claim 15, wherein the input is a multi-modal input.

28. The computer readable medium of claim 15, wherein the response is a multi-modal response.

29. A system for providing a voice user interface that generates cooperative conversational responses to user utterances, the system including at least one memory for storing shared knowledge about a current conversation and/or one or more past conversations, the system comprising:

at least one input device that receives an input, the input including at least one utterance; and

one or more processors collectively operable to:

update the shared knowledge to include knowledge about the at least one utterance;

determine an intended meaning for the at least one utterance based on the updated shared knowledge; and

generate a response based on the determined intended meaning.

30. The system of claim 29, wherein the one or more processors are operable to update the shared knowledge by populating a short-term context stack with information about the at least one utterance.

31. The system of claim 30, wherein the one or more processors are operable to expire the information about the at least one utterance in the short-term context stack after a predetermined amount of time.

32. The system of claim 29, wherein the one or more processors are operable to update the shared knowledge by updating one or more long-term profiles based on information about the at least one utterance.

33. The system of claim 29, wherein the one or more processors are operable to determine the intended meaning by identifying a conversation type for the at least one utterance, the conversation type based on one or more of a goal for the utterance, one or more participant roles for the utterance, and/or an allocation of information for the utterance.

34. The system of claim 29, wherein the one or more processors are operable to

determine the intended meaning by generating at least one hypothesis of an intended meaning for the at least one utterance, the at least one hypothesis including a degree of certainty.

35. The system of claim 34, wherein the one or more processors are operable to generate the at least one hypothesis by identifying contextual signifiers and/or qualifiers in the at least one utterance.

36. The system of claim 35, wherein the one or more processors are operable to infer missing contextual signifiers and/or qualifiers for the at least one utterance based on the shared knowledge.

37. The system of claim 36, wherein the degree of certainty for the at least one hypothesis is based on one or more of the identified contextual signifiers and/or qualifiers, the inferred contextual signifiers and/or qualifiers, and/or contextual signifiers and/or qualifiers necessary for generating a response.

38. The system of claim 29, wherein the one or more processors are operable to generate the response by adapting contextual signifiers and/or grammar in the response to reflect contextual signifiers and/or grammar in the at least one utterance.

39. The system of claim 29, wherein the one or more processors are operable to generate the response by framing the response to influence a subsequent utterance.

40. The system of claim 29, wherein the one or more processors are further operable to:

learn that the determined intended meaning is incorrect;

further update the shared knowledge to include knowledge that the determined intended meaning is incorrect; and

determine a second intended meaning for the at least one utterance based on the further updated shared knowledge, wherein the incorrect determined intended meaning is excluded as a possible second intended meaning.

41. The system of claim 29, wherein the input is a multi-modal input.

42. The system of claim 29, wherein the response is a multi-modal response.

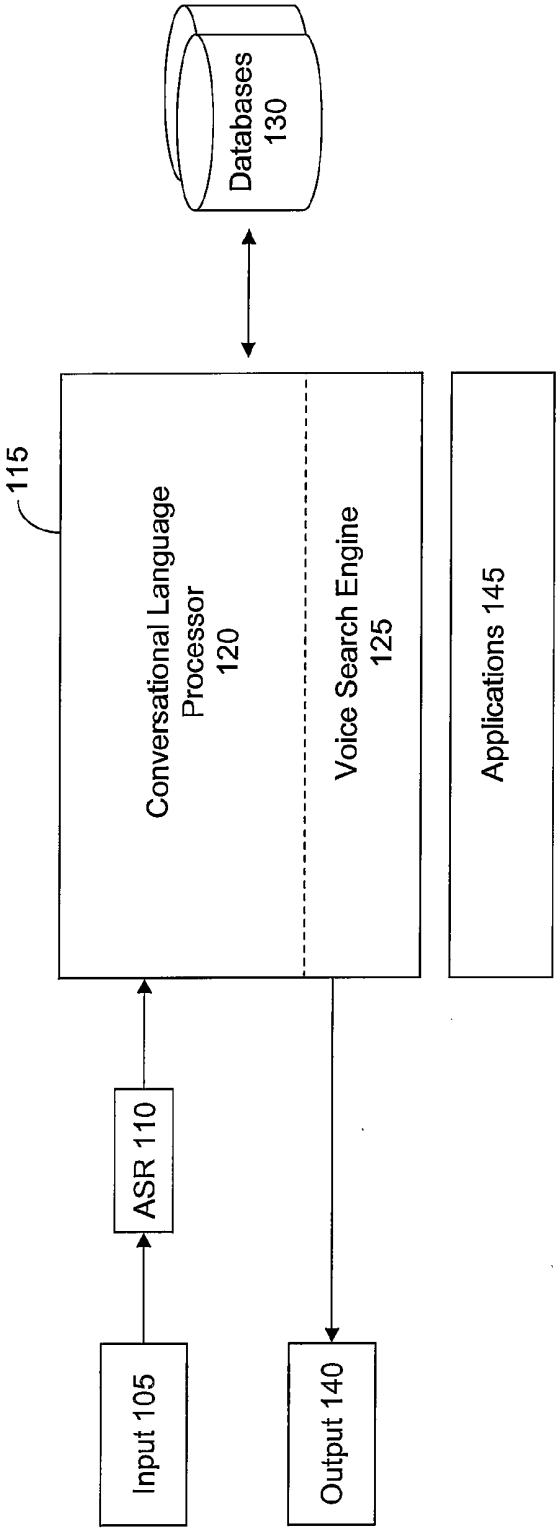

Databases
130
115
Conversational Language
Processor
120
Voice Search Engine
125
Applications 145
ASR 110
Input 105
Output 140

FIG. 1

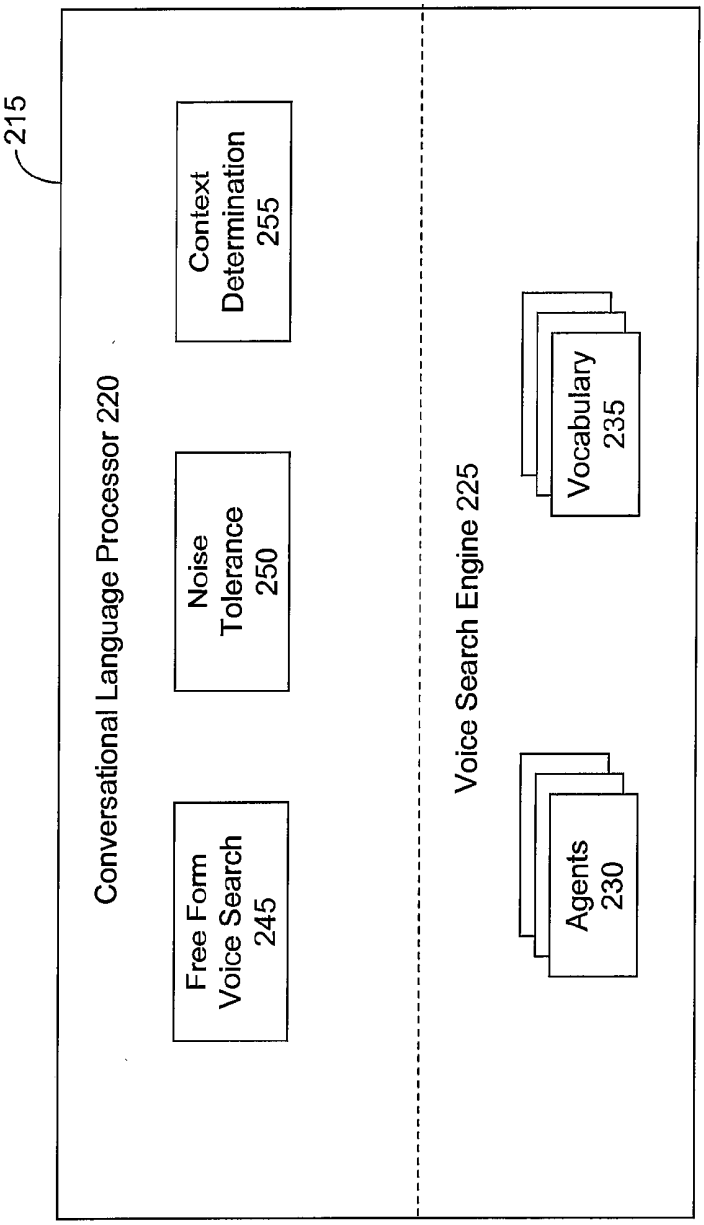

215
Conversational Language Processor 220
Free Form Voice Search 245
Noise Tolerance 250
Context Determination 255
Voice Search Engine 225
Agents 230
Vocabulary 235

FIG. 2

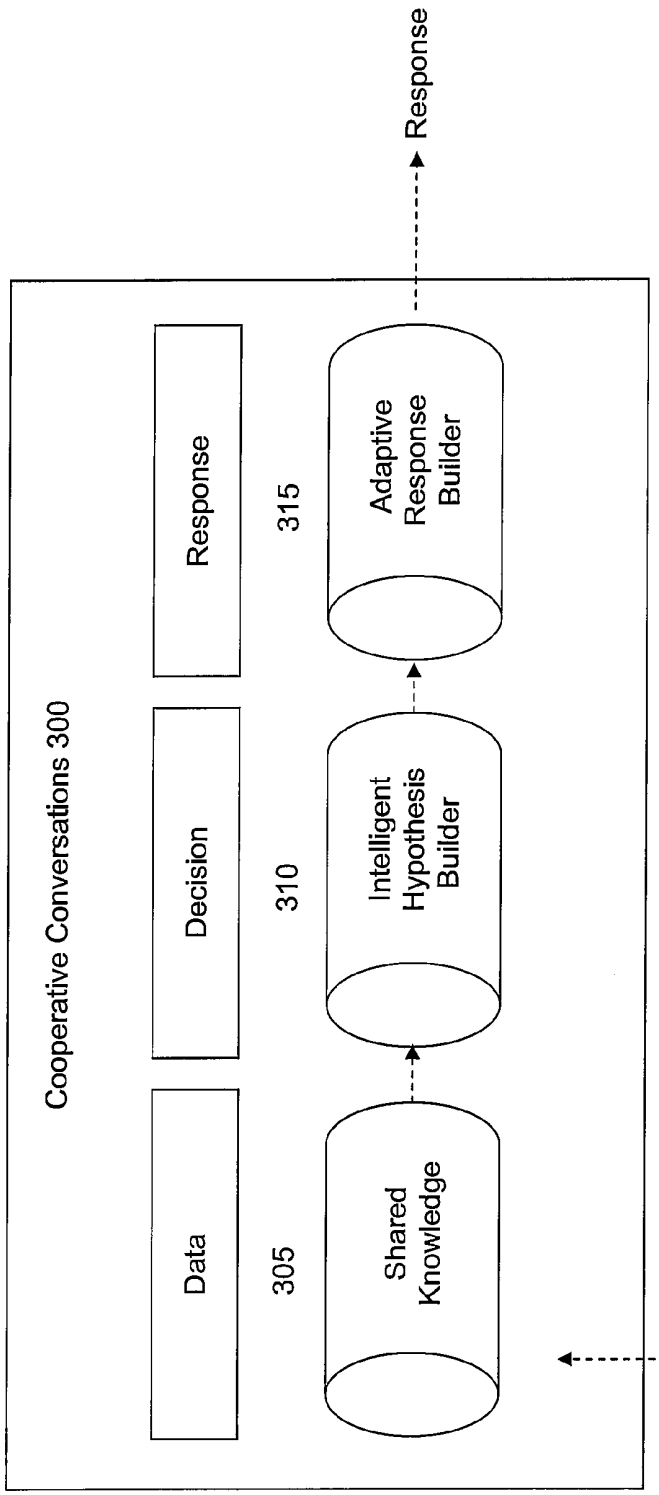

Response
Cooperative Conversations 300
Data
Decision
Response
305
310
315
Shared Knowledge
Intelligent Hypothesis Builder
Adaptive Response Builder
Input

FIG. 3

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 June 2009 (18.06.2009)

PCT

(10) International Publication Number
WO 2009/075912 A1

(51) International Patent Classification:
G06F 17/27 (2006.01)

(21) International Application Number:
PCT/US2008/069508

(22) International Filing Date: 9 July 2008 (09.07.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/954,064 11 December 2007 (11.12.2007) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [US/US]; 11980 NE 24th Street, Bellevue, Washington 98005 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **KENNEWICK, Michael** [US/US]; 12196 NE 24th Street, Bellevue, Washington 98005 (US). **CHEUNG, Catherine** [US/US]; 1639 207th Avenue NE, Sammamish, Washington 98074 (US). **BALDWIN, Larry** [US/US]; 23125 SE 243rd Place, Maple Valley, Washington 98038 (US). **SALOMON, Ari** [US/US]; 4133 96th Avenue SE, Mercer Island, Washington 98040 (US). **TJALVE, Michael** [DK/US]; 14904 SE 47th Court, Bellevue, Washington 98006 (US). **GUTTIGOLI, Sheetal** [IN/US]; 11314 Ohde Circle, Kirkland, Washington 98033 (US). **ARMSTRONG, Lynn** [US/US]; 19807 183rd Place NE, Woodinville, Washington 98077 (US). **DI CRISTO, Philippe** [FR/US]; 177 107th Avenue NE, #1007, Bellevue, Washington 98004 (US). **ZIMMERMAN, Bernie** [US/US]; 135 S. 107th Street, Seattle, Washington 98168 (US). **MENAKER, Sam** [US/US]; 480 150th Place NE, Bellevue, Washington 98007 (US).

(74) Agents: **GATTO, James G.** et al.; Pillsbury Winthrop Shaw Pittman LLP, P.O. Box 10500, McLean, Virginia 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),

[Continued on next page]

(54) Title: PROVIDING A NATURAL LANGUAGE VOICE USER INTERFACE IN AN INTEGRATED VOICE NAVIGATION SERVICES ENVIRONMENT

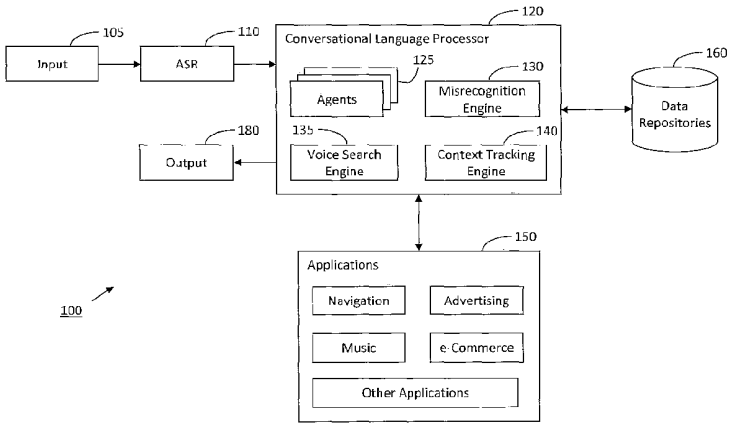


Figure 1

(57) Abstract: A conversational, natural language voice user interface may provide an integrated voice navigation services environment. The voice user interface may enable a user to make natural language requests relating to various navigation services, and further, may interact with the user in a cooperative, conversational dialogue to resolve the requests. Through dynamic awareness of context, available sources of information, domain knowledge, user behavior and preferences, and external systems and devices, among other things, the voice user interface may provide an integrated environment in which the user can speak conversationally, using natural language, to issue queries, commands, or other requests relating to the navigation services provided in the environment.

WO 2009/075912 A1

European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

PROVIDING A NATURAL LANGUAGE VOICE USER INTERFACE IN AN INTEGRATED VOICE NAVIGATION SERVICES ENVIRONMENT

FIELD OF THE INVENTION

[001] The present invention relates to a natural language voice user interface that facilitates cooperative, conversational interactions in an integrated voice navigation services environment, and in particular, to a natural language voice user interface in which users can request navigation services using conversational, natural language queries or commands.

BACKGROUND OF THE INVENTION

[002] As technology advances, consumer electronics tend to play increasingly significant roles in everyday life. As a result, users tend to expect greater functionality, mobility, and convenience from their electronic devices, as exemplified by modern mobile phones, navigation devices, personal digital assistants, portable media players, and other devices often providing a wealth of functionality beyond core applications. However, the greater functionality often tends to be accompanied by significant learning curves and other barriers that prevent users from fully exploiting device capabilities (e.g., features may often be buried within difficult to navigate menus or interfaces). Moreover, although increasing demand for mobility magnifies the need for simple on-the-go device interaction mechanisms, existing systems often have complex human to machine interfaces. For example, existing human to machine interfaces tend to primarily utilize various combinations of keyboards, keypads, point and click techniques, touch screen displays, or other interface mechanisms. However, these interfaces may often be unsuitable for mobile or vehicular devices (e.g., navigation devices), as they tend to be cumbersome in environments where speed of interaction and dangers of distraction pose significant issues. As such, existing systems often fall short in providing simple and intuitive interaction mechanisms, potentially inhibiting mass-market adoption for certain technologies. As such, there is an ever-growing demand for ways to exploit technology in intuitive ways.

[003] In response to these and other problems, various existing systems have turned to voice recognition software to simplify human to machine interactions. For example, voice recognition software can enable a user to exploit applications and features of a device that may otherwise be unfamiliar, unknown, or difficult to use. However, existing voice user interfaces, when they

actually work, still require significant learning on the part of the user. For example, existing voice user interfaces (e.g., command and control systems) often require users to memorize syntaxes, words, phrases, or other keywords or qualifiers in order to issue queries or commands. Similarly, when users may be uncertain of exactly what to request, or what a device may be capable of, existing systems cannot engage with the user in a productive, cooperative, natural language dialogue to resolve requests and advance conversations. Instead, many existing speech interfaces force users to use predetermined commands or keywords to communicate requests in ways that systems can understand. By contrast, cognitive research on human interaction demonstrates that a person asking a question or giving a command typically relies heavily on context and shared knowledge of an answering person. Similarly, the answering person also tends to rely on the context and shared knowledge to inform what may be an appropriate response. However, existing voice user interfaces do not adequately utilize context, shared knowledge, or other similar information to provide an environment in which users and devices can cooperate to satisfy mutual goals through conversational, natural language interaction.

[004] Furthermore, demand for global positional systems and other navigation-enabled devices has grown significantly in recent years. Navigation devices often tend to be used while a user may be driving, on-the-go, or in other environments where having a hands-free interface provides critical advantages. For example, a user may want to avoid being distracted by looking away from the road, yet the user may also want to interact with a navigation device, for example, to calculate a route to a destination, recalculate the route in response to traffic, find a local restaurant, gas station, or other point of interest, or perform another navigation related task. In these and other instances, efficiently processing a natural language voice-based input could enable the user to interact with the navigation device in a safer, simpler, and more effective way. However, existing systems often fall short in providing an integrated, conversational, natural language voice user interface that can provide such advantages in navigation and other mobile environments.

[005] Existing systems suffer from these and other problems.

SUMMARY OF THE INVENTION

[006] According to various aspects of the invention, various problems associated with existing systems may be addressed by a conversational, natural language voice user interface that provides an integrated voice navigation services environment.

[007] According to various aspects of the invention, the natural language voice user interface may resolve voice requests relating to navigation (e.g., calculating routes, identifying locations, displaying maps, etc.). The navigation application can provide a user with interactive, data-driven directions to a destination or waypoint, wherein the user can specify the destination or waypoint using free-form natural language (e.g., the user can identify full or partial destinations, including a specific address, a general vicinity, a city, a name or type of a place, a name or type of a business, a name of a person, etc.). As free form voice destination inputs may be provided in many different forms, post-processing may be performed on full or partial voice destination inputs to identify a suitable destination address for calculating a route (e.g., a closest address that makes "sense"). For example, an utterance containing a full or partial destination may be analyzed to identify one or more probable destinations (e.g., an N-best list of destinations). The N-best list may be post-processed to assign weights or rankings to the probable destinations (e.g., based on a degree of certainty that a given probable destination corresponds to an intended destination). Thus, a route may be calculated from a user's current location to a most heavily weighted one of the probable destinations in the N-best list. Further, when a voice destination entry includes a partial destination, a final destination may be successively refined over one or more subsequent voice destination entries. The navigation application may also provide dynamic, data-driven directions or routing to a destination. For instance, the navigation application may access data associated with various user-specific and environmental data sources to provide personalized data-driven directions along a route, which can be recalculated or modified based on information taken from the data sources. As such, data may be obtained dynamically to identify alternate routes, recalculate routes, or otherwise provide routing services. Further, possible answers or responses to a given utterance may be filtered according to a current route.

[008] According to various aspects of the invention, the natural language voice user interface may dynamically generate and/or load recognition grammars for interpreting what was said in an utterance (e.g., content of the utterance). Information contained in the dynamic recognition grammars may be used by a navigation agent, an Automatic Speech Recognizer, a context stack,

or various other components in the voice user interface that use grammar information. By efficiently generating, updating, loading, extending, or otherwise building dynamic grammars based on various factors, processing bottlenecks can be avoided, conflicts can be reduced, and other aspects of interpreting an utterance using a recognition grammar can be optimized. For example, a size of a generated grammar may be constrained by an amount of resources available in a system (e.g., in embedded devices or other devices having low amounts of dynamic memory, the constrained grammar size may limit a quantity of resources to be occupied). In another example, the size of the dynamic grammar can be reduced by eliminating redundant keywords, criteria, or other information available in the context stack, the shared knowledge, or other local sources. Thus, favorability of correct interpretations may be improved by reducing perplexity in the grammar (e.g., when two or more elements may likely be confused, one or more of the elements may be eliminated to reduce confusion).

[009] According to various aspects of the invention, the natural language voice user interface may generate dynamic recognition grammars using techniques of geographical chunking. A user's location can be determined at a given moment to determine one or more geographic proximities, which can be used to form an appropriate topological domain for the grammar. For example, the topological domains may reflect physical proximities (e.g., a distance from a current location), civil organization proximities (e.g., regions, states, cities, neighborhoods, subdivisions, localities, etc.), temporal proximities (e.g., amounts of travel time from the current location), directional proximities (e.g., based on directional travel vectors), or various combinations thereof. As a result, by mapping the user's geographic proximities to one or more topological domains, dynamic grammars may be pruned, extended, swapped in or out of memory, or otherwise generated and/or loaded to provide optimal recognition based on location, time, travel, or other factors (e.g., information may be swapped in and out of a grammar as a user moves from one area to another, ensuring that system resources utilize information presently relevant to a given location).

[010] According to various aspects of the invention, the natural language voice user interface may include dynamic grammars formed from one or more topological domains, which may be subdivided into a plurality of tiles, which may be further subdivided into a plurality of subtiles. Thus, information used to build the dynamic grammar can be subdivided or weighed in various ways to determine what information should be included in the grammar. Moreover,

geographical chunks based on physical, civil organization, temporal, directional, or other proximities may be extended into other domains in which a topological taxonomy can be placed. As a result, in addition to having relevance in navigation or other location-dependent systems, the geographical chunking techniques can be applied in other contexts or domains in which geography or location may be relevant. Further, a server operably coupled to the voice user interface may analyze various forms of information to build or refine a source of grammar information. For example, when various devices communicate with the server, information may be communicated to the server may be used to update proximities, topological domains, tiles, subtiles, peer-to-peer affinities, or other grammar information.

[011] According to various aspects of the invention, the natural language voice user interface may calculate routes, provide dynamic data-driven directions to a destination, provide dynamic routing to a destination, perform post-processing of full or partial destination entries, or otherwise provide various voice navigation services. Further, destinations and/or routes may be identified using techniques of successive refinement for voice destination entries, wherein context, agent adaptation, and shared knowledge, among other things, can help a user to narrow down a final destination using voice commands, multi-modal commands, or various combinations thereof. However, it will be apparent that the successive refinement techniques can be applied to various tasks in which generalized approximations can be successively refined through voice or multi-modal commands to narrow down information sought by a user, including various other domains, contexts, applications, devices, or other components that employ the techniques described herein.

[012] According to various aspects of the invention, the natural language voice user interface may enable successive refinement of a final destination by progressively narrowing the final destination. For example, successively refining the destination may be modeled after patterns of human interaction in which a route or a destination may be narrowed down or otherwise refined over a course of interaction. For example, a user may generally approximate a destination, which may result in a route being calculated along a preferred route to the approximated destination. While en route to the approximated destination, the user and the voice user interface may cooperatively refine the final destination through one or more subsequent interactions. Thus, a user may provide a full or partial destination input using free form natural language, for example, including voice commands and/or multi-modal commands. One or more

interpretations of a possible destination corresponding to the voice destination input may be organized in an N-best list. The list of possible destinations may be post-processed to assign weights or ranks to one or more of the entries therein, thus determining a most likely intended destination from a full or partial voice destination input. Thus, the post-processing operation may rank or weigh possible destinations according to shared knowledge about the user, domain-specific knowledge, dialogue history, or other factors. As a result, the full or partial destination input may be analyzed to identify an address to which a route can be calculated, for example, by resolving a closest address that makes "sense" relative to the input destination. Subsequent inputs may provide additional information relating to the destination, and the weighted N-best list may be iteratively refined until the final destination can be identified through successive refinement. As a result, when a suitable final destination has been identified, the route to the final destination may be completed.

[013] According to various aspects of the invention, the natural language voice user interface may include one or more advertising models for generating and/or detecting events relating to location dependent advertisements for navigation systems (e.g., as generated by a local or remote advertising engine, or via a data channel, or in other ways). For example, navigation systems typically include various mechanisms for determining a current location (e.g., a global positioning system, a radio frequency identification system, a system that determines location based on a distance to an identifiable wireless tower or access point, etc.). The location detection system may thus detect information associated with a radio frequency identifier over a data channel used by a marketer to provide advertisements. The marketer may broadcast the advertisement via the data channel, such that the navigation system triggers an event when within a suitable proximity of the RFIDs. Thus, information associated with the event may be filtered according to the current routing information or other contextual parameters to determine what action should be taken in response thereto. In other instances, advertisements may be uploaded to a server by one or more advertising partners, wherein the uploaded advertisements may be associated with metadata or other descriptive information that identifies a target audience, location-dependent information, or other criteria. In another example, a plurality of advertisements may be stored locally at the voice user interface, and an inferencing engine may determine appropriate circumstances in which an event should be generated to deliver one or more of the advertisements to a user. As a result, it will be apparent that advertising events may

be generated in a number of ways, and may be generated and/or detected locally, remotely, by detecting RFIDs, or in other ways.

[014] According to various aspects of the invention, the natural language voice user interface may track user interactions with delivered advertisements. In this way, affinity based models may be generated, for example, to ensure that promotions or advertisements will be delivered to a likely target audience. Thus, an event relating to a given advertisement may be generated and/or detected when shared knowledge about a user's behavior, preferences, or other characteristics match one or more criteria associated with peer-to-peer affinities associated with the advertisement. In other examples, an advertising model may include mobile pay-per-use systems, peer-to-peer local guides or recommendations, or other forms of advertising. Additionally, various aspects of the advertising model, such as the local guides and recommendations, may be generated according to a mapping applied to various topological domains. For example, certain types of advertisements may be dependent on geographic or topological characteristics, and such advertisements may be associated with a topological taxonomy based on geographical chunks. As a result, various advertising events may be generated and/or detected according to physical proximities, temporal proximities, directional proximities, civil organization proximities, or various combinations thereof.

[015] According to various aspects of the invention, the natural language voice user interface may enable a user to provide requests (e.g., queries, commands, or other requests) to the navigation device using natural language. As such, the user and the navigation device may engage in a cooperative, conversational dialogue to resolve the request. For example, the voice user interface may use prior context, dialogue histories, domain knowledge, short and long-term shared knowledge relating to user behavior and preferences, noise tolerance, and cognitive models, among various other things, to provide an integrated environment in which users can speak conversationally, using natural language, to issue queries, commands, or other requests that can be understood and processed by a machine. Accordingly, the voice user interface may understand free form human utterances, freeing the user from restrictions relating to how commands, queries, or other types of requests should be formulated. Instead, the user can use a natural or casual manner of speaking to request various voice services in an integrated environment, in which various devices can be controlled in a conversational manner, using natural language. For example, the voice user interface may be aware of data and services

associated with the navigation device, a media device, a personal computer, a personal digital assistant, a mobile phone, or various other computing devices or systems available in the environment.

[016] According to various aspects of the invention, the natural language voice user interface may include an input mechanism that receives a voice-based input, which includes at least an utterance or verbalization spoken by a user. The input mechanism may include a suitable device or combination of devices that can receive voice-based inputs (e.g., a directional microphone, an array of microphones, or other devices that encode speech). The input mechanism can be optimized to maximize gain in a direction of a user, cancel echoes, null point noise sources, perform variable rate sampling, filter out background conversations or environmental noise, or otherwise optimize fidelity of encoded speech. As such, the input mechanism may generate encoded speech generated in a manner that tolerates noise or other factors that could otherwise interfere with interpreting speech. Further, in various implementations, the input mechanism may include one or more other (non-voice) input modalities, which can be processed and/or correlated with one or more previous, current, or subsequent utterances or other voice-based inputs. As such, a user can provide other forms of input using a touch-screen interface, a stylus/tablet interface, a keypad or keyboard, or other input interfaces, for example, to clarify utterances or provide additional information about the utterances using other input modalities. For instance, the user could touch a stylus or other pointing device to a portion of a map displayed on a touch-screen interface, while also providing an utterance relating to the touched portion (e.g., "Show me restaurants around here."). In this example, the inputs can be correlated to interpret "around here" as likely referring to the touched portion of the map, as distinct from the user's current location or some other meaning.

[017] According to various aspects of the invention, the natural language voice user interface may include an Automatic Speech Recognizer that processes encoded speech to generate one or more preliminary interpretations of what was said in an utterance (e.g., content of the utterance). For example, the Automatic Speech Recognizer may generate the preliminary interpretations using phonetic dictation to recognize a stream of phonemes based on a dynamically adaptable recognition grammar. The dynamically adaptable recognition grammar may be based on dictionaries or phrases from various input domains (e.g., domains for languages, navigation, music, movies, weather, various temporal or geographic proximities, or various other domains).

Thus, the Automatic Speech Recognizer may generate one or more interpretations of an utterance, which may be represented as a series of phonemes or syllables. The one or more interpretations can be analyzed (e.g., utilizing phonotactic rules or models of human speech) to generate an N-best list of preliminary interpretations as to what was spoken by the user. The preliminary interpretations may then be provided to a conversational language processor, which utilizes shared knowledge, contextual information, and various other sources of information to generate an intelligent hypothesis as to an actual meaning, the user's intent, or other aspects of the utterance. By formulating the hypothesis using various features and components that model everyday human-to-human conversations, the conversational language processor may generate a hypothesis as to the meaning or intent of the utterance, which can inform a process of resolving one or more requests contained in the utterance.

[018] According to various aspects of the invention, the natural language voice user interface may include, among other things, a context tracking engine that establishes meaning for a given utterance. For example, the context tracking engine can manage competitions among one or more context-specific domain agents that establish the meaning (e.g., redistributable, adaptable engines or modules that provide functionality for a given context, domain, system, or application). The domain agents may analyze preliminary interpretations of an utterance to generate a domain-specific probable interpretation. For example, one or more of the agents may include adaptable vocabularies, concepts, available tasks, or other forms of information specific to the respective domain or context. In addition, the agents can use a voice search engine to search a network for information that may not be available within the system. The probable interpretation can be assigned a weighted ranking or score, which can be used to select a "winning" one of the agents. Thus, the winning one of the agents may be designated responsible for establishing or inferring further information (e.g., based on domain or context-specific information), updating the shared knowledge, or resolving requests in the utterance, among other things. The context tracking engine may also maintain a context stack to track conversation topics, track previously invoked agents, evaluate criteria, weigh parameters, or otherwise maintain information relating to a conversational context (e.g., the context stack may be traversed in light of recent contexts, frequently used contexts, or other information included therein to determine a most likely intent of the user). By identifying a context, or correlatively, by identifying capabilities, tasks, vocabularies, or other information within the context, the

context tracking engine can provide relevant information for establishing intent in addition to phonetic clues associated with the utterance (e.g., a word with multiple possible meanings may be disambiguated based on its meaning in a given context, previous usage in a dialogue, etc.).

[019] According to various aspects of the invention, the natural language voice user interface may utilize various forms of information to enable sharing of assumptions and expectations relating to a given utterance, conversation, or other human to machine interaction. For example, to inform decision making in the voice user interface, the voice user interface may include information sources containing short-term and long-term shared knowledge relating to user behavior, preferences, or other characteristics (e.g., short-term and long-term profiles of a specific user, global users, peer users, etc.). The short-term shared knowledge may accumulate information during a current conversation to dynamically establish awareness of a state of the voice user interface (e.g., recognition text for previous utterances, a cross-modal user interface manipulation history, a list of previously selected tasks or invoked queries, or other information). Storage of the short-term knowledge may be modeled after human interaction, such that certain information may be expired after a psychologically appropriate amount of time (e.g., to expunge stale data), whereas information with long-term significance can be added to the long-term shared knowledge (e.g., to establish persistent awareness of data likely to remain static over time). As such, the long-term shared knowledge may profile or otherwise model various characteristics, preferences, behavior, or other information relating to a user based on information accumulated over time (e.g., user-specific jargon, demographics, cognitive patterns, frequently requested tasks, favorite topics or concepts, etc.). Accordingly, the voice user interface may utilize various forms of information available via a context tracking engine, domain agents, a voice search engine, shared knowledge, internal or external databases, data associated with other devices, or other knowledge sources. As a result, a conversational type or goal associated with utterance can be identified. Based on the available information and the type or goal of the conversation, the voice user interface may attempt to resolve the utterance (e.g., by invoking an agent, which utilizes one or more applications to perform a requested task, retrieve requested information, etc.).

[020] According to various aspects of the invention, the natural language voice user interface may generate cross-modal intelligent responses sensitive to syntax, grammar, and context, which may provide a conversational feel to system-generated responses. When available, a generated

intelligent response may present results of a resolved request (e.g., feedback relating to a task, information retrieved as a result of a query, etc.). Further, the intelligent responses may be provided across modalities, for example, using various combinations of verbal and/or non-verbal outputs (e.g., information may be presented on a display device, via an audible alert, a verbalized output, another output mechanism, or various combinations thereof). Further, verbalized components of the cross-modal response may be adapted to the user's manner of speaking (e.g., by varying tone, pace, timing, or other variables), thus creating verbal responses having natural variation and personality. The intelligent responses may also be formulated to provide an output that guides the user towards a subsequent response (e.g., a subsequent utterance) that may be more likely to be recognized. For example, when an utterance cannot be resolved because of an ambiguous context, unrecognizable words or phrases, or other factors that can result in an ambiguous or unrecognized interpretation, the intelligent response can be framed to disambiguate context or request additional information from the user to clarify a meaning of the utterance. Further, when subsequent information indicates that a given interpretation or hypothesis was incorrect, one or more previous utterances may be reinterpreted to refine context and update short-term or long-term conversational models (e.g., an utterance of "No, I meant . . ." may indicate that a previous utterance was interpreted incorrectly, in which case various previous utterances may be reinterpreted based on a correct interpretation, thus building a more accurate conversational context). As a result, the voice user interface may tolerate full or partial failure through adaptive mechanisms.

[021] According to various aspects of the invention, the natural language voice user interface may provide voice navigation services within an agent-based architecture. The architecture may include a plurality of adaptable agents (e.g., specialized software, data, content, or other information that provide functionality, behavior, services, data, and other information in a plurality of respective contextual domains), at least one of which provides navigation services (e.g., route calculation, map control, location-sensitive information, data-driven directions, etc.). As the agents process requests, the agents may autonomously react, adapt, and otherwise reconfigure to provide optimal voice services in respective domains. For example, by building context over time (e.g., by generating short and long-term profiles of a user, conversations with the user, frequent topics or preferences, etc.), the agents may automatically combine knowledge, adapt preferences, remove conflicts, or perform other adaptations to refine or otherwise optimize

an operational framework thereof. Adaptation of the agents, which include at least a navigation agent, may occur across the plurality of agents, for example, in response to various ones of the agents resolving voice-based requests. As such, the adaptation may occur autonomously as a by-product of the agents providing voice services, generating inferences, identifying affinities (e.g., among users, peers, communities, etc.), receiving updates from external sources (e.g., using an update manager), or in other ways, as will be apparent.

[022] According to various aspects of the invention, the natural language voice user interface may include an agent-based architecture for providing voice navigation services. For example, the agent-based architecture may include one or more domain or context-specific agents, which include at least a navigation agent. The navigation agent may include, among other things, various navigation-specific content packages (e.g., dictionaries, available queries, tasks, commands, dynamic grammars, etc.), response lists (e.g., appropriate responses to commands, queries, or other requests), personality profiles (e.g., for creating a natural feel to system-generated speech), substitution lists (e.g., for substituting or transforming data into a structured form that can be understood by a target information source), or various other forms of navigation-specific information. Further, the navigation agent may be associated with pointers to local or remote data sources, parameters and operating data provided associated with other services in the architecture, or various other forms of information. For example, the data sources used by the navigation agent may include data relating to navigation, points-of-interest, traffic, events, parking, personal data, peer affinities, or various other sources of information. Further, the data sources may be populated, extended, pruned, or otherwise constructed through adaptation, analysis of various models, communication with a data service, or in other ways, as will be apparent.

[023] According to various aspects of the invention, the natural language voice user interface may include a navigation agent, which may be coupled with various sources of information, and may make use of context, communicating with various other adaptable agents and other system components to provide voice navigation services. For example, the navigation agent may use contextual information relating to a navigation domain, including tracked topics, user locations, routes traveled, previous requests, user interface states, user behaviors, preferences, demographics, or other characteristics, or various other types of contextual information. As a result, the navigation agent may have various sources of knowledge and resources available to

resolve voice navigation requests. For example, the navigation agent may generate inferences (e.g., using an inference engine) using the available knowledge and resources to apply various rules, policies, or other inferencing techniques to generate interpretations of an utterance (e.g., phonctic fuzzy matching, inductive logic, Bayesian probability analysis, monotonic or non-monotonic reasoning, etc.). As such, the navigation agent can infer keywords or criteria not explicitly provided in the utterance, determine suitable responses to subjective or indeterminate utterances, generate events, identify peer affinities, or otherwise generate inferences for resolving navigation-related requests.

[024] According to various aspects of the invention, the natural language voice user interface may include one or more inference engines, which can generate various inferences through awareness of previous context, short-term or long-term shared knowledge, command histories, states of vehicular systems, user interface states, and various other data sources. In various implementations, one or more of the agents may be associated with a respective inference engine that can generate inferences using domain-specific knowledge, rules, policies, or other criteria. For instance, the inference engines may identify keywords or criteria missing in an utterance, infer intended meanings, autonomously suggest available tasks, or otherwise assist an associated agent in identifying queries, commands, or other requests contained in an utterance. Moreover, when information cannot be suitably resolved using information sources associated with the navigation agent, or through generating inferences, the information may be requested from one or more other agents, other devices, network information sources (e.g., via a voice search engine), or in other ways, as will be apparent. Upon identifying the information through one or more of the other sources, the requesting agent may be adapted to make the information subsequently available. Thus, various devices, applications, systems, and other components of an architecture may cooperatively share available information and services (e.g., context, dialogue histories, shared knowledge, maps, points of interest, contact lists, user or peer affinities, dynamic grammars, available applications, command histories, etc.). Accordingly, the architecture may provide an integrated voice navigation services environment in which users can speak natural language requests relating to various available contexts, domains, applications, devices, information sources, or various combinations thereof.

[025] According to various aspects of the invention, the natural language voice user interface may accept natural language voice-based inputs to control an electronic device that can provide

navigational information, in addition to various other devices associated with an environment in which the voice user interface operates. Furthermore, various functional aspects of the voice user interface may reside at a client device, at a server, or various combinations thereof.

[026] According to various aspects of the invention, the natural language voice user interface may support multi-modal voice inputs. Thus, a given multi-modal voice input may include at least a voice component (e.g., an utterance) in addition to one or more non-voice input components (e.g., inputs provided via a keypad, a touch-screen, a stylus/tablet combination, a mouse, a keyboard, or other input modalities). As such, the non-voice input component can provide additional information or clarification relating to the utterance, adding to an amount of input information available when processing voice. For example, the user can use other input modalities to clarify a meaning of the utterance, provide additional information about the utterance, reduce a number of device interactions needed to make a given request, or otherwise provide additional in relation to a given utterance.

[027] According to various aspects of the invention, the natural language voice user interface may utilize various of cognitive models, contextual models, user-specific models, or other models to identify queries, commands, or other requests in a voice input. For example, a given input may include information relating to one or more contextual domains, one or more of which may be invoked to interpret and/or infer keywords, concepts, or other information contained in the input. Moreover, short-term and long-term shared knowledge about a user's behavior and preferences may be used in a hybrid recognition model that also considers semantic analysis and contextual inferences. For example, certain syllables, words, phrases, requests, queries, commands, or other information may be more likely to occur in a given context. Thus, the hybrid recognition model may analyze semantic patterns to resolve what was said by an utterance, and may further rely on contextual history or other information to resolve what was meant by the utterance. The hybrid recognition model may be used in conjunction with, or independently of, a peer to peer recognition model. For example, recognition models may include awareness of global usage patterns, preferences, or other characteristics of peer users, where certain keywords, concepts, queries, commands, or other aspects of a contextual framework may employed by peer users within the context.

[028] Other objects and advantages of the invention will be apparent based on the following drawings and detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

[029] Figure 1 illustrates a block diagram of an exemplary system for providing a navigation device with a conversational, natural language voice user interface according to various aspects of the invention.

[030] Figure 2 illustrates a block diagram of exemplary agent-based architecture for providing a navigation device with a conversational, natural language voice user interface according to various aspects of the invention.

[031] Figure 3 illustrates a flow diagram of an exemplary method for dynamically generating recognition grammars for a navigation device having a conversational, natural language voice user interface according to various aspects of the invention.

[032] Figure 4 illustrates a flow diagram of an exemplary method for processing multi-modal voice-based inputs provided to a navigation device having a conversational, natural language voice user interface according to various aspects of the invention.

[033] Figure 5 illustrates a flow diagram of an exemplary method for calculating routes based on voice-based inputs provided to a navigation device having a conversational, natural language voice user interface according to various aspects of the invention.

[034] Figure 6 illustrates a flow diagram of an exemplary method for providing voice services based on a current navigation route according to various aspects of the invention.

DETAILED DESCRIPTION

[035] According to various aspects of the invention, Figure 1 illustrates a block diagram of an exemplary system 100 for providing a navigation device with a conversational, natural language voice user interface. As illustrated in Figure 1, the system 100 may enable a user to engage in a natural language, cooperative, conversational dialogue with a voice-enabled navigation device. For example, the system 100 may understand free form human utterances, freeing the user from being restricted in how commands, queries, or other types of requests should be formulated. Instead, the user can use a casual or natural manner of speaking to request various voice navigation services, among various other voice services (e.g., services relating to telematics, communications, media, messaging, external systems, marketing, information retrieval, or various other computational services). As such, the user can use system 100 to control

navigation devices, media devices, personal computers, personal digital assistants, wireless phones, or other computing devices or systems in a conversational manner, using natural language. By making use of context, prior information, domain knowledge, short and long-term shared knowledge, noise tolerance, and cognitive models, among other things, system 100 can provide an integrated environment in which users can speak conversationally, using natural language, to issue queries, commands, or other requests that can be understood and processed by a machine. For example, in various implementations, system 100 may utilize one or more techniques as described in co-pending U.S. Patent Application 10/452,147, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed June 3, 2003, and co-pending U.S. Patent Application 10/618,633, entitled "Mobile Systems and Methods for Responding to Natural Language Speech Utterance," filed June 15, 2003, the disclosure of which are hereby incorporated by reference in their entirety.

[036] According to various aspects of the invention, system 100 may include an input mechanism 105 that receives a voice-based input, including at least one utterance or verbalization spoken by a user. The input mechanism 105 may include a device or combination of devices suitable for receiving a voice-based input (e.g., a directional microphone, an array of microphones, or other devices that encode speech). In various implementations, the input mechanism 105 can be optimized to receive the voice-based input, for example, by maximizing gain in a direction of a user, cancelling echoes, nulling point noise sources, performing variable rate sampling, filtering out background conversations or environmental noise, or performing various other techniques for maximizing fidelity of encoded speech. As such, the encoded speech generated by input mechanism 105 may be highly tolerant of noise or other factors that may potentially interfere with interpreting human speech.

[037] Further, in various implementations, the input mechanism 105 may be coupled to other input modalities, in that various forms of input other than voice can be processed and/or correlated with one or more previous, current, or subsequent voice-based inputs. For example, the input mechanism 105 may be coupled to a touch-screen interface, a stylus/tablet interface, a keypad or keyboard, or other devices or system interfaces, as will be apparent. As a result, an amount of input information potentially available for system 100 to process voice may be maximized, as the user can clarify utterances or provide additional information about utterances using other input modalities. For instance, the user could touch a stylus or other pointing device

to a portion of a map displayed on a touch-screen interface, while also providing an utterance relating to the touched portion (e.g., "Show me restaurants around here."). In this example, system 100 may correlate the inputs to interpret "around here" as likely referring to the touched portion of the map, as distinct from the user's current location or another meaning.

[038] System 100 may also include an Automatic Speech Recognizer 110 that receives the encoded voice input and generates one or more preliminary interpretations thereof. For example, the Automatic Speech Recognizer 110 may recognize the voice-based input using phonetic dictation to recognize a stream of phonemes based on a dynamically adaptable recognition grammar. As a result, the Automatic Speech Recognizer 110 may provide out-of-vocabulary capabilities, which may be tolerant of a user misspeaking, portions of a speech signal being dropped, or other factors that could interfere with interpreting an utterance. The dynamically adaptable recognition grammar may be based on dictionaries or phrases from various input domains (e.g., domains for different languages, navigation, music, movies, weather, various temporal or geographic proximities, or various other domains). Further, performance of the Automatic Speech Recognizer 110 may be improved, for example, by pruning a search space associated with the recognition grammar (e.g., the grammar can include a linking element, such as *schwa*, to represent an unstressed, central vowel that tends to be spoken frequently, even without the user's conscious awareness). Thus, using these and other techniques, Automatic Speech Recognizer 110 may analyze an incoming encoded utterance to represent portions of the utterance as a series of phonemes or syllables, which can be further broken down into core components of an onset, a nucleus, and a coda, among other sub-categories. The series of phonemes or syllables can then be analyzed (e.g., utilizing phonotactic rules that model human speech) to identify a plurality of preliminary interpretations or best guesses (e.g., an N-best list) as to what was actually spoken by the user. It will be apparent, however, that the Automatic Speech Recognizer 110 may use various techniques to generate the preliminary interpretations of the encoded utterance, including those described, for example, in co-pending U.S. Patent Application Ser. No. 11/513,269, entitled "Dynamic Speech Sharpening," filed August 31, 2006, the disclosure of which is hereby incorporated by reference in its entirety.

[039] The plurality of preliminary interpretations generated by the Automatic Speech Recognizer 110 may be provided to a conversational language processor 120, which utilizes shared knowledge to generate an intelligent hypothesis of an actual meaning, a user's intent, or

other aspect of the voice-based input. The conversational language processor 120 may formulate the hypothesis using various features and components that collectively operate to model everyday human-to-human conversations.

[040] For example, the conversational language processor 120 may include a context tracking engine 140 that establishes meaning for a given utterance by, among other things, managing a competition among one or more context-specific domain agents 125 (e.g., redistributable, adaptable packages or modules that provide functionality for a given context, domain, system, or application). For example, the preliminary interpretations may be ranked by the Automatic Speech Recognizer 110, and the agents 125 may further analyze the preliminary interpretations to generate a weighted ranking or score, which can be used to select a "winning" one of the agents 125. The winning one of the agents 125 may then be responsible for establishing or inferring further information, updating the shared knowledge, or performing other tasks to aid in generating the intelligent hypothesis. Moreover, the context tracking engine 140 may use a context stack to track conversation topics, evaluate criteria, weigh parameters, or otherwise maintain contextual information for generating the hypothesis based on a conversational context (e.g., a context stack may be traversed in light of recent contexts, frequently used contexts, or other information included therein to determine a most likely intent of the user). By identifying a context, or correlatively, by identifying capabilities, tasks, vocabularies, or other information within the context, the context tracking engine 140 can provide highly relevant information for establishing intent, apart from meager phonetic clues.

[041] Furthermore, the conversational language processor 120 may utilize various other forms of knowledge to inform the generation of the intelligent hypothesis. For example, various agents 125 may adaptively include domain-specific or context-specific vocabularies, concepts, available tasks, or other forms of information relevant to the respective domain or context. In addition, the various components associated with the conversational language processor 120 can invoke a voice search engine 135 (e.g., an engine that searches a network for information) to resolve information that may not be internally available (e.g., when an external knowledge source may be useful in clarifying an intent behind a particular word, command, query subject, or other aspect of an utterance).

[042] System 100 may also enable the user and the system 100 to share assumptions and expectations relating to a given utterance, conversation, or other interaction. For example, the

conversational language processor 120 may be coupled to one or more data repositories 160 that store short-term and long-term shared knowledge that inform decision making in the conversational language processor 120. The short-term shared knowledge may accumulate information during a current conversation (e.g., recognition text for previous utterances, a cross-modal user interface manipulation history, a list of previously selected tasks or invoked queries, or other information), thus dynamically establishing awareness of a cross-modal state of the voice user interface. Storage of the short-term knowledge may be modeled after human interaction, thus certain data may be expired after a psychologically appropriate amount of time (e.g., to expunge stale data), and information with long-term significance can be added to long-term shared knowledge (e.g., a new address of a long-term contact of the user). As such, the long-term shared knowledge may profile or otherwise model environmental, cognitive, historical, demographic, or other aspects of a user based on information accumulated over time (e.g., user-specific jargon, frequently requested tasks, favorite topics or concepts, etc.).

[043] Accordingly, the conversational language processor 120 includes various features that can be used to generate intelligent hypotheses as to a user's intent in a given utterance. For example, the hypothesis may be based on information provided via the context tracking engine 140, the agents 125, the voice search engine 135, the shared knowledge, or other knowledge sources. As a result, the conversational language processor 120 may attempt to identify a conversational type or goal of the utterance (e.g., a query for retrieving a discrete piece of information, a didactic interaction for clarifying information provided by the voice user interface, or an exploratory interaction in which conversational goals may be improvised as the conversation progresses, etc.). Based on the information available and the type or goal of the conversation, the generated hypothesis can be assigned a degree of certainty, which can inform how conversational language processor 120 resolves the utterance. For example, when the degree of certainty indicates that sufficient information has been identified, the conversational language processor may invoke one or more of the agents 125, which in turn can utilize one or more applications 150 to perform a requested task (e.g., a task relating to a navigation application, an advertising application, a music application, an electronic commerce application, or other suitable applications or tasks). In another example, one or more of the agents 125 may query the data repositories 160 or the voice search engine 135 to retrieve requested information, or otherwise take action to resolve a request for information contained in the utterance.

[044] Additionally, the conversational language processor 120 may generate cross-modal intelligent responses, which may be syntactically, grammatically, and contextually sensitive, thus providing a conversational feel to system-generated responses. When available, the intelligent response may present results of a performed task or an executed query to the user, and the response may be provided across modalities. For example, verbal and/or non-verbal outputs 180 may be used separately or in concert (e.g., by presenting information using a display device, an audible alert, a verbalized output, another output mechanism, or various combinations thereof). Further, a verbalized component of the cross-modal output 180 may be adapted to the user's manner of speaking (e.g., by varying tone, pace, timing, or other variables), thus creating verbal responses having natural variation and personality.

[045] The intelligent responses may also be formulated to provide an output 180 that guides the user towards a subsequent response likely to be favorable for recognition. For example, when the degree of certainty reflects an ambiguous context (e.g., when the competition results in a deadlock among the agents 125), an adaptive misrecognition engine 130 may identify the utterance as ambiguous or unrecognized, and the intelligent response can be framed to disambiguate context, or request a subsequent response from the user to clarify the meaning of the utterance, for example. Further, when the conversational language processor 120 or the misrecognition engine 130 determines that a given interpretation or hypothesis was incorrect, one or more previous utterances may be reinterpreted to refine context and build more accurate short-term or long-term conversational models (e.g., an utterance of "No, I meant . . ." may indicate that a previous utterance was interpreted incorrectly, in which case various previous utterances may be reinterpreted based on a correct interpretation, thus building a more accurate conversational context). As a result, the conversational language processor 120 may recover from full or partial failure. Additional techniques for adaptively responding to misrecognition or ambiguity may be included, such as those described in co-pending U.S. Patent Application Ser. No. 11/200,164, entitled "System and Method of Supporting Adaptive Misrecognition in Conversational Speech," filed August 10, 2005, the disclosure of which is hereby incorporated by reference in its entirety.

[046] Additional information relating to the various techniques described herein, as well as other techniques that system 100 uses to provide a conversational, natural language interaction may be provided, for example, in co-pending U.S. Patent Application Ser. No. 11/197,504,

entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed August 5, 2005, co-pending U.S. Patent Application Ser. No. 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," filed August 29, 2005, and co-pending U.S. Patent Application Ser. No. 11/580,926, entitled "System and Method for a Cooperative Conversational Voice User Interface," filed October 16, 2006, the disclosures of which are hereby incorporated by reference in their entirety.

[047] As such, system 100 may provide an environment where conversational, natural language interactions can occur between a user and the system 100. Further, as will be described in greater below, system 100 can be implemented to provide the conversational, natural language interactions, for example, as a voice user interface to one or more electronic devices or applications, including a navigation device. As will also be apparent from the further descriptions provided herein, the system 100 has capabilities for cross-device and cross-modal awareness, such that system 100 provides an environment in which the user can engage in a cooperative, natural language dialogue to request voice services relating to many different devices or applications.

[048] According to various aspects of the invention, Figure 2 illustrates a block diagram of exemplary agent-based architecture 200, which can provide a conversational, natural language voice user interface to a navigation device. The architecture 200 may include a plurality of agents 225a-n, which include specialized software, data, content, or other information that provide functionality, behavior, services, data, and other information in a plurality of respective contextual domains. The architecture 200 may be an integrated and dynamically adaptable environment, in which the agents 225 autonomously react, adapt, and reconfigure to provide optimal service in the environment. For example, by building context over time (e.g., by generating short and long-term profiles of a user, conversations with the user, frequent topics or preferences, etc.), the navigation agent 225a may automatically combine knowledge, adapt preferences, remove conflicts, or perform other adaptations to refine or otherwise optimize an operational framework thereof.

[049] For example, the agents 225, which include at least a navigation agent 225a, may adapt through ongoing use of short-term shared knowledge 270 and long-term shared knowledge 275 about a user's behavior, preferences, or other characteristics. Further, agent adaptation may occur across the plurality of agents 225, for example, in response to various ones of the agents

225 resolving voice-based requests (e.g., by invoking one or more applications 250, querying one or more data sources 260, cooperating with one or more other agents 225, or learning through a misrecognition engine 230, a voice search engine 235, a context tracking engine 24, etc.). In various implementations, adaptation may occur autonomously as a by-product of the agents 225 providing voice services, generating inferences, identifying affinities (e.g., among users, peers, communities, etc.), receiving updates from external sources (e.g., an update manager can update content, behavior, data, commands, domain knowledge, keywords, concepts, dictionaries, or other information for one or more of the agents 225), or in other ways, as will be apparent.

[050] As illustrated in Figure 2, the architecture 200 provides the conversational, natural language voice user interface to the navigation device by way of at least a navigation agent 225a. The navigation agent 225a can include, among other things, various navigation-specific content packages (e.g., dictionaries, available queries, tasks, commands, dynamic grammars, etc.), response lists (e.g., appropriate responses to commands, queries, or other requests), personality profiles (e.g., for creating a natural feel to system-generated speech), substitution lists (e.g., for substituting or transforming data into a structured form that can be understood by a target information source), or various other forms of navigation-specific information. Further, the navigation agent 225a may be associated with parameters and operating data provided by other services in the architecture 200 (e.g., available devices, applications 250, agents 225, etc.), pointers to local or remote data sources (e.g., short-term shared knowledge 270, long-term shared knowledge 275, data sources 260, voice search engine 235, etc.), among various other forms of information.

[051] The data sources 260 used by the navigation agent 225a may include, among other things, data relating to navigation (e.g., maps, addresses, street names, directories, alternate routes, etc.), points-of-interest (e.g., restaurants, hotels, museums, tourist attractions, gas stations, etc.), traffic (e.g., dynamic road conditions, traffic, detours, or other traffic-related information), events (e.g., sporting events, concerts, protests, etc.), parking (e.g., parking garages, restricted areas or times, street lots, street parking, etc.), personal data (e.g., telephone numbers and addresses associated with a user's contact lists), peer affinities (e.g., recommendations based on similarities among users having similar preferences, demographic profiles, or other characteristics). The data

sources 260 may be populated in various ways, such as being based on one or models, received via a data service, extended or refined through adaptation, or in other ways, as will be apparent.

[052] Further, contextual information relating to a navigation domain may be maintained via the context tracking engine 240, short-term shared knowledge 270, and long-term shared knowledge 275, among other things. For example, the contextual information may relate to tracked topics, user location, routes traveled, previous requests, user interface state, user behavior, preferences, demographics, or other characteristics, among other types of contextual information. For example, context may be built and utilized using techniques described in greater detail abovc with reference to Figure 1, or in the above-referenced co-pending U.S. Patent Applications, or various combinations thereof. Moreover, when navigation agent 225a cannot find information locally within architecture 200, and the information cannot be inferred, the information may be requested from onc or more other agents 225b-n (e.g., to request domain-specific information), other devices, the voice search engine 235 (e.g., to search a network), external sources (e.g., the update manager may be invoked to download available updates, which may have the information), or other sources, as will be apparent. If the information can be located through one or more of the other sources, the navigation agent 225a may be adapted, extended, or otherwise updated to make the information available subsequently.

[053] Accordingly, the navigation agent 225a may be coupled with various sources of information, and may make use of context, while communicating with various other adaptable agents 225b-n and other system components to provide voice navigation services. Thus, context, shared knowledge, maps, points of interest, contact lists, user or peer affinities, dynamic grammars, available applications, or various other aspects of the navigation architecture 200 can be shared or otherwise made available to various components, devices, or aspects of the architecture 200. As a result, an inference engine 255 included with the navigation agent 225a may various knowledge sources and other resources available to provide an integrated voice navigation services environment. For example, the inference engine 255 may generate inferences from the available knowledge and resources by applying one or more rules, policies, or other inferencing mechanisms to generate probable interpretations of utterances in the environment (e.g., phonetic fuzzy matching, inductive logic, Bayesian probability analysis, monotonic or non-monotonic reasoning, etc.).

[054] For instance, as described in greater detail above, the navigation agent 225a and one or more other agents 225b-n may compete in analyzing one or more preliminary interpretations of an utterance to generate one or more respective probable interpretations of the utterance. The navigation agent 225a may then be invoked to process the probable interpretation (e.g., by resolving information contained in the utterance) upon being identified as a "winning" one of the competing agents. For example, which of the competing agents wins may depend on context or other information contained in the utterance, whereby the navigation agent 225a may win and recalculate a route in response to an utterance of "This traffic is terrible, maybe back roads," while a music agent may win and change a radio channel in response to an utterance of "Traffic is terrible, maybe some Oldies," while a movie agent may win and search for movie show times in response to an utterance of "Traffic was terrible, maybe a comedy." Thus, when the navigation agent 225a generates a probable interpretation that results in being a winning one of the agents 225, the navigation agent 225a may manage processed for searching, inferring, or otherwise identifying information in light of context to provide voice navigation services using one or more applications 250 (e.g., navigation, grammar generation, location determination, response generation, advertising, device control, or other applications available within architecture 200).

[055] For example, the navigation agent 225a may utilize the inference engine 255 to infer keywords or criteria not explicitly provided in the utterance, determine suitable responses to subjective or indeterminate utterances (e.g., selecting a most likely answer to a query, asking a user for more information, etc.), generate events, identify peer affinities, or otherwise generate inferences for resolving navigation-related requests. The inference engine 255 may generate such inferences through awareness of previous context (e.g., through information provided by the context tracking engine 240), short-term or long-term shared knowledge 270, command histories, states of vehicular systems, user interfaces, or other devices, data sources 260, or other available information. For example, in an exemplary illustration, previous context, shared knowledge, and other information sources may be utilized in conjunction with the inference engine 255 and/or various other components of architecture 200 to enable a human to machine interaction that may occur as follows:

User:	"Show me restaurants around here."

Voice User Interface:	"All I can find is fast food, which you don't like."
User:	"Okay, how about closer to the hotel?"
Voice User Interface:	"Found an Indian restaurant with good reviews. Want me to call them?"
User:	"Yeah go ahead, that sounds good."

[056] In the above illustrated example, the first utterance establishes a context of "restaurants," and the voice user interface responds to the utterance in light of a shared knowledge inference indicating that the user does not like fast food. Furthermore, while the second utterance does not explicitly reference a restaurant, the context established in the first utterance can be used to infer the unreferenced information, forming an unambiguous request out of what would otherwise be ambiguous. Further still, long-term shared knowledge 275 about the user (e.g., the user may frequently go to Indian restaurants) and short-term knowledge 270 about the user (e.g., a hotel associated with a current route) can narrow a search space for resolving the second utterance. Moreover, the voice user interface may invoke the response generation application to advance the conversation towards a resolution (e.g., by prompting the user to call the restaurant). Because the integrated environment has cross-device awareness, the user's subsequent response can result in the device control application formulating an appropriate command structure to dial the user's mobile phone, which may or may not be distinct from the navigation device. Accordingly, as illustrated in this example, robust awareness of context, shared knowledge, available applications, and available systems or devices, among other things, has enabled the user to use natural language to locate a restaurant along a current route matching the user's preferences, while also enabling the user to dial a phone to call the restaurant, without the user having to manually interact with devices or otherwise interact with cumbersome human to machine interfaces.

[057] It will be understood, however, that the above illustrated dialogue provides but one example of a cooperative conversation that can occur between a user and a system that includes the above-described architecture 200. Because the architecture 200 supports free form interactions, and can tolerate serendipitous variations in human speech, vocabulary usage, or other factors, it will be understood that no two conversational interactions between the user and the system will necessarily be alike. As such, the user can seamlessly switch among contexts,

allowing subsequent utterances to occur in various available domains (e.g., a navigation-related utterance of "Take me to the Seahawks game" may be followed up with subsequent cross-domain utterances, such as "Are there tickets left?" or "When does it start?"). Thus, although various descriptions illustrate exemplary ways in which the agent-based architecture 200 can be used for voice navigation services, it will be apparent that, in a given utterance, the user can provide natural language voice-based requests relating to various available contexts, domains, applications, devices, information sources, or various combinations thereof.

[058] The navigation application available within architecture 200 may resolve natural language, voice-based requests relating to navigation (e.g., calculating routes, identifying locations, displaying maps, etc.). The navigation application can provide a user with interactive, data-driven directions to a destination or waypoint, wherein the user can specify the destination or waypoint using free-form natural language (e.g., the user can identify full or partial destinations, including a specific address, a general vicinity, a city, a name or type of a place, a name or type of a business, a name of a person, etc.). As a result, in various implementations, the navigation application may perform post-processing on the full or partial destinations to identify a suitable address for calculating a route (e.g., a closest address that makes "sense"). For example, an utterance containing a full or partial destination may be analyzed to identify one or more probable destinations (e.g., an N-best list of destinations), which may be subject to post-processing that weighs the identified probable destinations based on various factors. Thus, the probable destination most likely to provide a suitable (preliminary or complete) route may be identified from the N-best list.

[059] For example, in an illustrative conversation between a user and the voice user interface, the user may utter a voice-based input of "Take me to Seattle." Based on a current location of the user (e.g., as determined using one or more navigation sensors, radio frequency identifiers, local or remote map databases, etc.), the navigation application may select an address in Seattle that provides a route from the current location to Seattle (e.g., a central point in Seattle may be selected for long distance travel, whereas an address in northern Seattle may be selected for a current location being close to north Seattle). Further, when the requested destination does not specify a final destination (e.g., as in the above example), the user may successively refine the final destination in subsequent requests, and the post-processing may continue to select additional addresses based on the current location, a current route, or other factors. For instance,

continuing the above-provided example, a subsequent input of "Take me to Pike Street" may result in the post-processing analyzing the current route to determine an appropriate address on Pike Street, and possibly recalculating the route, as necessary. Thus, the current route may have the user driving north-bound on Interstate-5, such that an address may be selected on Pike Street on a same side as north-bound lanes of Interstate-5 (e.g., preserving current routing information). The final destination may continue to be successively refined, in combination with the post-processing. As a result, a user knowing only that a destination falls along a 500-block of Pike Street can specify, "501 Pike Street," and even if such an address does not exist, the post-processing may find a closest address that makes sense, thus routing the user to the 500-block of Pike Street.

[060] In addition, the navigation application may provide dynamic, data-driven directions to the destination or dynamically provide routing information, among other things. Furthermore, the response generation application may work together with the navigation application and/or one or more other applications to generate system responses personalized for the user. For instance, the navigation application may access data associated with various user-specific and environmental data sources to provide personalized data-driven directions along a route, which can be recalculated or modified based on information taken from the data sources. Thus, in one example, the response generation application may generate personalized data-driven directions as follows:

Voice User Interface:	"Go to Bob's House."
Voice User Interface:	"Take a left on 14th Street."
Voice User Interface:	"Make a right here to get on the highway, and prepare for heavy traffic."

[061] In the above-illustrated example, the response has been personalized by framing directions in terms of one of the user's contacts (i.e., Bob), while an available source of traffic data has been used to provide data-driven directions. In another example, data that may affect routing can be identified and used to recalculate a route or identify alternate routes (e.g., weather data may be used to avoid snowy roads, or event data may be used to avoid a route along which a protest may be occurring). As such, data may be obtained dynamically to identify alternate

routes, recalculate routes, or otherwise provide optimal routing service. In various implementations, the source of data may be based on persistent models, dynamically received data (e.g., a given highway may regularly have high volume during rush hour or sporting events), or other information. Routes may be recalculated automatically (e.g., user preferences may dictate that heavy traffic be avoided, or sporting event traffic may not be avoided when the user has tickets to the event), or the user can request recalculation through additional utterances (e.g., "I don't want to sit in traffic, can we go a different way?"). Further, possible answers or responses to a given utterance may be filtered according to a current route (e.g., possible results provided for an utterance of "Find me a gas station" may be filtered according to those gas stations within a given proximity of the route). Moreover, continuing the above-provided example, the user may later return to a point of origin, and personalized responses may include return directions, such as "Go back the way you came on 14th Street." Thus, the personalized responses may also create a natural feel to the directions, and may be based on context that builds over time.

[062] The navigation architecture 200 may be used in conjunction with an electronic device capable of accepting voice-based inputs and providing navigation-related information. For example, architecture 200 may be implemented within a handheld device (e.g., personal navigation device, mobile phone, etc.), a telematics devices (e.g., a vehicular navigation system), or various other suitable devices. Furthermore, various functional aspects of the architecture 200 may reside at a client device, at a server, or various combinations thereof. For example, a voice-based input may be received by the client device (e.g., a personal navigation device), which may communicate with the server to retrieve information, perform a task, process the input, or for other reasons. The server may subsequently return information to the client device in response to the request. In various implementations, however, the voice-based input may be processed within the system (e.g., processing delays may be reduced by minimizing communication with the server). Other suitable arrangements may be used, as will be apparent, and specific arrangements as to how information can be processed, retrieved, or returned to the user, among other things, may be highly dependent on context or specific implementations (e.g., what information can be obtained locally, remotely, or otherwise, what services a user subscribes to, etc.).

[063] According to various aspects of the invention, Figure 3 illustrates a flow diagram of an exemplary method 300 for dynamically generating and/or loading recognition grammars in a conversational, natural language voice user interface. For example, a grammar generator application (e.g., as illustrated in Figure 2) may dynamically generate and/or load recognition grammars for use by a navigation agent, an Automatic Speech Recognizer, a context stack, or other components in the voice user interface that incorporate a recognition grammar. While grammar generating method 300 may be used by various components of the natural language voice user interface, the method 300 may be particularly useful in reducing processing overhead and reducing perplexity or confusion. For example, by efficiently generating, updating, loading, extending, or otherwise building dynamic grammars based on various factors, as will described in greater detail herein, bottlenecks can be avoided, conflicts can be reduced, or other aspects of interpreting utterances using a grammar can be optimized.

[064] For example, in various implementations, grammar generation may be keyed to an amount of available resources in a system that includes the voice user interface. Thus, at an operation 310, the amount of system available resources may be calculated, and a subsequent operation 320 may constrain a size of the dynamic grammar based on the calculated resources. For instance, in embedded devices or other devices having low amounts of dynamic memory, the constrained grammar size may ensure that only as many resources as necessary will be occupied. Thus, throughput in the low-memory device will not be degraded by a large-list grammar over-occupying processing capacity. In another example, the size of the dynamic grammar can be reduced by eliminating redundant keywords, criteria, or other information available in the context stack, the shared knowledge, or other local sources. However, it will be apparent that various other techniques may be used to constrain grammar size, and that effectively considering available system resources can provide processing improvements in various environments.

[065] In a navigational context, in which a user's location can change from one moment to another, grammar generation may be further optimized using techniques of geographical chunking. For instance, a user's location can be determined at a given moment (e.g., via navigation sensors, global positioning systems, etc.), and an operation 330 can use the determined location to determine one or more geographic proximities associated with the location. Subsequently, at an operation 340, the determined proximities may be used to identify one or more topological domains appropriate for the determined proximities.

[066] As such, the proximities determined in operation 330 may be weighed or analyzed in various ways to determine how to utilize the topological domains identified in operation 340 in generating the dynamic grammar. For example, the topological domains may reflect physical proximities (e.g., distances from a current location, such as everything within a five-mile radius), civil organization proximities (e.g., regions, states, cities, neighborhoods, subdivisions, localities, etc.), temporal proximities (e.g., amounts of travel time from the current location, where road classes or other information may be used to estimate fastest/safest routes, such as highways and interstates being faster than city or surface roads), directional proximities (e.g., based on directional travel vectors, where a point ten miles ahead may be considered "closer" than a point one mile behind based on a direction of travel, an amount of time to turn around, etc.), or various combinations thereof. As a result, by mapping the user's geographic proximities to one or more topological domains, dynamic grammars may be pruned, extended, swapped in or out of memory, or otherwise generated and/or loaded to provide optimal recognition based on location, time, travel, or other factors (e.g., information may be swapped in and out of a grammar as a user drives from one area to another, optimizing the information along the drive to ensure that the system resources utilize information presently relevant for a given location).

[067] Furthermore, the topological domains may be subdivided into a plurality of tiles, which may in turn be subdivided into a plurality of subtiles. For example, a civil organization topological domain may include a tile representing grammar information for a state, and the tile may include one or more subtiles for counties within the state. Further, the county subtiles may include subtiles for various cities, neighborhoods, or other organizational boundaries within respective counties. Thus, the geographic proximities and topological domains used to build the dynamic grammar may be based on a variety of factors, and can be subdivided or weighed in various ways to determine what information to include in the grammar. Moreover, geographical chunks based on physical, civil organization, temporal, directional, or other proximities may be extended into various other domains in which a topological taxonomy can be placed. As a result, the geographical chunking techniques may have particular relevance in navigation or other location-dependent voice recognition systems, yet the techniques can be suitably applied in various contexts or domains in which geography or location may have relevance.

[068] As various implementations generate the recognition grammars dynamically, for example, to preserve system resources, information available to extend, update, or otherwise

update the grammar may be stored remotely (e.g., on a server). Thus, various implementations may include a system having a network connection, a data service, or another communication mechanism for establishing a link to the remote source. Based on the proximities and/or the topological domains identified for a given location, context, utterance, or otherwise, one or more grammar tiles and/or subtiles may be downloaded from the remote source at an operation 350. Further, the remote source may be adapted to store, analyze, or otherwise process certain information to refine the grammar information stored therein. For example, when a given device communicates with the remote source at operation 350, the remote source may receive identifying information relating to a user of the device (e.g., user preferences, user characteristics, request characteristics, etc.), requested information, or other information. Based on the received information, the remote source may dynamically update various tiles and/or subtiles, build affinities (e.g., among users, peers, communities, etc.), or perform other actions to refine a process by which relevant information can be identified (e.g., the remote source may determine that various different users, who share certain common demographics, may be interested in particular events, locations, etc.).

