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.

Amazon/VB Assets
Exhibit 1003
Page 710

[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.

Amazon/VB Assets
Exhibit 1003
Page 711

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.

Amazon/VB Assets
Exhibit 1003
Page 712

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

Amazon/VB Assets
Exhibit 1003
Page 713

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.

Amazon/VB Assets
Exhibit 1003
Page 714

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.

Amazon/VB Assets
Exhibit 1003
Page 715

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.

Amazon/VB Assets
Exhibit 1003
Page 716

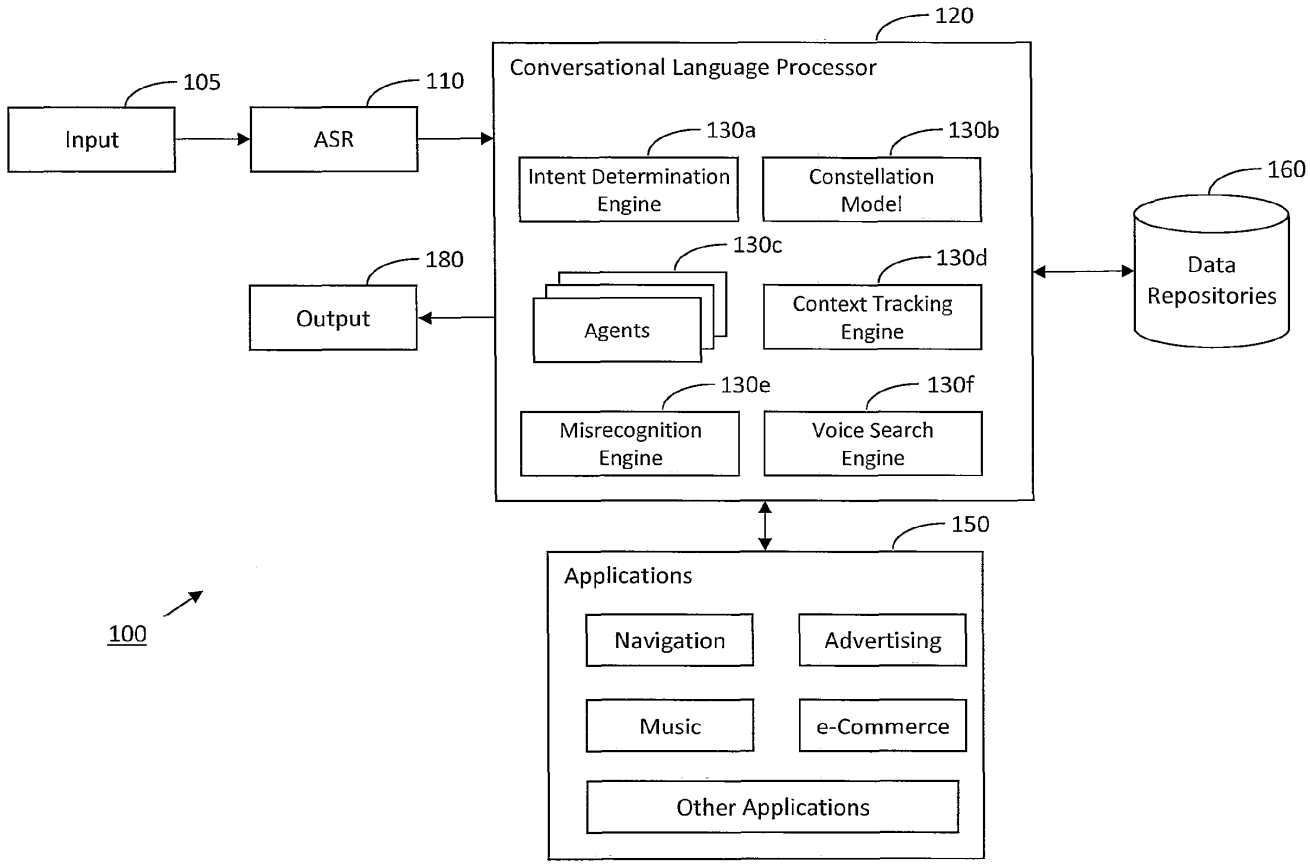


Figure 1

Amazon/VB Assets
Exhibit 1003
Page 717

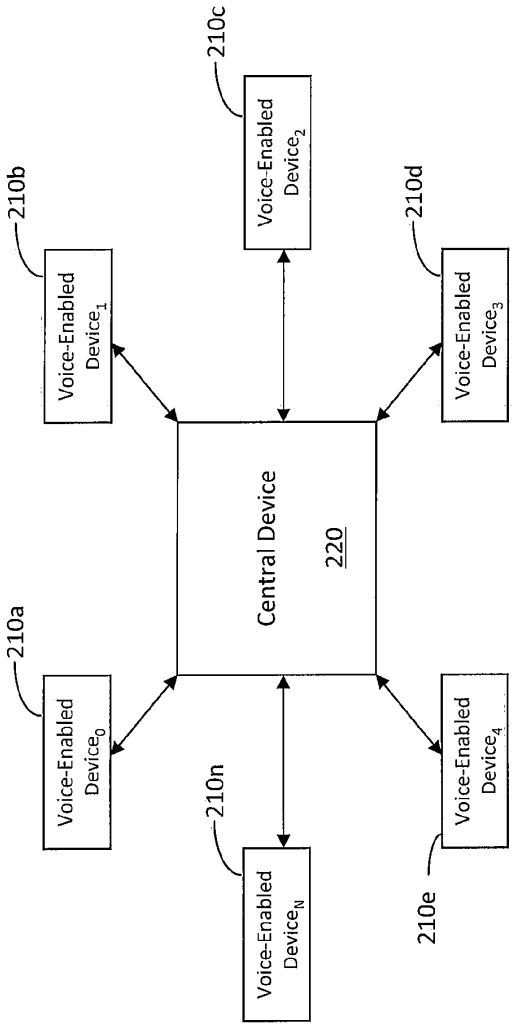

210c
Voice-Enabled Device$_2$
210b
Voice-Enabled Device$_1$
210d
Voice-Enabled Device$_3$
Central Device
220
210a
Voice-Enabled Device$_0$
210n
Voice-Enabled Device$_N$
Voice-Enabled Device$_4$
210e

Figure 2

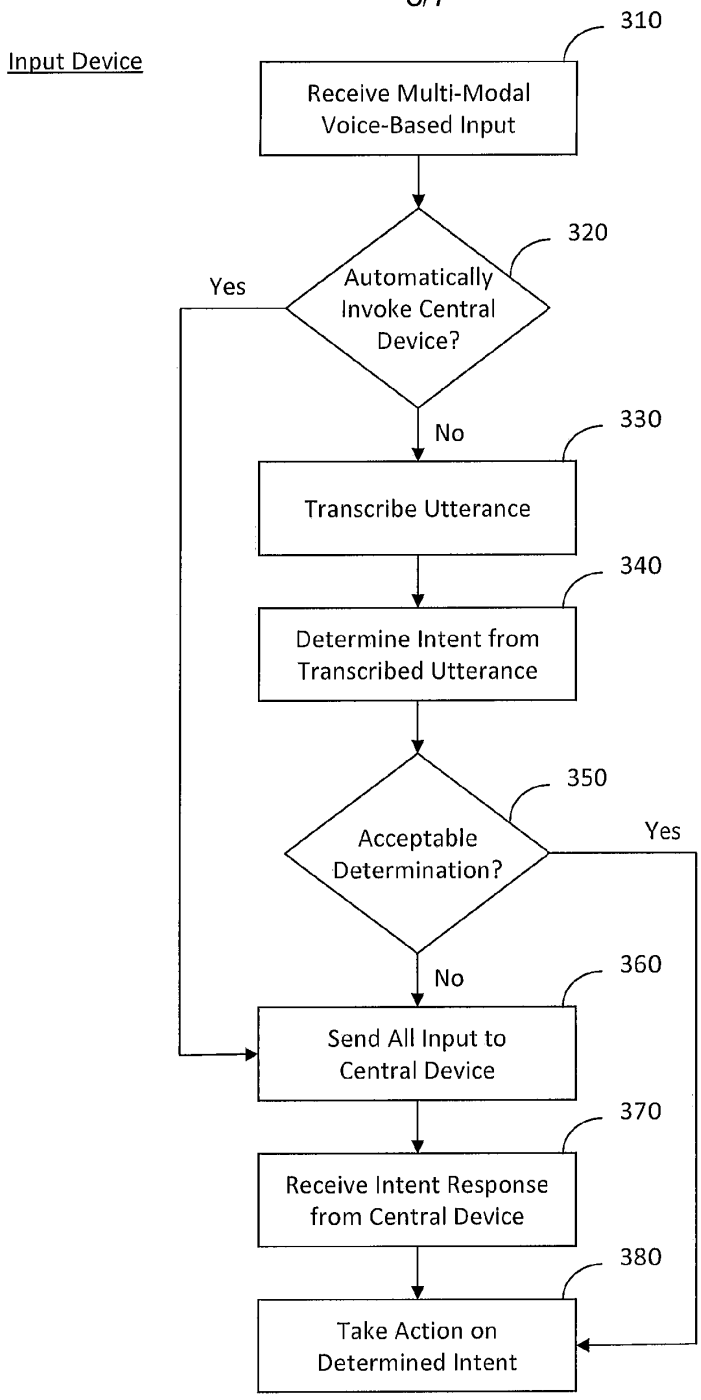

Input Device
310
Receive Multi-Modal Voice-Based Input
320
Automatically Invoke Central Device?
Yes
No
330
Transcribe Utterance
340
Determine Intent from Transcribed Utterance
350
Acceptable Determination?
Yes
No
360
Send All Input to Central Device
370
Receive Intent Response from Central Device
380
Take Action on Determined Intent

Figure 3

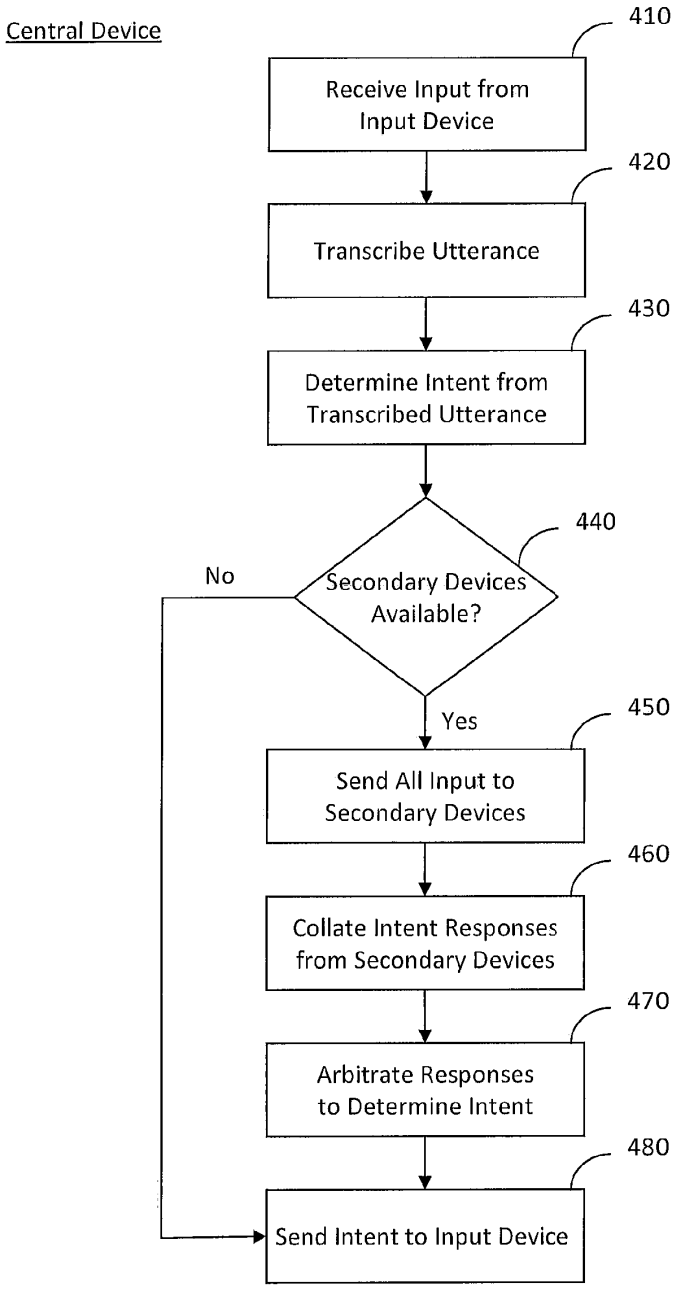

Central Device
410
Receive Input from Input Device
420
Transcribe Utterance
430
Determine Intent from Transcribed Utterance
440
Secondary Devices Available?
No
Yes
450
Send All Input to Secondary Devices
460
Collate Intent Responses from Secondary Devices
470
Arbitrate Responses to Determine Intent
480
Send Intent to Input Device

Figure 4

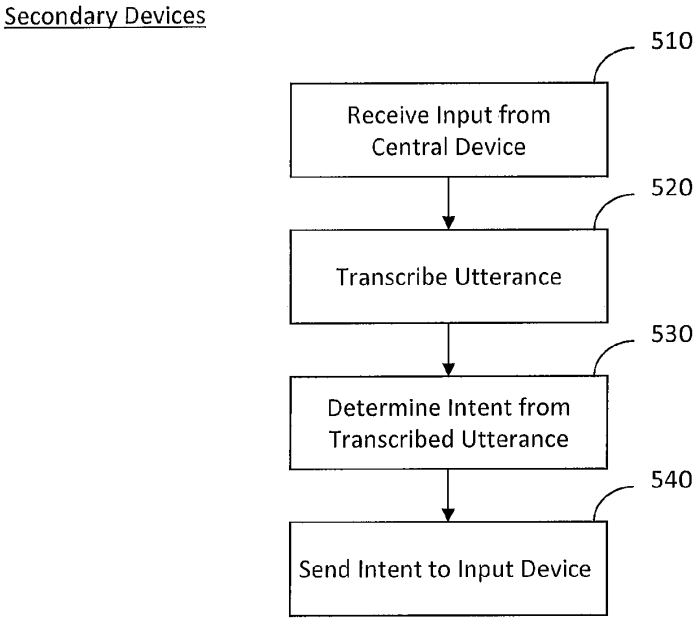

Secondary Devices
510
Receive Input from Central Device
520
Transcribe Utterance
530
Determine Intent from Transcribed Utterance
540
Send Intent to Input Device

Figure 5

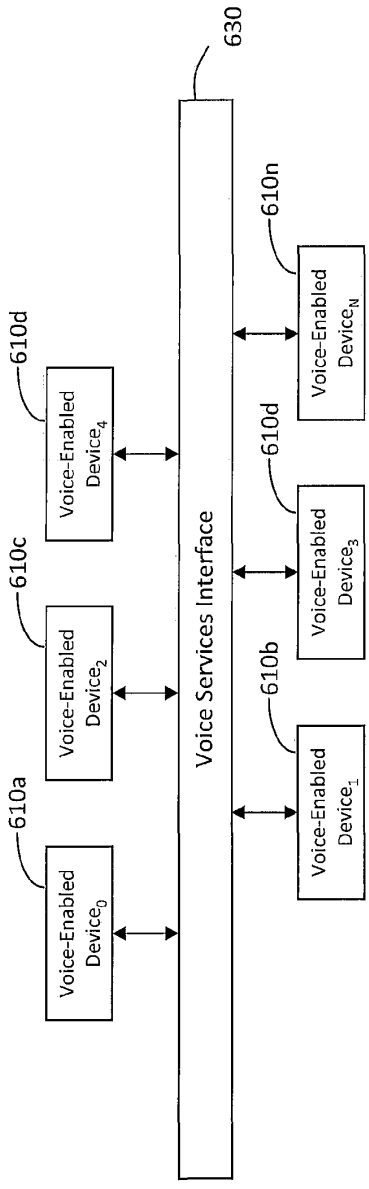

630
Voice Services Interface
610a
Voice-Enabled Device$_0$
610b
Voice-Enabled Device$_1$
610c
Voice-Enabled Device$_2$
610d
Voice-Enabled Device$_3$
610d
Voice-Enabled Device$_4$
610n
Voice-Enabled Device$_N$

Figure 6

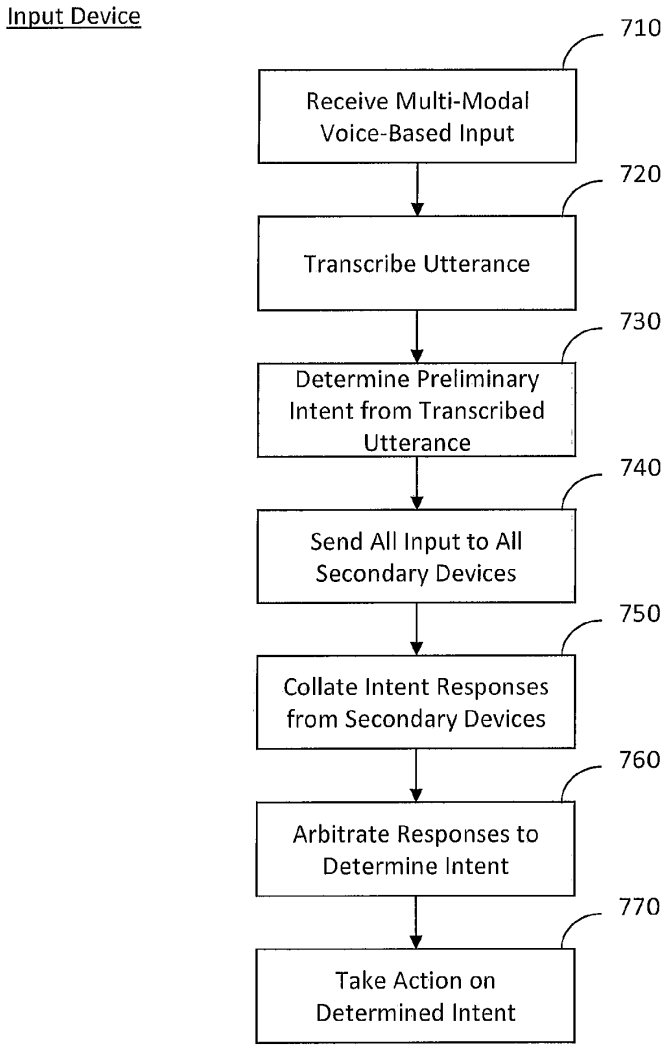

Input Device
710
Receive Multi-Modal Voice-Based Input
720
Transcribe Utterance
730
Determine Preliminary Intent from Transcribed Utterance
740
Send All Input to All Secondary Devices
750
Collate Intent Responses from Secondary Devices
760
Arbitrate Responses to Determine Intent
770
Take Action on Determined 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

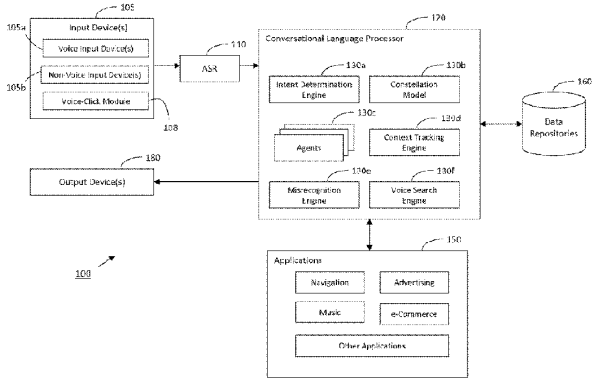


Figure 1

(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

Amazon/VB Assets
Exhibit 1003
Page 727

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

Amazon/VB Assets
Exhibit 1003
Page 728

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.

Amazon/VB Assets
Exhibit 1003
Page 729

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

Amazon/VB Assets
Exhibit 1003
Page 730

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

Amazon/VB Assets
Exhibit 1003
Page 731

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

Amazon/VB Assets
Exhibit 1003
Page 734

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.

Amazon/VB Assets
Exhibit 1003
Page 736

[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

Amazon/VB Assets
Exhibit 1003
Page 737

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

Amazon/VB Assets
Exhibit 1003
Page 738

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

Amazon/VB Assets
Exhibit 1003
Page 739

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

Amazon/VB Assets
Exhibit 1003
Page 740

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-

Amazon/VB Assets
Exhibit 1003
Page 741

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

Amazon/VB Assets
Exhibit 1003
Page 742

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

Amazon/VB Assets
Exhibit 1003
Page 743

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

Amazon/VB Assets
Exhibit 1003
Page 744

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

Amazon/VB Assets
Exhibit 1003
Page 745

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

Amazon/VB Assets
Exhibit 1003
Page 746

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.

Amazon/VB Assets
Exhibit 1003
Page 748

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.

Amazon/VB Assets
Exhibit 1003
Page 749

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;

Amazon/VB Assets
Exhibit 1003
Page 750

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.

Amazon/VB Assets
Exhibit 1003
Page 751

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.

Amazon/VB Assets
Exhibit 1003
Page 752

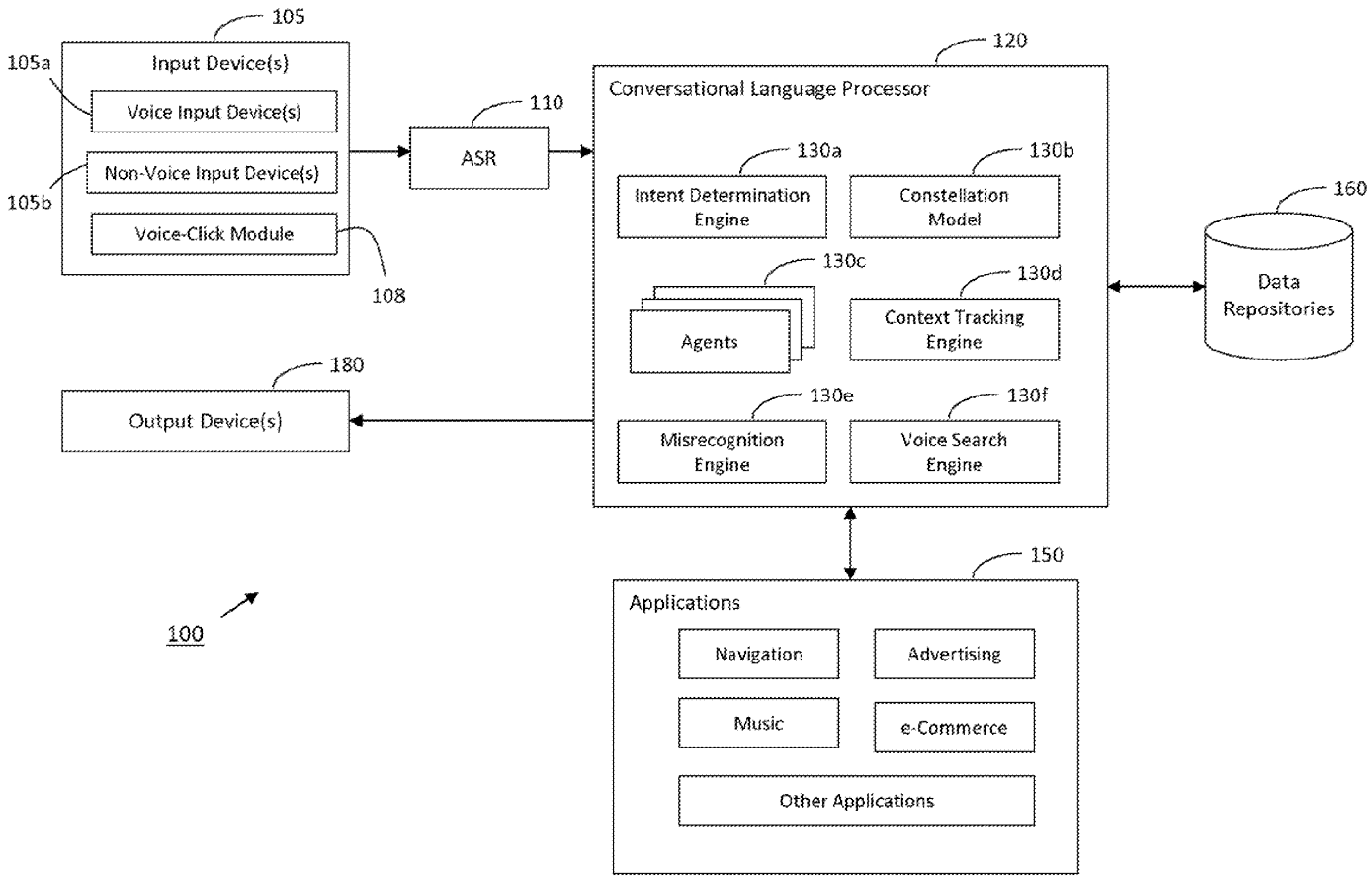

105
Input Device(s)
105a
Voice Input Device(s)
Non-Voice Input Device(s)
105b
Voice-Click Module
108
110
ASR
120
Conversational Language Processor
130a
Intent Determination Engine
130b
Constellation Model
130c
Agents
130d
Context Tracking Engine
130e
Misrecognition Engine
130f
Voice Search Engine
160
Data Repositories
180
Output Device(s)
150
Applications
Navigation
Advertising
Music
e-Commerce
Other Applications
100

Figure 1

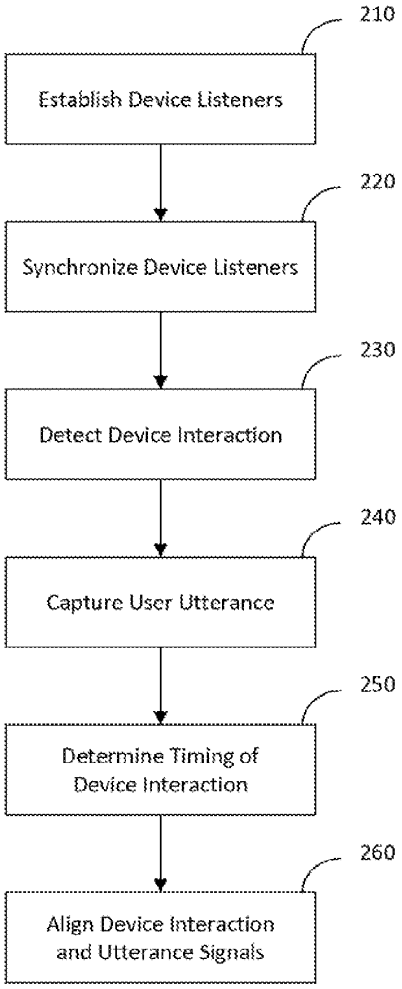

210
Establish Device Listeners
220
Synchronize Device Listeners
230
Detect Device Interaction
240
Capture User Utterance
250
Determine Timing of Device Interaction
260
Align Device Interaction and Utterance Signals

Figure 2

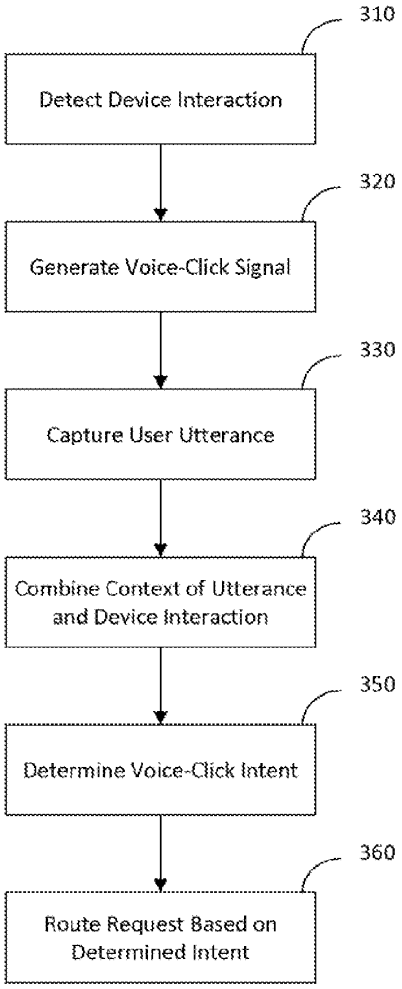

310
Detect Device Interaction
320
Generate Voice-Click Signal
330
Capture User Utterance
340
Combine Context of Utterance and Device Interaction
350
Determine Voice-Click Intent
360
Route Request Based on Determined Intent

Figure 3

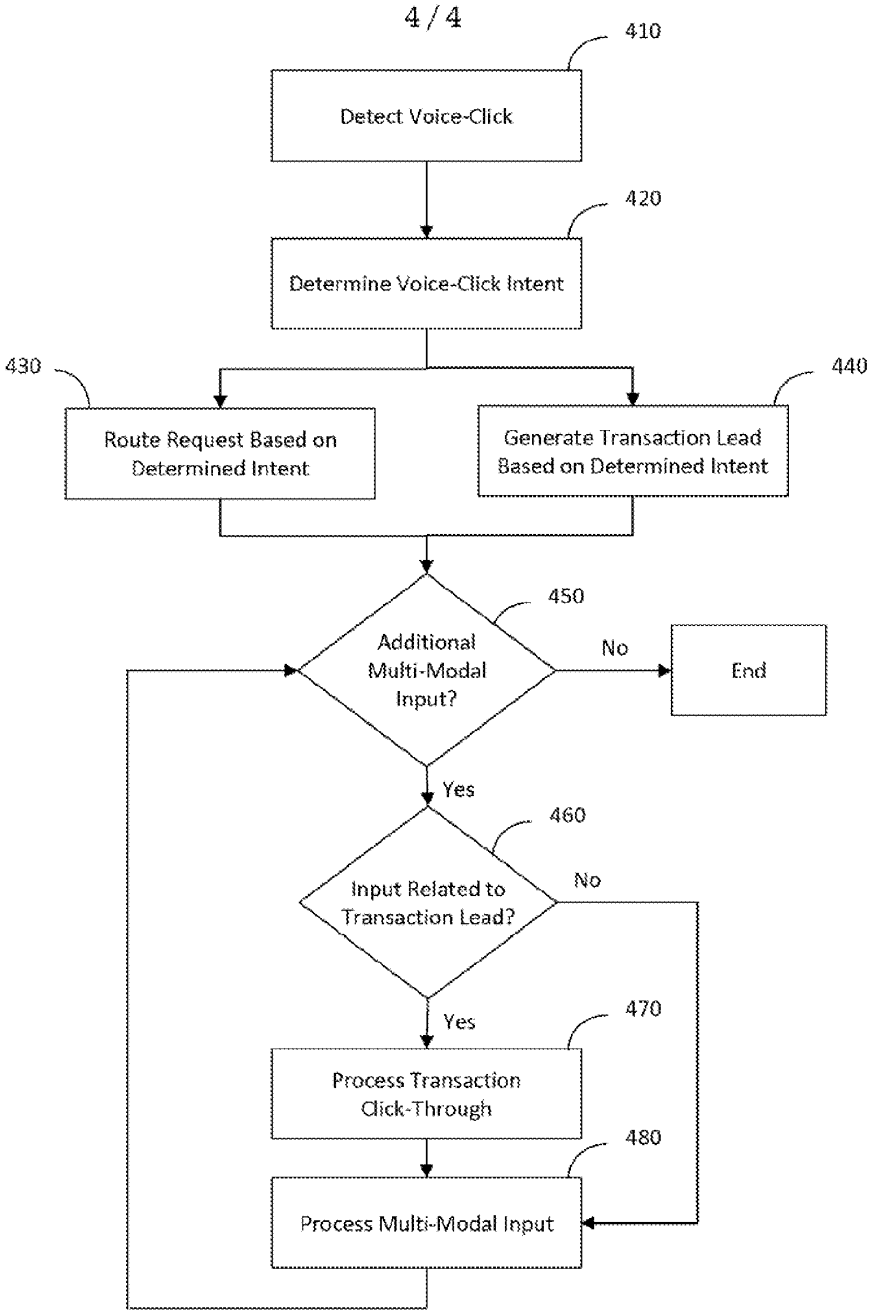

410
Detect Voice-Click
420
Determine Voice-Click Intent
430
Route Request Based on Determined Intent
440
Generate Transaction Lead Based on Determined Intent
450
Additional Multi-Modal Input?
No
End
Yes
460
Input Related to Transaction Lead?
No
Yes
470
Process Transaction Click-Through
480
Process Multi-Modal 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 Acknowledgement Receipt

EFS ID:	8972272
Application Number:	12943699
International Application Number:	
Confirmation Number:	4185
Title of Invention:	SYSTEM AND METHOD FOR PROVIDING A NATURAL LANGUAGE CONTENT DEDICATION SERVICE
First Named Inventor/Applicant Name:	Mike Kennewick
Customer Number:	00909
Filer:	Syed Jafar Ali/Susan Trader
Filer Authorized By:	Syed Jafar Ali
Attorney Docket Number:	090611-0384422
Receipt Date:	06-DEC-2010
Filing Date:	10-NOV-2010
Time Stamp:	14:37:39
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		0384422_SIDS.pdf	97781 b9f52041ee02bf43635fa1b9b18003ae63df8812	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:					
2	Foreign Reference	WO2007027546.pdf	3873530 5cc6e95751b76c1f016b92415f2860dd2cc39298	no	74
Warnings:					
Information:					
3	Foreign Reference	WO2007027989.pdf	1432119 6c33baceb03630bcc7fc958ccf493ab880e7e9e3	no	28
Warnings:					
Information:					
4	Foreign Reference	WO2008098039.pdf	1103202 fcab991006a05016a57e05548f6b00f78732c946	no	24
Warnings:					
Information:					
5	Foreign Reference	WO2008118195.pdf	2058896 635503176929e0d666e3e8683aa40f3a9643658e	no	39
Warnings:					
Information:					
6	Foreign Reference	WO2009075912.pdf	3688456 02c65abf9ea5c876fb8e4a654567b258513f4e25	no	69
Warnings:					
Information:					
7	Foreign Reference	WO2009145796.pdf	1934285 e71f0a22dfc12ce94396146b078264c9e5df68d9	no	39
Warnings:					
Information:					
8	Foreign Reference	WO2010096752.pdf	1502381 f4c719245f739ab6223fbd8f94c7b17cd9acf7b7	no	33
Warnings:					
Information:					
		Total Files Size (in bytes):	15690650		

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 :	Mike KENNEWICK *et al.*	CONFIRMATION NO.:	4185
SERIAL NUMBER :	12/943,699	EXAMINER:	*Unassigned*
FILING DATE :	November 10, 2010	ART UNIT:	2447
FOR :	SYSTEM AND METHOD FOR PROVIDING A NATURAL LANGUAGE CONTENT DEDICATION SERVICE		

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

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

Customer No.
00909

Dated: December 6, 2010

PATENT APPLICATION FEE DETERMINATION RECORD

Substitute for Form PTO-875

Application or Docket Number: 12/943,699

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	82		N/A	
SEARCH FEE (37 CFR 1.16(k), (i), or (m))	N/A	N/A	N/A	270		N/A	
EXAMINATION FEE (37 CFR 1.16(o), (p), or (q))	N/A	N/A	N/A	110		N/A	
TOTAL CLAIMS (37 CFR 1.16(i))	20 minus 20 =	*	x 26 =	0.00	OR		
INDEPENDENT CLAIMS (37 CFR 1.16(h))	2 minus 3 =	*	x 110 =	0.00			
APPLICATION SIZE FEE (37 CFR 1.16(s))	If the specification and drawings exceed 100 sheets of paper, the application size fee due is $270 ($135 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).			0.00			
MULTIPLE DEPENDENT CLAIM PRESENT (37 CFR 1.16(j))				0.00			
* If the difference in column 1 is less than zero, enter "0" in column 2.			TOTAL	462		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 =		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	

AMENDMENT B

	(Column 1) CLAIMS REMAINING AFTER AMENDMENT		(Column 2) HIGHEST NUMBER PREVIOUSLY PAID FOR	(Column 3) PRESENT EXTRA	RATE($)	ADDITIONAL FEE($)		RATE($)	ADDITIONAL FEE($)
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 found in the appropriate box in column 1.