[069] Thus, the grammar information downloaded at operation 350 may be based on a variety of factors, including information in a current voice-based input, current geographical proximities of the user, topological domains or geographical chunks related to the current geographical proximities, or various kinds of affinities, among other things. Based on these and other factors, appropriate information may be retrieved from various local or remote data sources (e.g., a local repository may contain a great deal of grammar information, a portion of which may be used by the dynamic grammar to interpret an utterance). The retrieved information may then be used to generate the dynamic grammars at an operation 360. As a result, the dynamic grammars may be adapted to include information favorable for correctly interpreting a given utterance. Furthermore, at an operation 370, the favorability of correct interpretations may further be improved by post-processing the generated grammars using one or more optimization techniques, such as reducing perplexity in the grammar. For example, when the grammar generated at operation 360 includes two or more elements likely to be confused (e.g., two or more streets having a same name), the optimization operation 370 may eliminate one or more of the elements to reduce confusion (e.g., selecting the element closest to a user's current location, or other criteria). In another example, when an interpretation turns out to be incorrect (e.g., a

subsequent utterance may be "No, 14th Street Northwest, not Northeast"), the dynamic grammar may be pruned to eliminate information associated with the incorrect interpretation (e.g., information associated with Northeast may be removed). As a result, the optimization operation 370 can reduce overhead, assure that mistakes will not be repeated, and/or improve accuracy of interpretations, among other things.

[070] Although various exemplary operations have been described for dynamically extending, pruning, loading, or otherwise generating recognition grammars, it will be apparent that in operation, the natural language voice user interface may use various forms of information or techniques to determine an optimal grammar. For instance, in addition to being based on available system resources, likelihood of perplexity, geographical proximities, or user affinities, the grammar may be generated in response to conversational contexts, shared knowledge about a user, and/or adaptable agent behavior, among other factors. Accordingly, the dynamic grammar can be structured to recognize queries, commands, or other information in an utterance based on information most likely to be relevant for the utterance, with minimal ambiguity. Further, when the utterance cannot be resolved using the information currently in the grammar, various information sources or techniques may be available to update the grammar in a way that will be favorable for recognition. Thus, the above-described aspects and features of the grammar generation method 300 should be regarded as exemplary only, as many variations will be possible without departing from the inventive concepts described.

[071] According to various aspects of the invention, Figure 4 illustrates a flow diagram of an exemplary method 400 for processing multi-modal voice inputs using a conversational, natural language voice user interface. As described in connection with Figure 1 above, method 400 may be performed in a voice user interface having a mechanism for receiving voice inputs (e.g., a directional microphone, an array of microphones, or other devices that encode speech), wherein the voice inputs include at least a voice component (e.g., an utterance). The utterance, which may be received in an operation 410, may be provided using natural language or other free form manner of expression.

[072] Furthermore, to enable processing of multi-modal voice inputs, the voice user interface may have one or more input mechanisms for receiving non-voice input components (e.g., inputs provided via a keypad, a touch-screen, a stylus/tablet combination, a mouse, a keyboard, or other input modalities). As such, in operation 410, the received voice input may include a non-voice

input component in addition to the utterance, such that an amount of input information may be maximized through usage of multi-modal inputs. For example, the user can use other input modalities to clarify a meaning of the utterance, provide additional information about the utterance, reduce a number of device interactions needed to make a given request, or otherwise maximize an amount of information provided in connection with a given utterance. For example, a multi-modal voice input may be used to efficiently interact with a map displayed on a navigation device (e.g., an utterance of "I only need to see this area" may be coupled with a stylus input that circles a given portion of a map, and the combination of inputs may result in a display of the circled portion of the map).

[073] Thus, for a given voice input, a decisional operation 420 may determine whether a non-voice component accompanies an utterance in the voice input. For example, in various implementations, the various input mechanisms associated with the voice user interface may be coupled to an Alternating Voice and Data (AVD) system. As a result, input data may alternate between voice and data, or the input data may be multiplexed together in a single stream, for example, to allocate processing resources where needed at a given moment.

[074] When decisional operation 420 determines that a non-voice input component has been received, the voice and non-voice components may be correlated in an operation 430. For example, the voice component may be processed using accurate speech recognition techniques, and the voice component may further be parsed and interpreted in light of context, shared knowledge, and available data sources, among other things, as described in greater detail above. As such, correlation operation 430 may generate additional information that can inform context, shared knowledge, or other information for processing the voice input. For instance, as discussed above, short-term shared knowledge may include awareness of various user interface states, which may be affected by non-voice components. Thus, correlation operation 430 may assist the voice user interface in developing one or more preliminary interpretations of the utterance, for example, by relating information in the utterance to information in the non-voice inputs, or vice versa. For example, a multi-modal input could include a user touching a stylus to one of a number of restaurant listings displayed on a touch-screen interface, while also providing an utterance relating to the touched listing (e.g., "Call this one."). In this example, correlation operation 430 can interpret "this one" to refer to the touched telephone listing, providing additional information for processing the voice input.

[075] However, it will be apparent that voice inputs need not necessarily include non-voice components. As such, when the voice input includes only utterances, verbalizations, or other spoken components, decisional operation 420 may advance method 400 directly to an operation 440. For instance, as will be apparent from the further descriptions provided elsewhere in this specification, the natural language voice user interface may efficiently resolve many requests that include only voice. As such, a received voice input can include stand-alone utterances, such that a navigation device, vehicular system, mobile phone, or other device can be controlled in one step using voice. For example, using existing navigation systems or other navigation interfaces, a user must often take multiple steps to configure a displayed map, even for relatively simple tasks (e.g., multiple button and/or stylus inputs may be required to display a desired portion of a map). By contrast, using the natural language voice user interface, the user can simplify controlling a device using one-step voice control (e.g., a map may be controlled using utterances of "Move the map up," "Move the map Northeast," "Show me downtown Seattle," or other utterances that provide all information necessary to control a device in one-step), which may substantially reduce a number of device interactions needed to control the device.

[076] In various instances, (e.g., whether the input received in operation 410 includes an utterance, a non-voice component, or various combinations thereof), one or more queries and/or commands contained in the voice input may be identified in operation 440. Once the one or more preliminary interpretations of the multi-modal input have been generated, appropriate contextual information, shared knowledge, external knowledge sources, or other information may be used to identify the queries and/or the commands contained in the utterance. The information sources relied upon may be associated with various devices, systems, or other components of the integrated voice user interface environment. Thus, information associated with a navigation device (e.g., maps, points-of-interest, etc.) may be distributed for sharing with other devices in the environment (e.g., a personal digital assistant capable of rendering maps may display a map associated with a navigation device, or a mobile phone may dial a telephone number of a point-of-interest associated with the navigation device, etc.). Moreover, the information may be shared in a bi-directional manner, such that information sources associated various devices can be shared with other devices in the integrated environment.

[077] Information may be associated persistently (e.g., by storing all information on a server, an online data backup system, or other data repositories), or may be built dynamically as a

function of processing multi-modal voice inputs. For example, a user may maintain an address book of contacts, e-mail addresses, street addresses, telephone numbers, or other information using a commercial service provider (e.g., GMail, Yahoo Mail, or another service provider capable of storing address books). When a user makes a request (e.g., via voice input, non-voice input, or multi-modal input) in which the service provider manages information needed to resolve the request, the information may be pulled and shared with a device that will be used to resolve the request. For example, a mobile phone may be integrated with the navigation voice user interface (e.g., a mobile phone having a GPS receiver), and the mobile phone may initially (e.g., at a first-time use) have no local contacts, addresses, telephone numbers, or other information. Thus, a user utterance of "I'm running late for dinner at Jane's house, call her to let her know I'm on the way" may result in a query being formulated to the service provider in order to retrieve an address and telephone number associated with Jane's house for the mobile phone to process accordingly. Moreover, information relating to being "late" may be processed to generate a route that avoids traffic, uses highways, or otherwise reduces an amount of travel time.

[078] A variety of cognitive models, contextual models, user-specific models, or other models may be used to identify the query and/or command in operation 440. For example, a given input may include information relating to one or more contextual domains, one or more of which may be invoked to interpret and/or infer certain keywords, concepts, or other information contained in the input. Moreover, short-term and long-term shared knowledge about a user's behavior and preferences may be used in a hybrid recognition model, for example, relating to navigation queries and/or commands. The hybrid recognition model may be used to generate probable or likely interpretations of an utterance using a combination of semantic analysis and contextual inferences. For example, certain syllables, words, phrases, requests, queries, commands, or other identifiable aspects of an utterance may be identified as being more likely to occur based on a given contextual history. The contextual history can be based on the short-term or the long-term shared knowledge about the user, previous utterances in a current conversation, common requests in a given environment (e.g., a navigation device may be used most often for displaying maps, calculating routes, and identifying locations), or other information. Thus, the hybrid recognition model for operation 440 may include various processes for analyzing semantic

patterns to resolve what was said by an utterance, in addition to various processes for relying on contextual history to resolve what was meant by the utterance.

[079] In addition, the hybrid recognition model may be used in conjunction with, or independently of, a peer to peer recognition model. For example, contextual histories may include various preferences or user characteristics, in addition to providing a basis for inferring information about a user based on patterns of usage or behavior, among others. As a result, the recognition models may include additional awareness relating to global usage patterns, preferences, or other characteristics of peer users on a global basis. For example, certain keywords, concepts, queries, commands, or other aspects of a contextual framework may be commonly employed by all users within a context. In another example, users of certain demographics may share common jargon, slang, or other semantic speech patterns. As a result, operation 440 may utilize various recognition models, which consider context and semantics in various dimensions, to identify queries or command. For example, in addition to information generally available within a given environment or context, a voice input may be recognized using context and semantics based on short-term or long-term behavior or preferences for a specific user, global users, peer users, or other meaningful user abstractions.

[080] Accordingly, upon identifying a suitable query and/or command contained in the utterance, a suitable response may be generated, or a task may be performed in an operation 450. As will be apparent from the descriptions already provided, the generated response and/or the performed task may be highly dependent on context and information contained in a given voice input. As a result, when certain voice inputs may be unclear, ambiguous, or otherwise unclear, the generated response may prompt a user for an additional input, and the response may be framed in a way favorable for recognition (e.g., an utterance of "Where is Joe's Pizza" provided while a user is in New York City may return too many results to be meaningful, such that results may be weighed and displayed for a user to select a correct one of the results).

[081] Furthermore, voice inputs may be used to perform compound requests, which could otherwise be impossible to perform using a single manual input. For example, a single voice-based input may include a compound map control request, such as "Show me downtown Seattle." As such, operation 450 may perform tasks of retrieving a map of Seattle and automatically zooming in on a downtown area of the Seattle map. Further, one or more responses or other outputs may be generated, such as suggesting a route to a frequent destination,

identifying possible points-of-interest to the user, or searching for traffic or event notifications, among other things. Many other variations will be apparent, including characteristics of the received inputs (e.g., utterances, non-voice inputs, etc.), the requested queries or commands, the generated responses, and the performed tasks, among other things. As such, it will be apparent that the method 400 illustrated herein may enable users to request many different navigation related tasks verbally, non-verbally, or various combinations thereof, such that the users can request various kinds of information or tasks available in the environment (e.g., relating to various devices, information sources, etc.). Thus, method 400 may operate in an environment in which the natural language voice user interface has been associated with one or more devices that provide navigation services.

[082] According to various aspects of the invention, Figure 5 illustrates a flow diagram of an exemplary method for calculating navigation routes from voice destination entries provided to a navigation device having a conversational, natural language voice user interface. As previously described, the voice user interface may calculate routes, provide dynamic data-driven directions to a destination, provide dynamic routing to a destination, perform post-processing of full or partial destination entries, or otherwise provide various voice navigation services. Thus, as illustrated in Figure 5, the method for calculating routes may support successive refinement of a voice destination entry, wherein context, agent adaptation, and shared knowledge, among other things, can help a user to narrow down a final destination using voice commands, multi-modal commands, or various combinations thereof. However, though Figure 5 specifically illustrates an exemplary implementation of successive refinement for voice destination entry, it will be apparent that the techniques described herein can be applied in performing various tasks in which generalized approximations can suitably be refined through successive voice or multi-modal commands to narrow down information sought by a user in various domains, contexts, applications, devices, or other components that employ the techniques described herein. As a result, successive refinement can be implemented in various domains that enable a user to "drill down" to a specific piece of information or data through specifying more and more specific criteria about the information sought.

[083] As illustrated in Figure 5, for example, successive refinement of a final destination may includes instances in which a user specifies the destination in a single utterance, over a plurality of utterances (e.g., which sequentially narrow down approximations of the destination), as well

as instances in which a system identifies the destination (e.g., through generating inferences, using context or shared knowledge, etc.). For example, a given approximation of a final destination may be associated with a finite number of possible destinations, and the system may be able to unambiguously identify the final destination from among the finite possible destinations by generating inferences, by relying on context, shared knowledge, and other information, or by other ways, as will be apparent. Moreover, in another example, successively refining the destination may be modeled after patterns of human interaction between passengers in a taxicab, or other similar situations in which a route or a destination may be narrowed down or otherwise refined over a course of interaction. For example, passengers in taxicabs sometimes specify a general approximation of a destination (e.g., a passenger picked up in Arlington, Virginia may state, "I'm going to the Foggy Bottom area of D.C."), which may result in a cab driver beginning to drive along a preferred route to the approximated destination (e.g., the driver may know to head towards Rock Creek Parkway when driving from Arlington to Foggy Bottom). While en route to the approximated destination, the passenger and/or the driver may cooperate to refine the final destination through one or more subsequent interactions (e.g., the cab driver may ask, "Where in Foggy Bottom are you headed?", and the passenger may specify "McFadden's," which may provide sufficient information for the driver to determine a route to the destination).

[084] However, it will be apparent that no two sequences of interactions will necessarily be alike. For example, the passenger may provide additional information relating to the final destination without being prompted by the driver, or the driver may not begin driving until the final destination has been identified (e.g., when several routes may be available, and the final destination has an impact on which route will be selected). As such, cooperative, conversational interactions may have a general goal of calculating a route to a final destination, yet the interactions may occur in many different ways, where a given interaction will likely be highly dependent on context, shared knowledge, agent adaptation, points of origin and/or presence, user preferences, dynamic data (e.g., relating to traffic, events, external systems, etc.), among many other factors.

[085] Accordingly, as further described herein, the method illustrated in Figure 5 may enable voice destination entry through successive refinement using the aforementioned factors, among other things, to inform route calculation based on full or partial destination inputs. For instance,

in similar fashion as described in the above cab driver example, the voice user interface (e.g., the cab driver) may have knowledge that certain routes may be preferable to reach a certain destination from a current point of presence (e.g., Arlington to Foggy Bottom via Rock Creek Parkway). Further, as will be described in greater detail in Figure 6, a search space associated with subsequent interactions may be constrained according to a current route, context, or other knowledge (e.g., locations named "McFadden's" that fall outside a certain proximity of Foggy Bottom may be excluded).

[086] Thus, in an operation 510, a user may provide a full or partial destination input using free form natural language, for example, including voice commands and/or multi-modal commands (e.g., an input may include an utterance of "I'm going here," coupled with a touched point on a display). The full or partial destination may be specified in various ways, including by specific address, place name, person's name, business name, neighborhood, city, region, or various other ways (e.g., a voice destination entry may be provided in an exploratory manner, such as when a user wants to visit a museum, but has yet to decide which one to visit). Furthermore, because successive refinement tolerates incomplete or partial addresses, voice destination inputs may be provided in a way that uses the successive refinement in reverse, for example, by narrowly specifying a desired destination in order to calculate a route to a broader vicinity of the desired destination (e.g., a route to South Philadelphia could be identified through an utterance of "Pat's Steaks in Philly," even though Pat's Steaks may not be a user's actual destination).

[087] The voice destination input may be parsed or otherwise analyzed using one or more dynamically adaptable recognition grammars, for example, as described above in reference to Figure 3. For example, recognition grammars may be loaded, generated, extended, pruned, or otherwise adapted based on various factors, including a proximity to a user's point of presence (e.g., as the user moves from one area to another, the recognition grammar may be optimized based on a current location, a direction of travel, temporal constraints, etc.), a contextual history (e.g., as the user interacts with the voice user interface, the grammar may adapt based on dictionaries, keywords, concepts, or other information associated with other contexts, domains, devices, applications, etc.), or other factors, as will be apparent. As such, an operation 520 may include generating one or more interpretations of the voice destination input, which may be analyzed using various data sources in order to generate an N-best list of possible destinations (e.g., a navigation agent may query a directory, a voice search engine, or other components to

identify one or more destinations that at least partially match criteria contained in the voice destination input).

[088] The generated list of possible destinations may be post-processed in an operation 530 in order to assign weights or ranks to one or more of the entries in the N-best list. The post-processing may include analyzing the destination list generated in operation 520 according to various factors in order to determine a most likely intended destination from a full or partial voice destination input. For example, post-processing operation 530 may rank or weigh possible destinations according to shared knowledge about the user, domain-specific knowledge, dialogue history, or other factors. Furthermore, the post-processing operation 530 may analyze the full or partial destination input in order to identify an address to which a route can be calculated, for example, by resolving a closest address that makes "sense" relative to the input destination. For example, a user may specify a partial destination that identifies a broad and approximated area (e.g., "Take me to Massachusetts"), and depending on a user's current location, direction of travel, preferences, or other information, post-processing operation 530 may select an address makes sense for calculating a route (e.g., an address in Cape Cod may be selected for a user having relatives that live in Cape Cod, whereas an address in Boston may be selected for a user who may be traveling to various popular sightseeing areas, etc.).

[089] As a result, the weighed list of N-best destinations may be evaluated in an operation 540 to determine a suitably identifiable destination exists in the list. For example, a full or partial voice destination entry may be ambiguous, or certain criteria or keywords in a voice destination entry may be unrecognizable, such that a highest ranked destination in the weighted list does not exceed a minimal confidence level needed to identify a destination. For instance, a user located in Oklahoma may utter a partial destination, as in "I'm heading to Washington," and the decisional operation 540 may return a negative indication when the voice user interface cannot disambiguate between Washington state, Washington, D.C., Washington University in Saint Louis, and a town of Washington located slightly south of Oklahoma City. In another example, a user originating in Saint Louis may provide a voice destination entry of "Take me to Springfield," which could result in an unidentifiable destination even though multiple destinations may satisfy the minimal confidence level. In this example, "Springfield" may be unidentifiable as a destination because Springfield, Illinois and Springfield, Missouri both reside within reasonable distances of Saint Louis, yet directions of travel to either destination may be

entirely opposite (e.g., a route to the Illinois Springfield includes traveling north on Interstate-55, whereas a route to the Missouri Springfield includes traveling southwest on Interstate-44). Thus, to avoid routing the user in a direction opposite from an intended destination, processing may instead branch to an operation 550 to generate a prompt for resolving the destination.

[090] For example, the prompt may ask the user a question or provide information that can clarify an intended meaning of the received voice destination input. For instance, in the Washington example given above, a multi-modal user prompt may display the four possible destinations on a display device (i.e., Washington state, Washington, D.C., Washington University, and Washington, Oklahoma). The multi-modal prompt may further include system-generated speech, for example, stating, "I found several possible Washingtons, did you mean one of these or something else?" As a result, processing then proceeds back to operation 510, in which the user can disambiguate the intended destination through another input (e.g., by touching one of the displayed destinations when the display device includes a touch-screen, or by verbalizing one of the destinations, or by providing additional information indicating that none of the displayed destinations were intended). Thereafter, subsequent N-best lists may be more likely to result in an identifiable destination, as an original voice destination input can be interpreted in view of a dialogue history.

[091] Alternatively, when at least one of the possible destinations can be identified unambiguously, while meeting the minimal confidence level, a positive indication may result in processing proceeding to another decisional operation 560, which controls how a route will be calculated to the identified destination (e.g., a highest ranking entry in the weighted N-best list of destinations). The decisional operation 560 may determine whether the identified destination provides a full or otherwise complete destination. For example, through successive refinement, a partial voice destination entry may result in an identifiable, yet incomplete, destination. As such, an operation 570 may calculate a preliminary route to the identified partial destination. Thereafter, subsequent processing returns to operation 510, where the voice user interface may await an additional voice destination input that refines the partial destination. Further, in various implementations, one or more system prompts may be generated to request the additional voice destination inputs (e.g., as the user approaches the partial destination and additional information will be needed to provide further routing, or as the user approaches a point in the preliminary

route where distinct routes may be taken to different topographical subtiles or points within the partial destination, etc.).

[092] As such, the processing operations 510-570 may be iteratively repeated until the final destination can be identified through successive refinement using one or more multi-modal voice destination entries. As a result, when a suitable final destination has been identified, operation 560 may be followed by an operation 580 that completes the route to the final destination. For example, the route may be completed by extrapolating the route to the partial destination into a complete route to the final destination. Further, it will be apparent that voice destination entries may be successively refined into a final destination en route (e.g., the final destination may be successively refined as a user proceeds along a preliminary route), in advance (e.g., the user may choose to drive upon a complete route to a final destination being identified), or in other ways.

[093] Moreover, as described in greater detail herein, the calculated route may be dynamically adjusted or rerouted based on subsequent inputs, generated inferences, dynamic data, or various other sources of information (e.g., inferences may be generated, which may result in dynamic routing, in response to dynamic data relating to traffic conditions, detours, events, weather conditions, or user preferences, or other factors). Thus, Figure 6 illustrates a flow diagram of an exemplary method for providing voice services based on a current navigation route according to various aspects of the invention. The voice user interface may use the method illustrated in Figure 6 to provide the voice services in a dynamic, data-driven manner, wherein the voice services may be provided in a customized manner, for example, based on context, shared knowledge about a user, or other sources of information.

[094] Accordingly, providing voice navigation services based on a current route includes identifying the current navigation route in an operation 610. The route may be identified from a preliminary route or a complete route, as calculated according to the techniques described above in reference to Figure 5. Thus, current routing information may be made available to various system components in order to process events or voice inputs, generate responses, formulate queries, resolve commands, or perform other processing services available in a voice navigation services environment. As such, the current routing information may provide further context relevant to interpreting voice inputs, taking action towards resolving the inputs, or responding to events, among other things. Thus, though not illustrated in Figure 6, various other sources of contextual information may be utilized to provide voice navigation services in view of the

current route. For example, user preferences may limit places to which a vehicle can travel, or for how long the vehicle can travel, among other things (e.g., for keeping a delivery or passenger route on a regular route, or for placing parental controls on teenage drivers, etc.). In another example, the voice user interface may have awareness of one or more external systems or devices, sources of dynamic data, or other information, which can be utilized to determine how to best provide the voice navigation services.

[095] Thus, upon identifying the current route in operation 610, along with other appropriate information relating to context, shared knowledge, or otherwise, a decisional operation 620 may determine whether the identified information results in an event being generated and/or detected. When no event can be generated and/or detected, processing proceeds to a decisional operation 630, which determines whether a voice input has been received from the user. As such, the method illustrated herein may loop through decisional operations 620 and 630 until one or more of an event has been detected and/or a voice input has been received.

[096] One type of event that can be detected and/or generated in decisional operation 620 may relate to location dependent advertisements for navigation systems (e.g., as generated by a local or remote advertising engine, or via a data channel, or in other ways). For example, in various implementations, the advertising events may be generated and/or detected using techniques described in co-pending U.S. Patent Application 11/671,526, entitled "System and Method for Selecting and Presenting Advertisements Based on Natural Language Processing of Voice-Based Input," filed February 6, 2007, the disclosure of which is hereby incorporated by reference in its entirety. For instance, as will be apparent, a navigation system will necessarily include one or more systems or mechanisms for determining the navigation system's current location (e.g., a global positioning system, a radio frequency identification system, a system that determines location based on a distance to an identifiable wireless tower or access point, etc.). Further, the natural language voice user interface may be able to determine location in a number of other ways, in addition to or apart from the location detection system (e.g., an adaptable prompt may affirmatively request the location information from a user, for example, upon failure of the location detection system, or other techniques utilizing shared knowledge or context may be used, as will be apparent).

[097] Thus, in one example, operation 620 may include detecting and/or generating events relating to a location dependent advertisement. A marketer or other entity seeking to reach a

target audience may simply scatter radio frequency identifiers (RFIDs) along a roadside, or another place in which a system may detect the RFIDs. The marketer or other entity may then broadcast one or more advertisements via a data channel associated with the RFIDs, such that the navigation system may trigger an event when within a suitable proximity of the RFIDs. Thus, processing would proceed to an operation 570, in which information associated with the event may be filtered according to the current routing information or other contextual parameters. As applied to this example, the filtering operation 650 may include monitoring the data channel to analyze information associated with the detected RFID, thus receiving the advertisements broadcasted over the data channel. Thus, an advertisement including speech, for example, may be analyzed in a similar manner to how user speech would be analyzed (e.g., maintaining awareness that the speech relates to an advertisement).

[098] Accordingly, the advertisement event data may be filtered in view of the current route, personal information (e.g., previous queries, preferences, history, etc.), or other factors to determine what action should be taken in an operation 660. For example, the advertisement may be associated with a local restaurant seeking to reach personal navigation devices that may not necessarily have a complete business database. Therefore, in one example, filtering operation 650 could discard the advertisement when shared knowledge indicates that the user prefers to reach destinations as fast as possible, without making unnecessary stops. In this case, operation 660 may effectively be a null operation, or may simply include displaying an image associated with the advertisement to allow the user to make an independent choice as to whether or not to stop. However, in a distinct example, the user may have provided a previous voice input indicating a desire to stop driving and eat lunch at some point along the route. Thus, in this example, operation 660 may include a spoken output that brings the restaurant to the user's attention, possibly asking the user whether the route should be temporarily diverted to stop for lunch at the restaurant.

[099] However, it should be appreciated that the above described RFID-based advertisement provides but one example of a possible advertising event that may be detected in operation 620. For instance, advertisements may be uploaded to a server by one or more advertising partners, wherein the uploaded advertisements may be associated with metadata or other descriptive information that identifies a target audience, location-dependent information, or other criteria. In another example, a plurality of advertisements may be stored locally at the voice user interface,

and an inferencing engine may determine appropriate circumstances in which an event should be generated to deliver one or more of the advertisements to a user. As a result, it will be apparent that advertising events may be generated in a number of ways, and may be generated and/or detected locally, remotely, by detecting RFIDs, or in other ways.

[0100] Further, as advertisements get delivered to users (e.g., in operation 660), the users' subsequent interaction with the delivered advertisements may be tracked. In this way, affinity based models may be generated, for example, to ensure that promotions or advertisements will be delivered to a likely target audience. Thus, an event relating to a given advertisement may be generated and/or detected in operation 620 when shared knowledge about a user's behavior, preferences, or other characteristics match one or more criteria associated with peer-to-peer affinities associated with the advertisement. For example, a marketer may be offering promotional coupons intended for a specific demographic audience, and the peer-to-peer affinities may be used to ensure that the coupons will only be provided to users likely to be in the intended audience (e.g., to avoid providing an entire market, including unlikely customers, with the coupons). Thus, when the user falls within the specified demographic audience, one or more events may be generated on a server, locally by an advertising engine, or in other ways to deliver one or more of the promotional coupons to the user.

[0101] In other examples, the advertising model may include mobile pay-per-use systems (e.g., network access may be provided through a wireless access point on a pay-per-use basis), peer-to-peer local guides or recommendations (e.g., affinities may indicate that various categories of peer users may be interested in similar points-of-interests, events, local restaurants, stores, activities, tourist attractions, or other aspects of a local area). Additionally, as described in greater detail above in reference to Figure 3, various aspects of the advertising model, such as the local guides and recommendations, may be generated according to a mapping applied to various topological domains. For example, local guides or recommendations may be dependent on topological characteristics (e.g., different guides for various cities, communities, localities, etc.). As such, the local guides, recommendations, or other aspects of the advertising model may be associated with a topological taxonomy based on geographical chunking. As a result, various advertising events may be generated and/or detected according to physical proximities, temporal proximities, directional proximities, civil organization proximities, or various combinations

thereof (e.g., peer-to-peer recommendations may be varied depending on a user's location, direction of travel, temporal constraints, etc.).

[0102] In addition to the advertising related events that can be generated and/or detected in operation 620, the natural language voice user interface may generate additional events through awareness of context, shared knowledge about a user, external systems and devices, or other information. For example, as discussed above in reference to Figure 2, the voice user interface may implemented within an agent-based architecture, in which one or more of a plurality of agents may include an inference engine. However, it will be apparent that the inference engine may be arranged within the voice user interface in various other ways. For example, the voice user interface may include the inference engine within a conversational language processor, where in such implementations, the inference engine may generate, detect, and distribute events to various components of the voice user interface. In another example, a managing inference engine may coordinate event generation and/or detection among the inference engines associated with the agents. As such, it will be apparent that the voice user interface may include various suitable arrangements of one or more inference engines, such that events and inferences can be detected, generated, and distributed to various system components as they may arise.

[0103] Thus, the one or more inference engines may utilize awareness of context, shared knowledge about a user, dynamic data sources, data and services associated with external or devices, among other information to generate and/or detect the events or other inferences identified in operation 620. When the events or other inferences occur when a current navigation route exists, the events or other inferences may be filtered in operation 650 prior to being a response being generated and/or a task being performed in operation 660.

[0104] For example, the inference engine may utilize personal information, such as a user frequently visiting a particular person in the user's address book, to generate an event that suggests rerouting the current route. Thus, in this example, the inference engine may generate the event (which may be detected in operation 620) when the current route or current location information reflects a certain proximity to an address associated with the person in the address book. Operation 650 could then filter the event by determining possible responses or courses of action for handling the event in a manner consistent with the current route. For example, information associated with the current route may indicate that the user was out running errands and may now be heading home. Further, a calendar in the user's address book may indicate that

the user has an appointment later in the day, but that no calendar entries exist for a few hours. Thus, an exemplary response generated in operation 660 may include a speech-based output of, "You are close to Jane's house, would you like to stop by?"

[0105] Although the above-provided example illustrates an event generated using personal information, it will be apparent that various available information sources may be utilized to generate and/or detect events. For example, events may be based on transient or dynamic data relating to communities, traffic, weather, and/or many other sources of data (e.g., a generated event may result in an output of "Tonight's soccer game is being played at a field near your house, do you want to go?", in which the soccer game may be a local community event).

[0106] In yet another example, various types of external system awareness may trigger events and resulting voice responses from the voice user interface. For example, the voice user interface may be coupled to various vehicular or telematics systems, providing awareness over a state of such systems. Thus, an event may be generated when the vehicular system indicates that the vehicle will soon run out of gas, resulting in a voice response prompting a user to stop for gas. Further, the external system awareness may be filtered according to the current routing information (e.g., in conjunction with the inference engine), wherein a calculation may be made indicating that a level of gas will be sufficient to get to a next gas station along the route, but that the level of gas will not be sufficient to get to a second gas station along the route. As a result, the route-based filtering operation 650 may result in a response being generated in operation 660 that provides a voice warning that the user must stop for gas at the next gas station.

[0107] Returning to operation 630, user-provided voice inputs may also be processed in view of current routing information. As previously discussed, interpretations of what was said in a voice input (e.g., content of an utterance) may be based on various phonetic models, dictionaries, context histories, dialogue histories, and other information that can form a dynamic recognition grammar. Moreover, context, shared knowledge, and available data and services, among other information, may be used to interpret what meant by the voice input (e.g., an intent of the utterance). As a result, using the various techniques described herein and in the incorporated co-pending U.S. Patent Applications, a query, command, or other request contained in the utterance may be identified in an operation 640. When a current route exists, a domain of possible tasks, commands, query answers, responses, and/or other system actions may be filtered according to the current route. For example, a voice input of "Where's the closest bathroom" may be

analyzed in view of the current route. Various proximities (e.g., a direction of travel, physical distances, temporal distances, etc.) may then be utilized to determine what bathrooms may be appropriate in view of the current route. Thus, when a route has the user traveling a long distance, for example, a bathroom at a rest area twenty miles ahead may be favored over a bathroom at a restaurant ten miles off a highway (e.g., to avoid diverting the user from the current route).

[0108] Furthermore, in many instances, an agent or other processing component may need to formulate a plurality of queries to various different data sources in order to resolve a request. As such, operation 660 for performing a task may determine what inferences, metadata, or other information may be needed to formulate queries to the data sources in order to resolve the request. As a result, voice requests relating to navigation may include information that may not necessarily be linked to navigation, or that may not necessarily be in a database or other source of navigation information. Even so, by being coupled with various available sources of information, complex or compound queries may be formulated to retrieve information from various data sources in order to resolve a given request. Furthermore, a given utterance may include multiple requests, queries, or commands, such that a given response or task performed in response to the utterance may include interactions with various different system components. Thus, in an illustrative example, an utterance of "Find the 7-11 closest to the Woodland Park Zoo" may include a first query to identify a location of Woodland Park Zoo, a second query to find one or more 7-11's within a certain radius of the Woodland Park Zoo location, a third query to determine a current location of the user, and a fourth query that invokes an inference engine to determine which 7-11 may be "closest" based on various user proximities. Thus, supposing that two 7-11's exist equidistantly from the Woodland Park Zoo, one in a direction east of the Zoo and one in a direction west of the Zoo, the user's current location may be used to determine which 7-11 to select; by contrast, a selection based purely on physical proximity to the Zoo could result in ambiguity.

[0109] Although the above-provided descriptions highlight various specific examples and illustrations of how the natural language voice user interface may provide voice navigation services, it will be understood that the descriptions, examples, and modes of operation provided above should be regarded as exemplary only. The natural language voice user interface may be highly tolerant of variations in how humans interact verbally, and may utilize context, shared

knowledge, and available data and systems to effectively interact with any given user in a conversational manner that advances mutual goals. As a result, interactions between the user and the voice user interface may be regarded as unique occurrences, as users may speak free form requests relating to various available data sources, applications, or devices in a given interaction. Thus, the voice user interface may dynamically adapt, learn, and otherwise refine an operational framework to provide optimal services in the voice navigation services environment.

[0110] Implementations of the invention may be made in hardware, firmware, software, or various combinations thereof. The invention may also be implemented as instructions stored on a machine-readable medium, which may be read and executed by one or more processors. A machine-readable medium may include a mechanism for storing or transmitting information in a form readable by a machine (e.g., a computing device). For example, a machine-readable storage medium may include read only memory, random access memory, magnetic disk storage media, optical storage media, flash memory devices, and others, and a machine-readable transmission media may include forms of propagated signals, such as carrier waves, infrared signals, digital signals, and others. Further, firmware, software, routines, or instructions may be described in the above disclosure in terms of specific exemplary aspects and implementations of the invention, and performing certain actions. However, those skilled in the art will recognize that such descriptions are merely for convenience and that such actions in fact result from computing devices, processors, controllers, or other devices executing the firmware, software, routines, or instructions.

[0111] Aspects and implementations may be described as including a particular feature, structure, or characteristic, but every aspect or implementation may not necessarily include the particular feature, structure, or characteristic. Further, when a particular feature, structure, or characteristic is described in connection with an aspect or implementation, it will be understood that one skilled in the art may be able to effect such feature, structure, or characteristic in connection with other aspects or implementations, whether or not explicitly described. Thus, various changes and modifications may be made to the provided description without departing from the scope or spirit of the invention. As such, the specification and drawings should be regarded as exemplary only, and the scope of the invention to be determined solely by the appended claims.

CLAIMS

What is claimed is:

1. A computer-implemented method for providing a natural language voice user interface for a navigation device, comprising:

receiving a navigation-related voice-based input at an input mechanism associated with the navigation device, the voice-based input including at least a natural language utterance provided by a user;

generating one or more preliminary interpretations of the natural language utterance using a dynamic recognition grammar associated with a speech recognition engine;

analyzing the preliminary interpretations using a conversational language processor, the conversational language processor using shared knowledge and information associated with a navigation context to determine a probable interpretation of the natural language utterance in the navigation context; and

communicating the probable interpretation of the natural language utterance to a navigation agent associated with the navigation context, the navigation agent operable to:

identify one or more navigation-related requests contained in the voice-based input based on the probable interpretation of the natural language utterance; and

resolve the requests using information associated with one or more of a plurality of information sources, the plurality of information sources including at least a navigation-specific information source.

2. The method of claim 1, the navigation agent further operable to determine that at least one of the requests includes an approximation of requested information, wherein the conversational language processor manages a dialogue with the user to resolve the approximated information, the managed dialogue including one or more interactions with the user that successively refine the approximated information.

3. The method of claim 2, each of the interactions including one or more of prompting the user to refine the approximated information or receiving a subsequent input from the user that refines the approximated information.

4. The method of claim 1, the navigation-related requests including a request for a route to a full or partial address, the navigation agent operable to resolve the request for the route by:

selecting a destination for the route, the selected destination having an address that best corresponds to the full or partial address;

calculating a route from a current location of the user to the selected destination; and

generating directions to guide the user on the calculated route to the selected destination, the generated directions dynamically driven by data from a plurality of information sources.

5. The method of claim 4, the dynamic data-driven directions including information about destinations, points of interest, traffic, parking, weather, events, or other information relevant to the calculated route.

6. The method of claim 4, further comprising:

receiving a subsequent voice-based input at the input mechanism associated with the navigation device, the subsequent voice-based input including at least one request provided by the user; and

invoking at least one of a plurality of domain agents to resolve the request included in the subsequent voice-based input, the at least one domain agent operable to filter results of the request according to the calculated route.

7. The method of claim 1, the dynamic recognition grammar including information associated with one or more topological domains.

8. The method of claim 7, the topological domains including physical, temporal, directional, and civil organizational proximities with respect to a current location of the user.

9. The method of claim 1, the conversational language processor coupled to an inferencing engine operable to generate one or more inferences, wherein the conversational language processor uses the generated inferences to determine the probable interpretation of the natural language utterance.

10. The method of claim 9, wherein the conversational language processor further uses the generated inferences to suggest one or more services available in the navigation context to the user.

11. The method of claim 1, the navigation-related requests including a multi-modal request to control a map display, the navigation agent operable to resolve the request to control the map display by:

associating a non-voice component of the voice-based input with the probable interpretation of the utterance, the non-voice component identifying a portion of the map display; and

issuing a command to control the identified portion of the map display in accordance with the probable interpretation of the utterance.

12. The method of claim 1, wherein the plurality of information sources include the navigation-specific information source, the shared knowledge, the information associated with the navigation context, and one or more information sources relating to maps, destinations, directories, points of interest, traffic, parking, weather, events, user address books, user devices and systems, a search engine, and a plurality of domain agents.

13. The method of claim 12, the shared knowledge including dialogue history information, request history information, user interface state information, short-term user profile information, long-term user profile information, peer user profile information, and user location information.

14. A system for providing a natural language voice user interface for a navigation device, comprising:

an input mechanism that receives a navigation-related voice-based input, the voice-based input including at least a natural language utterance provided by a user;

a speech recognition engine having a dynamic recognition grammar associated therewith, the speech recognition engine operable to generate one or more preliminary interpretations of the natural language utterance using the dynamic recognition grammar;

a plurality of domain agents associated with respective contexts, the plurality of domain agents including at least a navigation agent associated with a navigation context; and

a conversational language processor operable to:

analyze the preliminary interpretations using shared knowledge and information associated with the navigation context to determine a probable interpretation of the natural language utterance in the navigation context; and

communicate the probable interpretation of the natural language utterance to the navigation agent associated with the navigation context, the navigation agent operable to identify one or more navigation-related requests contained in the voice-based input based on the probable interpretation of the natural language utterance and resolve the requests using information associated with one or more of a plurality of information sources, the plurality of information sources including at least a navigation-specific information source.

15. The system of claim 14, the navigation agent further operable to determine that at least one of the requests includes an approximation of requested information; and

the conversational language processor further operable to manage a dialogue with the user to resolve the approximated information, the managed dialogue including one or more interactions with the user that successively refine the approximated information.

16. The system of claim 15, each of the interactions including one or more of prompting the user to refine the approximated information or receiving a subsequent input from the user that refines the approximated information.

17. The system of claim 14, the navigation-related requests including a request for a route to a full or partial address, the navigation agent operable to resolve the request for the route by:

selecting a destination for the route, the selected destination having an address that best corresponds to the full or partial address;

calculating a route from a current location of the user to the selected destination; and

generating directions to guide the user on the calculated route to the selected destination, the generated directions dynamically driven by data from a plurality of information sources.

18. The system of claim 17, the dynamic data-driven directions including information about destinations, points of interest, traffic, parking, weather, events, or other information relevant to the calculated route.

19. The system of claim 17, the one or more input mechanisms further receiving a subsequent voice-based input, the subsequent voice-based input including at least one request provided by the user, and the conversational language processor further operable to:

invoking at least one of a plurality of domain agents to resolve the request included in the subsequent voice-based input, the at least one domain agent operable to filter results of the request according to the calculated route.

20. The system of claim 14, the dynamic recognition grammar including information associated with one or more topological domains.

21. The system of claim 20, the topological domains including physical, temporal, directional, and civil organizational proximities with respect to a current location of the user.

22. The system of claim 14, further comprising an inferencing engine coupled to the conversational language processor that generates one or more inferences, wherein the conversational language processor uses the generated inferences to determine the probable interpretation of the natural language utterance.

23. The system of claim 22, further comprising an inferencing engine coupled to the conversational language processor, wherein the conversational language processor further uses the generated inferences to suggest one or more services available in the navigation context to the user.

24. The system of claim 14, the navigation-related requests including a multi-modal request to control a map display, the navigation agent operable to resolve the request to control the map display by:

associating a non-voice component of the voice-based input with the probable interpretation of the utterance, the non-voice component identifying a portion of the map display; and

issuing a command to control the identified portion of the map display in accordance with the intended meaning of the utterance.

25. The system of claim 14, wherein the plurality of information sources include the navigation-specific information source, the shared knowledge, the information associated with the navigation context, and one or more information sources relating to maps, destinations, directories, points of interest, traffic, parking, weather, events, user address books, user devices and systems, a search engine, and a plurality of domain agents.

26. The system of claim 25, the shared knowledge including dialogue history information, request history information, user interface state information, short-term user profile information, long-term user profile information, peer user profile information, and user location information.

27. A method for providing a natural language voice user interface in a voice navigation services environment, comprising:

receiving a voice-based destination input provided by a user, the voice-based destination input including at least a natural language utterance;

generating one or more preliminary interpretations of the natural language utterance using a dynamic recognition grammar associated with a speech recognition engine;

analyzing the preliminary interpretations using a conversational language processor, the conversational language processor using shared knowledge and information associated with a navigation context to determine an intended destination provided in the natural language utterance, the intended destination including an approximation of a destination; and

communicating the intended destination to a navigation agent, the navigation agent operable to provide a route to the intended destination by:

selecting a preliminary destination for the route, the selected preliminary destination having an address that best corresponds to the approximated destination;

calculating a route from a current location of the user to the selected preliminary destination; and

participating in a dialogue with the user to resolve a final destination for the calculated route, the dialogue includes one or more interactions with the user that successively refine the approximated destination until the final destination has been resolved.

28. The method of claim 27, each of the interactions including one or more of prompting the user to refine the approximated destination or receiving a subsequent input from the user that refines the approximated information.

29. The method of claim 27, the navigation agent further operable to dynamically recalculate the route in response to the successive refinements of the approximated destination.

30. The method of claim 27, the final destination being resolvable substantially later in time with respect to an initial calculation of the route.

31. The method of claim 27, the navigation agent operable to select the preliminary destination by identifying one or more addresses possibly corresponding to the approximated destination and selecting a highest ranked one of the identified addresses as the preliminary destination.

32. The method of claim 27, the identified addresses ranked according to a proximity to the current location of the user or the approximated destination.

33. The method of claim 32, the proximity to the based on one or more topological domains, including one or more of physical, temporal, directional, and civil organizational proximities with respect to the current location of the user or the approximated destination.

34. The method of claim 27, the navigation agent further operable to generate directions to guide the user on the calculated route, the generated directions dynamically driven by data from a plurality of information sources.

35. The method of claim 34, the plurality of information sources including information relating to destinations, points of interest, traffic, parking, weather, events, or other information relevant to the calculated route.

36. The method of claim 27, the shared knowledge including dialogue history information, request history information, user interface state information, short-term user profile information, long-term user profile information, peer user profile information, and user location information.

37. A method for providing a natural language voice user interface in a voice navigation services environment, comprising:

receiving a navigation-related voice-based input provided by a user, the voice-based input including at least a natural language utterance;

identifying a current location of a user using a location detection system;

determining one or more topological domains associated with the current location of the user;

generating a dynamic recognition grammar that includes grammar information associated with the determined topological domains; and

using the dynamic recognition grammar to generate one or more interpretations of the natural language utterance.

38. The method of claim 37, the dynamic recognition grammar organizing the grammar information according to geographic chunks associated with the topological domains.

39. The method of claim 38, the geographic chunks including physical proximities defined by a distance from the current location of the user.

40. The method of claim 38, the geographic chunks including temporal proximities defined by a time of travel from the current location of the user.

41. The method of claim 38, the geographic chunks including directional proximities defined by a directional vector of travel by the user.

42. The method of claim 38, the geographic chunks including civil organizational proximities defined by continents, countries, regions, states, cities, localities, neighborhoods, and communities.

43. The method of claim 38, further comprising dividing one or more of the geographic chunks into a plurality of tiles, the dynamic recognition grammar further organizing the grammar information according to the plurality of tiles.

44. The method of claim 43, further comprising dividing one or more of the plurality of tiles into a plurality of subtiles, the dynamic recognition grammar further organizing the grammar information according to the plurality of subtiles.

45. The method of claim 37, the dynamic recognition grammar having a size constrained according to memory or resource availability in the voice navigation services environment.

46. The method of claim 37, further comprising modifying the grammar information included in the dynamic recognition grammar in response to changes in the topological domains associated with the current location of the user.

47. The method of claim 37, further comprising determining affinities between the user and one or more peers of the user, the dynamic recognition grammar further organizing the grammar information according to the determined affinities.

48. A method for providing advertisements in a voice navigation services environment, comprising:

identifying a current location of a user using a location detection system;

retrieving shared knowledge and information associated with a navigation context, the retrieved information used to determine probable interpretations of natural language utterances received in the voice navigation services environment;

identifying one or more advertisements to provide to the user, the one or more advertisements identified based on an affinity to one or more of the current location of the user, the shared knowledge, or the information associated with the navigation context; and

generating a multi-modal output to provide the one or more identified advertisements to the user.

49. The method of claim 48, the multi-modal output including a system-generated speech utterance.

50. The method of claim 48, the location detection system receiving information relating to the one or more advertisements over a data channel.

51. The method of claim 50, the data channel associated with a radio frequency identifier detected by the location detection system.

52. The method of claim 48, the shared knowledge including dialogue history information, request history information, user interface state information, short-term user profile information, long-term user profile information, peer user profile information, and user location information.

53. The method of claim 48, further comprising:

receiving a voice-based input that includes at least a natural language utterance;

using the shared knowledge and the information associated with the navigation context to determine a probable interpretation of the received natural language utterance, the advertisements identified based on the probable interpretation of the natural language utterance.

54. The method of claim 48, the advertisement including a guide to a community local to the current location of the user.

55. The method of claim 54, the local community guide including recommendations for one or more points of interest, events, restaurants, stores, activities, or tourist attractions.

56. The method of claim 54, the local community guide including information associated with one or more of maps, destinations, directories, traffic, parking, or weather.

57. The method of claim 54, the local community guide based on affinities between the user and one or more of the current location of the user, one or more peers of the user, or one or more communities associated with the user.

58. The method of claim 48, the advertisement including one or more of an image, a banner, an audio message, a video message, a promotional offer, a coupon, and a data stream.

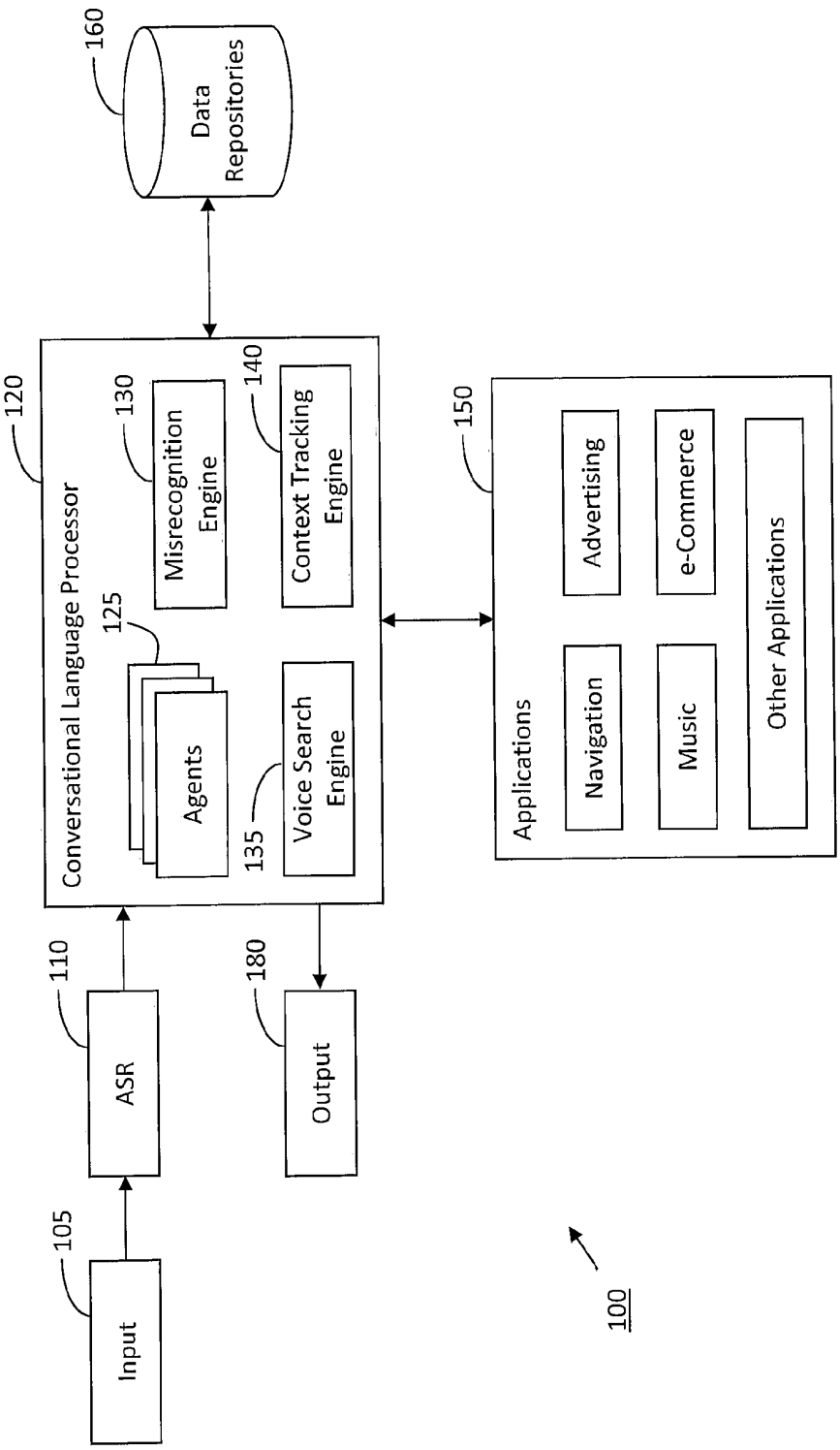

160
Data Repositories
120
Conversational Language Processor
130
Misrecognition Engine
140
Context Tracking Engine
125
Agents
135
Voice Search Engine
150
Applications
Advertising
e-Commerce
Navigation
Music
Other Applications
110
ASR
180
Output
105
Input
100

Figure 1

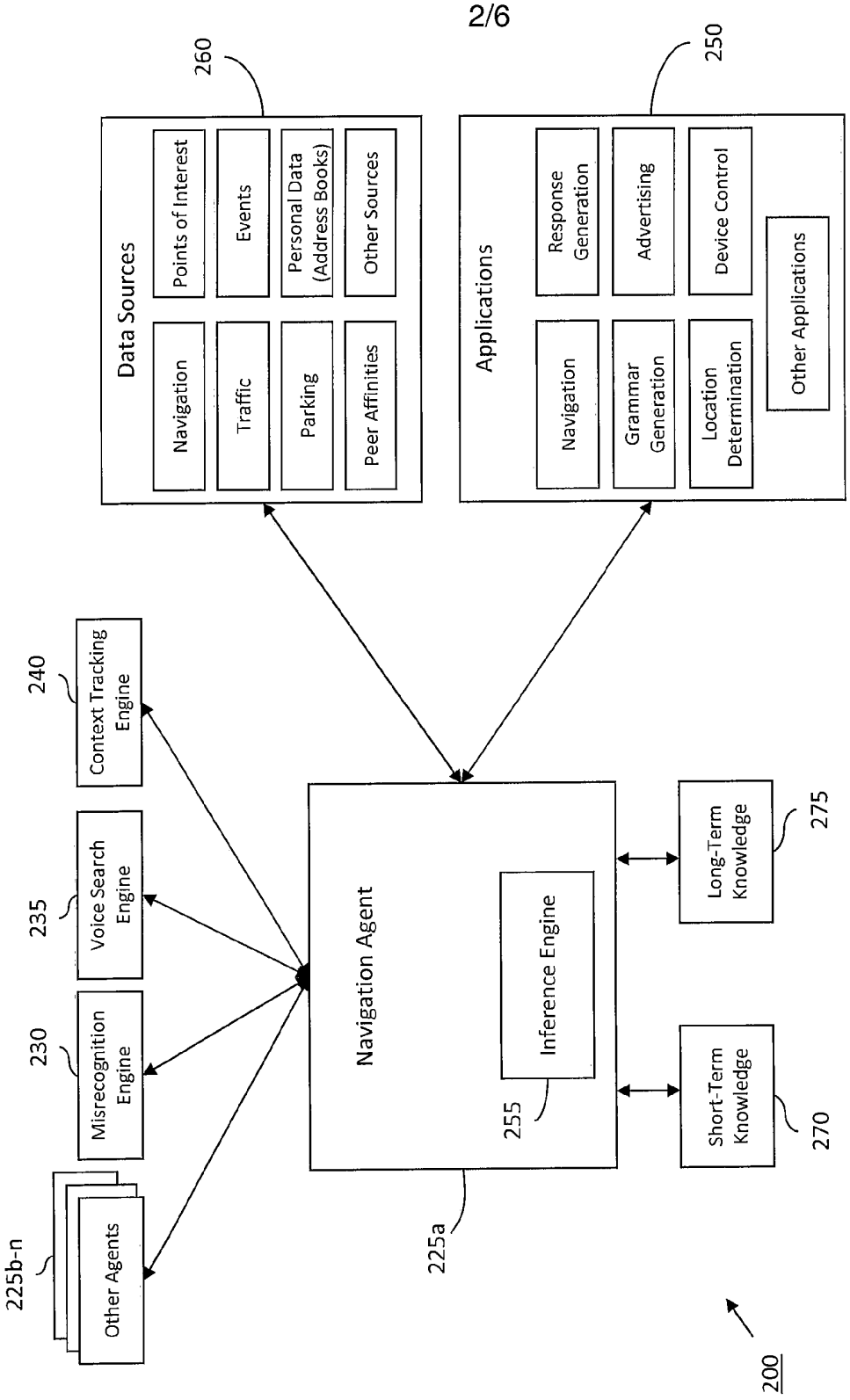

260
Data Sources
Points of Interest
Events
Personal Data (Address Books)
Other Sources
Navigation
Traffic
Parking
Peer Affinities
250
Applications
Response Generation
Advertising
Device Control
Navigation
Grammar Generation
Location Determination
Other Applications
240
Context Tracking Engine
235
Voice Search Engine
230
Misrecognition Engine
225b-n
Other Agents
Navigation Agent
Inference Engine
255
225a
Long-Term Knowledge
275
Short-Term Knowledge
270
200

Figure 2

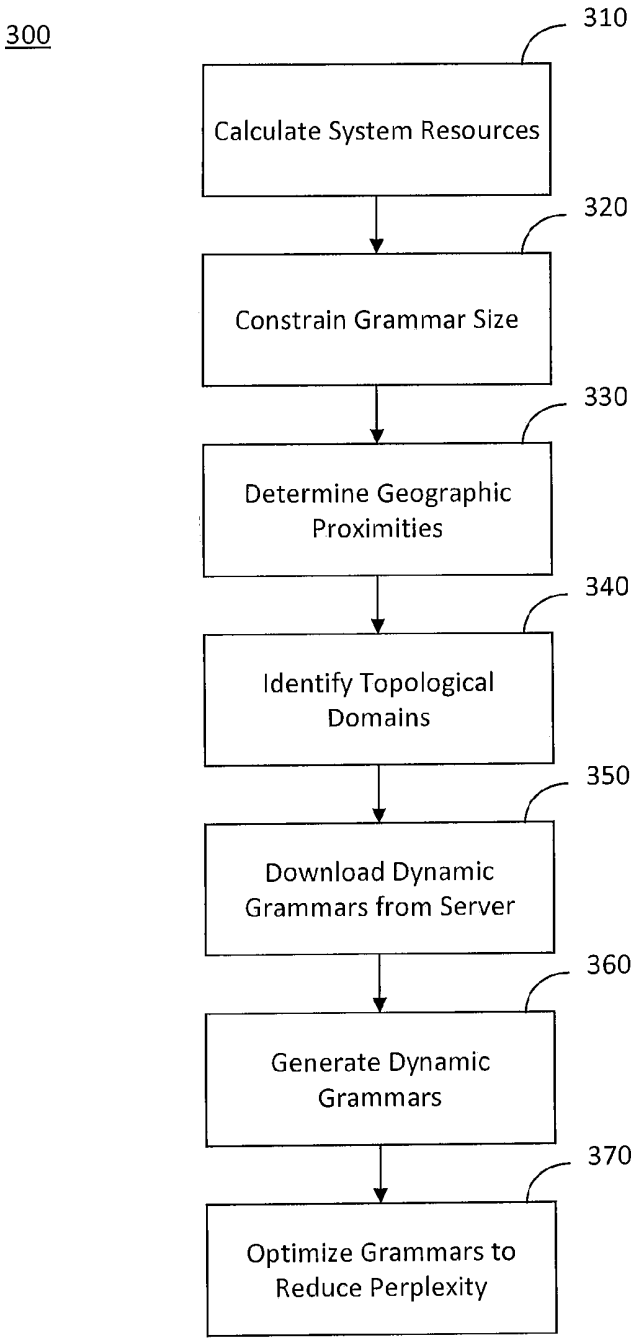

300
310
Calculate System Resources
320
Constrain Grammar Size
330
Determine Geographic Proximities
340
Identify Topological Domains
350
Download Dynamic Grammars from Server
360
Generate Dynamic Grammars
370
Optimize Grammars to Reduce Perplexity

Figure 3

400

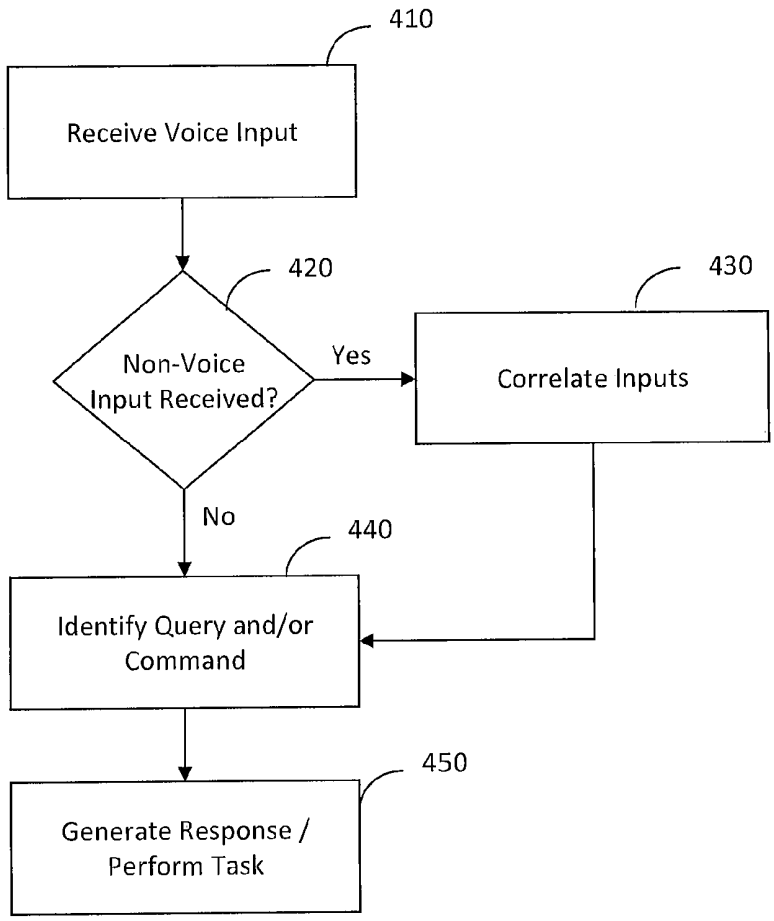

410
Receive Voice Input
420
Non-Voice Input Received?
Yes
430
Correlate Inputs
No
440
Identify Query and/or Command
450
Generate Response / Perform Task

Figure 4

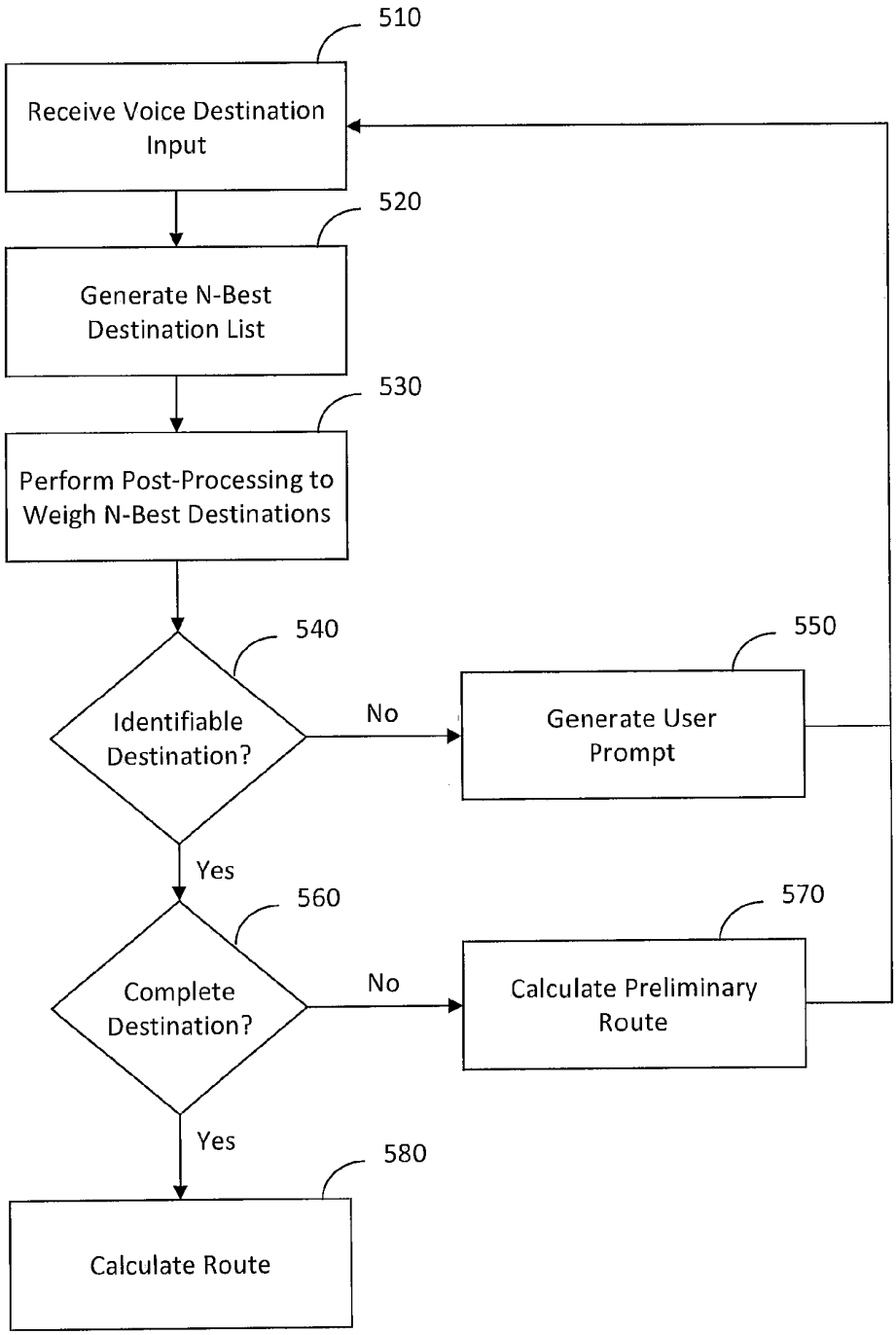

510
Receive Voice Destination Input
520
Generate N-Best Destination List
530
Perform Post-Processing to Weigh N-Best Destinations
540
Identifiable Destination?
No
550
Generate User Prompt
Yes
560
Complete Destination?
No
570
Calculate Preliminary Route
Yes
580
Calculate Route

Figure 5

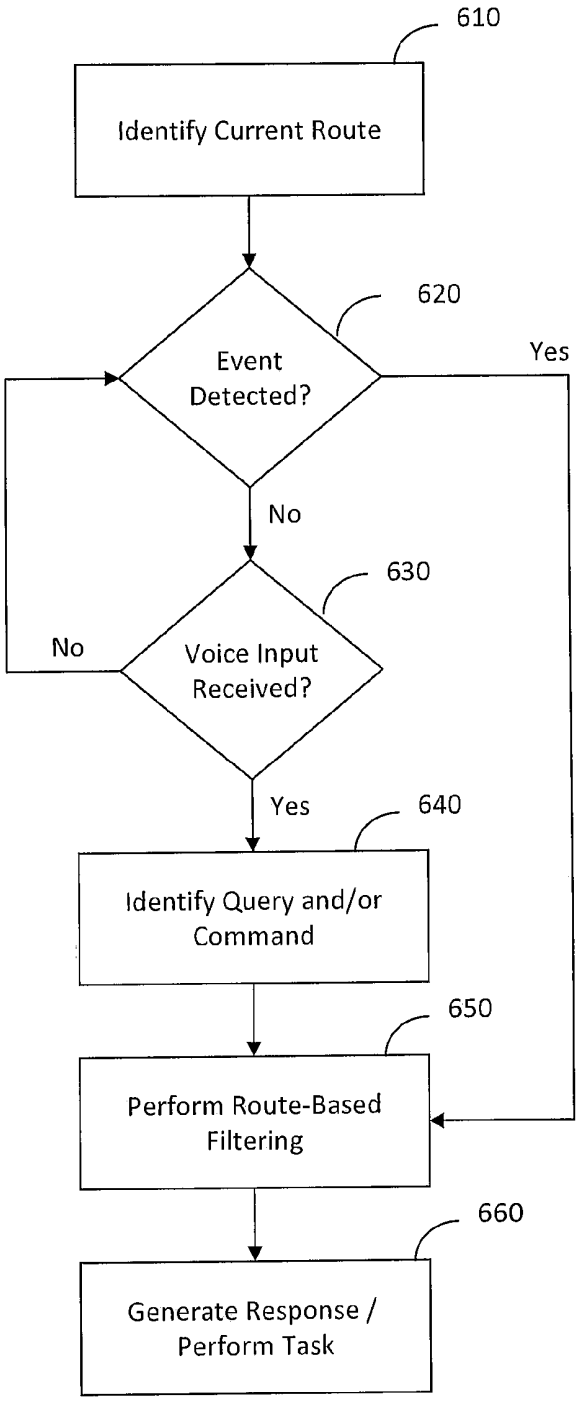

610
Identify Current Route
620
Event Detected?
Yes
No
630
Voice Input Received?
No
Yes
640
Identify Query and/or Command
650
Perform Route-Based Filtering
660
Generate Response / Perform Task

Figure 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 08/69508

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 17/27 (2008.04)
USPC - 704/9
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC - 704/9

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC - 704/1, 2, 8, 9, 200, 206, 221, E15.045

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PubWEST (PGPB, USPT, EPAB, JPAB); Google Scholar
Search Terms Used: map, gps, speech, voice, talk, navigation, natural, conversational, grammar, interpret, software, computer, input, interface, user interface, refine, zoom, route, directions

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X -- Y	US 2007/0050191 A1 (WEIDER et al.) 01 March 2007 (01.03.2007), entire document, especially paras. [0061]-[0073], [0100], [0121], [0123], [0126], [0133], [0140], [0142], [0147], [0153], [0167], [0189]-[0194], [0210]-[0211], [0215]	1-42, 45-58 ---------- 43-44
Y	US 5,902,347 A (BACKMAN et al.) 11 May 1999 (11.05.1999), entire document, especially col 1 ln 32-34, col 4 ln 33-44	43-44
A	US 5,274,560 A (LARUE) 28 December 1993 (28.12.1993), entire document	1-58
A	US 6,598,018 B1 (JUNQUA) 22 July 2003 (22.07.2003), entire document	1-58

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 September 2008 (25.09.2008)	02 OCT 2008
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (April 2007)

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
3 December 2009 (03.12.2009)

PCT

(10) International Publication Number
WO 2009/145796 A1

(51) International Patent Classification:
G06F 13/12 (2006.01)

(21) International Application Number:
PCT/US2008/069524

(22) International Filing Date:
9 July 2008 (09.07.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
12/127,343 27 May 2008 (27.05.2008) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [US/US]; 11980 NE 24th Street, Bellevue, Washington 98005 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **KENNEWICK, Robert** [US/US]; 13660 NE 42nd St., Bellevue, WA 98005 (US). **WEIDER, Chris** [US/US]; 10757 Greenwood Avenue N. #220, Seattle, Washington 98133 (US).

(74) Agents: **GATTO, James, G.** et al.; Pillsbury Winthrop Shaw Pittman LLP, P.O. Box 10500, McLean, Virginia 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) **Title**: SYSTEM AND METHOD FOR AN INTEGRATED, MULTI-MODAL, MULTI-DEVICE NATURAL LANGUAGE VOICE SERVICES ENVIRONMENT

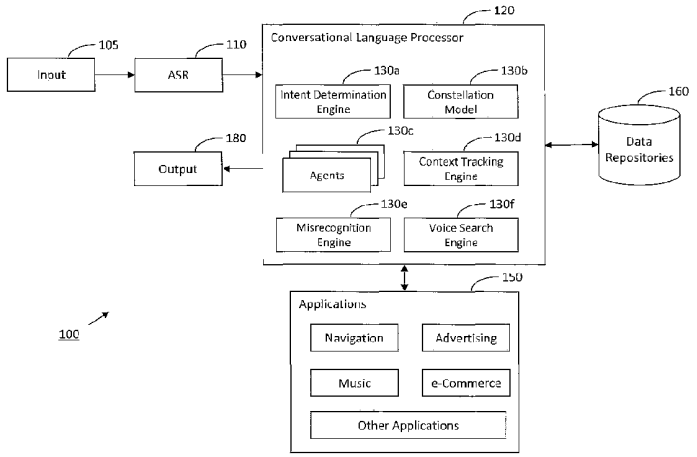


Figure 1

(57) **Abstract**: A system and method for an integrated, multi-modal, multi-device natural language voice services environment may be provided. In particular, the environment may include a plurality of voice-enabled devices each having intent determination capabilities for processing multi-modal natural language inputs in addition to knowledge of the intent determination capabilities of other devices in the environment. Further, the environment may be arranged in a centralized manner, a distributed peer-to-peer manner, or various combinations thereof. As such, the various devices may cooperate to determine intent of multi-modal natural language inputs, and commands, queries, or other requests may be routed to one or more of the devices best suited to take action in response thereto.

WO 2009/145796 A1

SYSTEM AND METHOD FOR AN INTEGRATED, MULTI-MODAL, MULTI-DEVICE NATURAL LANGUAGE VOICE SERVICES ENVIRONMENT

FIELD OF THE INVENTION

[001] The invention relates to an integrated voice services environment in which a plurality of devices can provide various voice services by cooperatively processing free form, multi-modal, natural language inputs, thereby facilitating conversational interactions between a user and one or more of the devices in the integrated environment.

BACKGROUND OF THE INVENTION

[002] As technology has progressed in recent years, consumer electronic devices have emerged to become nearly ubiquitous in the everyday lives of many people. To meet the increasing demand that has resulted from growth in the functionality and mobility of mobile phones, navigation devices, embedded devices, and other such devices, a wealth of features and functions are often provided therein in addition to core applications. Greater functionality also introduces the trade-offs, however, including learning curves that often inhibit users from fully exploiting all of the capabilities of their electronic devices. For example, many existing electronic devices include complex human to machine interfaces that may not be particularly user-friendly, which inhibits mass-market adoption for many technologies. Moreover, cumbersome interfaces often result in otherwise desirable features being buried (e.g., within menus that may be tedious to navigate), which has the tendency of causing many users to not use, or even know about, the potential capabilities of their devices.

[003] As such, the increased functionality provided by many electronic devices often tends to be wasted, as market research suggests that many users only use only a fraction of the features or applications available on a given device. Moreover, in a society where wireless networking and broadband access are increasingly prevalent, consumers tend to naturally desire seamless mobile capabilities from their electronic devices. Thus, as consumer demand intensifies for simpler mechanisms to interact with electronic devices, cumbersome interfaces that prevent quick and focused interaction can become an important concern. Accordingly, the ever-growing demand for mechanisms to use technology in intuitive ways remains largely unfulfilled.

[004] One approach towards simplifying human to machine interactions in electronic devices includes the use of voice recognition software, which can enable users to exploit features that

could otherwise be unfamiliar, unknown, or difficult to use. For example, a recent survey conducted by the Navteq Corporation, which provides data used in a variety of applications such as automotive navigation and web-based applications, demonstrates that voice recognition often ranks among the features most desired by consumers of electronic devices. Even so, existing voice user interfaces, when they actually work, still tend to require significant learning on the part of the user.

[005] For example, many existing voice user interface only support requests formulated according to specific command-and-control sequences or syntaxes. Furthermore, many existing voice user interfaces cause user frustration or dissatisfaction because of inaccurate speech recognition. Similarly, by forcing a user to provide pre-established commands or keywords to communicate requests in ways that a system can understand, existing voice user interfaces do not effectively engage the user in a productive, cooperative dialogue to resolve requests and advance a conversation towards a mutually satisfactory goal (e.g., when users may be uncertain of particular needs, available information, or device capabilities, among other things). As such, existing voice user interfaces tend to suffer from various drawbacks, including significant limitations on engaging users in a dialogue in a cooperative and conversational manner.

[006] Additionally, many existing voice user interfaces fall short in utilizing information distributed across various different domains or devices in order to resolve natural language voice-based inputs. Thus, existing voice user interfaces suffer from being constrained to a finite set of applications for which they have been designed, or to devices on which they reside. Although technological advancement has resulted in users often having several devices to suit their various needs, existing voice user interfaces do not adequately free users from device constraints. For example, users may be interested in services associated with different applications and devices, but existing voice user interfaces tend to restrict users from accessing the applications and devices as they see fit. Moreover, users typically can only practicably carry a finite number of devices at any given time, yet content or services associated with users' other devices that currently being used may be desired in various circumstances. Accordingly, although users tend to have varying needs, where content or services associated with different devices may be desired in various contexts or environments, existing voice technologies tend to fall short in providing an integrated environment in which users can request content or services associated with virtually any device or network. As such, constraints on information availability

and device interaction mechanisms in existing voice services environments tend to prevent users from experiencing technology in an intuitive, natural, and efficient way.

[007] Existing systems suffer from these and other problems.

SUMMARY OF THE INVENTION

[008] According to various aspects of the invention, a system and method for an integrated, multi-modal, multi-device natural language voice services environment may include a plurality of voice-enabled devices each having intent determination capabilities for processing multi-modal natural language inputs in addition to knowledge of the intent determination capabilities of other devices in the environment. Further, the environment may be arranged in a centralized manner, a distributed peer-to-peer manner, or various combinations thereof. As such, the various devices may cooperate to determine intent of multi-modal natural language inputs, and commands, queries, or other requests may be routed to one or more of the devices best suited to take action in response thereto.

[009] According to various aspects of the invention, the integrated natural language voice services environment arranged in the centralized manner includes an input device that receives a multi-modal natural language input, a central device communicatively coupled to the input device, and one or more secondary devices communicatively coupled to the central device. Each of the input device, the central device, and the one or more secondary devices may have intent determination capabilities for processing multi-modal natural language inputs. As such, an intent of a given multi-modal natural language input may be determined in the centralized manner by communicating the multi-modal natural language input from the input device to the central device. Thereafter, the central device may aggregate the intent determination capabilities of the input device and the one or more secondary devices and determine an intent of the multi-modal natural language input using the aggregated intent determination capabilities. The input device may then receive the determined intent from the central device and invoke at least one action at one or more of the input device, the central device, or the secondary devices based on the determined intent.

[010] According to various aspects of the invention, the integrated natural language voice services environment arranged in the distributed manner includes an input device that receives a multi-modal natural language input, a central device communicatively coupled to the input

device and one or more secondary devices communicatively coupled to the input device, wherein each of the input device and the one or more secondary devices may have intent determination capabilities for processing multi-modal natural language inputs, as in the centralized implementation. However, the distributed implementation may be distinct from the centralized implementation in that a preliminary intent of the multi-modal natural language input may be determined at the input device using local intent determination capabilities. The multi-modal natural language input may then be communicated to one or more of the secondary devices (e.g., when a confidence level of the intent determination at the input device falls below a given threshold). In such cases, each of the secondary devices determine an intent of the multi-modal natural language input using local intent determination capabilities. The input device collates the preliminary intent determination and the intent determinations of the secondary devices, and may arbitrate among the collated intent determinations to determine an actionable intent of the multi-modal natural input.

[011] According to various aspects of the invention, the integrated natural language voice services environment arranged in a manner that dynamically selects between a centralized model and a distributed model. For example, the environment includes an input device that receives a multi-modal natural language input one or more secondary devices communicatively coupled to the input device, each of which have intent determination capabilities for processing multi-modal natural language inputs. A constellation model may be accessible to each of the input device and the one or more secondary devices, wherein the constellation model describes the intent determination capabilities of the input device and the one or more secondary devices. The multi-modal natural language input can be routed for processing at one or more of the input device or the secondary devices to determine an intent thereof based on the intent determination capabilities described in the constellation model. For example, when the constellation model arranges the input device and the secondary devices in the centralized manner, one of the secondary devices may be designated the central device and the natural language input may be processed as described above. However, when the multi-modal natural language cannot be communicated to the central device, the constellation model may be dynamically rearranged in the distributed manner, whereby the input device and the secondary devices share knowledge relating to respective local intent determination capabilities and operate as cooperative nodes to

determine the intent of the multi-modal natural language input using the shared knowledge relating to local intent determination capabilities.

[012] Other objects and advantages of the invention will be apparent based on the following drawings and detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

[013] Figure 1 illustrates a block diagram of an exemplary multi-modal electronic device that may be provided in an integrated, multi-device natural language voice services environment, according to various aspects of the invention.

[014] Figure 2 illustrates a block diagram of an exemplary centralized implementation of the integrated, multi-modal, multi-device natural language voice service environment, according to various aspects of the invention.

[015] Figure 3 illustrates a flow diagram of an exemplary method for processing multi-modal, natural language inputs at an input device in the centralized implementation of the integrated, multi-modal, multi-device natural language voice service environment, according to various aspects of the invention.

[016] Figure 4 illustrates a flow diagram of an exemplary method for processing multi-modal, natural language inputs at a central device in the centralized implementation of the integrated, multi-modal, multi-device natural language voice service environment, according to various aspects of the invention.

[017] Figure 5 illustrates a flow diagram of an exemplary method for processing multi-modal, natural language inputs at a secondary device in the centralized implementation of the integrated, multi-modal, multi-device natural language voice service environment, according to various aspects of the invention.

[018] Figure 6 illustrates a block diagram of an exemplary distributed implementation of the integrated, multi-modal, multi-device natural language voice service environment, according to various aspects of the invention.

[019] Figure 7 illustrates a flow diagram of an exemplary method for processing multi-modal, natural language inputs at an input device in the distributed implementation of the integrated, multi-modal, multi-device natural language voice service environment, according to various aspects of the invention.

DETAILED DESCRIPTION

[020] According to various aspects of the invention, Figure 1 illustrates a block diagram of an exemplary multi-modal elcctronic device 100 that may be provided in a natural language voice services environment that includes one or more additional multi-modal devices (e.g., as illustrated in Figures 2 and 6). As will be apparent, the electronic device 100 illustrated in Figure 1 may be any suitable voicc-cnabled electronic device (e.g., a telematics device, a personal navigation device, a mobile phone, a VoIP node, a personal computer, a media device, an embedded device, a server, or another elcctronic device). The device 100 may include various components that collectively provide a capability to process conversational, multi-modal natural language inputs. As such, a user of the devicc 100 may engage in multi-modal conversational dialogues with the voice-enabled electronic dcvice 100 to resolve requests in a free form, cooperative manner.

[021] For example, the natural language processing components may support free form natural language utterances to liberate the user from restrictions relating to how commands, queries, or other requests should be formulated. Rather, the user may employ any manner of speaking that feels natural in order to request content or services available through the device 100 (e.g., content or services relating to telematics, communications, media, messaging, navigation, marketing, information rctrieval, etc.). For instance, in various implementations, the device 100 may process natural language utterances utilizing techniques described in co-pending U.S. Patent Application 10/452,147, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed June 3, 2003, and U.S. Patent Application 10/618,633, entitled "Mobilc Systems and Methods for Responding to Natural Language Speech Utterance," filed June 15, 2003, the disclosures of which are hereby incorporated by reference in their entirety.

[022] Moreover, because the dcvice 100 may be deployed in an integrated multi-device environment, the user may further request content or services available through other devices deployed in the environment. In particular, the integrated voice services environment may include a plurality of multi-modal devices, each of which include natural language components generally similar to those illustrated in Figure 1. Thc various devices in the environment may serve distinct purposes, however, such that available content, services, applications, or other capabilities may vary among the devices in the environment (e.g., core functions of a media

device may vary from those of a personal navigation device). Thus, each device in the environment, including device 100, may have knowledge of content, services, applications, intent determination capabilities, and other features available through the other devices by way of a constellation model 130b. Accordingly, as will be described in greater detail below, the electronic device 100 may cooperate with other devices in the integrated environment to resolve requests by sharing context, prior information, domain knowledge, short-term knowledge, long-term knowledge, and cognitive models, among other things.

[023] According to various aspects of the invention, the electronic device 100 may include an input mechanism 105 that can receive multi-modal natural language inputs, which include at least an utterance spoken by the user. As will be apparent, the input mechanism 105 may include any appropriate device or combination of devices capable of receiving a spoken input (e.g., a directional microphone, an array of microphones, or any other device that can generate encoded speech). Further, in various implementations, the input mechanism 105 can be configured to maximize fidelity of encoded speech, for example, by maximizing gain in a direction of the user, cancelling echoes, nulling point noise sources, performing variable rate sampling, or filtering environmental noise (e.g., background conversations). As such, the input mechanism 105 may generate encoded speech in a manner that can tolerate noise or other factors that could otherwise interfere with accurate interpretation of the utterance.

[024] Furthermore, in various implementations, the input mechanism 105 may include various other input modalities (i.e., the input mechanism 105 may be arranged in a multi-modal environment), in that non-voice inputs can be correlated and/or processed in connection with one or more previous, contemporaneous, or subsequent multi-modal natural language inputs. For example, the input mechanism 105 may be coupled to a touch-screen interface, a stylus and tablet interface, a keypad or keyboard, or any other suitable input mechanism, as will be apparent. As a result, an amount of information potentially available when processing the multi-modal inputs may be maximized, as the user can clarify utterances or otherwise provide additional information in a given multi-modal natural language input using various input modalities. For instance, in an exemplary illustration, the user could touch a stylus or other pointing device to a portion of a touch-screen interface of the device 100, while also providing an utterance relating to the touched portion of the interface (e.g., "Show me restaurants around here"). In this example, the natural language utterance may be correlated with the input received

via the touch-screen interface, resulting in "around here" being interpreted in relation to the touched portion of the interface (e.g., as opposed to the user's current location or some other meaning).

[025] According to various aspects of the invention, the device 100 may include an Automatic Speech Recognizer 110 that generates one or more preliminary interpretations of the encoded speech, which may be received from the input mechanism 105. For example, the Automatic Speech Recognizer 110 may recognize syllables, words, or phrases contained in an utterance using one or more dynamically adaptable recognition grammars. The dynamic recognition grammars may be used to recognize a stream of phonemes through phonetic dictation based on one or more acoustic models. Furthermore, as described in co-pending U.S. Patent Application 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed August 5, 2005, the disclosure of which is hereby incorporated by reference in its entirety, the Automatic Speech Recognizer 110 may be capable of multi-pass analysis, where a primary speech recognition engine may generate a primary interpretation of an utterance (e.g., using a large list dictation grammar) and request secondary transcription from one or more secondary speech recognition engines (e.g., using a virtual dictation grammar having decoy words for out-of-vocabulary words).

[026] Thus, the Automatic Speech Recognizer 110 may generate preliminary interpretations of an utterance in various ways, including exclusive use of a dictation grammar or virtual dictation grammar, or use of various combinations thereof (e.g., when the device 100 supports multi-pass analysis). In any event, the Automatic Speech Recognizer 110 may provide out-of-vocabulary capabilities and may tolerate portions of a speech signal being dropped, the user misspeaking, or other variables that may occur in natural language speech (e.g., stops and starts, stutters, etc.). Furthermore, the recognition grammars employed by the Automatic Speech Recognizer 110 may include vocabularies, dictionaries, syllables, words, phrases, or other information optimized according to various contextual or application-specific domains (e.g., navigation, music, movies, weather, shopping, news, languages, temporal or geographic proximities, or other suitable domains). Moreover, environmental knowledge (e.g., peer-to-peer affinities, capabilities of devices in the environment, etc.), historical knowledge (e.g., frequent requests, prior context, etc.), or other types of knowledge can be used to continually optimize the information contained in the recognition grammars on a dynamic basis.

[027] For example, information contained in the recognition grammars may be dynamically optimized to improve a likelihood of a given utterance being recognized accurately (e.g., following an incorrect interpretation of a word, the incorrect interpretation may be removed from the grammar to reduce a likelihood of the incorrect interpretation being repeated). Accordingly, thc Automatic Speech Recognizer 110 may use a number of techniques to generate preliminary interprctations of natural language utterances, including those described, for example, in co-pending U.S. Patent Application 11/513,269, entitled "Dynamic Speech Sharpening," filed August 31, 2006, thc disclosure of which is hereby incorporated by reference in its entirety. Furthermore, the techniques used by thc Automatic Speech Recognizer 110 associated with the device 100 may be considered in dcfining intent determination capabilities of the device 100, and such capabilities may be shared with other dcvices in the environment to enable convergence of speech recognition throughout the environmcnt (e.g., because various devices may employ distinct speech recognition techniques or have distinct grammars or vocabularies, the devices may share vocabulary translation mechanisms to enhance system-wide recognition).

[028] According to various aspects of the invention, the Automatic Speech Recognizer 110 may provide one or more preliminary interpretations of a multi-modal input, including an utterance contained therein, to a conversational language processor 120. The conversational language processor 120 may include various components that collectively operate to model everyday human-to-human conversations in order to engage in cooperative conversations with the user to resolve requests based on the user's intent. For example, the conversational language processor 120 may include, among other things, an intent determination engine 130a, a constellation model 130b, one or more domain agents 130c, a context tracking engine 130d, a misrecognition engine 130e, and a voice search engine 130f. Furthermore, the conversational language processor 120 may be coupled to one or more data repositories 160 and applications associated with one or more domains. Thus, the intent determination capabilities of the device 100 may bc defined based on the data and processing capabilities of the Automatic Speech Recognizer 110 and the conversational language processor 120.

[029] More particularly, the intent determination engine 130a may establish meaning for a given multi-modal natural language input basod on a consideration of the intent determination capabilities of the device 100 as wcll as the intent determination capabilities of other devices in the integrated voice services environment. For example, the intent determination capabilities of

the device 100 may be defined as a function of processing resources, storage for grammars, context, agents, or other data, and content or services associated with the device 100 (e.g., a media device with a small amount of memory may have a smaller list of recognizable songs than a device with a large amount of memory). Thus, the intent determination engine 130a may determine whether to process a given input locally (e.g., when the device 100 has intent determination capabilities that suggest favorable conditions for recognition), or whether to route information associated with the input to other devices, which may assist in determining the intent of the input.

[030] As such, to determine which device or combination of devices should process an input, the intent determination engine 130a may evaluate the constellation model 130b, which provides a model of the intent determination capabilities for each of the devices in the integrated voice services environment. For instance, the constellation model 130b may contain, among other things, knowledge of processing and storage resources available to each of the devices in the environment, as well as the nature and scope of domain agents, context, content, services, and other information available to each of the devices in the environment. As such, using the constellation model 130b, the intent determination engine 130a may be able to determine whether any of the other devices have intent determination capabilities that can be invoked to augment or otherwise enhance the intent determination capabilities of the device 100 (e.g., by routing information associated with a multi-modal natural language input to the device or devices that appear best suited to analyze the information and therefore determine an intent of the input). Accordingly, the intent determination engine 130a may establish the meaning of a given utterance by utilizing the comprehensive constellation model 130b that describes capabilities within the device 100 and across the integrated environment. The intent determination engine 130a may therefore optimize processing of a given natural language input based on capabilities throughout the environment (e.g., utterances may be processed locally to the device 100, routed to a specific device based on information in the constellation model 130b, or flooded to all of the devices in the environment in which case an arbitration may occur to select a best guess at an intent determination).

[031] Although the following discussion will generally focus on various techniques that can be used to determine the intent of multi-modal natural language inputs in the integrated multi-device environment, it will be apparent that the natural language processing capabilities of any

one of the devices may extend beyond the specific discussion that has been provided herein. As such, in addition to the co-pending U.S. Patent Applications referenced above, further natural language processing capabilities that may be employed include those described in co-pending U.S. Patent Application 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed August 5, 2005, U.S. Patent Application 11/200,164, entitled "System and Method of Supporting Adaptive Misrecognition in Conversational Speech," filed August 10, 2005, U.S. Patent Application 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," filed August 29, 2005, U.S. Patent Application 11/580,926, entitled "System and Method for a Cooperative Conversational Voice User Interface," filed October 16, 2006, U.S. Patent Application 11/671,526, entitled "System and Method for Selecting and Presenting Advertisements Based on Natural Language Processing of Voice-Based Input," filed February 6, 2007, and U.S. Patent Application 11/954,064, entitled "System and Method for Providing a Natural Language Voice User Interface in an Integrated Voice Navigation Services Environment," filed December 11, 2007, the disclosures of which are hereby incorporated by reference in their entirety.

[032] According to various aspects of the invention, Figure 2 illustrates a block diagram of an exemplary centralized implementation of the integrated, multi-modal, multi-device natural language voice service environment. As will be apparent from the further description to be provided herein, the centralized implementation of the integrated, multi-device voice services environment may enable a user to engage in conversational, multi-modal natural language interactions with any one of voice-enabled devices 210a-n or central voice-enabled device 220. As such, the multi-device voice services environment may collectively determine intent for any given multi-modal natural language input, whereby the user may request content or voice services relating to any device or application in the environment, without restraint.

[033] As illustrated in Figure 2, the centralized implementation of the multi-device voice service environment may include a plurality of voice-enabled devices 210a-n, each of which include various components capable of determining intent of natural language utterances, as described above in reference to Figure 1. Furthermore, as will be apparent, the centralized implementation includes a central device 220, which contains information relating to intent determination capabilities for each of the other voice-enabled devices 210a-n. For example, in various exemplary implementations, the central device 220 may be designated as such by virtue

of being a device most capable of determining the intent of an utterance (e.g., a server, home data center, or other device having significant processing power, memory resources, and communication capabilities making the device suitable to manage intent determination across the environment). In another exemplary implementation, the central device 220 may be dynamically selected based on one or more characteristics of a given multi-modal natural language input, dialogue, or interaction (e.g., a device may be designated as the central device 220 when a current utterance relates to a specific domain).

[034] In the centralized implementation illustrated in Figure 2, a multi-modal natural language input may be received at one of the voice-enabled devices 210a-n. Therefore, the receiving one of the devices 210a-n may be designated as an input device for that input, while the remaining devices 210a-n may be designated as secondary devices for that input. In other words, for any given multi-modal natural language input, the multi-device environment may include an input device that collects the input, a central device 220 that aggregates intent determination, inferencing, and processing capabilities for all of the devices 210a-n in the environment, and one or more secondary devices that may also be used in the intent determination process. As such, each device 210 in the environment may be provided with a constellation model that identifies all of the devices 210 having incoming and outgoing communication capabilities, thus indicating an extent to which other devices may be capable of determining intent for a given multi-modal natural language input. The constellation model may further define a location of the central device 220, which aggregates context, vocabularies, content, recognition grammars, misrecognitions, shared knowledge, intent determination capabilities, inferencing capabilities, and other information from the various devices 210a-n in the environment.

[035] Accordingly, as communication and processing capabilities permit, the central device 220 may be used as a recognizer of first or last resort. For example, because the central device 220 converges intent determination capabilities across the environment (e.g., by aggregating context, vocabularies, device capabilities, and other information from the devices 210a-n in the environment), inputs may be automatically routed to the central device 220 when used as a recognizer of first resort, or as a recognizer of last resort when local processing at the input device 210 cannot determine the intent of the input with a satisfactory level of confidence. However, it will also be apparent that in certain instances the input device 210 may be unable to make contact with the central device 220 for various reasons (e.g., a network connection may be

unavailable, or a processing bottleneck at the central device 220 may cause communication delays). In such cases, the input device 210 that has initiated contact with the central device 220 may shift into decentralized processing (e.g., as described in reference to Figure 6) and communicate capabilities with one or more of the other devices 210a-n in the constellation model. Thus, when the central device 220 cannot be invoked for various reasons, the remaining devices 210a-n may operate as cooperative nodes to determine intent in a decentralized manner.

[036] Additionally, in the multi-device voice services environment, the central device 220 and the various other devices 210a-n may cooperate to create a converged model of capabilities throughout the environment. For example, as indicated above, in addition to having intent determination capabilities based on processing resources, memory resources, and device capabilities, each of the devices 210a-n and the central device 220 may include various other natural language processing components. The voice services environment may therefore operate in an integrated manner by maintaining not only a complete model of data, content, and services associated with the various devices 210a-n, but also of other natural language processing capabilities and dynamic states associated with the various devices 210a-n. As such, the various devices 210a-n may operate with a goal of converging capabilities, data, states, and other information across the device, either on one device (e.g., the central device 220) or distributed among the various devices 210a-n (e.g., as in the decentralized implementation to be described in Figure 6).

[037] For example, as discussed above, each device 210 includes an Automatic Speech Recognizer, one or more dynamically adaptable recognition grammars, and vocabulary lists used to generate phonemic interpretations of natural language utterances. Moreover, each device 210 includes locally established context, which can range from information contained in a context stack, context and namespace variables, vocabulary translation mechanisms, short-term shared knowledge relating to a current dialogue or conversational interaction, long-term shared knowledge relating to a user's learned preferences over time, or other contextual information. Furthermore, each device 210 may have various services or applications associated therewith, and may perform various aspects of natural language processing locally. Thus, additional information to be converged throughout the environment may include partial or preliminary utterance recognitions, misrecognitions or ambiguous recognitions, inferencing capabilities, and

overall device state information (e.g., songs playing in the environment, alarms set in the environment, etc.).

[038] Thus, various data synchronization and referential integrity algorithms may be employed in concert by the various devices 210a-n and the central device 220 to provide a consistent worldview of the environment. For example, information may be described and transmitted throughout the environment for synchronization and convergence purposes using the Universal Plug and Play protocol designed for computer ancillary devices, although the environment can also operate in a peer-to-peer disconnected mode (e.g., when the central device 220 cannot be reached). However, in various implementations, the environment may also operate in a peer-to-peer mode regardless of the disconnected status, as illustrated in Figure 6, for example, when the devices 210a-n have sufficient commensurate resources and capabilities for natural language processing.

[039] In general, the algorithms for convergence in the environment can be executed at various intervals, although it may be desirable to limit data transmission so as to avoid processing bottlenecks. For example, because the convergence and synchronization techniques relate to natural language processing, in which any given utterance will typically be expressed over a course of several seconds, information relating to context and vocabulary need not be updated on a time frame of less than a few seconds. However, as communication capabilities permit, context and vocabulary could be updated more frequently to provide real-time recognition or the appearance of real-time recognition. In another implementation, the convergence and synchronization may be permitted to run until completion (e.g., when no requests are currently pending), or the convergence and synchronization may be suspended or terminated when a predetermined time or resource consumption limit has been reached (e.g., when the convergence relates to a pending request, an intent determination having a highest confidence level at the time of cut-off may be used).

[040] By establishing a consistent view of capabilities, data, states, and other information throughout the environment, an input device 210 may cooperate with the central device 220 and one or more secondary devices (i.e., one or more of devices 210a-n, other than the input device) in processing any given multi-modal natural language input. Furthermore, by providing each device 210 and the central device 220 with a constellation model that describes a synchronized state of the environment, the environment may be tolerant of failure by one or more of the

devices 210a-n, or of the central device 220. For example, if the input device 210 cannot communicate with the central device 220 (e.g., because of a server crash), the input device 210 may enter a disconnected peer-to-peer mode, whereby capabilities can be exchanged with one or more devices 210a-n with which communications remain available. To that end, when a device 210 establishes new information relating to vocabulary, context, misrecognitions, agent adaptation, intent determination capabilities, inferencing capabilities, or otherwise, the device 210 may transmit the information to the central device 220 for convergence purposes, as discussed above, in addition to consulting the constellation model to determine whether the information should be transmitted to one or more of the other devices 210a-n.

[041] For example, suppose the environment includes a voice-enabled mobile phone that has nominal functionality relating to playing music or other media, and which further has a limited amount of local storage space, while the environment further includes a voice-enabled home media system that includes a mass storage medium that provides dedicated media functionality. If the mobile phone were to establish new vocabulary, context, or other information relating to a song (e.g., a user downloads the song or a ringtone to the mobile phone while on the road), the mobile phone may transmit the newly established information to the home media system in addition to the central device 220. As such, by having a model of all of the devices 210a-n in the environment and transmitting new information to the devices where it will most likely be useful, the various devices may handle disconnected modes of operation when the central device 220 may be unavailable for any reason, while resources may be allocated efficiently throughout the environment.

[042] Thus, based on the foregoing discussion, it will be apparent that a centralized implementation of an integrated multi-device voice services environment may generally include a central device 220 operable to aggregate or converge knowledge relating to content, services, capabilities, and other information associated with various voice-enabled devices 210a-n deployed within the environment. In such centralized implementations, the central device 220 may be invoked as a recognizer of first or last resort, as will be described in greater detail with reference to Figures 3-5, and furthermore, the other devices 210a-n in the environment may be configured to automatically enter a disconnected or peer-to-peer mode of operation when the central device 220 cannot be invoked for any reason (i.e., devices may enter a decentralized or distributed mode, as will be described in greater detail with reference to Figures 6-7).

Knowledge and capabilities of each of the devices 210a-n may therefore be made available throughout the voice services environment in a centralized manner, a distributed manner, or various combinations thereof, thus optimizing an amount of natural language processing resources used to determine an intent of any given multi-modal natural language input.

[043] According to various aspects of the invention, Figure 3 illustrates a flow diagram of an exemplary method for processing multi-modal, natural language inputs at an input device in the centralized implementation of the integrated, multi-modal, multi-device natural language voice service environment. Similarly, Figures 4 and 5 illustrate corresponding methods associated with a central device and one or more secondary devices, respectively, in the centralized voice service environment. Furthermore, it will be apparent that the processing techniques described in relation to Figures 3-5 may generally be based on the centralized implementation illustrated in Figure 2 and described above, whereby the input device may be assumed to be distinct from the central device, and the one or more secondary devices may be assumed to be distinct from the central device and the input device. However, it will be apparent that various instances may involve a natural language input being received at the central device or at another device, in which case the techniques described in Figures 3-5 may be vary depending on circumstances of the environment (e.g., decisions relating to routing utterances to a specific device or devices may be made locally, collaboratively, or in other ways depending on various factors, such as overall system state, communication capabilities, intent determination capabilities, or otherwise).

[044] As illustrated in Figure 3, a multi-modal natural language input may be received at an input device in an operation 310. The multi-modal input may include at least a natural language utterance provided by a user, and may further include other input modalities such as audio, text, button presses, gestures, or other non-voice inputs. It will also be apparent that prior to receiving the natural language input in operation 310, the input device may be configured to establish natural language processing capabilities. For example, establishing natural language processing capabilities may include, among other things, loading an Automatic Speech Recognizer and any associated recognition grammars, launching a conversational language processor to handle dialogues with the user, and installing one or more domain agents that provide functionality for respective application domains or contextual domains (e.g., navigation, music, movies, weather, information retrieval, device control, etc.).

[045] The input device may also be configured to coordinate synchronization of intent determination capabilities, shared knowledge, and other information with the central device and the secondary devices in the environment prior to receiving the input at operation 310. For example, when the input device installs a domain agent, the installed domain agent may bootstrap context variables, semantics, namespace variables, criteria values, and other context related to that agent from other devices in the system. Similarly, misrecognitions may be received from the central device and the secondary devices in order to enable correction of agents that use information relevant to the received misrecognitions, and vocabularies and associated translation mechanisms may be synchronized among the devices to account for potential variations between the Automatic Speech Recognizers used by the various devices (e.g., each device in the environment cannot be guaranteed to use the same Automatic Speech Recognizer or recognition grammars, necessitating vocabulary and translation mechanisms to be shared among the devices that share intent determination capabilities).

[046] Upon establishing and synchronizing natural language processing capabilities and subsequently receiving a multi-modal natural language input in operation 310, the input device may determine whether the environment has been set up to automatically transmit the input to the central device in a decisional operation 320. In such a case, processing proceeds to an operation 360 for transmitting the input to the central device, which may then process the input according to techniques to be described in relation to Figure 4. If the environment has not been set up to automatically communicate the input to the central device, however, processing proceeds to an operation 330, where the input device performs transcription of the natural language utterance contained in the multi-modal input. For example, the input device may transcribe the utterance using the Automatic Speech Recognizer and recognition grammars associated therewith according to techniques described above and in the above-referenced U.S. Patent Applications.

[047] Subsequently, in an operation 340, an intent of the multi-modal natural language input may be determined at the input device using local natural language processing capabilities and resources. For example, any non-voice input modalities included in the input may be merged with the utterance transcription and a conversational language processor associated with the input device may utilize local information relating to context, domain knowledge, shared knowledge, context variables, criteria values, or other information useful in natural language

processing. As such, the input device may attempt to determine a best guess as to an intent of the user that provided the input, such as identifying a conversation type (e.g., query, didactic, or exploratory) or request that may be contained in the input (e.g., a command or query relating to one or more domain agents or application domains).

[048] The intent determination of the input device may be assigned a confidence level (e.g., a device having an Automatic Speech Recognizer that implements multi-pass analysis may assign comparatively higher confidence levels to utterance transcriptions created thereby, which may result in a higher confidence level for the intent determination). The confidence level may be assigned based on various factors, as described in the above-referenced U.S. Patent Applications. As such, a decisional operation 350 may include determining whether the intent determination of the input device meets an acceptable level of confidence. When the intent determination meets the acceptable level confidence, processing may proceed directly to an operation 380 where action may be taken in response thereto. For example, when the intent determination indicates that the user has requested certain information, one or more queries may be formulated to retrieve the information from appropriate information sources, which may include one or more of the other devices. In another example, when the intent determination indicates that the user has requested a given command (e.g., to control a specific device), the command may be routed to the appropriate device for execution.

[049] Thus, in cases where the input device can determine the intent of a natural language input without assistance from the central device or the secondary devices, communications and processing resources may be conserved by taking immediate action as may be appropriate. On the other hand, when the intent determination of the input device does not meet the acceptable level of confidence, decisional operation 350 may result in the input device requesting assistance from the central device in operation 360. In such a case, the multi-modal natural language input may be communicated to the central device in its entirety, whereby the central device processes the input according to techniques described in Figure 4. However, should transmission to the central device fail for some reason, the input device may shift into a disconnected peer-to-peer mode where one or more secondary devices may be utilized, as will be described below in relation to Figure 7. When transmission to the central device occurs without incident, however, the input device may receive an intent determination from the central device in an operation 370, and may further receive results of one or more requests that the central device was able to

resolve, or requests that the central device has formulated for further processing on the input device. As such, the input device may take action in operation 380 based on the information received from the central device in operation 370. For example, the input device may route queries or commands to local or remote information sources or devices based on the intent determination, or may present results of the requests processed by the central device to the user.

[050] Referring to Figure 4, the central device may receive the multi-modal natural language input from the input device in an operation 410. The central device, having aggregated context and other knowledge from throughout the environment, may thus transcribe the utterance in an operation 420 and determine an intent of the input from the transcribed utterance in an operation 430. As such, the central device may consider information relating to context, domain agents, applications, and device capabilities throughout the environment in determining the intent of the utterance, including identification of one or more domains relevant to the input. However, it will be apparent that utilizing information aggregated from throughout the environment may cause ambiguity or uncertainty in various instances (e.g., an utterance containing the word "traffic" may have a different intent in domains relating to movies, music, and navigation).

[051] As such, once the central device has attempted to determine the intent of the natural language input, a determination may be made in an operation 440 as to whether one or more secondary devices (i.e., other devices in the constellation besides the input device) may also be capable of intent determination in the identified domain or domains. When no such secondary devices can be identified, decisional operation 440 may branch directly to an operation 480 to send to the input device the determined intent and any commands, queries, or other requests identified from the determined intent.

[052] On the other hand, when one or more secondary devices in the environment have intent determination capabilities in the identified domain or domains, the natural language input may be sent to such secondary devices in an operation 450. The secondary devices may then determine an intent as illustrated in Figure 5, which may include techniques generally similar to those described above in relation to the input device and central device (i.e., the natural language input may be received in an operation 510, an utterance contained therein may be transcribed in an operation 520, and an intent determination made in an operation 530 may be returned to the central device in an operation 540).

[053] Returning to Figure 4, the central device may collate intent determination responses received from the secondary devices in an operation 460. For example, as indicated above, the central device may identify one or more secondary devices capable of determining intent in a domain that the central device has identified as being relevant to the natural language utterance. As will be apparent, the secondary devices invoked in operation 450 may often include a plurality of devices, and intent determination responses may be received from the secondary devices in an interleaved manner, depending on processing resources, communications throughput, or other factors (e.g., the secondary devices may include a telematics device having a large amount of processing power and a broadband network connection and an embedded mobile phone having less processing power and only a cellular connection, in which case the telematics device may be highly likely to provide results to the central device before the embedded mobile phone). Thus, based on potential variations in response time of secondary devices, the central device may be configured to place constraints on collating operation 460. For example, the collating operation 460 may be terminated as soon as an intent determination has been received from one of the secondary devices that meets an acceptable level of confidence, or the operation 460 may be cut off when a predetermined amount of time has lapsed or a predetermined amount of resources have been consumed. In other implementations, however, it will be apparent that collating operation 460 may be configured to run to completion, regardless of whether delays have occurred or suitable intent determinations have been received. Further, it will be apparent that various criteria may be used to determine whether or when to end the collating operation 460, including the nature of a given natural language input, dialogue, or other interaction, or system or user preferences, among other criteria, as will be apparent.

[054] In any event, when the collating operation 460 has completed, a subsequent operation 470 may include the central device arbitrating among the intent determination responses received from one or more of the secondary devices previously invoked in operation 450. For example, each of the invoked secondary devices that generate an intent determination may also assign a confidence level to that intent determination, and the central device may consider the confidence levels in arbitrating among the responses. Moreover, the central device may associate other criteria with the secondary devices or the intent determinations received from the secondary devices to further enhance a likelihood that the best intent determination will be used. For example, various ones of the secondary devices may only be invoked for partial recognition in

distinct domains, and the central device may aggregate and arbitrate the partial recognitions to create a complete transcription. In another example, a plurality of secondary devices may be invoked to perform overlapping intent determination, and the central device may consider capabilities of the secondary devices to weigh the respective confidence levels (e.g., when one of two otherwise identical secondary devices employs multi-pass speech recognition analysis, the secondary device employing the multi-pass speech recognition analysis may be weighed as having a higher likelihood of success). It will be apparent that the central device may be configured to arbitrate and select one intent determination from among all of the intent hypotheses, which may include the intent determination hypothesis generated by the central device in operation 430. Upon selecting the best intent determination hypothesis, the central device may then provide that intent determination to the input device in operation 480, as well as any commands, queries, or other requests that may be relevant thereto. The input device may then take appropriate action as described above in relation to Figure 3.

[055] According to various aspects of the invention, Figure 6 illustrates a block diagram of an exemplary distributed implementation of the integrated, multi-modal, multi-device natural language voice service environment. As described above, the distributed implementation may also be categorized as a disconnected or peer-to-peer mode that may be employed when a central device in a centralized implementation cannot be reached or otherwise does not meet the needs of the environment. The distributed implementation illustrated in Figure 6 may be generally operate with similar purposes as described above in relation to the centralized implementation (i.e., to ensure that the environment includes a comprehensive model of aggregate knowledge and capabilities of a plurality of devices 610a-n in the environment). Nonetheless, the distributed implementation may operate in a somewhat different manner, in that one or more of the devices 610a-n may be provided with the entire constellation model, or various aspects of the model may be distributed among the plurality of devices 610a-n, or various combinations thereof.

[056] Generally speaking, the plurality of voice-enabled devices 610a- may be coupled to one another by a voice services interface 630, which may include any suitable real or virtual interface (e.g., a common message bus or network interface, a service-oriented abstraction layer, etc.). The various devices 610a-n may therefore operate as cooperative nodes in determining intent for multi-modal natural language utterances received by any one of the devices 610. Furthermore,

the devices 610a-n may share knowledge of vocabularies, context, capabilities, and other information, while certain forms of data may be synchronized to ensure consistent processing among the devices 610a-n. For example, because natural language processing components used in the devices 610a-n may vary (e.g., different recognition grammars or speech recognition techniques may exist), vocabulary translation mechanisms, misrecognitions, context variables, criteria values, criteria handlers, and other such information used in the intent determination process should be synchronized to the extent that communication capabilities permit.

[057] By sharing intent determination capabilities, device capabilities, inferencing capabilities, domain knowledge, and other information, decisions as to routing an utterance to a specific one of the devices 610a-n may be made locally (e.g., at an input device), collaboratively (e.g., a device having particular capabilities relevant to the utterance may communicate a request to process the utterance), or various combinations thereof (e.g., the input device may consider routing to secondary devices only when an intent of the utterance cannot be determined). Similarly, partial recognition performed at one or more of the devices 610a-n may be used to determine routing strategies for further intent determination of the utterance. For example, an utterance that contains a plurality of requests relating to a plurality of different domains may be received at an input device that can only determine intent in one of the domains. In this example, the input device may perform partial recognition for the domain associated with the input device, and the partial recognition may also identify the other domains relevant to the utterance for which the input device does not have sufficient recognition information. Thus, the partial recognition performed by the input device may result in identification of other potentially relevant domains and a strategy may be formulated to route the utterance to other devices in the environment that include recognition information for those domains.

[058] As a result, multi-modal natural language inputs, including natural language utterances, may be routed among the various devices 610a-n in order to perform intent determination in a distributed manner. However, as the capabilities and knowledge held by any one of the devices 610a-n may vary, each of the devices 610a-n may be associated with a reliability factor for intent determinations generated by the respective devices 610a-n. As such, to ensure that final intent determinations can be relied upon with a sufficient level of confidence, knowledge may be distributed among the devices 610a-n to ensure that reliability metrics for intent determinations provided by each of the devices 610a-n are commensurable throughout the environment. For

example, additional knowledge may be provided to a device having a low intent determination reliability, even when such knowledge results in redundancy in the environment, to ensure commensurate reliability of intent determination environment-wide.

[059] Therefore, in distributed implementations of the integrated voice services environment, utterances may be processed in various ways, which may depend on circumstances at a given time (e.g., system states, system or user preferences, etc.). For example, an utterance may be processed locally at an input device and only routed to secondary devices when an intent determination confidence level falls below a given threshold. In another example, utterances may be routed to a specific device based on the modeling of knowledge and capabilities discussed above. In yet another example, utterances may be flooded among all of the devices in the environment, and arbitration may occur whereby intent determinations may be collated and arbitrated to determine a best guess at intent determination.

[060] Thus, utterances may be processed in various ways, including through local techniques, centralized techniques, distributed techniques, and various combinations thereof. Although many variations will be apparent, Figure 7 illustrates an exemplary method for combined local and distributed processing of multi-modal, natural language inputs in a distributed implementation of the voice service environment, according to various aspects of the invention. In particular, the distributed processing may begin in an operation 710, where a multi-modal natural language input may be received at an input device. The input device may then utilize various natural language processing capabilities associated therewith in an operation 720 to transcribe an utterance contained in the multi-modal input (e.g., using an Automatic Speech Recognizer and associated recognition grammars), and may subsequently determine a preliminary intent of the multi-modal natural language input in an operation 730. It will be apparent that operations 710 through 730 may generally be performed using local intent determination capabilities associated with the input device.

[061] Thereafter, the input device may invoke intent determination capabilities of one or more secondary devices in an operation 740. More particularly, the input device may provide information associated with the multi-modal natural language input to one or more of the secondary devices, which may utilize local intent determination capabilities to attempt to determine intent of the input using techniques as described in relation to Figure 5. It will also be apparent that, in various implementations, the secondary devices invoked in operation 740 may

include only devices having intent determination capabilities associated with a specific domain identified as being associated with the input. In any event, the input device may receive intent determinations from the invoked secondary devices in an operation 750, and the input device may then collate the intent determinations received from the secondary devices. The input device may then arbitrate among the various intent determinations, or may combine various ones of the intent determinations (e.g., when distinct secondary devices determine intent in distinct domains), or otherwise arbitrate among the intent determinations to determine a best guess at the intent of the multi-modal natural language input (e.g., based on confidence levels associated with the various intent determinations). Based on the determined intent, the input device may then take appropriate action in an operation 770, such as issuing one or more commands, queries, or other requests to be executed at one or more of the input device or the secondary devices.

[062] Furthermore, in addition to the exemplary implementations described above, various implementations may include a continuous listening mode of operation where a plurality of devices may continuously listen for multi-modal voice-based inputs. In the continuous listening mode, each of the devices in the environment may be triggered to accept a multi-modal input when one or more predetermined events occur. For example, the devices may each be associated with one or more attention words, such as "Phone, <multi-modal request>" for a mobile phone, or "Computer, <multi-modal request>" for a personal computer. When one or more of the devices in the environment recognize the associated attention word, keyword activation may result, where the associated devices trigger to accept the subsequent multi-modal request. Further, where a plurality of devices in a constellation may be listening, the constellation may use all available inputs to increase recognition rates.

[063] Moreover, it will be apparent that the continuous listening mode may be applied in centralized voice service environments, distributed centralized voice service environments, or various combinations thereof. For example, when each device in the constellation has a different attention word, any given device that recognizes an attention word may consult a constellation model to determine a target device or functionality associated with the attention word. In another example, when a plurality of devices in the constellation share one or more attention words, the plurality of devices may coordinate with one another to synchronize information for processing the multi-modal input, such as a start time for an utterance contained therein.

[064] Implementations of the invention may be made in hardware, firmware, software, or various combinations thereof. The invention may also be implemented as instructions stored on a machine-readable medium, which may be read and executed by one or more processors. A machine-readable medium may include various mechanisms for storing or transmitting information in a form readable by a machine (e.g., a computing device). For example, a machine-readable storage medium may include read only memory, random access memory, magnetic disk storage media, optical storage media, flash memory devices, and others, and a machine-readable transmission media may include forms of propagated signals, such as carrier waves, infrared signals, digital signals, and others. Further, firmware, software, routines, or instructions may be described in the above disclosure in terms of specific exemplary aspects and implementations of the invention, and performing certain actions. However, it will be apparent that such descriptions are merely for convenience and that such actions in fact result from computing devices, processors, controllers, or other devices executing the firmware, software, routines, or instructions.

[065] Aspects and implementations may be described as including a particular feature, structure, or characteristic, but every aspect or implementation may not necessarily include the particular feature, structure, or characteristic. Further, when a particular feature, structure, or characteristic has been described in connection with an aspect or implementation, it will be understood that such feature, structure, or characteristic may be included in connection with other aspects or implementations, whether or not explicitly described. Thus, various changes and modifications may be made to the preceding description without departing from the scope or spirit of the invention, and the specification and drawings should therefore be regarded as exemplary only, and the scope of the invention determined solely by the appended claims.

CLAIMS

What is claimed is:

1. A method for providing an integrated natural language voice services environment, the integrated voice services environment including an input device that receives a multi-modal natural language input, a central device communicatively coupled to the input device, and one or more secondary devices communicatively coupled to the central device, each of the input device, the central device, and the one or more secondary devices having intent determination capabilities for processing multi-modal natural language inputs, the method comprising:

communicating the multi-modal natural language input from the input device to the central device, the central device operable to:

aggregate the intent determination capabilities of the input device and the one or more secondary devices; and

determine an intent of the multi-modal natural language input received at the input device using the aggregated intent determination capabilities;

receiving the determined intent of the multi-modal natural language input at the input device; and

invoking at least one action at one or more of the input device, the central device, or the secondary devices based on the determined intent.

2. The method of claim 1, the intent determination capabilities for a given one of the input device, the central device, or the secondary devices based on at least one of processing power, storage resources, natural language processing capabilities, or local knowledge.

3. The method of claim 1, the central device further operable to:

communicate the multi-modal natural language input to each of the secondary devices, wherein each of the secondary devices determine an intent of the multi-modal natural language input received at the input device using local intent determination capabilities;

receive the intent determined by each of the secondary devices; and

arbitrate among the intent determinations of the secondary devices to determine the intent of the multi-modal natural input.

4. The method of claim 3, wherein the central device arbitrates among the intent determinations of the secondary devices based on respective confidence levels associated therewith.

5. The method of claim 4, the central device further operable to weigh the confidence level of the intent determinations of the secondary devices, wherein intent determinations of secondary devices that use multi-pass speech recognition techniques receive a higher weight.

6. The method of claim 1, further comprising sharing knowledge among the input device, the central device, and the secondary devices to update the intent determination capabilities thereof.

7. The method of claim 6, wherein the shared knowledge includes information relating to at least one of vocabularies, translation mechanisms, misrecognitions, context, or states of the devices in the environment.

8. The method of claim 1, the input device operable to:

determine an intent of the multi-modal natural language input received at the input device using local intent determination capabilities; and

invoke the at least one action without communicating the multi-modal natural language input to the central device when the intent determination of the input device meets a predetermined level of confidence.

9. The method of claim 1, wherein the voice services environment operates in a continuous listening mode.

10. A method for providing an integrated natural language voice services environment, the integrated voice services environment including an input device that receives a multi-modal natural language input and one or more secondary devices communicatively coupled to the input device, each of the input device and the one or more secondary devices having intent

determination capabilities for processing multi-modal natural language inputs, the method comprising:

determining a preliminary intent of the multi-modal natural language input, wherein the input device determines the preliminary intent using local intent determination capabilities;

communicating the multi-modal natural language input to one or more of the secondary devices, each of which determine an intent of the multi-modal natural language input using local intent determination capabilities;

collating the preliminary intent determination and the intent determinations of the secondary devices;

arbitrating among the collated intent determinations to determine an actionable intent of the multi-modal natural input; and

invoking at least one action at one or more of the input device or the secondary devices based on the determined actionable intent.

11. The method of claim 10, further comprising sharing knowledge among the input device and the secondary devices relating to respective local intent determination capabilities.

12. The method of claim 11, further comprising associating a reliability factor with the local intent determination capabilities of each of the input device and the one or more secondary devices, wherein an aggregation of the reliability factors for the local intent determination capabilities provide commensurable intent determination capabilities throughout the environment.

13. The method of claim 11, wherein the shared knowledge includes information relating to at least one of processing power, storage resources, natural language processing capabilities, or knowledge associated with the input device and the secondary devices.

14. The method of claim 10, further comprising sharing knowledge among the input device and the secondary devices to update the intent determination capabilities thereof.

15. The method of claim 14, wherein sharing the knowledge includes transmitting knowledge established on the input device to the one or more secondary devices.

16. The method of claim 14, wherein sharing the knowledge includes transmitting knowledge established on one of the secondary devices to the input device and to the secondary devices other than the device where the knowledge was established.

17. The method of claim 14, wherein the shared knowledge includes information relating to at least one of vocabularies, translation mechanisms, misrecognitions, context, or states of the devices in the environment.

18. The method of claim 10, wherein the voice services environment operates in a continuous listening mode.

19. A system for providing an integrated natural language voice services environment, the system comprising:

an input device that receives a multi-modal natural language input;

one or more secondary devices communicatively coupled to the input device, each of the input device and the one or more secondary devices having intent determination capabilities for processing multi-modal natural language inputs; and

a constellation model accessible to each of the input device and the one or more secondary devices, the constellation model describing the intent determination capabilities of the input device and the one or more secondary devices, the multi-modal natural language input routed for processing at one or more of the input device or the secondary devices to determine an intent thereof based on the intent determination capabilities described in the constellation model.

20. The system of claim 19, wherein the constellation model describes, for each of the input device and the one or more secondary devices, one or more of processing power, storage resources, natural language processing capabilities, or knowledge associated therewith.

21. The system of claim 19, wherein the constellation model dynamically selecting one of a centralized model or a distributed model for operating the environment.

22. The system of claim 21, the centralized model designating one of the secondary devices as a central device, the central device operable to:

aggregate the intent determination capabilities of the input device and the one or more secondary devices; and

determine the intent of the multi-modal natural language input using the aggregated intent determination capabilities.

23. The system of claim 22, wherein the constellation model selects the distributed model when the multi-modal natural language cannot be communicated to the central device.

24. The system of claim 23, wherein selecting the distributed model causes the input device and the one or more secondary devices to:

share knowledge relating to respective local intent determination capabilities; and

operate as cooperative nodes to determine the intent of the multi-modal natural language input using the shared knowledge relating to local intent determination capabilities.

25. The system of claim 24, wherein operating as cooperative nodes includes determining whether the intent of the multi-modal natural language input should be determined by the input device, by one or more of the secondary devices, or by each of the input device and the one or more secondary devices.

26. The system of claim 19, further comprising at least one domain agent associated with the input device, wherein the at least one domain agent bootstraps information relating to a context associated with the domain agent from one or more of the secondary devices.

27. The system of claim 19, wherein the voice services environment operates in a continuous listening mode.

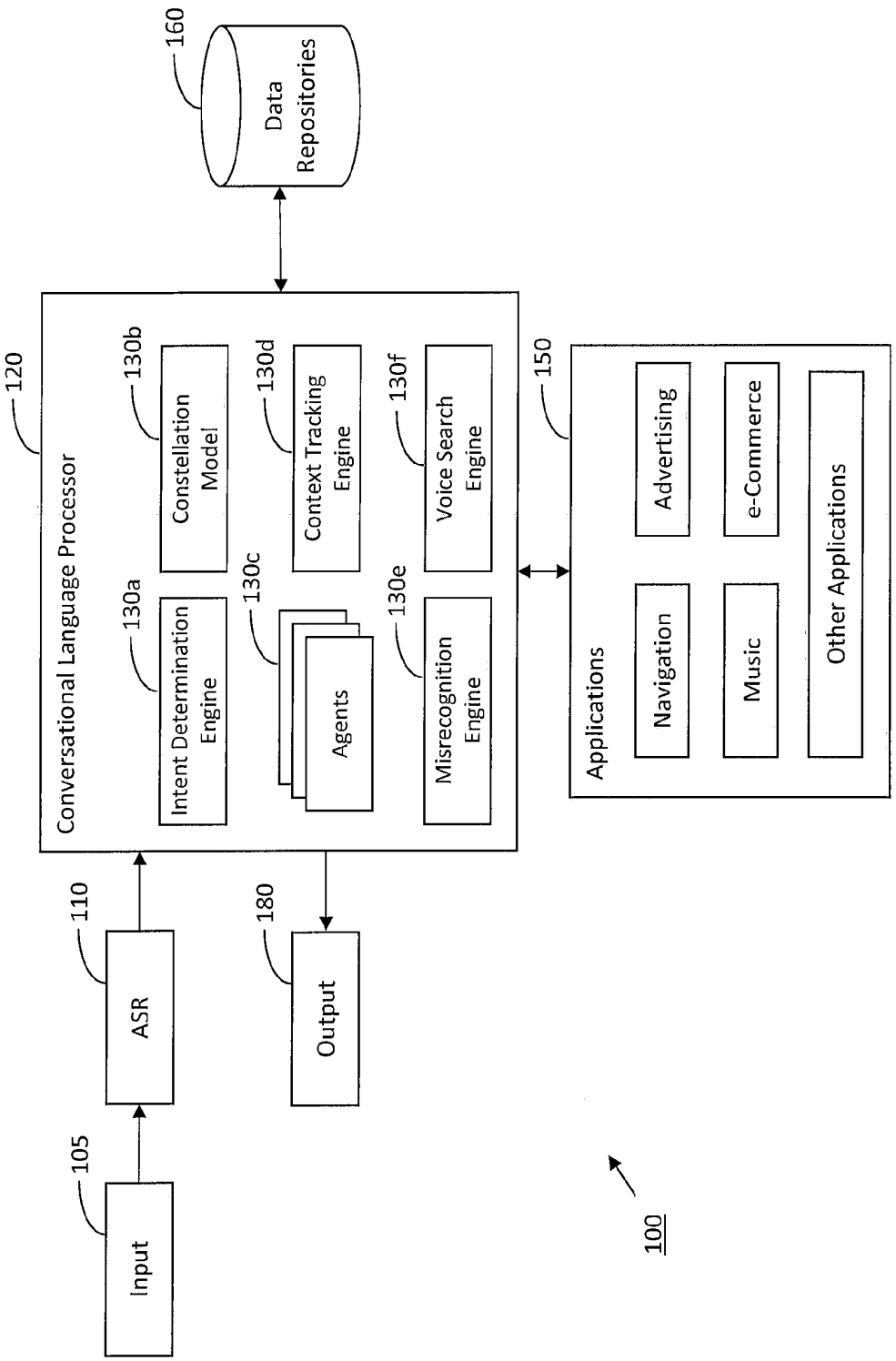

160
Data Repositories
120
Conversational Language Processor
130b
Constellation Model
130d
Context Tracking Engine
130f
Voice Search Engine
150
130a
Intent Determination Engine
130c
Agents
130e
Misrecognition Engine
Applications
Advertising
Navigation
e-Commerce
Music
Other Applications
110
ASR
180
Output
105
Input
100

Figure 1

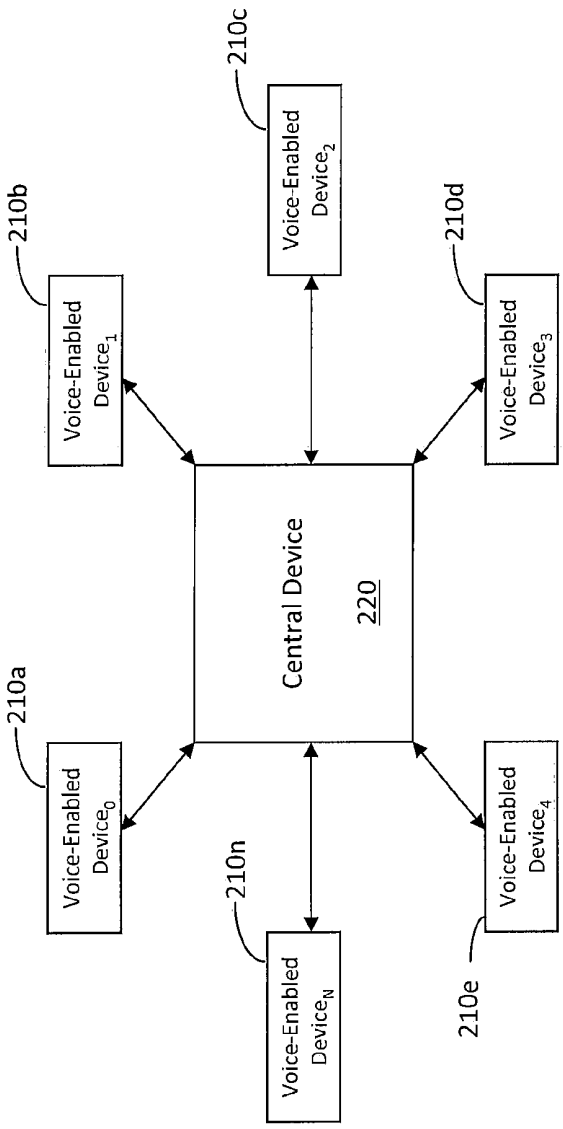

210c
Voice-Enabled Device$_2$
210b
Voice-Enabled Device$_1$
210d
Voice-Enabled Device$_3$
Central Device
220
210a
Voice-Enabled Device$_0$
210n
Voice-Enabled Device$_N$
Voice-Enabled Device$_4$
210e

Figure 2

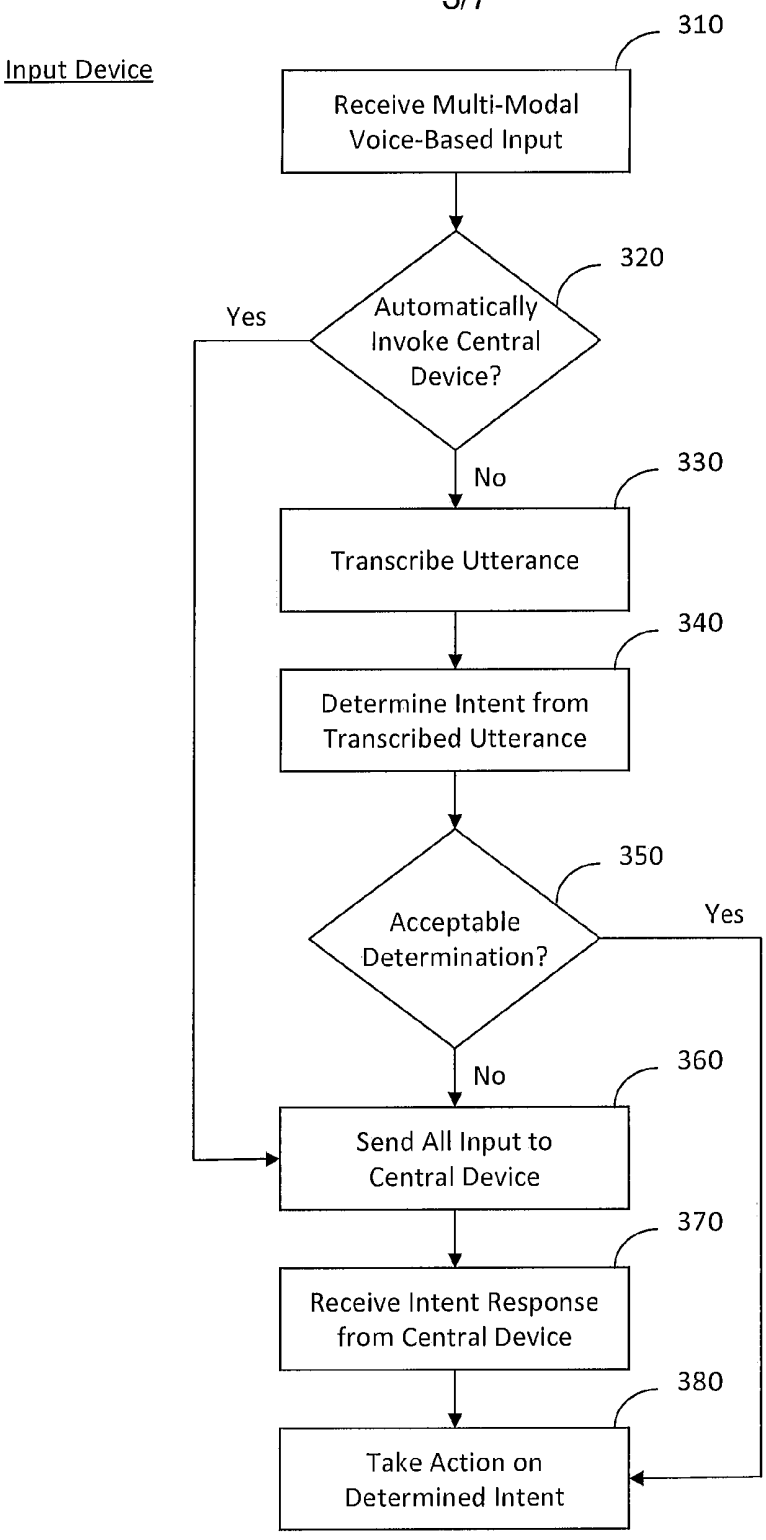

Input Device
310
Receive Multi-Modal Voice-Based Input
320
Automatically Invoke Central Device?
Yes
No
330
Transcribe Utterance
340
Determine Intent from Transcribed Utterance
350
Acceptable Determination?
Yes
No
360
Send All Input to Central Device
370
Receive Intent Response from Central Device
380
Take Action on Determined Intent

Figure 3

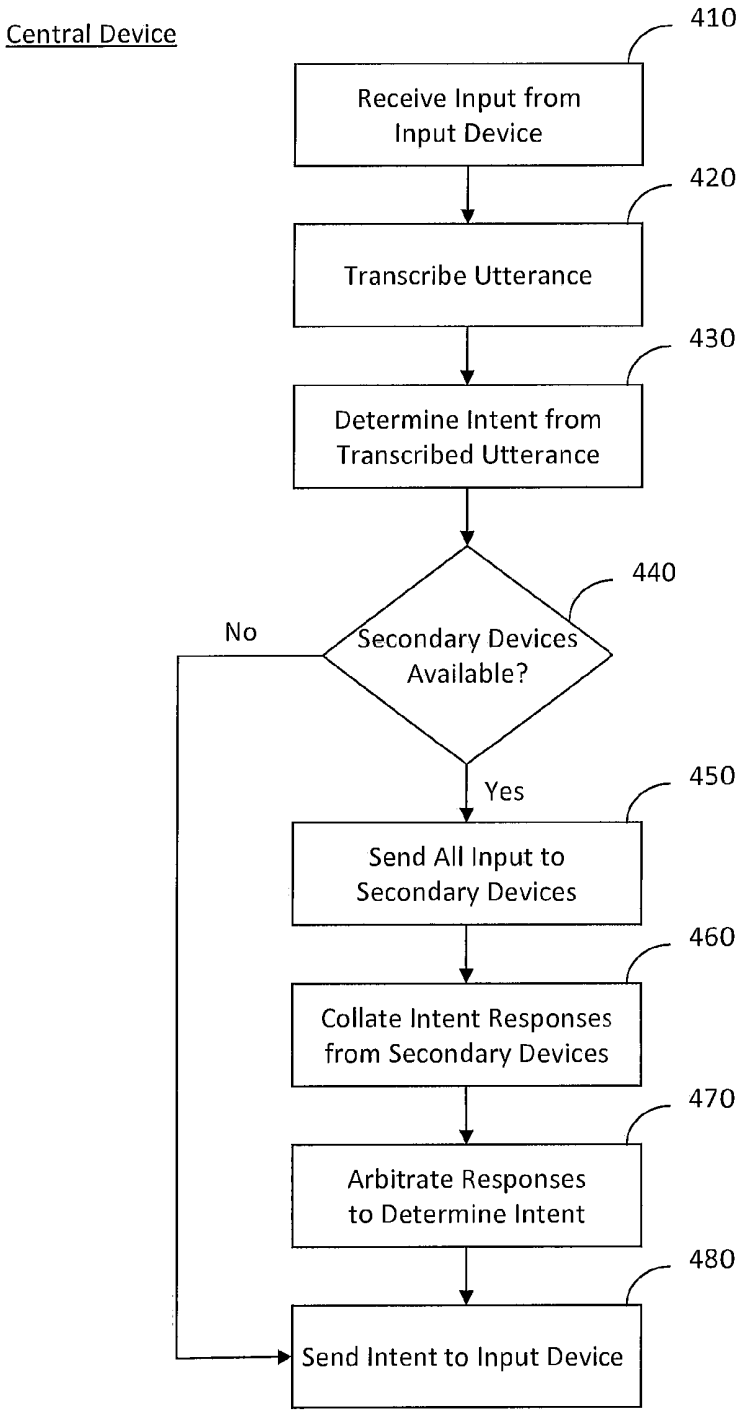

Central Device
410
Receive Input from Input Device
420
Transcribe Utterance
430
Determine Intent from Transcribed Utterance
440
Secondary Devices Available?
No
Yes
450
Send All Input to Secondary Devices
460
Collate Intent Responses from Secondary Devices
470
Arbitrate Responses to Determine Intent
480
Send Intent to Input Device

Figure 4

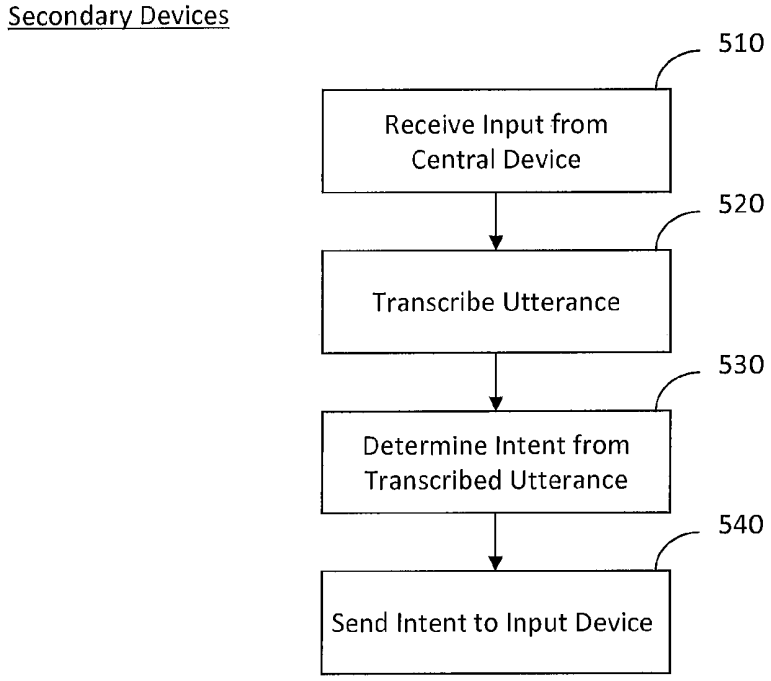

Secondary Devices
510
Receive Input from Central Device
520
Transcribe Utterance
530
Determine Intent from Transcribed Utterance
540
Send Intent to Input Device

Figure 5

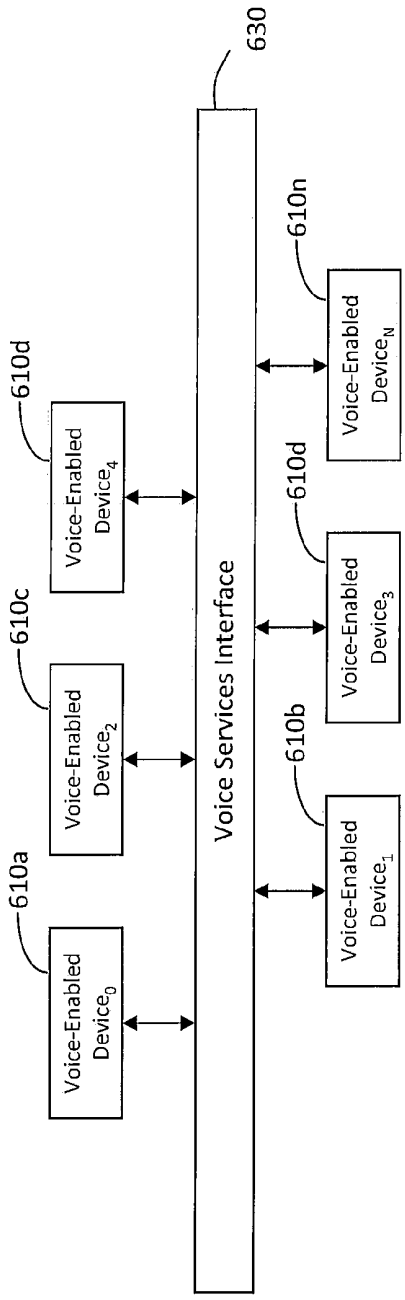

630
Voice Services Interface
610a
Voice-Enabled Device0
610b
Voice-Enabled Device1
610c
Voice-Enabled Device2
610d
Voice-Enabled Device3
610d
Voice-Enabled Device4
610n
Voice-Enabled DeviceN

Figure 6

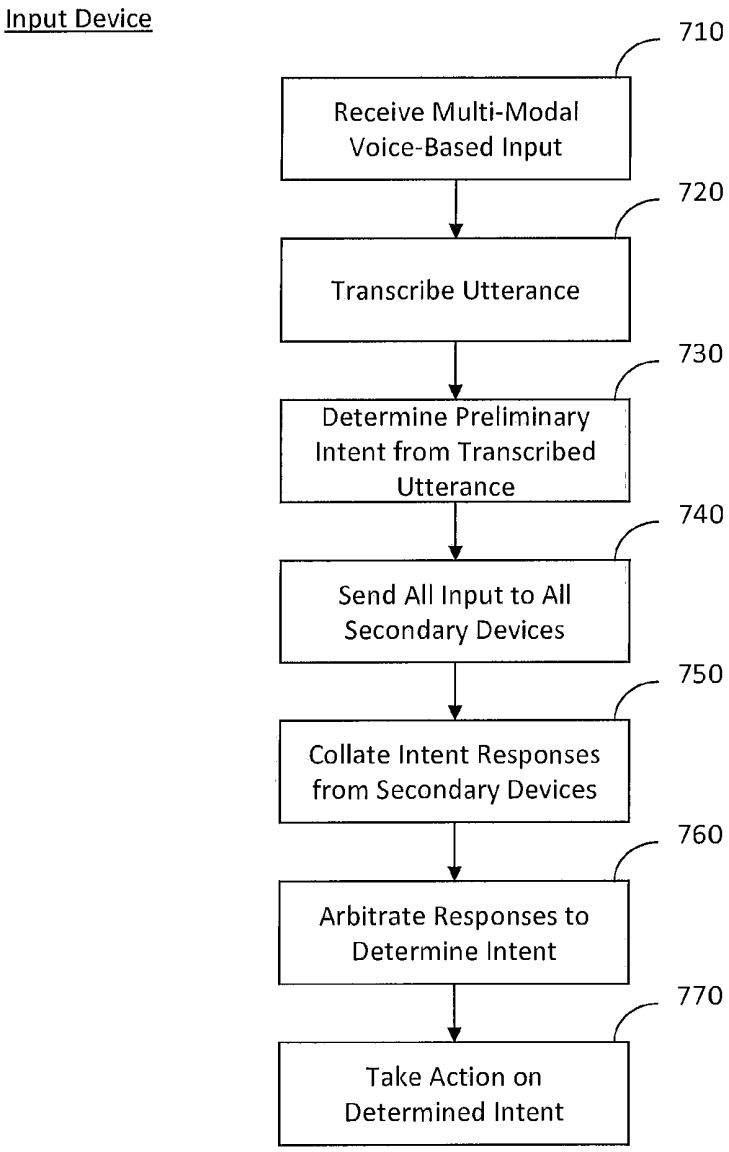

Input Device
710
Receive Multi-Modal Voice-Based Input
720
Transcribe Utterance
730
Determine Preliminary Intent from Transcribed Utterance
740
Send All Input to All Secondary Devices
750
Collate Intent Responses from Secondary Devices
760
Arbitrate Responses to Determine Intent
770
Take Action on Determined Intent

Figure 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 08/69524

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 13/12 (2008.04)
USPC - 710/65
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC: 710/65

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 710/11, 14, 37, 50, 73; 370/353, 458, 448; 379/907-908; 704/200, 246, 251-252 (see text terms below)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PubWEST (PGPB, USPT, USOC, EPAB, JPAB); DialogPRO; Google (incl. Patents, Scholar) - natural language, multi-modal, central, input, device, client, server, process, distributed, local, distributed system, local, context, speech

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 7,376,645 B2 (BERNARD) 20 May 2008 (20.05.2008), Abstract, FIGS.1-3; and Col 1, ln 19-30; Col 1, ln 35 to Col 2, ln 2; Col 2, ln 21 to Col 4, ln 16; Col 4, ln 21-56, 60; Col 5, ln 10; Col 6, ln 1-5, 7, 22-23; Col 6, ln 37 to Col 7, ln 4; Col 7, ln 13-26; Col 7, ln 52 to Col 10, ln 37	1-27
A	US 2004/0025115 A1 (SIENEL et al.) 5 February 2004 (05.02.2004)	1-27
A	US 2007/0050191 A1 (WEIDER et al.) 1 March 2007 (01.03.2007)	1-27
A	US 2004/0166832 A1 (PORTMAN et al.) 26 August 2004 (26.08.2004)	1-27
A	US 6,615,172 B1 (BENNETT et al.) 2 September 2003 (02.09.2003)	1-27

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
04 October 2008 (04.10.2008)	09 OCT 2008
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (April 2007)

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 August 2010 (26.08.2010)

PCT

(10) International Publication Number
WO 2010/096752 A1

(51) International Patent Classification:
G10L 15/00 (2006.01)

(21) International Application Number:
PCT/US2010/024895

(22) International Filing Date:
22 February 2010 (22.02.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
12/389,678 20 February 2009 (20.02.2009) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [US/US]; 11980 NE 24th Street, Suite 100, Bellevue, WA 98005 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **BALDWIN, Larry** [US/US]; 23125 SE 243rd Place, Maple Valley, WA 98038 (US). **WEIDER, Chris** [US/US]; 10757 Greenwood Avenue North #220, Seattle, WA 98133 (US).