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
12/943,699	11/10/2010	Mike Kennewick	090611-0384422

CONFIRMATION NO. 4185

909
PILLSBURY WINTHROP SHAW PITTMAN, LLP
P.O. BOX 10500
MCLEAN, VA 22102

FORMALITIES LETTER

OC000000044624812

Date Mailed: 11/29/2010

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 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:

- 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.

SUMMARY OF FEES DUE:

Total fee(s) required within **TWO MONTHS** from the date of this Notice is $**65** for a small entity
- $**65** Surcharge.

Amazon/VB Assets
Exhibit 1003
Page 764

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.

/atesfaye/

Office of Data Management, Application Assistance Unit (571) 272-4000, or (571) 272-4200, or 1-888-786-0101

Amazon/VB Assets
Exhibit 1003
Page 765

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	TOT CLAIMS	IND CLAIMS
12/943,699	11/10/2010	2447	462	090611-0384422	20	2

CONFIRMATION NO. 4185

909
PILLSBURY WINTHROP SHAW PITTMAN, LLP
P.O. BOX 10500
MCLEAN, VA 22102

FILING RECEIPT

OC000000044624811

Date Mailed: 11/29/2010

Receipt is acknowledged of this non-provisional patent application. The application will be taken up for examination in due course. Applicant will be notified as to the results of the examination. Any correspondence concerning the application must include the following identification information: the U.S. APPLICATION NUMBER, FILING DATE, NAME OF APPLICANT, and TITLE OF INVENTION. 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 submit a written request for a Filing Receipt Correction. 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**

Applicant(s)

Mike Kennewick, Bellevue, WA;
Lynn Elis Armstrong, Woodinville, WA;

Assignment For Published Patent Application

VoiceBox Technologies, Inc., Bellevue, WA

Power of Attorney: None

Domestic Priority data as claimed by applicant

This appln claims benefit of 61/259,820 11/10/2009

Foreign Applications

If Required, Foreign Filing License Granted: 11/22/2010

The country code and number of your priority application, to be used for filing abroad under the Paris Convention, is **US 12/943,699**

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

Non-Publication Request: No

Early Publication Request: No
** **SMALL ENTITY** **

Amazon/VB Assets
Exhibit 1003
Page 766

Title

SYSTEM AND METHOD FOR PROVIDING A NATURAL LANGUAGE CONTENT DEDICATION SERVICE

Preliminary Class

709

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).

PTO/SB/05 (08-08)
Approved for use through 06/30/2010. 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.	090611-0384422
First Inventor	MIKE KENNEWICK
Title	SYSTEM AND METHOD FOR PROVIDING A NATURAL LANGUAGE CONTENT DEDICATION SERVICE
Express Mail Label No.	

APPLICATION ELEMENTS

See *MPEP chapter 600 concerning utility patent application contents.*

1. ☐ **Fee Transmittal Form** (e.g., PTO/SB/17)
2. ☒ **Applicant claims small entity status.** See 37 CFR 1.27.
3. ☒ **Specification** [*Total Pages* 59]
   Both the claims and abstract must start on a new page
   (For information on the preferred arrangement, see MPEP 608.01(a))
4. ☒ **Drawing(s)** (35 U.S. C. *113)* [*Total Sheets* 8]
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. ☒ **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

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 citations attached
13. ☐ **Preliminary Amendment**
14. ☐ **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

☒ The address associated with Customer Number 00909 ***OR*** ☐ Correspondence address below

Name					
Address					
City		State		Zip Code	
Country		Telephone		Email	

Signature		Date	November 10, 2010
Name (Print/Type)	Syed Jafar Ali	Registration No. (Attorney/Agent)	58,780

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.

American LegalNet, Inc.
www.FormsWorkflow.com

Attorney Docket No. 090611-0384422 ***PATENT***

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of: Mike Kennewick, *et al.*

Application No.:	**New**	Confirmation No.:	Unassigned
Filed:	November 10, 2010	Group No.:	Unassigned
		Examiner:	Unassigned

For: SYSTEM AND METHOD FOR PROVIDING A NATURAL LANGUAGE CONTENT DEDICATION SERVICE

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

APPLICATION DATA SHEET
37 CFR § 1.76

BIBLIOGRAPHIC DATA

1. Applicant information

First Applicant: Mike Kennewick
Citizenship: The United States of America
Residence: 13407 37th Place, N.E., Bellevue, Washington 98005.

Second Applicant: Lynn Elise Armstrong
Citizenship: The United States of America
Residence: 19807 183rd Place, N.E., Woodinville, Washington 98077.

2. Correspondence information

Correspondence for this application should be addressed as follows:

Customer No.: 00909

3. Application information

Title of the Invention:	SYSTEM AND METHOD FOR PROVIDING A NATURAL LANGUAGE CONTENT DEDICATION SERVICE
Attorney Docket Number:	090611-0384422 Small Entity Status Claimed ☒
Total number of drawing sheets:	8
Type of application:	Utility

Amazon/VB Assets
Exhibit 1003
Page 770

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. Domestic Priority information.

Domestic priority for this application is claimed as follows:

35 U.S.C. § 119(e)

Application No.	Filed	Relationship
61/259,820	November 10, 2009	Priority

6. Assignee information

The assignee of this application is:

VoiceBox Technologies, Inc.
11980 N.E. 24th Street
Bellevue, Washington 98005.

Extent of interest of assignee in application: Entire right, title, and interest.

Date: November 10, 2010

PILLSBURY WINTHROP SHAW PITTMAN LLP
P.O. Box 10500
McLean, VA 22102
Main: (703) 770-7900
Fax: (703) 770-7901

Syed Jafar Ali
Registration No. 58,780
Direct Dial: (703) 770-7540

Customer No. 00909

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANTS :	Mike KENNEWICK *et al.*	CONFIRMATION NO.:	*Unassigned*
SERIAL NUMBER :	*Unassigned*	EXAMINER:	*Unassigned*
FILING DATE :	Herewith (November 10, 2010)	ART UNIT:	*Unassigned*
FOR :	SYSTEM AND METHOD FOR PROVIDING A NATURAL LANGUAGE CONTENT DEDICATION SERVICE		

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.

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.

Amazon/VB Assets
Exhibit 1003
Page 772

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).

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: November 10, 2010

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0384422	*Unassigned*
	Applicant(s)	
	Mike KENNEWICK *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	Herewith (November 10, 2010)	*Unassigned*

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	5,155,743	10/13/1992	Jacobs	375	28	
	1C	5,274,560	12/28/1993	LaRue	364	444	
	1D	5,377,350	12/27/1994	Skinner	395	600	
	1E	5,386,556	01/31/1995	Hedin et al.	395	600	
	1F	5,424,947	06/13/1995	Nagao et al.	364	419.08	
	1G	5,471,318	11/28/1995	Ahuja et al.	358	400	
	1H	5,475,733	12/12/1995	Eisdorfer et al.	379	52	
	1I	5,488,652	01/30/1996	Bielby *et al.*	379	88	
	1J	5,499,289	03/12/1996	Bruno et al.	379	220	
	1K	5,500,920	03/19/1996	Kupiec	395	2.79	
	1L	5,517,560	05/14/1996	Greenspan	379	114	
	1M	5,533,108	07/02/1996	Harris et al.	379	201	
	1N	5,537,436	07/16/1996	Bottoms et al.	375	222	
	1O	5,539,744	07/23/1996	Chu et al.	370	60	
	1P	5,557,667	09/17/1996	Bruno et al.	379	201	
	1Q	5,563,937	10/08/1996	Bruno et al.	379	201	
	1R	5,590,039	12/31/1996	Ikeda et al.	395	759	
	1S	5,617,407	04/01/1997	Bareis	369	275.3	
	1T	5,633,922	05/27/1997	August et al.	379	220	
	1U	5,675,629	10/07/1997	Raffel et al.	379	58	
	1V	5,696,965	12/09/1997	Dedrick	395	610	
	1W	5,708,422	01/13/1998	Blonder et al.	340	825.34	

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 01/78065	10/18/2001	International				
	1BB	WO 2004/072954	08/26/2001	International				

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

	1CC	REUTERS, "IBM to Enable Honda Drivers to Talk to Cars", Charles Schwab & Co., Inc., July 28, 2002, 1 page.
	1DD	LIN, Bor-shen, *et al.*, "A Distributed Architecture for Cooperative Spoken Dialogue Agents with Coherent Dialogue State and History", ASRU'99, 1999, 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.

402465373v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0384422	*Unassigned*
	Applicant(s)	
	Mike KENNEWICK *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	Herewith (November 10, 2010)	*Unassigned*

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	2A	5,721,938	02/24/1998	Stuckey	395	754	
	2B	5,722,084	02/24/1998	Chakrin et al.	455	551	
	2C	5,742,763	04/21/1998	Jones	395	200.3	
	2D	5,748,841	05/05/1998	Morin *et al.*	395	2.66	
	2E	5,752,052	05/12/1998	Richardson et al.	395	759	
	2F	5,754,784	05/19/1998	Garland et al.	395	200.49	
	2G	5,761,631	06/02/1998	Nasukawa	704	9	
	2H	5,774,859	06/30/1998	Houser *et al.*	704	275	
	2I	5,794,050	08/11/1998	Dahlgren et al.	395	708	
	2J	5,794,196	08/11/1998	Yegnanarayanan *et al.*	704	255	
	2K	5,797,112	08/18/1998	Komatsu *et al.*	701	201	
	2L	5,799,276	08/25/1998	Komissarchik *et al.*	704	251	
	2M	5,802,510	09/01/1998	Jones	707	2	
	2N	5,832,221	11/03/1998	Jones	375	200.36	
	2O	5,867,817	02/02/1999	Catallo *et al.*	704	255	
	2P	5,878,385	03/02/1999	Bralich *et al.*	704	9	
	2Q	5,878,386	03/02/1999	Coughlin	704	10	
	2R	5,892,813	04/06/1999	Morin *et al.*	379	88.01	
	2S	5,895,464	04/20/1999	Bhandari *et al.*	707	3	
	2T	5,895,466	04/20/1999	Goldberg et al.	707	5	
	2U	5,902,347	05/11/1999	Backman *et al.*	701	200	
	2V	5,911,120	06/08/1999	Jarett et al.	455	417	
	2W	5,918,222	06/29/1999	Fukui *et al.*	707	1	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO
	2X	WO 2007/019318	02/15/2007	International				
	2Y	WO 2007/021587	02/22/2007	International				

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

	2Z	KUHN, Thomas, *et al.*, "Hybrid In-Car Speech Recognition for Mobile Multimedia Applications", Vehicular Technology Conference, IEEE, July 1999, pages 2009-2013.
	2AA	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.

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.

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0384422	*Unassigned*
	Applicant(s)	
	Mike KENNEWICK *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	Herewith (November 10, 2010)	*Unassigned*

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	3A	5,926,784	07/20/1999	Richardson et al.	704	9	
	3B	5,933,822	08/03/1999	Braden-Harder et al.	707	5	
	3C	5,953,393	09/14/1999	Culbreth et al.	379	88.25	
	3D	5,963,894	10/05/1999	Richardson et al.	704	9	
	3E	5,963,940	10/05/1999	Liddy et al.	707	5	
	3F	5,987,404	11/16/1999	Della Pietra et al.	704	9	
	3G	5,991,721	11/23/1999	Asano et al.	704	257	
	3H	5,995,119	11/30/1999	Cosatto et al.	345	473	
	3I	6,009,382	12/28/1999	Martino et al.	704	1	
	3J	6,014,559	01/11/2000	Amin	455	413	
	3K	6,021,384	02/01/2000	Gorin et al.	704	1	
	3L	6,044,347	03/28/2000	Abella et al.	704	272	
	3M	6,049,602	04/11/2000	Foladare et al.	379	265	
	3N	6,058,187	05/02/2000	Chen	380	21	
	3O	6,078,886	06/20/2000	Dragosh et al.	704	270	
	3P	6,081,774	06/27/2000	De Hita et al.	704	9	
	3Q	6,101,241	08/08/2000	Boyce et al.	379	88.01	
	3R	6,119,087	09/12/2000	Kuhn et al.	704	270	
	3S	6,134,235	10/17/2000	Goldman et al.	370	352	
	3T	6,144,667	11/07/2000	Doshi et al.	370	401	
	3U	6,144,938	11/07/2000	Surace *et al.*	704	257	
	3V	6,154,526	11/28/2000	Dahlke *et al.*	379	88.03	
	3W	6,160,883	12/12/2000	Jackson et al.	379	230	

FOREIGN PATENT DOCUMENTS

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

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

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

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.

402465373v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0384422	*Unassigned*
	Applicant(s)	
	Mike KENNEWICK *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	Herewith (November 10, 2010)	*Unassigned*

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	4A	6,173,279	01/09/2001	Levin et al.	707	5	
	4B	6,175,858	01/16/2001	Bulfer et al.	709	206	
	4C	6,185,535	02/06/2001	Hedin *et al.*	704	270	
	4D	6,192,110	02/20/2001	Abella et al.	379	88.01	
	4E	6,192,338	02/20/2001	Haszto et al.	704	257	
	4F	6,195,634	02/27/2001	Dudemaine *et al.*	704	231	
	4G	6,195,651	02/27/2001	Handel *et al.*	707	2	
	4H	6,208,972	03/27/2001	Grant *et al.*	704	275	
	4I	6,219,346	04/17/2001	Maxemchuk	370	338	
	4J	6,219,643	04/17/2001	Cohen et al.	704	257	
	4K	6,226,612	05/01/2001	Srenger *et al.*	704	256	
	4L	6,233,556	05/15/2001	Teunen et al.	704	250	
	4M	6,233,559	05/15/2001	Balakrishnan	704	275	
	4N	6,246,981	06/12/2001	Papineni et al.	704	235	
	4O	6,272,455	08/07/2001	Hoshen et al.	704	1	
	4P	6,288,319	09/11/2001	Catona	84	609	
	4Q	6,292,767	09/18/2001	Jackson et al.	704	1	
	4R	6,308,151	10/23/2001	Smith	704	235	
	4S	6,314,402	11/06/2001	Monaco et al.	704	275	
	4T	6,366,886	04/02/2002	Dragosh et al.	704	270.1	
	4U	6,374,214	04/16/2002	Friedland *et al.*	704	235	
	4V	6,381,535	04/30/2002	Durocher *et al.*	701	202	
	4W	6,385,596	05/07/2002	Wiser *et al.*	705	51	

FOREIGN PATENT DOCUMENTS

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

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

	4X	CHEYER *et al.*, "Multimodal Maps: An Agent-Based Approach", International Conference on Cooperative Multimodal Communication (CMC/95), May 24-26, 1995, pages 111-121.
	4Y	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.
	4Z	TURUNEN, "Adaptive Interaction Methods in Speech User Interfaces", Conference on Human Factors in Computing Systems, Seattle, Washington, 2001, pages 91-92.

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.

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0384422	*Unassigned*
	Applicant(s)	
	Mike KENNEWICK *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	Herewith (November 10, 2010)	*Unassigned*

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	5A	6,385,646	05/07/2002	Brown et al.	709	217	
	5B	6,393,428	05/21/2002	Miller et al.	707	102	
	5C	6,404,878	06/11/2002	Jackson et al.	379	221.01	
	5D	6,408,272	06/18/2002	White *et al.*	704	270.1	
	5E	6,411,810	06/25/2002	Maxemchuk	455	453	
	5F	6,415,257	07/02/2002	Junqua *et al.*	704	275	
	5G	6,418,210	07/09/2002	Sayko	379	142.15	
	5H	6,420,975	07/16/2002	DeLine *et al.*	340	815.4	
	5I	6,430,285	08/06/2002	Bauer et al.	379	265.01	
	5J	6,430,531	08/06/2002	Polish	704	257	
	5K	6,434,523	08/13/2002	Monaco	704	257	
	5L	6,434,524	08/13/2002	Weber	704	257	
	5M	6,442,522	08/27/2002	Carberry et al.	704	257	
	5N	6,446,114	09/03/2002	Bulfer et al.	709	206	
	5O	6,453,153	09/17/2002	Bowker et al.	455	67.4	
	5P	6,453,292	09/17/2002	Ramaswamy et al.	704	235	
	5Q	6,456,711	09/24/2002	Cheung et al.	379	265.09	
	5R	6,466,654	10/15/2002	Cooper *et al.*	379	88.01	
	5S	6,466,899	10/15/2002	Yano *et al.*	704	1	
	5T	6,470,315	10/22/2002	Netsch *et al.*	704	256	
	5U	6,498,797	12/24/2002	Anerousis et al.	370	522	
	5V	6,499,013	12/24/2002	Weber	704	257	
	5W	6,501,833	12/31/2002	Phillips *et al.*	379	88.07	

FOREIGN PATENT DOCUMENTS

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

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

	5X	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.
	5Y	VANHOUCKE, Vincent, "Confidence Scoring and Rejection Using Multi-Pass Speech Recognition", Nuance Communications, Menlo Park, CA, 2005, 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.

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0384422	*Unassigned*
	Applicant(s)	
	Mike KENNEWICK *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	Herewith (November 10, 2010)	*Unassigned*

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	6A	6,501,834	12/31/2002	Milewski et al.	379	93.24	
	6B	6,513,006	01/28/2003	Howard *et al.*	704	257	
	6C	6,523,061	02/18/2003	Halverson et al.	709	202	
	6D	6,532,444	03/11/2003	Weber	704	257	
	6E	6,539,348	03/25/2003	Bond et al.	704	9	
	6F	6,553,372	04/22/2003	Brassell et al.	707	5	
	6G	6,556,970	04/29/2003	Sasaki *et al.*	704	257	
	6H	6,556,973	04/29/2003	Lewin	704	277	
	6I	6,560,576	05/06/2003	Cohen et al.	704	270	
	6J	6,567,778	05/20/2003	Chao Chang et al.	704	257	
	6K	6,567,797	05/20/2003	Schuetze *et al.*	707	2	
	6L	6,570,555	05/27/2003	Prevost *et al.*	345	156	
	6M	6,570,964	05/27/2003	Murveit et al.	379	67.1	
	6N	6,574,597	06/03/2003	Mohri *et al.*	704	251	
	6O	6,574,624	06/03/2003	Johnson et al.	707	5	
	6P	6,581,103	06/17/2003	Dengler	709	231	
	6Q	6,594,257	07/15/2003	Doshi et al.	370	352	
	6R	6,598,018	07/22/2003	Junqua	704	251	
	6S	6,604,075	08/05/2003	Brown *et al.*	704	270.1	
	6T	6,604,077	08/05/2003	Dragosh et al.	704	270.1	
	6U	6,611,692	08/26/2003	Raffel et al.	455	552	
	6V	6,614,773	09/02/2003	Maxemchuk	370	337	
	6W	6,615,172	09/02/2003	Bennett *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.)*

	6X	WENG, Fuliang, *et al.*, "Efficient Lattice Representation and Generation", Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, 1998, 4 pages.
	6Y	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.

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0384422	*Unassigned*
	Applicant(s)	
	Mike KENNEWICK *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	Herewith (November 10, 2010)	*Unassigned*

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	7A	6,629,066	09/30/2003	Jackson et al.	704	9	
	7B	6,631,346	10/07/2003	Karaorman *et al.*	704	9	
	7C	6,643,620	11/04/2003	Contolini *et al.*	704	270	
	7D	6,650,747	11/18/2003	Bala et al.	379	265.06	
	7E	6,658,388	12/02/2003	Kleindienst *et al.*	704	275	
	7F	6,678,680	01/13/2004	Woo	707	6	
	7G	6,691,151	02/10/2004	Cheyer et al.	709	202	
	7H	6,708,150	03/16/2004	Hirayama *et al.*	704	243	
	7I	6,721,001	04/13/2004	Berstis	348	231.3	
	7J	6,721,706	04/13/2004	Strubbe *et al.*	704	275	
	7K	6,735,592	05/11/2004	Neumann *et al.*	707	101	
	7L	6,741,931	05/25/2004	Kohut *et al.*	701	209	
	7M	6,742,021	05/25/2004	Halverson *et al.*	709	218	
	7N	6,751,591	06/15/2004	Gorin *et al.*	704	257	
	7O	6,757,718	06/29/2004	Halverson *et al.*	709	218	
	7P	6,795,808	09/21/2004	Strubbe *et al.*	704	275	
	7Q	6,801,604	10/05/2004	Maes *et al.*	379	88.17	
	7R	6,801,893	10/05/2004	Backfried *et al.*	704	257	
	7S	6,829,603	12/07/2004	Chai *et al.*	707	5	
	7T	6,833,848	12/21/2004	Wolff *et al.*	345	719	
	7U	6,856,990	02/15/2005	Barile *et al.*	707	10	
	7V	6,865,481	03/08/2005	Kawazoe *et al.*	701	211	
	7W	6,877,134	04/05/2005	Fuller *et al.*	715	500.1	

FOREIGN PATENT DOCUMENTS

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

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

	7X	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.
	7Y	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.

402465373v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0384422	*Unassigned*
	Applicant(s)	
	Mike KENNEWICK *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	Herewith (November 10, 2010)	*Unassigned*

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	8A	6,901,366	05/31/2005	Kuhn *et al.*	704	275	
	8B	6,937,977	08/30/2005	Gerson	704	201	
	8C	6,944,594	09/13/2005	Busayapongchai *et al.*	704	275	
	8D	6,950,821	09/27/2005	Faybishenko *et al.*	707	10	
	8E	6,973,387	12/06/2005	Masclet *et al.*	701	211	
	8F	6,980,092	12/27/2005	Turnbull *et al.*	340	425.5	
	8G	6,990,513	01/24/2006	Belfiore *et al.*	709	203	
	8H	6,996,531	02/07/2006	Korall *et al.*	704	270	
	8I	7,016,849	03/21/2006	Arnold *et al.*	704	275	
	8J	7,020,609	03/28/2006	Thrift *et al.*	704	270.1	
	8K	7,027,975	04/11/2006	Pazandak *et al.*	704	9	
	8L	7,062,488	06/13/2006	Reisman	707	8	
	8M	7,082,469	07/25/2006	Gold *et al.*	709	231	
	8N	7,092,928	08/15/2006	Elad *et al.*	706	60	
	8O	7,127,400	10/24/2006	Koch	704	270.1	
	8P	7,136,875	11/14/2006	Anderson *et al.*	707	104.1	
	8Q	7,137,126	11/14/2006	Coffman *et al.*	719	328	
	8R	7,146,319	12/05/2006	Hunt	704	254	
	8S	7,203,644	04/10/2007	Anderson *et al.*	704	246	
	8T	7,228,276	06/05/2007	Omote *et al.*	704	243	
	8U	7,231,343	06/12/2007	Treadgold *et al.*	704	9	
	8V	7,289,606	10/30/2007	Sibal *et al.*	379	52	
	8W	7,301,093	11/27/2007	Sater *et al.*	84	615	

FOREIGN PATENT DOCUMENTS

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

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

	8X	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.

402465373v1

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0384422	*Unassigned*
	Applicant(s)	
	Mike KENNEWICK *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	Herewith (November 10, 2010)	*Unassigned*

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	9A	7,340,040	03/04/2008	Saylor *et al.*	379	67.1	
	9B	7,376,645	05/20/2008	Bernard	707	3	
	9C	7,386,443	06/10/2008	Parthasarathy *et al.*	704	201	
	9D	7,398,209	07/08/2008	Kennewick *et al.*	704	255	
	9E	7,424,431	09/09/2008	Greene *et al.*	704	270	
	9F	7,461,059	12/02/2008	Richardson *et al.*	707	5	
	9G	7,487,088	02/03/2009	Gorin *et al.*	704	240	
	9H	7,493,559	02/17/2009	Wolff *et al.*	715	727	
	9I	7,502,738	03/10/2009	Kennewick *et al.*	704	257	
	9J	7,620,549	11/17/2009	Di Cristo *et al.*	704	257	
	9K	7,634,409	12/15/2009	Kennewick *et al.*	704	257	
	9L	7,640,160	12/29/2009	Di Cristo *et al.*	704	257	
	9M	7,693,720	04/06/2010	Kennewick *et al.*	704	275	
	9N	7,809,570	10/05/2010	Kennewick *et al.*	704	257	
	9O	7,818,176	10/19/2010	Freeman *et al.*			
	9P	2001/0041980	11/15/2001	Howard *et al.*	704	270	
	9Q	2002/0049805	04/25/2002	Yamada *et al.*	709	202	
	9R	2002/0065568	05/30/2002	Silfvast *et al.*	700	94	
	9S	2002/0069059	06/06/2002	Smith	704	257	
	9T	2002/0082911	06/27/2002	Dunn *et al.*	705	14	
	9U	2002/0087525	07/04/2002	Abbott *et al.*	707	3	
	9V	2002/0120609	08/29/2002	Lang *et al.*	707	1	
	9W	2002/0124050	09/05/2002	Middeljans	709	203	

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.

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0384422	*Unassigned*
	Applicant(s)	
	Mike KENNEWICK *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	Herewith (November 10, 2010)	*Unassigned*

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	10A	2002/0138248	09/26/2002	Corston-Oliver *et al.*	704	1	
	10B	2002/0143535	10/03/2002	Kist *et al.*	704	251	
	10C	2002/0188602	12/12/2002	Stubler *et al.*	707	3	
	10D	2002/0198714	12/26/2002	Zhou	704	252	
	10E	2003/0014261	01/16/2003	Kageyama	704	275	
	10F	2003/0046346	03/06/2003	Mumick *et al.*	709	205	
	10G	2003/0064709	04/03/2003	Gailey *et al.*	455	412	
	10H	2003/0088421	05/08/2003	Maes *et al.*	704	270.1	
	10I	2003/0097249	05/22/2003	Walker *et al.*	704	1	
	10J	2003/0112267	06/19/2003	Belrose	345	728	
	10K	2003/0135488	07/17/2003	Amir *et al.*	707	3	
	10L	2003/0182132	09/25/2003	Niemoeller	704	275	
	10M	2003/0236664	12/25/2003	Sharma	704	251	
	10N	2004/0006475	01/08/2004	Ehlen *et al.*	704	270.1	
	10O	2004/0025115	02/05/2004	Sienel *et al.*	715	513	
	10P	2004/0044516	03/04/2004	Kennewick *et al.*	704	5	
	10Q	2004/0166832	08/26/2004	Portman *et al.*	455	412.1	
	10R	2004/0193408	09/30/2004	Hunt	704	209	
	10S	2004/0193420	09/30/2004	Kennewick *et al.*	704	257	
	10T	2004/0199375	10/07/2004	Ehsani *et al.*	704	4	
	10U	2004/0243417	12/02/2004	Pitts, III *et al.*	704	276	
	10V	2005/0015256	01/20/2005	Kargman	704	272	
	10W	2005/0021334	01/27/2005	Iwahashi	704	240	

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.

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0384422	*Unassigned*
	Applicant(s)	
	Mike KENNEWICK *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	Herewith (November 10, 2010)	*Unassigned*

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	11A	2005/0021826	01/27/2005	Kumar	709	232	
	11B	2005/0033574	02/10/2005	Kim *et al.*	704	251	
	11C	2005/0043940	02/24/2005	Elder	704	9	
	11D	2005/0114116	05/26/2005	Fiedler	704	201	
	11E	2005/0137850	06/23/2005	Odell	704	4	
	11F	2005/0246174	11/03/2005	DeGolia	704	270	
	11G	2006/0206310	09/14/2006	Ravikumar *et al.*	704	9	
	11H	2007/0033005	02/08/2007	Cristo *et al.*	704	9	
	11I	2007/0038436	02/15/2007	Cristo *et al.*	704	9	
	11J	2007/0043574	02/22/2007	Coffman *et al.*	704	275	
	11K	2007/0050191	03/01/2007	Weider *et al.*	704	275	
	11L	2007/0055525	03/08/2007	Kennewick *et al.*	704	257	
	11M	2007/0179778	08/02/2007	Gong *et al.*	704	9	
	11N	2007/0186165	08/09/2007	Maislos *et al.*	715	728	
	11O	2007/0214182	09/13/2007	Rosenberg	707	104.1	
	11P	2007/0250901	10/25/2007	McIntire *et al.*	725	146	
	11Q	2007/0265850	11/15/2007	Kennewick *et al.*	704	257	
	11R	2007/0299824	12/27/2007	Pan *et al.*	707	3	
	11S	2008/0091406	04/17/2008	Baldwin *et al.*	704	4	
	11T	2008/0103761	05/01/2008	Printz *et al.*	704	9	
	11U	2008/0115163	05/15/2008	Gilboa *et al.*	725	34	
	11V	2008/0133215	06/05/2008	Sarukkai	704	2	
	11W	2008/0177530	07/24/2008	Cross *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.)*

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.

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0384422	*Unassigned*
	Applicant(s)	
	Mike KENNEWICK *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	Herewith (November 10, 2010)	*Unassigned*

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	12A	2008/0189110	08/07/2008	Freeman *et al.*	704	251	
	12B	2008/0235023	09/25/2008	Kennewick *et al.*	704	257	
	12C	2008/0235027	09/25/2008	Cross	704	270.1	
	12D	2008/0319751	12/25/2008	Kennewick *et al.*	704	257	
	12E	2009/0117885	05/07/2009	Roth	455	414.3	
	12F	2009/0144271	06/04/2009	Richardson *et al.*	707	5	
	12G	2009/0150156	06/11/2009	Kennewick *et al.*	704	257	
	12H	2009/0171664	07/02/2009	Kennewick *et al.*	704	257	
	12I	2009/0299745	12/03/2009	Kennewick *et al.*	704	257	
	12J	2010/0023320	01/28/2010	Di Cristo *et al.*	704	9	
	12K	2010/0049501	02/25/2010	Kennewick *et al.*	704	9	
	12L	2010/0049514	02/25/2010	Kennewick *et al.*	704	233	
	12M	2010/0057443	03/04/2010	Di Cristo *et al.*	704	9	
	12N	2010/0063880	03/11/2010	Atsmon *et al.*	705	14.53	
	12O	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.

SYSTEM AND METHOD FOR PROVIDING A NATURAL LANGUAGE CONTENT DEDICATION SERVICE

CROSS-REFERENCE TO RELATED APPLICATIONS

[001] This application claims the benefit of U.S. Provisional Patent Application Serial No. 61/259,820, entitled "System and Method for Providing a Natural Language Content Dedication Service," filed November 10, 2009, the contents of which are hereby incorporated by reference in their entirety.

FIELD OF THE INVENTION

[002] The invention generally relates to providing a natural language content dedication service in a voice services environment, and in particular, to detecting multi-modal device interactions that include requests to dedicate content, identifying the content requested for dedication from natural language utterances included in the multi-modal device interactions, processing transactions for the content requested for dedication, processing natural language to customize the content for recipients of the dedications, and delivering the customized content to the recipients of the dedications.

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' devices other than those 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 cooperative multi-modal interactions 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. In another example, users often listen to music or interact with other media in mobile environments, such that interest in purchasing music, media, or other content may be fleeting or often occur on an impulse basis. Whereas existing human to machine interfaces that lack voice recognition capabilities typically fall short in providing mechanisms that can readily meet

this demand, adding voice recognition to an electronic device can substantially simplify human to machine interaction in a manner that can meet user needs, improve experience, and satisfy potentially transient consumer interests. As such, interaction with electronic devices could be made far more efficient if users were provided with the ability to use natural language in order to exploit buried or otherwise difficult to use functionality.

SUMMARY OF THE INVENTION

[010] According to one aspect of the invention, a system and method for providing a natural language content dedication service may generally operate in a voice services environment that includes one or more electronic devices that can receive multi-modal natural language device interactions. In particular, providing the natural language content dedication service may generally include detecting multi-modal device interactions that include requests to dedicate content, identifying the content requested for dedication from natural language utterances included in the multi-modal device interactions, processing transactions for the content requested for dedication, processing natural language to customize the content for recipients of the dedications, and delivering the customized content to the recipients of the dedications.

[011] According to one aspect of the invention, the natural language content dedication service may operate in a hybrid processing environment, which may generally include a plurality of multi-modal devices configured to cooperatively interpret and process natural language utterances included in the multi-modal device interactions. For example, a virtual router may receive messages that include encoded audio corresponding to natural language utterances contained in the multi-modal device interactions, which may be received at one or more of the plurality of multi-modal devices in the hybrid processing environment. For example, the virtual router may analyze the encoded audio to select a cleanest sample of the natural language utterances and communicate with one or more other devices in the hybrid processing environment to determine an intent of the multi-modal device interactions. The virtual router may then coordinate resolving the multi-modal device interactions based on the intent of the multi-modal device interactions.

[012] According to one aspect of the invention, a method for providing the natural language content dedication service may comprise detecting a multi-modal device interaction at an

electronic device, wherein the multi-modal device interaction may include at least a natural language utterance. One or more messages containing information relating to the multi-modal device interaction may then be communicated to the virtual router through a messaging interface. The electronic device may then receive one or more messages (e.g., from the virtual router through message interface), wherein the messages may contain information relating to an intent of the multi-modal device interaction. As such, the multi-modal device interaction may be resolved at the electronic device based on the information contained in the one or more messages received from the virtual router.

[013] According to one aspect of the invention, a system for providing the natural language content dedication service may generally include a voice-enabled client device that can communicate with a content dedication system through the messaging interface. The content dedication system include a voice-enabled server, which may be configured to communicate with the virtual router through another messaging interface, or the content dedication system may alternatively include the virtual router. In addition, the content dedication system may further include a billing system for processing transactions relating to the content dedication service. The natural language content dedication service may be provided on any suitable voice-enabled client device having a suitable combination of input and output devices that can receive and respond to multi-modal device interactions that include natural language utterances, and the input and output devices may be further arranged to receive and respond to any other suitable type of input and output.

[014] According to one aspect of the invention, operating the natural language content dedication service may generally include a user of the voice-enabled client device listening to music, watching video, or otherwise interacting with content and providing a multi-modal natural language request to engage in a transaction to dedicate the music, video, or other content. Furthermore, the voice-enabled client device may be included within the hybrid processing environment that includes the plurality of multi-modal devices, whereby the content dedication request may relate to content played on a different device from the voice-enabled client device, although the content dedication request may relate to any suitable content (i.e., the request need not necessarily relate to played content, as users may provide natural

language to request content dedications for any suitable content, including a particular song or video that the user may be thinking about).

[015] According to one aspect of the invention, in response to the voice-enabled client device receiving a multi-modal interaction that includes a natural language utterance, the voice-enabled client device may invoke an Automatic Speech Recognizer (ASR) to generate a preliminary interpretation of the utterance. The ASR may then provide the preliminary interpretation of the utterance to a conversational language processor, which may attempt to determine an intent for the multi-modal interaction. For example, to determine the intent for the multi-modal interaction, the conversational language processor may determine a most likely context for the interaction from the preliminary interpretation of the utterance, any accompanying non-speech inputs in the multi-modal interaction that relate to the utterance, contexts associated with prior requests, short-term and long-term shared knowledge, or any other suitable information for interpreting the multi-modal interaction. Thus, in response to the conversational language processor determining that the intent for the multi-modal interaction relates to a content dedication request, a content dedication application may be invoked to resolve the content dedication request.

[016] According to one aspect of the invention, to resolve the intent of the content dedication request, the conversational language processor may search one or more data repositories that contain content information to identify content matching criteria contained in the content dedication request. Moreover, the conversational language processor may further cooperate with other devices in the hybrid processing environment to search for or otherwise identify the content requested for dedication (e.g., in response to a local data repository not yielding adequate results, another device in the hybrid processing environment having a larger content data repository than the client device may be invoked.). The conversational language processor may then receive appropriate results identifying the content requested for dedication and present the results to the user through the output device (e.g., displaying information about the content, playing a sample clip of the content, displaying options to purchase the content, recommending similar content, etc.). The results presented through the output device may further include an option to confirm the content dedication, wherein the user may confirm the

content dedication in a natural language utterance, a non-speech input, or any suitable combination thereof. The content dedication application may then be invoked to process the content dedication request.