(74) Agents: **ALI, Syed, Jafar** et al.; Pillsbury Winthrop Shaw Pittman LLP, P.O. Box 10500, McLean, VA 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

Published:

— *with international search report (Art. 21(3))*

(54) Title: SYSTEM AND METHOD FOR PROCESSING MULTI-MODAL DEVICE INTERACTIONS IN A NATURAL LANGUAGE VOICE SERVICES ENVIRONMENT

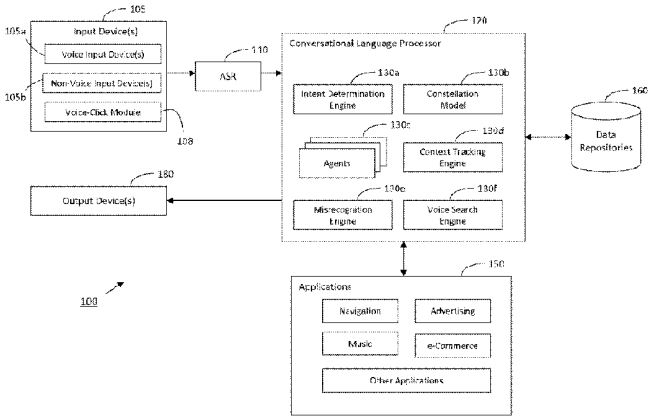


Figure 1

(57) Abstract: A system and method for processing multi-modal device interactions in a natural language voice services environment may be provided. In particular, one or more multi-modal device interactions may be received in a natural language voice services environment that includes one or more electronic devices. The multi-modal device interactions may include a non-voice interaction with at least one of the electronic devices or an application associated therewith, and may further include a natural language utterance relating to the non-voice interaction. Context relating to the non-voice interaction and the natural language utterance may be extracted and combined to determine an intent of the multi-modal device interaction, and a request may then be routed to one or more of the electronic devices based on the determined intent of the multi-modal device interaction.

WO 2010/096752 A1

SYSTEM AND METHOD FOR PROCESSING MULTI-MODAL DEVICE INTERACTIONS IN A NATURAL LANGUAGE VOICE SERVICES ENVIRONMENT

CROSS-REFERENCE TO RELATED APPLICATIONS

[001] This application claims the benefit of U.S. Patent Application Serial No. 12/389,678, entitled "SYSTEM AND METHOD FOR PROCESSING MULTI-MODAL DEVICE INTERACTIONS IN A NATURAL LANGUAGE VOICE SERVICES ENVIRONMENT", filed February 20, 2009, which is hereby incorporated by reference in its entirety.

FIELD OF THE INVENTION

[002] The invention relates to an integrated natural language voice services environment for processing multi-modal interactions with or one or more devices and/or applications, wherein the multi-modal interactions may provide additional context for cooperatively interpreting and otherwise processing a natural language utterance accompanying the multi-modal interactions.

BACKGROUND OF THE INVENTION

[003] As technology has progressed in recent years, consumer electronic devices have emerged to become nearly ubiquitous in the everyday lives of many people. To meet the increasing demand that has resulted from growth in the functionality and mobility of mobile phones, navigation devices, embedded devices, and other such devices, many devices offer a wealth of features and functions in addition to core applications. Greater functionality also introduces trade-offs, however, including learning curves that often inhibit users from fully exploiting all of the capabilities of their electronic devices. For example, many existing electronic devices include complex human to machine interfaces that may not be particularly user-friendly, which can inhibit mass-market adoption for many technologies. Moreover, cumbersome interfaces often result in otherwise desirable features being difficult to find or use (e.g., because of menus that are complex or otherwise tedious to navigate). As such, many users tend not to use, or even know about, many of the potential capabilities of their devices.

[004] As such, the increased functionality of electronic devices often tends to be wasted, as market research suggests that many users only use only a fraction of the features or applications available on a given device. Moreover, in a society where wireless networking and broadband access are increasingly prevalent, consumers tend to naturally desire seamless

mobile capabilities from their electronic devices. Thus, as consumer demand intensifies for simpler mechanisms to interact with electronic devices, cumbersome interfaces that prevent quick and focused interaction become an important concern. Nevertheless, the ever-growing demand for mechanisms to use technology in intuitive ways remains largely unfulfilled.

[005] One approach towards simplifying human to machine interactions in electronic devices has included the use of voice recognition software, which has the potential to enable users to exploit features that would otherwise be unfamiliar, unknown, or difficult to use. For example, a recent survey conducted by the Navteq Corporation, which provides data used in a variety of applications such as automotive navigation and web-based applications, demonstrates that voice recognition often ranks among the features most desired by consumers of electronic devices. Even so, existing voice user interfaces, when they actually work, still require significant learning on the part of the user.

[006] For example, many existing voice user interface only support requests formulated according to specific command-and-control sequences or syntaxes. Furthermore, many existing voice user interfaces cause user frustration or dissatisfaction because of inaccurate speech recognition. Similarly, by forcing a user to provide pre-established commands or keywords to communicate requests in ways that a system can understand, existing voice user interfaces do not effectively engage the user in a productive, cooperative dialogue to resolve requests and advance a conversation towards a satisfactory goal (e.g., when users may be uncertain of particular needs, available information, device capabilities, etc.). As such, existing voice user interfaces tend to suffer from various drawbacks, including significant limitations on engaging users in a dialogue in a cooperative and conversational manner.

[007] Additionally, many existing voice user interfaces fall short in utilizing information distributed across different domains, devices, and applications in order to resolve natural language voice-based inputs. Thus, existing voice user interfaces suffer from being constrained to a finite set of applications for which they have been designed, or to devices on which they reside. Although technological advancement has resulted in users often having several devices to suit their various needs, existing voice user interfaces do not adequately free users from device constraints. For example, users may be interested in services associated with different applications and devices, but existing voice user interfaces tend to restrict users from accessing the applications and devices as they see fit. Moreover, users typically can only practicably carry

a finite number of devices at any given time, yet content or services associated with users' other devices currently being used may be desired in various circumstances.

[008] Accordingly, although users tend to have varying needs, where content or services associated with different devices may be desired in various contexts or environments, existing voice technologies tend to fall short in providing an integrated environment in which users can request content or services associated with virtually any device or network. As such, constraints on information availability and device interaction mechanisms in existing voice services environments tend to prevent users from experiencing technology in an intuitive, natural, and efficient way. For instance, when a user wishes to perform a given function using a given electronic device but does not necessarily know how to go about performing the function, the user typically cannot engage in a multi-modal interaction with the device to simply utter words in natural language to request the function.

[009] Furthermore, relatively simple functions can often be tedious to perform using electronic devices that do not have voice recognition capabilities. For example, purchasing new ring-tones for a mobile phone tends to be a relatively straightforward process, but users must typically navigate several menus and press many different buttons in order to complete the process. As such, it becomes apparent that interaction with electronic devices could be far more efficient if users were able to use natural language to exploit buried or otherwise difficult to use functionality. Existing systems suffer from these and other problems.

SUMMARY OF THE INVENTION

[010] According to one aspect of the invention, a system and method for processing multi-modal device interactions in a natural language voice services environment may be provided. In particular, one or more multi-modal interactions may be received in a natural language voice services environment that includes one or more electronic devices. The multi-modal device interactions may include a user engaging in a non-voice interaction with one or more of the electronic devices or applications associated with the devices, while also providing a natural language utterance in relation to the non-voice interaction. For example, the non-voice device interaction may comprise the user selecting a particular segment, item, data, point of focus, or attention focus, or otherwise engaging in one or more unique and distinguishable interactions with the electronic devices or applications associated therewith. As such, context may be

extracted from the natural language utterance, and the non-voice device interaction may provide additional context for the natural language utterance. The context of the utterance and the non-voice device interaction may then be combined to determine an intent of the multi-modal device interaction, wherein one or more of the electronic devices may process a request based on the intent of the multi-modal device interaction.

[011] According to one aspect of the invention, at least one of the electronic devices may include an input device configured to receive voice-based inputs. In one implementation, the voice-based input device may be signaled to capture the natural language utterance in response to detecting the non-voice interaction with the one or more electronic devices or applications. Furthermore, the natural language voice services environment may include one or more listeners established for the electronic devices and the associated applications, wherein the listeners may be configured to detect the non-voice interaction with the electronic devices or applications. As such, information relating to the non-voice interaction and the accompanying natural language utterance may be aligned to enable cooperative processing of the utterance and the non-voice device interaction.

[012] According to one aspect of the invention, at least one transaction lead may be generated based on the intent of the multi-modal device interaction. For example, a further multi-modal device interaction may be received, wherein the further multi-modal device interaction may relate to the transaction lead generated for the first multi-modal device interaction. At least one request may then be routed to one or more of the electronic devices based on an intent determined for the further multi-modal device interaction, whereby a transaction click-through may be processed in response to receiving the device interaction relating to the generated transaction lead. For example, the transaction lead may include an advertisement or a recommendation that is selected based on the intent of the original multi-modal device interaction, while the further multi-modal device interaction may include a user selecting the advertisement or recommendation. Thus, the selection of the advertisement or recommendation may be considered a transaction click-through, which may generate revenue for a particular entity (e.g., a provider of the natural language voice services environment).

[013] Other objects and advantages of the invention will be apparent based on the following drawings and detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

[014] Figure 1 illustrates a block diagram of an exemplary system for processing multi-modal device interactions in a natural language voice services environment, according to various aspects of the invention.

[015] Figure 2 illustrates a block diagram of an exemplary method for synchronizing multi-modal devices in a natural language voice service environment, according to various aspects of the invention.

[016] Figure 3 illustrates a flow diagram of an exemplary method for processing multi-modal device interactions in a natural language voice service environment, according to various aspects of the invention.

[017] Figure 4 illustrates a flow diagram of an exemplary method for processing multi-modal device interactions to generate one or more transaction leads in a natural language voice service environment, according to various aspects of the invention.

DETAILED DESCRIPTION

[018] According to various aspects of the invention, Figure 1 illustrates a block diagram of an exemplary system 100 for processing multi-modal device interactions in a natural language voice services environment. As will be apparent from the further description to be provided herein, the system 100 illustrated in Figure 1 may include an input device 105, or a combination of input devices 105, which enable users to interact with the system 100 in a multi-modal manner. In particular, the system 100 may include various natural language processing components, including at least a voice-click module 108, which can collectively process a user's multi-modal interactions with one or more input devices 105. For example, in one implementation, the input devices 105 may include any suitable combination of at least one voice input device 105a (e.g., a microphone) and at least one non-voice input device 105b (e.g., a mouse, touch-screen display, wheel selector, etc.). As such, the input devices 105 may include any suitable combination of electronic devices having mechanisms for receiving both voice-based and non-voice-based inputs (e.g., a microphone coupled to one or more of a telematics device, personal navigation device, mobile phone, VoIP node, personal computer, media device, embedded device, server, or other electronic device). As such, the system 100 may enable users to engage in multi-modal conversational interactions with the one or more

electronic input devices 105 or applications associated with the electronic devices 105, wherein the system 100 may process the device interactions in a free-form and cooperative manner suitable for routing tasks or otherwise resolving requests.

[019] As indicated above, in one implementation, the system may include various natural language processing components that can support free-form utterances and/or other forms of device interactions, which may liberate users from restrictions relating to the manner of formulating commands, queries, or other requests. As such, users may interact with the input devices 105 using any manner of speaking into the voice input device 105a or otherwise interacting with the non-voice input devices 105b in order to request content or services available in the system 100. For instance, a user may request any content or services that may be available in the system 100 by providing a natural language utterance to the voice input device 105a. In one implementation, the utterance may then be processed using techniques described in U.S. Patent No. 7,398,209, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," issued July 8, 2008, and U.S. Patent Application 10/618,633, entitled "Mobile Systems and Methods for Responding to Natural Language Speech Utterance," filed June 15, 2003, the disclosures of which are hereby incorporated by reference in their entirety. In addition, the user may interact with one or more of the non-voice input devices 105b to provide further context or other information relating to the utterance and/or the requested content or services.

[020] In one implementation, the system 100 may be coupled to various other systems that include additional multi-modal devices, with the other systems having natural language processing capabilities similar to those shown in Figure 1. The system 100 may therefore provide an interface to a multi-device environment in which users may request content or services available through the various additional devices in the environment. For example, in one implementation, the system 100 may include a constellation model 130b that provides knowledge relating to content, services, applications, intent determination capabilities, and other features available through the other systems and devices in the environment. For example, in one implementation, the system 100 may interact with devices, applications, or other systems in the environment to cooperatively resolve requests, as described in co-pending U.S. Patent Application 12/127,343, entitled "System and Method for an Integrated, Multi-Modal, Multi-Device Natural Language Voice Services Environment," filed May 27, 2008, the

disclosure of which is hereby incorporated by reference in its entirety. For example, the multi-device environment may share information among various systems and devices to provide a cooperative environment for resolving requests, wherein the shared information may relate to aspects such as device capabilities, context, prior interactions, domain knowledge, short-term knowledge, long-term knowledge, and cognitive models, among other things.

[021] As indicated above, the system 100 illustrated in Figure 1 may include, among other things, one or more electronic input devices 105 that collectively provide an interface (or combination of interfaces) for receiving one or more multi-modal device interactions from a user, wherein the device interactions include at least a user-spoken utterance. Although the implementation illustrated in Figure 1 includes a distinct voice input device 105a and non-voice input device 105b, it will be apparent that in one or more implementations the voice input device 105a and the non-voice input device 105b may be components of the same or separate devices. For example, the input devices 105 may include a microphone coupled to a mobile phone (i.e., the voice input device 105a), and may further include one or more buttons, selectable displays, wheel selectors, or other components coupled to the mobile phone (i.e., the non-voice input devices 105b). In another example, the input devices 105 may include a combination of a microphone coupled to a telematics device (i.e., the voice input device 105a), and may further include buttons, a touch-screen display, a track wheel, or other non-voice input devices 105b coupled to a media player that is communicatively coupled to, yet distinct from, the telematics device. Thus, the input devices 105 may include any suitable combination of communicatively coupled electronic devices that includes at least one input device for receiving natural language utterance inputs and at least one input device for receiving multi-modal non-voice inputs.

[022] In one implementation, a voice-click module 108 communicatively coupled to the one or more input devices 105 may enable cooperative processing of multi-modal device interactions received at the voice input device 105a and one or more of the non-voice input devices 105b. For example, the voice-click module 108 may provide the system 100 with information that can be used to process a natural language utterance received via the voice input device 105a in view of one or more non-voice device interactions received via the non-voice input devices 105b. The voice-click module 108 therefore enables the user to interact with the various input devices 105 in an intuitive and free-form manner, whereby the user may provide various types

of information to the system 100 when seeking to initiate action, retrieve information, or otherwise request content or services available in the system 100.

[023] The voice input device 105a may comprise any appropriate device, or combination of devices, which have capabilities for receiving natural language utterances or other forms of spoken input. For example, in one implementation, the voice input device 105a may include a directional microphone, an array of microphones, or other devices capable of creating encoded speech. In one implementation, the voice input device 105a may be configured to maximize fidelity of the encoded speech. For example, the voice input device 105a may be configured to maximize gain in a direction of the user, cancel echoes, null point noise sources, perform variable rate sampling, filter environmental noise or background conversations, or use other techniques to maximize the fidelity of the encoded speech. As such, the voice input device 105a may create the encoded speech in a manner tolerant of noise or other factors that could otherwise interfere with the system 100 accurately interpreting natural language utterances.

[024] The non-voice input device 105b may comprise any appropriate device, or combination of devices, which have capabilities for supporting non-voice device interactions. For example, in one implementation, the non-voice input device 105b may include a combination stylus and touch-screen or tablet interface, a BlackBerry® wheel selector, an iPod® click wheel, a mouse, a keypad, buttons, or any other devices that supports distinguishable non-voice device interactions. The user may therefore use the non-voice input devices 105b to make data selections or identify a point of focus (or attention focus) to be processed in connection with a related natural language utterance provided via the voice input device 105a. For example, the user may point a stylus at a specific segment of a touch-screen display, highlight text using a mouse, click a button, interact with an application, or otherwise engage in any suitable device interaction for selecting data or otherwise identifying a point of focus (i.e., voice-activating or "voice-clicking" the selected data and/or the identified point of focus).

[025] Furthermore, in addition to being usable for making data selections, identifying points of focus, or otherwise activating data to be interpreted in association with one or more utterances, the user may further use the non-voice input devices 105b to engage in specialized device interactions that have meaning in the system 100. For example, the specialized device interactions (which may be referred to as "clicks" or "voice-clicks") may include clicks lasting a given duration, clicks continuously held for a given duration, clicks made in a predetermined

sequence, or any other interaction or sequence of interactions that the input devices 105 and/or the voice-click module 108 can identify, detect, or otherwise distinguish.

[026] In one implementation, the specialized device interactions may be associated with one or more actions, queries, commands, tasks, or other requests associated with applications or services available in the system 100. In one implementation, the specialized device interactions may further include one or more actions, queries, commands, tasks, or other requests associated with any of various devices deployed in the multi-device environment, as described in the above-referenced co-pending U.S. Patent Application 12/127,343, entitled "System and Method for an Integrated, Multi-Modal, Multi-Device Natural Language Voice Services Environment," filed May 27, 2008. For example, a distinct sequence of clicking a stylus on a particular segment or item displayed on a touch-screen display may be defined as a specialized device interaction or voice-click for initiating a telephone call on a mobile phone, calculating a route on a navigation device, purchasing a song to a media player, or another type of request.

[027] Thus, the voice-click module 108 coupled to the input device 105 may continually monitor the user's interactions with the non-voice input device 105b to detect occurrences of at least one non-voice device interaction, which may be referred to herein as a "voice-click." The detected voice-click may therefore provide further context for processing a multi-modal device interaction, which may include the at least one voice-click and one or more natural language utterances, each of which may provide context for task specification. Thus, the voice-click may generally signal the system 100 that a current utterance or other voice-based input is to be processed together with a current interaction with one or more of the devices 105. For example, in one implementation, the current device interaction may include the user selecting, highlighting, or otherwise identifying a particular point of attention, object, or other item associated with one or more of the devices 105. As such, the current device interaction may provide context for sharpening recognition, interpretation, and understanding of the accompanying utterance, and moreover, the current utterance may provide information to enhance the context provided by the accompanying device interaction.

[028] In one implementation, the voice-click module 108 may determine the various voice-click interactions to be detected based on particular characteristics of the non-voice input devices 105b (e.g., the voice-click interactions may include distinguishable interactions that the non-voice input devices 105b support). For example, a multi-touch display typically includes a

touch-screen display device configured to support various distinguishable gestures for interacting with information displayed therein (e.g., a user may zoom in, zoom out, rotate, or otherwise control graphical information displayed on a multi-touch screen using specific gestures or other interaction techniques). Thus, in one example, the non-voice input devices 105b may include a multi-touch display, in which case the voice-click module 108 may be configured to detect an occurrence of a voice-click when a user engages in one or more of the distinguishable gestures supported by the non-voice multi-touch display 105b.

[029] In one implementation, a user may customize or otherwise modify the voice-click interactions to be detected by the voice-click module 108. In particular, the specific device interactions detected by the voice-click module 108 may be removed or modified, or new device interactions may be added. As such, the voice-click device interactions detected by the voice-click module 108 may include any appropriate interaction or combination of interactions that the non-voice input devices 105b and/or the voice-click module 108 can distinguish.

[030] When the voice-click module 108 detects a user's engaging in a voice-click device interaction, the voice-click module 108 may extract context information associated with the voice-click device interaction for voice-activation. In particular, the voice-click module 108 may identify information relating to a segment, item, point of focus, attention focus, or other data selected by the user, or otherwise identify information relating to a particular device interaction or sequence of device interactions engaged in by the user. The voice-click module 108 thus extracts the identified information relating to the detected voice-click, which may be used as context information to be associated with one or more prior, contemporaneous, or subsequent natural language utterances.

[031] Thus, in response to the voice-click module 108 detecting a voice-click (e.g., selection of an icon, a section of text, specific coordinates on a map display, or other information), the voice-click module 108 may signal the system 100 to use a natural language utterance voice input (which may be received via the voice input device 105a) as further context for determining an action, query, command, task, or other request to perform in order to service the detected voice-click. As such, the various natural language processing components in the system 100 may use the combined context of the voice-click and the accompanying natural language utterance to determine the intent of the voice-click device interaction, and to

appropriately route one or more actions, queries, commands, tasks, or other requests to any of the various devices deployed in the multi-device environment.

[032] For instance, in one implementation, the multi-device environment may include a voice-enabled navigation device. Thus, an exemplary voice-click device interaction may include a user touching a stylus to a specific intersection displayed on a touch-screen display 105b associated with the voice-enabled navigation device, while also providing an utterance such as "What restaurants are around here?" into a microphone 105a. In this example, the voice-click module 108 may extract information relating to the voice-clicked intersection, which may be used as context for processing the accompanying utterance (i.e., the selected intersection may provide context for interpreting "around here," as opposed to the user's current location or some other meaning). Moreover, as indicated above, the voice input may be used as additional context for determining task specification. Thus, the utterance may be further processed for recognition and conversational interpretation using the various natural language processing components of the system 100, as described in greater detail below.

[033] In one implementation, an Automatic Speech Recognizer (ASR) 110 may generate one or more preliminary interpretations of the utterance received via the voice input device 105a. For example, the ASR 110 may recognize syllables, words, phrases, or other acoustic characteristics of the utterance using one or more dynamically adaptable recognition grammars. In one implementation, the dynamic recognition grammars may be used to recognize a stream of phonemes using phonetic dictation based on one or more acoustic models (e.g., as described in co-pending U.S. Patent Application 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed August 5, 2005, the disclosure of which is hereby incorporated by reference in its entirety).

[034] In one implementation, the ASR 110 may be configured to perform multi-pass speech recognition, where a primary speech recognition engine may generate a primary transcription of the utterance (e.g., using a large list dictation grammar), and may subsequently request one or more secondary transcriptions from one or more secondary speech recognition engines (e.g., using a virtual dictation grammar having decoy words for out-of-vocabulary words). In one implementation, the primary speech recognition engine may request the secondary transcriptions based on a confidence level for the primary transcription.

[035] The recognition grammars employed in the ASR 110 may include various vocabularies, dictionaries, syllables, words, phrases, or other information for recognizing utterances. In one implementation, information contained in the recognition grammars may be dynamically optimized to improve a likelihood of accurate recognition for a given utterance (e.g., following an incorrect interpretation of a word or phrase, the incorrect interpretation may be removed from the grammar to reduce a likelihood of repeating the incorrect interpretation). Additionally, various forms of knowledge can be used to continually optimize the information contained in the recognition grammars on a dynamic basis. For example, the system 100 may have knowledge that includes environmental knowledge (e.g., peer-to-peer affinities, capabilities of the various devices in the environment, etc.), historical knowledge (e.g., frequent requests, prior context, etc.), or short term shared knowledge relating to a current conversational dialogue or interaction, among other types of knowledge.

[036] In one implementation, the information in the recognition grammars may be further optimized according to context or application-specific domains. In particular, similar utterances may be interpreted differently depending on a context to which the utterance relates, including navigation, music, movies, weather, shopping, news, languages, temporal or geographic proximities, or other contexts or domains. For example, an utterance containing the word "traffic" may be subject to different interpretations depending on whether the context relates to navigation (i.e., road conditions), music (i.e., the 1960's rock band), or movies (i.e., the film directed by Steven Soderbergh). Accordingly, the ASR 110 may use various techniques to generate a preliminary interpretation of the natural language utterance, such as those described in the above-referenced co-pending U.S. Patent Applications and/or co-pending U.S. Patent Application 11/513,269, entitled "Dynamic Speech Sharpening," filed August 31, 2006, the disclosure of which is hereby incorporated by reference in its entirety.

[037] As such, the ASR 110 may provide one or more preliminary interpretations of the natural language utterance included in the voice-click to a conversational language processor 120. The conversational language processor 120 may include various natural language processing components collectively configured to model human-to-human conversations or interactions. For example, the conversational language processor 120 may include, among other things, an intent determination engine 130a, a constellation model 130b, one or more domain agents 130c, a context tracking engine 130d, a misrecognition engine 130e, and a voice

search engine 130f. Furthermore, the conversational language processor 120 may be coupled to one or more data repositories 160 and one or more applications 150 associated with various contexts or domains.

[038] Thus, the system 100 may use the various natural language processing components associated with the conversational language processor 120 in order to engage the user in a cooperative conversation and resolve voice-click device interactions based on the user's intent in initiating the voice-click. More particularly, the intent determination engine 130a may establish meaning for a given multi-modal device interaction based on capabilities of the system 100 as well as the capabilities of any other devices in the multi-device environment. For instance, referring to the example above where the user voice-clicked a particular intersection to determine "What restaurants are around here," the conversational language processor 120 may determine a conversational goal of the voice-click (e.g., "What" may indicate that the utterance relates to a query requesting data retrieval). In addition, the conversational language processor 120 may invoke the context tracking engine 130d to determine context for the voice-click. For example, to determine the voice-click context, the context tracking engine 130d may combine the context associated with the identified point of focus (i.e., the selected intersection) with context associated with the utterance (i.e., restaurants).

[039] As a result, the combined context of the voice-click (which includes both the device interaction and the accompanying utterance) may provide sufficient information for routing a specific query. For example, the query may include various parameters or criteria relating to restaurants and the identified intersection. The conversational language processor 120 may then select a particular device, application, or other component to which the query may be routed for processing. For example, in one implementation, the conversational language processor 120 may evaluate the constellation model 130b, which contains a model of capabilities for each device in the multi-device environment. In one implementation, the constellation model 130b may contain, among other things, knowledge of processing and storage resources available to each of the devices in the environment, as well as the nature and scope of domain agents, context, capabilities, content, services, or other information for each of the devices.

[040] As such, using the constellation model 130b and/or other information, the conversational language processor 120 may determine which device or combination of devices

has suitable capabilities that can be invoked to process a given voice-click device interaction. For instance, again referring to the example given above, the conversational language processor 120 may determine that the context of the voice-click relates to an interaction with a navigation device and therefore route the query for processing using a navigation application 150. Results of the query may then be processed (e.g., to weigh the results based on knowledge of the user, such as a preference for vegetarian restaurants) and returned to the user via an output device 180.

[041] According to various aspects of the invention, Figure 2 illustrates a block diagram of an exemplary method for synchronizing multi-modal devices in a natural language voice service environment. As described above, multi-modal device interactions (or "voice-clicks") may generally occur when a user engages in one or more interactions with one or more multi-modal devices while also providing one or more natural language utterances that relate to the interactions with the multi-modal devices. In one implementation, context information relating to the interactions with the multi-modal devices may be combined with context information relating to the natural language utterances to determine an intent of the voice-click (e.g., to initiate a particular action, query, command, task, or other request).

[042] In one implementation, various natural language processing components may be configured to continually listen or otherwise monitor the multi-modal devices to determine when voice-clicks occur. As such, the method illustrated in Figure 2 may be used to calibrate or otherwise configure the components responsible for continually listening or otherwise monitoring the multi-modal devices. For example, in one implementation, the natural language voice service environment may include a plurality of multi-modal devices that provide different capabilities or services, and the user may engage in one or more voice-clicks to request services relating to any of the various devices or capabilities in any given device interaction.

[043] To enable continuous listening for multi-modal device interactions or voice-clicks, each of the plurality of devices in the environment may be configured to receive information relating to a voice-click. Thus, in one implementation, an operation 210 may include establishing device listeners for each of the plurality of devices in the environment. Additionally, operation 210 may be performed in response to one or more new devices being added to the environment. The device listeners established in operation 210 may include any suitable combination of instructions, firmware, or other routines configured to be executed on one or more processing

devices or other hardware components. For each device in the environment, the associated device listener may communicate with the device to determine capabilities, features, supported domains, or other information associated with the device. In one implementation, the device listeners may be configured to communicate with the device using the Universal Plug and Play protocol designed for ancillary computer devices. It will be apparent, however, that any suitable mechanism for communicating with the multi-modal devices may be used.

[044] When device listeners have been established for each of the devices in the environment (or when device listeners have been established for new devices added to the environment), the various device listeners may be synchronized in an operation 220. In particular, each of the various devices may have different internal clocks or other timing mechanisms, wherein operation 220 may include synchronizing the various device listeners according to the respective internal clocks or timing mechanisms of the devices. In one implementation, synchronizing the device listeners may include each of the respective device listeners publishing information relating to the internal clock or timing of the associated device.

[045] Thus, when one or more multi-modal interactions or voice-clicks subsequently occur for one or more of the devices, the associated device listener may detect information associated with the voice-click in an operation 230. For example, in one implementation, the various device listeners established in operation 210 may be associated with the voice-click module described above and illustrated in Figure 1. Operation 230 may therefore include one or more of the device listeners or the voice-click module detecting an occurrence of a user interacting with one or more of the devices (e.g., selecting data associated with the device, identifying a point of focus or attention focus associated with the device, or otherwise engaging in one or more interactions or sequences of interactions with the device). Furthermore, an operation 240 may then comprise capturing an utterance from the user that relates to the device interaction detected in operation 230.

[046] For example, a user viewing a web page presented on a display device may see a product name on the web page and desire more information about purchasing the product. The user may select text from the web page that contains the product name (e.g., using a mouse or keyboard to highlight the text), and then initiate a voice-click to ask "Is this available on Amazon.com?" In this example, operation 230 may include a device listener associated with the display device detecting the selection of the text associated with the product name, while

operation 240 may include capturing the utterance inquiring about the availability of the product on Amazon.com.

[047] As described above, each device that receives an input from the user may have an internal clock or timing mechanism. In an operation 250, each device may therefore determine when the input was received from a local perspective, and notify the voice-click module that the input was received. In particular, a given voice-click may include at least a natural language utterance received via a voice input device in addition to one or more further interactions with one or more other devices. The utterance may be received prior to, contemporaneously with, or subsequent to the device interactions, whereby operation 250 includes determining the timing of the device interaction for correlation with the associated utterance. In particular, using the device listener signals synchronized as described in reference to operation 220, an operation 260 may include aligning the signals for the device interactions and the utterance. In matching the device interaction and utterance signals, a voice-click input may be generated that includes aligned voice and non-voice components. The voice-click input may then be subject to further natural language processing, as described in greater detail below.

[048] According to various aspects of the invention, Figure 3 illustrates a flow diagram of an exemplary method for processing multi-modal device interactions in a natural language voice service environment. As described above, multi-modal device interactions (or "voice-clicks") may generally occur when a user interacts with one or more multi-modal devices while also providing one or more natural language utterances related to the device interactions. As such, in one implementation, the method illustrated in Figure 3 may be performed when one or more natural language processing components continually listen or otherwise monitor the one or more multi-modal devices to determine when one or more voice-clicks occur.

[049] In one implementation, one or more device interactions may be defined as initiating a voice-click. For example, any given electronic device may generally support various distinguishable interactions, which may result in a given action, command, query, or other request being executed. Thus, in one implementation, any appropriate combination of device interactions that a given electronic device can uniquely recognize or otherwise use to generate a uniquely recognizable signal may be defined as a voice-click, wherein the voice-click may provide a signal indicating that a natural language utterance is to be processed together with context associated with the relevant device interactions. For example, devices that have a four-

way or a five-way navigation button may support specific distinguishable interactions, wherein pressing the navigation button in a particular way may cause a specific task or other action to be performed, such as controlling a map display or calculating a route. In another example, a BlackBerry® device having a wheel selector may support interactions such as wheeling a cursor over a particular point of focus or attention focus, pressing the wheel to select specific data or a given application, or various other interactions. Various other device interactions may be used to indicate when a natural language utterance is to be processed together with context associated with the device interactions, without limitation, wherein the specific device interactions may vary in any given implementation. For example, the relevant device interactions may further include one or more of gesturing with a pointing or drawing instrument on a touch-sensitive screen (e.g., drawing an ear-shaped squiggle), a unique interaction method such as a long touch or double-tap, and/or if the system is operating in the continuous listening mode described above, a predefined contextual command word may indicate that a current device context is to be processed with a portion of the voice-based input that follows the contextual command word (e.g., the command word be "OK," "Please," "Computer," or another suitable word, wherein a user may select a specific point on a map and say "Please zoom in," or say "OK read it" when an e-mail is displayed, etc.).

[050] As such, an operation 310 may include processing multi-modal device interactions in a natural language voice service environment to detect an occurrence of one or more device interaction that signal initiation of a voice-click. In particular, the device interactions detected in operation 310 may include any suitable interactions that cause an electronic device to generate a unique, recognizable, or otherwise distinguishable signal that relates to a user selecting data, identifying a point of focus or attention focus, invoking an application or task, or interacting with the device in another way, depending on the specific capabilities of the device.

[051] In addition to the specific signal that the device generates in response to the user interaction, the interaction detected in operation 310 may signal initiation of a voice-click, whereby a prior, contemporaneous, or subsequent natural language voice input will provide further context for interpreting the device interaction detected in operation 310. For example, a natural language processing system may generally be configured to accept a voice input when particular device interactions occur (e.g., pressing a button to turn on a microphone). Thus, in the method illustrated in Figure 3, the device interactions that signal an incoming voice input

may further include any suitable interaction or combination of interactions with an electronic device, including interactions associated with a user selecting data, identifying a point of focus or attention focus, invoking an application or task, or interacting with the device in another way, depending on the specific capabilities of the device.

[052] As such, when a voice-click device interaction has been detected in operation 310, a voice-click signal may be generated in an operation 320 to indicate that a natural language voice input should be associated with the interaction detected in operation 320. Subsequently, an operation 330 may include capturing a user utterance to be associated with the interaction detected in operation 310. In one implementation, the interaction detected in operation 310 may signal that a subsequent voice input will be provided, but it will be apparent that in one or more implementations the utterance captured in operation 330 may be provided prior to or contemporaneously with the interaction detected in operation 310 (e.g., a user may provide an utterance such as "Look up this artist on iTunes®" and subsequently voice-click the artist's name on a media player, or the user may provide the utterance while voice-clicking the artist's name, or the user may voice-click the artist's name and then provide the utterance).

[053] When information relating to the voice-click device interactions and associated natural language utterance have been received, an operation 340 may include extracting and combining context information for the device interactions and the associated utterance. In particular, the context information extracted from the voice-click device interactions may include information relating to a segment, item, point of focus, attention focus, or data selected by the user, or to a particular device interaction or sequence of device interactions engaged in by the user. The extracted context for the device interaction may then be combined with context extracted for the natural language utterance captured in operation 330, wherein the combined context information may be used to determine an intent of the voice-click in an operation 350.

[054] For example, in an exemplary voice-click device interaction, a user may be selectively copying a music collection from a media player to a backup storage device. While browsing the music on the media player, the user may encounter a particular song and voice-click the song while saying "Copy this whole album" (e.g., by pressing a particular button on the media player for an extended period of time while highlighting the song). In this example, operation 310 may include detecting the interaction of the extended button press, which causes the voice-click

signal to be generated in operation 320. The utterance of "Copy this whole album" may then be captured in operation 330, and context information associated with the voice-click device interaction and the utterance may be combined in operation 340. In particular, the context of the device interaction may include information relating to the selected song, among other things (e.g., the context may further include information contained in metadata associated with the song, such as an ID3 tag for music files). Furthermore, the context of the utterance may include information identifying a copy operation and an album that contains the selected song.

[055] As such, context information relating to voice-click interactions with the multi-modal devices may be combined with context information relating to the natural language utterances, whereby an operation 350 may determine the intent of the voice-click interaction. For instance, referring to the example above, the intent determined in operation 350 may include an intent to copy an album that contains the highlighted song from the media player to the backup storage device. Thus, in response to determining the intent of the voice-click in operation 350, one or more requests may be appropriately routed in an operation 360. In the example being discussed herein, operation 360 may include routing one or more requests to the media player to identify all of the data associated with the album that contains the highlighted song, and one or more requests to any appropriate combination of devices that can manage copying of the identified data from the media player to the backup storage device (e.g., a personal computer that interfaces with both the media player and the storage device).

[056] According to various aspects of the invention, Figure 4 illustrates a flow diagram of an exemplary method for processing multi-modal device interactions to generate a transaction lead or "click-through" in a natural language voice service environment. In particular, the method illustrated in Figure 4 may be used to generate a transaction lead or click-through in combination with one or more actions performed in response to one or more voice-click device interactions being detected.

[057] For example, an operation 410 may include detecting one or more voice-click device interactions received from a user, wherein the voice-click device interactions may comprise any suitable combination of one or more device interactions coupled with one or more related natural language utterances. The user's intent in engaging in the voice-click device interactions may then be determined in an operation 420, and a subsequent operation 430 may include routing one or more requests to one or more processing devices based on the determined

intent in order to resolve the voice-click interaction. In one implementation, operations 410, 420, and 430 may be performed in a manner similar to that described above in reference to Figure 2 and Figure 3, whereby signals for the device interactions may be aligned with signals for one or more natural language utterances and context information may be extracted from the signals to determine the intent of the voice-click device interaction.

[058] In addition to routing one or more requests based on the user's intent, the method illustrated in Figure 4 may further comprise generating one or more transaction leads that may result in one or more click-throughs. For example, a click-through may generally refer to an instance of a user clicking or otherwise selecting an electronic advertisement to access one or more services associated with the advertiser. In many electronic systems, click-throughs or click-through rates can provide mechanisms for measuring a user's interactions with an electronic advertisement, which may provide various measurements that advertisers can use to determine an amount to pay an entity that delivers the advertisement to users.

[059] As such, the method illustrated in Figure 4 may generate transaction leads, including advertisements or recommendations, whereby a user's voice-based input combined with a specific device interaction may provide further focus for generating transaction leads. In this manner, the advertisements or recommendations provided to the user may be more relevant to specific information with which the user may be interacting. Furthermore, using natural language cognitive models and shared knowledge relating to a user's preferences may provide further context for targeted transaction leads tailored to the specific user, which may thus be more likely to result in a click-through that can generate payment for a voice services provider.

[060] Thus, in addition to routing one or more requests based on the user's intent in engaging in the voice-click device interaction, an operation 440 may include generating one or more transaction leads based on the determined intent. In particular, based on the combined context of the device interaction and the associated natural language utterance, the transaction leads may be processed in a manner "closer" to the user, in that local voice and non-voice context can be used as state data in any appropriate system that performs targeted advertising. For instance, referring to the example given above where a user selects an intersection displayed on a navigation device while saying "Find restaurants are around here," the transaction leads generated in operation 440 may include one or more advertisements or recommendations for restaurants near the intersection, which may be targeted to the user

based on knowledge of the user's short-term and long-term preferences (e.g., a preferred type of restaurant, a preferred price range, etc.).

[061] The transaction leads may then be presented to the user (e.g., as selectable points on a map display). The user's subsequent multi-modal device interactions may then be monitored in an operation 450 to determine whether or when one or more further multi-modal device interactions occur. If no further interactions occur, a determination may be made that the user did not act on the transaction lead and the process may then end. If an additional multi-modal interaction does occur, however, the multi-modal input may be processed in an operation 480 to determine an intent of the input and route one or more requests accordingly. In addition, an operation 460 may include determining whether the multi-modal input relates to the transaction lead generated in operation 440. For example, the user may select one of the advertised or recommended restaurants by providing an utterance, a non-voice device interaction, or a voice-click device interaction requesting further action or information relating to the transaction lead. In such a case, an operation 470 may further include processing a transaction click-through in relation to the transaction lead generated in operation 440, wherein the transaction click-through may be used to determine payment or otherwise generate revenue for a particular entity (e.g., a provider of the voice services or another entity associated with the transaction lead or transaction click-through).

[062] Implementations of the invention may be made in hardware, firmware, software, or various combinations thereof. The invention may also be implemented as instructions stored on a machine-readable medium, which may be read and executed by one or more processors. A machine-readable medium may include various mechanisms for storing or transmitting information in a form readable by a machine (e.g., a computing device). For example, a machine-readable storage medium may include read only memory, random access memory, magnetic disk storage media, optical storage media, flash memory devices, or other storage media, and a machine-readable transmission media may include forms of propagated signals, such as carrier waves, infrared signals, digital signals, or other transmission media. Further, firmware, software, routines, or instructions may be described in the above disclosure in terms of specific exemplary aspects and implementations of the invention, and performing certain actions. However, it will be apparent that such descriptions are merely for convenience, and

that such actions in fact result from computing devices, processors, controllers, or other devices executing the firmware, software, routines, or instructions.

[063] Although the descriptions provided herein have generally focused on techniques for processing multi-modal device interactions in a natural language voice services environment, it will be apparent that various further natural language processing capabilities may be used in connection with, in addition to, or in lieu of those described in connection with the specific aspects and implementations discussed herein. For example, in addition to the techniques described in the co-pending U.S. Patent Applications referenced above, the systems and methods described herein may further utilize natural language processing capabilities as described in co-pending U.S. Patent Application 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed August 5, 2005, U.S. Patent Application 11/200,164, entitled "System and Method of Supporting Adaptive Misrecognition in Conversational Speech," filed August 10, 2005, U.S. Patent Application 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," filed August 29, 2005, U.S. Patent Application 11/580,926, entitled "System and Method for a Cooperative Conversational Voice User Interface," filed October 16, 2006, U.S. Patent Application 11/671,526, entitled "System and Method for Selecting and Presenting Advertisements Based on Natural Language Processing of Voice-Based Input," filed February 6, 2007, and U.S. Patent Application 11/954,064, entitled "System and Method for Providing a Natural Language Voice User Interface in an Integrated Voice Navigation Services Environment," filed December 11, 2007, the disclosures of which are hereby incorporated by reference in their entirety.

[064] Accordingly, aspects and implementations of the invention may be described herein as including a particular feature, structure, or characteristic, but it will be apparent that every aspect or implementation may or may not necessarily include the particular feature, structure, or characteristic. In addition, when a particular feature, structure, or characteristic has been described in connection with a given aspect or implementation, it will be understood that such feature, structure, or characteristic may be included in connection with other aspects or implementations, whether or not explicitly described. Thus, various changes and modifications may be made to the preceding description without departing from the scope or spirit of the

invention, and the specification and drawings should therefore be regarded as exemplary only, with the scope of the invention determined solely by the appended claims.

CLAIMS

What is claimed is:

1. A method for processing one or more multi-modal device interactions in a natural language voice services environment that includes one or more electronic devices, comprising:

detecting at least one multi-modal device interaction, wherein the multi-modal device interaction includes a non-voice interaction with at least one of the electronic devices or an application associated with at least one of the electronic devices, and wherein the multi-modal device interaction further includes at least one natural language utterance relating to the non-voice interaction;

extracting context information relating to the multi-modal device interaction, wherein the extracted context information includes context relating to the non-voice interaction, and wherein the extracted context information further include context relating to the natural language utterance;

combining the context relating to the non-voice interaction and the context relating to the natural language utterance;

determining an intent of the multi-modal device interaction based on the combined context relating to the non-voice interaction and the natural language utterance; and

routing at least one request to one or more of the electronic devices based on the determined intent of the multi-modal device interaction.

2. The method of claim 1, wherein at least one of the electronic devices includes an input device configured to receive the natural language utterance.

3. The method of claim 2, further comprising signaling the input device to capture the natural language utterance in response to the non-voice interaction being detected.

4. The method of claim 3, further comprising:

establishing one or more device listeners in the natural language voice services environment, the device listeners configured to detect the non-voice interaction; and

aligning information relating to the non-voice interaction detected by the device listeners and the natural language utterance captured by the input device.

5. The method of claim 1, further comprising:

generating at least one transaction lead based on the determined intent of the multi-modal device interaction;

receiving at least one further multi-modal device interaction relating to the generated transaction lead; and

processing a transaction click-through in response to receiving the multi-modal device interaction relating to the generated transaction lead.

6. The method of claim 5, wherein the generated transaction lead includes at least one of an advertisement or a recommendation relating to the determined intent of the multi-modal device interaction.

7. The method of claim 1, wherein the non-voice interaction comprises a selection of a segment, item, data, or application associated with one or more of the electronic devices.

8. The method of claim 1, wherein the non-voice interaction comprises an identification of a point of focus or an attention focus associated with one or more of the electronic devices.

9. The method of claim 1, wherein the non-voice interaction comprises one or more unique and distinguishable interactions associated with one or more of the electronic devices.

10. A system for processing one or more multi-modal device interactions in a natural language voice services environment that includes one or more electronic devices, wherein the system comprises one or more processing devices configured to:

detect at least one multi-modal device interaction, wherein the multi-modal device interaction includes a non-voice interaction with at least one of the electronic devices or an application associated with at least one of the electronic devices, and wherein the multi-modal device interaction further includes at least one natural language utterance relating to the non-voice interaction;

extract context information relating to the multi-modal device interaction, wherein the extracted context information includes context relating to the non-voice interaction, and wherein the extracted context information further include context relating to the natural language utterance;

combine the context relating to the non-voice interaction and the context relating to the natural language utterance;

determine an intent of the multi-modal device interaction based on the combined context relating to the non-voice interaction and the natural language utterance; and

route at least one request to one or more of the electronic devices based on the determined intent of the multi-modal device interaction.

11. The system of claim 10, wherein at least one of the electronic devices includes an input device configured to receive the natural language utterance.

12. The system of claim 11, the processing devices further configured to signal the input device to capture the natural language utterance in response to the non-voice interaction being detected.

13. The system of claim 12, the processing devices further configured to:

establish one or more device listeners in the natural language voice services environment, the device listeners configured to detect the non-voice interaction; and

align information relating to the non-voice interaction detected by the device listeners and the natural language utterance captured by the input device.

14. The system of claim 10, the processing devices further configured to:

at least one transaction lead based on the determined intent of the multi-modal device interaction;

receive at least one further multi-modal device interaction relating to the generated transaction lead; and

process a transaction click-through in response to receiving the multi-modal device interaction relating to the generated transaction lead.

15. The system of claim 14, wherein the generated transaction lead includes at least one of an advertisement or a recommendation relating to the determined intent of the multi-modal device interaction.

16. The system of claim 10, wherein the non-voice interaction comprises a selection of a segment, item, data, or application associated with one or more of the electronic devices.

17. The system of claim 10, wherein the non-voice interaction comprises an identification of a point of focus or an attention focus associated with one or more of the electronic devices

18. The system of claim 10, wherein the non-voice interaction comprises one or more unique and distinguishable interactions associated with one or more of the electronic devices.

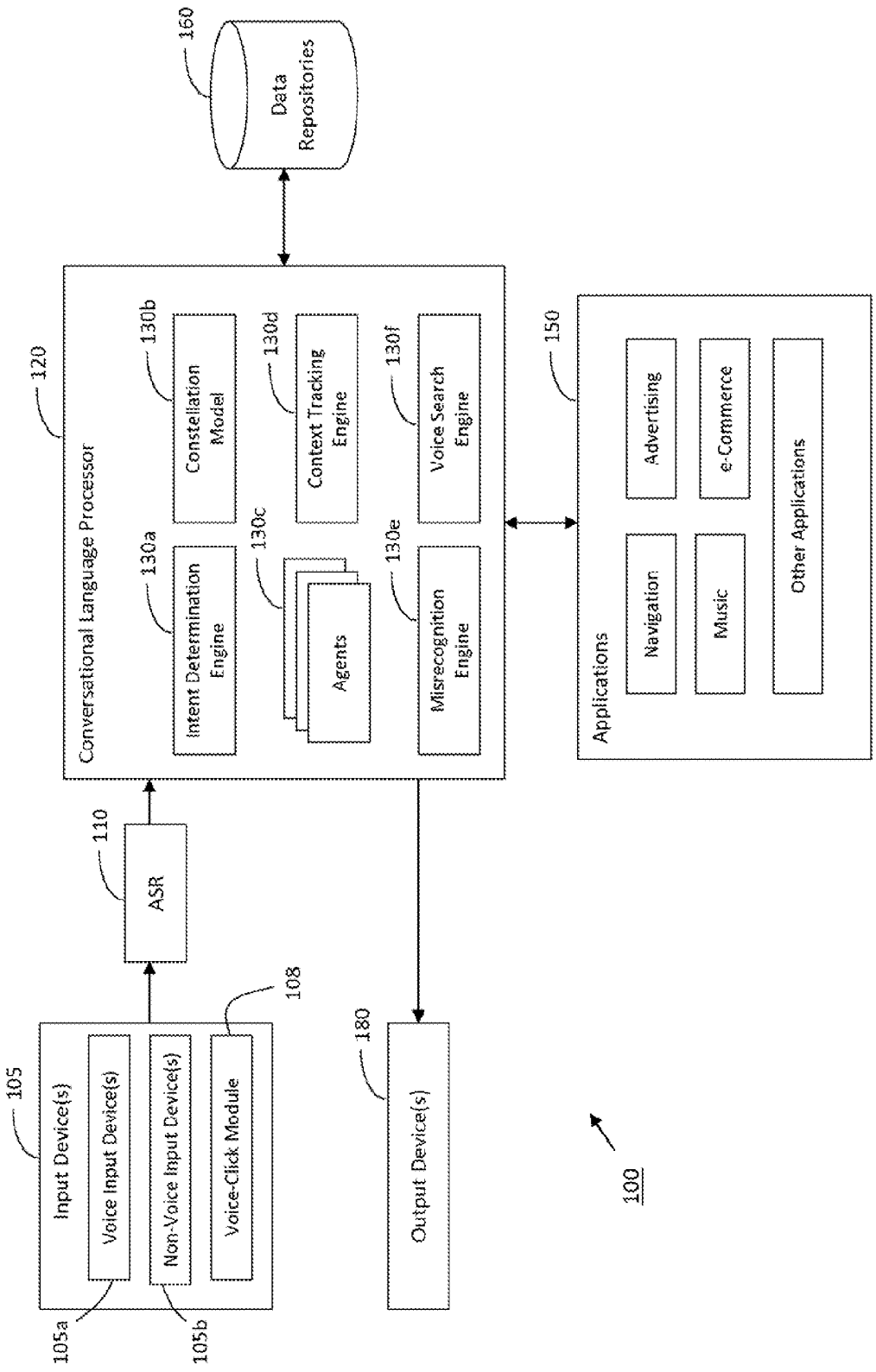

160
Data Repositories
120
Conversational Language Processor
130b
Constellation Model
130d
Context Tracking Engine
130f
Voice Search Engine
130a
Intent Determination Engine
130c
Agents
130e
Misrecognition Engine
150
Applications
Advertising
e-Commerce
Navigation
Music
Other Applications
110
ASR
105
Input Device(s)
Voice Input Device(s)
Non-Voice Input Device(s)
Voice-Click Module
108
105a
105b
180
Output Device(s)
100

Figure 1

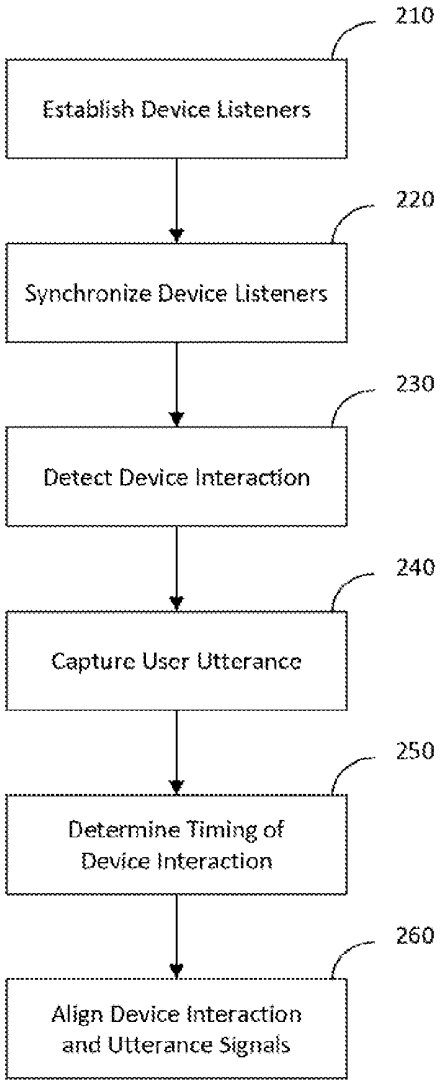

210
Establish Device Listeners
220
Synchronize Device Listeners
230
Detect Device Interaction
240
Capture User Utterance
250
Determine Timing of Device Interaction
260
Align Device Interaction and Utterance Signals

Figure 2

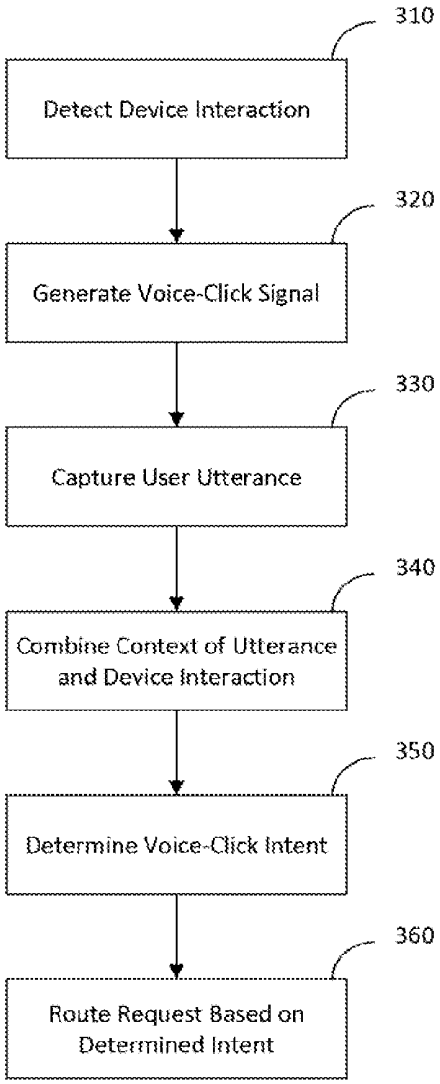

310
Detect Device Interaction
320
Generate Voice-Click Signal
330
Capture User Utterance
340
Combine Context of Utterance and Device Interaction
350
Determine Voice-Click Intent
360
Route Request Based on Determined Intent

Figure 3

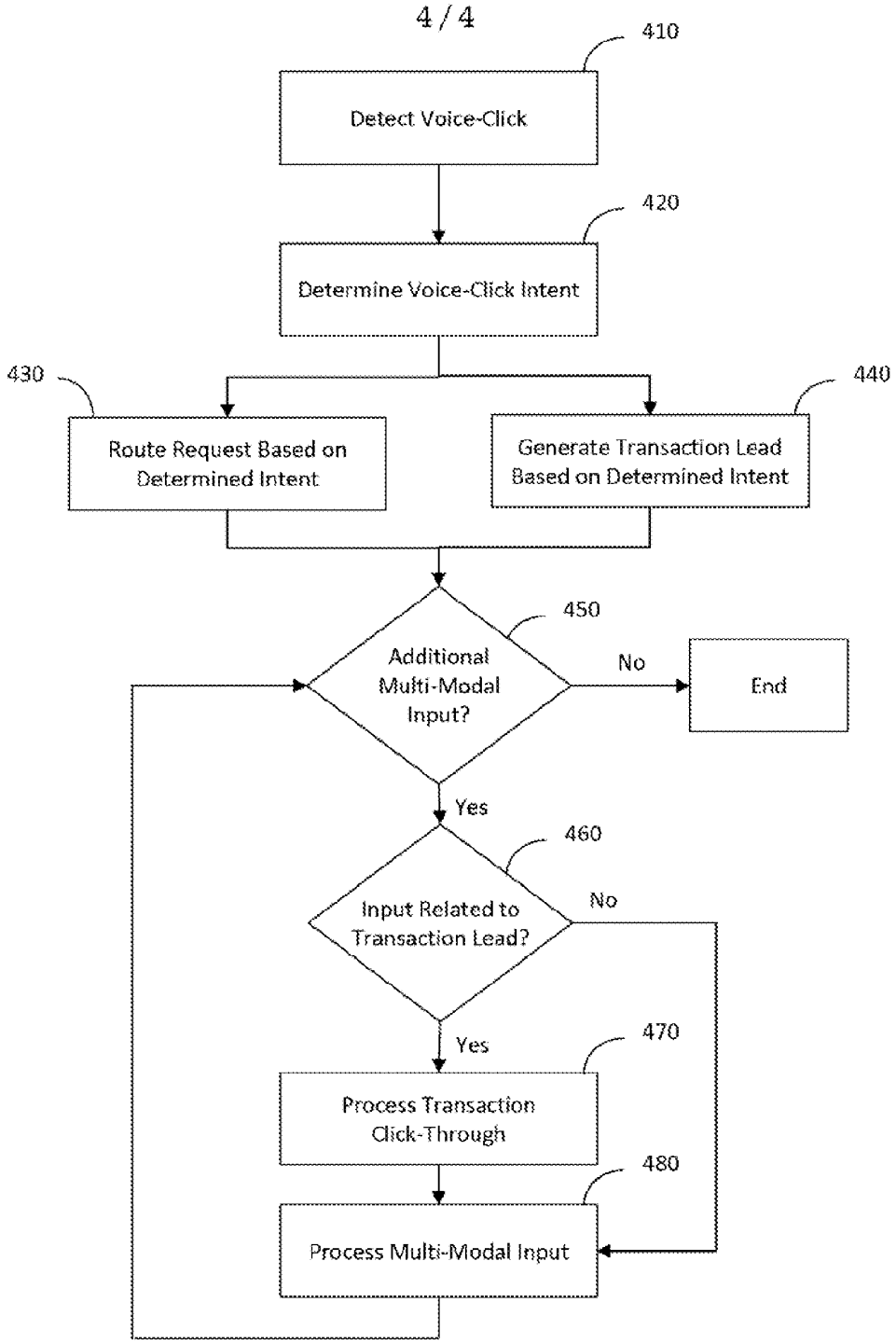

410
Detect Voice-Click
420
Determine Voice-Click Intent
430
Route Request Based on Determined Intent
440
Generate Transaction Lead Based on Determined Intent
450
Additional Multi-Modal Input?
No
End
Yes
460
Input Related to Transaction Lead?
No
Yes
470
Process Transaction Click-Through
480
Process Multi-Modal Input

Figure 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 10/24895

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G10L 15/00 (2010.01)
USPC - 704/257
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC: 704/257

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 704/270.1; 704/200; 704/9 (search term limited)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Electronic Databases Searched: pubWEST(PGPB,USPT,USOC,EPAB,JPAB); googleScholar
Search Terms Used: natural language, multimodal, speech/voice identification/recognition, voice/keyboard/mouse click, context/intent, marketing/advertising

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US 2007/0050191 A1 (WEIDER et al.) 01 March 2007 (01.03.2007) Entire document, especially: para [0001], [0016], [0018], [0020], [0031], [0043], [0116], [0133]; Fig. 7, 8	1-2, 7-11 and 16-18 --- 3-6 and 12-15
Y	US 2007/0043574 A1 (COFFMAN et al.) 22 February 2007 (22.02.2007) Entire document, especially: Abstract, para [0154], [0227]	3, 4, 12 and 13
Y	US 2008/0189110 A1 (FREEMAN et al.) 07 August 2008 (07.08.2008) Entire document, especially: para [0006], [0013], [0024], [0028] and Fig. 3	5, 6, 14 and 15
A	US 2008/0235027 A1 (CROSS) 25 September 2008 (25.09.2008)	1-18
A	US 2008/0091406 A1 (BALDWIN et al.) 17 April 2008 (14.04.2008)	1-18

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28 March 2010 (28.03.2010)	13 APR 2010
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (July 2009)

Electronic Patent Application Fee Transmittal

Application Number:	11580926
Filing Date:	16-Oct-2006
Title of Invention:	System and method for a cooperative conversational voice user interface
First Named Inventor/Applicant Name:	Larry Baldwin
Filer:	Syed Jafar Ali/Miriam E. Nelsen
Attorney Docket Number:	090611-0357945

Filed as Small Entity

Utility under 35 USC 111(a) Filing Fees

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Independent claims in excess of 3	2201	3	110	330
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Miscellaneous:				
Request for continued examination	2801	1	405	405
		Total in USD ($)		**735**

Electronic Acknowledgement Receipt

EFS ID:	9283901
Application Number:	11580926
International Application Number:	
Confirmation Number:	6670
Title of Invention:	System and method for a cooperative conversational voice user interface
First Named Inventor/Applicant Name:	Larry Baldwin
Customer Number:	00909
Filer:	Syed Jafar Ali/Miriam E. Nelsen
Filer Authorized By:	Syed Jafar Ali
Attorney Docket Number:	090611-0357945
Receipt Date:	21-JAN-2011
Filing Date:	16-OCT-2006
Time Stamp:	17:31:11
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	Deposit Account
Payment was successfully received in RAM	$735
RAM confirmation Number	4508
Deposit Account	033975
Authorized User	
The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows: Charge any Additional Fees required under 37 C.F.R. Section 1.21 (Miscellaneous fees and charges)	

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		090611-0357945_RCE_Amend_IDS.pdf	1459231 f93b279bb06136b38a8c6c61c2954c6772659cb7	yes	33

Multipart Description/PDF files in .zip description

Document Description	Start	End
Request for Continued Examination (RCE)	1	1
Amendment Submitted/Entered with Filing of CPA/RCE	2	2
Specification	3	6
Claims	7	21
Applicant Arguments/Remarks Made in an Amendment	22	29
Transmittal Letter	30	31
Information Disclosure Statement (IDS) Filed (SB/08)	32	33

Warnings:

Information:

2	Foreign Reference	WO1999046763.pdf	1247910 62eacd3541d472fdc4873f7f386c31e69fd9ea4e	no	28

Warnings:

Information:

3	Foreign Reference	WO2000021232.pdf	3963803 203ae5a44439b38f56ae96cdbaa5a64e9e3c570c	no	77

Warnings:

Information:

4	Foreign Reference	WO2000046792.pdf	1259047 705348c0646506c1bc55d53c27f1ae60013291b1	no	34

Warnings:

Information:

5	Foreign Reference	WO2004072954.pdf	1105346 c4acb3accfd8c87c21d4bf9b2cc1c8dfe629a50c	no	25

Warnings:

Information:					
6	Foreign Reference	WO2007019318.pdf	3196832 ff1325b971941c7a88c2c7871f5785b0154e9307	no	55
Warnings:					
Information:					
7	Foreign Reference	WO2007021587.pdf	3525793 196b60abf639ab219e2472fadc79b2bf65c0ee3a	no	61
Warnings:					
Information:					
8	Foreign Reference	WO2007027546.pdf	3873530 5cc6e95751b76c1f016b92415f2860dd2cc39298	no	74
Warnings:					
Information:					
9	Foreign Reference	WO2007027989.pdf	1432119 6c33baceb03630bcc7fc958ccf493ab880e7e9e3	no	28
Warnings:					
Information:					
10	Foreign Reference	WO2008098039.pdf	1103202 fcab991006a05016a57e05548f6b00f78732c946	no	24
Warnings:					
Information:					
11	Foreign Reference	WO2008118195.pdf	2058896 635503176929e0d666e3e8683aa40f3a9643658e	no	39
Warnings:					
Information:					
12	Foreign Reference	WO2009075912.pdf	3688456 02c65abf9ea5c876fb8e4a654567b258513f4e25	no	69
Warnings:					
Information:					
13	Foreign Reference	WO2009145796.pdf	1934285 e71f0a22dfc12ce94396146b078264c9e5df68d9	no	39
Warnings:					
Information:					
14	Foreign Reference	WO2010096752.pdf	1502381 f4c719245f739ab6223fbd8f94c7b17cd9acf7b7	no	33
Warnings:					

Information:					
15	NPL Documents	BAZZI.pdf	361077 67a2e53341c20cd2f0fecfdfc9ca971d47945b8c	no	4
Warnings:					
Information:					
16	NPL Documents	OSHAUGHNESSY.pdf	4563513 93d11484104e08a3f70ac1884e29bccb188a88f0	no	34
Warnings:					
Information:					
17	Fee Worksheet (PTO-875)	fee-info.pdf	31618 52f65f74cdd5af29c5989e1dae1f8d7b5c67bba3	no	2
Warnings:					
Information:					
			Total Files Size (in bytes):	36307039	

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

<u>New Applications Under 35 U.S.C. 111</u>
If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

<u>National Stage of an International Application under 35 U.S.C. 371</u>
If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

<u>New International Application Filed with the USPTO as a Receiving Office</u>
If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

PTO/SB/06 (07-06)
Approved for use through 1/31/2007. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

PATENT APPLICATION FEE DETERMINATION RECORD Substitute for Form PTO-875	Application or Docket Number 11/580,926	Filing Date 10/16/2006	☐ To be Mailed

APPLICATION AS FILED – PART I

	(Column 1)	(Column 2)	SMALL ENTITY ☒		OR	OTHER THAN SMALL ENTITY	
FOR	NUMBER FILED	NUMBER EXTRA	RATE ($)	FEE ($)		RATE ($)	FEE ($)
☐ BASIC FEE (37 CFR 1.16(a), (b), or (c))	N/A	N/A	N/A			N/A	
☐ SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A	N/A			N/A	
☐ EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A	N/A			N/A	
TOTAL CLAIMS (37 CFR 1.16(i))	minus 20 =	*	X $ =		OR	X $ =	
INDEPENDENT CLAIMS (37 CFR 1.16(h))	minus 3 =	*	X $ =			X $ =	
☐ APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $250 ($125 for small entity) for each additional 50 sheets or fraction thereof. See 35 U.S.C. 41(a)(1)(G) and 37 CFR 1.16(s).						
☐ MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))							
* If the difference in column 1 is less than zero, enter "0" in column 2.			TOTAL			TOTAL	

APPLICATION AS AMENDED – PART II

		(Column 1)		(Column 2)	(Column 3)	SMALL ENTITY		OR	OTHER THAN SMALL ENTITY	
	01/21/2011	CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA	RATE ($)	ADDITIONAL FEE ($)		RATE ($)	ADDITIONAL FEE ($)
AMENDMENT	Total (37 CFR 1.16(i))	* 42	Minus	** 42	= 0	X $26 =	0	OR	X $ =	
	Independent (37 CFR 1.16(h))	* 3	Minus	***3	= 0	X $110 =	0	OR	X $ =	
	☐ Application Size Fee (37 CFR 1.16(s))									
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE	0	OR	TOTAL ADD'L FEE	

		(Column 1)		(Column 2)	(Column 3)					
		CLAIMS REMAINING AFTER AMENDMENT		HIGHEST NUMBER PREVIOUSLY PAID FOR	PRESENT EXTRA	RATE ($)	ADDITIONAL FEE ($)		RATE ($)	ADDITIONAL FEE ($)
AMENDMENT	Total (37 CFR 1.16(i))	*	Minus	**	=	X $ =		OR	X $ =	
	Independent (37 CFR 1.16(h))	*	Minus	***	=	X $ =		OR	X $ =	
	☐ Application Size Fee (37 CFR 1.16(s))									
	☐ FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM (37 CFR 1.16(j))							OR		
						TOTAL ADD'L FEE		OR	TOTAL ADD'L FEE	

* If the entry in column 1 is less than the entry in column 2, write "0" in column 3.

** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 20, enter "20".

*** If the "Highest Number Previously Paid For" IN THIS SPACE is less than 3, enter "3".

The "Highest Number Previously Paid For" (Total or Independent) is the highest number found in the appropriate box in column 1.

Legal Instrument Examiner:
/MINNIE JACKSON/

This collection of information is required by 37 CFR 1.16. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

NOTICE OF ALLOWANCE AND FEE(S) DUE

909 7590 08/15/2011
PILLSBURY WINTHROP SHAW PITTMAN, LLP
P.O. BOX 10500
MCLEAN, VA 22102

EXAMINER	
YEN, ERIC L	
ART UNIT	PAPER NUMBER
2626	

DATE MAILED: 08/15/2011

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
11/580,926	10/16/2006	Larry Baldwin	090611-0357945	6670

TITLE OF INVENTION: SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

APPLN. TYPE	SMALL ENTITY	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	YES	$755	$300	$0	$1055	11/15/2011

THE APPLICATION IDENTIFIED ABOVE HAS BEEN EXAMINED AND IS ALLOWED FOR ISSUANCE AS A PATENT. PROSECUTION ON THE MERITS IS CLOSED. THIS NOTICE OF ALLOWANCE IS NOT A GRANT OF PATENT RIGHTS. THIS APPLICATION IS SUBJECT TO WITHDRAWAL FROM ISSUE AT THE INITIATIVE OF THE OFFICE OR UPON PETITION BY THE APPLICANT. SEE 37 CFR 1.313 AND MPEP 1308.

THE ISSUE FEE AND PUBLICATION FEE (IF REQUIRED) MUST BE PAID WITHIN THREE MONTHS FROM THE MAILING DATE OF THIS NOTICE OR THIS APPLICATION SHALL BE REGARDED AS ABANDONED. THIS STATUTORY PERIOD CANNOT BE EXTENDED. SEE 35 U.S.C. 151. THE ISSUE FEE DUE INDICATED ABOVE DOES NOT REFLECT A CREDIT FOR ANY PREVIOUSLY PAID ISSUE FEE IN THIS APPLICATION. IF AN ISSUE FEE HAS PREVIOUSLY BEEN PAID IN THIS APPLICATION (AS SHOWN ABOVE), THE RETURN OF PART B OF THIS FORM WILL BE CONSIDERED A REQUEST TO REAPPLY THE PREVIOUSLY PAID ISSUE FEE TOWARD THE ISSUE FEE NOW DUE.

HOW TO REPLY TO THIS NOTICE:

I. Review the SMALL ENTITY status shown above.

If the SMALL ENTITY is shown as YES, verify your current SMALL ENTITY status:

A. If the status is the same, pay the TOTAL FEE(S) DUE shown above.

B. If the status above is to be removed, check box 5b on Part B - Fee(s) Transmittal and pay the PUBLICATION FEE (if required) and twice the amount of the ISSUE FEE shown above, or

If the SMALL ENTITY is shown as NO:

A. Pay TOTAL FEE(S) DUE shown above, or

B. If applicant claimed SMALL ENTITY status before, or is now claiming SMALL ENTITY status, check box 5a on Part B - Fee(s) Transmittal and pay the PUBLICATION FEE (if required) and 1/2 the ISSUE FEE shown above.

II. PART B - FEE(S) TRANSMITTAL, or its equivalent, must be completed and returned to the United States Patent and Trademark Office (USPTO) with your ISSUE FEE and PUBLICATION FEE (if required). If you are charging the fee(s) to your deposit account, section "4b" of Part B - Fee(s) Transmittal should be completed and an extra copy of the form should be submitted. If an equivalent of Part B is filed, a request to reapply a previously paid issue fee must be clearly made, and delays in processing may occur due to the difficulty in recognizing the paper as an equivalent of Part B.

III. All communications regarding this application must give the application number. Please direct all communications prior to issuance to Mail Stop ISSUE FEE unless advised to the contrary.

IMPORTANT REMINDER: Utility patents issuing on applications filed on or after Dec. 12, 1980 may require payment of maintenance fees. It is patentee's responsibility to ensure timely payment of maintenance fees when due.