[017] According to one aspect of the invention, to process the content dedication request, the content dedication application may capture a natural language utterance that contains the dedication to accompany the content. The user may then provide the dedication utterance through the voice-enabled input device, and the dedication utterance may then be converted into an electronic signal that the content dedication application captures for the dedication. In addition, the content dedication application may prompt the user to provide any additional tags for the dedicated content (e.g., an image to insert as album art in the dedicated content, an utterance to insert or transcribe into metadata tags for the dedicated content, a non-speech or data input to insert in the metadata tags for the dedicated content, etc.). The content dedication application may further prompt the user to identify a recipient of the dedication, wherein the user may provide any suitable multi-modal input that includes information identifying the recipient of the content dedication. The content dedication application may then route the request to the content dedication system, which may process a transaction for the content dedication.

[018] According to one aspect of the invention, processing the transaction for the content dedication may include the content dedication system receiving encoded audio corresponding to the dedication utterance through the messaging interface. The content dedication system may then insert the encoded audio corresponding to the dedication utterance within the dedicated content, verbally annotate the dedicated content with the encoded audio, and/or transcribe the dedication utterance into a textual annotation for the dedicated content. Similarly, any utterances to insert into the metadata tags for the dedicated content may provide further verbal annotations for the dedicated content, and any such utterances may also be transcribed into text to provide further textual annotations for the dedicated content. The content dedication system may then invoke a content dedication application hosted on the voice-enabled server to process the request. In particular, the content dedication application hosted on the voice-enabled server may identify the content requested for dedication (e.g., if

the content dedication application on the voice-enabled client device was unable to suitably identify the requested content, or the information communicated from the voice-enabled client device may identify the requested content if the content dedication application on the voice-enabled client device was able to suitably identify the requested content, etc.). In response to identifying the content to be dedicated, the content dedication system may then communicate with the billing system to process an appropriate transaction for the dedication request based on a selected purchase option for the dedication request.

[019] According to one aspect of the invention, the content dedication system may support various purchase options to provide users with flexibility in requesting content dedications. For example, a buy-to-own purchase option may include the content dedication system purchasing full rights to the content from an appropriate content provider. The billing system may then charge the user of the voice-enabled client device an appropriate amount that encompasses the cost for purchasing the rights to the content from the content provider and a service charge for customizing the content with the dedication and any additional tags and subsequently delivering the customized content to the dedication recipient. The user may be charged in a similar manner under a pay-to-play purchase option, except that the rights purchased from the content provider may be limited (e.g., to a predetermined number of plays). Thus, the cost for purchasing the content from the content provider may be somewhat less under the pay-to-play purchase option, such that the billing system may charge the user somewhat less under the pay-to-play purchase option. Under a paid subscription purchase option, the user may pay a periodic service charge to the content dedication system that permits the user to make content dedications based on terms of the subscription (e.g., a predetermined number an unlimited number of content dedications may be made in a subscription period depending on the particular terms of the user's subscription). Furthermore, other purchase options may be suitably employed, as will be apparent.

[020] According to one aspect of the invention, a service provider associated with the content dedication system may negotiate agreements with content providers to determine the manner in which revenues for content transactions will be shared between the content dedication system and the content providers. For example, an agreement may permit a particular content

provider to keep all of the revenue for transactions that include purchasing content from the content provider, while the service provider associated with the content dedication system agrees to recoup any such costs from users. In another example, an agreement may share revenue for content transactions between a content provider and the service provider. Thus, the agreements may generally include any suitable arrangements that define the manner in which content providers and the service provider agree to divide revenue for transactions in which the content dedication system purchases content from the content provider, while the service provider associated with the content dedication system may be responsible for billing users for any natural language aspects of the content dedication service (e.g., for inserting dedication utterances, transcribing utterances into metadata tags, delivering the content to recipients, etc.).

[021] According to one aspect of the invention, in response to purchasing the requested content from an appropriate service provider and determining the appropriate billing options for the content dedication, the content dedication system may insert the natural language dedication utterance into the dedicated content, verbally annotate the dedicated content with the dedication utterance, or otherwise associate the content with the dedication utterance. Furthermore, the content dedication system may determine whether any additional tags have been specified for the dedicated content and insert such additional tags into the dedicated content, as appropriate (e.g., inserting an image or picture into metadata tags corresponding to album art for the dedicated content, transcribing natural language utterances, non-voice, and/or data inputs into text and inserting such text into the metadata tags for the dedicated content, etc.). The content dedication system may then send a content dedication message to the recipient of the dedication, wherein the message may include a link that the recipient can select to stream, download, or otherwise access the dedicated content, the dedication utterance, etc.). Thus, the content dedication message may generally notify the recipient that content has been dedicated to the recipient and provide various mechanisms for the recipient to access the content dedication, as will be apparent. Furthermore, if the purchase option selected for the content dedication includes the buy-to-own purchase option or another purchase option that confers full rights to the dedicated content, the recipient may then own

the full rights to the dedicated content, otherwise the recipient may own rights with respect to the dedicated content based on whatever terms the selected purchase option provides.

[022] Other objects and advantages of the invention will be apparent based on the following drawings and detailed description.

BRIEF DESCRIPTION OF THE DRAWINGS

[023] Figure 1 illustrates a block diagram of an exemplary voice-enabled device that can be used for hybrid processing in a natural language voice services environment, according to one aspect of the invention.

[024] Figure 2 illustrates a block diagram of an exemplary system for hybrid processing in a natural language voice service environment, according to one aspect of the invention.

[025] Figure 3 illustrates a flow diagram of an exemplary method for initializing various devices that cooperate to perform hybrid processing in a natural language voice services environment, according to one aspect of the invention.

[026] Figures 4-5 illustrate flow diagrams of exemplary methods for hybrid processing in a natural language voice services environment, according to one aspect of the invention.

[027] Figure 6 illustrates a block diagram of an exemplary system for providing a natural language content dedication service, according to one aspect of the invention.

[028] Figures 7-8 illustrate flow diagrams of exemplary methods for providing a natural language content dedication service, according to one aspect of the invention.

DETAILED DESCRIPTION

[029] According to one aspect of the invention, Figure 1 illustrates a block diagram of an exemplary voice-enabled device 100 that can be used for hybrid processing in a natural language voice services environment. As will be apparent from the further description to be provided herein, the voice-enabled device 100 illustrated in Figure 1 may generally include an input device 112, or a combination of input devices 112, which may enable a user to interact with the voice-enabled device 100 in a multi-modal manner. In particular, the input devices 112 may generally include any suitable combination of at least one voice input device 112 (e.g., a microphone) and at least one non-voice input device 112 (e.g., a mouse, touch-screen display,

wheel selector, etc.). As such, the input devices 112 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).

[030] In one implementation, the voice-enabled device 100 may enable the user to engage in various multi-modal conversational interactions, which the voice-enabled device 100 may process in a free-form and cooperative manner to execute various tasks, resolve various queries, or otherwise resolve various natural language requests included in the multi-modal interactions. For example, in one implementation, the voice-enabled device 100 may include various natural language processing components, including at least a voice-click module coupled to the one or more input devices 112, as described in further detail in co-pending 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, the contents of which are hereby incorporated by reference in their entirety. Thus, as will be described in further detail herein, the one or more input devices 112 and the voice-click module may be collectively configured to process various multi-modal interactions between the user and the voice-enabled device 100.

[031] For example, in one implementation, the multi-modal interactions may include at least one natural language utterance, wherein the natural language utterance may be converted into an electronic signal. The electronic signal may then be provided to an Automatic Speech Recognizer (ASR) 120, which may also be referred to as a speech recognition engine 120 and/or a multi-pass speech recognition engine 120. In response to receiving the electronic signal corresponding to the utterance, the ASR 120 may generate one or more preliminary interpretations of the utterance and provide the preliminary interpretation to a conversational language processor 130. Additionally, in one implementation, the multi-modal interactions may include one or more non-voice interactions with the one or more input devices 112 (e.g., button pushes, multi-touch gestures, point of focus or attention focus selections, etc.). As such, the voice-click module may extract context from the non-voice interactions and provide the

context to the conversational language processor 130 for use in generating an interpretation of the utterance (i.e., via the dashed line illustrated in Figure 1). As such, as described in greater detail below, the conversational language processor 130 may analyze the utterance and any accompanying non-voice interactions to determine an intent of the multi-modal interactions with the voice-enabled device 100.

[032] In one implementation, as noted above, the voice-enabled device 100 may include various natural language processing components that can support free-form utterances and/or other forms of non-voice device interactions, which may liberate the user from restrictions relating to the manner of formulating commands, queries, or other requests. As such, the user may provide the utterance to the voice input device 112 using any manner of speaking, and may further provide other non-voice interactions to the non-voice input device 112 to request any content or service available through the voice-enabled device 100. For instance, in one implementation, in response to receiving the utterance at the voice input device 112, the utterance may be processed using techniques described in U.S. Patent Application Serial No. 10/452,147, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," which issued as U.S. Patent No. 7,398,209 on July 8, 2008, and co-pending U.S. Patent Application Serial No. 10/618,633, entitled "Mobile Systems and Methods for Responding to Natural Language Speech Utterance," filed June 15, 2003, the contents 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 112 to provide buttons pushes, multi-touch gestures, point of focus or attention focus selections, or other non-voice device interactions, which may provide further context or other information relating to the natural language utterances and/or the requested content or service.

[033] In one implementation, the voice-enabled device 100 may be coupled to one or more additional systems that may be configured to cooperate with the voice-enabled device 100 to interpret or otherwise process the multi-modal interactions that include combinations of natural language utterances and/or non-voice device interactions. For example, as will be described in greater detail below in connection with Figure 2, the one or more additional systems may include one or more multi-modal voice-enabled devices having similar natural

language processing capabilities to the voice-enabled device 100, one or more non-voice devices having data retrieval and/or task execution capabilities, and a virtual router that coordinates interaction among the voice-enabled device 100 and the additional systems. As such, the voice-enabled device 100 may include an interface to an integrated natural language voice services environment that includes a plurality of multi-modal devices, wherein the user may request content or services available through any of the multi-modal devices.

[034] For example, in one implementation, the conversational language processor 130 may include a constellation model 132b that provides knowledge relating to content, services, applications, intent determination capabilities, and other features available in the voice services environment, as described in co-pending U.S. Patent Application Serial No. 12/127,343, entitled "System and Method for an Integrated, Multi-Modal, Multi-Device Natural Language Voice Services Environment," filed May 27, 2008, the contents of which are hereby incorporated by reference in their entirety. As such, the voice-enabled device 100 may have access to shared knowledge relating to natural language processing capabilities, context, prior interactions, domain knowledge, short-term knowledge, long-term knowledge, and cognitive models for the various systems and multi-modal devices, providing a cooperative environment for resolving the multi-modal interactions received at the voice-enabled device 100.

[035] In one implementation, the input devices 112 and the voice-click module coupled thereto may be configured to continually monitor for one or more multi-modal interactions received at the voice-enabled device 100. In particular, the input devices 112 and the voice-click module may continually monitor for one or more natural language utterances and/or one or more distinguishable non-voice device interactions, which may collectively provide the relevant context for retrieving content, executing tasks, invoking services or commands, or processing any other suitable requests. Thus, in response to detecting one or more multi-modal interactions, the input devices 112 and/or the voice-click module may signal the voice-enabled device 100 that an utterance and/or a non-voice interaction have been received. For example, in one implementation, the non-voice interaction may provide context for sharpening recognition, interpretation, and understanding of an accompanying utterance, and moreover, the utterance may provide further context for enhancing interpretation of the accompanying

non-voice interaction. Accordingly, the utterance and the non-voice interaction may collectively provide relevant context that various natural language processing components may use to determine an intent of the multi-modal interaction that includes the utterance and the non-voice interaction.

[036] In one implementation, as noted above, processing the utterance included in the multi-modal interaction may be initiated at the ASR 120, wherein the ASR 120 may generate one or more preliminary interpretations of the utterance. In one implementation, to generate the preliminary interpretations of the utterance, the ASR 120 may be configured to recognize one or more syllables, words, phrases, or other acoustic characteristics from the utterance using one or more dynamic recognition grammars and/or acoustic models. For example, in one implementation, the ASR 120 may use the dynamic recognition grammars and/or the acoustic models to recognize a stream of phonemes from the utterance based on phonetic dictation techniques, as described in U.S. Patent Application Serial No. 11/513,269, entitled "Dynamic Speech Sharpening," which issued as U.S. Patent No. 7,634,409 on December 15, 2009, the contents of which are hereby incorporated by reference in their entirety. In addition, the dynamic recognition grammars and/or the acoustic models may include unstressed central vowels (e.g., "schwa"), which may reduce a search space for recognizing the stream of phonemes for the utterance.

[037] Furthermore, in one implementation, the ASR 120 may be configured as a multi-pass speech recognition engine 120, as described in U.S. Patent Application Serial No. 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," which issued as U.S. Patent No. 7,640,160 on December 29, 2009, the contents of which are hereby incorporated by reference in their entirety. The multi-pass speech recognition 120 may be configured to initially invoke a primary speech recognition engine to generate a first transcription of the utterance, and further to optionally subsequently invoke one or more secondary speech recognition engines to generate one or more secondary transcriptions of the utterance. In one implementation, the first transcription may be generated using a large list dictation grammar, while the secondary transcriptions may be generated using virtual dictation grammars having decoy words for out-of-vocabulary words, reduced vocabularies derived from

a conversation history, or other dynamic recognition grammars. For example, in one implementation, if a confidence level for the first transcription does not meet or exceed a threshold, the secondary speech recognition engines may be invoked to sharpen the interpretation of the primary speech recognition engine. It will be apparent, however, that the multi-pass speech recognition engine 120 may interpret the utterance using any suitable combination of techniques that results in a preliminary interpretation derived from a plurality of transcription passes for the utterance (e.g., the secondary speech recognition engines may be invoked regardless of the confidence level for the first transcription, or the primary speech recognition engine and/or the secondary speech recognition engines may employ recognition grammars that are identical or optimized for a particular interpretation context, etc.).

[038] Accordingly, in one implementation, the dynamic recognition grammars used in the ASR 120 may be optimized for different languages, contexts, domains, memory constraints, and/or other suitable criteria. For example, in one implementation, the voice-enabled device 100 may include one or more applications 134 that provide content or services for a particular context or domain, such as a navigation application 134. As such, in response to the ASR 120 determining navigation as the most likely context for the utterance, the dynamic recognition grammars may be optimized for various physical, temporal, directional, or other geographical characteristics (e.g., as described in co-pending U.S. Patent Application Serial No. 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 contents of which are hereby incorporated by reference in their entirety). In another example, an utterance containing the word "traffic" may be subject to different interpretations depending on whether the user intended a navigation context (i.e., traffic on roads), a music context (i.e., the 1960's rock band), or a movie context (i.e., the Steven Soderbergh film). Accordingly, the recognition grammars used in the ASR 120 may be dynamically adapted to optimize accurate recognition for any given utterance (e.g., in response to incorrectly interpreting an utterance to contain a particular word or phrase, the incorrect interpretation may be removed from the recognition grammar to prevent repeating the incorrect interpretation).

[039] In one implementation, in response to the ASR 120 generating the preliminary interpretations of the utterance included in the multi-modal interaction using one or more of the techniques described above, the ASR 120 may provide the preliminary interpretations to the conversational language processor 130. The conversational language processor 130 may generally include various natural language processing components, which may be configured to model human-to-human conversations or interactions. Thus, the conversational language processor 130 may invoke one or more of the natural language processing components to further analyze the preliminary interpretations of the utterance and any accompanying non-voice interactions to determine the intent of the multi-modal interactions received at the voice-enabled device 100.

[040] In one implementation, the conversational language processor 120 may invoke an intent determination engine 130a configured to determine the intent of the multi-modal interactions received at the voice-enabled device 100. In one implementation, the intent determination engine 130a may invoke a knowledge-enhanced speech recognition engine that provides long-term and short-term semantic knowledge for determining the intent, as described in co-pending U.S. Patent Application Serial No. 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," filed August 29, 2005, the contents of which are hereby incorporated by reference in their entirety. For example, in one implementation, the semantic knowledge may be based on a personalized cognitive model derived from one or more prior interactions with the user, a general cognitive model derived from one or more prior interactions with various different users, and/or an environmental cognitive model derived from an environment associated with the user, the voice-enabled device 100, and/or the voice services environment (e.g., ambient noise characteristics, location sensitive information, etc.).

[041] Furthermore, the intent determination engine 132a may invoke a context tracking engine 132d to determine the context for the multi-modal interactions. For example, any context derived from the natural language utterance and/or the non-voice interactions in the multi-modal interactions may be pushed to a context stack associated with the context tracking engine 132d, wherein the context stack may include various entries that may be weighted or

otherwise ranked according to one or more contexts identified from the cognitive models and the context for the current multi-modal interactions. As such, the context tracking engine 132d may determine one or more entries in the context stack that match information associated with the current multi-modal interactions to determine a most likely context for the current multi-modal interactions. The context tracking engine 132d may then provide the most likely context to the intent determination engine 132a, which may determine the intent of the multi-modal interactions in view of the most likely context.

[042] In addition, based on the most likely context, the intent determination engine 132a may reference the constellation model 132b to determine whether to invoke any of the various systems or multi-modal devices in the voice services environment. For example, as noted above, the constellation model 132b may provide intent determination capabilities, domain knowledge, semantic knowledge, cognitive models, and other information available through the various systems and multi-modal devices. As such, the intent determination engine 132a may reference the constellation model 132b to determine whether one or more of the other systems and/or multi-modal devices should be engaged to participate in determining the intent of the multi-modal interactions. For example, in response to the constellation model 132b indicating that one or more of the other systems and/or multi-modal devices have natural language processing capabilities optimized for the most likely context, the intent determination engine 132a may forward information relating to the multi-modal interactions to such systems and/or multi-modal devices, which may then determine the intent of the multi-modal interactions and return the intent determination to the voice-enabled device 100.

[043] In one implementation, the conversational language processor 130 may be configured to engage the user in one or more cooperative conversations to resolve the intent or otherwise process the multi-modal interactions, as described in co-pending U.S. Patent Application Serial No. 11/580,926, entitled "System and Method for a Cooperative Conversational Voice User Interface," filed October 16, 2006, the contents of which are hereby incorporated by reference in their entirety. In particular, the conversational language processor 130 may generally identify a conversational goal for the multi-modal interactions, wherein the conversational goal may be identifying from analyzing the utterance, the non-voice interactions, the most likely

context, and/or the determined intent. As such, the conversational goal identified for the multi-modal interactions may generally control the cooperative conversation between the conversational language processor 130 and the user. For example, the conversational language processor 130 may generally engage the user in one or more query conversations, didactic conversations, and/or exploratory conversations to resolve or otherwise process the multi-modal interactions.

[044] In particular, the conversational language processor 130 may engage the user in a query conversation in response to identifying that the conversational goal relates to retrieving discrete information or performing a particular function. Thus, in a cooperative query conversation, the user may lead the conversation towards achieving the particular conversational goal, while the conversational language processor 130 may initiate one or more queries, tasks, commands, or other requests to achieve the goal and thereby support the user in the conversation. In response to ambiguity or uncertainty in the intent of the multi-modal interaction, the conversational language processor 130 may engage the user in a didactic conversation to resolve the ambiguity or uncertainty (e.g., where noise or malapropisms interfere with interpreting the utterance, multiple likely contexts cannot be disambiguated, etc.). As such, in a cooperative didactic conversation, the conversational language processor 130 may lead the conversation to clarify the intent of the multi-modal interaction (e.g., generating feedback provided through an output device 114), while the user may regulate the conversation and provide additional multi-modal interactions to clarify the intent. In response to determining the intent of the multi-modal interactions with suitable confidence, with the intent indicating an ambiguous or uncertain goal, the conversational language processor 130 may engage the user in an exploratory conversation to resolve the goal. In a cooperative exploratory conversation, the conversational language processor 130 and the user may share leader and supporter roles, wherein the ambiguous or uncertain goal may be improvised or refined over a course of the conversation.

[045] Thus, the conversational language processor 130 may generally engage in one or more cooperative conversations to determine the intent and resolve a particular goal for the multi-modal interactions received at the voice-enabled device 100. The conversational language

processor 130 may then initiate one or more queries, tasks, commands, or other requests in furtherance of the intent and the goal determined for the multi-modal interactions. For example, in one implementation, the conversational language processor 130 may invoke one or more agents 132c having capabilities for processing requests in a particular domain or application 134, a voice search engine 132f having capabilities for retrieving information requested in the multi-modal interactions (e.g., from one or more data repositories 136, networks, or other information sources coupled to the voice-enabled device 100), or one or more other systems or multi-modal devices having suitable processing capabilities for furthering the intent and the goal for the multi-modal interactions (e.g., as determined from the constellation model 132b).

[046] Additionally, in one implementation, the conversational language processor 130 may invoke an advertising application 134 in relation to the queries, tasks, commands, or other requests initiated to process the multi-modal interactions, wherein the advertising application 134 may be configured to select one or more advertisements that may be relevant to the intent and/or the goal for the multi-modal interactions, as described in co-pending U.S. Patent Application Serial No. 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 contents of which are hereby incorporated by reference in their entirety.

[047] In one implementation, in response to receiving results from any suitable combination of queries, tasks, commands, or other requests processed for the multi-modal interactions, the conversational language processor 130 may format the results for presentation to the user through the output device 114. For example, the results may be formatted into a natural language utterance that can be converted into an electronic signal and provided to the user through a speaker coupled to the output device 114, or the results may be visually presented on a display coupled to the output device 114, or in any other suitable manner (e.g., the results may indicate whether a particular task or command was successfully performed, or the results may include information retrieved in response to one or more queries, or the results may include a request to frame a subsequent multi-modal interaction if the results are ambiguous or otherwise incomplete, etc.).

[048] Furthermore, in one implementation, the conversational language processor 130 may include a misrecognition engine 132e configured to determine whether the conversational language processor 130 incorrectly determined the intent for the multi-modal interactions. In one implementation, the misrecognition engine 132e may determine that the conversational language processor 130 incorrectly determined the intent in response to one or more subsequent multi-modal interactions provided proximately in time to the prior multi-modal interactions, as described in U.S. Patent Application Serial No. 11/200,164, entitled "System and Method of Supporting Adaptive Misrecognition in Conversational Speech," which issued as U.S. Patent No. 7,620,549 on November 17, 2009, the contents of which are hereby incorporated by reference in their entirety. For example, the misrecognition engine 132e may monitor for one or more subsequent multi-modal interactions that include a stop word, override a current request, or otherwise indicate an unrecognized or misrecognized event. The misrecognition engine 132e may then determine one or more tuning parameters for various components associated with the ASR 120 and/or the conversational language processor 130 to improve subsequent interpretations.

[049] Accordingly, as described in further detail above, the voice-enabled device 100 may generally include various natural language processing components and capabilities that may be used for hybrid processing in the natural language voice services environment. In particular, the voice-enabled device 100 may be configured to determine the intent for various multi-modal interactions that include any suitable combination of natural language utterances and/or non-voice interactions and process one or more queries, tasks, commands, or other requests based on the determined intent. Furthermore, as noted above and as will be described in greater detail below, one or more other systems and/or multi-modal devices may participate in determining the intent and processing the queries, tasks, commands, or other requests for the multi-modal interactions to provide a hybrid processing methodology, wherein the voice-enabled device 100 and the various other systems and multi-modal devices may each perform partial processing to determine the intent and otherwise process the multi-modal interactions in a cooperative manner. For example, in one implementation, hybrid processing in the natural language voice services environment may include one or more techniques described in U.S.

Provisional Patent Application Serial No. 61/259,827, entitled "System and Method for Hybrid Processing in a Natural Language Voice Services Environment," filed on November 10, 2009, the contents of which are hereby incorporated by reference in their entirety.

[050] According to one aspect of the invention, Figure 2 illustrates a block diagram of an exemplary system for hybrid processing in a natural language voice service environment. In particular, the system illustrated in Figure 2 may generally include a voice-enabled client device 210 similar to the voice-enabled device described above in relation to Figure 1. For example, the voice-enabled client device 210 may include any suitable combination of input and output devices 215a respectively arranged to receive natural language multi-modal interactions and provide responses to the natural language multi-modal interactions. In addition, the voice-enabled client device 210 may include an Automatic Speech Recognizer (ASR) 220a configured to generate one or more preliminary interpretations of natural language utterances received at the input device 215a, and further configured to provide the preliminary interpretations to a conversational language processor 230a.

[051] In one implementation, the conversational language processor 230a on the voice-enabled client device 210 may include one or more natural language processing components, which may be invoked to determine an intent for the multi-modal interactions received at the voice-enabled client device 210. The conversational language processor 230a may then initiate one or more queries, tasks, commands, or other requests to resolve the determined intent. For example, the conversational language processor 230a may invoke one or more applications 234a to process requests in a particular domain, query one or more data repositories 236a to retrieve information requested in the multi-modal interactions, or otherwise engage in one or more cooperative conversations with a user of the voice-enabled client device 210 to resolve the determined intent. Furthermore, as noted above in connection with Figure 1, the voice-enabled client device 210 may also cooperate with one or more other systems or multi-modal devices having suitable processing capabilities for initiating queries, tasks, commands, or other requests to resolve the intent of the multi-modal interactions.

[052] In particular, to cooperate with the other systems or multi-modal devices in the hybrid processing environment, the voice-enabled client device 210 may use a messaging interface

250a to communicative with a virtual router 260, wherein the messaging interface 250a may generally include a light client (or thin client) that provides a mechanism for the voice-enabled client device 210 to transmit input to and receive output from the virtual router 260. In addition, the virtual router 260 may further include a messaging interface 250b providing a mechanism for communicating with one or more additional voice-enabled devices 270a-n, one or more non-voice devices 280a-n, and a voice-enabled server 240. Furthermore, although Figure 2 illustrates messaging interface 250a and messaging interface 250b as components that are distinct from the devices to which they are communicatively coupled, it will be apparent that such illustration is for ease of description only, as the messaging interfaces 250a-b may be provided as on-board components that execute on the various devices illustrated in Figure 2 to facilitate communication among the various devices in the hybrid processing environment.

[053] For example, in one implementation, the messaging interface 250a that executes on the voice-enabled client device 210 may transmit input from the voice-enabled client device 210 to the virtual router 260 within one or more XML messages, wherein the input may include encoded audio corresponding to natural language utterances, preliminary interpretations of the natural language utterances, data corresponding to multi-touch gestures, point of focus or attention focus selections, and/or other multi-modal interactions. In one implementation, the virtual router 260 may then further process the input using a conversational language processor 230c having capabilities for speech recognition, intent determination, adaptive misrecognition, and/or other natural language processing. Furthermore, the conversational language processor 230c may include knowledge relating to content, services, applications, natural language processing capabilities, and other features available through the various devices in the hybrid processing environment.

[054] As such, in one implementation, the virtual router 260 may further communicate with the voice-enabled devices 270, the non-voice devices 280, and/or the voice-enabled server 240 through the messaging interface 250b to coordinate processing for the input received from the voice-enabled client device 210. For example, based on the knowledge relating to the features and capabilities of the various devices in the hybrid processing environment, the virtual router 260 may identify one or more of the devices that have suitable features and/or capabilities for

resolving the intent of the input received from the voice-enabled client device 210. The virtual router 260 may then forward one or more components of the input to the identified devices through respective messaging interfaces 250b, wherein the identified devices may be invoked to perform any suitable processing for the components of the input forwarded from the virtual router 260. In one implementation, the identified devices may then return any results of the processing to the virtual router 260 through the respective messaging interfaces 250b, wherein the virtual router 260 may collate the results of the processing and return the results to the voice-enabled client device 210 through the messaging interface 250a.

[055] Accordingly, the virtual router 260 may communicate with any of the devices available in the hybrid processing environment through messaging interfaces 250a-b to coordinate cooperative hybrid processing for multi-modal interactions or other natural language inputs received from the voice-enabled client device 210. For example, in one implementation, the cooperative hybrid processing may be used to enhance performance in embedded processing architectures in which the voice-enabled client device 210 includes a constrained amount of resources (e.g., the voice-enabled client device 210 may be a mobile device having a limited amount of internal memory or other dedicated resources for natural language processing). As such, when the voice-enabled client device 210 has an embedded processing architecture, one or more components of the voice-enabled client device 210 may be configured to optimize efficiency of on-board natural language processing to reduce or eliminate bottlenecks, lengthy response times, or degradations in performance.

[056] For example, in one implementation, optimizing the efficiency of the on-board natural language processing may include configuring the ASR 220a to use a virtual dictation grammar having decoy words for out-of-vocabulary words, reduced vocabularies derived from a conversation history, or other dynamic recognition grammars (e.g., grammars optimized for particular languages, contexts, domains, memory constraints, and/or other suitable criteria). In another example, the on-board applications 234a and/or data repositories 236a may be associated with an embedded application suite providing particular features and capabilities for the voice-enabled client device 210. For example, the voice-enabled client device 210 may be embedded within an automotive telematics system, a personal navigation device, a global

positioning system, a mobile phone, or another device in which users often request location-based services. Thus, in such circumstances, the on-board applications 234a and the data repositories 236a in the embedded application suite may be optimized to provide certain location-based services that can be efficiently processed on-board (e.g., destination entry, navigation, map control, music search, hands-free dialing, etc.).

[057] Furthermore, although the components of the voice-enabled client device 210 may be optimized for efficiency in embedded architectures, a user may nonetheless request any suitable content, services, applications, and/or other features available in the hybrid processing environment, and the other devices in the hybrid processing environment may collectively provide natural language processing capabilities to supplement the embedded natural language processing capabilities for the voice-enabled client device 210. For example, the voice-enabled client device 210 may perform preliminary processing for a particular multi-modal interaction using the embedded natural language processing capabilities (e.g., the on-board ASR 220a may perform advanced virtual dictation to partially transcribe an utterance in the multi-modal interaction, the on-board conversational language processor 230a may determine a preliminary intent of the multi-modal interaction, etc.), wherein results of the preliminary processing may be provided to the virtual router 260 for further processing.

[058] In one implementation, the voice-enabled client device 210 may also communicate input corresponding to the multi-modal interaction to the virtual router 260 in response to determining that on-board capabilities cannot suitably interpret the interaction (e.g., if a confidence level for a partial transcription does not satisfy a particular threshold), or in response to determining that the interaction should be processed off-board (e.g., if a preliminary interpretation indicates that the interaction relates to a local search request requiring large computations to be performed on the voice-enabled server 240). As such, the virtual router 260 may capture the input received from the voice-enabled client device 210 and coordinate further processing among the voice-enabled devices 270 and the voice-enabled server 240 that provide natural language processing capabilities in addition to the non-voice devices 280 that provide capabilities for retrieving data or executing tasks. Furthermore, in response to the virtual router 260 invoking one or more of the voice-enabled devices 270, the

input provided to the voice-enabled devices 270 may be optimized to suit the processing requested from the invoked voice-enabled devices 270 (e.g., to avoid over-taxing processing resources, a particular voice-enabled device 270 may be provided a partial transcription or a preliminary interpretation and resolve the intent for a given context or domain).

[059] Alternatively, in response to the virtual router 260 invoking the voice-enabled server 240, the input provided to the voice-enabled devices 270 may further include encoded audio corresponding to natural language utterances and any other data associated with the multi-modal interaction. In particular, as shown in Figure 2, the voice-enabled server 240 may have a natural language processing architecture similar to the voice-enabled client device 210, except that the voice-enabled server 240 may include substantial processing resources that obviate constraints that the voice-enabled client device 210 may be subject to. Thus, when the voice-enabled server 240 cooperates in the hybrid processing for the multi-modal interaction, the encoded audio corresponding to the natural language utterances and the other data associated with the multi-modal interaction may be provided to the voice-enabled server 240 to maximize a likelihood of the voice-enabled server 240 correctly determining the intent of the multi-modal interaction (e.g., the ASR 220b may perform multi-pass speech recognition to generate an accurate transcription for the natural language utterance, the conversational language processor 230b may arbitrate among intent determinations performed in any number of different contexts or domains, etc.). Accordingly, in summary, the hybrid processing techniques performed in the environment illustrated in Figure 2 may generally include various different devices, which may or may not include natural language capabilities, cooperatively determining the intent of a particular multi-modal interaction and taking action to resolve the intent.

[060] Although the cooperative hybrid processing techniques described above have been particularly described in the context of an embedded processing architecture, such techniques are not necessarily limited to embedded processing architectures. In particular, the same techniques may be applied in any suitable voice services environment having various devices that can cooperate to initiate queries, tasks, commands, or other requests to resolve the intent of multi-modal interactions. Furthermore, in one implementation, the voice-enabled client device 210 may include a suitable amount of memory or other resources that can be dedicated

to natural language processing (e.g., the voice-enabled client device 210 may be a desktop computer or other device that can process natural language without substantially degraded performance). In such circumstances, one or more of the components of the voice-enabled client device 210 may be configured to optimize the on-board natural language processing in a manner that could otherwise cause bottlenecks, lengthy response times, or degradations in performance in an embedded architecture. For example, in one implementation, optimizing the on-board natural language processing may include configuring the ASR 220a to use a large list dictation grammar in addition to and/or instead of the virtual dictation grammar used in embedded processing architectures.

[061] Nonetheless, as will be described in greater detail below in connection with Figures 3-5, the cooperative hybrid processing techniques may be substantially similar regardless of whether the voice-enabled client device 210 has an embedded or non-embedded architecture. In particular, regardless of the architecture for the voice-enabled client device 210, cooperative hybrid processing may include the voice-enabled client device 210 optionally performing preliminary processing for a natural language multi-modal interaction and communicating input corresponding to the multi-modal interaction to the virtual router 260 for further processing through the messaging interface 250a. Alternatively (or additionally), the cooperative hybrid processing may include the virtual router 260 coordinating the further processing for the input among the various devices in the hybrid environment through messaging interface 250b, and subsequently returning any results of the processing to the voice-enabled client device 210 through messaging interface 250a.

[062] According to various aspects of the invention, Figure 3 illustrates a flow diagram of an exemplary method for initializing various devices that cooperate to perform hybrid processing in a natural language voice services environment. In particular, as noted above, the hybrid processing environment may generally include communication among various different devices that may cooperatively process natural language multi-modal interactions. For example, in one implementation, the various devices in the hybrid processing environment may include a virtual router having one or more messaging interfaces for communicating with one or more voice-enabled devices, one or more non-voice devices, and/or a voice-enabled server. As such, in one

implementation, the method illustrated in Figure 3 may be used to initialize communication in the hybrid processing environment to enable subsequent cooperative processing for one or more natural language multi-modal interactions received at any particular device in the hybrid processing environment.

[063] In one implementation, the various devices in the hybrid processing environment may be configured to continually listen or otherwise monitor respective input devices to determine whether a natural language multi-modal interaction has occurred. As such, the method illustrated in Figure 3 may be used to calibrate, synchronize, or otherwise initialize the various devices that continually listen for the natural language multi-modal interactions. For example, as described above in connection with Figure 2, the virtual router, the voice-enabled devices, the non-voice devices, the voice-enabled server, and/or other devices in the hybrid processing environment may be configured to provide various different capabilities or services, wherein the initialization method illustrated in Figure 3 may be used to ensure that the hybrid processing environment obtains a suitable signal to process any particular natural language multi-modal interaction and appropriately invoke one or more of the devices to cooperatively process the natural language multi-modal interaction. Furthermore, the method illustrated in Figure 3 and described herein may be invoked to register the various devices in the hybrid processing environment, register new devices added to the hybrid processing environment, publish domains, services, intent determination capabilities, and/or other features supported on the registered devices, synchronize local timing for the registered devices, and/or initialize any other suitable aspect of the devices in the hybrid processing environment.