PART B - FEE(S) TRANSMITTAL

Complete and send this form, together with applicable fee(s), to: Mail **Mail Stop ISSUE FEE**
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450
or Fax (571)-273-2885

INSTRUCTIONS: This form should be used for transmitting the ISSUE FEE and PUBLICATION FEE (if required). Blocks 1 through 5 should be completed where appropriate. All further correspondence including the Patent, advance orders and notification of maintenance fees will be mailed to the current correspondence address as indicated unless corrected below or directed otherwise in Block 1, by (a) specifying a new correspondence address; and/or (b) indicating a separate "FEE ADDRESS" for maintenance fee notifications.

CURRENT CORRESPONDENCE ADDRESS (Note: Use Block 1 for any change of address)

909 7590 08/15/2011
PILLSBURY WINTHROP SHAW PITTMAN, LLP
P.O. BOX 10500
MCLEAN, VA 22102

Note: A certificate of mailing can only be used for domestic mailings of the Fee(s) Transmittal. This certificate cannot be used for any other accompanying papers. Each additional paper, such as an assignment or formal drawing, must have its own certificate of mailing or transmission.

Certificate of Mailing or Transmission
I hereby certify that this Fee(s) Transmittal is being deposited with the United States Postal Service with sufficient postage for first class mail in an envelope addressed to the Mail Stop ISSUE FEE address above, or being facsimile transmitted to the USPTO (571) 273-2885, on the date indicated below.

(Depositor's name)
(Signature)
(Date)

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
11/580,926	10/16/2006	Larry Baldwin	090611-0357945	6670

TITLE OF INVENTION: SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

APPLN. TYPE	SMALL ENTITY	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	YES	$755	$300	$0	$1055	11/15/2011

EXAMINER	ART UNIT	CLASS-SUBCLASS
YEN, ERIC L	2626	704-009000

1. Change of correspondence address or indication of "Fee Address" (37 CFR 1.363).
- ❑ Change of correspondence address (or Change of Correspondence Address form PTO/SB/122) attached.
- ❑ "Fee Address" indication (or "Fee Address" Indication form PTO/SB/47; Rev 03-02 or more recent) attached. **Use of a Customer Number is required.**

2. For printing on the patent front page, list
(1) the names of up to 3 registered patent attorneys or agents OR, alternatively,
(2) the name of a single firm (having as a member a registered attorney or agent) and the names of up to 2 registered patent attorneys or agents. If no name is listed, no name will be printed.
1 _____
2 _____
3 _____

3. ASSIGNEE NAME AND RESIDENCE DATA TO BE PRINTED ON THE PATENT (print or type)

PLEASE NOTE: Unless an assignee is identified below, no assignee data will appear on the patent. If an assignee is identified below, the document has been filed for recordation as set forth in 37 CFR 3.11. Completion of this form is NOT a substitute for filing an assignment.

(A) NAME OF ASSIGNEE (B) RESIDENCE: (CITY and STATE OR COUNTRY)

Please check the appropriate assignee category or categories (will not be printed on the patent) : ❑ Individual ❑ Corporation or other private group entity ❑ Government

4a. The following fee(s) are submitted:
- ❑ Issue Fee
- ❑ Publication Fee (No small entity discount permitted)
- ❑ Advance Order - # of Copies _____

4b. Payment of Fee(s): **(Please first reapply any previously paid issue fee shown above)**
- ❑ A check is enclosed.
- ❑ Payment by credit card. Form PTO-2038 is attached.
- ❑ The Director is hereby authorized to charge the required fee(s), any deficiency, or credit any overpayment, to Deposit Account Number _____ (enclose an extra copy of this form).

5. **Change in Entity Status** (from status indicated above)
- ❑ a. Applicant claims SMALL ENTITY status. See 37 CFR 1.27.
- ❑ b. Applicant is no longer claiming SMALL ENTITY status. See 37 CFR 1.27(g)(2).

NOTE: The Issue Fee and Publication Fee (if required) will not be accepted from anyone other than the applicant; a registered attorney or agent; or the assignee or other party in interest as shown by the records of the United States Patent and Trademark Office.

Authorized Signature _____ Date _____

Typed or printed name _____ Registration No. _____

This collection of information is required by 37 CFR 1.311. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, Virginia 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, Virginia 22313-1450.

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
11/580,926	10/16/2006	Larry Baldwin	090611-0357945	6670

909 7590 08/15/2011

PILLSBURY WINTHROP SHAW PITTMAN, LLP
P.O. BOX 10500
MCLEAN, VA 22102

EXAMINER	
YEN, ERIC L	
ART UNIT	PAPER NUMBER
2626	

DATE MAILED: 08/15/2011

Determination of Patent Term Adjustment under 35 U.S.C. 154 (b)
(application filed on or after May 29, 2000)

The Patent Term Adjustment to date is 999 day(s). If the issue fee is paid on the date that is three months after the mailing date of this notice and the patent issues on the Tuesday before the date that is 28 weeks (six and a half months) after the mailing date of this notice, the Patent Term Adjustment will be 999 day(s).

If a Continued Prosecution Application (CPA) was filed in the above-identified application, the filing date that determines Patent Term Adjustment is the filing date of the most recent CPA.

Applicant will be able to obtain more detailed information by accessing the Patent Application Information Retrieval (PAIR) WEB site (http://pair.uspto.gov).

Any questions regarding the Patent Term Extension or Adjustment determination should be directed to the Office of Patent Legal Administration at (571)-272-7702. Questions relating to issue and publication fee payments should be directed to the Customer Service Center of the Office of Patent Publication at 1-(888)-786-0101 or (571)-272-4200.

Privacy Act Statement

The Privacy Act of 1974 (P.L. 93-579) requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (i.e., GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

Notice of Allowability	Application No.	Applicant(s)	
	11/580,926	BALDWIN ET AL.	
	Examiner	Art Unit	
	ERIC YEN	2626	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address--

All claims being allowable, PROSECUTION ON THE MERITS IS (OR REMAINS) CLOSED in this application. If not included herewith (or previously mailed), a Notice of Allowance (PTOL-85) or other appropriate communication will be mailed in due course. **THIS NOTICE OF ALLOWABILITY IS NOT A GRANT OF PATENT RIGHTS.** This application is subject to withdrawal from issue at the initiative of the Office or upon petition by the applicant. See 37 CFR 1.313 and MPEP 1308.

1. ☒ This communication is responsive to *Amendment filed 1/21/11*.

2. ☒ The allowed claim(s) is/are *1-42*.

3. ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
   a) ☐ All b) ☐ Some* c) ☐ None of the:
      1. ☐ Certified copies of the priority documents have been received.
      2. ☐ Certified copies of the priority documents have been received in Application No. _____ .
      3. ☐ Copies of the certified copies of the priority documents have been received in this national stage application from the International Bureau (PCT Rule 17.2(a)).

   * Certified copies not received: _____.

Applicant has THREE MONTHS FROM THE "MAILING DATE" of this communication to file a reply complying with the requirements noted below. Failure to timely comply will result in ABANDONMENT of this application.
THIS THREE-MONTH PERIOD IS NOT EXTENDABLE.

4. ☐ A SUBSTITUTE OATH OR DECLARATION must be submitted. Note the attached EXAMINER'S AMENDMENT or NOTICE OF INFORMAL PATENT APPLICATION (PTO-152) which gives reason(s) why the oath or declaration is deficient.

5. ☐ CORRECTED DRAWINGS (as "replacement sheets") must be submitted.
   (a) ☐ including changes required by the Notice of Draftsperson's Patent Drawing Review (PTO-948) attached
      1) ☐ hereto or 2) ☐ to Paper No./Mail Date _____.
   (b) ☐ including changes required by the attached Examiner's Amendment / Comment or in the Office action of Paper No./Mail Date _____.

   Identifying indicia such as the application number (see 37 CFR 1.84(c)) should be written on the drawings in the front (not the back) of each sheet. Replacement sheet(s) should be labeled as such in the header according to 37 CFR 1.121(d).

6. ☐ DEPOSIT OF and/or INFORMATION about the deposit of BIOLOGICAL MATERIAL must be submitted. Note the attached Examiner's comment regarding REQUIREMENT FOR THE DEPOSIT OF BIOLOGICAL MATERIAL.

Attachment(s)

1. ☒ Notice of References Cited (PTO-892)
2. ☐ Notice of Draftperson's Patent Drawing Review (PTO-948)
3. ☒ Information Disclosure Statements (PTO/SB/08), Paper No./Mail Date _____
4. ☐ Examiner's Comment Regarding Requirement for Deposit of Biological Material
5. ☐ Notice of Informal Patent Application
6. ☐ Interview Summary (PTO-413), Paper No./Mail Date _____ .
7. ☐ Examiner's Amendment/Comment
8. ☒ Examiner's Statement of Reasons for Allowance
9. ☐ Other _____.

DETAILED ACTION

Response to Amendment

1. In response to the Final Office Action mailed 11/1/10, applicant has submitted an amendment and Request for Continued Examination filed 1/21/11.

Claims 1-12, 14-26, 28-40, 42, have been amended.

Allowable Subject Matter

1. Claims 1-42 are allowed.

As per Claims 1, 15, 29:

The prior art does not teach the combination of limitations including determining context from short and long term knowledge, establishing an intended meaning for words in an utterance that have different meanings in different contexts, establishing the intended meaning within the identified context to disambiguate an intent that the user had in speaking the one or more words that have different meanings in different contexts, and generating a response that is grammatically or syntactically adapted based on the intended meaning established within the identified context identified by long and short-term shared knowledge about the user.

Kargman (US 2005/0015256) teaches a method (and corresponding computer-readable medium of Claim 15, and system of Claim 29) for providing a cooperativevoice user interface (“customer repeatedly orders… initial or supplemental prompt… individually tailored”, paragraph 50; where the consistent ordering of a particular item is

based on information shared between the user and system [i.e. what does the customer frequently order] to advance the mutual goal of taking down the customer's food order [see applicant's interpretation of "cooperative" in Specification, paragraph 2])

comprising:

receiving an utterance at a voice input device, wherein the voice input device receives the utterance in a current conversation with a user ("calls to place an order", paragraph 50; "speaking the word 'pizza", paragraph 44; "statistically… mushroom… shrimp", paragraph 48; speaking is uttering an input, and the speech itself is an utterance, which occurs during a current conversation when a user speaks with a machine to obtain a pizza order)

accumulating short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received at the voice input device in the current conversation ("whether they wish to order 'pizza'", paragraph 44; wishing to order pizza is knowledge about an utterance in the current conversation where the user speaks that he wants to order a pizza [i.e. short-term because the conversation occurs in the "short-term", i.e. that the conversation is current and not occurring a distant "long-term" time] and this knowledge is "shared" by the user to the machine, and is then information that both the user and machine possesses [i.e. it is "shared])

accumulating long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user ("whether they wish to order 'pizza'", paragraph 44; "order history", paragraph 50;

"prior orders", paragraph 50; "speaking the word 'pizza", paragraph 44; "statistically… mushroom… shrimp", paragraph 47-48; where prior orders are past conversations between the system and customer obtained in the "long-term" [i.e. at a distant/not-current time in the past]; where in Kargman the example for where statistical information is at least obviously accumulated and updated over time [since if people start ordering shrimp consistently it would be frustrating to the users to be constantly overridden by mushroom])

determining an intended meaning for the utterance, wherein a conversational speech engine determines the intended meaning for the utterance from the short-term shared knowledge and the long-term shared knowledge ("prior orders", paragraph 50; "speaking the word 'pizza", paragraph 44; "statistically… mushroom… shrimp", paragraph 48; "statistical database", paragraph 47; where the short term knowledge is the knowledge that the user is ordering some sort of pizza, and using the long-term information can be using the prior order to determine that the user "meant" pepperoni pizza when he said "pizza" because he repeatedly ordered pepperoni. Alternatively since it is obvious to continue to accumulate statistical information just in case things change and people do start ordering shrimp pizza frequently, it is obvious to use the statistics that obviously contain prior user orders as part of the statistics [since information about prior orders are the information represented in the statistics] to determine whether a user meant shrimp or mushroom)

generating a response to the utterance, wherein the conversational speech engine generates the response to the utterance based on the intended meaning

determined from the short-term shared knowledge and the long-term shared knowledge ("speak the order back… for their confirmation", paragraph 49; "tailored", paragraph 50; where the tailored prompt is based on the system determining that the user meant "pepperoni pizza" when the user said "pizza", or alternatively, where speaking the order back is a system response and depends on what the system interpreted the order to be [determined intended meaning] based on the statistics which as discussed above obviously include what the user says over time [to accurately represent the rate at which people order mushroom with pizza]).

Kargman does not, however, perform syntactic and grammatical adaptation of prompts and also does not describe words that have different meanings in different contexts (i.e. only similar-sounding DIFFERENT words, not the same word which has a different meaning and part-of-speech depending on context)

Kroeker et al. (US 2001/0049601) teaches semantic evaluation according to context used to determine the next application prompt (paragraph 36),

Kroeker does not, however, teach syntactic and grammatical adaptation of the prompt itself (i.e. only the selection of which prompt to present based on identified meaning of words in a given context [i.e. semantics]).

Franz et al. (US 6,278,968) teaches syntactically determines the type of words, understands the words semantically, and pragmatically determines the type of response to generate, and generates the response.

Franz similarly does not teach syntactic and grammatical adaptation of the prompts based on semantics (i.e. identified meaning).

An (US 7,536,374), Duan et al. (US 2004/0167771), Brun et al. (US 7,788,084), and Kasravi et al. (US 2007/0118357) generally describe the problems of words that have different parts of speech in different contexts, but also does not describe syntactic and grammatical adaptation of prompts/responses to the user.

As per Claims 8, 22, 36:

The prior art of record does not teach identifying a context for the current conversation from one or more inferred contextual signifiers or qualifiers missing from the utterance, inferred in response to determining that the utterance contains insufficient information to generate the response, as in Claims 8, 22, 36, and also does not teach these limitations in Claims 9, 23, and 37, which incorporate the limitations based on their dependence on Claims 8, 22, and 36.

Chai et al. (US 6,829,603) suggests wherein determining the intended meaning for the utterance includes inferring one or more contextual signifiers or qualifiers missing from the utterance in response to determining that the utterance contains insufficient information to generate the response (“instantiated if the values of slots… item name… quantity… model… missing slot 807 values… inferred from context”, col. 11, line 57 - col. 12, line 27; where determining there are missing slots determines that there is insufficient information in the utterance when only one utterance has been spoken at the time [thus filling only a limited number of slots] and the slots signify certain types of information and are "contextual" in the sense that they are inferred from context and thus have some attribute associated with that context).

However, Chai does not teach where a context is identified from the inferred missing information.

Conclusion

Any inquiry concerning this communication or earlier communications from the examiner should be directed to ERIC YEN whose telephone number is (571)272-4249. The examiner can normally be reached on M-F 7:30-4:00.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Richemond Dorvil can be reached on 571-272-7602. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

EY 8/9/11

/Eric Yen/
Primary Examiner, Art Unit 2626

Notice of References Cited	Application/Control No. 11/580,926	Applicant(s)/Patent Under Reexamination BALDWIN ET AL.	
	Examiner ERIC YEN	Art Unit 2626	Page 1 of 1

U.S. PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Name	Classification
*	A	US-2001/0049601	12-2001	Kroeker et al.	704/254
*	B	US-6,278,968	08-2001	Franz et al.	704/3
*	C	US-7,536,374	05-2009	Au, Lawrence	706/55
*	D	US-2004/0167771	08-2004	Duan et al.	704/010
*	E	US-7,788,084	08-2010	Brun et al.	704/7
*	F	US-2007/0118357	05-2007	Kasravi et al.	704/010
	G	US-			
	H	US-			
	I	US-			
	J	US-			
	K	US-			
	L	US-			
	M	US-			

FOREIGN PATENT DOCUMENTS

*		Document Number Country Code-Number-Kind Code	Date MM-YYYY	Country	Name	Classification
	N					
	O					
	P					
	Q					
	R					
	S					
	T					

NON-PATENT DOCUMENTS

*		Include as applicable: Author, Title Date, Publisher, Edition or Volume, Pertinent Pages)
	U	
	V	
	W	
	X	

*A copy of this reference is not being furnished with this Office action. (See MPEP § 707.05(a).)
Dates in MM-YYYY format are publication dates. Classifications may be US or foreign.

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
	11580926	BALDWIN ET AL.
	Examiner	Art Unit
	ERIC YEN	2626

SEARCHED			
Class	Subclass	Date	Examiner
704	270-275	6/15/10	EY
704	9, 270-275	8/9/11	EY

SEARCH NOTES		
Search Notes	Date	Examiner
EAST search (attached)	6/15/10	EY
EAST Search (attached)	10/27/10	EY
EAST Search (attached)	8/9/11	EY

INTERFERENCE SEARCH			
Class	Subclass	Date	Examiner
704	9, 270-275	8/9/11	EY

EAST Search History

EAST Search History (Prior Art)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
L2	0	"704".clas. and ((context$1 near5 meaning$1) near5 verb near5 noun)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 09:42
L3	3	"704".clas. and ((context$1 near5 meaning$1) with verb with noun)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 09:42
L4	5	"704".clas. and ((context$1 near5 disambiguat$4) with verb with noun)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:02
L5	11	"704".clas. and ((context$1 with disambiguat$4) with verb with noun)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:02
L6	0	"704".clas. and (prompt$1 with context $1 with verb$1 with noun$1)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:05
L7	1	"704".clas. and (prompt$1 same (context$1 with verb $1 with noun$1))	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:05
L8	0	"704".clas. and ((meaning near5 prompt$1) and (context$1 with verb $1 with noun$1))	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:06
L9	0	"704".clas. and ((meaning$1 near5 prompt$1) and (context$1 with verb $1 with noun$1))	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:06
L10	0	"704".clas. and ((meaning$1 with prompt$1) and (context$1 with verb $1 with noun$1))	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:07

L11	0	"704".clas. and ((semantic$4 with prompt$1) and (context$1 with verb $1 with noun$1))	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:07
L12	17	"704".clas. and ((semantic$4 with (respond$4 response $1)) and (context$1 with verb$1 with noun $1))	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:07
L13	6	"704".clas. and ((semantic$4 with prompt$1 with context $1))	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:15
L14	1	"20010049601".pn. and prompt$1	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:19
L15	1	"20010049601".pn. and prompt$1 and semantic$4	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:21
L16	2	11/580926 and (grammatic$4 near5 syntactic$4)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:26
L17	0	11/580926 and (context$1 near5 verb $1 near5 noun$1) and (prompt$1 near5 context$1)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:29
L18	0	"704".clas. and (context$1 near5 verb $1 near5 noun$1) and (prompt$1 near5 context$1)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:29
L19	0	"704".clas. and (context$1 with verb $1 with noun$1) and (prompt$1 near5 context$1)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:29
L20	1	"704".clas. and (context$1 same verb $1 same noun$1) and (prompt$1 near5 context$1)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:29

L21	6	"704".clas. and (context$1 same verb $1 same noun$1) and (prompt$1 same context$1)	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:31
L22	9	11/580926 and response$1	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:35
L23	2	11/580926 and psychologically	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 10:48
L24	9072	704/9,270-275.ccls.	US-PGPUB; USPAT; USOCR; EPO; JPO	OR	OFF	2011/08/09 12:17

EAST Search History (Interference)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
L25	5252	704/9,270-275.ccls.	USPAT; UPAD	OR	OFF	2011/08/09 12:17

8/9/2011 12:17:53 PM
C:\Documents and Settings\eyen\My Documents\EAST\Workspaces\11580926A.wsp

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
/E.Y./	1A	5,794,196	08/11/1998	Yegnanarayanan *et al.*	704	255	
	1B	5,867,817	02/02/1999	Catallo *et al.*	704	255	
	1C	6,470,315	10/22/2002	Netsch *et al.*	704	256	
	1D	6,604,075	08/05/2003	Brown *et al.*	704	270.1	
	1E	6,622,119	09/16/2003	Ramaswamy *et al.*	704	9	
	1F	6,658,388	12/02/2003	Kleindienst *et al.*	704	275	
	1G	6,708,150	03/16/2004	Hirayama *et al.*	704	243	
	1H	7,231,343	06/12/2007	Treadgold *et al.*	704	9	
	1I	7,386,443	06/10/2008	Parthasarathy *et al.*	704	201	
	1J	7,415,414	08/19/2008	Azara *et al.*	704	270	
	1K	7,809,570	10/05/2010	Kennewick *et al.*	704	257	
	1L	7,818,176	10/19/2010	Freeman *et al.*	704	270	
	1M	2002/0087525	07/04/2002	Abbott *et al.*	707	3	
	1N	2003/0046346	03/06/2003	Mumick *et al.*	709	205	
	1O	2003/0064709	04/03/2003	Gailey *et al.*	455	412	
	1P	2003/0236664	12/25/2003	Sharma	704	251	
	1Q	2004/0006475	01/08/2004	Ehlen *et al.*	704	270.1	
	1R	2004/0193408	09/30/2004	Hunt	704	209	
	1S	2004/0243417	12/02/2004	Pitts, III *et al.*	704	276	
	1T	2005/0043940	02/24/2005	Elder	704	9	
	1U	2005/0137850	06/23/2005	Odell	704	4	
	1V	2007/0179778	08/02/2007	Gong *et al.*	704	9	
/E.Y./	1W	2008/0140385	06/12/2008	Mahajan *et al.*	704	9	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO
/E.Y./	1X	WO 99/46763	09/16/1999	International				
	1Y	WO 00/21232	04/13/2000	International				
	1Z	WO 00/46792	08/10/2000	International				
	1AA	WO 2004/072954	08/26/2001	International				
	1BB	WO 2007/019318	02/15/2007	International				
	1CC	WO 2007/021587	02/22/2007	International				
/E.Y./	1DD	WO 2007/027546	03/08/2007	International				

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

/E.Y./	1EE	BAZZI, Issam *et al.*, "Heterogeneous Lexical Units for Automatic Speech Recognition: Preliminary Investigations", Processing of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 3, June 5-9, 2000, XP010507574, pages 1257-1260.

EXAMINER	DATE CONSIDERED
/E. Yen/	08/09/2011

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

402593690v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
/E.Y./	2A	2008/0177530	07/24/2008	Cross *et al.*	704	4	
	2B	2009/0216540	08/27/2009	Tessel *et al.*	704	275	
	2C	2010/0204986	08/12/2010	Kennewick *et al.*	704	226	
	2D	2010/0204994	08/12/2010	Kennewick *et al.*	704	257	
	2E	2010/0217604	08/26/2010	Baldwin *et al.*	704	275	
	2F	2010/0286985	11/11/2010	Kennewick *et al.*	704	257	
/E.Y./	2G	2010/0299142	11/25/2010	Freeman *et al.*	704	9	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO
/E.Y./	2H	WO 2007/027989	03/08/2007	International				
	2I	WO 2008/098039	08/14/2008	International				
	2J	WO 2008/118195	10/02/2008	International				
	2K	WO 2009/075912	06/18/2009	International				
	2L	WO 2009/145796	12/03/2009	International				
/E.Y./	2M	WO 2010/096752	08/26/2010	International				

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

/E.Y./	2N	O'SHAUGHNESSY, Douglas, "Interacting with Computers by Voice: Automatic Speech Recognition and Synthesis", Proceedings of the IEEE, Volume 91, No. 9, September 1, 2003, XP011100665, pages 1272-1305.

EXAMINER	DATE CONSIDERED
/E. Yen/	08/09/2011

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

402593690v1

Issue Classification

Application/Control No.	Applicant(s)/Patent Under Reexamination
11580926	BALDWIN ET AL.
Examiner	**Art Unit**
ERIC YEN	2626

ORIGINAL		INTERNATIONAL CLASSIFICATION										
CLASS	SUBCLASS	CLAIMED					NON-CLAIMED					
704	9	G	0	6	F	17 / 27 (2006.01.01)						
		G	1	0	L	21 / 00 (2006.01.01)						
CROSS REFERENCE(S)		G	1	0	L	11 / 00 (2006.01.01)						

CLASS	SUBCLASS (ONE SUBCLASS PER BLOCK)				
704	270	275			

☐ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original
1	1	15	17	29	33										
2	2	16	18	30	34										
3	3	17	19	31	35										
4	4	18	20	41	36										
5	5	19	21	42	37										
6	6	39	22	32	38										
7	7	40	23	33	39										
37	8	20	24	34	40										
38	9	21	25	35	41										
8	10	22	26	36	42										
9	11	23	27												
10	12	24	28												
11	13	25	29												
12	14	26	30												
13	15	27	31												
14	16	28	32												

NONE		**Total Claims Allowed:**	
(Assistant Examiner)	(Date)	42	
/ERIC YEN/ Primary Examiner.Art Unit 2626	8/9/11	O.G. Print Claim(s)	O.G. Print Figure
(Primary Examiner)	(Date)	1	3

U.S. Patent and Trademark Office

Part of Paper No. 20110809

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	A	5,839,107	11/17/1998	Gupta *et al.*	704	270	
	B	5,897,613	04/27/1999	Chan	704	210	
	C	5,960,399	09/28/1999	Barclay *et al.*	704	270	
	D	6,049,607	04/11/2000	Marash *et al.*	379	410	
	E	6,085,186	07/04/2000	Christianson *et al.*	707	3	
	F	6,397,181	05/28/2002	Li *et al.*	704	256	
	G	6,429,813	08/06/2002	Feigen	342	357.13	
	H	6,522,746	02/18/2003	Marchok *et al.*	379	406.03	
	I	6,549,629	04/15/2003	Finn *et al.*	381	92	
	J	6,587,858	07/01/2003	Strazza	707	102	
	K	6,975,993	12/13/2005	Keiller	704	275	
	L	7,024,364	04/04/2006	Guerra *et al.*	704	270	
	M	7,236,923	06/26/2007	Gupta	704	9	
	N	7,472,020	12/30/2008	Brulle-Drews	701	211	
	O	7,493,259	02/17/2009	Jones *et al.*	704	257	
	P	7,536,297	05/19/2009	Byrd *et al.*	704	10	
	Q	7,831,433	11/09/2010	Belvin *et al.*	704	275	
	R	7,983,917	07/19/2011	Kennewick *et al.*	704	257	
	S	8,015,006	09/06/2011	Kennewick *et al.*	704	236	
	T	2003/0144846	07/31/2003	Denenberg *et al.*	704	277	
	U	2003/0212550	11/13/2003	Ubale	704	215	
	V	2007/0033020	02/08/2007	(Kelleher) Francois *et al.*	704	226	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

403110745v1

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANTS :	Larry BALDWIN *et al.*	CONFIRMATION NO.:	6670
SERIAL NUMBER :	11/580,926	EXAMINER:	Eric L. Yen
FILING DATE :	October 16, 2006	ART UNIT:	2626
FOR :	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE		

INFORMATION DISCLOSURE STATEMENT UNDER 37 C.F.R. § 1.97(i)

Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313-1450

Sir:

Applicants submit the following documents listed on the attached Form PTO-1449 in accordance with 37 C.F.R. §1.97(i). While Applicants do not consider these documents to be material to the patentability of any claims in the above-identified application, Applicants nonetheless submit them to comply with the spirit of 37 C.F.R. §1.56.

Documents 1AA, 1BB, and 1CC were cited in search reports by foreign patent offices in counterpart foreign applications. Submission of an English language version of each search report that indicates the degree of relevance found by the foreign office is provided.

It is not believed that any fees are necessary for submission of this Information Disclosure Statement. However, if any fees are required, they are hereby authorized to be charged to Deposit Account No. 03-3975, Reference No. 090611-0357945.

Respectfully submitted,

PILLSBURY WINTHROP SHAW PITTMAN LLP

By _____
Syed Jafar Ali
Registration No.: 58,780

Customer No.
00909

P. O. Box 10500
McLean, Virginia 22102
Telephone: (703) 770-7900
Telefax: (703) 770-7901

Dated: September 9, 2011

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	1A	5,027,406	06/25/1991	Roberts *et al.*	381	43	
	1B	5,960,447	09/28/1999	Holt *et al.*	707	500	
	1C	5,995,928	11/30/1999	Nguyen *et al.*	704	251	
	1D	6,018,708	01/25/2000	Dahan *et al.*	704	244	
	1E	6,108,631	08/22/2000	Ruhl	704	270	
	1F	6,167,377	12/26/2000	Gillick *et al.*	704	240	
	1G	6,173,266	01/09/2001	Marx *et al.*	704	270	
	1H	6,208,964	03/27/2001	Sabourin	704	244	
	1I	6,269,336	07/31/2001	Ladd *et al.*	704	270	
	1J	6,366,882	04/02/2002	Bijl *et al.*	704	235	
	1K	6,377,913	04/23/2002	Coffman *et al.*	704	8	
	1L	6,405,170	06/11/2002	Phillips *et al.*	704	270	
	1M	6,606,598	08/12/2003	Holthouse *et al.*	704	275	
	1N	6,681,206	01/20/2004	Gorin *et al.*	704	243	
	1O	6,701,294	03/02/2004	Ball *et al.*	704	257	
	1P	6,704,708	03/09/2004	Pickering	704	235	
	1Q	6,751,612	06/15/2004	Schuetze *et al.*	707	4	
	1R	6,832,230	12/14/2004	Zilliacus *et al.*	707	203	
	1S	6,912,498	06/28/2005	Stevens *et al.*	704	235	
	1T	6,954,755	10/11/2005	Reisman	707	10	
	1U	7,003,463	02/21/2006	Maes *et al.*	704	270.1	
	1V	7,058,890	06/06/2006	George *et al.*	715	728	
	1W	7,197,069	03/27/2007	Agazzi *et al.*	375	233	
	1X	7,277,854	10/02/2007	Bennett *et al.*	704	257	
	1Y	7,366,669	04/29/2008	Nishitani *et al.*	704	256	
	1Z	7,472,060	12/30/2008	Gorin *et al.*	704	240	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO
	1AA	1 320 043 A2	06/18/2003	Europe				
	1BB	1 646 037	04/12/2006	Europe				

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

	1CC	"Statement in Accordance with the Notice from the European Patent Office dated 1 October 2007 Concerning Business Methods (OJ EPO 11/2007, 592-593), XP002456252.

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

403110786v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	2A	7,558,730	07/07/2009	Davis *et al.*	704	235	
	2B	7,676,365	03/09/2010	Hwang *et al.*	704	240	
	2C	7,873,523	01/18/2011	Potter *et al.*	704	275	
	2D	7,917,367	03/29/2011	Di Cristo *et al.*	704	270.1	
	2E	7,949,529	05/24/2011	Weider *et al.*	704	270	
	2F	2002/0035501	03/21/2002	Handel *et al.*	705	10	
	2G	2003/0120493	06/26/2003	Gupta	704	270.1	
	2H	2009/0271194	10/29/2009	Davis *et al.*	704	235	
	2I	2011/0112827	05/12/2011	Kennewick *et al.*	704	9	
	2J	2011/0112921	05/12/2011	Kennewick *et al.*	705	26.1	
	2K	2011/0131036	06/02/2011	Di Cristo *et al.*	704	9	
	2L	2011/0131045	06/02/2011	Cristo *et al.*	704	249	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

403110786v1

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 320 043 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
18.06.2003 Bulletin 2003/25

(51) Int Cl.[7]: **G06F 17/30**

(21) Application number: **02258126.8**

(22) Date of filing: **26.11.2002**

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR IE IT LI LU MC NL PT SE SK TR
Designated Extension States:
AL LT LV MK RO SI

(30) Priority: **13.12.2001 GB 0129788**

(71) Applicant: **Hewlett-Packard Company**
Palo Alto, CA 94304 (US)

(72) Inventor: **Belrose, Guillaume**
Bristol BS2 8HT (GB)

(74) Representative: **Squibbs, Robert Francis**
Hewlett-Packard Limited,
IP Section,
Building 3
Filton Road
Stoke Gifford, Bristol BS 34 8QZ (GB)

(54) Multi-modal picture

(57) A system for presenting a multi-modal picture includes picture presentation equipment (37) for displaying an image (38) of the picture and for enabling a user to interact with the picture by selecting specific picture features and asking a specific query relating to the feature. A voice browser system (3) controlled according to dialog scripts (46) associated with the picture, determines an appropriate response having regard to the spoken user query and the picture feature being pointed to. Each picture can have multiple narrators associated with it and the can choose which narrator is currently active. Picture authoring apparatus is also provided.

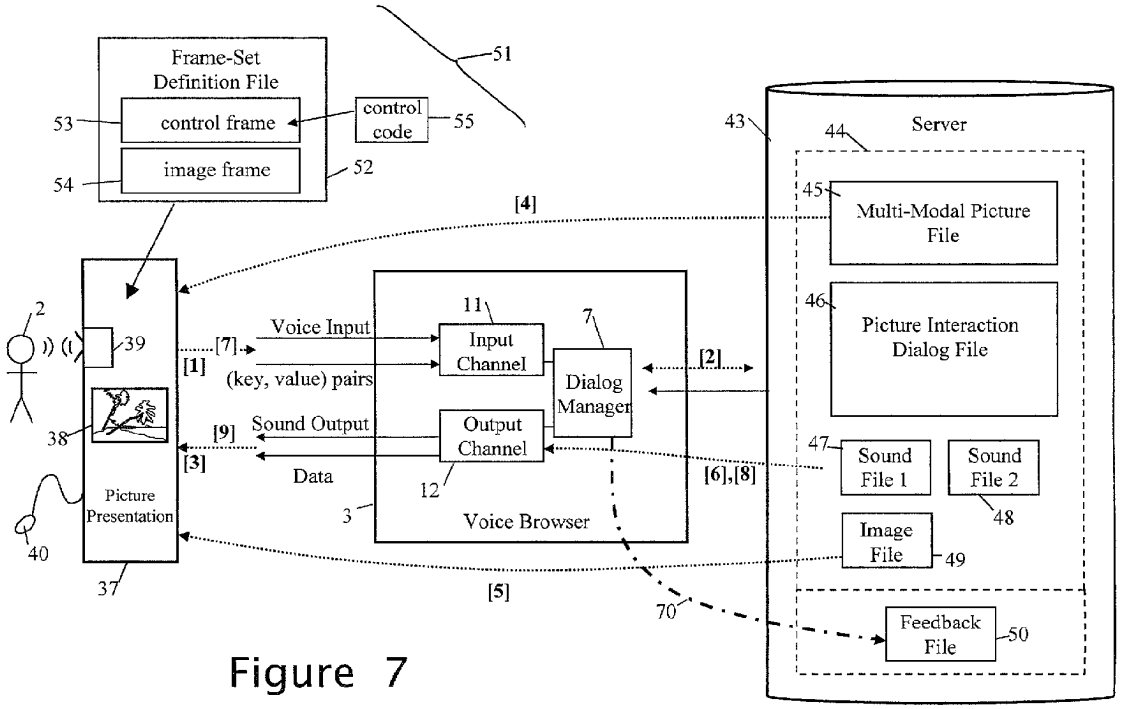


Figure 7

EP 1 320 043 A2

Printed by Jouve, 75001 PARIS (FR)

Description

Field of the Invention

[0001] The present invention relates to multi-modal pictures with which a use can interact by spoken dialog exchanges.

Background of the Invention

[0002] In recent years there has been an explosion in the number of services available over the World Wide Web on the public internet (generally referred to as the "web"), the web being composed of a myriad of pages linked together by hyperlinks and delivered by servers on request using the HTTP protocol. Each page comprises content marked up with tags to enable the receiving application (typically a GUI browser) to render the page content in the manner intended by the page author; the markup language used for standard web pages is HTML (HyperText Markup Language).

[0003] However, today far more people have access to a telephone than have access to a computer with an Internet connection. Sales of cellphones are outstripping PC sales so that many people have already or soon will have a phone within reach where ever they go. As a result, there is increasing interest in being able to access web-based services from phones. 'Voice Browsers' offer the promise of allowing everyone to access web-based services from any phone, making it practical to access the Web any time and any where, whether at home, on the move, or at work.

[0004] Voice browsers allow people to access the Web using speech synthesis, pre-recorded audio, and speech recognition. Figure 1 of the accompanying drawings illustrates the general role played by a voice browser. As can be seen, a voice browser is interposed between a user 2 and a voice page server 4. This server 4 holds voice service pages (text pages) that are marked-up with tags of a voice-related markup language (or languages). When a page is requested by the user 2, it is interpreted at a top level (dialog level) by a dialog manager 7 of the voice browser 3 and output intended for the user is passed in text form to a Text-To-Speech (TTS) converter 6 which provides appropriate voice output to the user. User voice input is converted to text by speech recognition module 5 of the voice browser 3 and the dialog manager 7 determines what action is to be taken according to the received input and the directions in the original page. The voice input / output interface can be supplemented by keypads and small displays.

[0005] In general terms, therefore, a voice browser can be considered as a largely software device which interprets a voice markup language and generate a dialog with voice output, and possibly other output modalities, and / or voice input, and possibly other modalities (this definition derives from a working draft, dated September 2000, of the Voice browser Working Group of the World Wide Web Consortium).

[0006] Voice browsers may also be used together with graphical displays, keyboards, and pointing devices (e.g. a mouse) in order to produce a rich "multimodal voice browser". Voice interfaces and the keyboard, pointing device and display maybe used as alternate interfaces to the same service or could be seen as being used together to give a rich interface using all these modes combined.

[0007] Some examples of devices that allow multimodal interactions could be multimedia PC, or a communication appliance incorporating a display, keyboard, microphone and speaker/headset, an in car Voice Browser might have display and speech interfaces that could work together, or a Kiosk.

[0008] Some services may use all the modes together to provide an enhanced user experience, for example, a user could touch a street map displayed on a touch sensitive display and say "Tell me how I get here?". Some services might offer alternate interfaces allowing the user flexibility when doing different activities. For example while driving speech could be used to access services, but a passenger might used the keyboard.

[0009] Figure 2 of the accompanying drawings shows in greater detail the components of an example voice browser for handling voice pages 15 marked up with tags related to four different voice markup languages, namely:

- tags of a dialog markup language that serves to specify voice dialog behaviour;
- tags of a multimodal markup language that extends the dialog markup language to support other input modes (keyboard, mouse, etc.) and output modes (large and small screens);
- tags of a speech grammar markup language that serve to specify the grammar of user input; and
- tags of a speech synthesis markup language that serve to specify voice characteristics, types of sentences, word emphasis, etc.

[0010] When a page 15 is loaded into the voice browser, dialog manager 7 determines from the dialog tags and multimodal tags what actions are to be taken (the dialog manager being programmed to understand both the dialog and multimodal languages 19). These actions may include auxiliary functions 18 (available at any time during page processing) accessible through APIs and including such things as database lookups, user identity and validation, telephone call control etc. When speech output to the user is called for, the semantics of the output is passed, with any

associated speech synthesis tags, to output channel 12 where a language generator 23 produces the final text to be rendered into speech by text-to-speech converter 6 and output to speaker 17. In the simplest case, the text to be rendered into speech is fully specified in the voice page 15 and the language generator 23 is not required for generating the final output text; however, in more complex cases, only semantic elements are passed, embedded in tags of a natural language semantics markup language (not depicted in Figure 2) that is understood by the language generator. The TTS converter 6 takes account of the speech synthesis tags when effecting text to speech conversion for which purpose it is cognisant of the speech synthesis markup language 25.

[0011] User voice input is received by microphone 16 and supplied to an input channel of the voice browser. Speech recogniser 5 generates text which is fed to a language understanding module 21 to produce semantics of the input for passing to the dialog manager 7. The speech recogniser 5 and language understanding module 21 work according to specific lexicon and grammar markup language 22 and, of course, take account of any grammar tags related to the current input that appear in page 15. The semantic output to the dialog manager 7 may simply be a permitted input word or may be more complex and include embedded tags of a natural language semantics markup language. The dialog manager 7 determines what action to take next (including, for example, fetching another page) based on the received user input and the dialog tags in the current page 15.

[0012] Any multimodal tags in the voice page 15 are used to control and interpret multimodal input/output. Such input/output is enabled by an appropriate recogniser 27 in the input channel 11 and an appropriate output constructor 28 in the output channel 12.

[0013] Whatever its precise form, the voice browser can be located at any point between the user and the voice page server. Figures 3 to 5 illustrate three possibilities in the case where the voice browser functionality is kept all together; many other possibilities exist when the functional components of the voice browser are separated and located in different logical/physical locations.

[0014] In Figure 3, the voice browser 3 is depicted as incorporated into an end-user system 8 (such as a PC or mobile entity) associated with user 2. In this case, the voice page server 4 is connected to the voice browser 3 by any suitable data-capable bearer service extending across one or more networks 9 that serve to provide connectivity between server 4 and end-user system 8. The data-capable bearer service is only required to carry text-based pages and therefore does not require a high bandwidth.

[0015] Figure 4 shows the voice browser 3 as co-located with the voice page server 4. In this case, voice input/output is passed across a voice network 9 between the end-user system 8 and the voice browser 3 at the voice page server site. The fact that the voice service is embodied as voice pages interpreted by a voice browser is not apparent to the user or network and the service could be implemented in other ways without the user or network being aware.

[0016] In Figure 5, the voice browser 3 is located in the network infrastructure between the end-user system 8 and the voice page server 4, voice input and output passing between the end-user system and voice browser over one network leg, and voice-page text data passing between the voice page server 4 and voice browser 3 over another network leg. This arrangement has certain advantages; in particular, by locating expensive resources (speech recognition, TTS converter) in the network, they can be used for many different users with user profiles being used to customise the voice-browser service provided to each user.

[0017] It is known to enhance pictures by providing associated speech annotations. It is an object of the present invention to provide further enhancements to such pictures.

Summary of the Invention

[0018] According to one aspect of the present invention, there is provided a system for presenting information concerning a picture to a user, the system comprising:

- a data store for holding responses, specific to said picture, in respect of specific user queries concerning particular picture features;
- a manually-operable feature-selection arrangement for enabling a user to select a feature in a displayed image of the picture, and for providing an output indication regarding what said particular feature, if any, the user has thereby selected;
- a voice dialog input-output subsystem including a speech recogniser for interpreting queries from a user;
- a control arrangement responsive to a user selecting a said particular feature and asking a specific query regarding that feature, to output the corresponding stored response.

[0019] Advantageously, each response is associated with a specified narrator and the user can opt to receive only the response of a selected narrator for either a current query or for all queries.

[0020] Preferably, provision is made for providing feedback to an author of the picture or its associated responses in the event that a user makes a query about a picture feature that is not a said particular picture feature or, although

a query about such a feature, is not a said specific query for which a response is available.

[0021] According to another aspect of the present invention, there is provided a multi-modal picture specified by data held on at least one data carrier, this data comprising:

- picture image data for displaying a picture image;
- response data indicative of voice responses intended to be given to specific user queries concerning particular picture features of the picture;
- first control data for enabling a determination to be made as to which said particular feature in the picture image, if any, a user is selecting when using a selection arrangement to indicate a feature in the displayed image; and
- second control data for determining, on the basis of a spoken user query and on which said particular picture feature is selected by the user using the selection arrangement, which said response is to be used to reply to the user query.

[0022] According to a further aspect of the present invention, there is provided a multi-modal picture comprising a hard-copy picture image, and data held on at least one data carrier, this data comprising:

- response data indicative of voice responses intended to be given to specific user queries concerning particular picture features;
- first control data for enabling a determination to be made as to which said particular feature in the picture image, if any, a user is selecting when using a selection arrangement to indicate a feature of the image; and
- second control data for determining, on the basis of a spoken user query and on which said particular picture feature is selected by the user using the selection arrangement, which said response is to be used to reply to the user query.

[0023] The present invention also encompasses the overall process of generating a multi-modal picture and using it to convey information about particular picture features to a user. Thus, according to a still further aspect of the present invention, there is provided a method of conveying information about particular features in a picture, the method comprising the steps of:

(a) creating the following specifically-associated data:

- picture image data for displaying a picture image;
- response data indicative of voice responses intended to be given to specific user queries concerning particular picture features of the picture;
- first control data for enabling a determination to be made as to which said particular feature in the picture image, if any, a user is selecting when using a selection arrangement to indicate a feature in the displayed image; and
- second control data for determining, on the basis of a spoken user query and on which said particular picture feature is selected by the user, which said response is to be used to reply to the user query;

(b) using the image data to display an image of the picture;
(c) having a user use a manually-operated selection arrangement to select a feature in the displayed image and using the first control data to determine which said particular feature in the picture image, if any, the user is selecting;
(d) receiving and interpreting a spoken query from the user to determine if a said specific query is being asked; and
(e) using the second control data to determine, on the basis of the said particular feature determined as being selected in step (c) and the said specific query determined as being asked in step (d), which said response is to be used to reply and thereupon using the response data to output the corresponding voice response.

[0024] Similarly, according to a yet further aspect of the present invention, there is provided a method of conveying information about particular features in a hard-copy picture, the method comprising the steps of:

(a) creating the following specifically-associated data:

- response data indicative of voice responses intended to be given to specific user queries concerning particular picture features in said picture;
- first control data for enabling a determination to be made as to which said particular feature in the picture, if any, a user is selecting when using a selection arrangement to indicate a feature of the picture; and
- second control data for determining, on the basis of a spoken user query and on which said particular picture

feature is selected by the user, which said response is to be used to reply to the user query;

(b) making the picture and data available to a user;
(c) having the user use a manually-operated selection arrangement to select a feature in the picture and using the first control data to determine which said particular feature in the picture, if any, the user is selecting;
(d) receiving and interpreting a spoken query from the user to determine if a said specific query is being asked; and
(e) using the second control data to determine, on the basis of the said particular feature determined as being selected in step (c) and the said specific query determined as being asked in step (d), which said response is to be used to reply and thereupon using the response data to output the corresponding voice response.

[0025] According to another aspect of the present invention, there is provided apparatus for authoring a multi-modal picture, comprising:

- a first tool for defining image hotspots associated with particular picture-image features;
- a second tool with speech recognition capability, for recording user responses input by voice, to user-specified queries each associated with a particular said picture-image feature; and
- means for automatically generating control data for determining, on the basis of a spoken user query and on which said particular picture feature is selected by a user, which said response is to be used to reply to the user query.

[0026] With respect to the above-mentioned feature regarding the provision of feedback on user queries for which responses are not available, a further independent aspect of the present invention provides a method of collecting user-interest information regarding a picture, the method comprising the steps of:

(a) presenting a picture image to a user, particular picture features in the image each having associated information that is presented to the user upon the user requesting such information;
(b) recognising a user information request, in respect of a picture feature, made by the user at least selecting that picture feature using a feature-selection arrangement; and
(c) in response to recognition of a user information request in step (b) concerning a picture feature that is other than a said particular picture feature, outputting an identifier of the concerned picture feature independently of the presented image.

[0027] This aspect of the invention may be implemented in the form of a system for collecting user-interest information regarding a picture, the system comprising:

- a data store for holding information concerning particular picture features;
- a feature-selection arrangement for enabling a user to select a feature in a displayed view of the picture, and for providing an output indication regarding what said particular feature, if any, the user has thereby selected;
- an information-output subsystem for outputting information about picture features to the user;
- a control arrangement for recognising a user information request, in respect of a picture feature, made by the user at least selecting that picture feature using the feature-selection arrangement, the control arrangement including:

  - first output means arranged to respond to a user making an information request concerning a said particular feature, by outputting the corresponding stored information via the information-output subsystem; and
  - second output means arranged to respond to a user making an information request concerning a picture feature that is other than a said particular picture feature, by outputting an identifier of the concerned picture feature for informing an author of the picture or its associated information regarding the unsatisfied information request.

Brief Description of the Drawings

[0028] A method and apparatus embodying the invention will now be described, by way of nonlimiting example, with reference to the accompanying diagrammatic drawings, in which:

. **Figure 1** is a diagram illustrating the role of a voice browser;
. **Figure 2** is a diagram showing the functional elements of a voice browser and their relationship to different types of voice markup tags;
. **Figure 3** is a diagram showing a voice service implemented with voice browser functionality located in an end-user system;

- **Figure 4** is a diagram showing a voice service implemented with voice browser functionality co-located with a voice page server;
- **Figure 5** is a diagram showing a voice service implemented with voice browser functionality located in a network between the end-user system and voice page server; and
- **Figure 6** shows an example picture image of a multi-modal picture embodying the invention;
- **Figure 7** is a diagram of a system for presenting a multi-modal picture to a user;
- **Figure 8** is a diagram showing constituent dialog blocks of a picture interaction dialog file of the Figure 6 multi-modal picture;
- **Figure 9** is a diagram of apparatus for authoring a multi-modal picture;
- **Figure 10** shows, in respect of the Figure 6 image, a user seeking information from two picture features for which there is no associated information; and
- **Figure 11** shows the Figure 6 picture image enhanced with upper and lower information bars.

Best Mode of Carrying Out the Invention

[0029] In the following description, voice dialog interaction with a user is described based on a voice page server serving a dialog page with embedded voice markup tags to a multi-modal voice browser. Unless otherwise indicated, the foregoing description of voice browsers, and their possible locations and access methods is to be taken as applying also to the described embodiments ofthe invention. Furthermore, although voice-browser based forms of voice dialog services are preferred, the present invention in its widest conception, is not limited to these forms of voice dialog service system and other suitable systems will be apparent to persons skilled in the art.

Multi-Modal Picture

[0030] Figure 6 depicts a multi-modal picture comprising a displayed picture image 30 (here shown as being of a holiday island taken, for example, whilst the author of the picture was on holiday) with which a recipient (referred to below as the "user") can interact using multiple modalities and, in particular, by spoken dialogues and the use of a pointing arrangement, such as a cursor controlled by a mouse, stylus or keyboard keys, a touch-screen detection arrangement, etc. Thus, the user can query the picture using speech input, the query being either a general query about the picture or a specific query about a particular feature (item or area) of the picture indicated by use of the pointing arrangement (for example, the user uses a mouse to move a cursor over a picture feature and then clicks a mouse button). Of course, only certain picture features will have been set up to be queried and in the Figure 6 example, there are three such features, namely a coconut tree 31, a first person 32, and a second person 33; the dashed boxes around these features represent the "hotspot" image areas set up to encompass the features, the dashed lines generally not being visible (though displaying a hotspot boundary can be used as one way of indicating the hotspot to the user - a more typical way would be, in the case of the pointing arrangement being a mouse-controlled cursor, to change the cursor image as it moved into and out of a hotspot).

[0031] Processing functionality associated with the multi-modal picture is arranged to recognize one or more one or more general queries and, for each picture feature set up for selection, one or more specific queries. For each query set to be recognized, there is a corresponding response which is output to the user when the query is recognized, this output generally being in spoken form, generated using either speech synthesis or pre-recorded audio content. Thus, for example:

General Queries - available general queries can include picture location, the date of the picture was taken, and a description of the general subject of the picture. For instance:

User: "Describe the picture"
System: "This is a picture of the XYZ beach"
User: "What is the date?"
System: "The picture was taken November last year"
User: "What is the location?"
System: "XYZ beach is in Martinique, an island that is part of the French West Indies."

(In this and other dialog examples given in the present specification, the "System" is the functionality used to present the multi-modal picture to the user and to provide for multi-modal interaction).

Specific Queries - typically, the same specific queries are available for all the selectable picture features; example specific queries are: "What it is?", "Any story?", etc. For instance:

User: "What is it"? "
System: "This is my cousin John".
User: "Any story?"
System: "John goes fishing quite often in the Caribbean Sea. One day, he just escaped a white shark attack."

[0032] The responses from the system can, of course, be more than just a simple single turn response and may involve a multi-turn structured dialog. Provision may also be made for different levels of content, for instance an initial level of public information and a lower level of private information only accessible under password control (or by some other security check procedure). Alternatively, there may be both a brief and a much fuller response .

[0033] A further possibility is to have responses from more than one narrator. Thus, a user can first obtain a response from one person associated with the picture and then a response from a different person also associated with the picture. For example, the user receives a picture from his/her parents. In the picture, the user's mother appears to be talking to someone, but the user's father looks bored. The user interaction might therefore proceed as follows:

User: "Mother, who is she?" [User uses pointing arrangement to indicate a person in the picture].
System: [In pre-recorded voice of user's mother] "This is my colleague from work. We had such a good time."
User: "Father, can you please describe the picture?"
System: [In pre-recorded voice of user's father] "We went to this rather boring evening with your mother's colleagues. They ended up talking about clothes all the time."

[0034] From the foregoing general description of a multi-modal picture, it will be appreciated that the term "multi-modal picture" covers more than just the picture image and includes also the associated dialog and the behaviour of the picture in response to user input.

Example Implementation

[0035] Figure 7 depicts an example implementation of a system for presenting a multi-modal picture.

[0036] The multi-modal picture is specified by a set 44 of related files held on a server 43, these files specifying both the image associated with the multi-modal picture and also the picture interaction dialogue associated with the picture. At least the dialogue and image files of set 44 are specifically associated with each other- that is, the association of these files is prespecified (in advance of user interaction) with the contents of the dialogue file being specifically tailored to the picture represented by the image file of the same set 44.

[0037] User 2 has local picture-presentation equipment 37; in the present case, this equipment is standard processing platform, such as a PC or portable computer, arranged to run a graphical web browser such as Microsoft's Internet Explorer. Other implementations of the picture presentation equipment 37 can alternatively be provided such as a mobile phone with a graphical display and suitable programs.

[0038] The picture-presentation equipment 37 displays the picture image of the multi-modal picture on display 38 and also provides for the input and output audio data via audio functionality 39, this audio input and output being transferred to and from a voice browser 3 (here shown as located separately from the picture picture-presentation equipment - for example, a network-based voice browser - but alternatively, integrated with the picture-presentation equipment 37). The picture-presentation equipment also has an associated pointing arrangement 40 depicted in Figure 7 as a mouse for controlling a cursor movable over the displayed picture image; other forms of pointing arrangements for selecting features of the picture image are, of course, also possible such as the use of touch pads, tracker balls or joysticks for moving an image cursor; touch-sensitive displays and other arrangements (such as a matrix of infra-red beams immediately overlying the display) for enabling a user to use a finger or stylus to point directly at a feature of the displayed image; etc..

[0039] The voice browser 3 comprises input channel 11, dialog manager 7 and output channel 12. The voice browser provides interactive dialog with the user, via the picture-presentation equipment 37, in respect of the currently-presented multi-modal picture, this dialog being in accordance with picture interaction dialog data retrieved from server 43.

[0040] The Figure 7 voice browser 3, as well as providing for voice input and output exchanges with the picture-presentation equipment 37, is also arranged to receive input data from the equipment in the form of key,value pairs indicating, for example, the selection of a particular picture feature by the user using the pointing arrangement 40 of the picture-presentation equipment. This data input is used by the dialog manager 7 in determining the course of the dialog with the user. The voice browser 3 can also provide data back to the picture-presentation equipment 37.

[0041] The picture-presentation equipment 37, voice browser 3 and server 3 inter-communicate via any suitable communications infrastructure or direct links.

[0042] Considering in more detail the files involved in the presentation of a multi-modal picture, these comprise, in addition to the set 44 of files that specify a multi-modal picture, a set of generic files 51. The multi-modal picture files

44 comprise:

- the picture file 45, this file including a source reference for the picture image file 49 to be displayed and map data defining the image hotspots;
- a picture interaction dialog file 46 containing dialog scripts;
- one or more sound files 47, 48 (such as " .wav" files) containing audio data;
- the image file 49 (such as a ".jpg" or ".gif' file) containing the image data to be displayed.

[0043] The generic files 51 can be stored locally in the picture-presentation equipment 37 or retrieved from a remote location such as the server 43. The generic files comprise:

- a frame-set definition file 52 defining two frames 53, 54 into which page files can be independently loaded; one frame 53 is used to hold a file 55 containing control code (the contents of this frame not being visible), and the other frame 54 being used to hold the picture file 45 for the multi-modal picture to be presented.
- the control code file 55 to be loaded into frame 53, the control code being in the form of a number of scripts the main purpose of which is to provide key,value pairs to the voice browser according to events detected by the browser software run by the picture-presentation equipment 37 - in particular, clicking on an image hotspot as defined in file 45 is arranged to trigger a corresponding script in the control code file 55 whereby to cause a corresponding key,value pair to be passed to the voice browser 3 to inform it that a particular picture feature (corresponding to the activated hotspot) has been selected by the user.

[0044] To use the generic files to present a particular multi-modal picture, it is necessary to provide a reference to the multi-modal picture. This reference can, for example, be manually input by the user into an initial form displayed in the image frame 54 and then used by a script to request the loading of a file into frame 54. Alternatively, the multi-modal picture reference can be included as data in a query string attached to the URL of the frame-set definition file 52 (this URL and query string being, for example, provided to the user by the author of the multi-modal picture); in this case, in response to a request for the frame-set definition file 52, server-side code could, for example, extract the data from the query string and place it in the file source reference in the definition line for frame 54 in the frame-set definition file before that file is returned to the user.

[0045] In the present example, the multi-modal picture reference used to initiate presentation of the multi-modal picture is a reference to the picture interaction dialog file 46 to be loaded into the voice browser, rather than a reference to the picture file 45 that is to be loaded into the image frame 54. Thus, the multi-modal picture reference, however obtained, is passed in a key,value pair to the voice browser 3; voice browser 3 thereupon retrieves the picture interaction dialog file 46 to the dialog manager 7 of the voice browser. The file 46 includes a reference to the picture image file 45 to be loaded into the frame 54 and this reference is returned to the picture-presentation equipment 37 where it is used to retrieve the picture file 45. Where the picture-presentation equipment 37 comprises a standard web browser, one way of achieving the above is to have the dialog file reference sent (either as a source reference in the frame-set definition file or by a script in, for example, file 55) to the voice browser in a request for a file to load into frame 54, the voice browser sending back the picture file reference as a redirection.

[0046] It will be appreciated that the multi-modal picture reference passed into the generic files could have been that of the picture file 45, the latter then being retrieved into frame 54 and including an "onLoad" event script for passing to the voice browser a reference to the interaction dialog file.

[0047] The various steps involved in presenting a multi-modal picture according to the Figure 7 implementation are summarized below with reference to the reference numerals in square brackets in Figure 7:

[1] the generic files 51 have been loaded into the picture-presentation equipment 37 and a multi-modal picture reference has been provided; as a result, a key,value pair including a reference to the picture interaction dialog file 46 is passed to the voice browser 3;

[2] voice browser retrieves the dialog file 46 from the server 43 and dialog manger 7 uses the file to control further interaction with the user;

[3] the first action taken by the dialog manager 7 under the control of dialog file 46 is to pass to the equipment the reference for the picture file 45 along with a voice greeting to the user;

[4] the picture file 45 (including a reference to the image file 49) is retrieved from server 43 and loaded into the image frame 54;

[5] the image file 49 is retrieved from the server 43 and displayed on display 38;

[6] meanwhile, the dialog manager 7 causes a sound (in sound file 47) to be played to the user to indicate that the picture is ready to receive user input (this sound can simply be an appropriate background sound such as, for the Figure 6 picture image, the sound of the sea);

[7] the user queries the picture by voice input (and possibly also by pointing to a particular area of the picture, this being indicated by a corresponding key,value pair sent to the voice browser along with the user voice input);
[8] dialog manager 7 acknowledges the receipt of the user query by causing an acknowledgement sound (in sound file 48) to be played back to the user;
[9] the dialog manager 7, having determined the appropriate response to the user query, outputs this response.

Steps [8] and [9] are repeated as many times as required by the user. In due course the user asks to exit and the dialog is terminated by the dialog manager.

[0048] Figure 8 illustrates the contents of the picture interaction dialog file 46. This file contains a number of dialog blocks 60 to 73 that contain dialog elements and/or control structures relating to dialog progression. Thus, dialog block 60 provides the initial greeting and causes the picture file reference to be passed to the equipment 37 (in step [3] above). Block 61 defines the query grammar and represents a waiting state for the dialog pending the receipt of a query from the user.

[0049] Block 62 carries out an analysis of a recognized query to determine whether it is an exit request (if so, an exit dialog block 63 is entered), a generic request, or a specific request; generic and specific requests are further analyzed to determine the nature of the query (that is, what "action" - type of information - is being requested). For a general query, the available actions are, in the present example, "date", "description", and "location; for a specific query, the action types are, in the present example, "what" and "story". Depending on the outcome of the action analysis, the dialog manager proceeds to one of blocks 64 - 66 (for a general query) or one of blocks 67 and 68 (for a specific query). The analysis carried out by dialog block 62 is on the basis of voice input only.

[0050] If the query was a general one, then block 64 is used to answer a date query, block 65 is used to respond to a description query, and block 66 is used to respond to a location query.

[0051] If the query was a specific "what" query, block 67 determines the identity of the picture feature (object) involved using the key,value pair provided to the voice browser; depending on the object identity, the dialog manager proceeds either to a "what" dialog block 70 for a coconut tree or to a "what" dialog block 71 for a French tourist. Similarly if the query was a specific "story" query, block 68 determines the identity of the picture feature (object) involved using the key,value pair provided to the voice browser; depending on the object identity, the dialog manager proceeds either to a "story" dialog block 72 for a coconut tree or to a "story" dialog block 73 for a French tourist.

[0052] After a response is provided by any one of the dialog blocks 64-66 or 70-73, the dialog manager returns to dialog block 61.

[0053] An example dialog interaction based on the Figure 6 picture and the Figure 8 dialog structure is given below; in this example, there is one narrator called Roger.

Example Interaction:

[0054]

System: "This picture was sent to you by Guillaume. You can talk to Roger to ask for information about it."
User: "Roger, Describe the picture"
System: "This is a picture of the XYZ beach."
User: "What is the location?"
System: "This picture was taken in Martinique, an island located in the French West Indies."
User: "What is it" [and clicks on feature 31]
System: "This is a coconut tree. This kind of trees is very common in the Caribbean."
User: "Any story?"
System: "When I was little, I used to get coconuts from my garden. There is a sweet liquid inside that is very tasty. We use coconut to make ice creams and all sorts of nice desserts."
User: "What is it?" [and clicks on 32]
System: "I guess this is a French tourist"
User: "Any story?"
System: "Tourism is one of the main source of revenue in Martinique. A lot of people come from France when it is winter over there."
User: "Exit"
System: "Thanks for using this digital picture frame."

[0055] The Appendix to the present description includes a detailed script example of the Figure 8 dialog interaction as well as the HTML source of a picture file 45.

Multi-modal picture authoring tool.

[0056] Figure 9 illustrates apparatus for enabling an author 80 to author a multi-modal picture. The apparatus comprises a computing platform 81 providing a graphical user interface and running a tool, such as Microsoft's FrontPage product, for authoring mark-up language pages and for creating image hotspot maps. The apparatus further comprises a speech interface system 82 (here shown as a voice browser though other forms of speech interface system can be used). The speech interface system 82 permits the author 80 to interact with the apparatus by voice and is set up to recognize command words such as "Record".

[0057] Different people, known as narrators, can author different aspects of the same picture. The apparatus keeps a record of narrators known to it.

[0058] The apparatus is arranged to interact with one or more narrators to build up, in memory, the set of files 44 that specify a multi-modal picture, it being assumed that the picture image file 47 around which the multi-modal picture is to be built has already been loaded into memory (for example, from a digital camera) and is displayed via the graphical user interface (GUI) of platform 81. The process of building the required files is controlled by a top-level authoring program 90 that has three main steps 91-93 as will be more fully explained below.

Identifying the narrator - (step 91)

[0059] The first step of the authoring program is to identify the current narrator. The narrator speaks his/her name into the speech interface system 82; if the name is known to the apparatus, the system replies with a recognition greeting. However, if the narrator's name is not already known to the apparatus, the system asks the narrator to create a new profile (basically, input his/her name and any other required information), using appropriate data collection screens displayed on the graphical user interface of the computing platform 81.

[0060] Example: the apparatus knows the following names: "Lawrence" and "Marianne".

Apparatus:	"What is your name?"
Narrator:	"Steve".
Apparatus:	"Sorry, I do not know this name, please write it down."
Narrator:	[inputs "Steve" via a data collection screen of the GUI].

[0061] The authoring program uses the narrator's name to customize a greeting dialog block of a template picture interaction dialog file 46.

Adding general information about the picture - (step 92)

[0062] After identification, the narrator can input general information concerning the picture image such as the date, the location or the description, via a spoken dialogue. In the following example, the command words recognized by the speech interface system 82 are shown in bold whilst the nature of information being recorded (corresponding to the query "action" type of Figure 8) is indicated by underlining. The words indicating the nature of the information are either pre-designated to the system (effectively limiting the classification of information to be input) or else the system can be arranged to analyze narrator "Record" commands to determine the nature of the information to be recorded.

Example:

[0063]

Narrator:	**"Record** description"
Apparatus:	[Plays a beep].
Narrator:	"This is a picture of me and John fishing in the Caribbean sea". [The apparatus records this input, either directly as sound data or as a text data after the input has been subject to speech recognition by the system 82]
Narrator:	**"Write** date".
Apparatus:	[displays date capture screen on GUI].
Narrator:	[inputs date information via GUI].
Narrator:	**"Record** story"
Apparatus:	[Plays a beep.]
Narrator:	"This day, John was attacked by a white shark." [The apparatus records this input]

[0064] The authoring program uses the input from the narrator to create corresponding dialog blocks, similar to those described above with reference to Figure 8, in dialog file 46.

Adding specific information - (step 93)

[0065] The narrator can also input information concerning a specific feature of the picture image. To do this, the narrator indicates the picture feature of interest by using the GUI to draw a "hotspot" boundary around the feature. The apparatus responds by asking the narrator to input a label for the feature via an entry screen displayed on the GUI.. The authoring program uses the input from the narrator to create the multi-media picture file 45 holding an image hotspot map with appropriate links to the control code scripts.

[0066] The narrator can then enter further information using the speech interface system or the GUI. The narrator can record or write multiple descriptions or stories for a single area of the picture, for example, to give different level of details.

Narrator: **"Record** description".
Apparatus: [Plays beep.]
Narrator: "It is a whale" [The apparatus records this input]
Narrator: **"Record** story"
Apparatus: [Plays beep.]
Narrator: "We saw this whale on the way to Dominica." [The apparatus records this input]
Narrator: **"Record next"** (indicates that further details for the story are to be recorded).
Narrator: "We were crossing the Guadeloupe channel when we saw it." [The apparatus records this input]

[0067] Again, the authoring program uses the input from the narrator to create corresponding dialog blocks; thus, for the above example, where a "whale" hotspot has been designated by the user, the authoring program generates a set of dialogs blocks: 'whaleDescription', 'whaleStoryl', 'whaleStory2', etc.

[0068] After the first narrator has finished inputting information, other narrators can enter information in the same manner.

User Feedback

[0069] It will be appreciated that the authoring of a multi-modal picture, and in particular the adding of the dialog data, can be quite involved. As a result, it is quite likely that the author will not always include information that a recipient may be interested in.

[0070] It is therefore useful to be able to monitor a user's interaction with a multi modal picture to see if the user tries to access missing information. For example, a user receives the Figure 6 multi modal picture from a friend. Two objects 35,36 (a building and an island) in the picture image intrigues the user (see Figure 9) who therefore clicks on each picture feature concerned and asks for more information ("What is it?"). Unfortunately, there are no hotspots associated with either picture feature and therefore in each case the voice browser comes back with the response "Sorry, there is no information about this item."

[0071] However, the coordinates corresponding to each picture feature the user clicked are known to the browser used to display the picture image and the control script (for example, in file 55) can be used to pass the coordinates as key,value pairs to the voice browser 3. At the same time, a user verbal query is also passed to the voice browser. The voice browser first determines whether the query is a general one and if it is, the voice browser ignores the received coordinate values; however, if the voice browser determines that the query is a specific one, then it determines from the key,value pairs received that the user has indicated a picture feature for which there is no corresponding information available. In this case, the voice browser logs the received coordinate values and the associated "action" type (in the example given above, "what" or "story") in a feedback file 50 that forms part of the set 44 of related files associated with the multimodal picture (see arrow 70 in Figure 7). For example, upon the user clicking on the island feature 36 in Figure 9 and asking "what is this?", the data (action='what'; coordx =400;coordy=300) is logged to file 50. Such logging functionality is, for example, provided by a further dialog block of Figure 8. The logged coordinates provide, together with an indication of the picture concerned, a picture-feature identifier that identifies the picture feature about which information has been requested by the user.

[0072] Alternatively or additionally to logging the "desired-information" feedback data in file 50, the author of the multi-modal picture can be sent a message (for example, an e-mail message) explaining the query from the user such as "John wants a description of this object.". This message includes a picture feature identifier that identifies the picture feature concerned. The picture feature identifier can take the form of explicit data items indicative of the picture concerned and the coordinates of the feature in the picture or may more directly indicate the feature by including either

image data representing the relevant portion of the picture image or image data showing the picture image with a marking indicating the position in the image of the feature concerned (both such forms of picture-feature indication can also be included the feedback file 50 additionally or alternatively to the feature coordinates). The picture-feature indication need not be sent to the file 50 (or included in a message to the author) at the time the system detects that the user has asked for information about a non-hotspot picture feature; instead, the indication and related query can be temporarily stored and output, together with other accumulated similar indications and queries, either at a specified time or event (such as termination of viewing of the picture by the user) or upon request from the picture author.

[0073] The author can then decide whether or not to add in further information by adding an additional hotspot and additional dialogs (for example, using the above-described authoring apparatus).

[0074] Of course, the same general feedback process can be used where although a selected picture feature is associated with an existing hotspot, there is no existing query "action" corresponding to a user's query. Furthermore, a similar feedback process can be used where user queries are input by means other than speech (such as, for example, via a keyboard or by a hand-writing or stroke recognition system) and, indeed, where there are no explicit user queries such as when the selecting of picture feature is taken as a request for information about that feature.

Variants

[0075] Many variants are, of course, possible to the arrangements described above. For example, Figure 11 illustrates a variant form of the multi-modal picture in which the picture image 30 is accompanied by upper and lower information bars 100 and 101 respectively. The upper information bar 100 indicates the narrators associated with the picture whilst the lower information bar 101 indicates what types of general and specific queries are available for use. These information bars assists the user in appreciating what queries can be put and to whom. To further assist the user, speaking (or clicking on) a narrator's name is preferably arranged to indicate what hotspots are associated with that narrator and where these hotspots are located in the image - thus, in Figure 11, narrator "Vivian" has been selected and hotspot 32 is indicated on the picture image by a dashed hotspot boundary line. Query types used by "Vivian" can also be indicated by highlighting these types (in the Figure 11 example, "Vivian" has used general query type 'description' and specific query types 'what' and 'story'). Preferably, once a narrator has been selected, that narrator remains selected until a different (or all) narrators is subsequently selected, whereby with a narrator selected, only the responses of that narrator will be used in responding to the users' queries. However, it is also possible to arrange for selection of a narrator to be effective only for a single query. In this latter case, a convenient way of providing for user selection of a narrator to answer a query about a particular feature is for a list of narrators of responses about that feature to be displayed whenever the user points to that feature.

[0076] Instead of identifying narrators by name as shown in Figure 11, other forms of identifier can be used such as an image of each narrator.

[0077] As regards the generic control code, this can be provided in the form of a Java applet or any other suitable form and is not limited to the use of client-side scripts of the form described above. Furthermore, rather than the frame-set definition file 52 and control code file 55 being generic in form, they can be made specific to each multi-modal picture.

[0078] As already indicated, picture invocation can be initiated in a variety of ways. As a further example, picture invocation can be arranged to be effected by the creator sending the user a reference to the picture interaction dialog file, the user sending this reference to the voice browser to enable the latter to retrieve the dialog file; the dialog file is arranged to cause the return to the user of the frame-set definition file (or a reference to it) with the latter already including the correct source reference to the picture file as well as to the control code file.

[0079] Persons skilled in the art will appreciate that there are very many ways of implementing multi-modal pictures and the supporting functionality ranging from the multiple file approach described above to having just a single data file containing all the necessary data and arranged to be executed by specifically adapted software. It will also be appreciated that whilst in the described embodiments the course of the system-user interaction is controlled by the control logic embedded in the picture interaction dialog file and interpreted by the dialog manager 7 of the voice browser, it is possible to provide this functionality separate from the voice browser and the response scripts (thus, in general terms, the dialog block 62, 67 and 68 would be used by a separate control arrangement for determining, on the basis of the user voice and pointing-arrangement inputs, which of multiple stored responses is the appropriate one to play back to the user). Furthermore, it is not necessary to explicitly identify a selected feature from the coordinates output by the pointing arrangement as a separate step to choosing an appropriate response to a particular user query; thus, a picture feature could simply be specified in the control logic of the dialog interaction file by its coordinate values (of range of values) whereby this control logic tests coordinate values output by the pointing arrangement against coordinate values of particular features in the course of selecting a response to a user query.