[064] In one implementation, initializing the various devices in the hybrid processing environment may include an operation 310, wherein a device listener may be established for each of the devices in the hybrid processing environment. The device listeners established in operation 310 may generally include any suitable combination of instructions, firmware, or other routines that can be executed on the various devices to determine capabilities, features, supported domains, or other information associated with the devices. For example, in one implementation, the device listeners established in operation 310 may be configured to communicate with the respective devices using the Universal Plug and Play protocol designed

for ancillary computer devices, although it will be apparent that any appropriate mechanism for communicating with the various devices may be suitably substituted.

[065] In response to establishing the device listeners for each device registered in the hybrid processing environment (or in response to establishing device listeners for any device newly registered in the hybrid processing environment), the device listeners may then be synchronized in an operation 320. In particular, each of the registered devices may have an internal clock or other timing mechanism that indicates local timing for an incoming natural language multi-modal interaction, wherein operation 320 may be used to synchronize the device listeners established in operation 310 according to the internal clocks or timing mechanisms for the respective devices. Thus, in one implementation, synchronizing the device listeners in operation 320 may include each device listener publishing information relating to the internal clock or local timing for the respective device. For example, the device listeners may publish the information relating to the internal clock or local timing to the virtual router, whereby the virtual router may subsequently coordinate cooperative hybrid processing for natural language multi-modal interactions received at one or more of the devices in the hybrid processing environment. It will be apparent, however, that the information relating to the internal clock or local timing for the various devices in the hybrid processing environment may be further published to the other voice-enabled devices, the other non-voice devices, the voice-enabled server, and/or any other suitable device that may participate in cooperative processing for natural language multi-modal interactions provided to the hybrid processing environment.

[066] In one implementation, in response to establishing and synchronizing the device listeners for the various devices registered in the hybrid processing environment, the device listeners may continually listen or otherwise monitor respective devices on the respective registered devices in an operation 330 to detect information relating to one or more natural language multi-modal interactions. For example, the device listeners may be configured to detect occurrences of the natural language multi-modal interactions in response to detecting an incoming natural language utterance, a point of focus or attention focus selection associated with an incoming natural language utterance, and/or another interaction or sequence of interactions that relates to an incoming natural language multi-modal interaction. In addition,

operation 330 may further include the appropriate device listeners capturing the natural language utterance and/or related non-voice device interactions that relate to the natural language utterance.

[067] In one implementation, the captured natural language utterance and related non-voice device interactions may then be analyzed in an operation 340 to manage subsequent cooperative processing in the hybrid processing environment. In one implementation, for example, operation 340 may determine whether one device listener or multiple device listeners captured information relating to the natural language multi-modal interaction detected in operation 330. In particular, as noted above, the hybrid processing environment may generally include various different devices that cooperate to process natural language multi-modal interactions, whereby the information relating to the natural language multi-modal interaction may be provided to one or a plurality of the devices in the hybrid processing environment. As such, operation 340 may determine whether one device listener or multiple device listeners captured the information relating to the natural language multi-modal interaction in order to determine whether the hybrid processing environment needs to synchronize signals among various device listeners that captured information relating to the multi-modal interaction.

[068] For example, a user interacting with the hybrid processing environment may view a web page presented on a non-voice display device and provide a natural language multi-modal interaction that requests more information about purchasing a product displayed on the web page. The user may then select text on the web page containing the product name using a mouse, keyboard, or other non-voice input device and provide a natural language utterance to a microphone or other voice-enabled device such as "Is this available on Amazon.com?" In this example, a device listener associated with the non-voice display device may detect the text selection for the product name in operation 330, and a device listener associated with the voice-enabled device may further detect the natural language utterance inquiring about the availability of the product in operation 330. Furthermore, in one implementation, the user may be within a suitable range of multiple voice-enabled devices, which may result in multiple device listeners capturing different signals corresponding to the natural language utterance (e.g., the interaction may occur within range of a voice-enabled mobile phone, a voice-enabled

telematics device, and/or other voice-enabled devices, depending on the arrangement and configuration of the various devices in the hybrid processing environment).

[069] Accordingly, as will be described in greater detail herein, a sequence of operations that synchronizes different signals relating to the multi-modal interaction received at the multiple device listeners may be initiated in response to operation 340 determining that multiple device listeners captured information relating to the natural language multi-modal interaction. On the other hand, in response to operation 340 determining that only one device listener captured information relating to the natural language multi-modal interaction, the natural language multi-modal interaction may be processed in an operation 390 without executing the sequence of operations that synchronizes different signals (i.e., the one device listener provides all of the input information relating to the multi-modal interaction, such that hybrid processing for the interaction may be initiated in operation 390 without synchronizing different input signals). However, in one implementation, the sequence of synchronization operations may also be initiated in response to one device listener capturing a natural language utterance and one or more non-voice interactions to align different signals relating to the natural language multi-modal interaction, as described in greater detail herein.

[070] As described above, each device listener that receives an input relating to the natural language multi-modal interaction detected in operation 330 may have an internal clock or other local timing mechanism. As such, in response to determining that one or more device listeners captured different signals relating to the natural language multi-modal interaction in operation 340, the sequence of synchronization operations for the different signals may be initiated in an operation 350. In particular, operation 350 may include the one or more device listeners determining local timing information for the respective signals based on the internal clock or other local timing mechanism associated with the respective device listeners, wherein the local timing information determined for the respective signals may then be synchronized.

[071] For example, in one implementation, synchronizing the local timing information for the respective signals may be initiated in an operation 360. In particular, operation 360 may generally include notifying each device listener that received an input relating to the multi-modal interaction of the local timing information determined for each respective signal. For

example, in one implementation, each device listener may provide local timing information for a respective signal to the virtual router, and the virtual router may then provide the local timing information for all of the signals to each device listener. As such, in one implementation, operation 360 may result in each device listener receiving a notification that includes local timing information for each of the different signals that relate to the natural language multi-modal interaction detected in operation 330. Alternatively (or additionally), the virtual router may collect the local timing information for each of the different signals from each of the device listeners and further synchronize the local timing information for the different signals to enable hybrid processing for the natural language multi-modal interaction.

[072] In one implementation, any particular natural language multi-modal interaction may include at least a natural language utterance, and may further include one or more additional device interactions relating to the natural language utterance. As noted above, the utterance may generally be received prior to, contemporaneously with, or subsequent to the additional device interactions. As such, the local timing information for the different signals may be synchronized in an operation 370 to enable hybrid processing for the natural language multi-modal interaction. In particular, operation 370 may include aligning the local timing information for one or more signals corresponding to the natural language utterance and/or one or more signals corresponding to any additional device interactions that relate to the natural language utterance. In addition, operation 370 may further include aligning the local timing information for the natural language utterance signals with the signals corresponding to the additional device interactions.

[073] Thus, in matching the utterance signals and the non-voice device interaction signals, any devices that participate in the hybrid processing for the natural language multi-modal interaction may be provided with voice components and/or non-voice components that have been aligned with one another. For example, in one implementation, operation 370 may be executed on the virtual router, which may then provide the aligned timing information to any other device that may be invoked in the hybrid processing. Alternatively (or additionally), one or more of the other devices that participate in the hybrid processing may locally align the timing information (e.g., in response to the virtual router invoking the voice-enabled server in

the hybrid processing, resources associated with the voice-enabled server may be employed to align the timing information and preserve communication bandwidth at the virtual router).

[074] Furthermore, in one implementation, the virtual router and/or other devices in the hybrid processing environment may analyze the signals corresponding to the natural language utterance in an operation 380 to select the cleanest sample for further processing. In particular, as noted above, the virtual router may include a messaging interface for receiving an encoded audio sample corresponding to the natural language utterance from one or more of the voice-enabled devices. For example, the audio sample received at the virtual router may include the natural language utterance encoded in the MPEG-1 Audio Layer 3 (MP3) format or another lossy format to preserve communication bandwidth in the hybrid processing environment. However, it will be apparent that the audio sample may alternatively (or additionally) be encoded using the Free Lossless Audio Codec (FLAC) format or another lossless format in response to the hybrid processing environment having sufficient communication bandwidth for processing lossless audio that may provide a better sample of the natural language utterance.

[075] Regardless of whether the audio sample has been encoded in a lossy or lossless format, the signal corresponding to the natural language utterance that provides the cleanest sample may be selected in operation 380. For example, one voice-enabled device may be in a noisy environment or otherwise associated with conditions that interfere with generating a clean audio sample, while another voice-enabled device may include a microphone array or be configured to employ techniques that maximize fidelity of encoded speech. As such, in response to multiple signals corresponding to the natural language utterance being received in operation 330, the cleanest signal may be selected in operation 380 and hybrid processing for the natural language utterance may then be initiated in an operation 390.

[076] Accordingly, the synchronization and initialization techniques illustrated in Figure 3 and described herein may ensure that the hybrid processing environment synchronizes each of the signals corresponding to the natural language multi-modal interaction and generates an input for further processing in operation 390 most likely to result in a correct intent determination. Furthermore, in synchronizing the signals and selecting the cleanest audio sample for the

further processing in operation 390, the techniques illustrated in Figure 3 and described herein may ensure that none of the devices in the hybrid processing environment take action on a natural language multi-modal interaction until the appropriate signals to be used in operation 390 have been identified. As such, hybrid processing for the natural language multi-modal interaction may be initiated in operation 390, as described in further detail herein.

[077] According to one aspect of the invention, Figure 4 illustrates a flow diagram of am exemplary method for performing hybrid processing at one or more client devices in a natural language voice services environment. In particular, as will be described in greater below with reference to Figure 5, the one or more client devices may perform the hybrid processing in cooperation with a virtual router through a messaging interface that communicatively couples the client devices and the virtual router. For example, in one implementation, the messaging interface may generally include a light client (or thin client) that provides a mechanism for the client devices to transmit input relating to a natural language multi-modal interaction to the virtual router, and that further provides a mechanism for the client devices to receive output relating to the natural language multi-modal interaction from the virtual router.

[078] For example, in one implementation, the hybrid processing at the client devices may be initiated in response to one or more of the client devices receiving a natural language multi-modal interaction in an operation 410. In particular, the natural language multi-modal interaction may generally include a natural language utterance received at a microphone or other voice-enabled input device coupled to the client device that received the natural language multi-modal interaction, and may further include one or more other additional input modalities that relate to the natural language utterance (e.g., text selections, button presses, multi-touch gestures, etc.). As such, the natural language multi-modal interaction received in operation 410 may include one or more queries, commands, or other requests provided to the client device, wherein the hybrid processing for the natural language multi-modal interaction may then be initiated in an operation 420.

[079] As described in greater detail above, the natural language voice services environment may generally include one or more voice-enabled client devices, one or more non-voice devices, a voice-enabled server, and a virtual router arranged to communicate with each of the

voice-enabled client devices, the non-voice devices, and the voice-enabled server. In one implementation, the virtual router may therefore coordinate the hybrid processing for the natural language multi-modal interaction among the voice-enabled client devices, the non-voice devices, and the voice-enabled server. As such, the hybrid processing techniques described herein may generally refer to the virtual router coordinating cooperative processing for the natural language multi-modal interaction in a manner that involves resolving an intent of the natural language multi-modal interaction in multiple stages.

[080] In particular, as described above in connection with Figure 3, the various devices that cooperate to perform the hybrid processing may be initialized to enable the cooperative processing for the natural language multi-modal interaction. As such, in one implementation, in response to initializing the various devices, each of the client devices that received an input relating to the natural language multi-modal interaction may perform initial processing for the respective input in an operation 420. For example, in one implementation, a client device that received the natural language utterance included in the multi-modal interaction may perform initial processing in operation 420 that includes encoding an audio sample corresponding to the utterance, partially or completely transcribing the utterance, determining a preliminary intent for the utterance, or performing any other suitable preliminary processing for the utterance. In addition, the initial processing in operation 420 may also be performed at a client device that received one or more of the additional input modalities relating to the utterance. For example, the initial processing performed in operation 420 for the additional input modalities may include identifying selected text, selected points of focus or attention focus, or generating any other suitable data that can be used to further interpret the utterance. In one implementation, an operation 430 may then include determining whether the hybrid processing environment has been configured to automatically route inputs relating to the natural language multi-modal interaction to the virtual router.

[081] For example, in one implementation, operation 430 may determine that automatic routing has been configured to occur in response to multiple client devices receiving the natural language utterance included in the multi-modal interaction in operation 410. In this example, the initial processing performed in operation 420 may include the multiple client devices

encoding respective audio samples corresponding to the utterance, wherein messages that include the encoded audio samples may then be sent to the virtual router in an operation 460. The virtual router may then select one of the encoded audio samples that provides a cleanest signal and coordinate subsequent hybrid processing for the natural language multi-modal interaction, as will be described in greater detail below with reference to Figure 5. In another example, operation 430 may determine that automatic routing has been configured to occur in response to the initial processing resulting in a determination that the multi-modal interaction relates to a request that may be best suited for processing on the voice-enabled server (e.g., the request may relate to a location-based search query or another command or task that requires resources managed on the voice-enabled server, content, applications, domains, or other information that resides on one or more devices other than the client device that received the request, etc.). However, it will be apparent that the hybrid processing environment may be configured for automatic routing in response to other conditions and/or regardless of whether any attendant conditions exist, as appropriate.

[082] In one implementation, in response to the virtual router coordinating the hybrid processing for the natural language multi-modal interaction, the virtual router may provide results of the hybrid processing to the client device in an operation 470. For example, the results provided to the client device in operation 470 may include a final intent determination for the natural language multi-modal interaction, information requested in the interaction, data generated in response to executing a command or task requested in the interaction, and/or other results that enable the client device to complete processing for the natural language request in operation 480. For example, in one implementation, operation 480 may include the client device executing a query, command, task, or other request based on the final intent determination returned from the virtual router, presenting the requested information returned from the virtual router, confirming that the requested command or task has been executed, and/or performing any additional processing to resolve the natural language request.

[083] Referring back to operation 430, in response to determining that the conditions that trigger automatic routing have not been satisfied or that automatic router has otherwise not been configured, the client device may further process the natural language multi-modal

interaction in an operation 440. In one implementation, the further processing in operation 440 may include the client device attempting to determine an intent for the natural language multi-modal interaction using local natural language processing capabilities. For example, the client device may merge any non-voice input modalities included in the multi-modal interaction a transcription for the utterance included in the multi-modal interaction. The conversational language processor on the client device may then determine the intent for the multi-modal interaction utilizing local information relating to context, domains, shared knowledge, criteria values, or other information. The client device may then generate one or more interpretations for the utterance to determine the intent for the multi-modal interaction (e.g., identifying a conversation type, one or more requests contained in the interactions, etc.).

[084] In one implementation, operation 440 may further include determining a confidence level for the intent determination generated on the client device (e.g., the confidence level may be derived in response to whether the client devices includes a multi-pass speech recognition engine, whether the utterance contained any ambiguous words or phrases, whether the intent differs from one context to another, etc.). In one implementation, an operation 450 may then determine whether or not to invoke off-board processing depending on the confidence level determined in operation 440. For example, operation 450 may generally include determining whether the intent determined in operation 440 satisfies a particular threshold value that indicates an acceptable confidence level for taking action on the determined intent. As such, in response to the confidence level for the intent determination satisfying the threshold value, operation 450 may determine to not invoke off-board processing. In particular, the confidence level satisfying the threshold value may indicate that the client device has sufficient information to take action on the determined intent, whereby the client device may then process one or more queries, commands, tasks, or other requests to resolve the multi-modal interaction in operation 480.

[085] Alternatively, in response to the confidence level for the intent determination failing to satisfy the threshold value, operation 450 may invoke off-board processing, which may include sending one or more messages to the virtual router in operation 460. The one or more messages may cause the virtual router to invoke additional hybrid processing for the multi-

modal interaction in a similar manner as noted above, and as will be described in greater detail herein with reference to Figure 5.

[086] According to one aspect of the invention, Figure 5 illustrates a flow diagram of am exemplary method for performing hybrid processing at a virtual router in a natural language voice services environment. In particular, as the virtual router may coordinate the hybrid processing for natural language multi-modal interactions received at one or more client devices. In one implementation, in an operation 510, the virtual router may receive one or more messages relating to a natural language multi-modal interaction received at one or more of the client devices in the voice services environment. For example, the virtual router may include a messaging interface that communicatively couples the virtual router to the client devices and a voice-enabled server, wherein the messaging interface may generally include a light client (or thin client) that provides a mechanism for the virtual router to receive input from one or more the client devices and/or the voice-enabled server, and further to transmit output to one or more the client devices and/or the voice-enabled server. The messages received in operation 510 may generally include any suitable processing results for the multi-modal interactions, whereby the virtual router may coordinate hybrid processing in a manner that includes multiple processing stages that may occur at the virtual router, one or more of the client devices, the voice-enabled server, or any suitable combination thereof.

[087] In one implementation, the virtual router may analyze the messages received in operation 510 to determine whether to invoke the hybrid processing in a peer-to-peer mode. For example, one or more of the messages may include a preliminary intent determination that the virtual router can use to determine whether to invoke one or more of the client devices, the voice-enabled server, or various combinations thereof in order to execute one or more of the multiple processing stages for the multi-modal interaction. In another example, one or more of the messages may include an encoded audio sample that the virtual router forwards to one or more of the various devices in the hybrid processing environment. As such, in one implementation, the virtual router may analyze the messages received in operation 510 to determine whether or not to invoke the voice-enabled server to process the multi-modal interaction (e.g., the messages may include a preliminary intent determination that indicates

that the multi-modal interaction includes a location-based request that requires resources residing on the server).

[088] In response to the virtual router determining to invoke the voice-enabled server, the virtual router may forward the messages to the server in an operation 530. In particular, the messages forwarded to the server may generally include the encoded audio corresponding to the natural language utterance and any additional information relating to other input modalities relevant to the utterance. For example, as described in greater detail above with reference to Figure 2, the voice-enabled server may include various natural language processing components that can suitably determine the intent of the multi-modal interaction, whereby the messages sent to the voice-enabled server may include the encoded audio in order to permit the voice-enabled server to determine the intent independently of any preliminary processing on the client devices that may be inaccurate or incomplete. In response to the voice-enabled server processing the messages received from the virtual router, results of the processing may then be returned to the virtual router in an operation 570. For example, the results may include the intent determination for the natural language multi-modal interaction, results of any queries, commands, tasks, or other requests performed in response to the determined intent, or any other suitable results, as will be apparent.

[089] Alternatively, in response to the virtual router determining to invoke the peer-to-peer mode in operation 520, the virtual router may coordinate the hybrid processing among one or more the client devices, the voice-enabled server, or any suitable combination thereof. For example, in one implementation, the virtual router may determine a context for the natural language multi-modal interaction in an operation 540 and select one or more peer devices based on the determined context in an operation 550. For example, one or more of the client devices may be configured to provide content and/or services in the determined context, whereby the virtual router may forward one or more messages to such devices in an operation 560 in order to request such content and/or services. In another example, the multi-modal interaction may include a compound request that relates to multiple contexts supported on different devices, whereby the virtual router may forward messages to each such device in operation 560 in order to request appropriate content and/or services in the different contexts.

[090] In still another example, the interaction may include a request to be processed on the voice-enabled server, yet the request may require content and/or services that reside on one or more of the client devices (e.g., a location-based query relating to an entry in an address book on one or more of the client devices). As such, the virtual router may generally forward various messages to the selected peer devices in operation 560 to manage the multiple stages in the hybrid processing techniques described herein. For example, the virtual router may send messages to one or more voice-enabled client devices that have intent determination capabilities in a particular context, one or more non-voice client devices that have access to content, services, and/or other resources needed to process the multi-modal interaction, or any appropriate combination thereof. The virtual router may therefore send messages to the client devices and/or the voice-enabled server in operation 560 and receive responsive messages from the client devices and/or the voice-enabled server in operation 570 in any appropriate manner (e.g., in parallel, sequentially, iteratively, etc.). The virtual router may then collate the results received in the responsive messages in operation 580 and return the results to one or more of the client devices for any final processing and/or presentation of the results.

[091] According to one aspect of the invention, Figure 6 illustrates a block diagram of an exemplary system for providing a natural language content dedication service. In particular, as shown in Figure 6, the system for providing the natural language content dedication service may generally include a voice-enabled client device 610 that can communicate with a content dedication system 665 through a messaging interface 665. In one implementation, the voice-enabled client device 610 may communicate with the content dedication system 665 in a similar manner as described above with reference to Figures 2 through 5. For example, the content dedication system 665 may include a virtual router 660 and a voice-enabled server 640 that can service multi-modal natural language requests provided to the voice-enabled client device 610 in a similar manner as described above, and may further include a billing system 638 that may be used to process transactions relating to the content dedication service. Further, although Figure 6 illustrates the content dedication system 665 as including the virtual router 660, the voice-enabled server 640, and the billing system 638 within one component, such illustration will be understood to be for ease of description only, and that the virtual router 660,

the voice-enabled server 640, and the billing system 638 may in fact be arranged within any number of components that can suitably communicate with one another to process multi-modal natural language requests relating to the content dedication service (e.g., the virtual router 660 may communicate with the voice-enabled server 640 through another messaging interface distinct from the messaging interface 650 for communicating with the voice-enabled client device 610, as shown in the exemplary system illustrated in Figure 2).

[092] In one implementation, the system shown in Figure 6 may provide the natural language content dedication service to any suitable voice-enabled client device 610 having a suitable combination of input and output devices 615a that can receive natural language multi-modal interactions and provide responses to the natural language multi-modal interactions, wherein the input and output devices 615a may be further arranged to receive any other suitable type of input and provide any other suitable type of output. For example, in one implementation, the voice-enabled client device 610 may comprise a mobile phone that includes a keypad input device 615a, a touch screen input device 615a, or other input mechanisms 615a in addition to any input microphones or other suitable input devices 615a that can receive voice signals. As such, in this example, the mobile phone may further include an output display device 615a in addition to any output microphones or other suitable output devices 615a that can output audible signals. Thus, a user of the voice-enabled client device 610 may be listening to music, watching video, or otherwise interacting with content through the input and output devices 615a and provide a multi-modal natural language request to engage in a transaction to dedicate the music, video, or other content, as will be described in greater detail below. Furthermore, in one implementation, the voice-enabled client device 610 may be included within a hybrid processing environment that may include a plurality of different devices, whereby the content dedication request may relate to content played on a different device from the voice-enabled client device 610, although the content dedication request may relate to any suitable content (i.e., the request need not necessarily relate to played content, as users may provide natural language to request content dedications for any suitable content).

[093] Thus, in one implementation, in response to the voice-enabled client device 710 receiving a multi-modal interaction that includes a natural language utterance, the voice-

enabled client device 610 may invoke an Automatic Speech Recognizer (ASR) 620a to generate one or more preliminary interpretations of the utterance. The ASR 620a may then provide the preliminary interpretations to a conversational language processor 630a, which may attempt to determine an intent for the multi-modal interaction. In one implementation, to determine the intent for the multi-modal interaction, the conversational language processor 630a may determine a most likely context for the interaction from the preliminary interpretations of the utterance, any accompanying non-speech inputs in the multi-modal interaction that relate to the utterance, contexts associated with prior requests, short-term and long-term shared knowledge, or any other suitable information for interpreting the multi-modal interaction. Thus, in response to the conversational language processor 630a determining that the intent for the multi-modal interaction relates to a content dedication request, a content dedication application 634a may be invoked to resolve the content dedication request.

[094] For example, in one implementation, an initial multi-modal interaction may include the utterance "Find 'Superstylin'' by Groove Armada." In response to the initial multi-modal interaction, the ASR 620a may generate a preliminary interpretation that includes the words "Find" and "Superstylin'" and the phrase "Groove Armada." The ASR 620a may then provide the preliminary interpretation to the conversational language processor 630a, which may determine that the word "Find" indicates that the most likely intent of the interaction includes a search, while the word "Superstylin'" and the phrase "Groove Armada" provide criteria for the search. Furthermore, in response to determining the most likely intent of the interaction, the conversational language processor 630a may establish a music context for the interaction and attempt to resolve the search request. For example, the conversational language processor 630a may search one or more data repositories 636a that contain music information to identify music having a song title of "Superstylin'" and an artist name of "Groove Armada," and the conversational language processor 630a may further cooperate with other devices in the hybrid processing environment to search for the song (e.g., in response to the local data repositories 636a not yielding adequate results, another device in the environment having a larger music data repository than the client device 610 may be invoked). The conversational language processor 630a may then receive appropriate results for the search and present the

results to the user through the output device 615a (e.g., displaying information about the song, playing a sample audio clip of the song, displaying options to purchase the song, recommending similar songs or similar artists, etc.).

[095] Continuing with the above example, a subsequent multi-modal interaction may include the utterance "Share this with my wife," "Dedicate it to Charlene," "That's the one, I want to pass along to some friends," or another suitable utterance that reflects a request to dedicate the content. Alternatively (or additionally), in response to the intent of the initial interaction including a request to search for content, the conversational language processor 630a may invoke the content dedication application 634a, which may present an option to dedicate the content through the output device 615a together with the results of the search. As such, the request to dedicate the content may also be provided in a non-speech input, such as a button press or touch screen selection of the option to dedicate the content. Accordingly, in response to detecting a content dedication request (e.g., through the ASR 620a and the conversational language processor 630a processing a suitable utterance, the input device 615a receiving a suitable non-speech input, or any suitable combination thereof), the content dedication application 634a may be invoked to process the content dedication request.

[096] In one implementation, to process the content dedication request, the content dedication application 634a may capture a natural language utterance that contains the dedication. For example, the content dedication application 634a may provide a prompt through the output device 615a that instructs the user to provide the dedication utterance (e.g., a visual or audible prompt instructing the user to begin speaking, to speak after an audible beep, etc.). The user may then provide the dedication utterance through the voice-enabled input device 615a, and the dedication utterance may then be converted into an electronic signal that the content dedication application 634a captures for the dedication (e.g., "Dear Charlene, I was listening to this song and I thought of you. Enjoy!"). In addition, the content dedication application 634a may further prompt the user to provide any additional tags for the dedicated content (e.g., an image or a picture to be inserted as album art for the dedicated content, a natural language utterance that includes information to be inserted as voice-tags for the dedicated content, a non-speech input or data input that includes information to insert in

tags for the dedicated content, etc.). The content dedication application 634a may then prompt the user to identify a recipient 690 of the dedication, wherein the user may provide any suitable multi-modal input that includes an e-mail address, a telephone number, an address book entry, or other information identifying the recipient 690 of the dedication.

[097] In one implementation, the content dedication application 634a may then route the request, including the dedication utterance, the additional tags (if any), and the information identifying the recipient 690 to the content dedication system 665 through the messaging interface 650. For example, the dedication utterance may be converted into encoded audio that can be communicated through the messaging interface 650, whereby the content dedication system 665 can insert the encoded audio corresponding to the dedication utterance within the dedicated content and/or verbally annotate the dedicated content with the encoded audio. Alternatively (or additionally), the dedication utterance may be interpreted and parsed to transcribe one or more words or phrases from the dedication utterance, wherein the transcribed words or phrases may provide a textual annotation for the dedicated content (e.g., the textual annotation may be inserted within metadata for the dedicated content, such as an ID3 Comments tag). Similarly, any utterances to be inserted as voice-tags for the dedicated content may provide further verbal annotations for the dedicated content, or such utterances may be transcribed to provide further textual annotations for the dedicated content. In one implementation, verbal annotations, textual annotations, and other types of annotations may be created and associated with the dedicated content using techniques described in co-pending U.S. Patent Application Serial No. 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," filed August 29, 2005, the contents of which are hereby incorporated by reference in their entirety.

[098] In one implementation, in response to the content dedication system 665 receiving the content dedication request, the dedication utterance, and any additional tags from the content dedication application 634a, the content dedication system 665 may invoke a similar content dedication application 634b on the voice-enabled server 640 to process the request. In particular, the content dedication application 634b on the voice-enabled server 640 may identify the content requested for dedication and initiate a transaction for the content

requested for dedication. For example, in one implementation, if the content dedication application 634a on the voice-enabled client device 610 was able to identify the requested content, the content dedication request received at the content dedication system 665 may include an identification of the content to be dedicated. Alternatively, because the content dedication system 665 can cooperate in resolving the multi-modal interactions involved in the dedication request, the content dedication system 665 may use shared knowledge relating to the dedication request to identify the content to be dedicated. For example, the content dedication system 665 may invoke one or more local natural language processing components (e.g., ASR 620b, conversational language processor 630b or 630c, etc.), search one or more local data repositories 636b, interact with one or more content providers 680 over a network 670, pull data from a satellite radio system that played the requested content at the voice-enabled client device 610 or another device in the hybrid processing environment, or otherwise consult available resources that can be used to identify the content to be dedicated.

[099] In one implementation, in response to identifying the content to be dedicated, the content dedication system 665 may communicate with a billing system 638 to identify one or more purchase options for the dedication request and process an appropriate transaction for the dedication request. In particular, the content dedication system 665 may generally support various different purchase options to provide users with flexibility in the manner of requesting content dedications, including a buy-to-own purchase option, a pay-to-play purchase option, a paid subscription purchase option, or other appropriate options. In one implementation, the purchase options may be modeled on techniques for providing natural language services and subscriptions described in U.S. Patent Application Serial No. 10/452,147, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," which issued as U.S. Patent No. 7,398,209 on July 8, 2008, and co-pending U.S. Patent Application Serial No. 10/618,633, entitled "Mobile Systems and Methods for Responding to Natural Language Speech Utterance," filed June 15, 2003, the contents of which are hereby incorporated by reference in their entirety. For example, in the buy-to-own purchase option, the content dedication system 665 may purchase full rights to the content from an appropriate content provider 680, wherein the billing system 638 may then charge the user of the voice-enabled client device 610 a particular

amount that encompasses the cost for purchasing the content from the content provider 680 plus a service charge for dedicating the content, tagging the dedicated content, delivering the dedicated content to the recipient 690, or any other appropriate services rendered for the content dedication request. The billing system 638 may charge the user in a similar manner under the pay-to-play purchase option, except that the rights purchased from the content provider 680 may be limited to a single play, such that the cost for purchasing the content from the content provider 680 may be somewhat less under the pay-to-play purchase option.

[0100] Under the paid subscription purchase option, however, the user may pay a periodic service charge to the content dedication system 665 that permits the user to make a predetermined number of content dedications or an unlimited number of content dedications in a subscription period, or the subscription purchase option may permit the user to make content dedications in any other suitable manner (e.g., different subscription levels having different content dedication options may be offered, such that the user may select a subscription level that meets the user's particular needs). For example, under the paid subscription purchase option, the billing system 638 may only charge the user any costs for obtaining the rights to the content, which may be purchased under either the buy-to-own option or the pay-to-play option, or alternatively the user may be charged nothing if the user already owns the dedication content. In another example, a first subscription level may cost a first amount to permit the user to make a particular number of content dedications in a subscription period, while a second subscription level may cost a higher amount to permit the user to make an unlimited number of content dedications in the subscription period, while still other subscription levels having different terms may be offered, as will be apparent.

[0101] Furthermore, in one implementation, a service provider associated with the content dedication system 665 may negotiate an agreement with the content provider 680 to determine the manner in which revenues for content transactions will be shared between the content dedication system 665 and the content provider 680. For example, the agreement may provide that the content provider 680 may keep all of the revenue for transactions that include purchasing content from the content provider 680 and that the service provider associated with the content dedication system 665 may recoup costs for such transactions from users. In

another example, the agreement may provide that the content provider 680 and the service provider associated with the content dedication system 665 may share the revenue for the transactions that include purchasing content from the content provider 680. Thus, the agreement may generally include any suitable arrangement that defines the manner in which the content provider 680 and the service provider associated with the content dedication system 665 manage the revenue for the transactions that include purchasing content from the content provider 680, while the service provider associated with the content dedication system 665 may manage billing users for the natural language aspects of the content dedication service according to the techniques described in further detail above.

[0102] According to one aspect of the invention, Figure 7 illustrates a flow diagram of an exemplary method for providing a natural language content dedication service. In particular, the method for providing a natural language content dedication service, as shown in Figure 7, may be performed on a voice-enabled client device that can communicate with a content dedication system through a messaging interface, wherein the voice-enabled client device may communicate with the content dedication system in a similar manner as described above with reference to Figures 2 through 6. For example, the content dedication system may include a virtual router and a voice-enabled server that can service multi-modal natural language requests provided to the voice-enabled client device, and may further include a billing system for processing transactions relating to the content dedication service.

[0103] In one implementation, the method shown in Figure 7 may be used to provide the natural language content dedication service to any suitable voice-enabled client device having a suitable combination of input and output devices that can receive natural language multi-modal interactions and provide responses to the natural language multi-modal interactions, wherein the input and output devices may be further arranged to receive any other suitable type of input and provide any other suitable type of output. For example, in one implementation, the voice-enabled client device may comprise a mobile phone that includes a keypad input device, a touch screen input device, or other input mechanisms in addition to any input microphones or other suitable input devices that can receive voice signals, and may further include an output display device in addition to any output microphones or other suitable output devices that can

output audible signals. Thus, a user of the voice-enabled client device may be listening to music, watching video, or otherwise interacting with content through the input and output devices, wherein an operation 710 may include the voice-enabled client device receiving a multi-modal natural language interaction to engage in a transaction to dedicate the music, video, or other content.

[0104] Thus, in one implementation, in response to the voice-enabled client device receiving the multi-modal interaction in operation 710 that includes a natural language utterance, the voice-enabled client device may invoke an Automatic Speech Recognizer (ASR) to generate one or more preliminary interpretations of the utterance. The ASR may then provide the preliminary interpretations to a conversational language processor, which may attempt to determine an intent for the multi-modal interaction. In one implementation, to determine the intent for the multi-modal interaction, the conversational language processor may determine a most likely context for the interaction from the preliminary interpretations of the utterance, any accompanying non-speech inputs in the multi-modal interaction that relate to the utterance, contexts associated with prior requests, short-term and long-term shared knowledge, or any other suitable information for interpreting the multi-modal interaction. Thus, in one implementation, an operation 720 may include the conversational language processor detecting a content dedication request in response to determining that the intent for the multi-modal interaction relates to a content dedication request. The conversational language processor may then invoke a content dedication application to resolve the content dedication request.

[0105] For example, in one implementation, the multi-modal interaction received in operation 720 may include the utterance "Buy that song and send it to Michael." In response to the initial multi-modal interaction, the ASR may generate a preliminary interpretation that includes words and/or phrases such as "Buy," "that song," "send it," and "to Michael." The ASR may then provide the preliminary interpretation to the conversational language processor, which may determine that the word and/or phrase combination of "Buy" and "send it" indicates that the most likely intent of the interaction includes a content dedication request, while the word and/or phrase combination of "that song" and "to Michael" provide criteria for the intended

content and recipient for the dedication. Furthermore, in response to determining the most likely intent of the interaction, the conversational language processor may establish a device context, a dedication context, a content or music context, an address book context, or other suitable contexts in an attempt to resolve the request.

[0106] For example, the device context may enable the conversational language processor to retrieve data from the voice-enabled client device or another suitable device that provides the user's intended meaning for the phrase "that song" (e.g., the user may be referring to a song playing on the user's satellite radio device, such that the conversational language processor can identify the song that was playing when the device interaction was received in operation 710). Furthermore, the address book context may enable the conversational language processor to identify "Michael," the intended recipient of the dedication request. The conversational language processor may then receive appropriate results for resolving the intent of the request and present the results to the user through the output device (e.g., displaying information requesting that the user confirm that the content and recipient was correctly identified, playing a sample audio clip of the song, displaying options to purchase the song, recommending similar songs or similar artists, etc.). Accordingly, in response to detecting the content dedication request in operation 720 and identifying the relevant criteria identifying the content to be dedicated and the intended recipient, the content dedication application may be invoked to process the content dedication request.