[0080] With respect to the selection of a picture feature of interest to a user, in the described embodiments this has been carried out by means of a user-operated pointing arrangement whereby data input about the picture feature of interest is generated through a manual operation. In addition to the described arrangements for manually effecting

feature selection, it is possible to use an arrangement in which the coordinates of the feature of interest are manually input using a keypad (a coordinate system being displayed as part of the image presented to the user, or being presented around the boundary of the display area). Another possible feature-selection arrangement is one based on specifying, via keypad input, a particular area of the display where either the display is divided into labelled areas or there is a direct mapping between keypad keys and display areas. It is also possible to label each feature of interest in the image with a reference and have the user effect feature selection by keypad input of the appropriate reference.

[0081] Further as regards the determination of which picture feature is being selected by a user, it should be noted that determining the picture feature of interest from the image coordinates or image area identified by whatever selection arrangement is being used, can be done in ways other than that described above in which image coordinates generated by the feature-selection arrangement are mapped to picture features using predetermined mapping data as described above. For example, the image can have data encoded into it that labels particular picture features, the pointing arrangement being arranged to read this label data when over the corresponding picture feature. Technology for embedding auxiliary data in picture image data in a way that does not degrade the image for the average human viewer is already known in the art.

[0082] The picture image can be a hard-copy image carrying markings (such as infra-red ink markings) intended to be read by a sensor of a suitable pointing arrangement whereby to determine what particular picture feature is being pointed to; the markings can be a pattern of markings enabling the pointing arrangement to determine the position of its sensor on the image (in which case, an image map can be used to translate the coordinates into picture features) or the markings can be feature labels appropriately located on the image. Even without special markings added to the image, the image can still be a hard-copy image provided the image is located in a reference position relative to which the pointing arrangement can determine the position of a pointing element (so that an image map can be used to translate pointing-element position into picture features). Other manually-operated selection arrangements, such as those based on explicit coordinate input via a keypad, can also be used.

[0083] Where image map data is used to translate image coordinates to picture features, the image map data can be held separately from the picture file and only accessed when needed; this facilitates updating of the image map data. A reference (e.g. URL) to the image map data can be included in the picture file or, where the image is a hard-copy image, in markings carried by the image.

[0084] Preferably, the described embodiments are applied to pictures of scenes and places in the real world such as a tourist might take with a camera. However, it is also possible to apply the embodiments to topographic pictures that are primarily intended to convey map-type information.

APPENDIX

This Appendix forms an integral part of the description.

Example of Multi-Modal Picture file.

The following HTML file was generated using Microsoft FrontPage and is the Multi-Modal Picture file 45 for the picture depicted in Figure 6; the HTML code pulls in image file "beachPicture.jpg" and defines three "hotspots" which when clicked result in a call to a function (method) of the control code file 55 with the passing of a parameter value indicating which hotspot has been clicked on.

File Name: MMPicture1.htm

```
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=windows-1252">
<meta name="GENERATOR" content="Microsoft FrontPage 4.0">
<meta name="ProgId" content="FrontPage.Editor.Document">
<title>Frame Page</title>
</head>
<body>
<p><map name="FPMap0">
<area href="http://localhost/mmpic1/dpfBlank.htm"
onclick="parent.parent.scriptFrame.pushParams('object=coconutTree')"
target="invisible" coords="128, 84, 253, 236" shape="rect">
<area href="http://localhost/mmpic1/dpfBlank.htm"
onclick="parent.parent.scriptFrame.pushParams('object=frenchTourist')"
target="invisible" shape="rect" coords="204, 266, 219, 282">
<area href="http://localhost/mmpic1/dpfBlank.htm"
onclick="parent.parent.scriptFrame.pushParams('object=frenchTourist')"
target="invisible" shape="rect" coords="309, 257, 324, 269"></map>
<img border="0" src="beachPicture.jpg" align="center" usemap="#FPMap0"
width="569" height="375"></p>
</body>
</html>
```

Example of Picture Interaction Dialogue file

[0085] The following is an example implementation of the Figure 8 picture interaction dialog file 46. In this example, there is one narrator called Roger. The user can say things like:"Roger, what is the date?".

```
<talkml>
    <!--variables declaration-->
    <create name="object"          expr="'undefined'"/>
    <create name="action"          expr="'what'"/>
    <create name="type"        expr="'generic'"/>
    <!-- Specify here the url of the display webserver-->
    <display-server url="localhost"/>

    <dialog name="welcome">
        <say> Welcome to the digital picture frame. </say>
        <!-- Display the Multi Modal picture of beach. -->
        <display-show url="MMPicture1.htm"/>
        <say>
            This picture was sent to you by Guillaume. You can talk to Roger to
            ask for information about it.
        </say>
        <goto next="getUserRequest"/>
    </dialog>

    <dialog name="getUserRequest">
        <play src="sound1.wav"/>
        <display-listen wait_after="2000"/>
        <speech wait_after="2000">
            sentence =  narrator question | question | exit {action='exit'};
            question =  genericQuestion {type='generic'} | specificQuestion
                        {type='specific'};
```

```
            narrator = roger {narrator='roger'};
            roger    = "roger";
            genericQuestion = date {action='date'}| location {action='location'}|
                              description {action='description'};
            date     = "what is the date";
            location = "what is the location";
            description = "describe the picture";
            specificQuestion= what {action='what'} | story {action='story'};
            what     = "what is it";
            story    = "any story";
            exit     = "exit";
        </speech>

        <recognised>
            <play src="sound2.wav"/>
            <goto next="filterRequest"/>
        </recognised>

        <unrecognised>
            <play src="sound2.wav"/>
            <say> I am sorry, I have some trouble understanding you.</say>
            <retry/>
        </unrecognised>

</dialog>


<dialog name="filterRequest">
        <if cond="action='exit'">
            <goto next="exit"/>
        </if>
```

```
        <if cond="type='generic'">
            <if cond="action='description'"><goto next="description"/></if>
            <if cond="action='date'"><goto next="date"/></if>
            <if cond="action='location'"><goto next="location"/></if>
        </if>

        <if cond="type='specific'">
            <if cond="action='what'">    <goto next="what"/>  </if>
            <if cond="action='story'">    <goto next="story"/></if>
        </if>

</dialog>


<dialog name="description">
        <say>
            This is a picture of the beach.
        </say>
        <goto next="getUserRequest"/>
</dialog>


<dialog name="date">
        <say>
            Guillaume took this picture during his last trip home to
            Martinique last November
        </say>
        <goto next="getUserRequest"/>
</dialog>


<dialog name="location">
```

```
        <say>
            This picture was taken in Martinique, an island located in the
            French West Indies.
        </say>
        <goto next="getUserRequest"/>
    </dialog>

    <dialog name="what">
        <if cond="object='coconutTree'"><goto next="coconutTreeWhat"/></if>
        <if cond="object='frenchTourist'"><goto next="frenchTouristWhat"/></if>
        <say>
            Apparently, you did not click on any parts of the picture. Try
            again.
        </say>
        <goto next="getUserRequest"/>
    </dialog>


    <dialog name="story">
        <if cond="object='coconutTree'"><goto next="coconutTreeStory"/></if>
        <if cond="object='frenchTourist'"><goto next="frenchTouristStory"/></if>
        <say>
            Apparently, you did not click on any parts of the picture. Try
            again.
        </say>
        <goto next="getUserRequest"/>
    </dialog>

    <dialog name="frenchTouristWhat">
        <say>
            I guess this is a French tourist.
        </say>
```

```
        <goto next="getUserRequest"/>
    </dialog>

    <dialog name="coconutTreeWhat">
        <say>
            This is a coconut tree. This kind of trees is very common in the
            Caribbean.
        </say>
        <goto next="getUserRequest"/>
    </dialog>

    <dialog name="frenchTouristStory">
        <say>
            Tourism is one of the main source of revenue in Martinique. A lot
            of people come from France when it is winter over there.
        </say>
        <goto next="getUserRequest"/>
    </dialog>

    <dialog name="coconutTreeStory">
        <say>
            When I was little,I used to get coconuts from my garden. There is
            a sweet liquid inside that is very tasty. We use coconut to make
            ice creams and all sorts of nice desserts.
        </say>
        <goto next="getUserRequest"/>
    </dialog>
```

```
<dialog name="exit">
    <say> Thanks for using this digital picture frame. </say>
</dialog>

</talkml>
```

Claims

1. A system for presenting information concerning a picture to a user, the system comprising:

   - a data store for holding responses, specific to said picture, in respect of specific user queries concerning particular picture features;
   - a manually-operable feature-selection arrangement for enabling a user to select a feature in a displayed image of the picture, and for providing an output indication regarding what said particular feature, if any, the user has thereby selected;
   - a voice dialog input-output subsystem including a speech recogniser for interpreting queries from a user;
   - a control arrangement responsive to a user selecting a said particular feature and asking a specific query regarding that feature, to output the corresponding stored response.

2. A system according to claim 1, wherein image-map data is associated with the picture image for mapping image coordinates to said particular features, the selection arrangement being arranged to use the image-map data to determine what picture feature is selected by the user.

3. A system according to claim 1, wherein said image includes label data positioned in the region of a said particular picture feature to indicate the identity of that feature, the selection arrangement being arranged to read the label data to determine what picture feature is selected by the user.

4. A system according to any one of claims 1 to 3, further comprising a display subsystem for displaying said image provided to it in the form of digital image data.

5. A system according to any one of claims 1 to 3, wherein the picture image is a hard-copy image.

6. A system according to any one of the preceding claims, wherein the data store is arranged also to hold responses concerning general queries that are not associated with any particular picture feature, the control arrangement being arranged to respond to the user voice input of a general query by returning the appropriate response.

7. A system according to any one of the preceding claims, wherein the voice interface subsystem and control arrangement are arranged to cooperate in recognising multiple different queries in respect of a particular picture feature selected by the user.

8. A system according to claim 1, wherein each of at least some of the responses is associated with a specified narrator, the system being arranged to permit a user to receive only the response of a user-selected narrator in respect of at least one query.

9. A system according to claim 8, wherein user selection of a narrator is arranged to be effected by voice input, the control arrangement being arranged to respond to the selection of a particular specified narrator by user voice input, by using only the said responses associated with that narrator in providing a response to a user query.

10. A system according to claim 9, further comprising means for displaying, along with said picture image, identifiers of narrators associated with the picture.

11. A system according to claim 9, further comprising a display subsystem for displaying said image provided to it in

the form of digital image data and identifiers of narrators associated with the picture; the display subsystem being arranged to respond to selection of a particular specified narrator by user voice input by indicating on the displayed image the said particular features for which responses are available concerning that narrator.

12. A system according to claim 8, further comprising a display subsystem for displaying said image provided to it in the form of digital image data, and identifiers of narrators associated with the picture; user selection of a narrator being arranged to be effected by the user using said selection arrangement to select a displayed narrator identifier, and the control arrangement being arranged to respond to the selection of a particular narrator by using only the said responses associated with that narrator in providing a response to a user query.

13. A system according to claim 12, wherein the display subsystem is arranged to respond to selection of a particular specified narrator to indicate on the displayed image the said particular features for which responses are available concerning the currently-selected narrator.

14. A system according to any one of the preceding claims, wherein said selection arrangement is a pointing arrangement usable by the user to point to a feature of interest in a displayed view of the picture.

15. A system according to any one of the preceding claims, further comprising a feedback arrangement arranged to respond to a user query concerning a picture feature that is other than a said particular picture feature, by outputting at least an identifier of the concerned picture feature for informing an author of the picture or its associated responses regarding the unsatisfied query.

16. A system according to any one of the preceding claims, further comprising a feedback arrangement arranged to respond to a user query that although concerning a said particular picture feature is not a said specific query, by outputting an indication of the query and an identifier of the concerned picture feature for informing an author of the picture or its associated responses regarding the unsatisfied query.

17. A system according to claim 15, wherein the feedback arrangement is further arranged to output an indication of the user query together with said picture feature identifier.

18. A system according to any one of claims 15 to 17, wherein the picture feature identifier arranged to be output by the feedback arrangement comprises an indication of the picture concerned and image coordinates of the picture feature.

19. A system according to any one of claims 15 to 17, further comprising means for converting digital picture data into said displayed image of the picture, the picture feature identifier arranged to be output by the feedback arrangement comprising digital data representing a corresponding portion of the image.

20. A system according to any one of claims 15 to 17, further comprising means for converting digital picture data into said displayed image of the picture, the picture feature identifier arranged to be output by the feedback arrangement comprising digital data representing the image with a marking indicating the position of the feature in the image.

21. A system according to any one of claims 15 to 20, wherein the feedback arrangement is operative to output said picture feature identifier to a log file.

22. A system according to any one of claims 15 to 20, wherein the feedback arrangement is operative to output said picture feature identifier in the form of an electronic message addressed to a person associated with the generation of the picture or its associated responses.

23. A multi-modal picture specified by data held on at least one data carrier, this data comprising:

- picture image data for displaying a picture image;
- response data indicative of voice responses intended to be given to specific user queries concerning particular picture features of the picture;
- first control data for enabling a determination to be made as to which said particular feature in the picture image, if any, a user is selecting when using a selection arrangement to indicate a feature in the displayed image; and
- second control data for determining, on the basis of a spoken user query and on which said particular picture

feature is selected by the user using the selection arrangement, which said response is to be used to reply to the user query.

24. A multi-modal picture according to claim 23, wherein the first control data comprises image-map data mapping image coordinates to said particular features.

25. A multi-modal picture according to claim 23, wherein the first control data comprises label data arranged to be positioned in the displayed image in the region of each said particular picture feature to indicate the identity of that feature.

26. A multi-modal picture according to claim 23, wherein at least some of the said responses are associated with a narrator identified in the response data, the second control data enabling the determination of which said response is to be used to reply to the user query to be restricted to those responses associated with a said narrator that has been selected by the user.

27. A multi-modal picture according to any one of claims 23 to 26, wherein the picture is of a non-topographic real-world scene.

28. A multi-modal picture comprising a hard-copy picture image, and data held on at least one data carrier, this data comprising:

- response data indicative of voice responses intended to be given to specific user queries concerning particular picture features;
- first control data for enabling a determination to be made as to which said particular feature in the picture image, if any, a user is selecting when using a selection arrangement to indicate a feature of the image; and
- second control data for determining, on the basis of a spoken user query and on which said particular picture feature is selected by the user using the selection arrangement, which said response is to be used to reply to the user query.

29. A multi-modal picture according to claim 28, wherein the first control data comprises image-map data mapping image coordinates to said particular features.

30. A multi-modal picture according to claim 28, wherein the first control data comprises label data arranged to be positioned in or on the image in the region of each said particular picture feature to indicate the identity of that feature.

31. A multi-modal picture according to claim 28, wherein at least some of the said responses are associated with a narrator identified in the response data, the second control data enabling the determination of which said response is to be used to reply to the user query to be restricted to those responses associated with a said narrator that has been selected by the user.

32. A multi-modal picture according to any one of claims 28 to 31, wherein the picture is of a non-topographic real-world scene.

33. Apparatus for authoring a multi-modal picture, comprising:

- a first tool for defining image hotspots associated with particular picture-image features;
- a second tool with speech recognition capability, for recording user responses input by voice, to user-specified queries each associated with a particular said picture-image feature; and
- means for automatically generating control data for determining, on the basis of a spoken user query and on which said particular picture feature is selected by a user, which said response is to be used to reply to the user query.

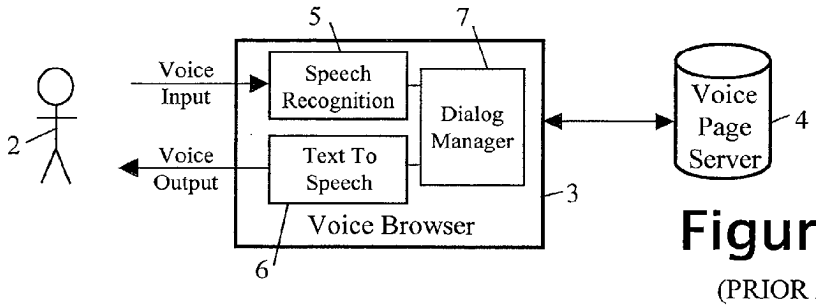


Figure 1
(PRIOR ART)

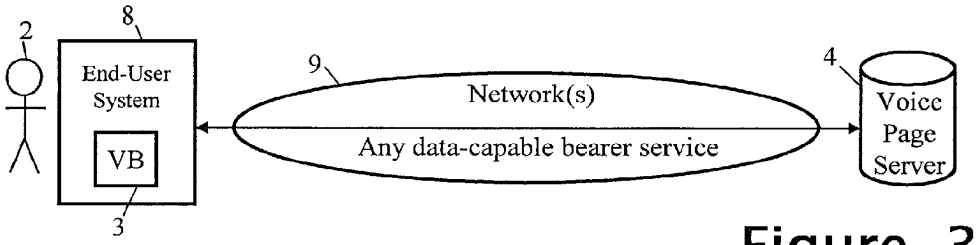


Figure 3
(PRIOR ART)

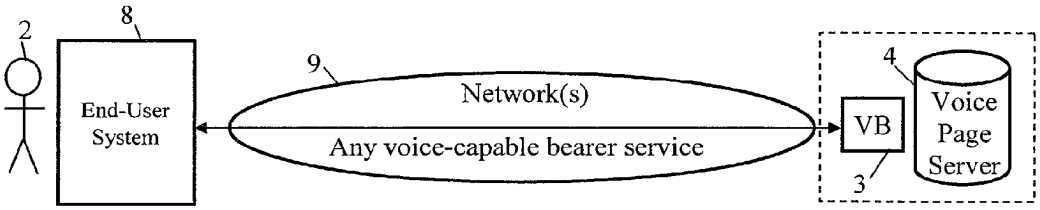


Figure 4
(PRIOR ART)

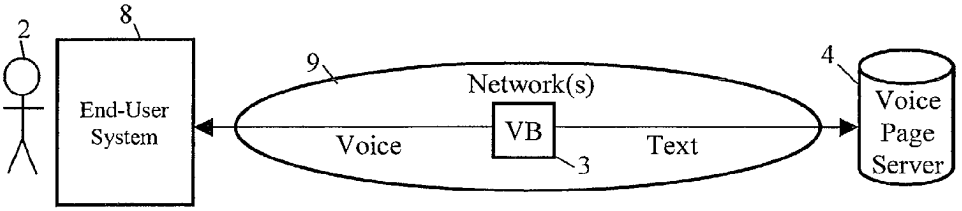


Figure 5
(PRIOR ART)

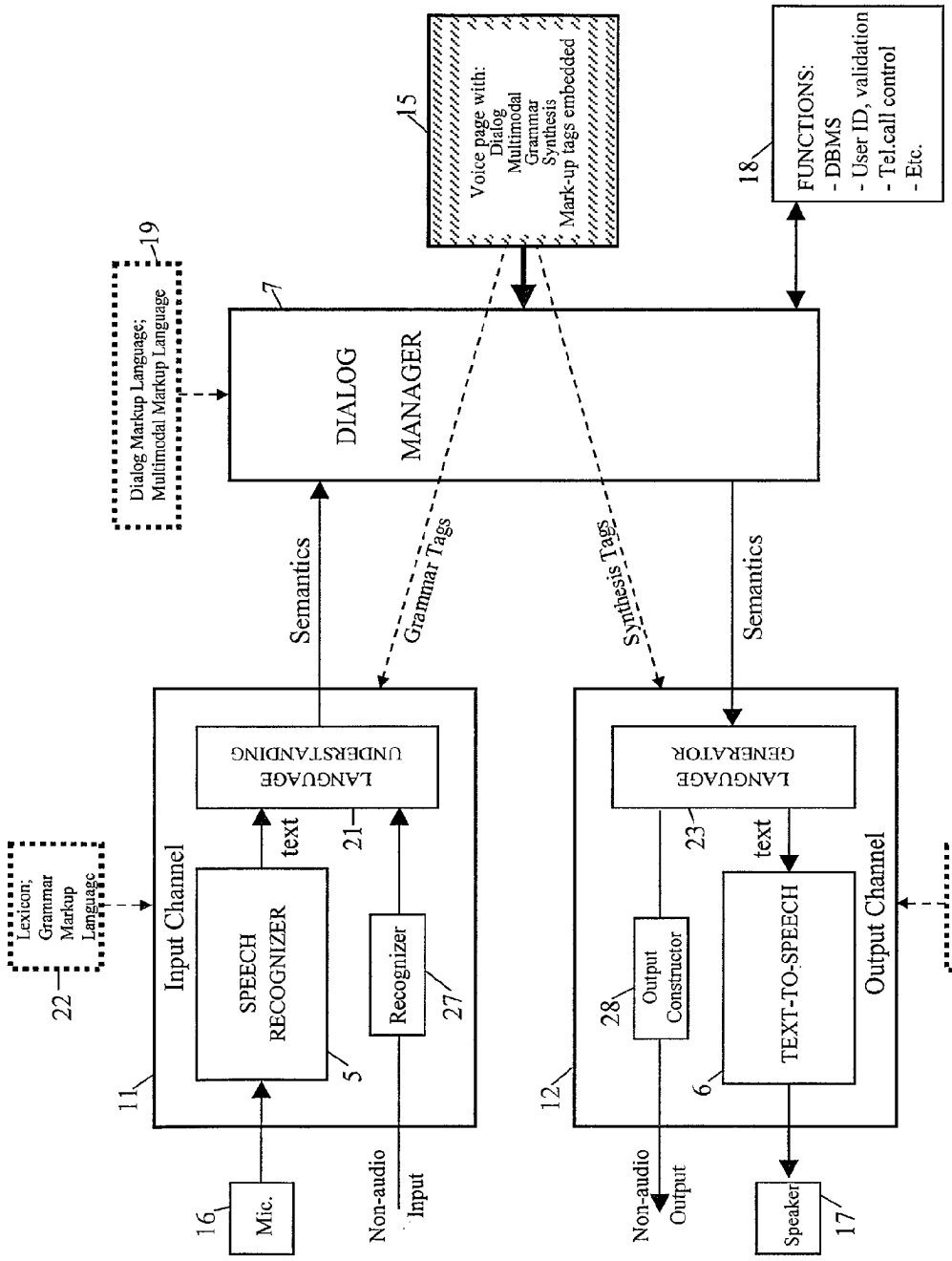


Figure 2
(PRIOR ART)

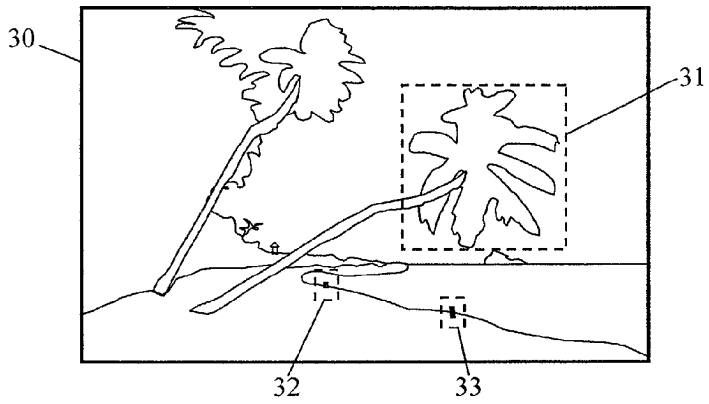


Figure 6

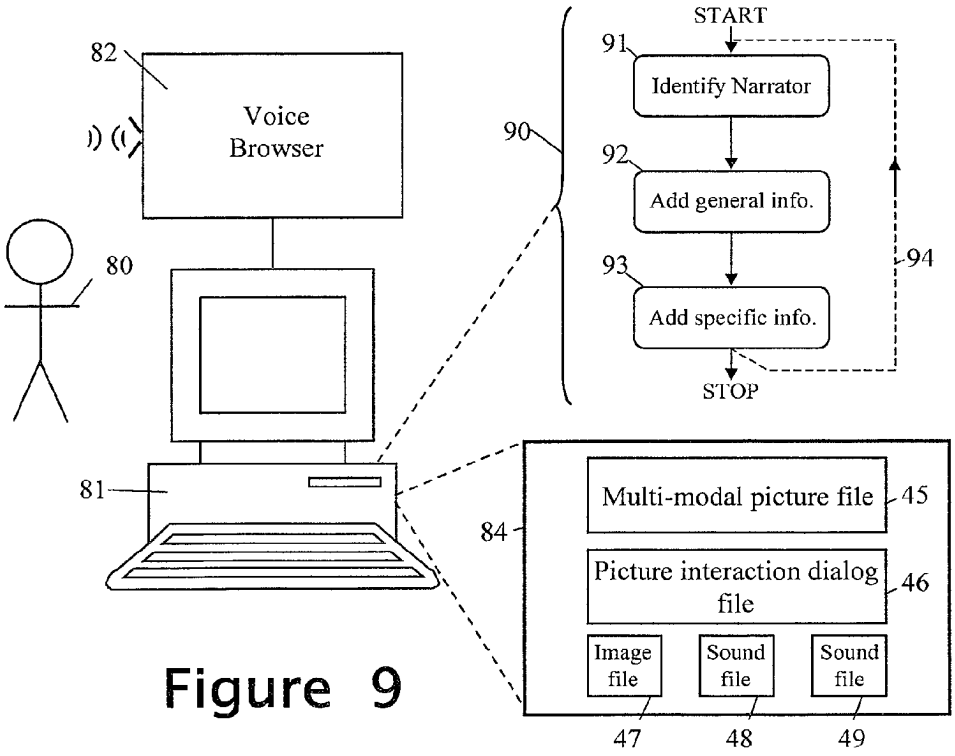


Figure 9

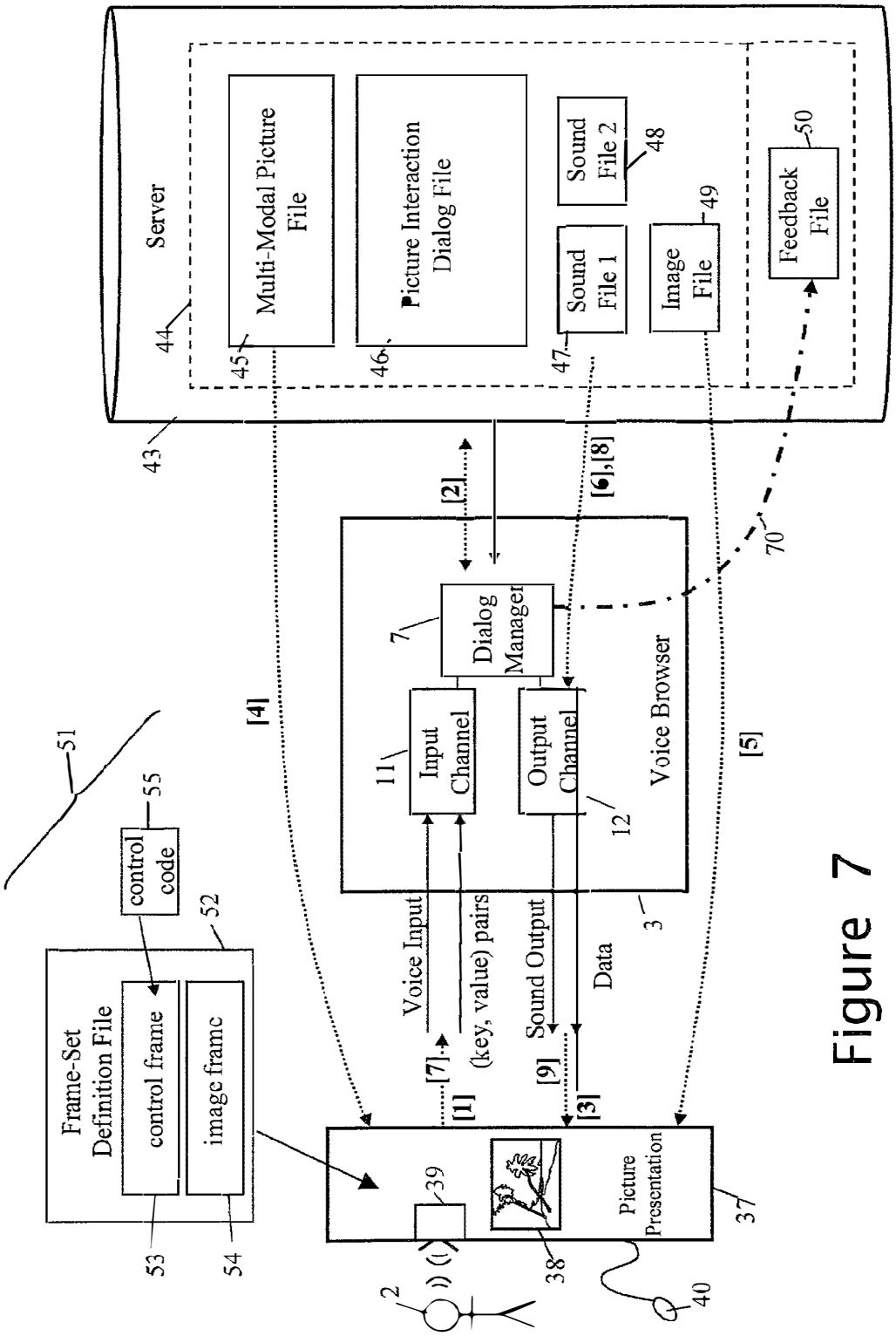
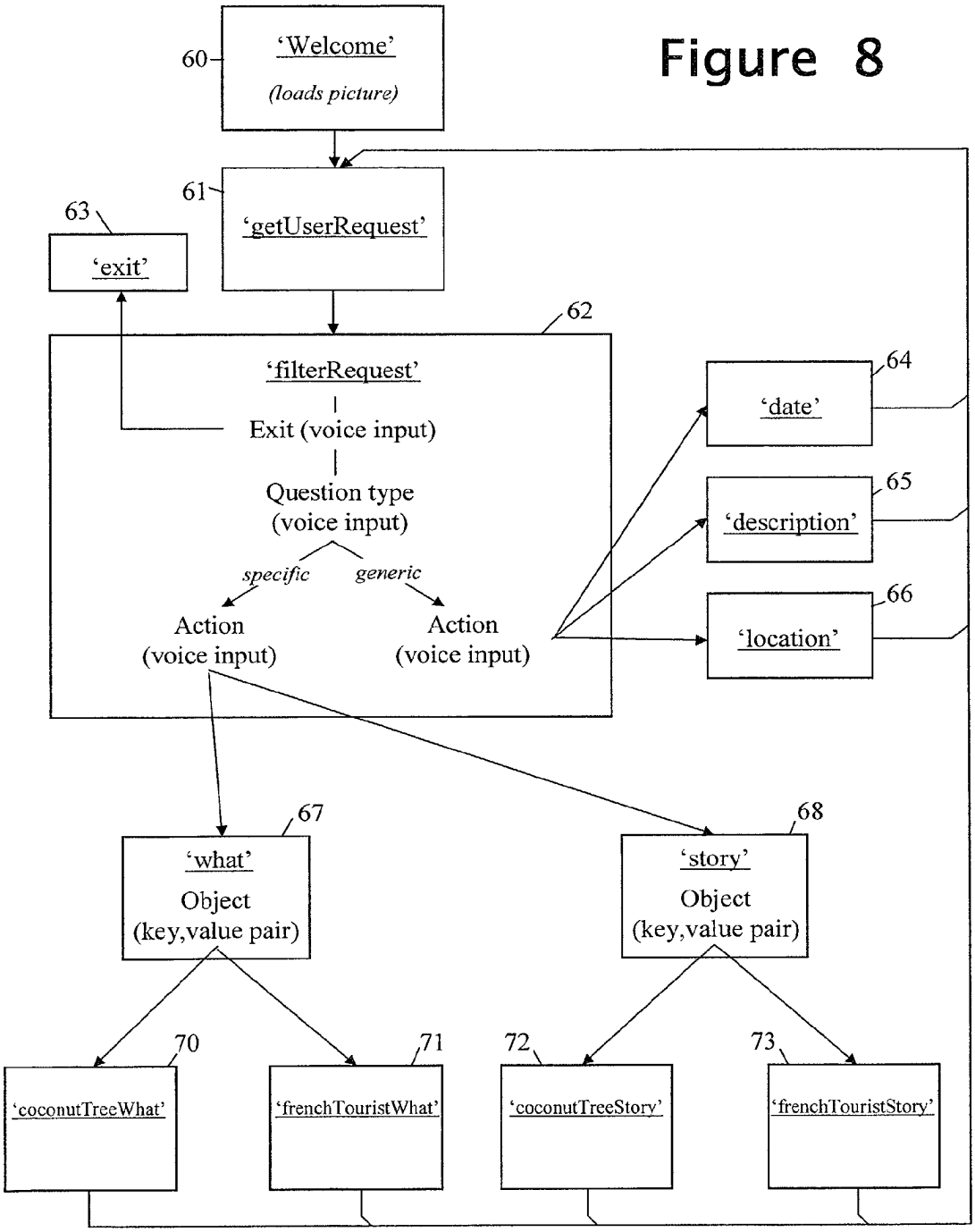

Server
44
Multi-Modal Picture File
45
Picture Interaction Dialog File
46
Sound File 1
47
Sound File 2
48
Image File
49
Feedback File
50
43
[2]
[6],[8]
70
Dialog Manager
7
Input Channel
11
Output Channel
12
Voice Browser
3
[4]
[5]
51
control code
55
Frame-Set Definition File
52
control frame
image frame
53
54
Voice Input
(key, value) pairs
Sound Output
Data
[7]
[1]
[9]
[3]
39
38
Picture Presentation
37
2
40

Figure 7

Figure 8


60
'Welcome'
(loads picture)
61
'getUserRequest'
63
'exit'
62
'filterRequest'
Exit (voice input)
Question type
(voice input)
specific
generic
Action
(voice input)
Action
(voice input)
64
'date'
65
'description'
66
'location'
67
'what'
Object
(key,value pair)
68
'story'
Object
(key,value pair)
70
'coconutTreeWhat'
71
'frenchTouristWhat'
72
'coconutTreeStory'
73
'frenchTouristStory'

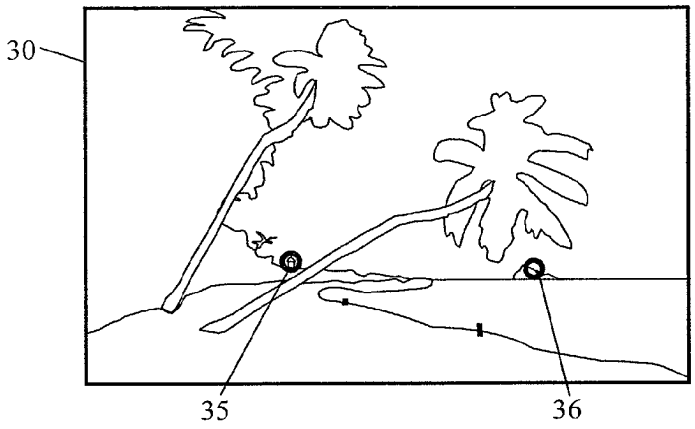


Figure 10

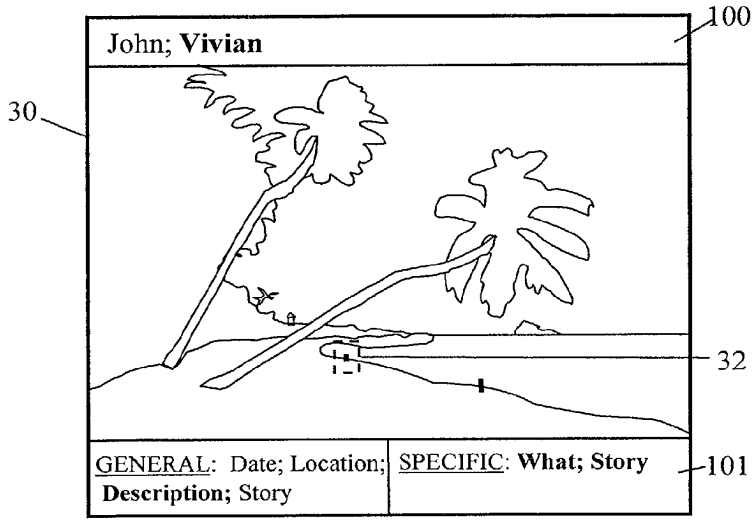


Figure 11

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 1 646 037 B1

(12) EUROPEAN PATENT SPECIFICATION

(45) Date of publication and mention of the grant of the patent:
26.03.2008 Bulletin 2008/13

(51) Int Cl.:
G10L 15/26 *(2006.01)*

(21) Application number: **05291989.1**

(22) Date of filing: **23.09.2005**

(54) Method and apparatus for enhancing speech recognition accuracy by using geographic data to filter a set of words

Verfahren und Vorrichtung für erhöhte Spracherkennungspräzision unter Verwendung geografischer Daten zur Filterung einer Wortgruppe

Procédé et appareil pour améliorer la précision de la reconnaissance vocale en utilisant des données géographiques pour filtrer un lot de mots

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HU IE IS IT LI LT LU LV MC NL PL PT RO SE SI SK TR

(30) Priority: **29.09.2004 US 953671**

(43) Date of publication of application:
12.04.2006 Bulletin 2006/15

(73) Proprietor: **FRANCE TELECOM**
75015 Paris (FR)

(72) Inventors:
- **Waters, Keith**
  West Newton
  MA 02465 (US)
- **Simon, Benoît Benjamin**
  Boston
  MA 02114 (US)

(56) References cited:
EP-A- 0 661 688
EP-A- 1 233 407
US-B1- 6 230 132

Note: Within nine months from the publication of the mention of the grant of the European patent, any person may give notice to the European Patent Office of opposition to the European patent granted. Notice of opposition shall be filed in a written reasoned statement. It shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

EP 1 646 037 B1

Description

FIELD OF THE INVENTION

[0001] The invention relates generally to speech recognition systems and methods and, more particularly, to systems and methods for enhancing speech recognition accuracy by using geographic data to reduce the size of a set of words that must be checked to effect speech recognition.

BACKGROUND OF THE INVENTION

[0002] While visiting an unfamiliar city, a mobile handset user may wish to locate a business, landmark, street, or other place of interest. It would be quite advantageous if the user could place a voice call to a remote server at a designated telephone number, speaking his or her request for information into the handset. The handset would execute a speech recognition procedure, transforming the spoken request into a server query, and receiving the requested information from the remote server in the form of an electronic signal for producing audible speech. For example, a user in Manhattan may wish to locate the closest Neiman Marcus store. She dials a designated telephone number and speaks a request such as "where is the closest Neiman Marcus?" The handset uses speech recognition to convert the request into an electronic query for receipt by the remote server. Based upon the current location of the handset, the remote server accesses a database to retrieve the address and phone number of the nearest Neiman Marcus. The address and phone number are sent from the server to the user's handset in the form of an electronic signal, whereupon the handset converts the electronic signal into audible speech.

[0003] Unfortunately, there are significant hurdles to overcome if one wishes to use existing speech recognition algorithms to carry out the foregoing procedures. More specifically, it is a challenge to develop speech recognition algorithms, which provide a desired level of accuracy in certain operational environments, such as that of mobile telephonic devices. In general, the processing power and memory capabilities of mobile telephonic devices are limited by size and cost constraints, thereby placing limits on the extent to which the complexity of existing algorithms can be increased. Moreover, the existence of ambient noise severely degrades the performance of many speech recognition algorithms. In view of these shortcomings, prior art techniques have been developed to improve the performance of speech recognition algorithms in applications where processing power may be limited. These prior art techniques are directed toward enhancing a recognition model by utilizing improved acoustic models. The following definitions are presented herein for purposes of understanding and clarity:

[0004] *Lexical Model:* A lexical model breaks down individual words of a syntactic model (defined hereinafter) into a sequence of phonemes (also defined hereinafter).

[0005] *Acoustic Model:* An acoustic model breaks down the sequence of phonemes determined by the lexical model into a Markov model of cepstrum coefficients, also termed vector sequences. Cepstrum coefficients are generated by taking the Fourier Transform of the power spectrum of a speech signal.

[0006] *Recognition Model:* A recognition model is a set of rules and operations that are used by a recognition engine to recognize individual words. The recognition model includes a fully deployed speech recognition system which encompasses the acoustic model, the lexical model and the syntactic model as follows: recognition model = (acoustic model) x (lexical model) x (syntactic model).

[0007] U.S. Patent Application Publication No. US2003236099 discloses a technique for improving the accuracy of a speech recognition algorithm by selecting an appropriate acoustic model in accordance with the location of a mobile telephonic device ("mobile handset", or just "handset" for the sake of convenience). As indicated above, an acoustic model includes a set of cepstrum coefficient statistics for each phoneme in a language. Several different acoustic models can be precompiled, each representing a unique acoustical environment, such that a speech recognition system can select an acoustic model appropriate for a given set of environmental conditions. For example, each acoustic model is prepared by individually recording several hundred speakers in a controlled, model-specific laboratory environment that has known acoustical properties illustratively specified in terms of noise, echo, reverberation, and frequency response. Each model-specific laboratory environment is intended to represent a real-world environment, such as a warehouse, an office, a shopping mall, an airport, or the interior of an automobile. Hence, acoustic models incorporate properties of specific environments as well as characteristics of the human voice. The laboratory environment may include artificially generated sounds to simulate, for example, traffic, conversations, computer keyboard clicks, and forklift trucks.

[0008] Pursuant to this prior art technique, position information is retrieved from a location-based service (LBS), as defined hereinafter, based upon the current location of the mobile handset. The retrieved position information is then employed to select the acoustical properties of the environment surrounding the handset. An acoustic model that approximates the acoustical properties of the type of environment for the current location of the handset is selected from a set of previously prepared acoustic models. The selected acoustic model is then utilized to improve the accuracy of the speech recognition algorithm.

[0009] The technique disclosed in U.S. Patent Application Publication No. US2003236099 presents many practical shortcomings. Existing location-based services, using the Global Positioning System (GPS), will not

determine a user's position to a requisite accuracy of, say, one to two meters, nor will such services determine a user's altitude with sufficient accuracy for many speech recognition applications. The impact of insufficient position accuracy is that, if a user is situated near the entrance of a building, it is not possible to accurately determine whether the user is, in fact, inside the building. The impact of insufficient altitude accuracy is that it is not possible to determine with certainty whether a user is within the building or standing on the roof. An acoustic model that is appropriate for an interior building lobby is unlikely to be totally appropriate for use in an outdoor setting. At the very least, selection of an inappropriate acoustic model based upon an inaccurate determination of a user's actual position will degrade the accuracy of many speech recognition algorithms.

[0010] Another shortcoming of the technique disclosed in U.S. Patent Application Publication No. US2003236099 is the need for creating an acoustic model. First of all, it is not feasible to create a unique acoustic model for every type of location. Secondly, even if an acoustic model is developed for a certain type of location, an actual location may not possess the acoustical characteristics embodied in the model. For example, due to variations in construction materials, furnishings, and room configurations, an acoustic model that attempts to represent a typical conference room, for example, may not adequately represent a specific conference room in a specific office building. Accordingly, what is needed is an improved technique for improving the accuracy of speech recognition algorithms.

[0011] EP1233407 recites a navigation system with speech recognition capacities for a moving vehicle. Whenever the distance between the vehicle and known active locations becomes greater then a given threshold, a pre-existing word list, corresponding to the active location, is acquired by the navigation system for subsequent use with the speech recognition engine.

[0012] US6230132 recites another navigation system with speech recognition functions. A lexicon is generated at periodic intervals with a limited number of words so as to facilitate the speech recognition functions. Furthermore, when a user needs to use the navigation system, he/she has to make a vocal request with the name of a location so that a lexicon can be generated.

[0013] EP0661688 recites a single device navigation apparatus, on board a car, with speed recognition functions. Pre-existing libraries are used and updated at regular intervals to facilitate the speech recognition engine.

SUMMARY OF THE INVENTION

[0014] The present invention is defined by the appended independent claims.

[0015] Other aspects and features of the present invention will become apparent from the following detailed description considered in conjunction with the accompanying drawings. It is to be understood, however, that the drawings are designed solely for purposes of illustration and not as a definition of the limits of the invention, for which reference should be made to the appended claims. It should be further understood that, unless otherwise indicated, the drawings are merely intended to conceptually illustrate the structures and procedures described herein.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] In the drawings:

[0017] FIG. 1 is a diagrammatic representation of a geographic area of interest.

[0018] FIG. 2 is a data structure diagram setting forth a portion of an illustrative element database.

[0019] FIG. 3 is a data structure diagram setting forth an illustrative subset of words selected from the element database of FIG. 2.

[0020] FIG. 4 is a generalized information flow diagram showing a set of fundamental operations for the present invention.

[0021] FIG. 5 is a detailed block diagram setting forth information flow for a speech recognition system constructed in accordance with a preferred embodiment of the present invention.

DETAILED DESCRIPTION OF THE DRAWINGS

[0022] The following terms used herein are defined as follows:

[0023] *Geographic Area of Interest:* Illustratively defined in terms of a zip code, group of zip codes, cell identification (Cell-ID) information, caller ID (Call-ID) information, the boundaries of a state, county, city, town, place, or other governmental or municipal subdivision, standard USGS (United States Geological Survey) grids, locations within a fixed range of latitude and longitude, or locations within a specified geometric boundary. The geographic area of interest may, but need not, be set automatically based upon the present location of a handset. The geographic area of interest may, but need not, be designated by a user based upon manual, verbal, or user-specified input. The geographic area of interest can be a coarse geographic area of interest or a refined geographic area of interest, with the latter being a portion of the former.

[0024] *E911:* As mandated by the FCC (Federal Communications Commission), a standard for a system by which wireless carriers are able to pinpoint the location of a mobile device in an emergency.

[0025] Element: Name of at least one of a street, a business, a merchant, an individual's name, a place, a point of interest, a transportation facility, an individual household, an activity, and a landmark.

[0026] *GPS*: Global Positioning System -- Enables geographic position determination at a land-based portable device equipped to receive signals from three or more orbiting GPS satellites.

[0027] *LBS:* Location-Based Service -- Any service

that obtains information concerning the current geographical location of a mobile device user. Illustrative examples of location-based services include maps displayed in an automobile that track a motorists' position using GPS, or automatic localization services provided by dialing 911 on an E911-compliant mobile telephone network.

[0028] *Mobile Device:* Any portable device enabled for wireless communication, such as a wireless telephone, a personal digital assistant, or a laptop computer equipped with a wireless modem.

[0029] *Phoneme:* A phoneme represents a unique human uttered sound. Individual words, as pronounced by a human speaker, consist of a sequence of phonemes.

[0030] *Syntactic Model:* A set of rules and operations that specify allowed word sequences (phrases) of one or more words for subsequent use in a speech recognition procedure. Allowed word sequences are set forth in a list of static allowed phrases. In addition, the phrases can be optionally weighted by occurrence probabilities. Otherwise, word sequences can be in the form of stochastic or probabilistic lists, such as a Markov model of word sequences. Syntactic models are usually designed to be specific to a particular speech engine.

[0031] *Word:* A unit of language that native speakers can identify, and which communicates a meaning. Words are the building blocks from which phrases and sentences are constructed. Illustratively, for elements, "The Empire State Building" is the name of an element that contains four words. A street with the name "125th Street" includes the six words "one hundred and twenty fifth street". The street name "K Street" has the words "K" (which is a word for purposes of this invention) and "street".

[0032] The invention presumes that a very large element database is available for some sizeable, designated geographic area, such as a city, state, county, or country. One example of a known element database is TIGER (Topologically Integrated Geographic Encoding and Referencing system), available from the United States Census Bureau. TIGER compounds localized street names, rivers and railroads into a series of large regional databases.

[0033] The element database contains a set of words for elements at locations that are within the designated geographic area. In accordance with an aspect of the invention, enhanced speech recognition accuracy is provided by using geographic data to automatically select a subset of the words in the element database for use with a speech recognition procedure. The subset of words is selected by selecting only those words in the element database that are situated within a geographic area of interest. The geographic area of interest, either coarse or refined (depending on the circumstances, as explained below) is illustratively defined in terms of a zip code, group of zip codes, cell identification (Cell-ID) information, caller ID (Call-ID) information, the boundaries of a state, county, city, town, place, or other governmental or municipal subdivision, standard USGS (United States Geological Survey) grids, locations within a fixed range of latitude and longitude, or locations within a specified geometric boundary.

[0034] By selecting a subset of words, the total number of words used in the speech recognition procedure is substantially reduced. For example, a large element database containing street and business names used throughout all five boroughs of New York City can be reduced to a substantially shorter list of street and business names covering only the Grand Central Station area by specifying zip code 10017. Such a reduction in the total number of potential street names to be recognized, significantly improves speech recognition accuracy.

[0035] Pursuant to a further aspect of the invention, position information is obtained from an LBS based upon the current location of a mobile handset, illustratively using any of GPS coordinates, zip codes, and cellular base station triangulation. The obtained position information can be utilized to automatically define a coarse geographic area of interest. Let us say, as an example, that zip codes are used for this purpose. Thus, once the current location of the handset is determined, the zip code which includes that location automatically becomes the coarse geographic area of interest. Typically, the system can determine which localization method to use based on a knowledge of call configuration. For example, when a call is initiated, the Cell-ID can be used to determine the approximate location of the device. Alternatively, the application itself can choose which methods to employ, or the application can prompt the user to select which method to use. An illustrative prompt might include "*Do you want to use Cell-ID, or GPS to determine you location?*"

[0036] In accordance with a further aspect of the invention, an interface enables a user to refine the coarse geographic area of interest. For example, the user can select a specified geometrical shape within the coarse geographic area centered about the current location of the handset. This can be accomplished using key-based entries. Also, a graphical user interface can be used to draw a geometrical shape within the coarse geographic area. Additionally or alternatively, a Cell-ID can be specified within the geographic area zip code, thus defining a refined geographic area of interest for which the subset of words is to be selected. Combining automatic definition of a coarse geographic area of interest with additional, user-selectable geometric refinement is advantageous because it further reduces the total number of words that must be used in the speech recognition procedure. For example, a user-selectable rectangular region of 300 x 300 feet may contain only a few dozen elements that need to be recognized.

[0037] FIG. 1 is a diagrammatic representation of a geographic region of interest contained within a large area 100, such as a city. The current geographic location of a mobile device is determined to be point 401. In this illustrative example, the coarse geographic area of inter-

est is defined by the zip code. Thus, the coarse geographic area of interest is defined by Zip code region 470 within which point 401 is located. A mobile device user has selected a specific radius 407 around point 401 for defining the refined geographic area of interest. Elements contained (in whole or in part) within this coarse geographic area include Street A 420, Street B 422, and a first building 406. Elements contained within this coarse geographic area of interest but which are not within this refined geographic area include Street C 424, second building 404, and third building 402.

[0038] An element database (FIG. 2, 500) includes a list of element names for at least one of streets, businesses, merchants, places, points of interest, transportation facilities, individual households, activities, or landmarks in large area 100 (FIG. 1). This database can be part of the LBS database. Each element name is associated with one or more corresponding parameters for identifying or associating the element with a refined geographic area of interest, and a parameter for identifying or associating the element with a coarse geographic area of interest. A set of phonemes corresponding to the element name can be computed in real time by recognition model compiler 213, using the outputted element name (501) in conjunction with a phonetizer. A phonetizer is a mechanism that provides phonemic transcriptions for each of a plurality of element names. Phonemic transcriptions comprise a phoneme, or a sequences of two or more phonemes, which indicate the manner in which an element name may be spoken. Illustratively, the element name "Madison" has a corresponding phonemic transcription "mäd è sun".

[0039] Element names are stored in an element name field 501. Each name in element name field 501 is associated with a corresponding parameter field (x,y) for the refined geographic area of interest 503, and a corresponding parameter field for the coarse geographic area of interest 505. These element names may or may not include elements that are within whatever refined geographic area of interest is selected by the user during operation of the invention (in this example, defined by radius 407 around point 401 of FIG. 1).

[0040] Refined geographic parameters are stored in the parameter field for the refined geographic area of interest 503 (FIG. 2). A refined geographic parameter sets forth a characteristic of an element in terms of how the refined geographic area of interest is defined. For example, if the refined geographic area is defined by the user as a shape (e.g., a circle) around the current location of the handset, the refined geographic parameters comprise (x,y) coordinates. In the illustrative example of FIG. 2, the refined geographic parameters are (x,y) coordinates determined with respect to an arbitrarily assigned reference point. When a first set and a second set of (x, y) coordinates are present in the parameter field for the refined geographic area of interest 503, this represents a line segment between the first set of (x,y) coordinates and the second set of (x,y) coordinates. Such a segment is used, for example, to define a street that extends from the first set of (x,y) coordinates to the second set of (x, y) coordinates.

[0041] Coarse geographic parameters are stored in the parameter field for the coarse geographic area of interest 505. A coarse geographic parameter sets forth a characteristic of an element in terms of how the coarse geographic area of interest is defined. For example, if the coarse geographic area of interest is defined in terms of one or more zip codes, the coarse geographic parameters in the parameter field for the coarse geographic area of interest 505 comprise one or more zip codes. For example, if Street A 420 (FIG. 1) crosses into two different zip code regions, then the parameter field for the coarse geographic area of interest 505 includes two zip codes, such as 10017 and 10036. However, other types of coarse geographic parameters could be employed to relate the element to a zip code system, such as standard USGS (United States Geological System) grids, regions within a fixed range of latitude and longitude, state boundaries, county boundaries, city boundaries, and/or boundaries of urbanized areas. For example, a lookup table could be used to determine the foregoing relationship.

[0042] Returning now to FIG. 1, and in accordance with one feature of the invention, only words corresponding to elements within the geographic area of interest are extracted from element database 500 (FIG. 2) to form word subset list 600 (FIG. 3). If the user defines the refined geographic area of interest as shown in FIG. 1, the limited set of words does not include Street C 424, second building 404, and third building 402. Use of this limited set of words enhances the accuracy of a subsequently performed speech recognition procedure because words pertaining to each and every element of large area 100 of FIG. 1 need not be considered.

[0043] FIG. 3 is a data structure diagram setting forth a portion of an illustrative word subset list 600. The names of one or more elements within the geographic area of interest are stored in an element name field 601. Optionally, word subset list 600 may include additional information, such as the probability that a particular element will be incorporated into a user's query based upon the distance of that element from the current position of the mobile device, so as to provide an enhanced level of performance.

[0044] FIG. 4 is a generalized information flow diagram describing a set of fundamental operations for the present invention. Geographic data acquisition mechanism 101 (the details of which are shown in FIG. 5) is equipped to collect geographic data corresponding to a mobile device's current geographic position and, based thereon, to determine and output a geographic area of interest. Geographic data includes, but is not limited to, any of: global positioning system (GPS) coordinates in latitude and longitude, and also altitude if needed, [x,y, z] coordinates, city name, place name, zip code area, geographic area, type of area (urban, suburban or coun-

try), one or more cell identification (Cell-ID) parameters from a cellular network, or a caller identification (Call-ID) parameter from a terrestrial telephone network.

[0045] Word filtering mechanism 102 is equipped to receive a geographic area of interest from geographic data acquisition mechanism 101 and, based thereupon, select a subset of words from element database 500 (FIGs. 2 and 4). If a coarse geographic area of interest is being used, then the subset of words is selected by comparing coarse geographic parameters in the parameter field for the coarse geographic area of interest 505 corresponding to the elements in element name field 501 with coarse geographic parameters used to characterize the geographic area of interest. More specifically, if the coarse geographic area of interest is zip code 10017, then all the element names in element name field 501 will be selected which are associated with zip code 10017 in the parameter field for the coarse geographic area of interest 505. A similar operation is performed if a refined geographic area of interest is used. Of course, the parameter field for the refined geographic area of interest 503 will be used instead of the parameter field for the coarse geographic area of interest 505. The refined geographic parameters are combined with the coarse geographic parameters to enhance the word filtering mechanism 102 by using a two-step filtering process. A first step uses the parameter field for the coarse geographic area of interest 505, and a second step uses the parameter field for the refined geographic area of interest 503.

[0046] Word filtering mechanism 102 places the selected subset of words into a word subset list 600 (FIGs. 3 and 4). Speech recognition mechanism 103 (FIG. 4) is equipped to receive word subset list 600 (FIGS. 3 and 4) from word filtering mechanism 102 (FIG. 4) and, based thereupon, perform a speech recognition procedure. Illustratively, speech recognition mechanism 103 may be equipped to implement a speech recognition procedure by utilizing a (syntactic model) x (lexical model) x (acoustic model) paradigm in conjunction with Markov models. Alternatively or additionally, speech recognition mechanism 103 may utilize a neural network. More generally, the invention is applicable to any speech recognition mechanism that accepts and processes words or groups of words referred to as "grammar". Accordingly, word filtering mechanism 102 includes a grammar formatting mechanism that receives the word subset list 600 and formats the list into a grammar format usable by recognition model compiler 213.

[0047] Recognition model compiler 213 may or may not be required, depending upon the specific design characteristics of speech recognition mechanism 103. Recognition model compiler 213 parses (or analyzes) all of the entries in a formatted grammar, one after the other, and then, in one or more successive stages or "passes", constructs an output code readable by speech recognition mechanism 103. Since the speech recognition procedure is based upon a geographically refined subset of words, the overall accuracy of this procedure is enhanced relative to prior art designs. Speech recognition mechanism 103 can be implemented by a server or by a mobile handset.

[0048] A speech query from a mobile device user is received by an acoustic input mechanism 104 at the mobile device. Acoustic input mechanism 104 may be implemented using a microphone or other transducer equipped to transform acoustical vibrations into electronic signals. Acoustic input mechanism 104 feeds an electronic representation of user speech to speech recognition mechanism 103. Speech recognition mechanism 103 attempts to recognize one or more words in the electronic representation of user speech that are included in word subset list 600 retrieved from element database 500. If one or more words are recognized, speech recognition mechanism 103 sends an electronic message to an output mechanism 105 at the mobile device. The electronic message includes any recognized words in the speech query received from the mobile device user that were also included in word subset list 600. Output mechanism 105 converts the electronic message into audible and/or visual form for perception by the mobile device user. Illustratively, output mechanism 105 may be implemented using a loudspeaker coupled to audio amplifier circuitry and voice synthesizer circuitry for producing an audio message including one or more recognized words. Alternatively or additionally, output mechanism 105 may be implemented using a visual display device for displaying alphanumeric characters including one or more recognized words.

[0049] FIG. 5 is a detailed block diagram setting forth information flow for a speech recognition system constructed in accordance with an embodiment of the present invention. FIG. 5 shows details of geographic data acquisition mechanism 101. The depicted operations are performed either by a remote server 240 or by mobile device 250. Information flow commences upon receipt of a "start" signal at input mechanism 201 of mobile device 250. In practice, this "start" signal can be a request issued from a multimodal web page, an HTML page, or just a phone call. Alternatively, a user may place a call on a mobile handset to a designated telephone number to access a zip code form displayed on the mobile handset. The user fills in the zip code form using DTMF keypresses on the handset and validates the entered zip code by pressing DTMF digit "1" as a "start" signal. Basically, the "start" signal triggers a request for information.

[0050] Next, at mechanism 202, a decision is made as to whether the geographic area of interest will be determined locally (i.e., at mobile device 250), or at a remote location (i.e., at remote server 240). Depending upon the specifics of a given system application, this decision can be made by issuing a query to mobile device 250 in order to ascertain the configuration and capabilities of mobile device 250.

[0051] In cases where the geographic area of interest is determined locally, any one of two approaches may be employed. Pursuant to a first approach, mobile device

250 may include an optional GPS receiver 207, whereupon acquired GPS data are forwarded to a geographic-area-of-interest defining mechanism 204 (server side, see FIG. 5). Mechanism 204 uses the GPS data corresponding to the current location of mobile device 250 to obtain coarse geographic parameters for characterizing the coarse geographic area of interest. According to a second approach, zip code data, for example, can be entered by the user at input mechanism 201 and forwarded to geographic-area-of-interest defining mechanism 204. Illustratively, zip code data are entered into input mechanism 201 by sequentially pressing keys on a DTMF (dual-tone, multifrequency) keypad. Mechanism 204 uses the entered zip code data as a coarse geographic parameter for characterizing the geographic area of interest.

[0052] In cases where the geographic area of interest is determined remotely, mechanism 202 forwards the request received from input mechanism 201 to a location-based service 203 at server 240. Alternatively or additionally, GPS data from GPS receiver 207 are conveyed to location-based service 203, if needed. In response to the received request, location-based service 203 retrieves, for example, zip code information or cell identification (Cell ID) data for mobile device 250. Cell-IDs data can be retrieved in a mobile communications network using the Automatic Number Identification (ANI) retrieved from the mobile device 250. ANI is a service that provides the receiver of a telephone call with the number of the calling phone. The method of providing this information is determined by the service provider. Cell ID data specifies the geographic location of the wireless base station currently in communication with mobile device 250 by identifying at least one of the telephone number and the Internet Protocol (IP) address assigned to mobile device 250. For example, mobile device 250 may send to location-based service 203 Cell ID data in the form of an automated number identification (ANI) parameter. Optionally or alternatively, location-based service 203 determines the geographic location of mobile device 250 by performing a triangulation procedure among a plurality of wireless base stations. The Cell ID data, Caller ID data, and/or triangulation data retrieved by location-based service 203 (ultimately through the telecommunication provider) are forwarded to geographic-area-of-interest defining mechanism 204.

[0053] Optionally, in cases where the geographic area of interest is determined remotely, LBS service 203 utilizes at least one of Caller ID data and Cell ID data to access a zip code database 243. Based upon the Caller ID data or the Cell ID data, a zip code is retrieved from zip code database 243 and then sent to geographic-area-of-interest defining mechanism 204. Although the illustrative embodiment of FIG. 5 shows zip code database 243 and location-based service 203 as separate elements, those skilled in the relevant art will easily recognize that zip code database 243 could, but need not, be incorporated into location-based service 203. Likewise, although element database 500 and location-based service 203 are shown as separate elements, those skilled in the relevant art will easily recognize that element database 500 could, but need not, be incorporated into location-based service 203.

[0054] Geographic-area-of-interest defining mechanism 204 processes incoming geographic data, illustratively comprising at least one of incoming GPS data, zip code data, Caller ID data, or Cell ID data, to generate one or more geographic parameters defining the geographic area of interest for mobile device 250. For example, as discussed above, if the incoming data are based upon zip code data, then the output of mechanism 204 is a zip code. This parameter, e.g., a zip code, is used to specify a geographic area of interest. The defined geographic area of interest may combine a coarse geographic area of interest with a refined geographic area of interest, as is illustrated in the following example. One or more zipcode(s) may be deduced from GPS coordinates, thus defining a coarse geographic area of interest. The GPS coordinates themselves, associated with an area of uncertainty, define a refined geographic area of interest. Such a defined geographic area of interest may be used to implement the two-step filtering process described above.

[0055] Optionally, as indicated by the dotted lines of FIG. 5, a user at mobile device 250 may wish to further refine the geographic area of interest defined by geographic-area-of-interest defining mechanism 204. If so, the geographic area of interest defined by geographic-area-of-interest defining mechanism 204 is referred to as a coarse geographic area of interest, and the area of interest as refined by the user is referred to as a refined geographic area of interest. The refined geographic area of interest comprises a subset of the coarse geographic area of interest. Input mechanism 201 accepts user input setting forth a geometric specification of a refined geographic area of interest. User input is illustratively received over connection line 275 in the form of keypad entries, graphical objects, sketches, textual descriptions, or vocal input. Examples of geometric specifications include squares, rectangles, octagons, triangles, circles, and freehand sketches. Geometry specification refinement mechanism 205 accepts user input from input mechanism 201, and refines the set of geographic parameters generated by geographic-area-of-interest defining mechanism 204.

[0056] The one or more geographic parameters generated by geographic-area-of-interest defining mechanism 204, as optionally refined by geometry specification refinement mechanism 205, are received by word filtering mechanism 102. Word filtering mechanism 102 is enabled to carry out its functions when two conditions are met. Firstly, mobile device 250 should be executing a speech-enabled application where the user has issued a request for information. This requirement is met upon generation of a request from input mechanism 201. Secondly, the geographic area of interest should be defined.

This requirement is met once geographic-area-of-interest defining mechanism 204 generates the above-described geographic parameters.

[0057] When enabled, word filtering mechanism 102 uses the one or more geographic parameters received from mechanism 204 or 205 to determine whether or not each of a plurality of elements in element database 500 is situated within the geographic area of interest. The element names stored in element name field 501 (FIG. 2) may include elements that are within the geographic area of interest, as well as elements that do not lie within the geographic area of interest. Accordingly, word filtering mechanism 102 (FIG. 5) compares the one or more received parameters defining the geographic area of interest against one or more parameters stored in at least one of parameter field for the refined geographic area of interest 503 or parameter field for the coarse geographic area of interest 505 (FIG. 2), to determine whether or not each of a plurality of elements in element database 500 (FIGs. 2 and 5) is situated within the geographic area of interest. For example, Street A is characterized in the parameter field for the refined geographic area of interest 503 by a first and a second set of (x,y) coordinates. Accordingly, Street A is conceptualized as including a straight line segment connecting the first and second sets of (x,y) coordinates. If any portion of this straight line segment intersects the geographic area of interest, then Street A is included in word subset list 600 (FIGs. 3 and 5).

[0058] Word filtering mechanism 102 appends to word subset list 600 only those element names within the geographic area of interest. Optionally, if the user has made any refinements to the coarse geographic area of interest, only elements that are within the refined geographic area of interest are appended to word subset list 600.

[0059] The word subset list 600 of FIGs. 3 and 5 represents an illustrative output format for word filtering mechanism 102. Alternatively or additionally, the output of word filtering mechanism 102 may be conceptualized as a computer-readable stream of words representing one or more elements in a geographic area of interest. This stream, illustratively in text format, is compatible with at least one of a speech recognition engine or a voice browser embedding speech recognition mechanism 103 (FIG. 5). Optionally, word filtering mechanism 102 may reformat at least one of word subset list 600 or the computer-readable stream of words into a format such as an XML document, to provide a subset of words compatible with speech recognition mechanism 103.

[0060] Speech recognition mechanism 103 may be implemented at server 240 or at mobile device 250. In cases where speech recognition is performed at mobile device 250, the output from word filtering mechanism 102 (illustratively in the form of a stream of words corresponding to grammar subset list 600, FIG. 3) is sent from server 240 to mobile device 250. Word subset list 600 is then compiled by the recognition model compiler 213 at mobile device 250 to provide a speech recognition model for use by a speech recognition mechanism 103 responsive to the acoustic input mechanism 104. Speech recognition mechanism 103 then forwards recognized speech to the output mechanism 105 for generating an electronic output signal representing the recognized speech. Such an arrangement is advantageous because it reduces the burden on wireless communication resources. Transmissions between the server and the mobile device are reduced because only word subset list 600 needs to be communicated to the mobile device. The speech recognition steps can be performed at the mobile device with reference to word subset list 600, thereby enhancing the speed and efficiency of the speech recognition process.

[0061] If speech recognition is performed at server 240, then recognition model compiler 213 and speech recognition mechanism 103 are implemented at server 240 instead of at mobile device 250, wherein the output of speech recognition mechanism 103 is sent from server 240 to output mechanism 105 at mobile device 250.

[0062] The operational principles of the configurations shown in FIGs. 4 and 5 are best illustrated with reference to several illustrative scenarios. These scenarios clarify the location resolution process that will govern the filtering process performed by word filtering mechanism 102 in accordance with location data to be resolved. The scenarios consider use of any of several different types of inputs by which a location context will be determined.

[0063] Scenario One: Geographic area of interest to be determined by entered zip code and/or city. Zip code and/or city are entered into mobile device 250 by the user. Possible clients used by mobile device 250 to accept the entered zip code and/or city include a voice browser and/or a multimodal client. In the case of a phone call, the client is not the phone itself, but the voice browser (from the server point of view). Location data are provided in the form of a city and/or a zip code. The geographic area of interest corresponding to this location data is determined in the form of a city or zip code. In this case, location resolution is performed locally. The zip code and/or city are entered into mobile device 250 using DTMF keypresses and/or voice. An HTTP request is issued in response to the entered information after the information is received at input mechanism 201. In this case, word subset list 600 is generated by selecting a set of words and/or grammar pertaining to the entered city and/or zip code. The following steps are performed:

[0064] 1. User connects to geographic data acquisition mechanism 101 using mobile device 250. If mobile device 250 represents a multimodal client, an HTTP request is made by the client to a document server of word filtering mechanism 102 (FIGs. 4 and 5). If mobile device 250 represents a wireless telephone, an http request is made by the voice browser ("speaking" with the telephone) to the document server of word filtering mechanism 102.

[0065] 2. A first page is loaded into mobile device 250 or loaded onto the voice browser. The page is prompting the user to enter a zip code. The zip code by either be entered by DTMF (touch tones) or by voice. The grammar used to recognize the zip code is static.

[0066] 3. The user enters the zip code into input mechanism 201. A second HTTP request is made by the client, with the zip code value as a parameter. As location-based service 203 is not employed in this scenario to resolve the location, location resolution is "local". The geographic area of interest is evaluated at geographic area of interest defining mechanism 204, and no further refinement of the geographical area of interest takes place.

[0067] 4. A second page is loaded into mobile device 250 (the "reduced" page). The document server of word filtering mechanism 102 computes reduced grammars associated with the page in the form of word subset list 600.

[0068] 5. The user can vocally interact with the page over mobile device 250, as the page is optimized for its location

[0069] Note: The geometry specification refinement mechanism 205 is not used in this scenario. If this mechanism were to be used, a "remote" location resolution must be performed so as to enable sending a map (for example) back to the multimodal client of mobile device 250.

[0070] Scenario Two: Geographic area of interest to be determined automatically using Caller ID or Cell ID.

[0071] Possible clients used by mobile device 250 for Caller Id or Cell ID include a voice browser and/or a multimodal client. Location data are provided in the form of a caller ID or cell ID parameter. The geographic area of interest is determined as a zip code or as an area ID that identifies a geographic area. Location resolution is performed remotely, not locally. Input mechanism 201 is not employed for purposes of location resolution. Mobile device 250 is programmed to issue an HTTP request while a first page is retrieved from a document server of word filtering mechanism 102. In this scenario, words and/or grammar from element database 500 are filtered to prepare word subset list 600 based upon ANI or Cell-ID information. Input mechanism 201 is completely transparent, in the sense that the request that trims the word filtering process appears at the first stage of the interaction, and there is no need of a data collecting stage, as in the previous scenario. The HTTP request can be issued by a voice browser (that provides ANI info) or a multimodal client at mobile device 250.

[0072] The following operational sequence is performed:

[0073] 1. User connects to geographic data acquisition mechanism 101 using mobile device 250. If the user is using a multimodal client at mobile device 250, an HTTP request is made by the client to the document server of word filtering mechanism 102. If mobile device 250 is a wireless telephone, an http request is made by the voice browser ("speaking" with the telephone) to the document server of word filtering mechanism 102. The ANI and/or the Cell ID are passed to the document server as part of the HTTP request. Location-based service 203 determines the geographic area of interest, in this case following the remote path indicated in FIG. 5 which commences at mechanism 202. Location-based service also determines geographic position of the user, an area of uncertainty with respect to the user's position, and one or more associated zip codes.

[0074] 2. A page is downloaded to mobile device 250 or loaded onto the voice browser. The geographic context is used by word filtering mechanism 102 to compute a word subset list 600 associated with the page in accordance with the previously determined geographic area of interest.

[0075] 3. The user can vocally interact with the page using mobile device 250. The page is optimized for the user's geographic area of interest.

[0076] Scenario Three: Geographic area of interest to be determined automatically using GPS data. Possible clients used by mobile device 250 for GPS include a multimodal client. Location data are provided in the form of GPS coordinates. The geographic area of interest is determined as a zip code, as (x,y) or (x,y,z) coordinates, optionally incorporating an area of uncertainty. Location resolution is performed locally, not remotely. Input mechanism 201 is not employed to determine geographic area of interest. An HTTP request is issued by mobile device 250 during the first page fetch. In this case, word filtering mechanism 102 is preparing a word subset list 600 by filtering element database 500 using GPS data. Input mechanism 201 is transparent, in the sense that the HTTP request that results in word filtering appears at the first stage of the interaction. Unlike the first scenario, there is no need for a data collecting stage. The following steps are performed:

[0077] 1. User connects to geographic data acquisition mechanism 101 using mobile device 250. If mobile device 250 utilizes a multimodal client, an HTTP request is made by the client to a document server of word filtering mechanism 102. The GPS information is sent as part of the http request. Location-based service 203 determines the geographic area of interest by adding coarse localization to the GPS information, and the coarse localization is sent to geographic area of interest defining mechanism 204.