[0107] For example, in one implementation, processing the content dedication request may include the content dedication application capturing a natural language utterance that contains the dedication for the requested content in an operation 730. For example, the content dedication application may provide a prompt through the output device that instructs the user to provide the dedication utterance (e.g., a visual or audible prompt instructing the user to begin speaking, to speak after an audible beep, etc.). The user may then provide the dedication utterance through the voice-enabled input device, and the dedication utterance may then be converted into an electronic signal that the content dedication application captures in operation 730. In addition, an operation 740 may include the content dedication application further prompting the user to provide any additional tags for the dedicated content (e.g., an

image or a picture to be inserted as album art for the dedicated content, one or more natural language utterances to insert as voice-tags for the dedicated content, one or more natural language utterances to be transcribed into text that can be inserted in tags for the dedicated content, a non-speech input or data input that include information to insert in tags for the dedicated content, etc.).

[0108] Thus, in response to determining that the user has provided additional information to insert in tags for the dedication content in operation 740, the content dedication application may then capture the tags in an operation 750. For example, operation 750 may include the content dedication application capturing an image or picture that the user identifies for album art to be inserted in the dedicated content, capturing any natural language utterances to insert as voice-tags for the dedicated content, communicating with the ASR and/or conversational language processor to transcribe any natural language utterances to be inserted as text within tags for the dedicated content, capturing any non-speech inputs or data inputs that include information to insert as text within the tags for the dedicated content, or otherwise capturing information that relates to the additional tags.

[0109] In one implementation, the content dedication application may then process the content dedication request in an operation 760, which may include prompting the user to identify the recipient of the dedication. For example, the content dedication application may request information identifying the recipient of the dedication in response to the user not having already identified the recipient, in response to the user identifying the recipient in a manner that includes ambiguity or other criteria that cannot be resolved without further information, to distinguish among different contact information known for the recipient, or in response to other appropriate circumstances. Thus, the user may provide any suitable multi-modal input that includes an e-mail address, a telephone number, an address book entry, or other criteria that can be used to uniquely identify the information for contacting the recipient of the dedication. Furthermore, in one implementation, processing the content dedication request in operation 760 may further include the content dedication application routing the request, including the dedication utterance, the additional tags (if any), and the information for contacting the recipient to the content dedication system through the messaging interface,

wherein the content dedication system may then process a transaction for the content dedication request, as will be described in greater detail below.

[0110] According to one aspect of the invention, Figure 8 illustrates a flow diagram of an exemplary method for providing a natural language content dedication service. In particular, the method for providing a natural language content dedication service, as shown in Figure 8, may include an operation 810 in which a content dedication system may receive a natural language content dedication request through a messaging interface. For example, in one implementation, the content dedication system may receive the natural language content dedication request from a content dedication application that executes on a voice-enabled client device. In one implementation, the natural language content dedication request received in operation 810 may generally include a natural language dedication utterance, information to be inserted within one or more tags for content to be dedicated, and information identifying a recipient of the dedicated content.

[0111] For example, in one implementation, the dedication utterance received in operation 810 may include encoded audio that the content dedication system can insert within the dedicated content or use to verbally annotate the dedicated content. Alternatively (or additionally), the content dedication system may interpret and parse the dedication utterance to transcribe one or more words or phrases from the dedication utterance, wherein the transcribed words or phrases may provide a textual annotation for the dedicated content (e.g., the textual annotation may be inserted within metadata for the dedicated content, such as an ID3 Comments tag). Similarly, any utterances to be inserted as voice-tags for the dedicated content may provide further verbal annotations for the dedicated content, or such utterances may be transcribed to provide further textual annotations for the dedicated content. In one implementation, verbal annotations, textual annotations, and other types of annotations may be created and associated with the dedicated content using techniques described in co-pending U.S. Patent Application Serial No. 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," filed August 29, 2005, the contents of which are hereby incorporated by reference in their entirety.

[0112] In one implementation, in response to the content dedication system receiving the content dedication request, the dedication utterance, and any additional tags from the content dedication application in operation 810, the content dedication system may invoke a content dedication application on a voice-enabled server to process the content dedication request. In particular, an operation 820 may include the content dedication application on the voice-enabled server identifying the content requested for dedication. In one implementation, the request received from the content dedication application in operation 810 may include information identifying the requested content and/or information that the content dedication system can use to identify the requested content. For example, the request may include a multi-modal natural language input that includes a natural language utterance and/or a non-voice input, wherein the content dedication system may use shared knowledge relating to the dedication request to identify the content to be dedicated.

[0113] Thus, operation 820 may include the content dedication system invoking one or more natural language processing components (e.g., an ASR, a conversational language processor, etc.), searching one or more data repositories, interacting with one or more content providers, pulling data from a satellite radio system that played the requested content at the voice-enabled client device or another device in a hybrid processing environment, or otherwise consulting available resources that can be used to identify the content to be dedicated. For example, in one implementation, the hybrid processing environment may include a device having an application that can identify played content (e.g., a Shazam® listening device that a user can hold near a speaker to identify content playing through the speaker). Thus, operation 820 may generally include the content dedication system communicating with any suitable device, system, application, or other resource to identify the content requested for dedication.

[0114] In one implementation, in response to identifying the content to be dedicated, an operation 830 may include the content dedication system identifying one or more purchase options for the dedication request. In one implementation, operation 830 may include the content dedication system communicating with a billing system that supports various purchase options to provide users with flexibility in the manner of requesting content dedications. For example, the billing system may support content dedication purchase options that include a

buy-to-own purchase option, a pay-to-play purchase option, a paid subscription purchase option, or other appropriate options. In response to identifying the purchase option for the content dedication request, an operation 840 may include the content dedication system processing a transaction for the content identified in operation 820.

[0115] For example, in response to operation 830 indicating that the user has selected the buy-to-own purchase option, the transaction processed in operation 840 may include purchasing full rights to the requested content from an appropriate content provider, wherein the billing system may then charge the user of the voice-enabled client device an appropriate amount that includes costs for purchasing the content from the content provider, adding the natural language utterance dedicating the content, tagging the dedicated content, delivering the dedicated content to the recipient, or any other appropriate services rendered for the content dedication request. Alternatively, in response to operation 830 indicating that the user has selected the pay-to-play purchase option, the transaction processed in operation 840 may include purchasing rights from the content provider that permits the purchased content to be played a predetermined number of times, wherein the billing system may charge the user in a similar manner as described above, except that the cost for purchasing limited rights to the content under the pay-to-play purchase option may be somewhat less than the cost for purchasing ownership rights to the content, as in the buy-to-own purchase option.

[0116] In another alternative implementation, in response to determining that the requested content dedication includes a selection of the paid subscription purchase option, the user may pay a periodic service charge to the content dedication system that permits the user to make a predetermined number of content dedications in a subscription period, an unlimited number of content dedications in the subscription period, or otherwise make content dedications in accordance with terms defined in a subscription (e.g., different subscription levels having different content dedication options may be offered, such as a subscription level that only permits utterance dedications, another subscription level that further permits interpreting and parsing utterance dedications, etc.). Thus, the content transaction may be processed in operation 840 according to the purchase options identified in operation 830, and the content

dedication system may then further process the dedication request to customize the dedicated content according to criteria provided in the request previously received in operation 810.

[0117] For example, in one implementation, an operation 850 may include the content dedication system inserting the natural language dedication utterance into the dedicated content, verbally annotating the dedicated content with the dedication utterance, or otherwise associating the dedicated content with the dedication utterance. Furthermore, an operation 860 may include the content dedication system determining whether any additional tags have been specified for the dedicated content. For example, as noted above, the dedication request may identify an image or picture to insert as album art for the dedicated content, one or more natural language utterances, non-voice, and/or data inputs to be transcribed into text that can be inserted in tags for the dedicated content, or any other suitable information that can be inserted within or associated with metadata for the dedicated content. Thus, in response to determining that information for any additional tags has been provided, an operation 870 may include inserting such additional tags into the dedicated content.

[0118] In one implementation, in response to having purchased the content requested for dedication, associating the dedication utterance with the dedicated content, and associating the additional tags (if any) with the dedicated content, the content dedication system may send a content dedication message to the recipient of the dedication in an operation 880. For example, the content dedication message may include a Short Message Service (SMS) text message, an electronic mail message, an automated telephone call managed by a text-to-speech engine, or any other appropriate message that can be appropriately delivered to the recipient (e.g., the message may include a link that the recipient can select to stream, download, or otherwise access the dedicated content, the dedication utterance, etc.). Thus, the content dedication message may generally notify the recipient that content has been dedicated to the recipient and provide various mechanisms for the recipient to access the content dedication, as will be apparent.

[0119] 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.

[0120] 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 system for providing a natural language content dedication service, comprising:

an electronic device configured to:

receive a multi-modal interaction that includes a request to dedicate content, wherein the multi-modal interaction includes a natural language utterance relating to the request to dedicate the content; and

capture an utterance dedicating the content for a recipient; and

a content dedication system in communication with the electronic device through a message interface, wherein the content dedication system is configured to:

identify the content to dedicate for the recipient based on an interpretation of the natural language utterance included in the multi-modal interaction;

process a transaction for the identified content to dedicate for the recipient; and

send a content dedication message to the recipient, wherein the content dedication message includes information for the recipient to access the content, and wherein the content dedication message further includes information for the recipient to access the utterance dedicating the content for the recipient.

2. The system of claim 1, wherein the content dedication system is further configured to receive one or more messages from the electronic device through the messaging interface, wherein one or more messages contain information relating to the multi-modal interaction that includes the request to dedicate the content, and wherein the one or more messages further contain information corresponding to the utterance dedicating the content for the recipient.

3. The system of claim 2, wherein the electronic device is further configured to receive information corresponding to one or more tags dedicating the content for the recipient.

4. The system of claim 3, wherein the content dedication system is further configured to:

insert the utterance dedicating the content for the recipient within the content; and

insert the tags dedicating the content for the recipient within metadata for the content.

5. The system of claim 1, wherein the tags dedicating the content for the recipient include one or more of an image or a picture to insert within album art metadata for the content.

6. The system of claim 1, wherein the content dedication system is further configured to annotate the content with the utterance dedicating the content for the recipient.

7. The system of claim 1, wherein the content dedication system is further configured to purchase the identified content from a content provider to process the transaction for the identified content.

8. The system of claim 7, wherein the content dedication system is further configured to:

identify a purchase option for the request to dedicate the content, wherein the purchase option includes one or more of a buy-to-own purchase option, a pay-to-pay purchase option, or a paid subscription purchase option; and

bill a user of the electronic device for the natural content dedication service based on identified purchase option.

9. The system of claim 1, wherein the content dedication message sent to the recipient includes an electronic link for the recipient to access the content and the utterance dedicating the content for the recipient.

10. The system of claim 1, wherein the content includes one or more of a song or a video.

11. A method for providing a natural language content dedication service, comprising:

receiving, at an electronic device, a multi-modal interaction that includes a request to dedicate content, wherein the multi-modal interaction includes a natural language utterance relating to the request to dedicate the content;

capturing an utterance dedicating the content for a recipient;

identifying the content to dedicate for the recipient based on an interpretation of the natural language utterance included in the multi-modal interaction;

processing, at a content dedication system, a transaction for the identified content to dedicate for the recipient; and

sending a content dedication message from the content dedication system to the recipient, wherein the content dedication message includes information for the recipient to access the content, and wherein the content dedication message further includes information for the recipient to access the utterance dedicating the content for the recipient.

12. The method of claim 11, further comprising receiving, at the content dedication system, one or more messages from the electronic device through a messaging interface, wherein one or more messages contain information relating to the multi-modal interaction that includes the request to dedicate the content, and wherein the one or more messages further contain information corresponding to the utterance dedicating the content for the recipient.

13. The method of claim 12, further comprising receiving, at the electronic device, information corresponding to one or more tags dedicating the content for the recipient.

14. The method of claim 13, further comprising:

inserting the utterance dedicating the content for the recipient within the content; and

inserting the tags dedicating the content for the recipient within metadata for the content.

15. The method of claim 11, wherein the tags dedicating the content for the recipient include one or more of an image or a picture to insert within album art metadata for the content.

16. The method of claim 11, further comprising annotating the content with the utterance dedicating the content for the recipient.

17. The method of claim 11, further comprising purchasing, by the content dedication system, the identified content from a content provider to process the transaction for the identified content.

18. The method of claim 17, further comprising:

identifying, at the content dedication system, a purchase option for the request to dedicate the content, wherein the purchase option includes one or more of a buy-to-own purchase option, a pay-to-pay purchase option, or a paid subscription purchase option; and

billing, by the content dedication system, a user of the electronic device for the natural content dedication service based on identified purchase option.

19. The method of claim 11, wherein the content dedication message sent to the recipient includes an electronic link for the recipient to access the content and the utterance dedicating the content for the recipient.

20. The method of claim 11, wherein the content includes one or more of a song or a video.

ABSTRACT

The system and method described herein may provide a natural language content dedication service in a voice services environment. In particular, providing the natural language content dedication service may generally include detecting multi-modal device interactions that include requests to dedicate content, identifying the content requested for dedication from natural language utterances included in the multi-modal device interactions, processing transactions for the content requested for dedication, processing natural language to customize the content for recipients of the dedications, and delivering the customized content to the recipients of the dedications.

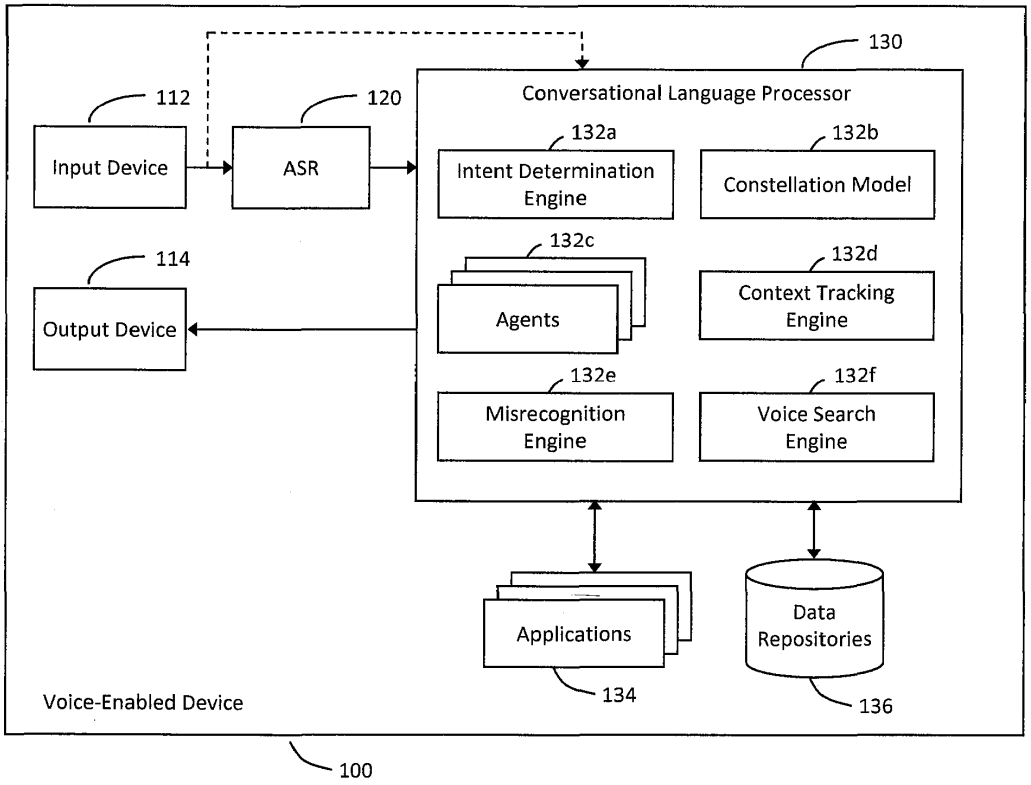

130
Conversational Language Processor
112
Input Device
120
ASR
132a
Intent Determination Engine
132b
Constellation Model
132c
Agents
114
Output Device
132d
Context Tracking Engine
132e
Misrecognition Engine
132f
Voice Search Engine
Applications
134
Data Repositories
136
Voice-Enabled Device
100

Figure 1

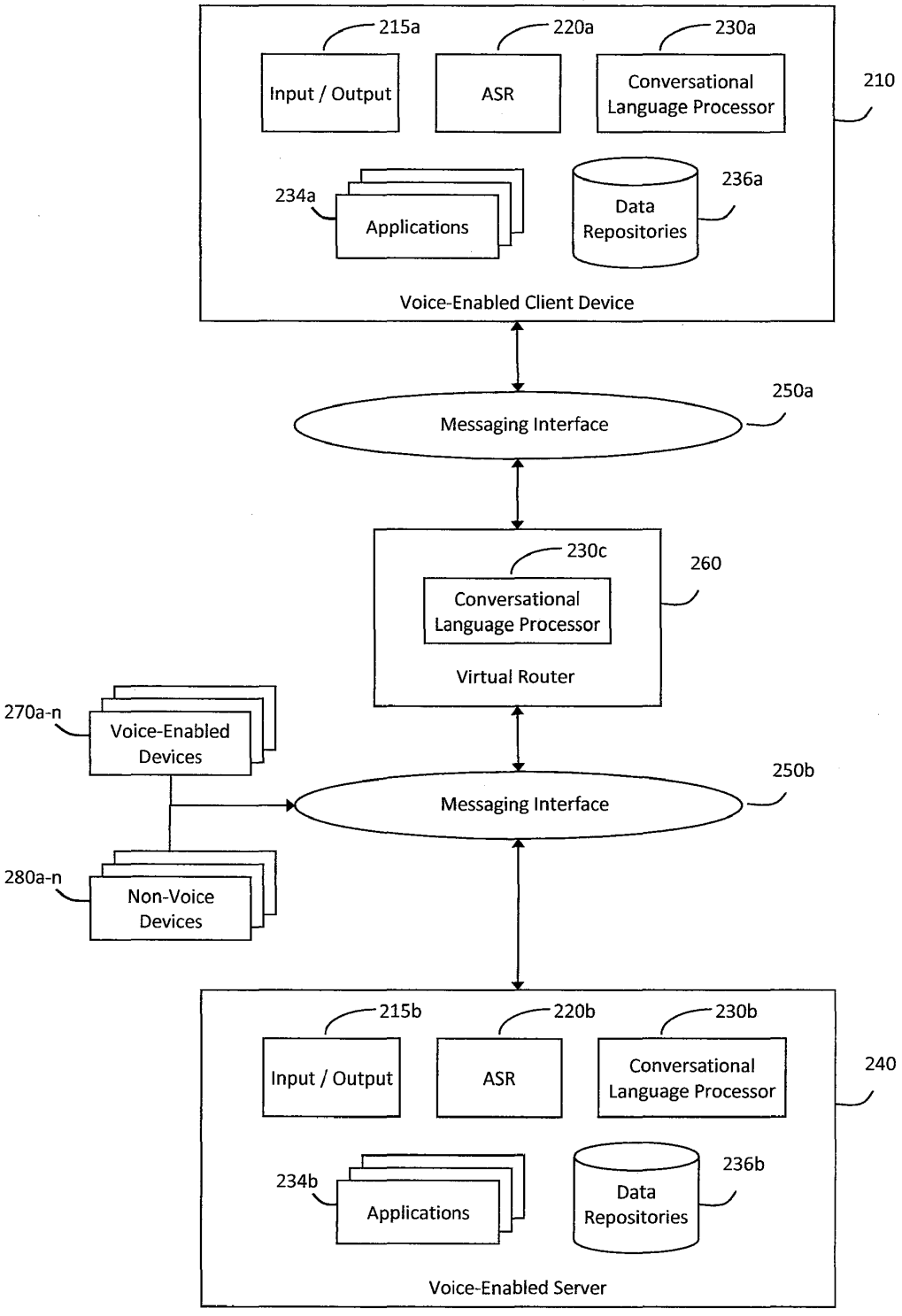

215a
220a
230a
Input / Output
ASR
Conversational Language Processor
210
234a
Applications
Data Repositories
236a
Voice-Enabled Client Device
250a
Messaging Interface
230c
260
Conversational Language Processor
Virtual Router
270a-n
Voice-Enabled Devices
250b
Messaging Interface
280a-n
Non-Voice Devices
215b
220b
230b
Input / Output
ASR
Conversational Language Processor
240
234b
Applications
Data Repositories
236b
Voice-Enabled Server

Figure 2

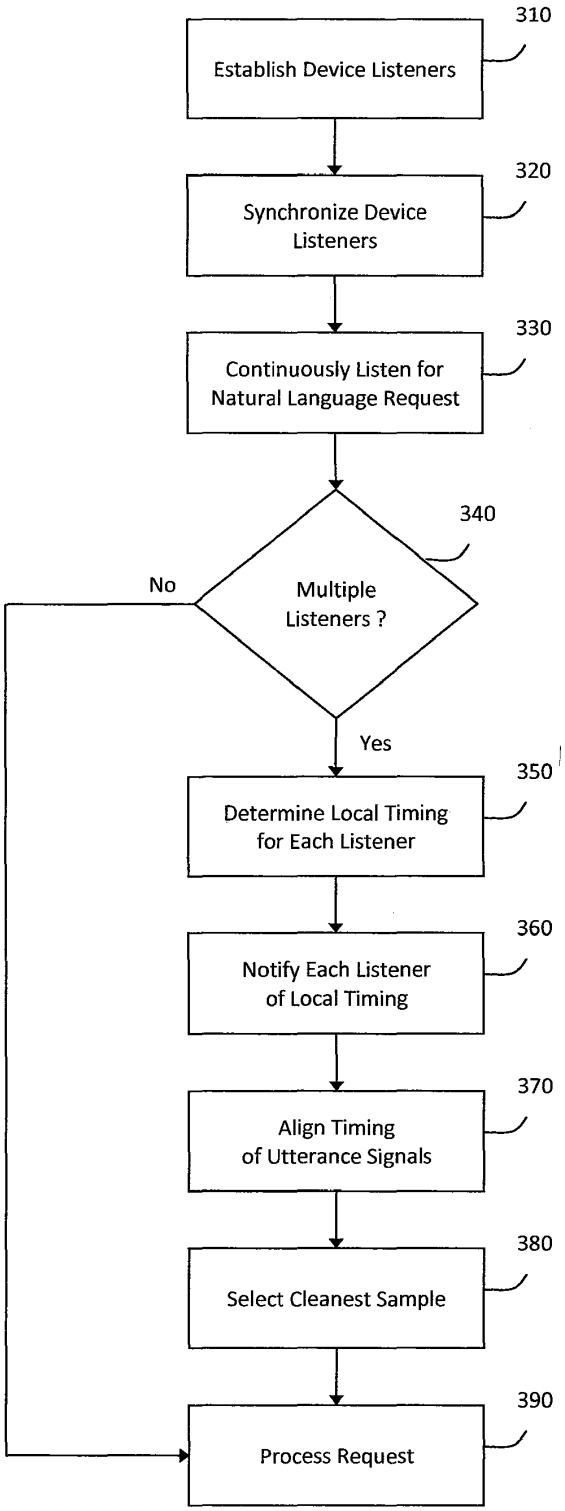

310
Establish Device Listeners
320
Synchronize Device Listeners
330
Continuously Listen for Natural Language Request
340
Multiple Listeners ?
No
Yes
350
Determine Local Timing for Each Listener
360
Notify Each Listener of Local Timing
370
Align Timing of Utterance Signals
380
Select Cleanest Sample
390
Process Request

Figure 3

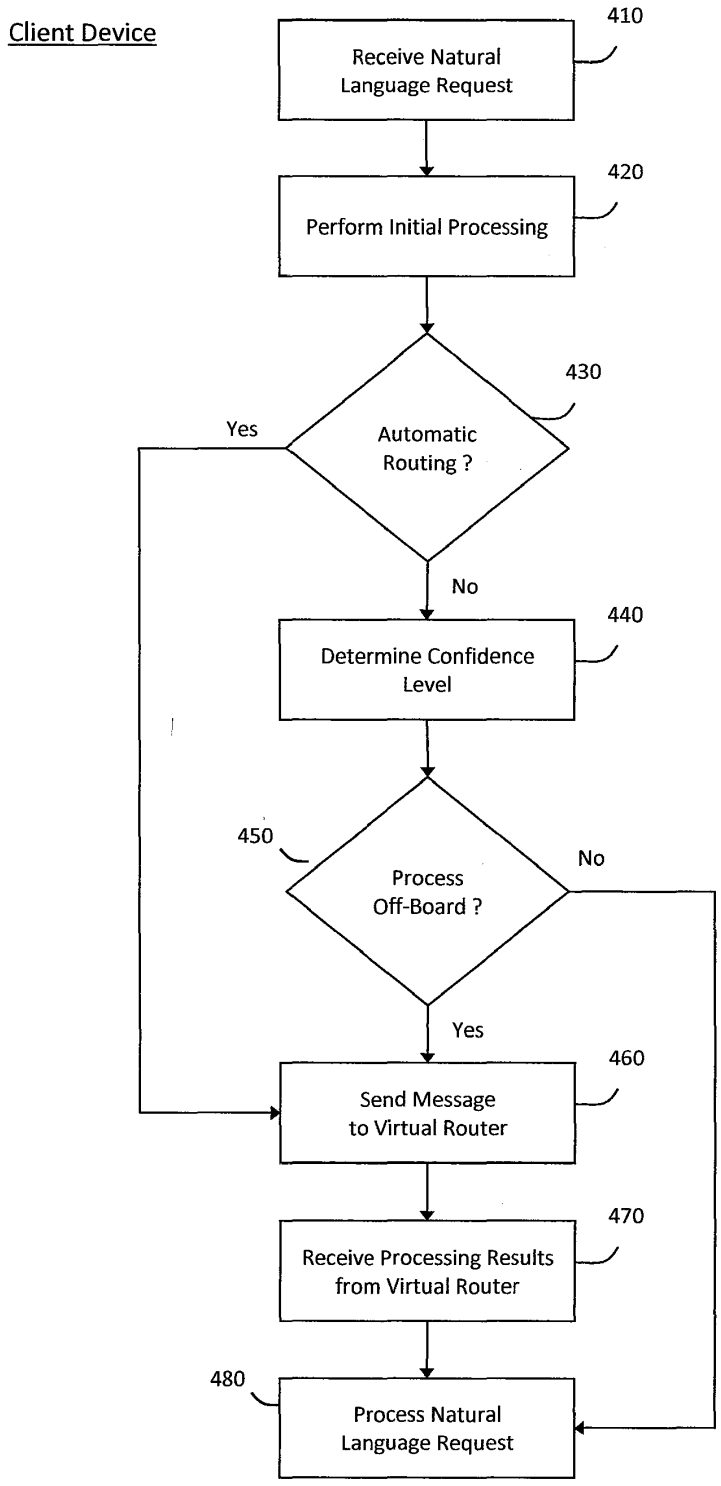

Client Device
410
Receive Natural Language Request
420
Perform Initial Processing
430
Automatic Routing ?
Yes
No
440
Determine Confidence Level
450
Process Off-Board ?
No
Yes
460
Send Message to Virtual Router
470
Receive Processing Results from Virtual Router
480
Process Natural Language Request

Figure 4

Virtual Router

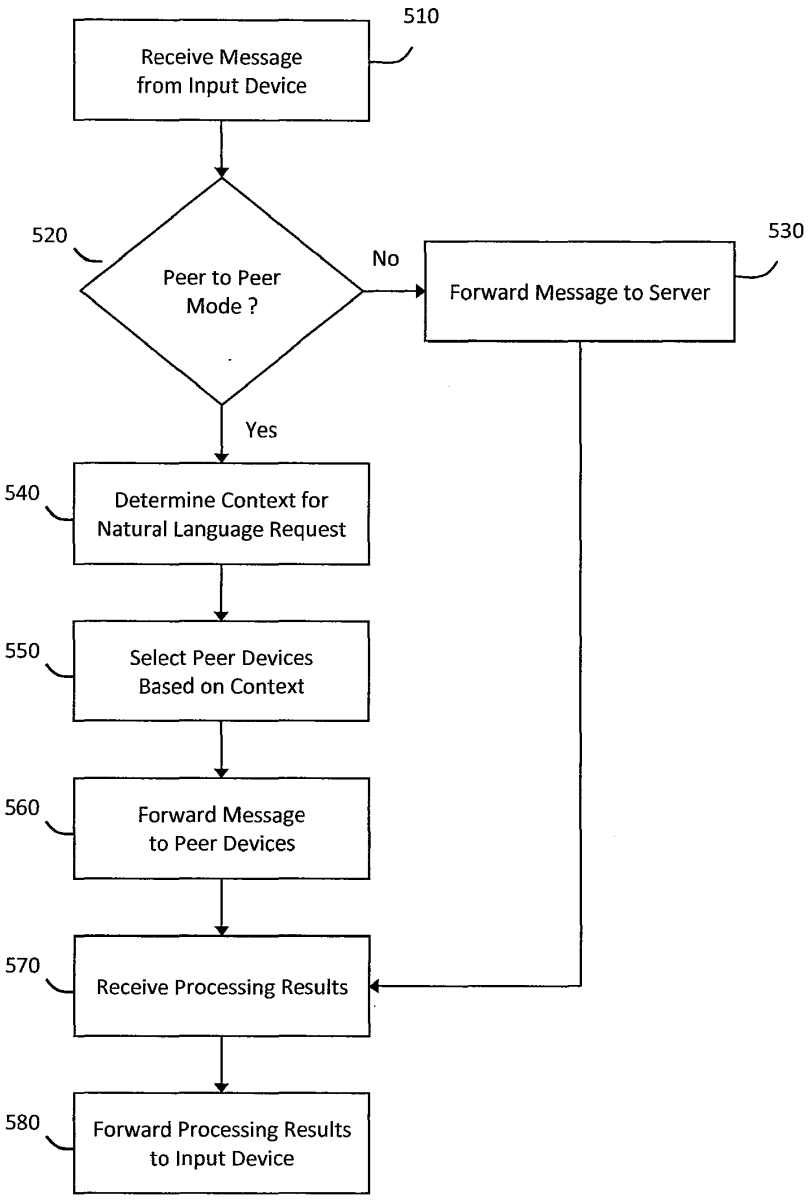


Figure 5

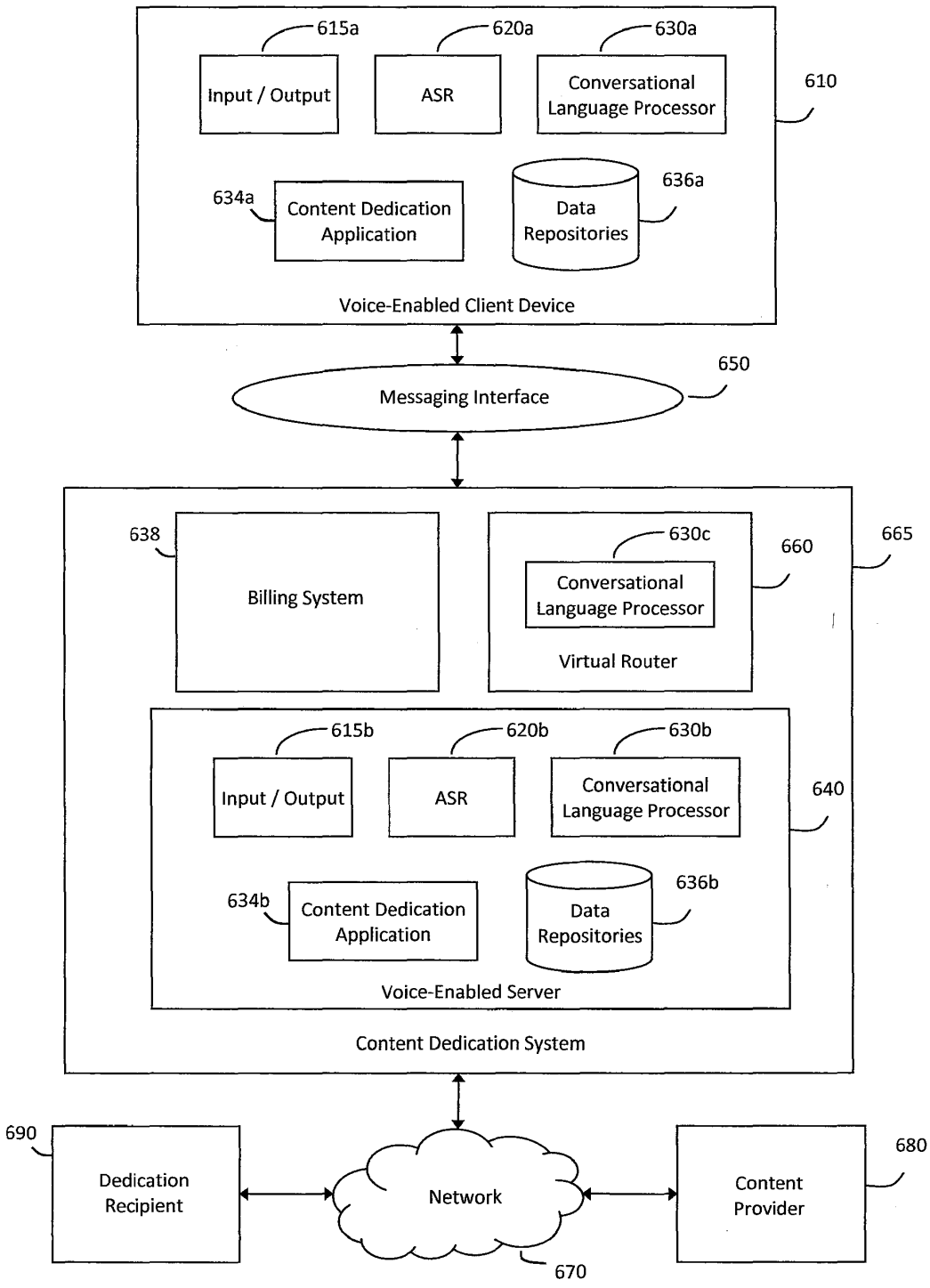

615a
620a
630a
Input / Output
ASR
Conversational Language Processor
610
634a
Content Dedication Application
Data Repositories
636a
Voice-Enabled Client Device
650
Messaging Interface
638
Billing System
630c
Conversational Language Processor
660
665
Virtual Router
615b
620b
630b
Input / Output
ASR
Conversational Language Processor
640
634b
Content Dedication Application
Data Repositories
636b
Voice-Enabled Server
Content Dedication System
690
Dedication Recipient
Network
Content Provider
680
670

Figure 6

Client Device

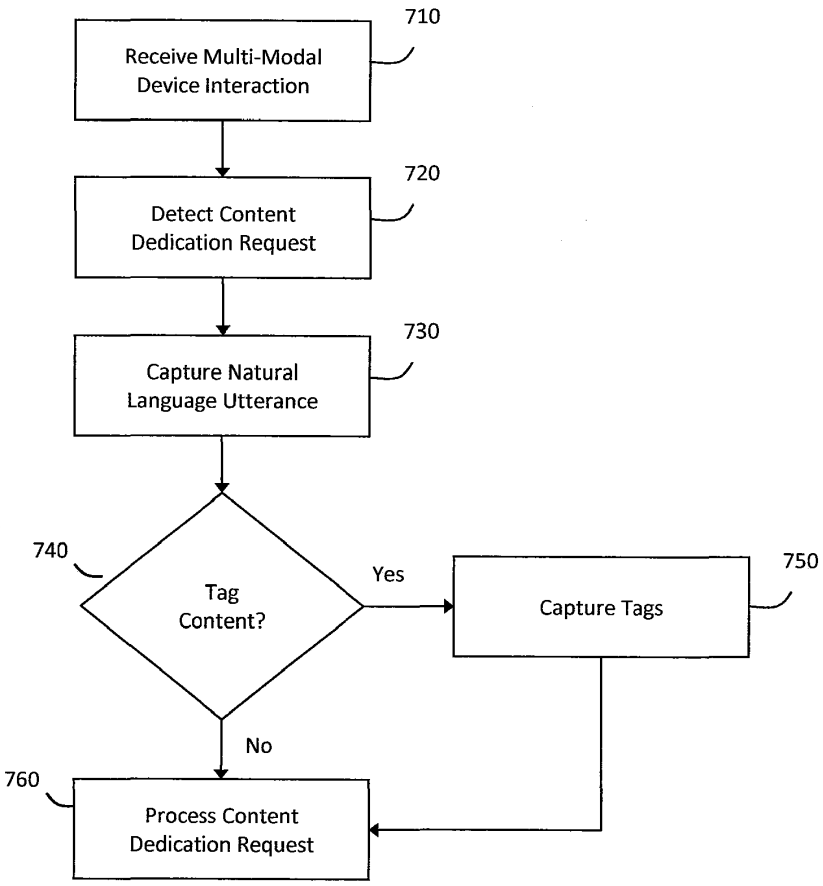


Figure 7

Content Dedication System

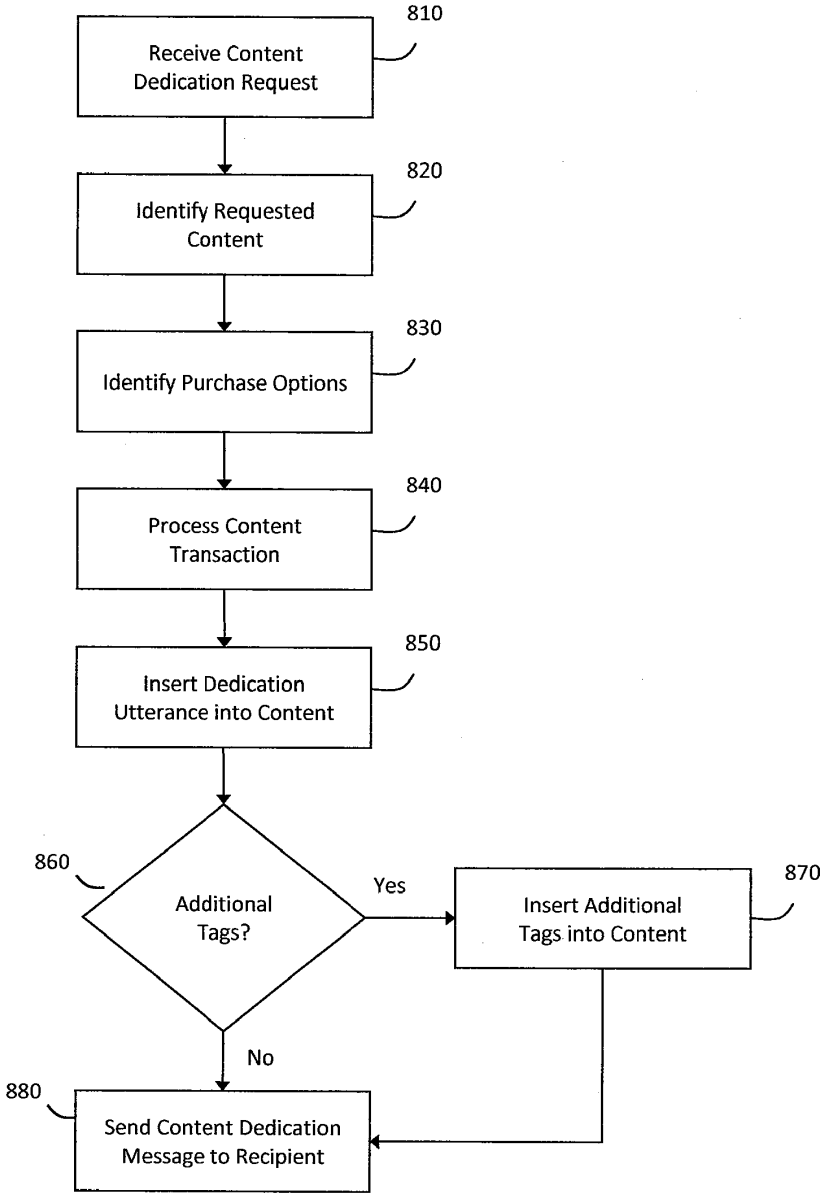


Figure 8

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G10L 5/06	A1	(11) International Publication Number: **WO 99/46763**
		(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).