[0078] 2. A page is downloaded to mobile device 250: The geographic area of interest in the form of the coarse and/or refined localization received at geographic area of interest defining mechanism 204 is used by word filtering mechanism 102 to compute a word subset list 600 and to place the word subset list into a page that is sent to mobile device 250, using the two-step filtering process described above.

[0079] 3. The user can vocally interact with the page, optimized for the geographic area of interest. Note:

[0080] 1. In cases where geometry specification refinement mechanism 205 is employed, an intermediate page is downloaded to mobile device 250, asking to the user to precise its area of interest (before step 2):

[0081] 1.a - An intermediate (i.e., "localization") page is downloaded to mobile device 250.

[0082] 1.b - The user inputs an area of interest into

mobile device 250.

[0083] 1.c - A refined geographic area of interest is sent by mobile device 250 and incorporated into the geographic area of interest defined by geographic area of interest defining mechanism 204.

[0084] In contrast to prior art techniques, the present invention does not employ acoustic models to improve speech recognition accuracy. Instead, the invention reduces the total number of words that need to be considered by a speech recognition algorithm. The number of words is reduced by excluding words that do not relate to a geographic area of interest. In this manner, the reduced processing burden makes it possible to handle speech recognition at the handset. Moreover, the accuracy of the speech recognition algorithm is improved.

[0085] To summarize, the advantages of the invention include more accurate speech recognition due to utilization of a database that has been reduced in size (i.e., word subset list 600). The burden on communication resources is decreased by sending the reduced database to the handset, whereupon the handset performs speech recognition using only the reduced size database instead of a much larger database. The decreased burden on communication resources results in faster, more efficient speech recognition.

[0086] Thus, while there have been shown and described novel features of the invention as applied to embodiments thereof, it will be understood that various omissions and substitutions and changes in the form and details of the devices illustrated, and in their operation, may be made by those skilled in the art without departing from the scope of the invention, which is solely defined by the appended claims. For example, although an embodiment of the invention was described in terms of a user refining an automatically determined geographic area of interest, it is also possible for a user to augment an automatically determined coarse geographic area of interest.

[0087] In view of the foregoing, it is expressly intended that all combinations of those elements and/or method steps which perform substantially the same function in substantially the same way to achieve the same results may be incorporated in any other disclosed or described or suggested form or embodiment as a general matter of design choice in as much as they fall within the scope of the claims appended hereto.

Claims

1. A method for use with a mobile handset equipped to communicate with a remote server over a wireless communications network, the remote server equipped to access an element database (500) including words that describe elements at each of a plurality of geographically defined locations, the method comprising the steps of:

   determining (202) geographic data for the mobile handset following a request; and
   using (102) the determined geographic data to automatically select by the remote server a subset (600) of words for use with a speech recognition procedure;

   wherein the subset of words is selected from the element database, thereby reducing the number of words used in the speech recognition procedure.

2. The method of claim 1 wherein the determined geographic data includes at least one of global positioning system (GPS) position data, cell identity (Cell-ID) data, caller identification (Caller ID) data, place name data, or zip code data.

3. The method of claim 1 wherein the geographic data are derived from a signal received from the mobile handset, said signal being related to a current position of said mobile handset or to a manual input entered into said mobile handset.

4. The method of any preceding claim wherein the elements include at least one of street names, businesses, merchants, points of interest, transportation facilities, individual households, activities, or landmarks.

5. The method of any preceding claim wherein the mobile handset receives the subset of words from the remote server and executes the speech recognition procedure based upon the received subset of words, thereby reducing use of the communications network and the remote server.

6. A speech recognition system comprising:

   a mobile handset equipped to communicate with a remote server over a wireless network,
   a data acquisition mechanism (101) for determining geographic data for the mobile handset following a request input; and
   a selection mechanism (102) for using the determined geographic data to automatically select a subset of words (106) for use with a speech recognition procedure;

   wherein the subset of words is selected by the remote server from an element database (500) said remote server has access to, said element database including words that describe elements at each of a plurality of geographically defined locations, thereby reducing the number of words used in the speech recognition procedure.

7. The speech recognition system of claim 6 wherein the geographic data are derived from a signal re-

ceived from the mobile handset, said signal being related to a current position of said mobile handset or to a manual input entered into said mobile handset.

8. The speech recognition system of claim 6 wherein the determined geographic data includes at least one of global positioning system (GPS) position data, cell identity (Cell-ID) data, caller identification (Caller ID) data, place name data, or zip code data.

9. The speech recognition system of any of claims 6 to 8 wherein the elements include at least one of street names, businesses, merchants, points of interest, transportation facilities, individual households, activities, or landmarks.

10. A speech recognition system for associating received speech with words retrieved from an element database, the system comprising:

means for determining a geographic area of interest following a request wherein speech is to be received, said means further comprising:

(i) a portable location determining mechanism associated with a mobile handset for generating an indication signal indicative of current geographic location, and
(ii) a server, in communication with the portable location determining mechanism, programmed to determine a geographic area of interest from the indication signal;

means for selecting a subset of words from the element database based upon the geographic area of interest; said means comprising the server programmed to extract from the element database only words that are associated with the geographic area of interest; and,
means for associating the received speech with words selected from the subset of words, said means being implemented on the mobile handset.

11. The speech recognition system of claim 10 further comprising:

means for accepting a signal from at least one graphical user interface for selecting a portion of the determined geographic area of interest to thereby specify a further limited geographic area of interest; and
means for further reducing the subset of words according to the further limited geographic area of interest to generate a further subset of words, such that the means for associating received speech with words only selects words from the further subset of words.

12. The speech recognition system of claim 11 wherein the means for accepting a signal includes:

an electronic display for displaying a map of the determined geographic area of interest; and
a processing mechanism for combining the accepted signal with the map of the determined geographic area of interest so as to cause a display of the further limited geographic area of interest on the electronic display.

13. A mobile handset (250) comprising speech recognition means for association of received speech with words received from a remote database, the mobile handset comprising:

means for receiving a request;
means for providing a signal representing a geographic area of interest following said request;
means for transmitting said signal to a remote server and for receiving from the remote server a subset of words from the remote database based upon the geographic area of interest; and
means for associating the received speech with words selected from the subset of words.

14. The mobile handset of claim 13 wherein the means for providing a signal comprises means for determining a geographic area of interest wherein speech is to be received.

15. The mobile handset of claim 13 or 14 further comprising:

means (201) for accepting at least one input used to determine a refined geographic area of interest in the geographic area of interest;
means (205) for reducing the subset of words according to the refined geographic area of interest;

and wherein:

the means for associating the received speech is arranged for association of speech only with the reduced subset of words.

Patentansprüche

1. Verfahren zur Verwendung mit einem mobilen Handapparat, der dafür ausgestattet ist, über ein drahtloses Kommunikationsnetz mit einem abgesetzten Server zu kommunizieren, wobei der abgesetzte Server dafür ausgestattet ist, auf eine Elementdatenbank (500) zuzugreifen, die Wörter enthält, die

Elemente an jedem von mehreren geographisch definierten Orten beschreiben, wobei das Verfahren die folgenden Schritte umfaßt:

Bestimmen (202) geographischer Daten für den mobilen Handapparat nach einer Anforderung und Verwenden (102) der bestimmten geographischen Daten zum automatischen Auswählen einer Teilmenge (600) von Wörtern zur Verwendung mit einer Spracherkennungsprozedur durch den abgesetzten Server;

wobei die Teilmenge von Wörtern aus der Elementdatenbank ausgewählt wird, wodurch die Anzahl der in der Spracherkennungsprozedur verwendeten Wörter reduziert wird.

2. Verfahren nach Anspruch 1, wobei die bestimmten geographischen Daten Positionsdaten des globalen Navigationssystems (GPS) und/oder Daten der Zellenidentität (Cell-ID) und/oder Daten der Anruferidentifikation (Caller ID) und/oder Städtenamendaten und/oder Postleitzahldaten umfassen.

3. Verfahren nach Anspruch 1, wobei die geographischen Daten aus einem aus dem mobilen Handapparat empfangenen Signal abgeleitet werden, wobei das Signal eine aktuelle Position des mobilen Handapparats oder eine in den mobilen Handapparat eingegebene manuelle Eingabe betrifft.

4. Verfahren nach einem der vorhergehenden Ansprüche, wobei die Elemente Straßennamen und/oder Unternehmen und/oder Händler und/oder interessierende Punkte und/oder Transporteinrichtungen und/oder individuelle Haushalte und/oder Aktivitäten und/oder Sehenswürdigkeiten umfassen.

5. Verfahren nach einem der vorhergehenden Ansprüche, wobei der mobile Handapparat die Teilmenge von Wörtern von dem abgesetzten Server empfängt und auf der Basis der empfangenen Teilmenge von Wörtern die Spracherkennungsprozedur ausführt und **dadurch** die Benutzung des Kommunikationsnetzes und des abgesetzten Servers reduziert.

6. Spracherkennungssystem, umfassend:

einen mobilen Handapparat, der dafür ausgestattet ist, über ein drahtloses Netz mit einem abgesetzten Server zu kommunizieren,
einen Datenbeschaffungsmechanismus (101) zum Bestimmen von geographischen Daten für den mobilen Handapparat nach einer Anforderungseingabe und
einen Auswahlmechanismus (102) zum Verwenden der bestimmten geographischen Daten zum automatischen Auswählen einer Teilmenge von Wörtern (106) zur Verwendung mit einer Spracherkennungsprozedur;

wobei die Teilmenge von Wörtern durch den abgesetzten Server aus einer Elementdatenbank (500) ausgewählt wird, zu der der abgesetzte Server Zugang hat, wobei die Elementdatenbank Wörter enthält, die Elemente an jedem von mehreren geographisch definierten Orten beschreiben, wodurch die Anzahl der bei der Spracherkennungsprozedur verwendeten Wörter reduziert wird.

7. Spracherkennungssystem nach Anspruch 6, wobei die geographischen Daten aus einem aus dem mobilen Handapparat empfangenen Signal abgeleitet werden, wobei das Signal eine aktuelle Position des mobilen Handapparats oder eine in den mobilen Handapparat eingegebene manuelle Eingabe betrifft.

8. Spracherkennungssystem nach Anspruch 6, wobei die bestimmten geographischen Daten Positionsdaten des globalen Navigationssystems (GPS) und/oder Daten der Zellenidentität (Cell-ID) und/oder Daten der Anruferidentifikation (Caller ID) und/oder Städtenamendaten und/oder Postleitzahldaten umfassen.

9. Spracherkennungssystem nach einem der Ansprüche 6 bis 8, wobei die Elemente Straßennamen und/oder Unternehmen und/oder Händler und/oder interessierende Punkte und/oder Transporteinrichtungen und/oder individuelle Haushalte und/oder Aktivitäten und/oder Sehenswürdigkeiten umfassen.

10. Spracherkennungssystem zum Assoziieren von empfangener Sprache mit aus einer Elementdatenbank abgerufenen Wörtern, wobei das System folgendes umfaßt:

Mittel zum Bestimmen eines interessierenden geographischen Gebiets nach einer Aufforderung, wobei Sprache zu empfangen ist, wobei die Mittel ferner folgendes umfassen:

i) einen mit einem mobilen Handapparat assoziierten tragbaren Ortsbestimmungsmechanismus zum Erzeugen eines Anzeigesignals, das einen aktuellen geographischen Ort anzeigt und
ii) einen mit dem tragbaren Ortsbestimmungsmechanismus kommunizierenden Server, der dafür programmiert ist, aus dem Anzeigesignal ein interessierendes geographisches Gebiet zu bestimmen;

Mittel zum Auswählen einer Teilmenge von Wörtern aus der Elementdatenbank auf der Ba-

sis des interessierenden geographischen Gebiets; wobei die Mittel den Server umfassen, der dafür programmiert ist, aus der Elementdatenbank nur Wörter zu extrahieren, die mit dem interessierenden geographischen Gebiet assoziiert sind und
Mittel zum Assoziieren der empfangenen Sprache mit aus der Teilmenge von Wörtern ausgewählten Wörtern, wobei die Mittel auf dem mobilen Handapparat implementiert sind.

11. Spracherkennungssystem nach Anspruch 10, ferner umfassend:

Mittel zum Annehmen eines Signals von mindestens einer graphischen Benutzeroberfläche zum Auswählen eines Teils des bestimmten interessierenden geographischen Gebiets, um **dadurch** ein weiter eingeschränktes interessierendes geographisches Gebiet zu spezifizieren und
Mittel zum weiteren Reduzieren der Teilmenge von Wörtern gemäß dem weiter eingeschränkten interessierenden geographischen Gebiet, um eine weitere Teilmenge von Wörtern zu erzeugen, dergestalt, daß die Mittel zum Assoziieren empfangener Sprache mit Wörtern nur Wörter aus der weiteren Teilmenge von Wörtern auswählen.

12. Spracherkennungssystem nach Anspruch 11, wobei die Mittel zum Annehmen eines Signals folgendes umfassen:

ein elektronisches Display zum Anzeigen einer Karte des bestimmten interessierenden geographischen Gebiets und
einen Verarbeitungsmechanismus zum Kombinieren des angenommenen Signals mit der Karte des bestimmten interessierenden geographischen Gebiets, um so eine Anzeige des weiter eingeschränkten interessierenden geographischen Gebiets auf dem elektronischen Display zu bewirken.

13. Mobiler Handapparat (150) mit Spracherkennungsmitteln zum Assoziieren empfangener Sprache mit aus einer abgesetzten Datenbank empfangenen Wörtern, wobei der mobile Handapparat folgendes umfaßt:

Mittel zum Empfangen einer Anforderung;
Mittel zum Bereitstellen eines Signals, das ein interessierendes geographisches Gebiet repräsentiert, nach der Anforderung;
Mittel zum Senden des Signals zu einem abgesetzten Server und zum Empfangen einer Teilmenge von Wörtern aus der abgesetzten Datenbank von dem abgesetzten Server auf der Basis des interessierenden geographischen Gebiets und
Mittel zum Assoziieren der empfangenen Sprache mit aus der Teilmenge von Wörtern ausgewählten Wörtern.

14. Mobiler Handapparat nach Anspruch 13, wobei die Mittel zum Bereitstellen eines Signals Mittel zum Bestimmen eines interessierenden geographischen Gebiets umfaßt, worin Sprache empfangen werden soll.

15. Mobiler Handapparat nach Anspruch 13 oder 14, ferner umfassen:

Mittel (210) zum Annehmen mindestens einer Eingabe, mit der ein verfeinertes interessierendes geographisches Gebiet in dem interessierenden geographischen Gebiet bestimmt wird;
Mittel (205) zum Reduzieren der Teilmenge von Wörtern gemäß dem verfeinerten interessierenden geographischen Gebiet und wobei

die Mittel zum Assoziieren der empfangenen Sprache dafür ausgelegt sind, Sprache nur mit der reduzierten Teilmenge von Wörtern zu assoziieren.

Revendications

1. Procédé pour une utilisation avec un combiné mobile équipé pour communiquer avec un serveur distant sur un réseau de communication sans fil, le serveur distant étant équipé pour accéder à une base de données d'éléments (500) comprenant des mots qui décrivent des éléments pour chacun d'une pluralité d'emplacements définis géographiquement, le procédé comprenant les étapes de :

détermination (202) de données géographiques pour le combiné mobile suite à une demande ; et utilisation (102) des données géographiques déterminées pour sélectionner automatiquement par le serveur distant un sous-ensemble (600) de mots pour une utilisation avec une procédure de reconnaissance vocale ;

dans lequel le sous-ensemble de mots est sélectionné parmi la base de données d'éléments, réduisant de ce fait le nombre de mots utilisés dans la procédure de reconnaissance vocale.

2. Procédé selon la revendication 1 dans lequel les données géographiques déterminées comprennent au moins une parmi des données de position du système de positionnement mondial (GPS), des données d'identité de cellule (identifiant de cellule), des

données d'identification d'appelant (identifiant d'appelant), des données de nom de lieu ou des données de code postal.

3. Procédé selon la revendication 1 dans lequel les données géographiques sont déduites d'un signal reçu du combiné mobile, ledit signal étant lié à une position actuelle dudit combiné mobile ou à une entrée manuelle entrée dans ledit combiné mobile.

4. Procédé selon une revendication précédente quelconque dans lequel les éléments comprennent au moins un parmi des noms de rues, des entreprises, des commerçants, des centres d'intérêt, des installations de transport, des maisons individuelles, des activités ou des points de repère.

5. Procédé selon une revendication précédente quelconque dans lequel le combiné mobile reçoit le sous-ensemble de mots du serveur distant et exécute la procédure de reconnaissance vocale sur la base du sous-ensemble reçu de mots, réduisant de ce fait l'utilisation du réseau de communication et du serveur distant.

6. Système de reconnaissance vocale comprenant :

un combiné mobile équipé pour communiquer avec un serveur distant sur un réseau sans fil ;
un mécanisme d'acquisition de données (101) pour déterminer des données géographiques pour le combiné mobile suite à une entrée de demande ; et
un mécanisme de sélection (102) pour utiliser les données géographiques déterminées afin de sélectionner automatiquement un sous-ensemble de mots (106) pour une utilisation avec une procédure de reconnaissance vocale ;

dans lequel le sous-ensemble de mots est sélectionné par le serveur distant à partir d'une base de données d'éléments (500) à laquelle ledit serveur distant a accès, ladite base de données d'éléments comprenant des mots qui décrivent des éléments pour chacun d'une pluralité d'emplacements définis géographiquement, réduisant de ce fait le nombre de mots utilisés dans la procédure de reconnaissance vocale.

7. Système de reconnaissance vocale selon la revendication 6 dans lequel les données géographiques sont déduites d'un signal reçu du combiné mobile, ledit signal étant lié à une position actuelle dudit combiné mobile ou à une entrée manuelle entrée dans ledit combiné mobile.

8. Système de reconnaissance vocale selon la revendication 6 dans lequel les données géographiques déterminées comprennent au moins une parmi des données de position du système de positionnement mondial (GPS), des données d'identité de cellule (identifiant de cellule), des données d'identification d'appelant (identifiant d'appelant), des données de nom de lieu ou des données de code postal.

9. Système de reconnaissance vocale selon l'une quelconque des revendications 6 à 8 dans lequel les éléments comprennent au moins un parmi des noms de rues, des entreprises, des commerçants, des centres d'intérêt, des installations de transport, des maisons individuelles, des activités ou des points de repère.

10. Système de reconnaissance vocale pour associer une voix reçue à des mots récupérés d'une base de données d'éléments, le système comprenant :

des moyens pour déterminer une zone géographique d'intérêt suite à une demande dans laquelle une voix doit être reçue, lesdits moyens comprenant en outre :

(i) un mécanisme de détermination d'emplacement de portable associé à un combiné mobile pour générer un signal d'indication révélateur d'un emplacement géographique actuel, et
(ii) un serveur, en communication avec le mécanisme de détermination d'emplacement de portable, programmé pour déterminer une zone géographique d'intérêt à partir du signal d'indication ;

des moyens pour sélectionner un sous-ensemble de mots parmi la base de données d'éléments sur la base de la zone géographique d'intérêt ; lesdits moyens comprenant le serveur programmé pour extraire de la base de données d'éléments seulement des mots qui sont associés à la zone géographique d'intérêt ; et
des moyens pour associer la voix reçue à des mots sélectionnés parmi le sous-ensemble de mots, lesdits moyens étant mis en oeuvre sur le combiné mobile.

11. Système de reconnaissance vocale selon la revendication 10 comprenant en outre :

des moyens pour accepter un signal d'au moins une interface utilisateur graphique pour sélectionner une partie de la zone géographique d'intérêt déterminée afin de spécifier de ce fait une zone géographique d'intérêt limitée supplémentaire ; et
des moyens pour réduire en outre le sous-ensemble de mots selon la zone géographique

d'intérêt limitée supplémentaire afin de générer un sous-ensemble supplémentaire de mots, de telle sorte que les moyens pour associer une voix reçue à des mots sélectionnent seulement des mots parmi le sous-ensemble supplémentaire de mots.

12. Système de reconnaissance vocale selon la revendication 11 dans lequel les moyens pour accepter un signal comprennent :

un affichage électronique pour afficher une carte de la zone géographique d'intérêt déterminée ; et
un mécanisme de traitement pour combiner le signal accepté avec la carte de la zone géographique d'intérêt déterminée de façon à provoquer un affichage de la zone géographique d'intérêt limitée supplémentaire sur l'affichage électronique.

13. Combiné mobile (250) comprenant des moyens de reconnaissance vocale pour une association d'une voix reçue à des mots reçus d'une base de données distante, le combiné mobile comprenant :

des moyens pour recevoir une demande ;
des moyens pour fournir un signal représentant une zone géographique d'intérêt suite à ladite demande ;
des moyens pour transmettre ledit signal à un serveur distant et recevoir du serveur distant un sous-ensemble de mots provenant de la base de données distante sur la base de la zone géographique d'intérêt ; et
des moyens pour associer la voix reçue à des mots sélectionnés parmi le sous-ensemble de mots.

14. Combiné mobile selon la revendication 13 dans lequel les moyens pour fournir un signal comprennent des moyens pour déterminer une zone géographique d'intérêt dans laquelle une voix doit être reçue.

15. Combiné mobile selon la revendication 13 ou 14 comprenant en outre:

des moyens (201) pour accepter au moins une entrée utilisée pour déterminer une zone géographique affinée d'intérêt dans la zone géographique d'intérêt ;
des moyens (205) pour réduire le sous-ensemble de mots selon la zone géographique affinée d'intérêt;

et dans lequel :

les moyens pour associer la voix reçue sont agencés pour une association d'une voix seulement avec le sous-ensemble réduit de mots.

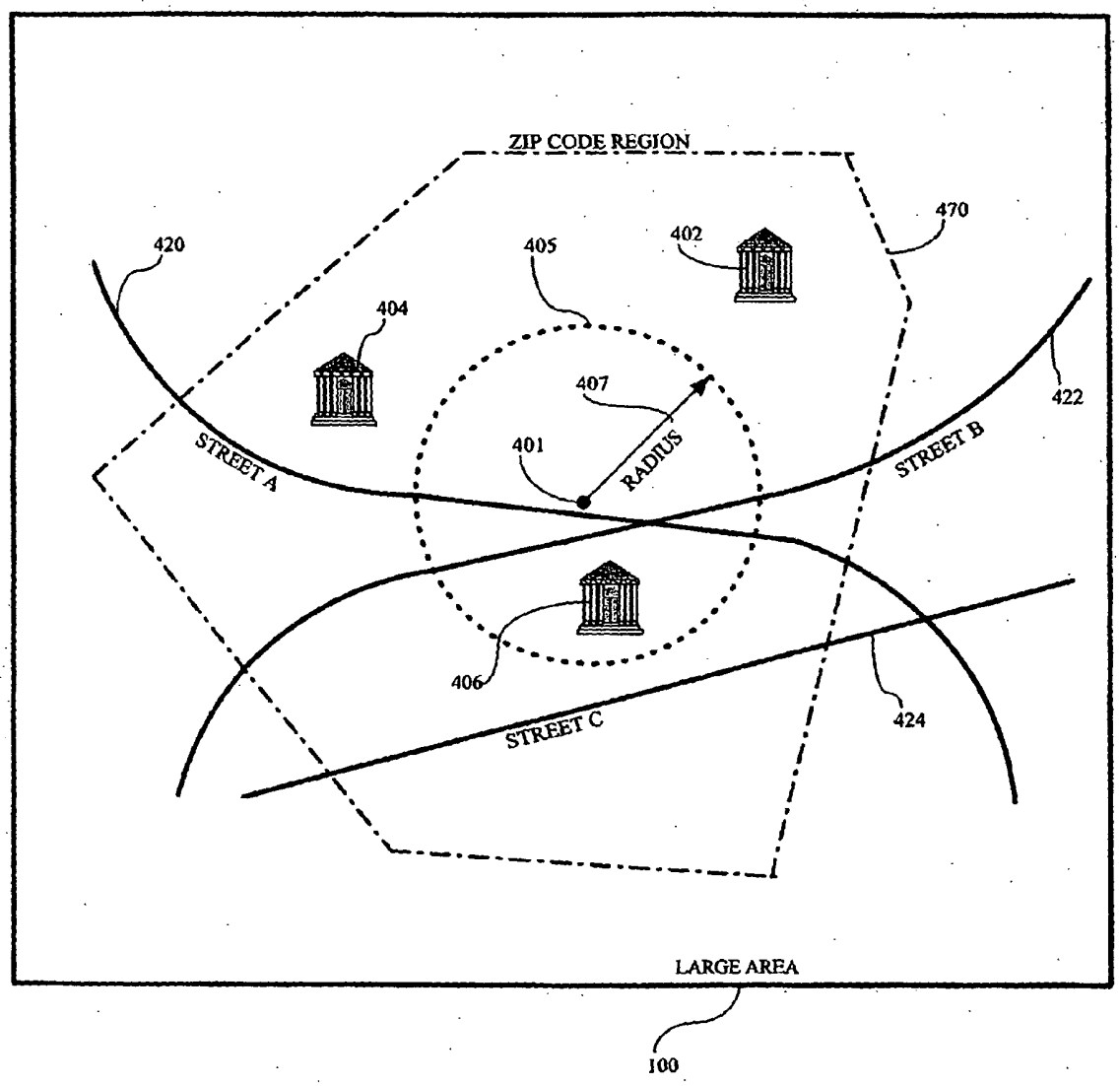

ZIP CODE REGION
470
420
405
402
404
407
422
401
RADIUS
STREET A
STREET B
406
424
STREET C
LARGE AREA
100

FIG. 1

500

ELEMENT DATABASE

ELEMENT NAME FIELD 501	PARAMETER FIELD FOR THE REFINED GEOGRAPHIC AREA OF INTEREST 503	PARAMETER FIELD FOR THE COARSE GEOGRAPHIC AREA OF INTEREST 505
FIRST BUILDING	{[603001, 2428833]}	10017
SECOND BUILDING	{[602298, 2428830]}	10017
THIRD BUILDING	{[603010, 2428836]}	10017
STREET A	{[603012, 2428836],[603013, 24428836]}	10017, 10036
STREET B	{[603014, 2428836],[603015, 24428836]}	10017, 10055
STREET C	{[679010, 2428836],[680000, 24428836]}	10017, 12533
...	...	...
DEVON FARMS ROAD	{[673010, 2428836],[675000, 24428836]}	12582

FIG. 2

600

WORD SUBSET LIST

ELEMENT NAME FIELD 601
FIRST BUILDING
STREET A
STREET B

FIG. 3

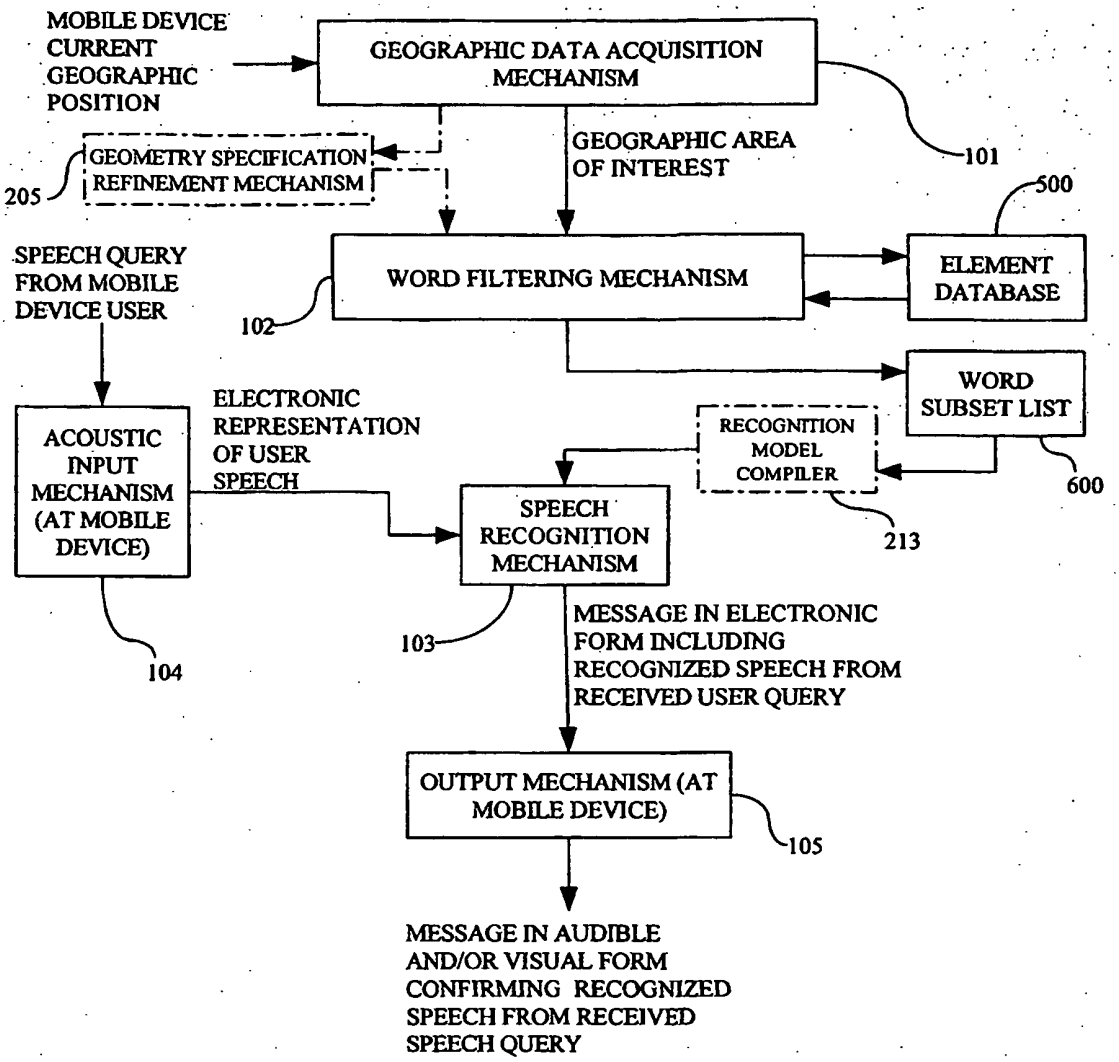

MOBILE DEVICE CURRENT GEOGRAPHIC POSITION
GEOGRAPHIC DATA ACQUISITION MECHANISM
101
GEOGRAPHIC AREA OF INTEREST
GEOMETRY SPECIFICATION REFINEMENT MECHANISM
205
500
SPEECH QUERY FROM MOBILE DEVICE USER
WORD FILTERING MECHANISM
102
ELEMENT DATABASE
WORD SUBSET LIST
600
ELECTRONIC REPRESENTATION OF USER SPEECH
ACOUSTIC INPUT MECHANISM (AT MOBILE DEVICE)
RECOGNITION MODEL COMPILER
213
SPEECH RECOGNITION MECHANISM
103
104
MESSAGE IN ELECTRONIC FORM INCLUDING RECOGNIZED SPEECH FROM RECEIVED USER QUERY
OUTPUT MECHANISM (AT MOBILE DEVICE)
105
MESSAGE IN AUDIBLE AND/OR VISUAL FORM CONFIRMING RECOGNIZED SPEECH FROM RECEIVED SPEECH QUERY

FIG. 4

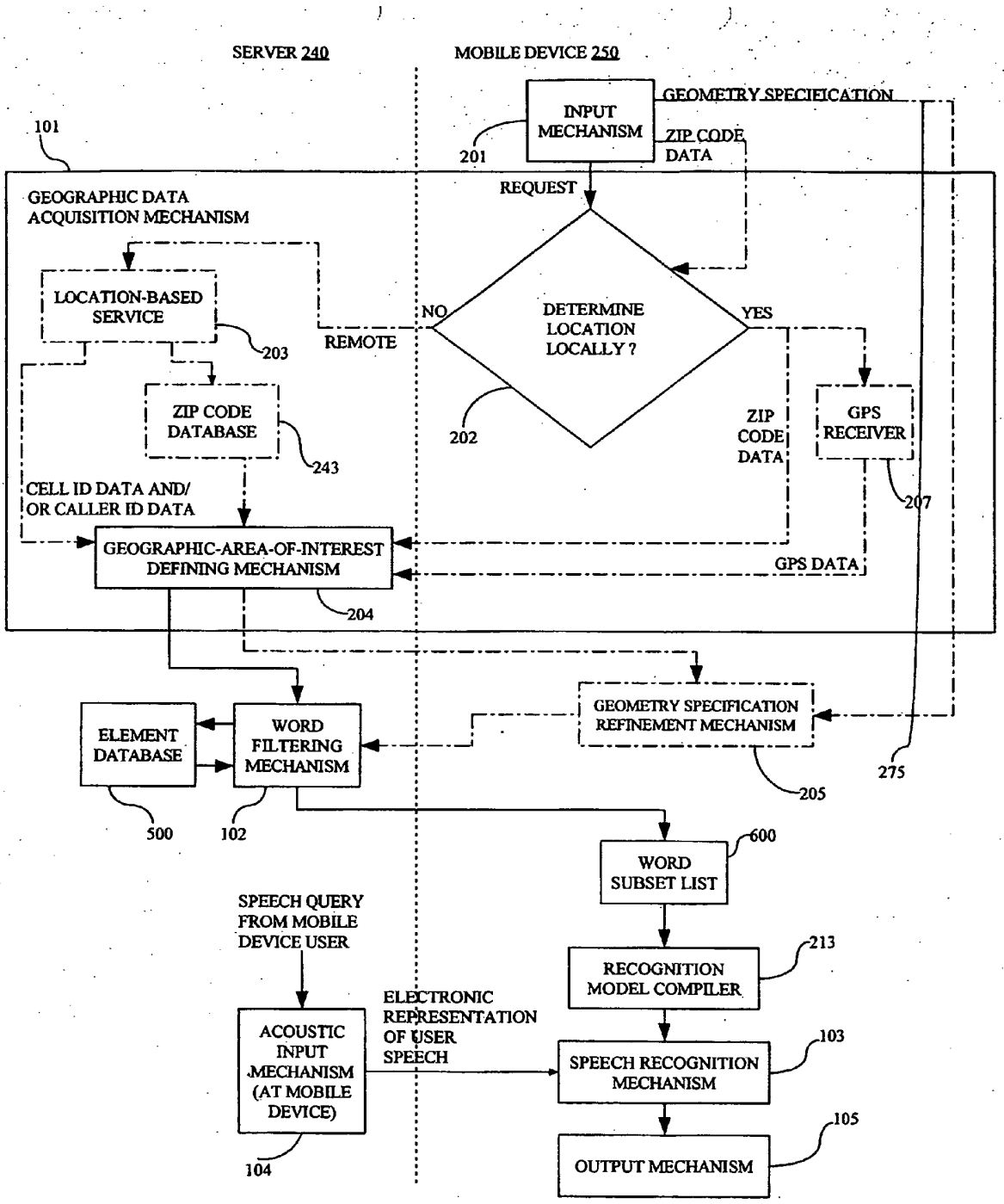

SERVER 240
MOBILE DEVICE 250
INPUT MECHANISM
201
GEOMETRY SPECIFICATION
ZIP CODE DATA
101
REQUEST
GEOGRAPHIC DATA ACQUISITION MECHANISM
LOCATION-BASED SERVICE
203
REMOTE
NO
DETERMINE LOCATION LOCALLY?
YES
202
ZIP CODE DATABASE
243
ZIP CODE DATA
GPS RECEIVER
207
CELL ID DATA AND/ OR CALLER ID DATA
GEOGRAPHIC-AREA-OF-INTEREST DEFINING MECHANISM
GPS DATA
204
GEOMETRY SPECIFICATION REFINEMENT MECHANISM
275
205
ELEMENT DATABASE
WORD FILTERING MECHANISM
500
102
600
WORD SUBSET LIST
SPEECH QUERY FROM MOBILE DEVICE USER
213
RECOGNITION MODEL COMPILER
ELECTRONIC REPRESENTATION OF USER SPEECH
ACOUSTIC INPUT MECHANISM (AT MOBILE DEVICE)
103
SPEECH RECOGNITION MECHANISM
105
OUTPUT MECHANISM
104

FIG. 5

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- US 2003236099 A **[0007] [0009] [0010]**
- EP 1233407 A **[0011]**
- US 6230132 B **[0012]**
- EP 0661688 A **[0013]**

Electronic Patent Application Fee Transmittal

Application Number:	11580926
Filing Date:	16-Oct-2006
Title of Invention:	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE
First Named Inventor/Applicant Name:	Larry Baldwin
Filer:	Syed Jafar Ali/Susan Trader
Attorney Docket Number:	090611-0357945

Filed as Small Entity

Utility under 35 USC 111(a) Filing Fees

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Miscellaneous:				
Submission- Information Disclosure Stmt	1806	1	180	180
		Total in USD ($)		**180**

Electronic Acknowledgement Receipt

EFS ID:	10914913
Application Number:	11580926
International Application Number:	
Confirmation Number:	6670
Title of Invention:	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE
First Named Inventor/Applicant Name:	Larry Baldwin
Customer Number:	00909
Filer:	Syed Jafar Ali/Susan Trader
Filer Authorized By:	Syed Jafar Ali
Attorney Docket Number:	090611-0357945
Receipt Date:	09-SEP-2011
Filing Date:	16-OCT-2006
Time Stamp:	15:02:15
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	Deposit Account
Payment was successfully received in RAM	$180
RAM confirmation Number	1545
Deposit Account	033975
Authorized User	

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)

1		0357945_IDS.pdf	979411 bda38d28279c3a12c6334798b7237eed8493561f	yes	27

	Multipart Description/PDF files in .zip description		
	Document Description	**Start**	**End**
	Transmittal Letter	1	3
	Information Disclosure Statement (IDS) Form (SB08)	4	4
	Transmittal Letter	5	25
	Information Disclosure Statement (IDS) Form (SB08)	26	27

Warnings:					
Information:					
2	Foreign Reference	EP1320043A2.pdf	1416212 2ec7d64c9ab7063100d70c545035fd72a13ded77	no	28
Warnings:					
Information:					
3	Foreign Reference	EP1646037.pdf	262902 56537431a1754beec4b827cf69e772a2da14c1a8	no	20
Warnings:					
Information:					
4	Non Patent Literature	Statement_in_Accordance_with_EPO_Notice.pdf	86084 a41f2c134966e666f56e1c9c6e421bc9a3281ef0	no	1
Warnings:					
Information:					
5	Fee Worksheet (SB06)	fee-info.pdf	30363 cf0dd86e23a44a8e58242ae7d639d455f96c61c5	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	2774972		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANTS :	Larry BALDWIN *et al.*	CONFIRMATION NO.:	6670
SERIAL NUMBER :	11/580,926	EXAMINER:	Eric L. Yen
FILING DATE :	October 16, 2006	ART UNIT:	2626
FOR :	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE		

SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT

Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313-1450

Sir:

Pursuant to the duty of disclosure under 37 C.F.R. §§ 1.56, 1.97 and 1.98, Applicants hereby make of record in the above-identified application the documents listed on the attached Form PTO-1449.

Applicants reserve the right to establish the patentability of the claimed invention over any of the information provided herewith, and/or to prove that this information may not be prior art, and/or to prove that this information may not be enabling for the teachings purportedly offered.

This Supplemental Information Disclosure Statement is being filed more than three months after the filing date of a national application other than a continued prosecution application under 37 C.F.R. § 1.53(d), or more than three months after the date of entry of the national stage as set forth in 37 C.F.R. § 1.491 in an international application, and after the

mailing date of any of a Final Action under 37 C.F.R. § 1.113, a Notice of Allowance under 37 C.F.R. § 1.311, or an action that otherwise closes prosecution in the application, but on or before payment of the Issue Fee. 37 C.F.R. § 1.97(d).

Documents B, C, D, H, I, U, and V were cited by the U.S. Patent Office in co-pending Application Serial No. 12/765,753, filed April 22, 2010, in an Office Action dated June 9, 2011. Documents G, L, N, O, and Q were cited by the U.S. Patent Office in co-pending Application Serial No. 11/954,064, filed December 11, 2007, in an Office Action dated June 16, 2011. Documents A, F, and T were cited by the U.S. Patent Office in co-pending Application Serial No. 12/365,516, filed February 4, 2009, in an Office Action dated June 23, 2011. Document K was cited by the U.S. Patent Office in co-pending Application Serial No. 12/127,343, filed May 27, 2008, in an Office Action dated July 20, 2011. Documents E, J, M, and P were cited by the U.S. Patent Office in co-pending Application Serial No. 12/838,982, filed July 19, 2010, in an Office Action dated August 10, 2011. I hereby certify that no item of information contained in this Supplemental Information Disclosure Statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to my knowledge after making reasonable inquiry, no item of information contained in this Supplemental Information Disclosure Statement was known to any individual designated in 37 C.F.R. § 1.56(c) more than three months prior to the filing of this Supplemental Information Disclosure Statement. 37 C.F.R. § 1.97(e)(2).

Please charge the $180.00 fee [37 C.F.R. § 1.17(p)] to Deposit Account No. 03-3975. In addition, please charge any fee deficiencies, or credit any overpayment, to Deposit Account No.

03-3975 as needed to ensure consideration of the Supplemental Information Disclosure Statement.

It is respectfully requested that the Examiner initial and return a copy of the enclosed PTO-1449.

Respectfully submitted,

PILLSBURY WINTHROP SHAW PITTMAN LLP

By_____
Syed Jafar Ali
Registration No.: 58,780

Customer No.
00909

P. O. Box 10500
McLean, Virginia 22102
Telephone: (703) 770-7900
Telefax: (703) 770-7901

Dated: September 9, 2011

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
11/580,926	10/16/2006	Larry Baldwin	090611-0357945	6670

909 7590 09/22/2011

PILLSBURY WINTHROP SHAW PITTMAN, LLP (NV)
P.O. BOX 10500
MCLEAN, VA 22102

EXAMINER	
YEN, ERIC L	

ART UNIT	PAPER NUMBER
2626	

NOTIFICATION DATE	DELIVERY MODE
09/22/2011	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

docket_ip@pillsburylaw.com
margaret.drosos@pillsburylaw.com

UNITED STATES DEPARTMENT OF COMMERCE
U.S. Patent and Trademark Office
Address : COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450

APPLICATION NO./ CONTROL NO.	FILING DATE	FIRST NAMED INVENTOR / PATENT IN REEXAMINATION	ATTORNEY DOCKET NO.
11/580,926	16 October, 2006	BALDWIN ET AL.	090611-0357945

PILLSBURY WINTHROP SHAW PITTMAN, LLP (NV)
P.O. BOX 10500
MCLEAN, VA 22102

EXAMINER	
ERIC YEN	
ART UNIT	PAPER
2626	20110913

DATE MAILED:

Please find below and/or attached an Office communication concerning this application or proceeding.

Commissioner for Patents

IDS has been considered

/Eric Yen/
Primary Examiner, Art Unit 2626

PTO-90C (Rev.04-03)

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /E.Y./

Cited in the Information Disclosure Statement Under 37 C.F.R. § 1.97(i)

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	1A	5,027,406	06/25/1991	Roberts *et al.*	381	43	
	1B	5,960,447	09/28/1999	Holt *et al.*	707	500	
	1C	5,995,928	11/30/1999	Nguyen *et al.*	704	251	
	1D	6,018,708	01/25/2000	Dahan *et al.*	704	244	
	1E	6,108,631	08/22/2000	Ruhl	704	270	
	1F	6,167,377	12/26/2000	Gillick *et al.*	704	240	
	1G	6,173,266	01/09/2001	Marx *et al.*	704	270	
	1H	6,208,964	03/27/2001	Sabourin	704	244	
	1I	6,269,336	07/31/2001	Ladd *et al.*	704	270	
	1J	6,366,882	04/02/2002	Bijl *et al.*	704	235	
	1K	6,377,913	04/23/2002	Coffman *et al.*	704	8	
	1L	6,405,170	06/11/2002	Phillips *et al.*	704	270	
	1M	6,606,598	08/12/2003	Holthouse *et al.*	704	275	
	1N	6,681,206	01/20/2004	Gorin *et al.*	704	243	
	1O	6,701,294	03/02/2004	Ball *et al.*	704	257	
	1P	6,704,708	03/09/2004	Pickering	704	235	
	1Q	6,751,612	06/15/2004	Schuetze *et al.*	707	4	
	1R	6,832,230	12/14/2004	Zilliacus *et al.*	707	203	
	1S	6,912,498	06/28/2005	Stevens *et al.*	704	235	
	1T	6,954,755	10/11/2005	Reisman	707	10	
	1U	7,003,463	02/21/2006	Maes *et al.*	704	270.1	
	1V	7,058,890	06/06/2006	George *et al.*	715	728	
	1W	7,197,069	03/27/2007	Agazzi *et al.*	375	233	
	1X	7,277,854	10/02/2007	Bennett *et al.*	704	257	
	1Y	7,366,669	04/29/2008	Nishitani *et al.*	704	256	
	1Z	7,472,060	12/30/2008	Gorin *et al.*	704	240	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO
	1AA	1 320 043 A2	06/18/2003	Europe				
	1BB	1 646 037	04/12/2006	Europe				

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

	1CC	"Statement in Accordance with the Notice from the European Patent Office dated 1 October 2007 Concerning Business Methods (OJ EPO 11/2007, 592-593), XP002456252.

EXAMINER	DATE CONSIDERED
/E. Yen/	09/13/2011

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

403110786v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	2A	7,558,730	07/07/2009	Davis *et al.*	704	235	
	2B	7,676,365	03/09/2010	Hwang *et al.*	704	240	
	2C	7,873,523	01/18/2011	Potter *et al.*	704	275	
	2D	7,917,367	03/29/2011	Di Cristo *et al.*	704	270.1	
	2E	7,949,529	05/24/2011	Weider *et al.*	704	270	
	2F	2002/0035501	03/21/2002	Handel *et al.*	705	10	
	2G	2003/0120493	06/26/2003	Gupta	704	270.1	
	2H	2009/0271194	10/29/2009	Davis *et al.*	704	235	
	2I	2011/0112827	05/12/2011	Kennewick *et al.*	704	9	
	2J	2011/0112921	05/12/2011	Kennewick *et al.*	705	26.1	
	2K	2011/0131036	06/02/2011	Di Cristo *et al.*	704	9	
	2L	2011/0131045	06/02/2011	Cristo *et al.*	704	249	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	/E. Yen/	DATE CONSIDERED	09/13/2011

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

403110786v1

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /E.Y./

INFORMATION DISCLOSURE CITATION IN AN APPLICATION

Form PTO-1449

Docket Number	Application Number
090611-0357945	11/580,926
Applicant(s)	
Larry BALDWIN *et al.*	
Filing Date	Group Art Unit
October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	A	5,839,107	11/17/1998	Gupta *et al.*	704	270	
	B	5,897,613	04/27/1999	Chan	704	210	
	C	5,960,399	09/28/1999	Barclay *et al.*	704	270	
	D	6,049,607	04/11/2000	Marash *et al.*	379	410	
	E	6,085,186	07/04/2000	Christianson *et al.*	707	3	
	F	6,397,181	05/28/2002	Li *et al.*	704	256	
	G	6,429,813	08/06/2002	Feigen	342	357.13	
	H	6,522,746	02/18/2003	Marchok *et al.*	379	406.03	
	I	6,549,629	04/15/2003	Finn *et al.*	381	92	
	J	6,587,858	07/01/2003	Strazza	707	102	
	K	6,975,993	12/13/2005	Keiller	704	275	
	L	7,024,364	04/04/2006	Guerra *et al.*	704	270	
	M	7,236,923	06/26/2007	Gupta	704	9	
	N	7,472,020	12/30/2008	Brulle-Drews	701	211	
	O	7,493,259	02/17/2009	Jones *et al.*	704	257	
	P	7,536,297	05/19/2009	Byrd *et al.*	704	10	
	Q	7,831,433	11/09/2010	Belvin *et al.*	704	275	
	R	7,983,917	07/19/2011	Kennewick *et al.*	704	257	
	S	8,015,006	09/06/2011	Kennewick *et al.*	704	236	
	T	2003/0144846	07/31/2003	Denenberg *et al.*	704	277	
	U	2003/0212550	11/13/2003	Ubale	704	215	
	V	2007/0033020	02/08/2007	(Kelleher) Francois *et al.*	704	226	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED
/E. Yen/	09/13/2011

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

403110745v1

PART B - FEE(S) TRANSMITTAL

Complete and send this form, together with applicable fee(s), to: Mail **Mail Stop ISSUE FEE**
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450
or Fax (571)-273-2885

INSTRUCTIONS: This form should be used for transmitting the ISSUE FEE and PUBLICATION FEE (if required). Blocks 1 through 5 should be completed where appropriate. All further correspondence including the Patent, advance orders and notification of maintenance fees will be mailed to the current correspondence address as indicated unless corrected below or directed otherwise in Block 1, by (a) specifying a new correspondence address; and/or (b) indicating a separate "FEE ADDRESS" for maintenance fee notifications.

CURRENT CORRESPONDENCE ADDRESS (Note: Use Block 1 for any change of address)

909 7590 08/15/2011
PILLSBURY WINTHROP SHAW PITTMAN, LLP
P.O. BOX 10500
MCLEAN, VA 22102

Note: A certificate of mailing can only be used for domestic mailings of the Fee(s) Transmittal. This certificate cannot be used for any other accompanying papers. Each additional paper, such as an assignment or formal drawing, must have its own certificate of mailing or transmission.

Certificate of Mailing or Transmission
I hereby certify that this Fee(s) Transmittal is being deposited with the United States Postal Service with sufficient postage for first class mail in an envelope addressed to the Mail Stop ISSUE FEE address above, or being facsimile transmitted to the USPTO (571) 273-2885, on the date indicated below.

(Depositor's name)
(Signature)
(Date)

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
11/580,926	10/16/2006	Larry Baldwin	090611-0357945	6670

TITLE OF INVENTION: SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

APPLN. TYPE	SMALL ENTITY	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	YES	$755	$300	$0	$1055	11/15/2011

EXAMINER	ART UNIT	CLASS-SUBCLASS
YEN, ERIC L	2626	704-009000

1. Change of correspondence address or indication of "Fee Address" (37 CFR 1.363).
☐ Change of correspondence address (or Change of Correspondence Address form PTO/SB/122) attached.
☐ "Fee Address" indication (or "Fee Address" Indication form PTO/SB/47; Rev 03-02 or more recent) attached. **Use of a Customer Number is required.**

2. For printing on the patent front page, list
(1) the names of up to 3 registered patent attorneys or agents OR, alternatively,
(2) the name of a single firm (having as a member a registered attorney or agent) and the names of up to 2 registered patent attorneys or agents. If no name is listed, no name will be printed.
1 Pillsbury Winthrop Shaw Pittman LLP
2
3

3. ASSIGNEE NAME AND RESIDENCE DATA TO BE PRINTED ON THE PATENT (print or type)

PLEASE NOTE: Unless an assignee is identified below, no assignee data will appear on the patent. If an assignee is identified below, the document has been filed for recordation as set forth in 37 CFR 3.11. Completion of this form is NOT a substitute for filing an assignment.

(A) NAME OF ASSIGNEE: VoiceBox Technologies, Inc.
(B) RESIDENCE: (CITY and STATE OR COUNTRY): Bellevue, Washington

Please check the appropriate assignee category or categories (will not be printed on the patent): ☐ Individual ☑ Corporation or other private group entity ☐ Government

4a. The following fee(s) are submitted:
☑ Issue Fee
☑ Publication Fee (No small entity discount permitted)
☐ Advance Order - # of Copies _____

4b. Payment of Fee(s): **(Please first reapply any previously paid issue fee shown above)**
☐ A check is enclosed.
☐ Payment by credit card. Form PTO-2038 is attached.
☑ The Director is hereby authorized to charge the required fee(s), any deficiency, or credit any overpayment, to Deposit Account Number 03-3975 (enclose an extra copy of this form).

5. **Change in Entity Status** (from status indicated above)
☐ a. Applicant claims SMALL ENTITY status. See 37 CFR 1.27.
☐ b. Applicant is no longer claiming SMALL ENTITY status. See 37 CFR 1.27(g)(2).

NOTE: The Issue Fee and Publication Fee (if required) will not be accepted from anyone other than the applicant; a registered attorney or agent; or the assignee or other party in interest as shown by the records of the United States Patent and Trademark Office.

Authorized Signature _____ Date October 4, 2011
Typed or printed name Syed Jafar Ali Registration No. 58,780

This collection of information is required by 37 CFR 1.311. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, Virginia 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, Virginia 22313-1450.

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

Electronic Patent Application Fee Transmittal

Application Number:	11580926
Filing Date:	16-Oct-2006
Title of Invention:	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE
First Named Inventor/Applicant Name:	Larry Baldwin
Filer:	Syed Jafar Ali/Diane Bogusz
Attorney Docket Number:	090611-0357945

Filed as Small Entity

Utility under 35 USC 111(a) Filing Fees

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Utility Appl issue fee	2501	1	870	870
Publ. Fee- early, voluntary, or normal	1504	1	300	300

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Extension-of-Time:				
Miscellaneous:				
			Total in USD ($)	**1170**

Electronic Acknowledgement Receipt

EFS ID:	11106955
Application Number:	11580926
International Application Number:	
Confirmation Number:	6670
Title of Invention:	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE
First Named Inventor/Applicant Name:	Larry Baldwin
Customer Number:	909
Filer:	Syed Jafar Ali/Diane Bogusz
Filer Authorized By:	Syed Jafar Ali
Attorney Docket Number:	090611-0357945
Receipt Date:	04-OCT-2011
Filing Date:	16-OCT-2006
Time Stamp:	13:43:50
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	Deposit Account
Payment was successfully received in RAM	$1170
RAM confirmation Number	10
Deposit Account	033975
Authorized User	
The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows: Charge any Additional Fees required under 37 C.F.R. Section 1.20 (Post Issuance fees)	

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Issue Fee Payment (PTO-85B)	0357945.pdf	90721 4a19005736a7753111a2390a16405a0c23a6a8ef	no	1
Warnings:					
Information:					
2	Fee Worksheet (SB06)	fee-info.pdf	32135 b1efec2ba0be71434ceaccf78506d439a62bd455	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	122856		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111
If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371
If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office
If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
/E.Y./	1A	5,794,196	08/11/1998	Yegnanarayanan *et al.*	704	255	
	1B	5,867,817	02/02/1999	Catallo *et al.*	704	255	
	1C	6,470,315	10/22/2002	Netsch *et al.*	704	256	
	1D	6,604,075	08/05/2003	Brown *et al.*	704	270.1	
	1E	6,622,119	09/16/2003	Ramaswamy *et al.*	704	9	
	1F	6,658,388	12/02/2003	Kleindienst *et al.*	704	275	
	1G	6,708,150	03/16/2004	Hirayama *et al.*	704	243	
	1H	7,231,343	06/12/2007	Treadgold *et al.*	704	9	
	1I	7,386,443	06/10/2008	Parthasarathy *et al.*	704	201	
	1J	7,415,414	08/19/2008	Azara *et al.*	704	270	
	1K	7,809,570	10/05/2010	Kennewick *et al.*	704	257	
	1L	7,818,176	10/19/2010	Freeman *et al.*	704	270	
	1M	2002/0087525	07/04/2002	Abbott *et al.*	707	3	
	1N	2003/0046346	03/06/2003	Mumick *et al.*	709	205	
	1O	2003/0064709	04/03/2003	Gailey *et al.*	455	412	
	1P	2003/0236664	12/25/2003	Sharma	704	251	
	1Q	2004/0006475	01/08/2004	Ehlen *et al.*	704	270.1	
	1R	2004/0193408	09/30/2004	Hunt	704	209	
	1S	2004/0243417	12/02/2004	Pitts, III *et al.*	704	276	
	1T	2005/0043940	02/24/2005	Elder	704	9	
	1U	2005/0137850	06/23/2005	Odell	704	4	
	1V	2007/0179778	08/02/2007	Gong *et al.*	704	9	
/E.Y./	1W	2008/0140385	06/12/2008	Mahajan *et al.*	704	9	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO
/E.Y./	1X	WO 99/46763	09/16/1999	International				
	1Y	WO 00/21232	04/13/2000	International				
	1Z	WO 00/46792	08/10/2000	International				
	1AA	WO 2004/072954	08/26/~~2004~~	International 2004				
	1BB	WO 2007/019318	02/15/2007	International				
	1CC	WO 2007/021587	02/22/2007	International				
/E.Y./	1DD	WO 2007/027546	03/08/2007	International				

Change(s) applied to document, /G.H./ 9/15/2011

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

/E.Y./	1EE	BAZZI, Issam *et al.*, "Heterogeneous Lexical Units for Automatic Speech Recognition: Preliminary Investigations", Processing of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 3, June 5-9, 2000, XP010507574, pages 1257-1260.

EXAMINER	DATE CONSIDERED
/E. Yen/	08/09/2011

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

402593690v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0357945	11/580,926
	Applicant(s)	
	Larry BALDWIN *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	October 16, 2006	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
EY	5A	6,567,778	05/20/2003	Chao Chang et al.	704	257	
	5B	6,570,964	05/27/2003	Murveit et al.	379	67.1	
	5C	6,574,624	06/03/2003	Johnson et al.	707	5	
	5D	6,594,257	07/15/2003	Doshi et al.	370	352	
	5E	6,604,077	08/05/2003	Dragosh et al.	704	270.1	
	5F	6,611,692	08/26/2003	Raffel et al.	455	552	
	5G	6,614,773	09/02/2003	Maxemchuk	370	337	
	5H	6,615,172	09/02/2003	Bennett *et al.*	704	257	
	5I	6,629,066	09/30/2003	Jackson et al.	704	9	
	5J	6,650,747	11/18/2003	Bala et al.	379	265.06	
	5K	6,691,151	02/10/2004	Cheyer et al.	709	202	
	5L	6,741,931	05/25/2004	Kohut *et al.*	701	209	
	5M	6,937,977	08/30/2005	Gerson	704	201	
	5N	6,980,092	12/27/2005	Turnbull *et al.*	340	425.5	
	5O	6,996,531	02/07/2006	Korall *et al.*	704	270	
	5P	7,020,609	03/28/~~2003~~ 2006	Thrift *et al.*	704	270.1	
	5Q	7,027,975	04/11/2006	Pazandak *et al.*	704	9	
	5R	7,062,488	06/13/2006	Reisman	707	8	
	5S	7,092,928	08/15/2006	Elad *et al.*	706	60	
	5T	7,137,126	11/14/06	Coffman *et al.*	719	328	
	5U	2002/0143535	10/03/2002	Kist *et al.*	704	251	
	5V	2003/0182132	09/25/2003	Niemoeller	704	275	
	5W	2004/0044516	03/04/2004	Kennewick *et al.*	704	5	
	5X	2004/0193420	09/30/2004	Kennewick *et al.*	704	257	
EY	5Y	2005/0021826	01/27/2005	Kumar	709	232	

Change(s) applied to document,
/G.H./
9/15/2011

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO

OTHER DOCUMENTS *(Including Author, Title, Date, Pertinent Pages, Etc.)*

EXAMINER	DATE CONSIDERED
/Eric Yen/	10/26/10

EXAMINER: Initial if citation considered, whether or not citation is in conformance with MPEP Section 609; Draw line through citation if not in conformance and not considered. Include copy of this form with next communication to Applicant.

400646104v1

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	ISSUE DATE	PATENT NO.	ATTORNEY DOCKET NO.	CONFIRMATION NO.
11/580,926	12/06/2011	8073681	090611-0357945	6670

909 7590 11/16/2011

Pillsbury Winthrop Shaw Pittman, LLP (NV)
PO Box 10500
McLean, VA 22102

ISSUE NOTIFICATION

The projected patent number and issue date are specified above.

Determination of Patent Term Adjustment under 35 U.S.C. 154 (b)
(application filed on or after May 29, 2000)

The Patent Term Adjustment is 1217 day(s). Any patent to issue from the above-identified application will include an indication of the adjustment on the front page.

If a Continued Prosecution Application (CPA) was filed in the above-identified application, the filing date that determines Patent Term Adjustment is the filing date of the most recent CPA.

Applicant will be able to obtain more detailed information by accessing the Patent Application Information Retrieval (PAIR) WEB site (http://pair.uspto.gov).

Any questions regarding the Patent Term Extension or Adjustment determination should be directed to the Office of Patent Legal Administration at (571)-272-7702. Questions relating to issue and publication fee payments should be directed to the Application Assistance Unit (AAU) of the Office of Data Management (ODM) at (571)-272-4200.

APPLICANT(s) (Please see PAIR WEB site http://pair.uspto.gov for additional applicants):

Larry Baldwin, Maple Valley, WA;
Tom Freeman, Mercer Island, WA;
Michael Tjalve, Bellevue, WA;
Blane Ebersold, Seattle, WA;
Chris Weider, Everett, WA;

IR103 (Rev. 10/09)

PTO/SB/80 (11-08)
Approved for use through 11/30/2011. OMB 0651-0035
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

POWER OF ATTORNEY TO PROSECUTE APPLICATIONS BEFORE THE USPTO

I hereby revoke all previous powers of attorney given in the application identified in the attached statement under 37 CFR 3.73(b).

I hereby appoint:

☑ Practitioners associated with the Customer Number: 133759

OR

☐ Practitioner(s) named below (if more than ten patent practitioners are to be named, then a customer number must be used):

Name	Registration Number	Name	Registration Number

as attorney(s) or agent(s) to represent the undersigned before the United States Patent and Trademark Office (USPTO) in connection with any and all patent applications assigned only to the undersigned according to the USPTO assignment records or assignment documents attached to this form in accordance with 37 CFR 3.73(b).

Please change the correspondence address for the application identified in the attached statement under 37 CFR 3.73(b) to:

☑ The address associated with Customer Number: 133759

OR

☐ Firm or Individual Name			
Address			
City		State	Zip
Country			
Telephone		Email	

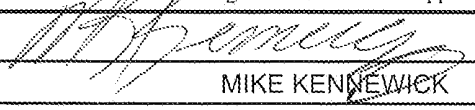
Assignee Name and Address:

VOICEBOX TECHNOLOGIES CORPORATION
11980 NE 24th Street, Suite 100
Bellevue, WA 98005

A copy of this form, together with a statement under 37 CFR 3.73(b) (Form PTO/SB/96 or equivalent) is required to be filed in each application in which this form is used. The statement under 37 CFR 3.73(b) may be completed by one of the practitioners appointed in this form if the appointed practitioner is authorized to act on behalf of the assignee, and must identify the application in which this Power of Attorney is to be filed.

SIGNATURE of Assignee of Record

The individual whose signature and title is supplied below is authorized to act on behalf of the assignee

Signature		Date
Name	MIKE KENNEWICK	Telephone
Title	CEO OF VOICEBOX TECHNOLOGIES CORPORATION	

This collection of information is required by 37 CFR 1.31, 1.32 and 1.33. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Electronic Acknowledgement Receipt

EFS ID:	24024205
Application Number:	11580926
International Application Number:	
Confirmation Number:	6670
Title of Invention:	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE
First Named Inventor/Applicant Name:	Larry Baldwin
Customer Number:	909
Filer:	James G. Gatto/Todd Yarborough
Filer Authorized By:	James G. Gatto
Attorney Docket Number:	090611-0357945
Receipt Date:	09-NOV-2015
Filing Date:	16-OCT-2006
Time Stamp:	14:30:35
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		VoiceBox_373b_POA_45HV_218162.pdf	736053 1e276f23ff7a4010eb0695d9304fcb59ec7e9f56	yes	4

	Multipart Description/PDF files in .zip description		
	Document Description	**Start**	**End**
	Assignee showing of ownership per 37 CFR 3.73	1	3
	Power of Attorney	4	4
Warnings:			
Information:			
	Total Files Size (in bytes):	736053	

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111
If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371
If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office
If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

PTO/AIA/96 (08-12)
Approved for use through 01/31/2013. OMB 0651-0031
U.S. Patent and Trademark Office;U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of1995, no persons arerequired to respond to a collection of information unless it displays a valid OMB control number.

STATEMENT UNDER 37 CFR 3.73(c)

Applicant/Patent Owner: Larry BALDWIN et al.

Application No./Patent No.: 11/580,926 Filed/Issue Date: October 16, 2006

Titled: SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

Voicebox Technologies Corporation, a corporation

(Name of Assignee) (Type of Assignee, e.g., corporation, partnership, university, government agency, etc.)

states that, for the patent application/patent identified above, it is (choose **one** of options 1, 2, 3 or 4 below):

1. ☑ The assignee of the entire right, title, and interest.

2. ☐ An assignee of less than the entire right, title, and interest (check applicable box):

   ☐ The extent (by percentage) of its ownership interest is _____%. Additional Statement(s) by the owners holding the balance of the interest must be submitted to account for 100% of the ownership interest.

   ☐ There are unspecified percentages of ownership. The other parties, including inventors, who together own the entire right, title and interest are:

   Additional Statement(s) by the owner(s) holding the balance of the interest must be submitted to account for the entire right, title, and interest.

3. ☐ The assignee of an undivided interest in the entirety (a complete assignment from one of the joint inventors was made). The other parties, including inventors, who together own the entire right, title, and interest are:

   Additional Statement(s) by the owner(s) holding the balance of the interest must be submitted to account for the entire right, title, and interest.

4. ☐ The recipient, via a court proceeding or the like (*e.g.*, bankruptcy, probate), of an undivided interest in the entirety (a complete transfer of ownership interest was made). The certified document(s) showing the transfer is attached.

The interest identified in option 1, 2 or 3 above (not option 4) is evidenced by either (choose **one** of options A or B below):

A. ☐ An assignment from the inventor(s) of the patent application/patent identified above. The assignment was recorded in the United States Patent and Trademark Office at Reel _____, Frame _____, or for which a copy thereof is attached.

B. ☑ A chain of title from the inventor(s), of the patent application/patent identified above, to the current assignee as follows:

1. From: Larry BALDWIN et al. To: Voicebox Technologies, Inc.

   The document was recorded in the United States Patent and Trademark Office at

   Reel 018774, Frame 0247, or for which a copy thereof is attached.

2. From: Voicebox Technologies, Inc. To: Voicebox Technologies Corporation

   The document was recorded in the United States Patent and Trademark Office at

   Reel 032620, Frame 0956, or for which a copy thereof is attached.

This collection of information is required by37 CFR3.73(b). The information is required toobtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentialityis governed by35 U.S.C. 122and 37 CFR1.11 and1.14. Thiscollection is estimated to take 12 minutes to complete, including gathering, preparing, and submittingthe completed application form to the USPTO.Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent tothe Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS.**SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

STATEMENT UNDER 37 CFR 3.73(c)

3. From: _____ To: _____

The document was recorded in the United States Patent and Trademark Office at

Reel _____, Frame _____, or for which a copy thereof is attached.

4. From: _____ To: _____

The document was recorded in the United States Patent and Trademark Office at

Reel _____, Frame _____, or for which a copy thereof is attached.

5. From: _____ To: _____

The document was recorded in the United States Patent and Trademark Office at

Reel _____, Frame _____, or for which a copy thereof is attached.

6. From: _____ To: _____

The document was recorded in the United States Patent and Trademark Office at

Reel _____, Frame _____, or for which a copy thereof is attached.

☐ Additional documents in the chain of title are listed on a supplemental sheet(s).

☐ As required by 37 CFR 3.73(c)(1)(i), the documentary evidence of the chain of title from the original owner to the assignee was, or concurrently is being, submitted for recordation pursuant to 37 CFR 3.11.

[NOTE: A separate copy (i.e., a true copy of the original assignment document(s)) must be submitted to Assignment Division in accordance with 37 CFR Part 3, to record the assignment in the records of the USPTO. See MPEP 302.08]

The undersigned (whose title is supplied below) is authorized to act on behalf of the assignee.

/James G. Gatto/	November 9, 2015
Signature	Date
James G. Gatto	32,694
Printed or Typed Name	Title or Registration Number

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that yoube given certain informationin connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, pleasebe advised that: (1) the general authority forthe collection of thisinformation is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and(3) the principal purpose forwhich the information isused by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent applicationor patent. If you do not furnish the requested information,the U.S. Patent and Trademark Office may not be able to process and/or examineyour submission,which may result in termination of proceedings or abandonment of the applicationor expiration of the patent.

The informationprovided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the informationin order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an InternationalApplication filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, arecord may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from thissystem of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
11/580,926	10/16/2006	Larry Baldwin	090611-0357945

CONFIRMATION NO. 6670

909
Pillsbury Winthrop Shaw Pittman, LLP (NV)
PO Box 10500
McLean, VA 22102

POWER OF ATTORNEY NOTICE

OC000000078826394

Date Mailed: 11/25/2015

NOTICE REGARDING CHANGE OF POWER OF ATTORNEY

This is in response to the Power of Attorney filed 11/09/2015.

- The Power of Attorney to you in this application has been revoked by the assignee who has intervened as provided by 37 CFR 3.71. Future correspondence will be mailed to the new address of record(37 CFR 1.33).

Questions about the contents of this notice and the requirements it sets forth should be directed to the Office of Data Management, Application Assistance Unit, at **(571) 272-4000** or **(571) 272-4200** or **1-888-786-0101**.

/ytdemisse/

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
11/580,926	10/16/2006	Larry Baldwin	090611-0357945

CONFIRMATION NO. 6670

133759
Sheppard Mullin Richter & Hampton LLP
12275 El Camino Real, Suite 200
San Diego, CA 92130

POA ACCEPTANCE LETTER

OC000000078826419

Date Mailed: 11/25/2015

NOTICE OF ACCEPTANCE OF POWER OF ATTORNEY

This is in response to the Power of Attorney filed 11/09/2015.

The Power of Attorney in this application is accepted. Correspondence in this application will be mailed to the above address as provided by 37 CFR 1.33.

Questions about the contents of this notice and the requirements it sets forth should be directed to the Office of Data Management, Application Assistance Unit, at **(571) 272-4000** or **(571) 272-4200** or **1-888-786-0101**.

/ytdemisse/

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

INVENTORS :	Larry BALDWIN *et al.*	CONFIRMATION NO.:	6670
SERIAL NUMBER :	11/580,926	PATENT NO.:	8,073,681
FILING DATE :	October 16, 2006	ISSUE DATE:	December 6, 2011
FOR :	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE		

DEFICIENCY PAYMENT UNDER 37 C.F.R. § 1.28(c)

Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313-1450

Sir:

This paper is being submitted to address a good faith error in payment of small entity fees. In this application, small entity status was established in good faith, and fees as a small entity were paid in good faith. It was later discovered that such status as a small entity was established in error. As such, Applicant believes small entity status is no longer appropriate in the instant application. Applicant notes that the erroneous payments of small entity fees were made inadvertently, without deceptive intent.

Pursuant to 37 C.F.R. § 1.28(c)(1)-(2), the fees paid in error and an itemization of the deficiencies owed for each previous fee erroneously paid as a small entity is attached as **APPENDIX A**, including an itemization of the requirements of 37 C.F.R. § 1.28(c)(2)(ii).

Accordingly, pursuant to 37 C.F.R. § 1.28(c), Applicant requests that the Commissioner excuse the good faith error in payment of small entity fees, and charge the deficiency payment owed, which is calculated to be **$7,640.00,** to Deposit Account No. 504561 **(Ref. No. 62KC-218162).**

Respectfully submitted,

SHEPPARD MULLIN RICHTER & HAMPTON LLP

By /Mark A. Patrick/
Mark A. Patrick
Registration No.: 72,958

2099 Pennsylvania Avenue, NW
Suite 100
Washington, D.C. 20006-6801
Telephone: (202) 747-2192

Customer No.
133759

Dated: June 5, 2019

Deficiency Payment Under 37 C.F.R. § 1.28(c)
Application Serial No.: 11/580,926
Patent No.: 8,073,681
Attorney Docket No.: 62KC-218162

APPENDIX A
ITEMIZATION OF DEFICIENCY PAYMENTS

Deficient Payment	Type of Fee Erroneously Paid as Small Entity 37 C.F.R. 1.28(c)(2)(ii)(A)	Current Fee for Non-Small Entity 37 C.F.R. 1.28(c)(2)(ii)(A)	Small Entity Fee Actually Paid 37 C.F.R. 1.28(c)(2)(ii)(B)	Date Small Entity Fee Paid 37 C.F.R. 1.28(c)(2)(ii)(B)	Deficiency Owed 37 C.F.R. 1.28(c)(2)(ii)(C)
First Deficient Payment	Basic Filing Fee – Utility – Fee Code 2011	$300.00	$150.00	Application filed October 16, 2006	$300.00 - $150.00 = **$150.00**
Second Deficient Payment	Utility Search Fee – Fee Code 2111	$660.00	$250.00	Application filed October 16, 2006	$660.00 - $250.00 = **$410.00**
Third Deficient Payment	Utility Examination Fee – Fee Code 2311	$760.00	$100.00	Application filed October 16, 2006	$760.00 - $100.00 = **$660.00**
Fourth Deficient Payment	22 Claims in excess of 20 – Fee Code 2202	$2,200.00	$550.00	Application filed October 16, 2006	$2,200.00 - $550.00 = **$1,650.00**
Fifth Deficient Payment	Surcharge – Late filing fee, search fee, examination fee, inventor's oath or declaration – Fee Code 2051	$160.00	$65.00	Response to Notice to File Missing Parts filed January 5, 2007	$160.00 - $65.00 = **$95.00**
Sixth Deficient Payment	Request for Continued Examination – Fee Code 2801	$1,300.00	$405.00	Response filed January 21, 2011	$1,300.00 - $405.00 = **$895.00**
Seventh Deficient Payment	3 Independent claims in excess of 3 – Fee Code 2201	$1,380.00	$330.00	Response filed January 21, 2011	$1,380.00 - $330.00 = **$1,050.00**
Eighth Deficient Payment	Utility Issue Fee – Fee Code 2501	$1,000.00	$870.00	Issue Fee paid October 4, 2011	$1,000.00 - $870.00 = **$130.00**

SMRH:4827-8384-7064.1

Deficiency Payment Under 37 C.F.R. § 1.28(c)
Application Serial No.: 11/580,926
Patent No.: 8,073,681
Attorney Docket No.: 62KC-218162

NINTH DEFICIENT PAYMENT	3.5-Year Maintenance Fee – Fee Code 2551	$1,600.00	$800.00	Maintenance Fee paid May 20, 2015	$1,600.00 - $800.00 = **$800.00**
TENTH DEFICIENT PAYMENT	7.5-Year Maintenance Fee – Fee Code 2552	$3,600.00	$1,800.00	Maintenance Fee paid May 23, 2019	$3,600.00 - $1,800.00 = **$1,800.00**
TOTAL DEFICIENCY PAYMENT OWED **37 C.F.R. 1.28(c)(2)(ii)(D)**	$150.00 + $410.00 + $660.00 + $1,650.00 + $95.00 + $895.00 + $1,050.00 + $130.00 + 800.00 + $1,800.00 = **$7,640.00** Total Deficiency Owed				

SMRH:4827-8384-7064.1

United States Patent and Trademark Office

- Sales Receipt -

06/07/2019 DALLEN 00000021 504561 11580926

01 FC:1599 5040.00 DA

United States Patent and Trademark Office

- Sales Receipt -

06/07/2019 DALLEN 00000022 504561 8073681

01 FC:1599 2600.00 DA

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY.DOCKET NO./TITLE	REQUEST ID
11/580,926	10/16/2006	Larry Baldwin	62KC-218162	89169

Acknowledgement of Loss of Entitlement to Entity Status Discount

The entity status change request below filed through Private PAIR on 06/05/2019 has been accepted.

CERTIFICATIONS:

Change of Entity Status:

☒ Applicant changing to regular undiscounted fee status.

NOTE: Checking this box will be taken to be notification of loss of entitlement to small or micro entity status, as applicable.

This portion must be completed by the signatory or signatories making the entity status change in accordance with 37 CFR 1.4(d)(4).

Signature:	/Mark A. Patrick/
Name:	Mark A. Patrick
Registration Number:	72958

Document code: WFEE

United States Patent and Trademark Office
Sales Receipt for Accounting Date: 07/19/2019

CKHLOK ADJ #00000011 Mailroom Dt: 06/05/2019
Seq No: 22 Sales Acctg Dt: 06/07/2019 11580926
01 FC : 1599 2600.00 CR

Document code: WFEE

United States Patent and Trademark Office
Sales Receipt for Accounting Date: 07/19/2019

CKHLOK ADJ #00000010 Mailroom Dt: 06/05/2019
Seq No: 21 Sales Acctg Dt: 06/07/2019 504561 11580926
01 FC : 1599 5040.00 CR

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

INVENTORS :	Larry BALDWIN *et al.*	CONFIRMATION NO.:	6670
SERIAL NUMBER :	11/580,926	PATENT NO.:	8,073,681
FILING DATE :	October 16, 2006	ISSUE DATE:	December 6, 2011
FOR :	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE		

PETITION TO EXPEDITE UNDER 37 C.F.R.§ 1.182

Office of Petitions
P. O. Box 1450
Alexandria, VA 22313-1450

Sir:

This is a Petition to Expedite Under 37 C.F.R. § 1.182 to expedite the processing of the Deficiency Payment Under 37 C.F.R. § 1.28(c) filed herewith.

Please charge the processing fee required under 37 C.F.R. § 1.17(f) in the amount of $400.00 for this Petition to Expedite to Deposit Account No. 504561 (Ref. No. **62KC-218162**). Please contact the undersigned by telephone if there are any questions regarding this matter.

Respectfully submitted,

SHEPPARD MULLIN RICHTER & HAMPTON LLP

By /Mark A. Patrick/
Mark A. Patrick
Registration No.: 72,958

Customer No.
133759

2099 Pennsylvania Avenue, NW
Suite 100
Washington, D.C. 20006-6801
Telephone: (202) 747-2192

Dated: June 5, 2019

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

INVENTORS :	Larry BALDWIN *et al.*	CONFIRMATION NO.:	6670
SERIAL NUMBER :	11/580,926	PATENT NO.:	8,073,681
FILING DATE :	October 16, 2006	ISSUE DATE:	December 6, 2011
FOR :	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE		

DEFICIENCY PAYMENT UNDER 37 C.F.R. § 1.28(c)

Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313-1450

Sir:

This paper is being submitted to address a good faith error in payment of small entity fees. In this application, small entity status was established in good faith, and fees as a small entity were paid in good faith. It was later discovered that such status as a small entity was established in error. As such, Applicant believes small entity status is no longer appropriate in the instant application. Applicant notes that the erroneous payments of small entity fees were made inadvertently, without deceptive intent.

Pursuant to 37 C.F.R. § 1.28(c)(1)-(2), the fees paid in error and an itemization of the deficiencies owed for each previous fee erroneously paid as a small entity is attached as **APPENDIX A**, including an itemization of the requirements of 37 C.F.R. § 1.28(c)(2)(ii).

Accordingly, pursuant to 37 C.F.R. § 1.28(c), Applicant requests that the Commissioner excuse the good faith error in payment of small entity fees, and charge the deficiency payment owed, which is calculated to be **$7,640.00,** to Deposit Account No. 504561 **(Ref. No. 62KC-218162).**

Respectfully submitted,

SHEPPARD MULLIN RICHTER & HAMPTON LLP

By /Mark A. Patrick/
Mark A. Patrick
Registration No.: 72,958

2099 Pennsylvania Avenue, NW
Suite 100
Washington, D.C. 20006-6801
Telephone: (202) 747-2192

Customer No.
133759

Dated: June 5, 2019

Deficiency Payment Under 37 C.F.R. § 1.28(c)
Application Serial No.: 11/580,926
Patent No.: 8,073,681
Attorney Docket No.: 62KC-218162

APPENDIX A
ITEMIZATION OF DEFICIENCY PAYMENTS

Deficient Payment	Type of Fee Erroneously Paid as Small Entity 37 C.F.R. 1.28(c)(2)(ii)(A)	Current Fee for Non-Small Entity 37 C.F.R. 1.28(c)(2)(ii)(A)	Small Entity Fee Actually Paid 37 C.F.R. 1.28(c)(2)(ii)(B)	Date Small Entity Fee Paid 37 C.F.R. 1.28(c)(2)(ii)(B)	Deficiency Owed 37 C.F.R. 1.28(c)(2)(ii)(C)
First Deficient Payment	Basic Filing Fee – Utility – Fee Code 2011	$300.00	$150.00	Application filed October 16, 2006	$300.00 - $150.00 = **$150.00**
Second Deficient Payment	Utility Search Fee – Fee Code 2111	$660.00	$250.00	Application filed October 16, 2006	$660.00 - $250.00 = **$410.00**
Third Deficient Payment	Utility Examination Fee – Fee Code 2311	$760.00	$100.00	Application filed October 16, 2006	$760.00 - $100.00 = **$660.00**
Fourth Deficient Payment	22 Claims in excess of 20 – Fee Code 2202	$2,200.00	$550.00	Application filed October 16, 2006	$2,200.00 - $550.00 = **$1,650.00**
Fifth Deficient Payment	Surcharge – Late filing fee, search fee, examination fee, inventor's oath or declaration – Fee Code 2051	$160.00	$65.00	Response to Notice to File Missing Parts filed January 5, 2007	$160.00 - $65.00 = **$95.00**
Sixth Deficient Payment	Request for Continued Examination – Fee Code 2801	$1,300.00	$405.00	Response filed January 21, 2011	$1,300.00 - $405.00 = **$895.00**
Seventh Deficient Payment	3 Independent claims in excess of 3 – Fee Code 2201	$1,380.00	$330.00	Response filed January 21, 2011	$1,380.00 - $330.00 = **$1,050.00**
Eighth Deficient Payment	Utility Issue Fee – Fee Code 2501	$1,000.00	$870.00	Issue Fee paid October 4, 2011	$1,000.00 - $870.00 = **$130.00**

SMRH:4827-8384-7064.1

NINTH DEFICIENT PAYMENT	3.5-Year Maintenance Fee -- Fee Code 2551	$1,600.00	$800.00	Maintenance Fee paid May 20, 2015	$1,600.00 - $800.00 = **$800.00**
TENTH DEFICIENT PAYMENT	7.5-Year Maintenance Fee -- Fee Code 2552	$3,600.00	$1,800.00	Maintenance Fee paid May 23, 2019	$3,600.00 - $1,800.00 = **$1,800.00**
TOTAL DEFICIENCY PAYMENT OWED **37 C.F.R. 1.28(c)(2)(ii)(D)**	$150.00 + $410.00 + $660.00 + $1,650.00 + $95.00 + $895.00 + $1,050.00 + $130.00 + 800.00 + $1,800.00 = **$7,640.00** Total Deficiency Owed				

Electronic Patent Application Fee Transmittal

Application Number:	11580926
Filing Date:	16-Oct-2006
Title of Invention:	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE
First Named Inventor/Applicant Name:	Larry Baldwin
Filer:	Mark Andrew Patrick/Diane Bogusz
Attorney Docket Number:	62KC-218162

Filed as Large Entity

Filing Fees for Utility under 35 USC 111(a)

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
PETITION FEE- 37 CFR 1.17(F) (GROUP I)	1462	1	400	400
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Extension-of-Time:				
Miscellaneous:				
			Total in USD ($)	**400**

Electronic Acknowledgement Receipt

EFS ID:	36213069
Application Number:	11580926
International Application Number:	
Confirmation Number:	6670
Title of Invention:	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE
First Named Inventor/Applicant Name:	Larry Baldwin
Customer Number:	133759
Filer:	Mark Andrew Patrick/Diane Bogusz
Filer Authorized By:	Mark Andrew Patrick
Attorney Docket Number:	62KC-218162
Receipt Date:	05-JUN-2019
Filing Date:	16-OCT-2006
Time Stamp:	16:08:13
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	DA
Payment was successfully received in RAM	$400
RAM confirmation Number	060619INTEFSW00002905504561
Deposit Account	
Authorized User	
The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:	

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Petition for review by the Office of Petitions	62KC-218162_Petition_to_Expedite.pdf	111963 ebfb292db0edae3000f27c4aedaa4c1e9a9e9cfd	no	1
Warnings:					
Information:					
2	Notification of loss of entitlement to small entity status	62KC-218162_Deficiency_Payment.pdf	148943 7bd1f7d1fd124423912229b53a9a3e189ebeda41	no	4
Warnings:					
Information:					
3	Fee Worksheet (SB06)	fee-info.pdf	30757 94d32db33a52242627d8782384e4928ed439a9ee	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	291663		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
11/580,926	10/16/2006	Larry Baldwin	62KC-218162	6670

133759 7590 07/05/2019

Sheppard Mullin Richter & Hampton LLP
650 Town Center Drive, 10th Floor
Costa Mesa, CA 92626

EXAMINER
YEN, ERIC L

ART UNIT	PAPER NUMBER
2626	

NOTIFICATION DATE	DELIVERY MODE
07/05/2019	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

SheppardMullin_Pair@firsttofile.com
dmipdocketing@sheppardmullin.com

PTOL-90A (Rev. 04/07)

UNITED STATES PATENT AND TRADEMARK OFFICE

Commissioner for Patents
United States Patent and Trademark Office
P.O. Box 1450
Alexandria, VA 22313-1450
www.uspto.gov

In re Application of	:	
Baldwin et al.	:	
Application No. 11/580,926	:	
Filed: 16 Oct 2006	:	DECISION ON PETITION
For: SYSTEM AND METHOD FOR A	:	
COOPERATIVE CONVERSATIONAL	:	
VOICE USER INTERFACE	:	

This is a decision in response to the petition under 37 CFR 1.182, filed June 5, 2019 to expedite consideration of a simultaneously filed petition, for acceptance of a fee deficiency submission under 37 CFR 1.28(c).

The file does not indicate a change of address has been submitted, although the address given on the petition differs from the address of record. If appropriate, a change of address should be filed in accordance with MPEP 601.03. A courtesy copy of this decision is being mailed to the address given on the petition; however, the Office will mail all future correspondence solely to the address of record.

The petition under 37 CFR 1.182 is **GRANTED**. A decision on the request for acceptance of a fee deficiency under 37 CFR 1.28(c) follows.

Your fee deficiency submission under 37 CFR 1.28(c) is hereby **DISMISSED**.

It is noted that an itemization of the fee deficiency was submitted; however, the calculation of the deficiency owed is not in compliance with the requirements of 37 CFR 1.28(c)(2)(i), which reads:

> (i) *Calculation of the deficiency owed.* The deficiency owed for each previous fee erroneously paid as a small entity is the difference between the current fee amount (for other than a small entity) on the date the deficiency is paid in full and the amount of the previous erroneous (small entity) fee payment. The total deficiency payment owed is the sum of the individual deficiency owed amounts for each fee amount previously erroneously paid as a small entity. Where a fee paid in error as a small entity was subject to a fee decrease between the time the fee was paid in error and the time the deficiency is paid in full, the deficiency owed is equal to the amount (previously) paid in error;" Emphasis added

37 CFR 1.28(c)(2)(ii) goes on to say:

> (ii) *Itemization of the deficiency payment.* An itemization of the total deficiency payment is required. The itemization must include the following information:
>
> (A) Each particular type of fee that was erroneously paid as a small entity, (*e.g.*, basic statutory filing fee, two-month extension of time fee) along with the current fee amount for a non-small entity;
>
> (B) The small entity fee actually paid, and when. This will permit the Office to differentiate, for example, between two one-month extension of time fees erroneously paid as a small entity but on different dates;
>
> (C) The deficiency owed amount (for each fee erroneously paid); and
>
> (D) The total deficiency payment owed, which is the sum or total of the individual deficiency owed amounts set forth in paragraph (c)(2)(ii)(C) of this section.

In reviewing the application it is noted, that the issue fee is incorrect since it was paid before the fees was reduce, it is not entitle to the lower amount. ($1740-$1000 = $740 deficient).

Applicant is given **ONE MONTH or THIRTY DAYS**, whichever is longer, from the mailing date of this decision to submit the fees required pursuant to 37 CFR 1.28(2)(i) and a replacement itemized listing of each fee erroneously paid as a small entity in compliance with 37 CFR 1.28(c)(2)(ii). **NO EXTENSION OF TIME UNDER 37 CFR 1.136 IS PERMITTED.**

Further correspondence with respect to this matter should be delivered through one of the following mediums:

By mail:	Mail Stop PETITIONS Commissioner for Patents Post Office Box 1450 Alexandria, VA 22313-1450
By hand:	Customer Service Window Mail Stop Petitions Randolph Building 401 Dulany Street Alexandria, VA 22314
By fax:	(571) 273-8300 ATTN: Office of Petitions
By Internet:	EFS-Web[1]

[1] www.uspto.gov/ebc/efs_help.html (for help using EFS-Web call the Patent Electronic Business Center at (866) 217-9197)

Any questions concerning this matter may be directed to the undersigned at (571) 272-1058.

/Angela Walker/
Angela Walker
Paralegal Specialist
Office of Petitions

cc: Mark A. Patrick
2099 Pennsylvania Ave., NW
Ste. 100
Washington, DC 20006-6801

Transmittal Communication on Petition	Application/Control No. 11/580,926	Applicant(s)/Patent Under Reexamination Baldwin et al.	
	Deciding Official YEN, ERIC L	Office of Petitions OPET	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address. --

(ADDITIONAL PARTY'S CORRESPONDENCE ADDRESS)

Mark A. Patrick
2099 Pennsylvania Ave., NW
Ste. 100
Washington, DC 20006-6801

Enclosed is a copy of the latest communication from the United States Patent and Trademark Office in the above-identified Application/Patent.

Document code: WFEE

United States Patent and Trademark Office
Sales Receipt for Accounting Date: 07/19/2019

CKHLOK SALE #00000010 Mailroom Dt: 07/09/2019 504561 11580926
01 FC : 1461 5,780.00 DA

Document code: WFEE

United States Patent and Trademark Office
Sales Receipt for Accounting Date: 07/19/2019

CKHLOK SALE #00000011 Mailroom Dt: 07/09/2019 11580926

01	FC : 1559	2,600.00	DA

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

INVENTORS :	Larry BALDWIN *et al.*	CONFIRMATION NO.:	6670
SERIAL NUMBER :	11/580,926	PATENT NO.:	8,073,681
FILING DATE :	October 16, 2006	ISSUE DATE:	December 6, 2011
FOR :	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE		

CORRECTED DEFICIENCY PAYMENT UNDER 37 C.F.R. § 1.28(c)

Commissioner for Patents
P. O. Box 1450
Alexandria, VA 22313-1450

Sir:

In response to the Decision on Petition mailed July 5, 2019, this paper is being submitted to address a good faith error in payment of small entity fees. In this application, small entity status was established in good faith, and fees as a small entity were paid in good faith. It was later discovered that such status as a small entity was established in error. As such, Applicant believes small entity status is no longer appropriate in the instant application. Applicant notes that the erroneous payments of small entity fees were made inadvertently, without deceptive intent.

Pursuant to 37 C.F.R. § 1.28(c)(1)-(2), the fees paid in error and an itemization of the deficiencies owed for each previous fee erroneously paid as a small entity is attached as **APPENDIX A**, including an itemization of the requirements of 37 C.F.R. § 1.28(c)(2)(ii).

Accordingly, pursuant to 37 C.F.R. § 1.28(c), Applicant requests that the Commissioner excuse the good faith error in payment of small entity fees, and charge the remaining deficiency payment owed, which is calculated to be **$740.00,** to Deposit Account No. 504561 **(Ref. No. 62KC-218162).**

Respectfully submitted,

SHEPPARD MULLIN RICHTER & HAMPTON LLP

By /Mark A. Patrick/
Mark A. Patrick
Registration No.: 72,958

Customer No.
133759

2099 Pennsylvania Avenue, NW
Suite 100
Washington, D.C. 20006-6801
Telephone: (202) 747-2192

Dated: July 9, 2019

APPENDIX A
ITEMIZATION OF DEFICIENCY PAYMENTS

Deficient Payment	Type of Fee Erroneously Paid as Small Entity 37 C.F.R. 1.28(c)(2)(ii)(A)	Current Fee for Non-Small Entity 37 C.F.R. 1.28(c)(2)(ii)(A)	Small Entity Fee Actually Paid 37 C.F.R. 1.28(c)(2)(ii)(B)	Date Small Entity Fee Paid 37 C.F.R. 1.28(c)(2)(ii)(B)	Deficiency Owed 37 C.F.R. 1.28(c)(2)(ii)(C)
First Deficient Payment	Basic Filing Fee – Utility – Fee Code 2011	$300.00	$150.00	Application filed October 16, 2006	$300.00 - $150.00 = **$150.00**
Second Deficient Payment	Utility Search Fee – Fee Code 2111	$660.00	$250.00	Application filed October 16, 2006	$660.00 - $250.00 = **$410.00**
Third Deficient Payment	Utility Examination Fee – Fee Code 2311	$760.00	$100.00	Application filed October 16, 2006	$760.00 - $100.00 = **$660.00**
Fourth Deficient Payment	22 Claims in excess of 20 – Fee Code 2202	$2,200.00	$550.00	Application filed October 16, 2006	$2,200.00 - $550.00 = **$1,650.00**
Fifth Deficient Payment	Surcharge – Late filing fee, search fee, examination fee, inventor's oath or declaration – Fee Code 2051	$160.00	$65.00	Response to Notice to File Missing Parts filed January 5, 2007	$160.00 - $65.00 = **$95.00**
Sixth Deficient Payment	Request for Continued Examination – Fee Code 2801	$1,300.00	$405.00	Response filed January 21, 2011	$1,300.00 - $405.00 = **$895.00**
Seventh Deficient Payment	3 Independent claims in excess of 3 – Fee Code 2201	$1,380.00	$330.00	Response filed January 21, 2011	$1,380.00 - $330.00 = **$1,050.00**
Eighth Deficient Payment	Utility Issue Fee – Fee Code 2501	~~$1,000.00~~ $1,740.00	$870.00 $130.00	Issue Fee paid October 4, 2011 Deficiency paid June 7, 2019	$1,740.00 - $870.00 - $130.00 = **$740.00**

NINTH DEFICIENT PAYMENT	3.5-Year Maintenance Fee -- Fee Code 2551	$1,600.00	$800.00	Maintenance Fee paid May 20, 2015	$1,600.00 - $800.00 = **$800.00**
TENTH DEFICIENT PAYMENT	7.5-Year Maintenance Fee -- Fee Code 2552	$3,600.00	$1,800.00	Maintenance Fee paid May 23, 2019	$3,600.00 - $1,800.00 = **$1,800.00**
TOTAL DEFICIENCY PAYMENT OWED **37 C.F.R. 1.28(c)(2)(ii)(D)**	$150.00 + $410.00 + $660.00 + $1,650.00 + $95.00 + $895.00 + $1,050.00 + $130.00 + $740.00 + 800.00 + $1,800.00 = ~~**$7,640.00**~~ **$8,380.00** Total Deficiency Owed Please note that $5,040.00 and $2,600.00 were previously charged on June 7, 2019. Accordingly, the remaining deficiency payment owed is **$740.00.**				

Electronic Acknowledgement Receipt

EFS ID:	36534518
Application Number:	11580926
International Application Number:	
Confirmation Number:	6670
Title of Invention:	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE
First Named Inventor/Applicant Name:	Larry Baldwin
Customer Number:	133759
Filer:	Mark Andrew Patrick/Susan Trader
Filer Authorized By:	Mark Andrew Patrick
Attorney Docket Number:	62KC-218162
Receipt Date:	09-JUL-2019
Filing Date:	16-OCT-2006
Time Stamp:	16:49:49
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Petition for review by the Office of Petitions	62KC-218162_Corrected_Deficiency_Payment.pdf	151254 c362faeed9045569cf1a804ca02b294f4d0fe5c1	no	4

Warnings:

Information:	
Total Files Size (in bytes):	151254

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111

If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.

National Stage of an International Application under 35 U.S.C. 371

If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.

New International Application Filed with the USPTO as a Receiving Office

If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

UNITED STATES PATENT AND TRADEMARK OFFICE

UNITED STATES DEPARTMENT OF COMMERCE
United States Patent and Trademark Office
Address: COMMISSIONER FOR PATENTS
P.O. Box 1450
Alexandria, Virginia 22313-1450
www.uspto.gov

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
11/580,926	10/16/2006	Larry Baldwin	62KC-218162	6670

133759 7590 07/25/2019

Sheppard Mullin Richter & Hampton LLP
650 Town Center Drive, 10th Floor
Costa Mesa, CA 92626

EXAMINER
YEN, ERIC L

ART UNIT	PAPER NUMBER
2626	

NOTIFICATION DATE	DELIVERY MODE
07/25/2019	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

SheppardMullin_Pair@firsttofile.com
dmipdocketing@sheppardmullin.com

PTOL-90A (Rev. 04/07)

UNITED STATES PATENT AND TRADEMARK OFFICE

Commissioner for Patents
United States Patent and Trademark Office
P.O. Box 1450
Alexandria, VA 22313-1450
www.uspto.gov

In re Application of	:	
Baldwin et al.	:	
Application No. 11/580,926	:	
Filed: 16 Oct 2006	:	DECISION ON PETITION
For: SYSTEM AND METHOD FOR A	:	
COOPERATIVE CONVERSATIONAL	:	
VOICE USER INTERFACE	:	

This is a notice regarding your renewed request filed July 9, 2019, for acceptance of a fee deficiency submission under 37 CFR 1.28(c).

The file does not indicate a change of address has been submitted, although the address given on the petition differs from the address of record. If appropriate, a change of address should be filed in accordance with MPEP 601.03. A courtesy copy of this decision is being mailed to the address given on the petition; however, the Office will mail all future correspondence solely to the address of record.

The Office no longer investigates or rejects original or reissue applications under 37 CFR 1.56. **1098 Off. Gaz. Pat. Office 502 (January 3, 1989)**. Therefore, nothing in this Notice is intended to imply that an investigation was done.

Your fee deficiency submission under 37 CFR 1.28(c) is hereby **ACCEPTED**.

Status a small entity has been removed. Accordingly, all future fees paid in this patent must be paid at the undiscounted entity rate.

Inquiries related to this communication should be directed to the undersigned at (571) 272-1058.

/Angela Walker/
Angela Walker
Paralegal Specialist
Office of Petitions

cc: Mark A. Patrick
2099 Pennsylvania Ave., NW
Suite 100
Washington, DC 2006-6801

Transmittal Communication on Petition	Application/Control No. 11/580,926	Applicant(s)/Patent Under Reexamination Baldwin et al.	
	Deciding Official YEN, ERIC L	Office of Petitions OPET	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address. --

(ADDITIONAL PARTY'S CORRESPONDENCE ADDRESS)

Mark A. Patrick
2099 Pennsylvania Ave., NW
Suite 100
Washington, DC 2006-6801

Enclosed is a copy of the latest communication from the United States Patent and Trademark Office in the above-identified Application/Patent.

AO 120 (Rev. 08/10)

TO:	REPORT ON THE FILING OR DETERMINATION OF AN ACTION REGARDING A PATENT OR TRADEMARK
Mail Stop 8 Director of the U.S. Patent and Trademark Office P.O. Box 1450 Alexandria, VA 22313-1450	

In Compliance with 35 U.S.C. § 290 and/or 15 U.S.C. § 1116 you are hereby advised that a court action has been filed in the U.S. District Court for the District of Delaware on the following

☐ Trademarks or ☑ Patents. (☐ the patent action involves 35 U.S.C. § 292.):

DOCKET NO.	DATE FILED	U.S. DISTRICT COURT for the District of Delaware
PLAINTIFF VB Assets, LLC		DEFENDANT Amazon.com, Inc, Amazon.com LLC; Amazon Web Services, Inc.; A2Z Development Center, Inc. d/b/a Lab126; Rawles LLC; AMZN Mobile LLC; AMZN Mobile 2 LLC; Amazon.com Services, Inc. f/k/a Amazon, et al.

	PATENT OR TRADEMARK NO.	DATE OF PATENT OR TRADEMARK	HOLDER OF PATENT OR TRADEMARK
1	US 8,073,681 B2	12/6/2011	VB Assets, LLC
2	US 9,015,049 B2	4/21/2015	VB Assets, LLC
3	US 9,626,703 B2	4/18/2017	VB Assets, LLC
4	US 7,818,176 B2	10/19/2010	VB Assets, LLC
5	US 8,886,536 B2	11/11/2014	VB Assets, LLC

In the above—entitled case, the following patent(s)/ trademark(s) have been included:

DATE INCLUDED	INCLUDED BY ☐ Amendment ☐ Answer ☐ Cross Bill ☐ Other Pleading		
	PATENT OR TRADEMARK NO.	DATE OF PATENT OR TRADEMARK	HOLDER OF PATENT OR TRADEMARK
1			
2			
3			
4			
5			

In the above—entitled case, the following decision has been rendered or judgement issued:

DECISION/JUDGEMENT

CLERK	(BY) DEPUTY CLERK	DATE

Copy 1—Upon initiation of action, mail this copy to Director **Copy 3—Upon termination of action, mail this copy to Director**
Copy 2—Upon filing document adding patent(s), mail this copy to Director **Copy 4—Case file copy**

AO 120 (Rev. 08/10)

TO:	REPORT ON THE FILING OR DETERMINATION OF AN ACTION REGARDING A PATENT OR TRADEMARK
Mail Stop 8 Director of the U.S. Patent and Trademark Office P.O. Box 1450 Alexandria, VA 22313-1450	

In Compliance with 35 U.S.C. § 290 and/or 15 U.S.C. § 1116 you are hereby advised that a court action has been filed in the U.S. District Court for the District of Delaware on the following

☐ Trademarks or ☑ Patents. (☐ the patent action involves 35 U.S.C. § 292.):

DOCKET NO.	DATE FILED	U.S. DISTRICT COURT for the District of Delaware
PLAINTIFF VB Assets, LLC		DEFENDANT Amazon.com, Inc, Amazon.com LLC; Amazon Web Services, Inc.; A2Z Development Center, Inc. d/b/a Lab126; Rawles LLC; AMZN Mobile LLC; AMZN Mobile 2 LLC; Amazon.com Services, Inc. f/k/a Amazon, et al.

	PATENT OR TRADEMARK NO.	DATE OF PATENT OR TRADEMARK	HOLDER OF PATENT OR TRADEMARK
1	US 9,269,097 B2	2/23/2016	VB Assets, LLC
2			
3			
4			
5			

In the above—entitled case, the following patent(s)/ trademark(s) have been included:

DATE INCLUDED	INCLUDED BY ☐ Amendment ☐ Answer ☐ Cross Bill ☐ Other Pleading	
PATENT OR TRADEMARK NO.	DATE OF PATENT OR TRADEMARK	HOLDER OF PATENT OR TRADEMARK
1		
2		
3		
4		
5		

In the above—entitled case, the following decision has been rendered or judgement issued:

DECISION/JUDGEMENT

CLERK	(BY) DEPUTY CLERK	DATE

Copy 1—Upon initiation of action, mail this copy to Director **Copy 3—Upon termination of action, mail this copy to Director**
Copy 2—Upon filing document adding patent(s), mail this copy to Director **Copy 4—Case file copy**

PTO/SB/43 (07-09)
Approved for use through 11/30/2020. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

DISCLAIMER IN PATENT UNDER 37 CFR 1.321(a)

Name of Patentee	Docket Number (Optional)
VB Assets, LLC	62KC-218162
Patent Number	**Date Patent Issued**
8,073,681	12-06-2011

Title of Invention
SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

I hereby disclaim the following complete claims in the above identified patent: claims 37-42

The extent of my interest in said patent is (if assignee of record, state liber and page, or reel and frame, where assignment is recorded): Interest: 100%; Reel/Frame Nos. : 046456/0128 and 053851/0873

The fee for this disclaimer is set forth in 37 CFR 1.20(d).

- [ ] Patentee claims small entity status. See 37 CFR 1.27.
- [ ] Small entity status has already been established in this case, and is still proper.
- [ ] A check in the amount of the fee is enclosed.
- [ ] Payment by credit card. Form PTO-2038 is attached.
- [X] The Director is hereby authorized to charge any fees which may be required or credit any overpayment to Deposit Account No. 504561 .

WARNING: Information on this form may become public. Credit card information should not be included on this form. Provide credit card information and authorization on PTO-2038.

Signed at Bellevue, State of Washington this 15th day of May 2021.

MKennewick

Signature

Registration Number, if applicable

VB Assets, LLC Mike Kennewick, Manager

Typed or printed name of patentee/ attorney or agent of record

(425) 891-6537

Telephone Number

1229 120th Ave NE Suite A

Address

Bellevue, WA. 98005

City, State, Zip Code or Foreign Country as applicable

This collection of information is required by 37 CFR 1.321. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

If you need assistance in completing the form, call 1-800-PTO-9199 and select option 2.

Electronic Patent Application Fee Transmittal

Application Number:	11580926
Filing Date:	16-Oct-2006
Title of Invention:	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE
First Named Inventor/Applicant Name:	Larry Baldwin
Filer:	Timothy P. Cremen/Diane Bogusz
Attorney Docket Number:	62KC-218162

Filed as Large Entity

Filing Fees for Utility under 35 USC 111(a)

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Miscellaneous:				
STATUTORY OR TERMINAL DISCLAIMER	1814	1	170	170
		Total in USD ($)		**170**

Electronic Acknowledgement Receipt

EFS ID:	42741834
Application Number:	11580926
International Application Number:	
Confirmation Number:	6670
Title of Invention:	SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE
First Named Inventor/Applicant Name:	Larry Baldwin
Customer Number:	133759
Filer:	Timothy P. Cremen/Diane Bogusz
Filer Authorized By:	Timothy P. Cremen
Attorney Docket Number:	62KC-218162
Receipt Date:	17-MAY-2021
Filing Date:	16-OCT-2006
Time Stamp:	16:27:44
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	DA
Payment was successfully received in RAM	$170
RAM confirmation Number	E20215GG28112121
Deposit Account	
Authorized User	
The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:	

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Terminal Disclaimer Filed	62KC-218162_Disclaimer.pdf	454581 42a3273ccdacd979f6b0e3b0b4dccfd4ef990a07	no	1
Warnings:					
Information:					
2	Fee Worksheet (SB06)	fee-info.pdf	30534 3c632d83e18d6b2f765887ec699693ccea4c9bfe	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	485115		

This Acknowledgement Receipt evidences receipt on the noted date by the USPTO of the indicated documents, characterized by the applicant, and including page counts, where applicable. It serves as evidence of receipt similar to a Post Card, as described in MPEP 503.

New Applications Under 35 U.S.C. 111
If a new application is being filed and the application includes the necessary components for a filing date (see 37 CFR 1.53(b)-(d) and MPEP 506), a Filing Receipt (37 CFR 1.54) will be issued in due course and the date shown on this Acknowledgement Receipt will establish the filing date of the application.
National Stage of an International Application under 35 U.S.C. 371
If a timely submission to enter the national stage of an international application is compliant with the conditions of 35 U.S.C. 371 and other applicable requirements a Form PCT/DO/EO/903 indicating acceptance of the application as a national stage submission under 35 U.S.C. 371 will be issued in addition to the Filing Receipt, in due course.
New International Application Filed with the USPTO as a Receiving Office
If a new international application is being filed and the international application includes the necessary components for an international filing date (see PCT Article 11 and MPEP 1810), a Notification of the International Application Number and of the International Filing Date (Form PCT/RO/105) will be issued in due course, subject to prescriptions concerning national security, and the date shown on this Acknowledgement Receipt will establish the international filing date of the application.

Trials@uspto.gov
571-272-7822

Paper 27
Date: March 7, 2022

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

AMAZON.COM, INC., AMAZON.COM LLC, AMAZON WEB SERVICES, INC., A2Z DEVELOPMENT CENTER, INC. D/B/A LAB126, RAWLES LLC, AMZN MOBILE LLC, AMZN MOBILE 2 LLC, AMAZON.COM SERVICES, INC. F/K/A AMAZON FULFILLMENT SERVICES, INC., and AMAZON.COM SERVICES LLC (formerly AMAZON DIGITAL SERVICES LLC),
Petitioner,

v.

VB ASSETS, LLC,
Patent Owner.

IPR2020-01367
Patent 8,073,681 B2

Before MICHELLE N. WORMMEESTER, SCOTT C. MOORE, and SEAN P. O'HANLON, *Administrative Patent Judges.*

MOORE, *Administrative Patent Judge.*

JUDGMENT
Final Written Decision
Determining No Challenged Claims Unpatentable
35 U.S.C. § 318(a)

I. INTRODUCTION

A. Background and Summary

Amazon.com, Inc., Amazon.com LLC, Amazon Web Services, Inc., A2Z Development Center, Inc. d/b/a Lab126, Rawles LLC, AMZN Mobile LLC, AMZN Mobile 2 LLC, Amazon Services, Inc. f/k/a Amazon Fulfillment Services, Inc., and Amazon.com Services LLC (formerly Amazon Digital Services LLC) (collectively "Petitioner") filed a Petition requesting an *inter partes* review of claims 1–42 of U.S. Patent No. 8,073,681 B2 ("the '681 patent"). Paper 1 ("Petition" or "Pct."). VB Assets, LLC ("Patent Owner") filed a Preliminary Response. Paper 6. We instituted an *inter partes* review as to all claims and grounds set forth in the Petition. Paper 7 ("Institution Decision").

On May 21, 2021, Patent Owner filed a statutory disclaimer disclaiming claims 37–42 of the '681 Patent. Ex. 2005. As explained below, this disclaimer eliminated claims 37–42 from the scope of this proceeding. The same day, Patent Owner filed a Response to the Petition (Paper 13, "Response" or "Resp."). Thereafter, Petitioner filed a Reply to the Response (Paper 16, "Reply"), and Patent Owner filed a Sur-Reply (Paper 23, "Sur-Reply"). An oral hearing was held on November 4, 2021, and a transcript of the hearing is in the record. Paper 26 ("Tr.").

We have jurisdiction under 35 U.S.C. § 6. This Final Written Decision is issued pursuant to 35 U.S.C. § 318(a). For the reasons that follow, we determine that Petitioner has not shown by a preponderance of the evidence that any of claims 1–36 is unpatentable.

B. Related Matters

Patent Owner asserts the '681 patent against Petitioner in the following litigation: *VB Assets, LLC v. Amazon.com Inc., et al.*, Case No. 1:19-cv-01410-MN (D. Del.) (filed July 29, 2019). Pet. 2; Paper 5, 2.

C. The '681 Patent

The '681 patent is directed to "a cooperative conversational model for a human to machine voice user interface." Ex. 1001, 1:7–8. The disclosed cooperative conversational voice user interface understands free-form human utterances, "freeing users from being restricted to a fixed set of commands and/or requests." *Id.* at 1:63–67.

The '681 patent discloses a system that receives an input, which may include a human utterance (i.e., words, syllables, phonemes, or any other audible sound made by a human being), where the utterance includes one or more requests (i.e., command, directive, other instruction for a device, computer or other machine, to retrieve information, perform a task, or take some other action). Ex. 1001, 2:6–14. The utterance component of the input is processed by a speech recognition engine to generate one or more preliminary interpretations of the utterance. *Id.* at 2:16–20. The one or more preliminary interpretations are then provided to a conversational speech engine for further processing, where the conversational speech engine communicates with one or more databases to generate an adaptive conversational response that may then be returned to the user as an output. *Id.* at 2:20–25.

Figure 1 of the '681 patent, reproduced below, depicts an embodiment of a cooperative conversational voice user interface. Ex. 1001, 6:64–65, 7:7–9.

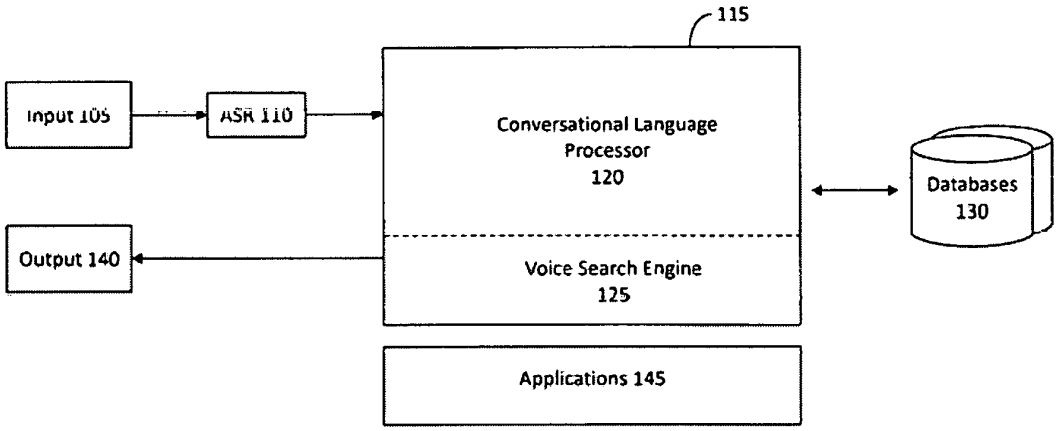


Figure 1

Figure 1 depicts a "system architecture for implementing a cooperative conversational voice user interface." Ex. 1001, 7:7–9. The depicted system receives an input 105 from a user, where input 105 is an utterance received by an input device (e.g., a microphone), and where the utterance includes one or more requests. *Id.* at 7:10–13. The utterance component of input 105 is processed by speech recognition engine 110 (i.e., Automatic Speech Recognizer ("ASR") 110) to generate one or more preliminary interpretations of the utterance. *Id.* at 7:24–28. The one or more preliminary interpretations generated by speech recognition engine 110 are then provided to conversational speech engine 115 for further processing. *Id.* at 7:37–40. Conversational speech engine 115 communicates with one or more databases 130 to generate an adaptive conversational response, which is returned to the user as output 140. *Id.* at 7:43–46.

D. Challenged Claims

Petitioner challenges claims 1–36 of the '681 patent. Petitioner initially also challenged claims 37–42, but Patent Owner filed a statutory disclaimer disclaiming those additional claims. Ex. 2005. This disclaimer

effectively eliminated claims 37–42 from the '681 Patent, leaving the patent as if those claims never existed. *See Sanofi-Aventis U.S., LLC v. Dr. Reddy's Labs., Inc.*, 933 F.3d 1367, 1373 (Fed. Cir. 2019). Thus, claims 37–42 are no longer part of this proceeding.

Challenged claims 1, 13, and 25 are independent. Claims 2–12 depend from claim 1, claims 14–24 depend from claim 13, and claims 26–36 depend from claim 25.

Claim 1 is reproduced below, with bracketed references added.

> 1. A method for providing a cooperative conversational voice user interface, comprising:
>
> [1.a] receiving an utterance at a voice input device during a current conversation with a user, wherein the utterance includes one or more words that have different meanings in different contexts;
>
> [1.b] accumulating short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received during the current conversation;
>
> [1.c] accumulating long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;
>
> [1.d] determining an intended meaning for the utterance, wherein determining the intended meaning for the utterance includes:
>
> [1.e] identifying, at a conversational speech engine, a context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge; and
>
> [1.f] establishing the intended meaning within the identified context, wherein the conversational speech engine establishes the intended meaning within the identified context to disambiguate an intent that the user had in

speaking the one or more words that have the different meanings in the different contexts; and

[1.g] generating a response to the utterance, wherein the conversational speech engine grammatically or syntactically adapts the response based on the intended meaning established within the identified context.

Ex. 1001, 19:35–65.

Independent claims 13 and 25 are similar to independent method claim 1, except that claims 13 and 25 are directed to a non-transitory computer readable medium and a system, respectively. *See* Ex. 1001, 19:35–65, 21:6–35, 22:9–17.

E. Asserted Grounds

Petitioner asserts the following grounds of unpatentability:

Ground	Claims Challenged	35 U.S.C. §	Reference(s)/Basis
1	1, 2, 9, 13, 14, 21, 25, 26, 33	103(a)[1]	Kennewick[2]
2	5, 8, 17, 20, 29, 32	103(a)	Kennewick, Huang[3]
3	10, 22, 34	103(a)	Kennewick, Seneff[4]
4	6, 7, 18, 19, 30, 31	103(a)	Kennewick, Huang, Seneff

[1] The Leahy-Smith America Invents Act ("AIA"), Pub. L. No. 112-29, 125 Stat. 284, 287–88 (2011), amended 35 U.S.C. § 103, effective March 16, 2013. Because the application from which the '681 patent issued was filed before this date, the pre-AIA version of § 103 applies.
[2] U.S. Patent App. Pub. No. 2004/0193420 A1 (pub. Sept. 30, 2004) (Ex. 1003, "Kennewick").
[3] Huang, Xuedong et al., "Spoken Language Processing: A Guide to Theory, Algorithm and System Development," Prentice Hall (2001) (Ex. 1011, "Huang").
[4] Seneff, et al., "Hypothesis Selection and Resolution in the Mercury Flight Reservation System," Spoken Language Systems Group, MIT (2001) (Ex. 1025, "Seneff").

Ground	Claims Challenged	35 U.S.C. §	Reference(s)/Basis
5	11, 12, 23, 24, 35, 36	103(a)	Kennewick, Coffman[5]
6	3, 4, 15, 16, 27, 28	103(a)	Kennewick, Mitsuyoshi[6]

Pet. 4. Petitioner also relies on the Declaration of Padhraic Smyth, Ph.D. (Ex. 1002) and the Supplemental Declaration of Padhraic Smyth, Ph.D. (Ex. 1033). Patent Owner relies on the Declaration of Anatole Gershman, Ph.D. (Ex. 2001).

II. ANALYSIS

A. Principles of Law

1. Burden

In an *inter partes* review, the burden of proof is on the petitioner to show that the challenged claims are unpatentable, and that burden never shifts to the patentee. *See* 35 U.S.C. § 316(e); *In re Magnum Oil Tools Int'l, Ltd.*, 829 F.3d 1364, 1375 (Fed. Cir. 2016) (citing *Dynamic Drinkware, LLC v. Nat'l Graphics, Inc.*, 800 F.3d 1375, 1378 (Fed. Cir. 2015)).

2. Obviousness under 35 U.S.C. § 103(a)

A patent claim is unpatentable under 35 U.S.C. § 103(a) if the differences between the claimed subject matter and the prior art are such that the subject matter, as a whole, would have been obvious at the time the invention was made to a person of ordinary skill in the art to which said subject matter pertains. *KSR Int'l Co. v. Teleflex Inc.*, 550 U.S. 398, 406 (2007). The question of obviousness is resolved on the basis of underlying factual determinations including: (1) the scope and content of the prior art; (2) any differences between the claimed subject matter and the prior art;

[5] U.S. Patent No. 6,839,896 B2 (iss. Jan. 4, 2005) (Ex. 1027, "Coffman").
[6] U.S. Patent App. Pub. No. 2003/0182123 A1 (pub. Sept. 25, 2003) (Ex. 1028, "Mitsuyoshi").

(3) the level of ordinary skill in the art; and (4) when in evidence, objective evidence of nonobviousness. *Graham v. John Deere Co.*, 383 U.S. 1, 17–18 (1966).

In determining obviousness when all elements of a claim are found in various pieces of prior art, "the factfinder must further consider the factual questions of whether a person of ordinary skill in the art would be motivated to combine those references, and whether in making that combination, a person of ordinary skill would have had a reasonable expectation of success." *Dome Patent L.P. v. Lee*, 799 F.3d 1372, 1380 (Fed. Cir. 2015); *see also WMS Gaming, Inc. v. Int'l Game Tech.*, 184 F.3d 1339, 1355 (Fed. Cir. 1999) ("When an obviousness determination relies on the combination of two or more references, there must be some suggestion or motivation to combine the references.").

B. Level of Ordinary Skill in the Art

The level of skill in the art is "a prism or lens" through which we view the prior art and the claimed invention. *Okajima v. Bourdeau*, 261 F.3d 1350, 1355 (Fed. Cir. 2001). Petitioner asserts that a person of ordinary skill in the art (a "POSITA") "would have at least a Bachelor-level degree in computer science, computer engineering, electrical engineering, or a related field in computing technology, and two years of experience with automatic speech recognition and natural language understanding, or equivalent education, research experience, or knowledge." Pet. 4–5 (citing Ex. 1002 ¶¶ 28–31). Patent Owner does not dispute Petitioner's assertion, and Patent Owner and its declarant, Dr. Gershman, both apply Petitioner's definition of a POSITA. Resp. 3.

On this record and given the absence of any dispute, we adopt and apply Petitioner's proposed formulation regarding the level of ordinary skill

in the art, which we find is consistent with the level of skill reflected in the cited prior art references. *See Okajima*, 261 F.3d at 1355.

C. Claim Construction

We generally give the words of a claim their ordinary and customary meaning. *See Phillips v. AWH Corp.*, 415 F.3d 1303, 1312 (Fed. Cir. 2005) (en banc); 37 C.F.R. § 42.100(b). "[T]he ordinary and customary meaning of a claim term is the meaning that the term would have to a person of ordinary skill in the art in question at the time of the invention." *Id.* at 1313. "[T]he person of ordinary skill in the art is deemed to read the claim term not only in the context of the particular claim in which the disputed term appears, but in the context of the entire patent, including the specification." *Id.*

"[W]e need only construe terms 'that are in controversy, and only to the extent necessary to resolve the controversy.'" *Nidec Motor Corp. v. Zhongshan Broad Ocean Motor Co.*, 868 F.3d 1013, 1017 (Fed. Cir. 2017) (quoting *Vivid Techs., Inc. v. Am. Sci. & Eng'g, Inc.*, 200 F.3d 795, 803 (Fed. Cir. 1999)).

In its Petition, Petitioner asks us to construe the "implicit hypothesis" term of claims 38, 40, and 42. Pet. 7–8. As discussed above, however, Patent Owner filed a disclaimer that eliminated these claims from this proceeding. Petitioner does not ask us to construe any terms in the remaining challenged claims. *See id.*; *see generally* Reply. Patent Owner also does not ask us to construe any claim terms. *See generally* Response.

On this record, we decline to adopt constructions of any claims. *See Nidec*, 868 F.3d at 1017.

D. Overview of the Asserted References

1. Kennewick

Kennewick is directed to "the retrieval of online information and processing of commands through a speech interface in a vehicle environment." Ex. 1003 ¶ 2. Kennewick discloses "a mobile interactive natural language speech system . . . that includes a speech unit." *Id.* ¶ 12. The system includes "a speech interface device that receives spoken natural language queries, commands and/or other utterances from a user, and a computer device or system that receives input from the speech unit[,] processes the input (e.g., retrieves information responsive to the query, takes action consistent with the command, [etc.]), and responds to the user with a natural language speech response." *Id.* ¶ 14.

Figure 5, reproduced below, illustrates an interactive natural language speech processing system. Ex. 1003 ¶ 92.

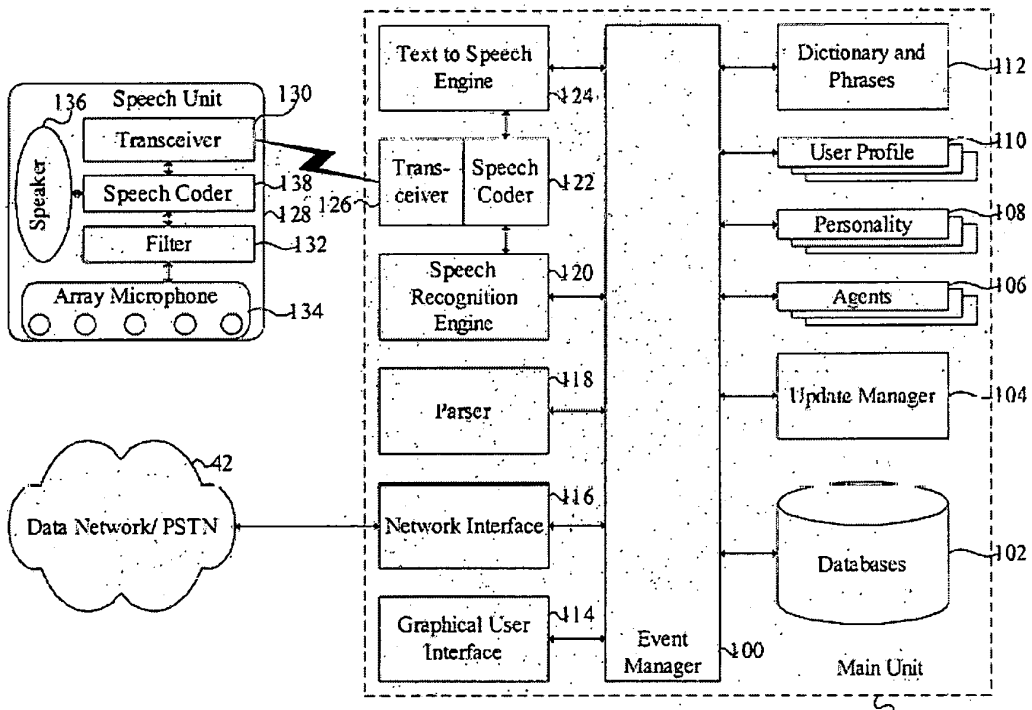


Figure 5. Speech Processing System Block Diagram

Figure 5 depicts components of an "interactive natural language speech processing system." Ex. 1003 ¶ 118. The main user interface for the system is speech unit 128. *Id.* ¶ 121. Speech unit 128 includes one or more microphones (e.g., array microphone 134) to receive the utterances of a user. *Id.* Speech unit 128 encodes the received speech via speech coder 138 and transmits the coded speech via transceiver module 130 to the main unit 98. *Id.* The coded speech is then passed to speech coder 122 for decoding. *Id.* ¶ 123. The decoded speech is then processed by the speech recognition engine 120 using data in the dictionary and phrases module 112 and data received from agents 106. *Id.* The recognized words and phrases are then "processed by parser 118, which transforms them into complete commands and questions using data supplied by the agents," where the agents "then process the commands or questions." *Id.* The agents are used to determine context for the questions and commands. *Id.* ¶ 128.

2. Other References

Petitioner also cites Huang, Seneff, Coffman, and Mitsuyoshi as secondary references. *See* Pet. 4. For the reasons discussed below, the issues in dispute all turn on the teachings and suggestions of Kennewick. Accordingly, this Decision does not address the substance of the Huang, Seneff, Coffman, or Mitsuyoshi references.

E. Obviousness under 35 U.S.C. § 103(a)

1. Objective Evidence of Nonobviousness

Patent Owner does not rely on objective evidence of nonobviousness in this case. *See generally* Resp. (not alleging that any such objective evidence is present).

2. *Scope and Content of the Prior Art; Differences between Claimed Subject Matter and Prior Art; Obviousness*

a) Analysis

Step [1.a] of independent method claim 1 recites "receiving an utterance" that "includes one or more words that have different meanings in different contexts." Steps [1.b] through [1.d] recite "accumulating short-term shared knowledge about the current conversation" and "accumulating long-term shared knowledge about the user" and "determining an intended meaning for the utterance." Steps [1.e] and [1.f] additionally recite "identifying . . . a context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge" and "establishing the intended meaning within the identified context . . . to disambiguate an intent that the user had in speaking the one or more words that have different meanings in the different contexts." Ex. 1001, 19:37–61.

Independent claims 13 and 25 recite substantively similar limitations, but are directed to a non-transitory computer-readable medium and a system, respectively. *See* Ex. 1001, 21:6–36, 22:55–23:16; Pet. 25 ("Claim 13 is substantially similar to claim 1"), 26 ("Claim 25 is a system variant of claim 1"). The remaining challenged claims all depend from one of these three independent claims. Thus, all challenged claims require either the steps of, computer-executable instructions operable to, or a system configured to: receive an utterance including "one or more words that have different meanings in different contexts," identify a context associated with the utterance from short-term shared knowledge about the conversation and long-term shared knowledge about the user, and establish the intended meaning within the identified context to disambiguate the user's intent.

Addressing step [1.a] and the similar limitations in claims 13 and 25, Petitioner contends that Kennewick "discloses receiving utterances that 'include[] one or more words that have different meanings in different contexts.'" Pet. 16–17 (alteration in original); *see also id.* at 25–27 (regarding claims 13 and 25). Petitioner cites to two disclosures in Kennewick of receiving words that allegedly have different meanings in different contexts: the receipt of the word "temperature" described in paragraph 160, and the receipt of the phrase "flight one hundred and twenty too" described in paragraph 163. Pet. 16–17.

Regarding steps [1.b] through [1.d] and the similar limitations of claims 13 and 25, Petitioner contends that Kennewick's dialog history (including tags) constitutes "short-term shared knowledge about the current conversation," and that Kennewick's user profiles constitute "long-term shared knowledge about the user." Pet. 18–19; 25–27.

With respect to limitations [1.e] through [1.f] and the similar limitations of claims 13 and 25, Petitioner contends that Kennewick's system "determine(s) a context for an utterance by applying prior probabilities or fuzzy possibilities to keyword matching, user profile 110, dialog history, and context stack contents." Pet. 20 (citing Ex. 1003 ¶ 160). Thus, according to Petitioner, context is determined at least in part based on short-term shared knowledge (i.e., dialog history) and long-term shared knowledge (i.e., user profile 110). *See id.* at 19–20. Petitioner further contends that Kennewick's system could, for example, "use a weather context to understand that the user is requesting a forecast and not their body temperature," or determine based on "the flight reservation context" that there is a greater likelihood that the word "too" was intended to mean "two." *Id.* at 20–21.

Patent Owner argues in response that Kennewick does not teach or suggest the specific type of ambiguity required by the challenged claims: "ambiguity where 'one or more words . . . ***have different meanings in different contexts***.'" Resp. 4. According to Patent Owner, Petitioner cites "a 'temperature' example that does not have two different meanings, and a 'flight' example that relates to determining which of two different words that sound the same is the correct word ***in a single context***." *Id.* at 4–5. On this record, and having reviewed the evidence and arguments put forth by both parties, we agree with Patent Owner's arguments.

Though neither party has proposed a construction of "one or more words that have different meanings in different contexts," the '681 patent specification provides examples of this type of ambiguity. For example, the '681 patent discloses that the word "traffic" could have different meanings in different contexts. Ex. 1001, 11:10–20. In particular, a media player user might intend the word traffic to refer to "a Rock and Roll band," a user searching for films might intend the word traffic to refer to "a film directed by Steven Soderbergh," and a user seeking navigation directions might intend the word traffic to refer to "conditions on roads along a route." *Id.* The '681 patent also discloses that the word "Portland" has different meanings in different "environmental" (i.e., geographic) contexts because it can refer to a city in Oregon or a city in Maine. *See id.* at 11:62–12:13. For example, the '681 patent discloses that a user may ask "What's the weather in Seattle," causing the system to establish weather as a context and Seattle as an environmental context. *See id.* If the same user then utters "and Portland?" the system may use the environmental context of Seattle to determine that the user intended to ask about the weather in Portland,

Oregon (based on the environmental proximity of the two cities) rather than Portland, Maine. *See id.*

In both of these examples from the specification, the ambiguous word itself has different potential meanings in different contexts. The word "traffic" can mean a band, a movie, or road conditions, depending on context. The word "Portland" can mean a city in Oregon, or an entirely different city in Maine, depending on environmental context. Nothing in the '681 patent suggests that the patentee intended for other types of ambiguity to fall within the scope of the challenged claims.

Petitioner argues in its Reply that we should interpret the claim limitation "one or more words that have different meanings in different contexts" more broadly than discussed above, so as to encompass other types of ambiguity. *See* Reply 2–6. Specifically, Petitioner argues that the word "Portland" in the '681 patent specification does not have different meanings in different contexts because it always refers to a city. *Id.* at 5. Petitioner then reasons that a POSITA would have been led by this example to interpret the claim limitation "words that have different meanings in different contexts" to include "words like 'Portland,' which clearly refers to a city but, absent context, would still be ambiguous." *Id.* This argument is not persuasive because it misinterprets the disclosure of the challenged patent. As discussed above, the '681 patent discloses that the word "Portland" has two separate possible meanings in two different environmental contexts: a city in Oregon, or an entirely different city in Maine. Petitioner's own expert admits that the word "Portland" has different meanings in different contexts. *See* Ex. 1033 ¶ 6 ("In other words, the word 'Portland' in the utterance 'and Portland?' has a different meaning depending on the context determined for the dialog which affects how the

system responds."). Thus, we are not persuaded by Petitioner's argument that a POSITA would interpret the claim limitation "one or more words that have different meanings in different contexts" to encompass types of ambiguity not resulting from a word or words that have different meanings in different contexts.

We now turn back to Petitioner's contentions regarding Kennewick. We begin with Petitioner's contention that paragraph 160 of Kennewick teaches or suggests receiving an utterance including the word "temperature," and that this word can have different meanings in different contexts, as required by limitation [1.a] and similar limitations of claims 13 and 25. The relevant portion of paragraph 160 is reproduced below.

> The parser may determine a context for an utterance by applying prior probabilities or fuzzy possibilities to keyword matching, user profile 110, dialog history, and context stack contents. The context of a question or command may determine the domain and thereby, the domain agent 156, if any, to be invoked. For example, a question with the keyword[] "temperature" implies a context value of weather for the question. Within a different dialog, the keyword "temperature" can imply a context for a measurement. The parser dynamically receives keyword and associated prior probability or fuzzy possibility updates from the system agent 150 or an already active domain agent 156. Based on these probabilities or possibilities the possible contexts are scored and the top one or few are used for further processing.

Ex. 1003 ¶ 160 (cmphasis addcd).

As Patent Owner points out, and as we observed in our Institution Decision, the word "temperature" does not have different meanings when used to describe weather or a measurement. *See* Resp. 4; Paper 7, 14. In both example situations, the word "temperature" has the same meaning—a measurement of warmth or coolness. *See* Ex. 2001 ¶ 26. Paragraph 160, thus, does not describe an ambiguity resulting from the word "temperature"

having different meanings in different contexts. The ambiguity results from not knowing if the user wants to know the temperature of the outside air (i.e., the weather) or the temperature of something else. *See id.*

Dr. Smyth's declaration testimony about words having different meanings in different contexts is unpersuasive. Dr. Smyth testifies that the keyword "'temperature' could be interpreted as an outdoor temperature or as a body temperature depending on context." *See* Ex. 1002 ¶ 69. This testimony is not persuasive because the word "temperature" refers to a measurement of warmth or coolness in both examples. That is, "temperature" has the same meaning in both contexts. Moreover, as Dr. Gershman points out, Kennewick does not teach or suggest a body temperature. Ex. 2001 ¶ 26. Dr. Smyth's testimony is unpersuasive for this additional reason. *See* Ex. 1002 ¶ 69.

Petitioner argues in its Reply that Kennewick describes different types of "temperature information, e.g., temperature measured by a sensor," or temperature from "a weather website." Reply 6–7. Thus, Petitioner reasons, merely "recognizing that temperature is a 'measurement of warmth or coolness' is not enough information to 'disambiguate an intent that the user had . . . .'" *Id.* at 7. But the fact that additional information may be necessary to "disambiguate an intent" does not demonstrate that the ambiguity is the result of the word "temperature" having different meanings in different contexts. As Petitioner concedes, the word "temperature" means a "measurement of warmth or coolness" in all of these examples. *See id.* Thus, we are not persuaded that this portion of Kennewick teaches or suggests an ambiguity resulting from the word "temperature" having different meanings in different contexts.

We now address Petitioner's alternative argument that the phrase "what about flight one hundred and twenty too" in paragraph 163 of Kennewick teaches or suggests receiving an utterance that includes "one or more words that have different meanings in different contexts." The relevant portion of paragraph 163 is reproduced below.

> In another example, the user asks, "what about flight one hundred and twenty too?" The parser and domain agent use flight information in the database and network information along with context to determine the most plausible interpretation among; flight 100 and flight 20 also, flight 100 and flight 22, flight 122, etc.

Ex. 1003 ¶ 163.

As Patent Owner points out, and as we observed in our Institution Decision, the ambiguity in the term "flight one hundred and twenty too" is an ambiguity that exists within a single context, and is not the result of the term having different meanings in different contexts. Resp. 6; Paper 7, 14–15. In particular, we agree with Patent Owner that paragraph 163 of Kennewick describes an ambiguity that exists after Kennewick's system has already determined context. *See* Resp. 7 (citing Ex. 2001 ¶ 30). The preceding paragraph of Kennewick discloses that "[o]nce the context for the question or command has been determined, the parser 118 can invoke the correct agent." Ex. 1003 ¶ 162. Reading paragraph 163 in the context of the preceding paragraph, it is apparent that when the user asks "what about flight one hundred and twenty too?" Kennewick's system has already identified the context and selected the appropriate domain agent to respond to the user's request. *See id.* ¶¶ 162–163; Resp. 7; Ex. 2001 ¶¶ 29–30. Thus, we agree with Patent Owner that a POSITA would not interpret paragraph 163 of Kennewick to teach or suggest an ambiguity resulting from

the phrase "what about flight one hundred and twenty too?" having different meanings in different contexts.

Petitioner argues that "[a] context in which 'flight 122' exists would result in a different meaning than a context where the system just listed several flight options including 'flight 100' and 'flight 20.'" Pet. 16–17 (citing Ex. 1002 ¶ 70). We agree with Patent Owner that this argument is unpersuasive because Dr. Smyth does not provide any support for his assertion that these hypothetical situations would constitute different contexts. *See* Ex. 1002 ¶ 70; Ex. 2001 ¶ 29. Moreover, limitation 1[e] and the corresponding limitations in claims 13 and 25 require identifying a "context associated with the utterance from the short-term shared knowledge [which Petitioner contends is Kennewick's dialog history] and the long-term shared knowledge [which Petitioner contends is Kennewick's user profiles]." Ex. 1001, 19:52–55, 21:22–25, 23:5–7; Pet. 18–19; 25–27. A system of the type described by Dr. Smyth, which determines context by checking whether a specific flight exists, would not determine context based on what Petitioner has identified as Kennewick's short-term and long-term shared knowledge. Thus, even if Dr. Smyth were correct, Kennewick would still not teach or suggest identifying a context in the manner required by claims 1, 13, or 25.

In the Reply, Petitioner, supported by Dr. Smyth, attempts to tweak its argument, positing that a user described in paragraph 163 of Kennewick might have previously reserved a seat on "flight 122," that this prior search information might have been stored in Kennewick's dialog history, and that Kennewick's system might use this dialog history to determine whether or not flight 122 does, in fact, exist. *See* Reply 9–10; Ex. 1033 ¶ 15. But this argument, and Dr. Smyth's accompanying testimony, are unsupported

speculation. Petitioner and Dr. Smyth do not identify anything in Kennewick that teaches, suggests, or even implies, that the user described in paragraph 163 previously conducted such a search for "flight 122," that information reflecting the search was stored in the user's dialog history or user profile, or that this information was then used to disambiguate the term "flight one hundred and twenty too." *See id.*

b) Legal Conclusion regarding Obviousness

On this record, and having considered the evidence and arguments put forth by both parties, we are not persuaded that Kennewick teaches or suggests that the word "temperature" has different meanings in different contexts, as required by independent claims 1, 13, and 25. *See* Resp. 4–6; Ex. 2001 ¶¶ 23–27. As detailed above, Patent Owner's arguments and the supporting testimony of Dr. Gershman are consistent with the ordinary and customary language of the claims, the specification of the '681 patent, and the disclosure of Kennewick. *See id.* In contrast, Petitioner's arguments are inconsistent with the ordinary and customary meaning of the claim language and the specification of the '681 patent, and are not adequately supported by the disclosure of Kennewick.

On this record, and having considered the evidence and arguments put forth by both parties, we also are not persuaded that Kennewick teaches or suggests that the phrase "what about flight one hundred and twenty too?" has different meanings in different contexts, as required by independent claims 1, 13, and 25. *See* Resp. 6–7; Ex. 2001 ¶¶ 28–30. Moreover, even if the "flight" example did describe two contexts (one in which "flight 122" exists, and one in which it does not), Petitioner's argument would still fail because Petitioner has offered no persuasive evidence that Kennewick teaches or suggests identifying one of these two alleged contexts based on

Kennewick's short-term shared information (dialog history) and long-term shared information (user profiles) as required by independent claims 1, 13, and 25. For the reasons explained above, Patent Owner's arguments and the supporting testimony of Dr. Gershman are consistent with and supported by the claims and specification of the '681 patent and the disclosure of Kennewick. *See* Resp. 4–6; Ex. 2001 ¶¶ 23–27. In contrast, Petitioner's arguments are not consistent with or supported by the ordinary and customary meaning of the claim language, the specification of the '681 patent, or the disclosure of Kennewick.

All remaining challenged claims depend from either claim 1, claim 13, or claim 25, and Petitioner's arguments regarding the challenged claims and the six asserted grounds of patentability all depend on Petitioner's contentions that the "temperature" and "flight" examples in Kennewick teach or suggest the claim limitations discussed above in independent claims 1, 13, and 25. *See* Pet. 15–23, 32, 41, 50, 53, 58. Accordingly, we find on this record that Petitioner has not shown by a preponderance of the evidence that any claim of the '681 patent is unpatentable under 35 U.S.C. § 103(a).

III. CONCLUSION

For the foregoing reasons, Petitioner has not shown by a preponderance of the evidence that any of challenged claims 1–36 is unpatentable.

In summary:

Claim(s) Challenged	35 U.S.C. §	Reference(s)	Claims Shown Unpatentable	Claims Not Shown Unpatentable
1, 2, 9, 13, 14, 21, 25, 26, 33	103(a)	Kennewick		1, 2, 9, 13, 14, 21, 25, 26, 33
5, 8, 17, 20, 29, 32	103(a)	Kennewick, Huang		5, 8, 17, 20, 29, 32
10, 22, 34	103(a)	Kennewick, Seneff		10, 22, 34
6, 7, 18, 19, 30, 31	103(a)	Kennewick, Huang, Seneff		6, 7, 18, 19, 30, 31
11, 12, 23, 24, 35, 36	103(a)	Kennewick, Coffman		11, 12, 23, 24, 35, 36
3, 4, 15, 16, 27, 28	103(a)	Kennewick, Mitsuyoshi		3, 4, 15, 16, 27, 28
Overall Outcome				1 36

IV. ORDER

In consideration of the foregoing, it is hereby:

ORDERED that claims 1–36 of U.S. Patent 8,073,681 B2 have not been shown to be unpatentable; and

FURTHER ORDERED that because this is a final written decision, parties to the proceeding seeking judicial review of this decision must comply with the notice and service requirements of 37 C.F.R. § 90.2.

FOR PETITIONER:

J. David Hadden
Saina Shamilov
Allen Wang
Brian Hoffman
FENWICK & WEST LLP
dhadden-ptab@fenwick.com
sshamilov-ptab@fenwick.com
allen.wang@fenwick.com
bhoffman@fenwick.com

FOR PATENT OWNER:

Matthew A. Argenti
Michael T. Rosato
James C. Yoon
Ryan Smith
WILSON SONSINI GOODRICH & ROSATI
margenti@wsgr.com
mrosato@wsgr.com
jyoon@wsgr.com
rsmith@wsgr.com

Disclaimer

8,073,681 B2 - Larry Baldwin, Maple Valley, WA (US); Tom Freeman, Mercer Island, WA (US); Michael Tjalve, Bellevue, WA (US); Blanc Ebersold, Seattle, WA (US); Chris Weider, Everett, WA (US). SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE. Patent dated December 6, 2011. Disclaimer filed May 17, 2021, by the assignee, VB Assets, LLC.

I hereby disclaim the following complete claims 37-42 of said patent.

(Official Gazette, November 22, 2022)

(12) **INTER PARTES REVIEW CERTIFICATE** (2941st)
United States Patent
Baldwin et al.

(10) **Number: US 8,073,681 K1**
(45) **Certificate Issued: Dec. 21, 2022**

(54) **SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE**

(75) Inventors: **Larry Baldwin**; **Michael Tjalve**; **Blane Ebersold**; **Tom Freeman**; **Chris Weider**

(73) Assignee: **NUANCE COMMUNICATIONS, INC.**

Trial Number:

IPR2020-01367 filed Jul. 28, 2020

Inter Partes Review Certificate for:

Patent No.: **8,073,681**
Issued: **Dec. 6, 2011**
Appl. No.: **11/580,926**
Filed: **Oct. 16, 2006**

The results of IPR2020-01367 are reflected in this inter partes review certificate under 35 U.S.C. 318(b).

AS A RESULT OF THE INTER PARTES REVIEW PROCEEDING, IT HAS BEEN DETERMINED THAT:

Claims **1-36** are found patentable.

Claims **37-42** are disclaimed.

* * * * *