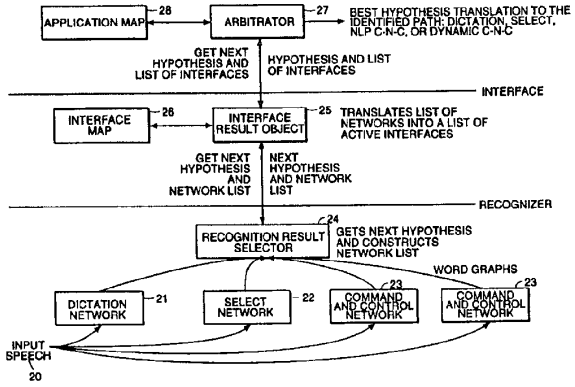
(72) Inventors: GRABHERR, Manfred; 42 Electric Avenue, Somerville, MA 02144 (US). SARUKKAI, Ramesh; Apartment #K–19, 158 Concord Road, Billerica, MA 01821 (US). GANONG, William, F., III; 128 Davis Avenue, Brookline, MA 02146 (US).

(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**	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		

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

Amazon/VB Assets
Exhibit 1003
Page 855

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

Amazon/VB Assets
Exhibit 1003
Page 856

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

Amazon/VB Assets
Exhibit 1003
Page 857

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

Amazon/VB Assets
Exhibit 1003
Page 858

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

Amazon/VB Assets
Exhibit 1003
Page 859

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

Amazon/VB Assets
Exhibit 1003
Page 860

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

Amazon/VB Assets
Exhibit 1003
Page 861

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

Amazon/VB Assets
Exhibit 1003
Page 862

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

Amazon/VB Assets
Exhibit 1003
Page 863

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

Amazon/VB Assets
Exhibit 1003
Page 864

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

Amazon/VB Assets
Exhibit 1003
Page 865

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

Amazon/VB Assets
Exhibit 1003
Page 866

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

Amazon/VB Assets
Exhibit 1003
Page 867

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

Amazon/VB Assets
Exhibit 1003
Page 868

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

Amazon/VB Assets
Exhibit 1003
Page 869

contained in the above description or shown in the accompanying drawings shall be interpreted as illustrative and not in a limiting sense.

Amazon/VB Assets
Exhibit 1003
Page 870

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.

Amazon/VB Assets
Exhibit 1003
Page 871

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:

Amazon/VB Assets
Exhibit 1003
Page 872

(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

Amazon/VB Assets
Exhibit 1003
Page 873

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.

Amazon/VB Assets
Exhibit 1003
Page 874

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

Amazon/VB Assets
Exhibit 1003
Page 875

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]

Amazon/VB Assets
Exhibit 1003
Page 876

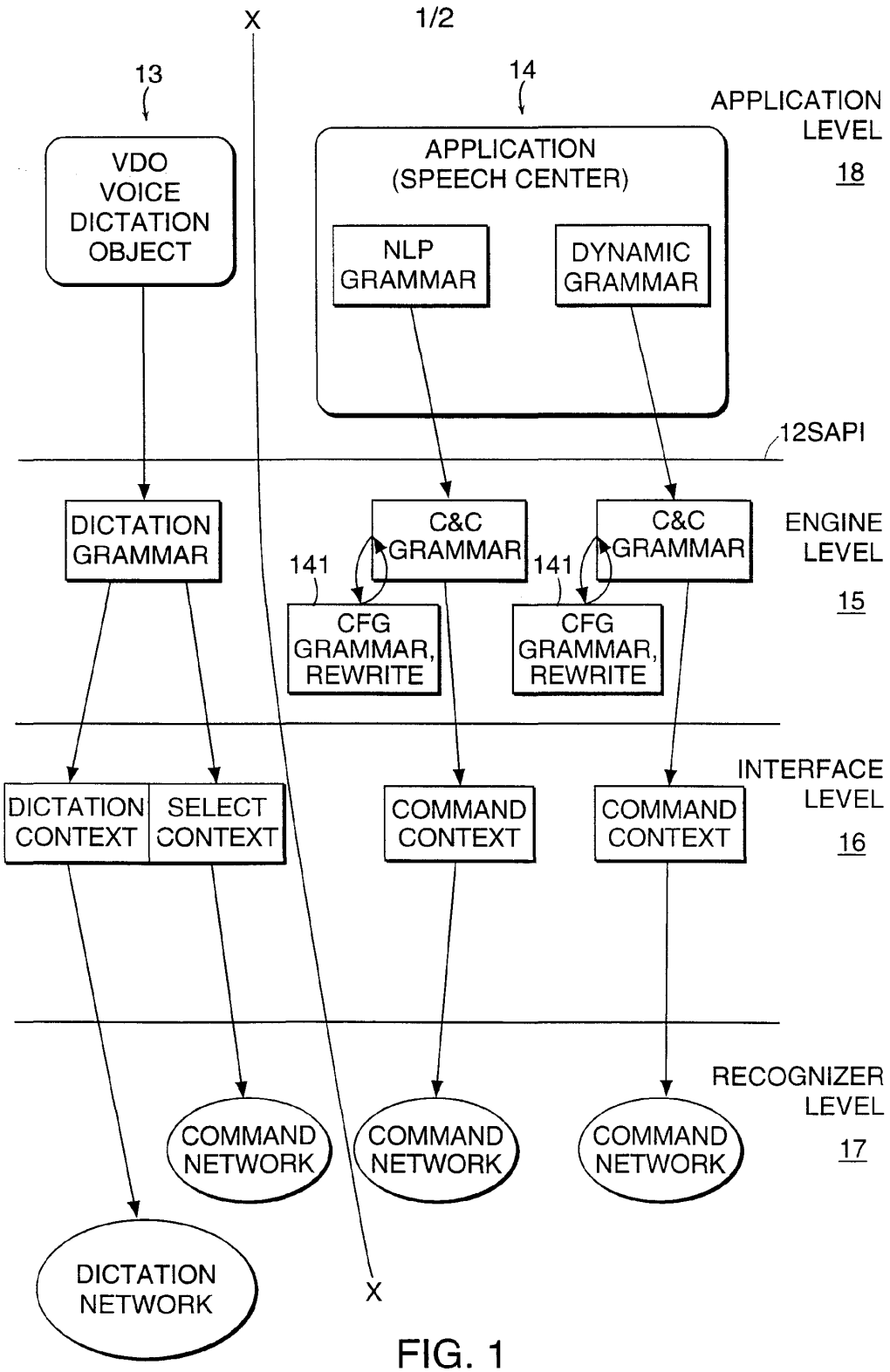


FIG. 1

Amazon/VB Assets
Exhibit 1003
Page 877

SUBSTITUTE SHEET (RULE 26)

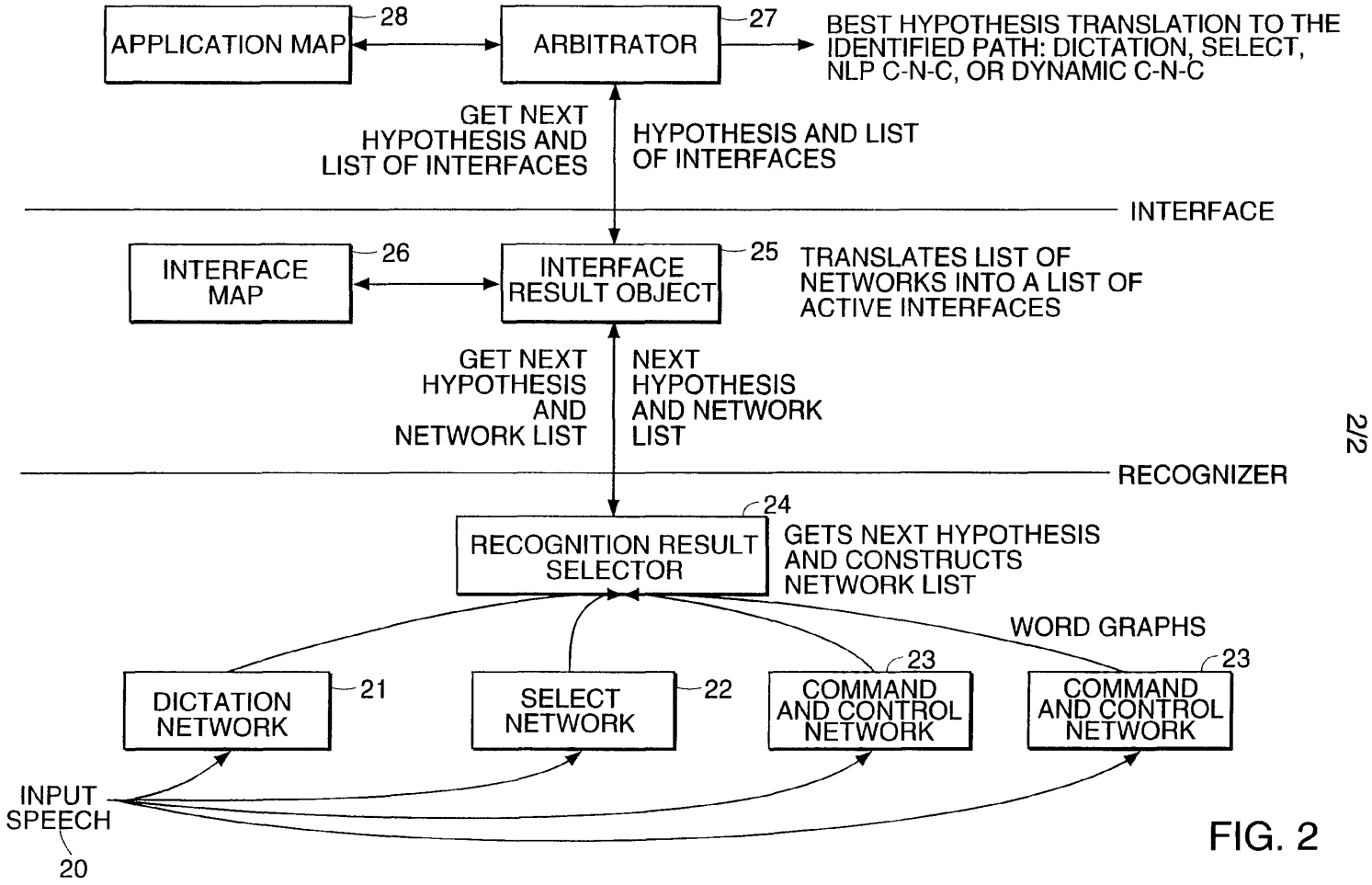


FIG. 2

Amazon/VB Assets
Exhibit 1003
Page 878

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).

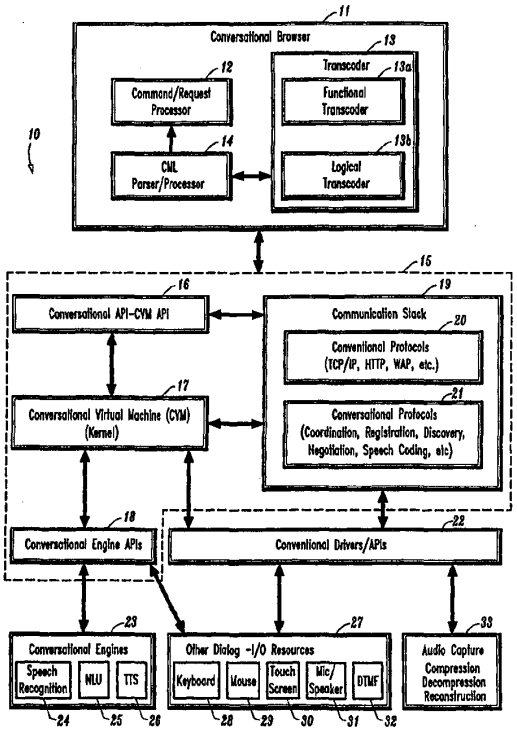
(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: GOPALAKRISHNAN, Ponani [IN/US]; 3073 Radcliff Drive, Yorktown Heights, NY 10598 (US). LUCAS, Bruce, D. [US/US]; 2408 Mill Pond Road, Yorktown Heights, NY 10598 (US). MAES, Stephane, H. [BE/US]; 1 Wintergreen Hill Road, Danbury, CT 06811 (US). 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. Am 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

Amazon/VB Assets
Exhibit 1003
Page 883

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

Amazon/VB Assets
Exhibit 1003
Page 884

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,

Amazon/VB Assets
Exhibit 1003
Page 885

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

Amazon/VB Assets
Exhibit 1003
Page 886

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.

Amazon/VB Assets
Exhibit 1003
Page 887

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

Amazon/VB Assets
Exhibit 1003
Page 888

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,

Amazon/VB Assets
Exhibit 1003
Page 889

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

Amazon/VB Assets
Exhibit 1003
Page 890

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.

Amazon/VB Assets
Exhibit 1003
Page 891

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

Amazon/VB Assets
Exhibit 1003
Page 892

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

Amazon/VB Assets
Exhibit 1003
Page 893

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

Amazon/VB Assets
Exhibit 1003
Page 894

communication mechanisms can all be built directly into, and made part of, the 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

Amazon/VB Assets
Exhibit 1003
Page 895

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

Amazon/VB Assets
Exhibit 1003
Page 896

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

Amazon/VB Assets
Exhibit 1003
Page 897

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

Amazon/VB Assets
Exhibit 1003
Page 898

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

Amazon/VB Assets
Exhibit 1003
Page 899

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

Amazon/VB Assets
Exhibit 1003
Page 900

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.

Amazon/VB Assets
Exhibit 1003
Page 901

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 be 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

Amazon/VB Assets
Exhibit 1003
Page 902

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.

Amazon/VB Assets
Exhibit 1003
Page 903

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

Amazon/VB Assets
Exhibit 1003
Page 904

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

Amazon/VB Assets
Exhibit 1003
Page 905

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

Amazon/VB Assets
Exhibit 1003
Page 906

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,

Amazon/VB Assets
Exhibit 1003
Page 907

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

Amazon/VB Assets
Exhibit 1003
Page 908

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

Amazon/VB Assets
Exhibit 1003
Page 909

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,

Amazon/VB Assets
Exhibit 1003
Page 910

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

Amazon/VB Assets
Exhibit 1003
Page 911

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

Amazon/VB Assets
Exhibit 1003
Page 912

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).

Amazon/VB Assets
Exhibit 1003
Page 913

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.

Amazon/VB Assets
Exhibit 1003
Page 914

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

Amazon/VB Assets
Exhibit 1003
Page 915

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

Amazon/VB Assets
Exhibit 1003
Page 916

<**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

Amazon/VB Assets
Exhibit 1003
Page 917

(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."

Amazon/VB Assets
Exhibit 1003
Page 918

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)

Amazon/VB Assets
Exhibit 1003
Page 919

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).

Amazon/VB Assets
Exhibit 1003
Page 920

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.

Amazon/VB Assets
Exhibit 1003
Page 921

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

Amazon/VB Assets
Exhibit 1003
Page 922

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.

Amazon/VB Assets
Exhibit 1003
Page 923

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

Amazon/VB Assets
Exhibit 1003
Page 924

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

Amazon/VB Assets
Exhibit 1003
Page 925

becomes active when the page is first loaded. Advantageously, 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.

Amazon/VB Assets
Exhibit 1003
Page 926

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

Amazon/VB Assets
Exhibit 1003
Page 927

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>

...

Amazon/VB Assets
Exhibit 1003
Page 928

</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 language. 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

Amazon/VB Assets
Exhibit 1003
Page 929

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).

Amazon/VB Assets
Exhibit 1003
Page 930

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 the 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

Amazon/VB Assets
Exhibit 1003
Page 931

<**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

Amazon/VB Assets
Exhibit 1003
Page 932

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

Amazon/VB Assets
Exhibit 1003
Page 933

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

Amazon/VB Assets
Exhibit 1003
Page 934

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

Amazon/VB Assets
Exhibit 1003
Page 935

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).

Amazon/VB Assets
Exhibit 1003
Page 936

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 these 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

Amazon/VB Assets
Exhibit 1003
Page 937

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.

Amazon/VB Assets
Exhibit 1003
Page 938

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.

Amazon/VB Assets
Exhibit 1003
Page 939

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.

Amazon/VB Assets
Exhibit 1003
Page 940

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.

Amazon/VB Assets
Exhibit 1003
Page 941

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.

Amazon/VB Assets
Exhibit 1003
Page 942

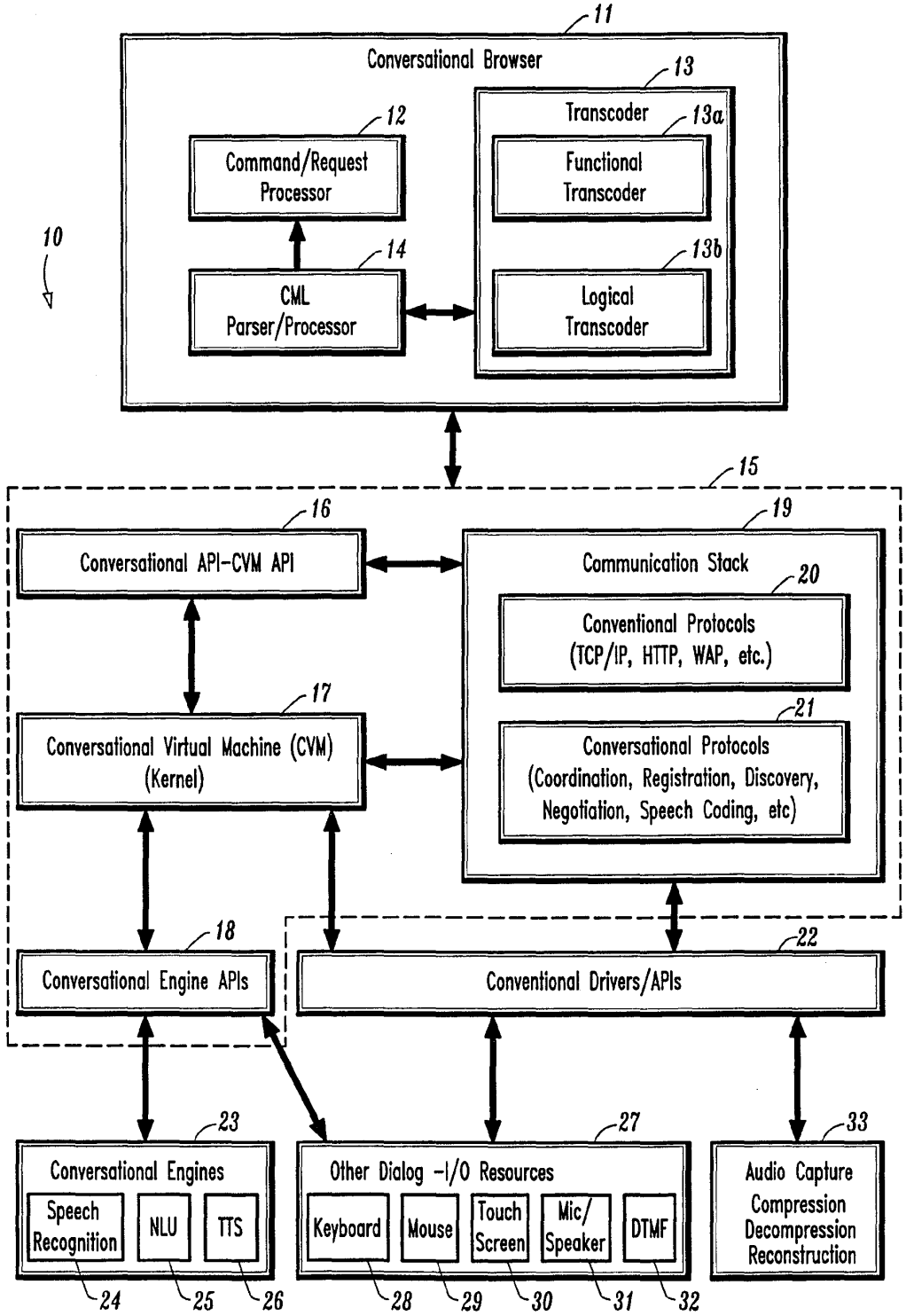


FIG. 1

SUBSTITUTE SHEET (RULE 26)

Amazon/VB Assets
Exhibit 1003
Page 943

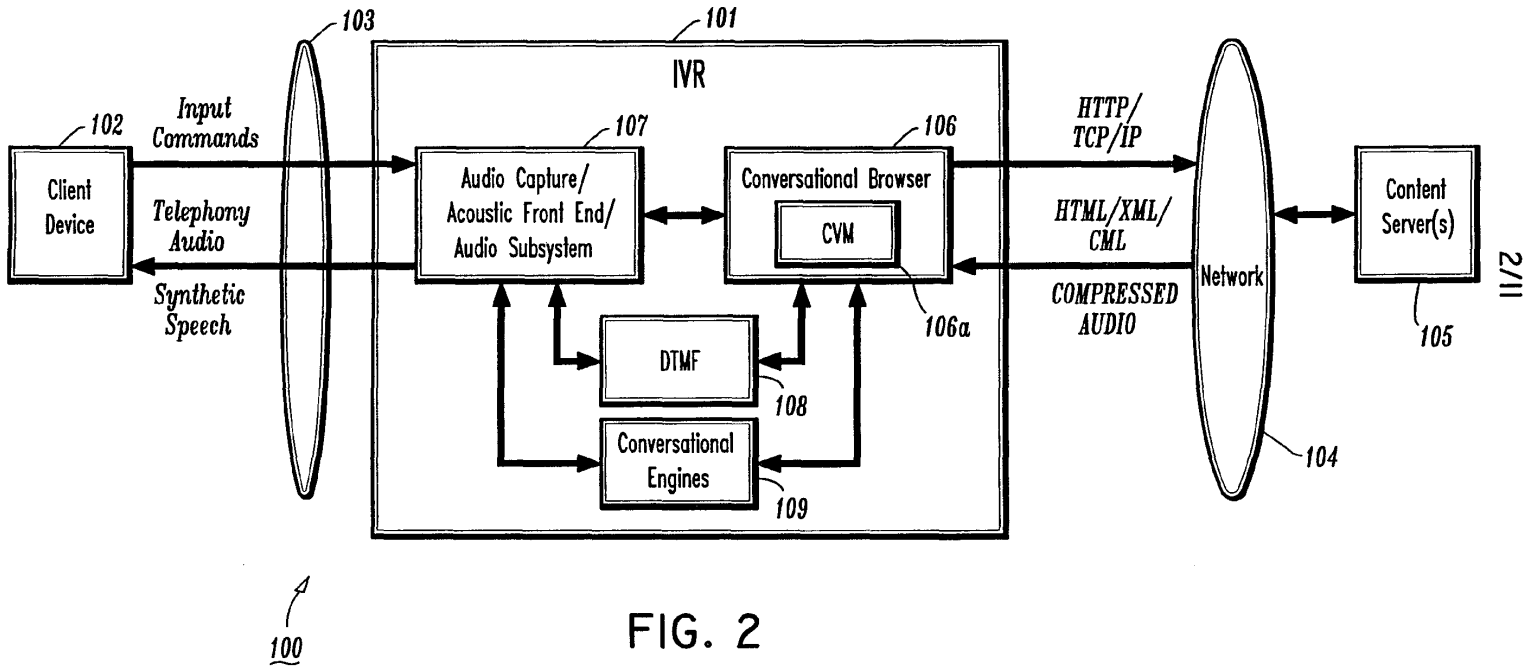


FIG. 2

Amazon/VB Assets
Exhibit 1003
Page 944

Amazon/VB Assets
Exhibit 1003
Page 945

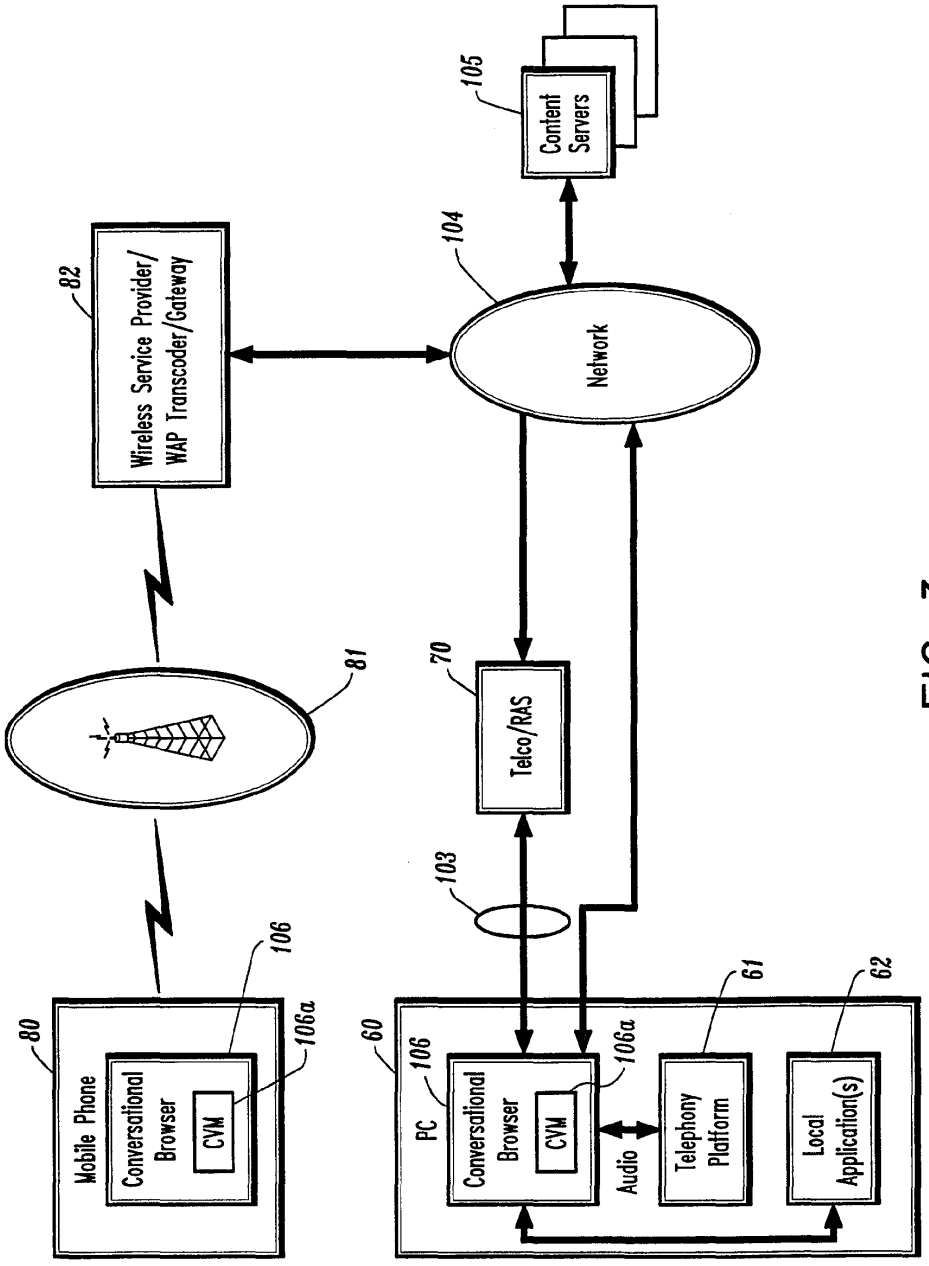


FIG. 3

Amazon/VB Assets
Exhibit 1003
Page 946

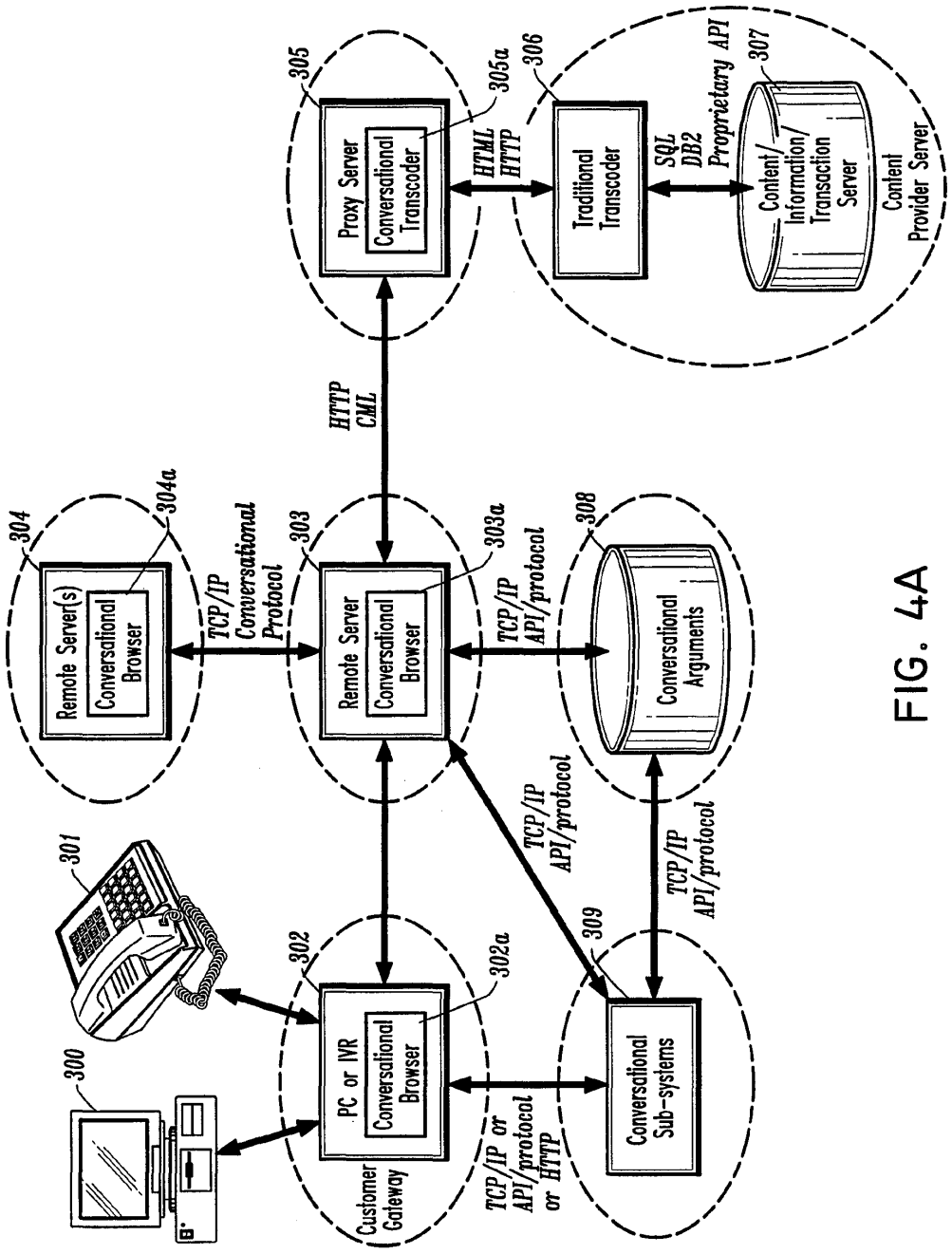


FIG. 4A

SUBSTITUTE SHEET (RULE 26)

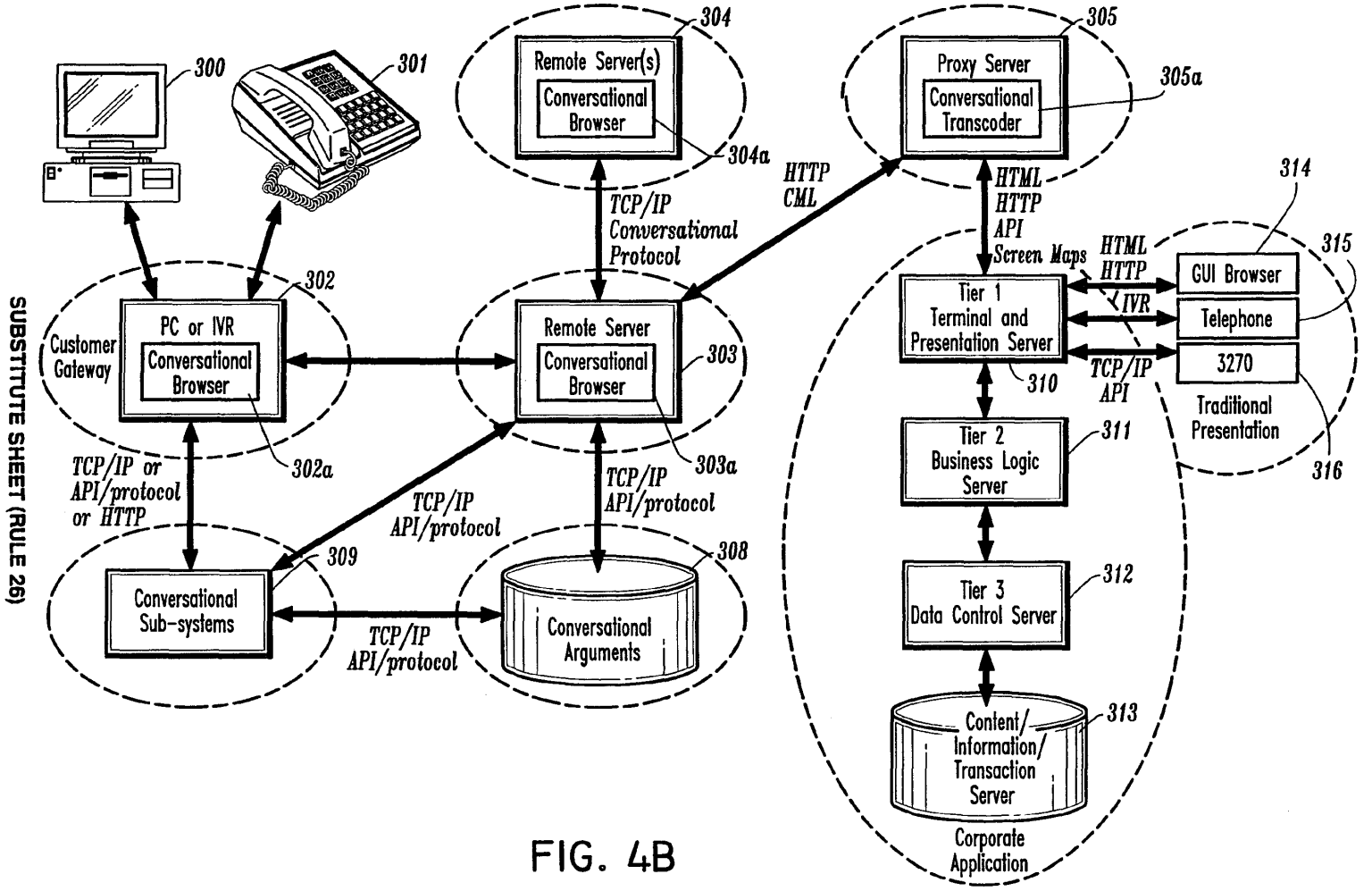


FIG. 4B

Amazon/VB Assets
Exhibit 1003
Page 947

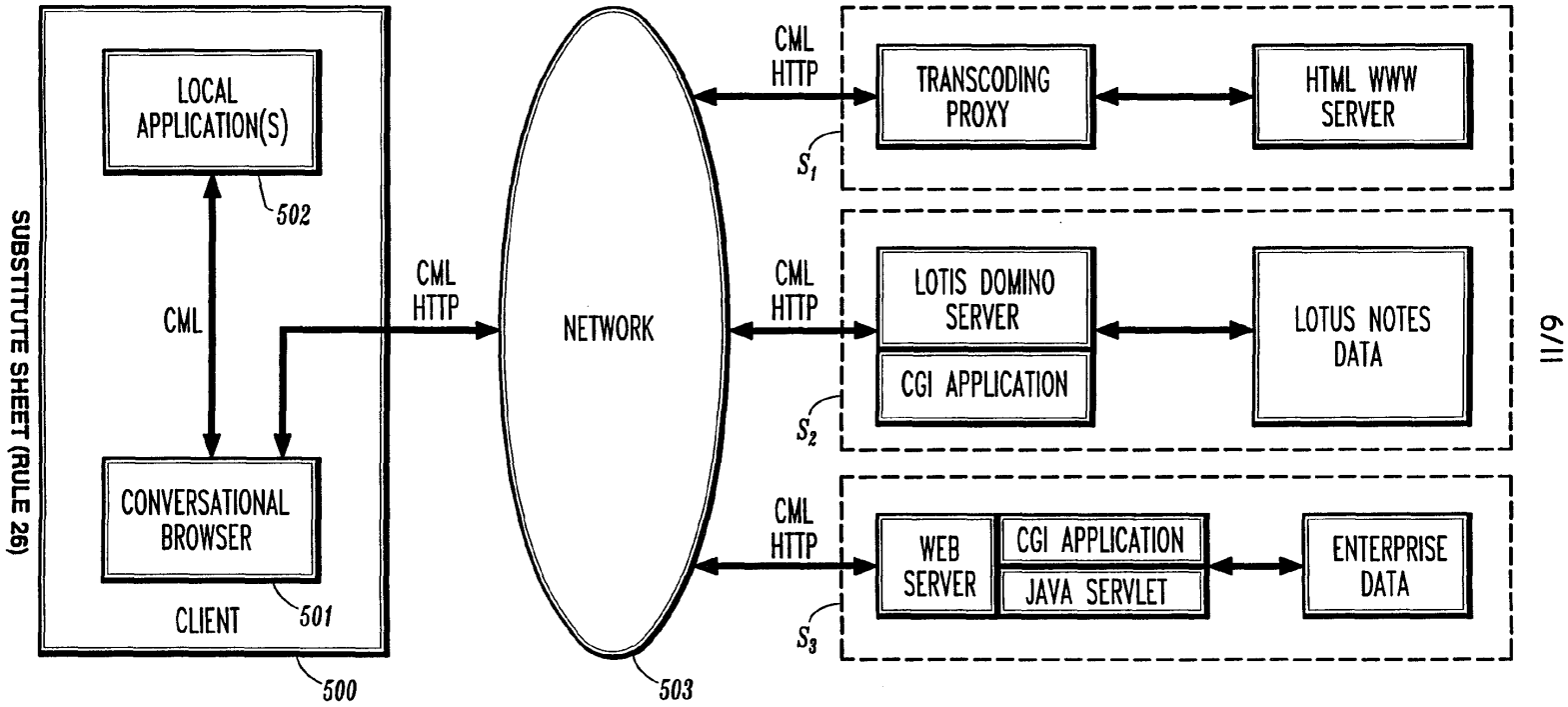


FIG. 5

Amazon/VB Assets
Exhibit 1003
Page 948

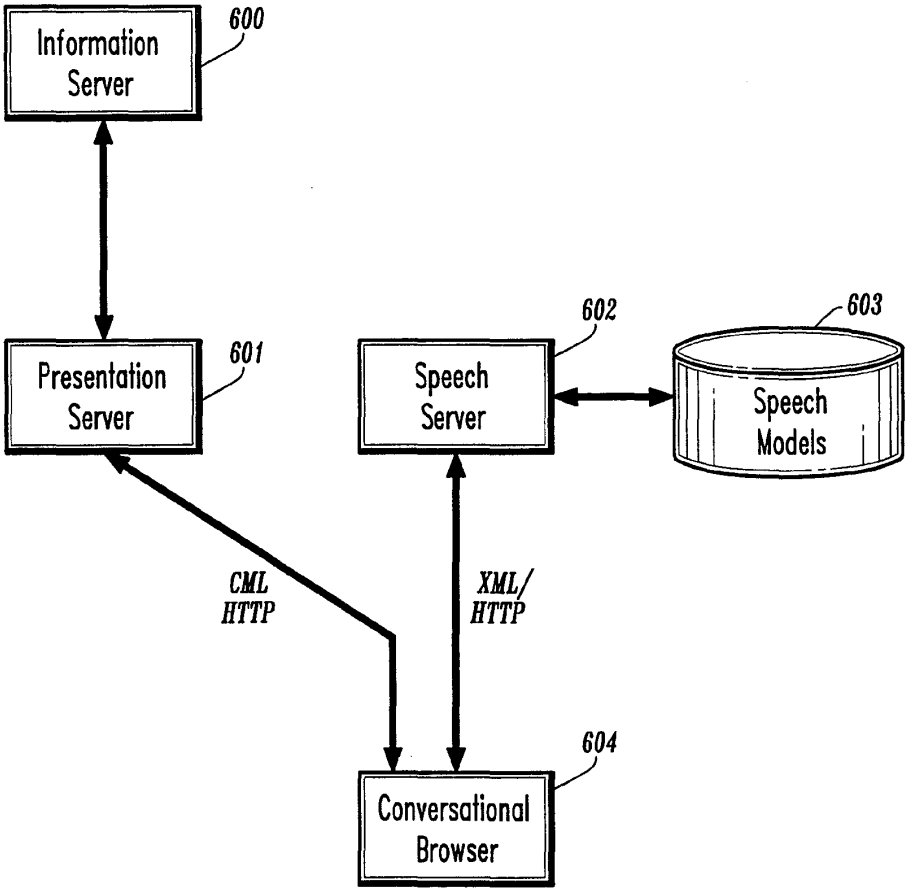


FIG. 6

Amazon/VB Assets
Exhibit 1003
Page 949

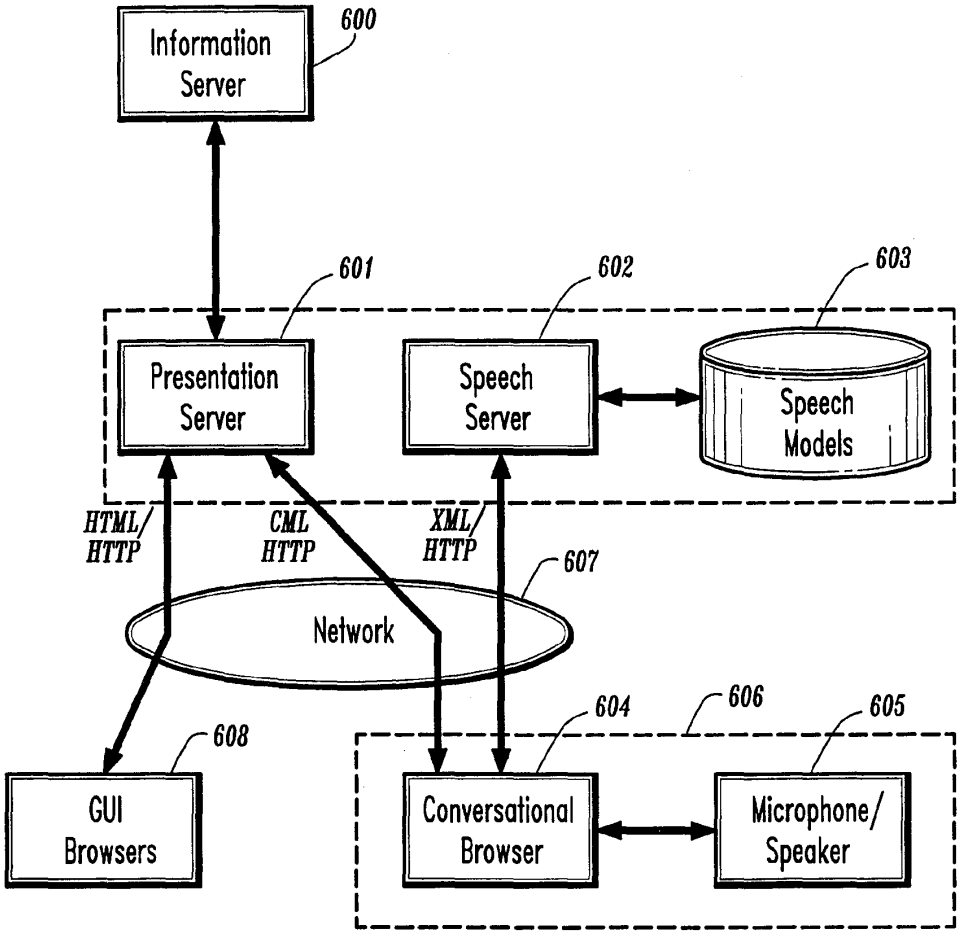


FIG. 7

Amazon/VB Assets
Exhibit 1003
Page 950

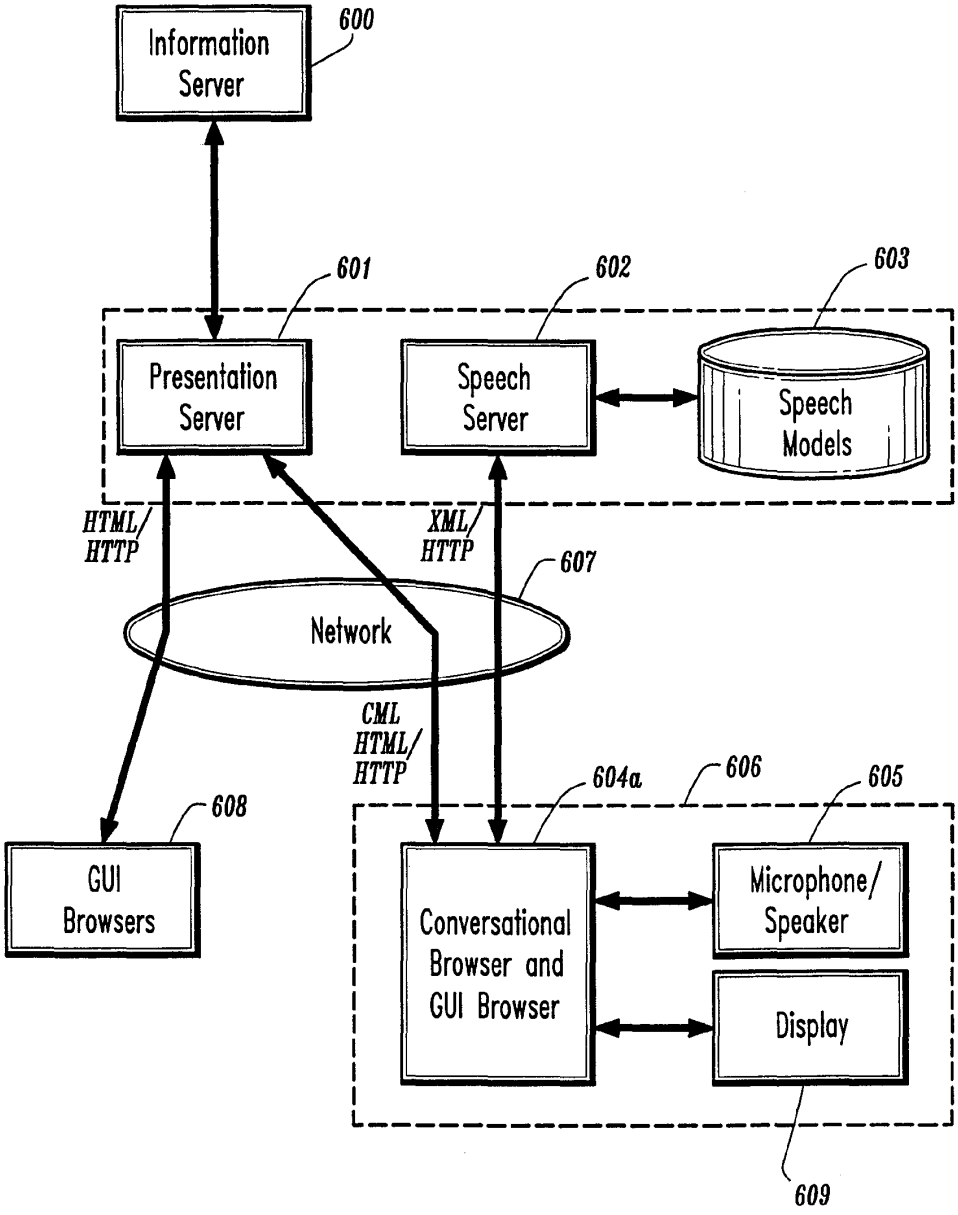


FIG. 8

Amazon/VB Assets
Exhibit 1003
Page 951

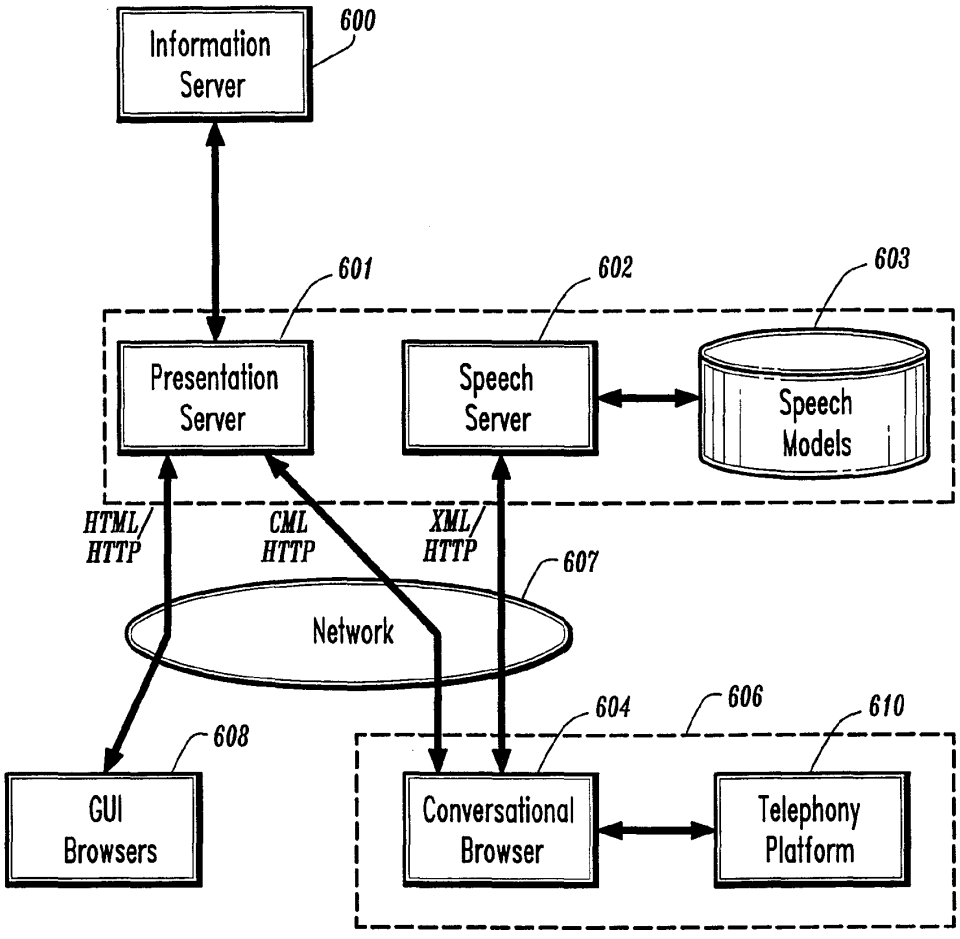


FIG. 9

Amazon/VB Assets
Exhibit 1003
Page 952

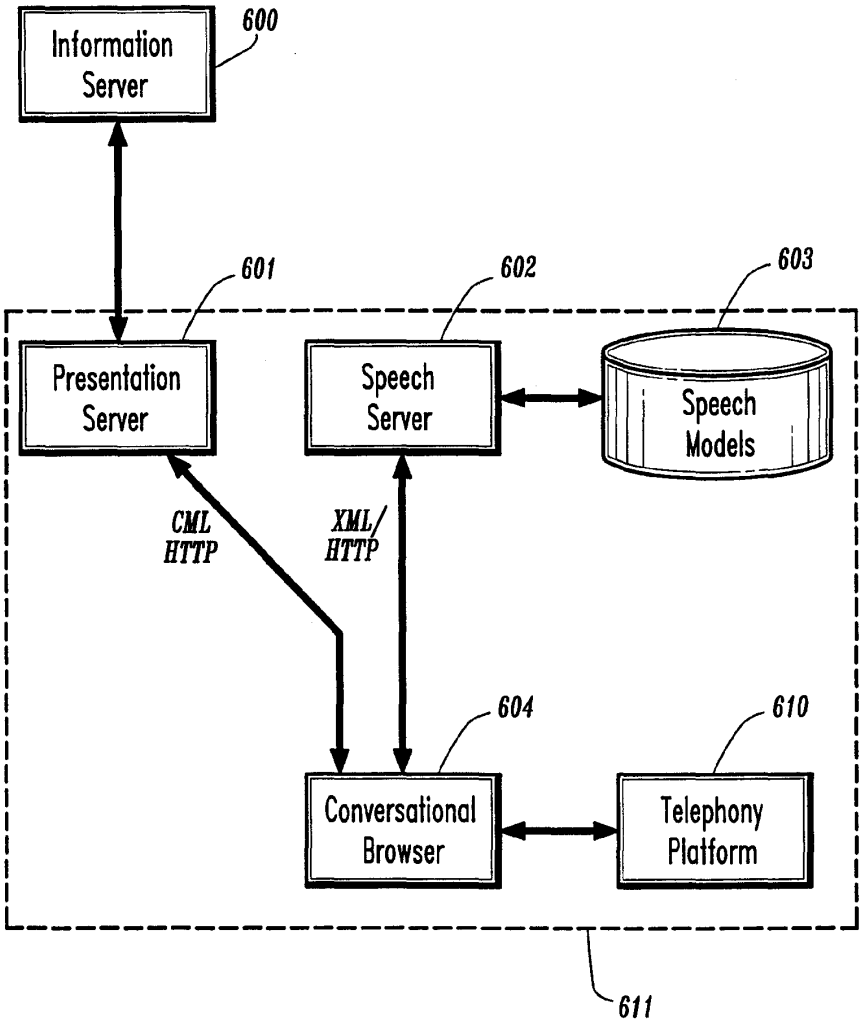


FIG. 10

Amazon/VB Assets
Exhibit 1003
Page 953

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).

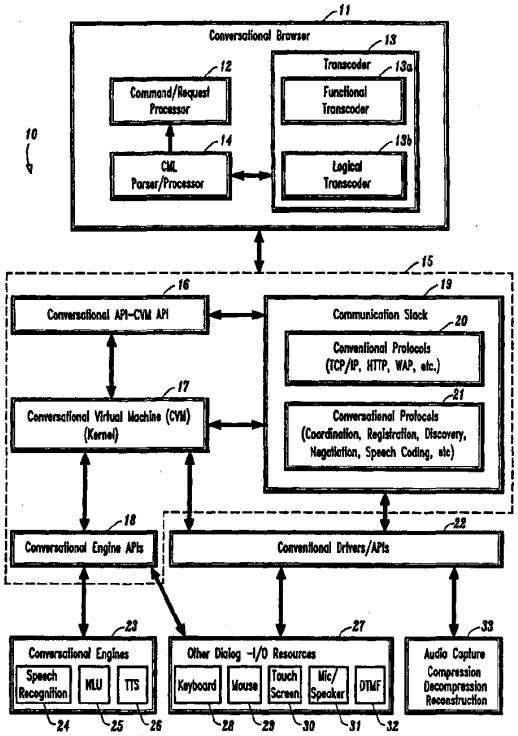
Published
With international search report.

(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.



Amazon/VB Assets
Exhibit 1003
Page 954

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

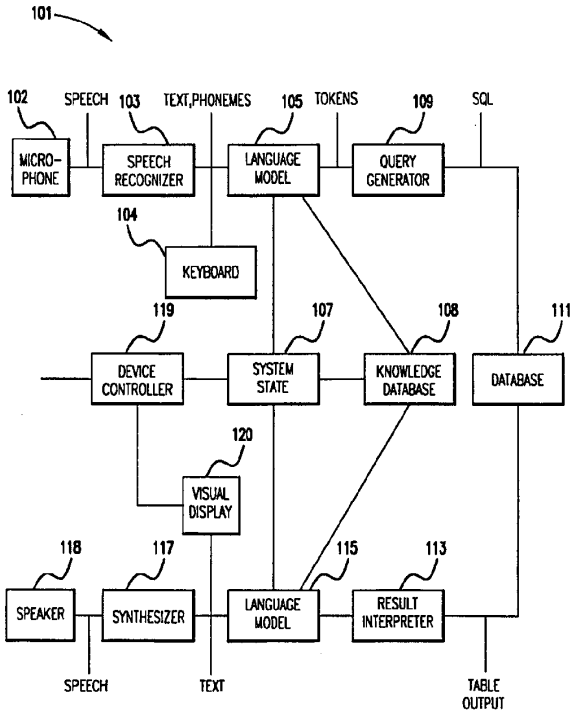
With international search report.

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.

Amazon/VB Assets
Exhibit 1003
Page 960

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

Amazon/VB Assets
Exhibit 1003
Page 961

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.

Amazon/VB Assets
Exhibit 1003
Page 962

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

Amazon/VB Assets
Exhibit 1003
Page 963

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?",

Amazon/VB Assets
Exhibit 1003
Page 964

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:

Amazon/VB Assets
Exhibit 1003
Page 965

"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

Amazon/VB Assets
Exhibit 1003
Page 966

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.

Amazon/VB Assets
Exhibit 1003
Page 967

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",

Amazon/VB Assets
Exhibit 1003
Page 968

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:

Amazon/VB Assets
Exhibit 1003
Page 969

"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."

Amazon/VB Assets
Exhibit 1003
Page 970

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

Amazon/VB Assets
Exhibit 1003
Page 971

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

Amazon/VB Assets
Exhibit 1003
Page 972

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

Amazon/VB Assets
Exhibit 1003
Page 973

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.

Amazon/VB Assets
Exhibit 1003
Page 974

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?

Amazon/VB Assets
Exhibit 1003
Page 975

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.

Amazon/VB Assets
Exhibit 1003
Page 976

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:

Amazon/VB Assets
Exhibit 1003
Page 977

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.

Amazon/VB Assets
Exhibit 1003
Page 978

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;

Amazon/VB Assets
Exhibit 1003
Page 979

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.

Amazon/VB Assets
Exhibit 1003
Page 980

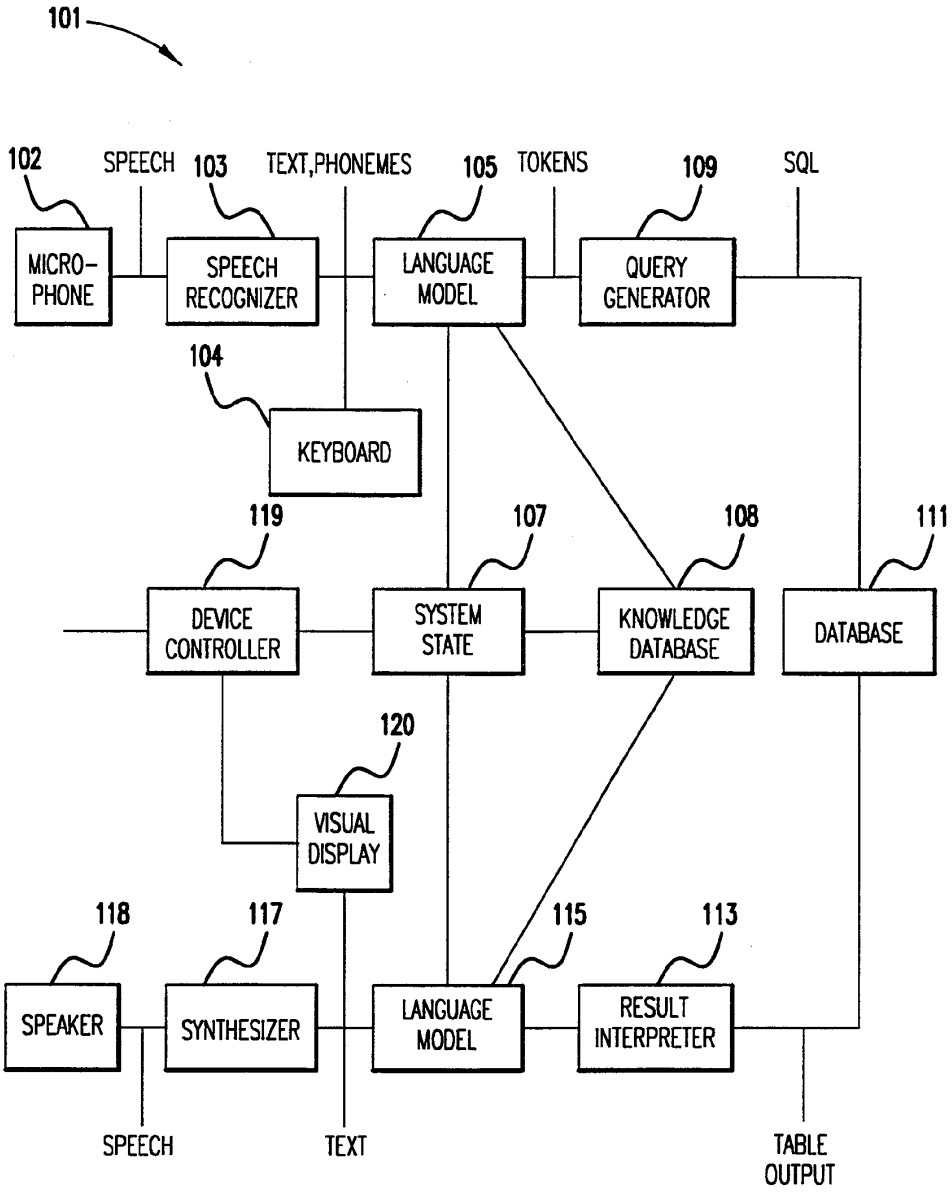


FIG.1A

Amazon/VB Assets
Exhibit 1003
Page 981

Amazon/VB Assets
Exhibit 1003
Page 982

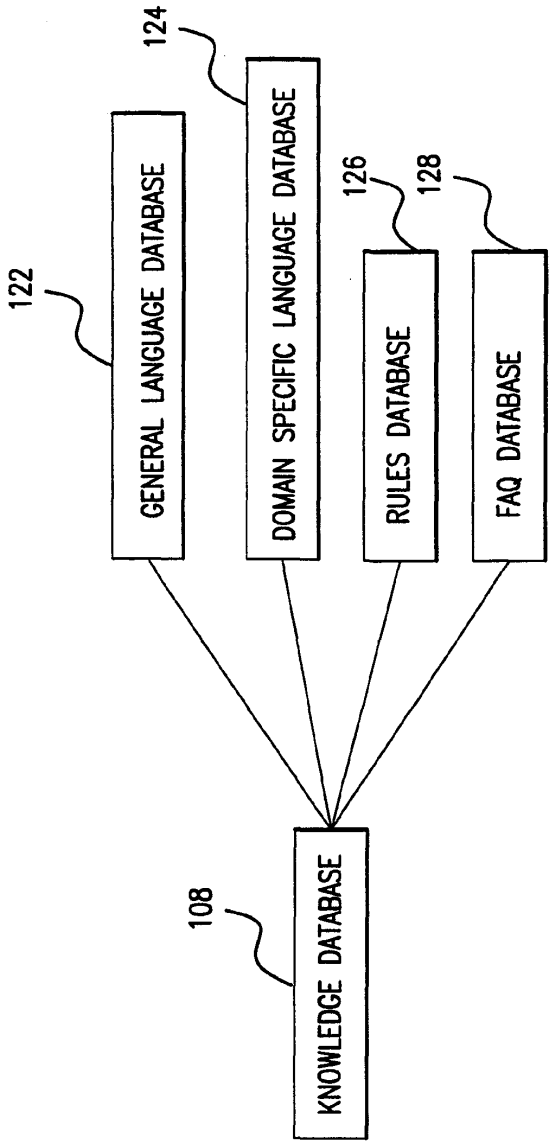


FIG.1B

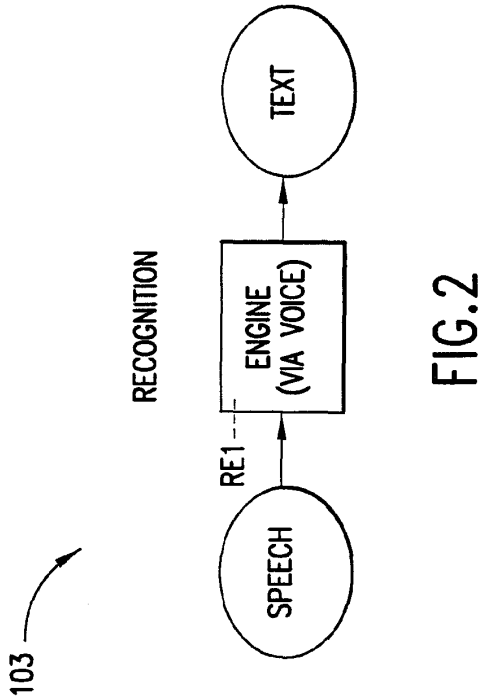


FIG.2

SUBSTITUTE SHEET (RULE26)

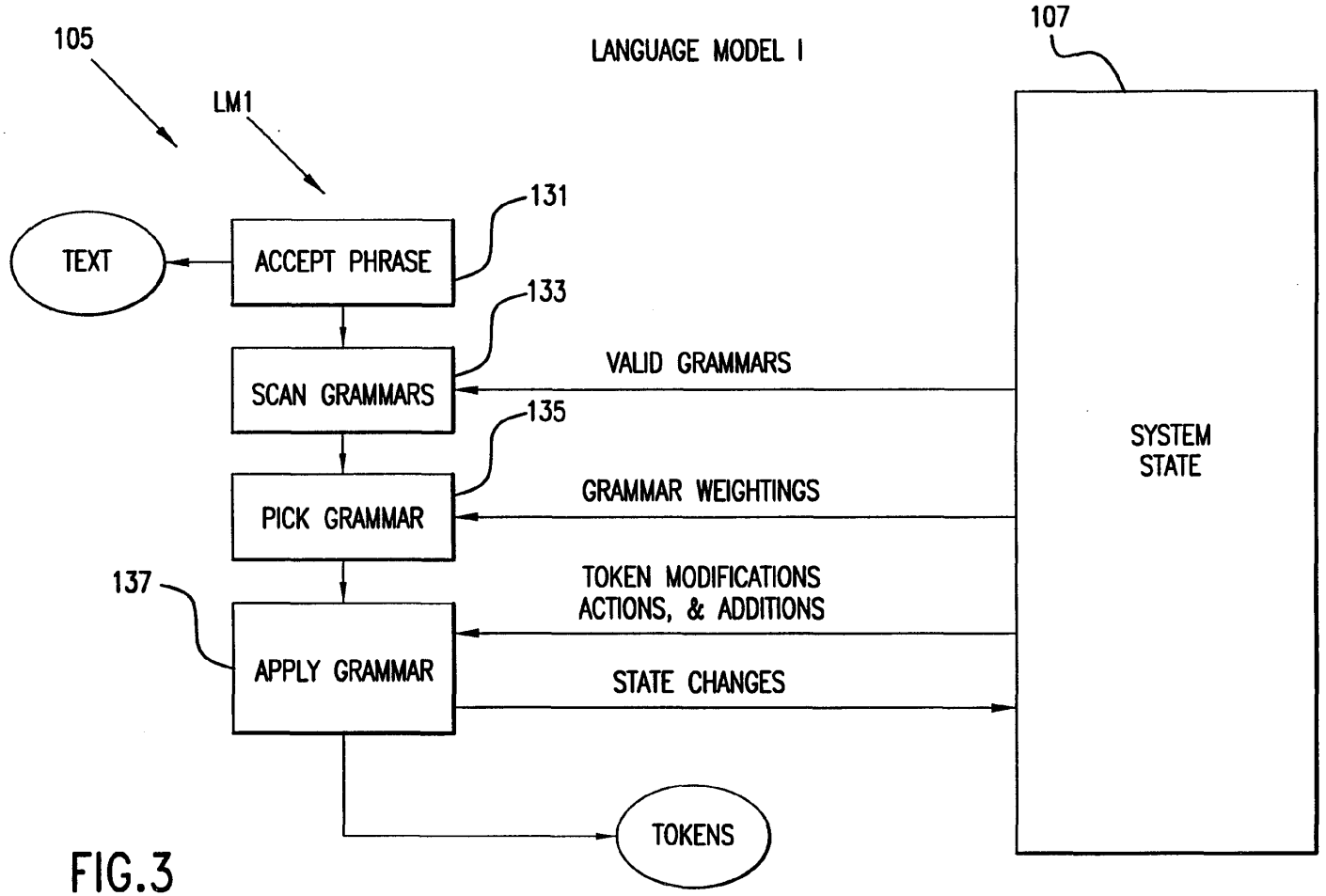


FIG.3

Amazon/VB Assets
Exhibit 1003
Page 984

Amazon/VB Assets
Exhibit 1003
Page 985

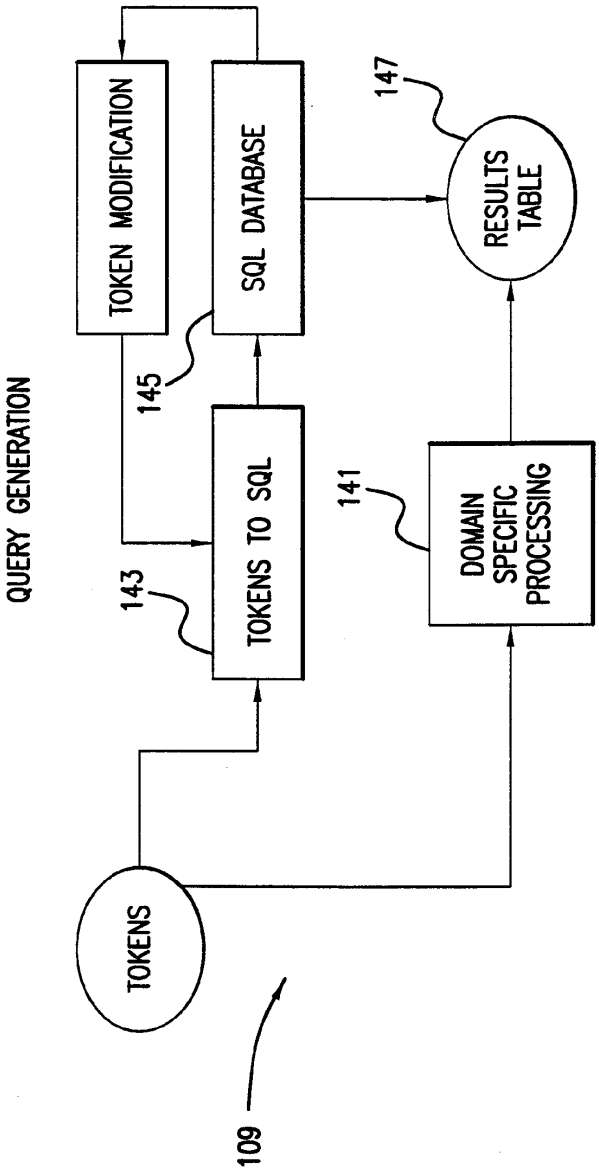


FIG.4

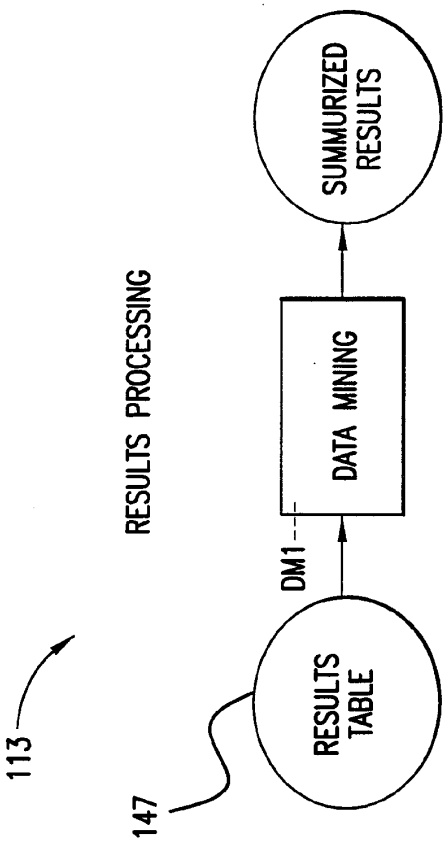


FIG.5

Amazon/VB Assets
Exhibit 1003
Page 986

Amazon/VB Assets
Exhibit 1003
Page 987

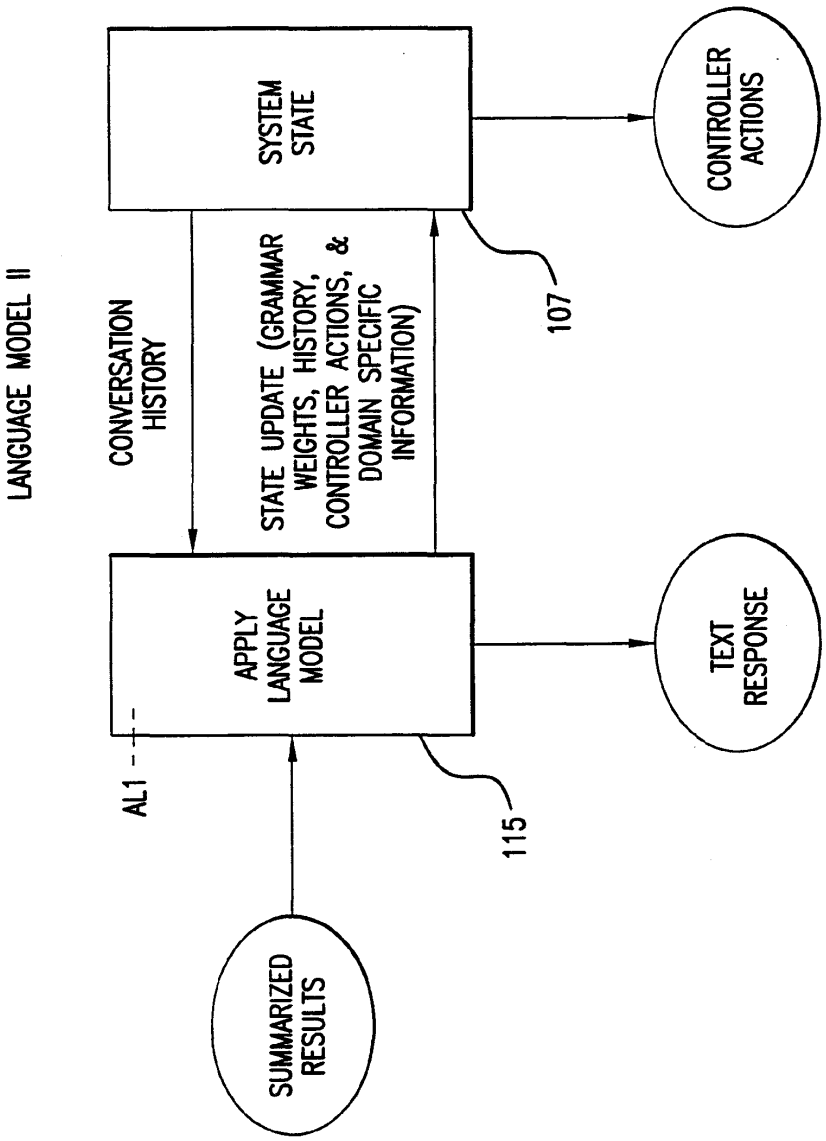


FIG.6

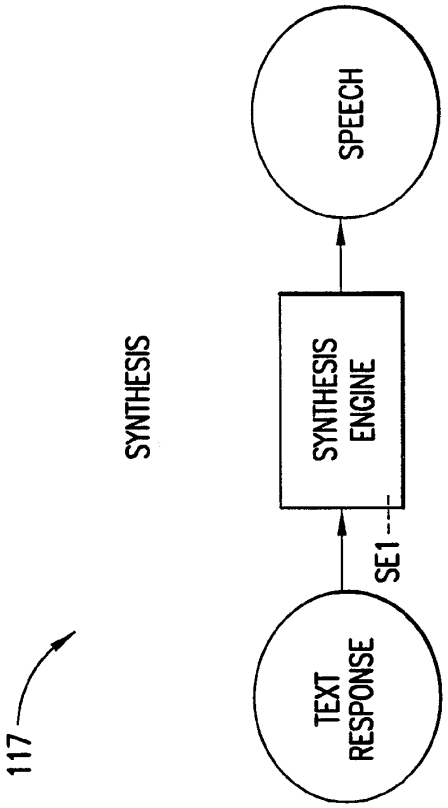


FIG.7

Amazon/VB Assets
Exhibit 1003
Page 988

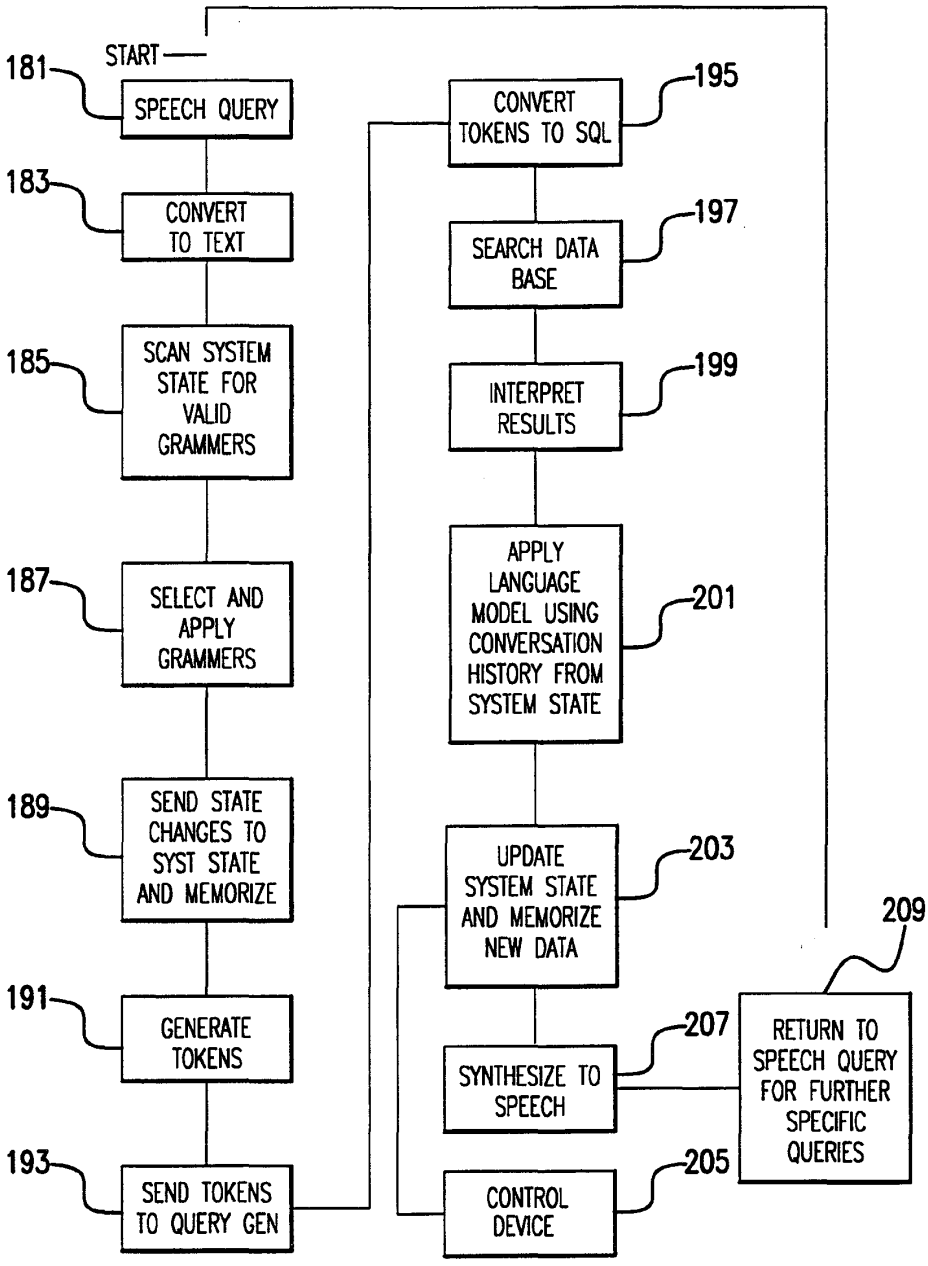


FIG.8

Amazon/VB Assets
Exhibit 1003
Page 989

Amazon/VB Assets
Exhibit 1003
Page 990

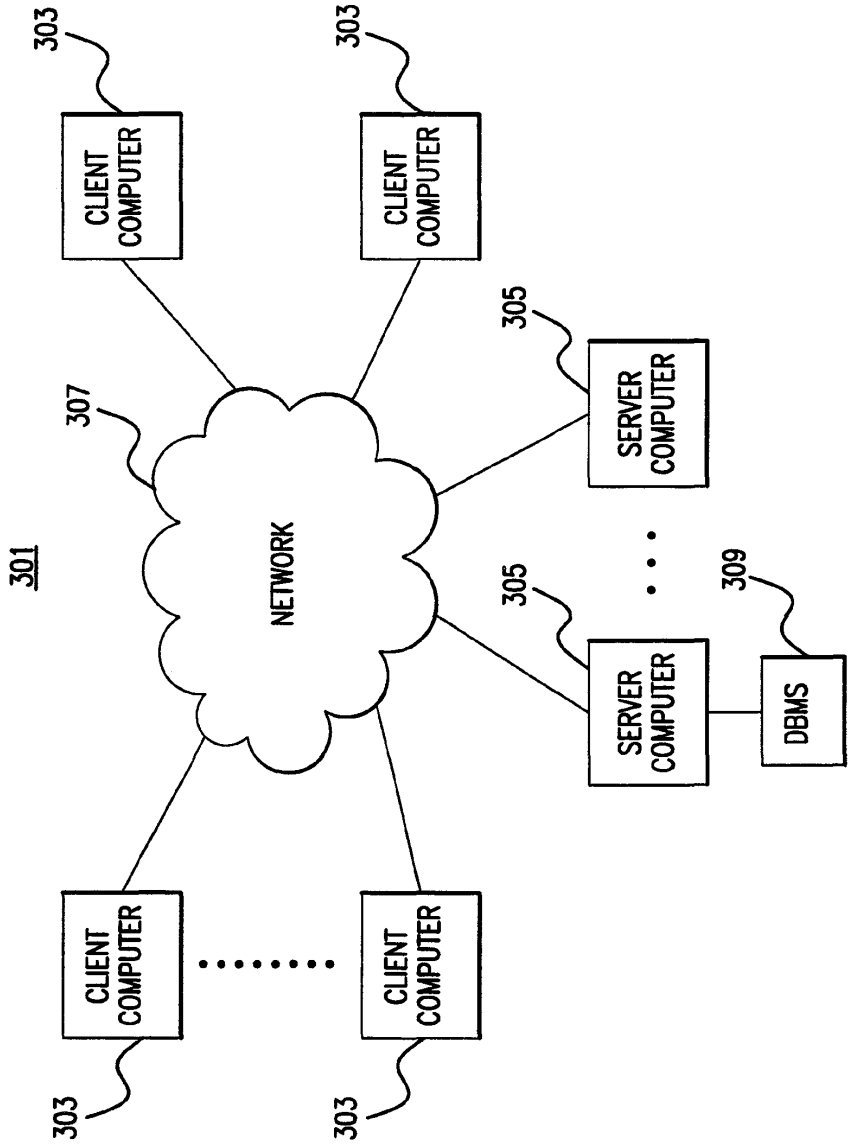


FIG.9

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	Date of mailing of the international search report
20 MAY 2000	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) 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

Amazon/VB Assets
Exhibit 1003
Page 994

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

Amazon/VB Assets
Exhibit 1003
Page 995

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

Amazon/VB Assets
Exhibit 1003
Page 996

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

Amazon/VB Assets
Exhibit 1003
Page 997

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

Amazon/VB Assets
Exhibit 1003
Page 998

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

Amazon/VB Assets
Exhibit 1003
Page 999

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

Amazon/VB Assets
Exhibit 1003
Page 1000

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

Amazon/VB Assets
Exhibit 1003
Page 1001

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

Amazon/VB Assets
Exhibit 1003
Page 1002

"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

Amazon/VB Assets
Exhibit 1003
Page 1003

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

Amazon/VB Assets
Exhibit 1003
Page 1004

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

Amazon/VB Assets
Exhibit 1003
Page 1005

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.

Amazon/VB Assets
Exhibit 1003
Page 1006

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.

Amazon/VB Assets
Exhibit 1003
Page 1007

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

Amazon/VB Assets
Exhibit 1003
Page 1008

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

Amazon/VB Assets
Exhibit 1003
Page 1009

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

Amazon/VB Assets
Exhibit 1003
Page 1010

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

Amazon/VB Assets
Exhibit 1003
Page 1011

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

Amazon/VB Assets
Exhibit 1003
Page 1012

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,

Amazon/VB Assets
Exhibit 1003
Page 1013

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

Amazon/VB Assets
Exhibit 1003
Page 1014

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

Amazon/VB Assets
Exhibit 1003
Page 1015

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")

Amazon/VB Assets
Exhibit 1003
Page 1016

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

Amazon/VB Assets
Exhibit 1003
Page 1017

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**

Amazon/VB Assets
Exhibit 1003
Page 1018

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

Amazon/VB Assets
Exhibit 1003
Page 1019

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

Amazon/VB Assets
Exhibit 1003
Page 1020

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">
-->
```


Amazon/VB Assets
Exhibit 1003
Page 1021

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

Amazon/VB Assets
Exhibit 1003
Page 1022

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

Amazon/VB Assets
Exhibit 1003
Page 1023

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.

Amazon/VB Assets
Exhibit 1003
Page 1024

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.

Amazon/VB Assets
Exhibit 1003
Page 1025

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:

Amazon/VB Assets
Exhibit 1003
Page 1026

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.

Amazon/VB Assets
Exhibit 1003
Page 1027

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

Amazon/VB Assets
Exhibit 1003
Page 1028

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:

Amazon/VB Assets
Exhibit 1003
Page 1029

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.

Amazon/VB Assets
Exhibit 1003
Page 1030

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

Amazon/VB Assets
Exhibit 1003
Page 1031

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

Amazon/VB Assets
Exhibit 1003
Page 1032

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.

Amazon/VB Assets
Exhibit 1003
Page 1033

Amazon/VB Assets
Exhibit 1003
Page 1034

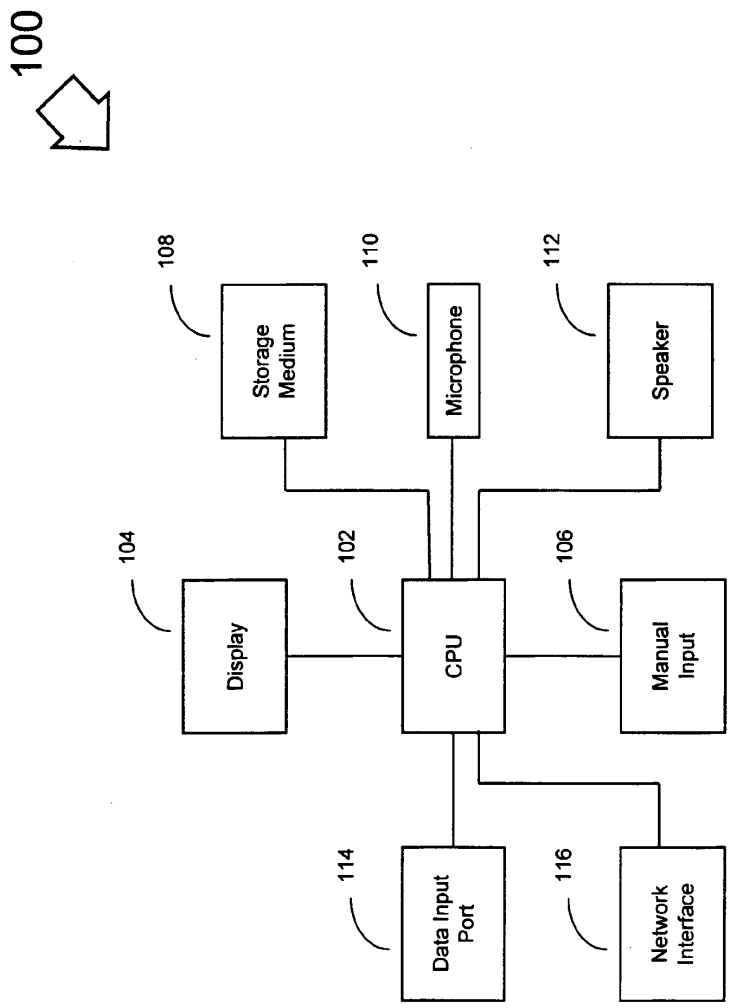


FIG. 1

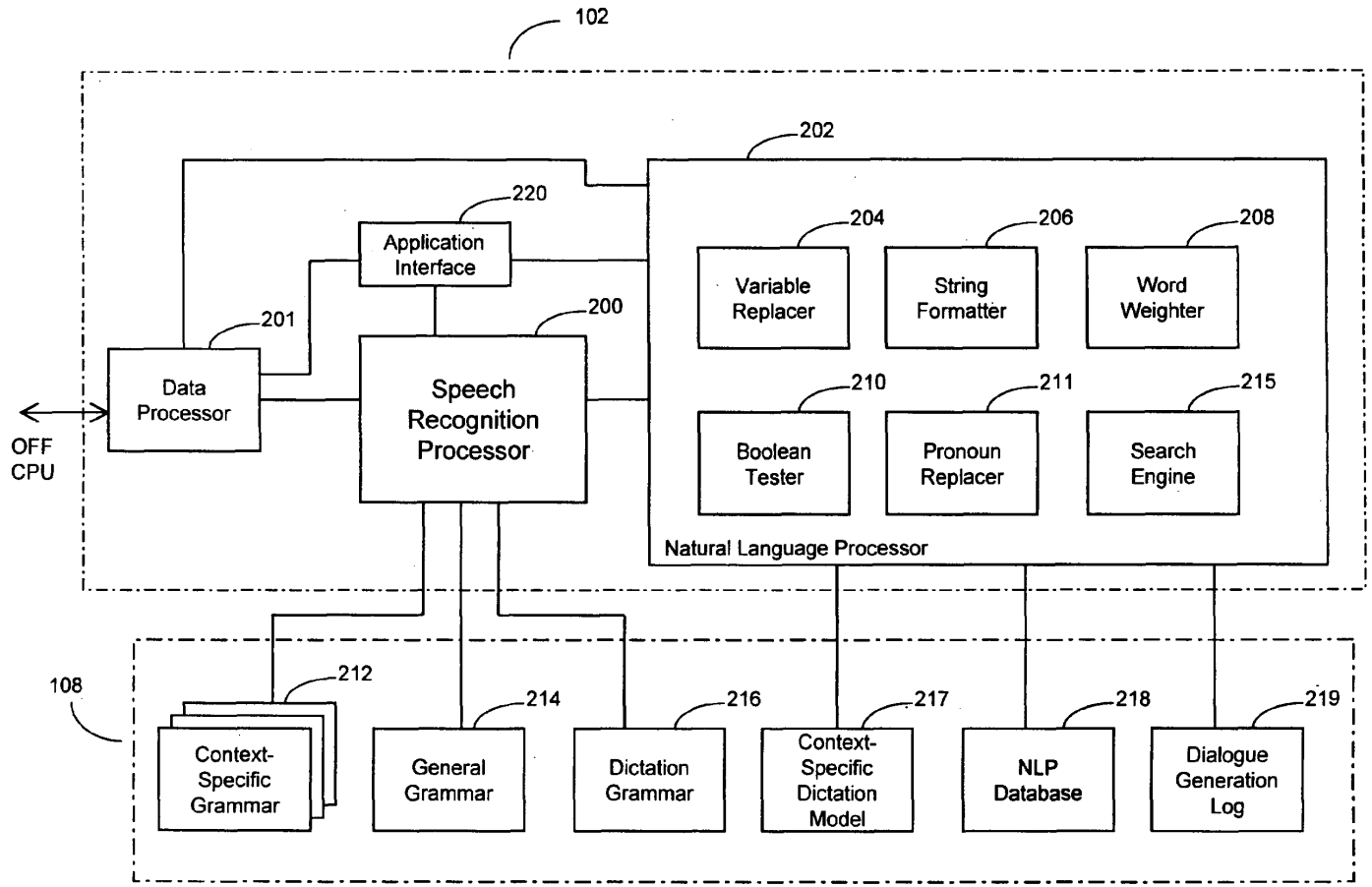

102
202
220
Application Interface
204
206
208
Variable Replacer
String Formatter
Word Weighter
201
200
Data Processor
Speech Recognition Processor
210
211
215
Boolean Tester
Pronoun Replacer
Search Engine
OFF CPU
Natural Language Processor
108
212
214
216
217
218
219
Context-Specific Grammar
General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

FIG. 2

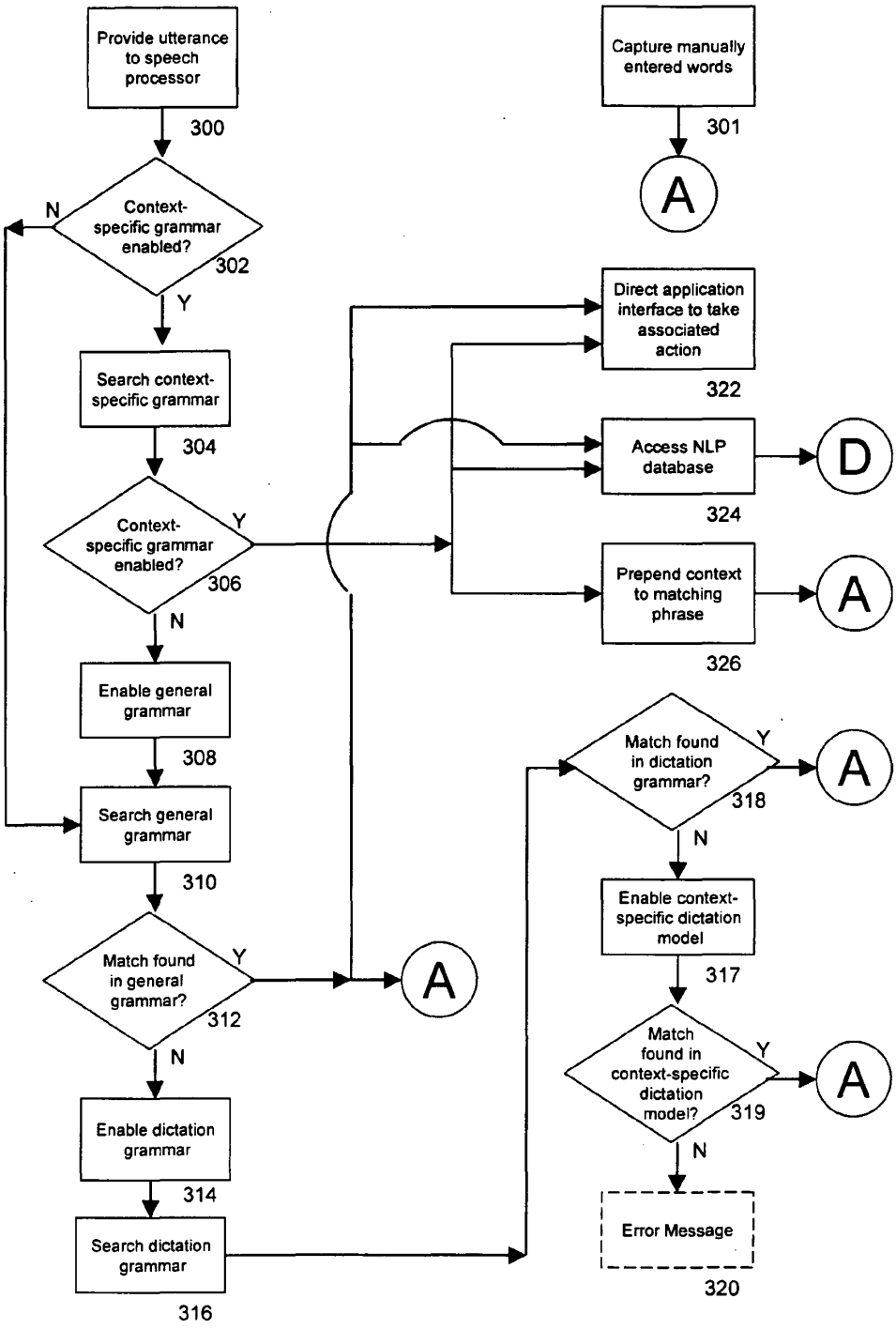

Provide utterance to speech processor
300
Context-specific grammar enabled?
302
N
Y
Search context-specific grammar
304
Context-specific grammar enabled?
306
Y
N
Enable general grammar
308
Search general grammar
310
Match found in general grammar?
312
Y
N
Enable dictation grammar
314
Search dictation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

FIG. 3A

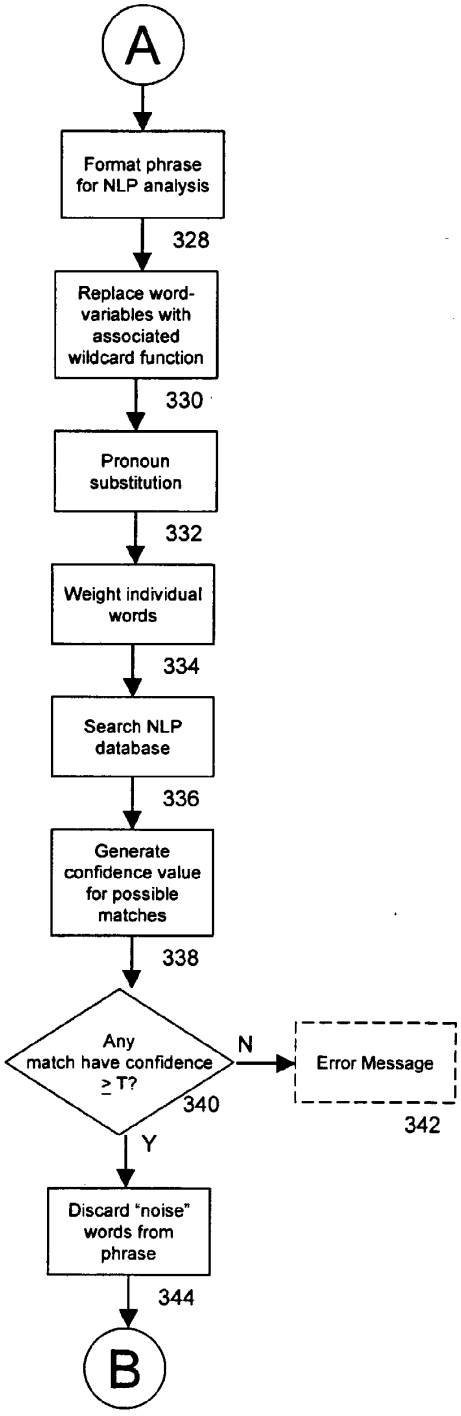

A
Format phrase for NLP analysis
328
Replace word-variables with associated wildcard function
330
Pronoun substitution
332
Weight individual words
334
Search NLP database
336
Generate confidence value for possible matches
338
Any match have confidence ≥ T?
340
N
Error Message
342
Y
Discard "noise" words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

FIG. 3B

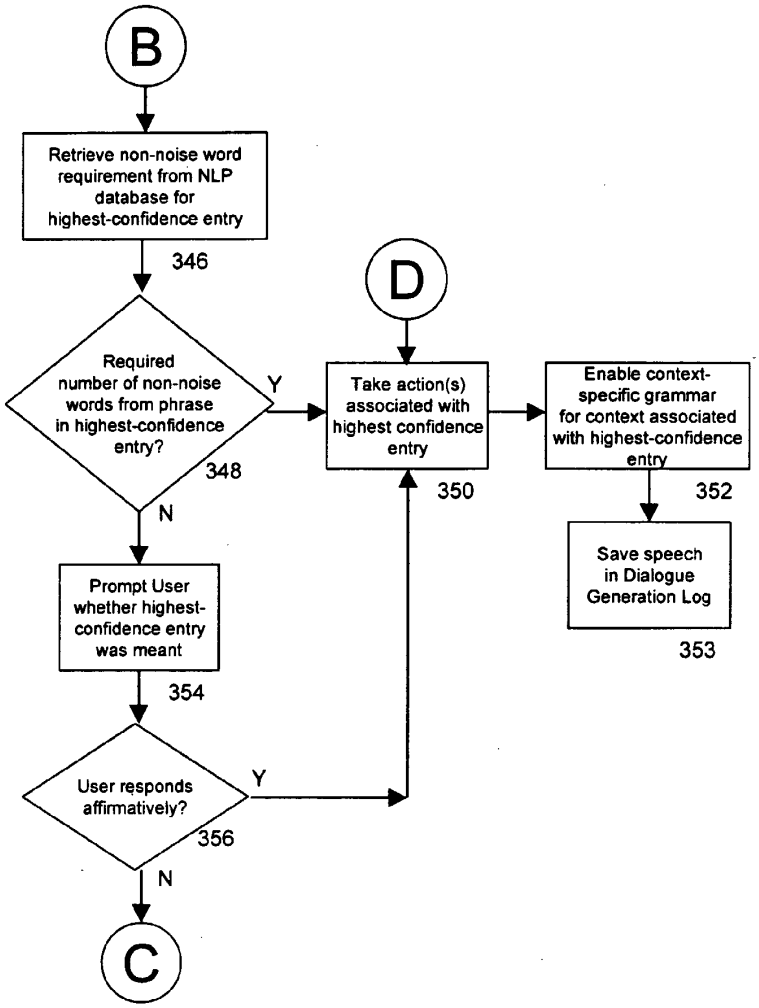

B
Retrieve non-noise word requirement from NLP database for highest-confidence entry
346
D
Required number of non-noise words from phrase in highest-confidence entry?
348
Y
N
Take action(s) associated with highest confidence entry
350
Enable context-specific grammar for context associated with highest-confidence 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

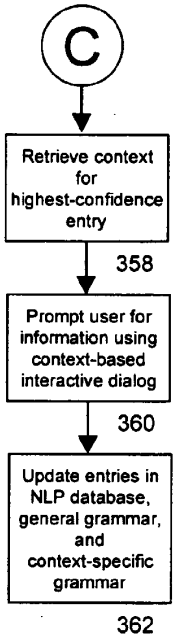

C
Retrieve context for highest-confidence entry
358
Prompt user for information using context-based interactive dialog
360
Update entries in NLP database, general grammar, and context-specific grammar
362

FIG. 3D



412A-*n*

400 Phrase	402 Required words	404 Context/Subcontext	408 Action 1	410 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

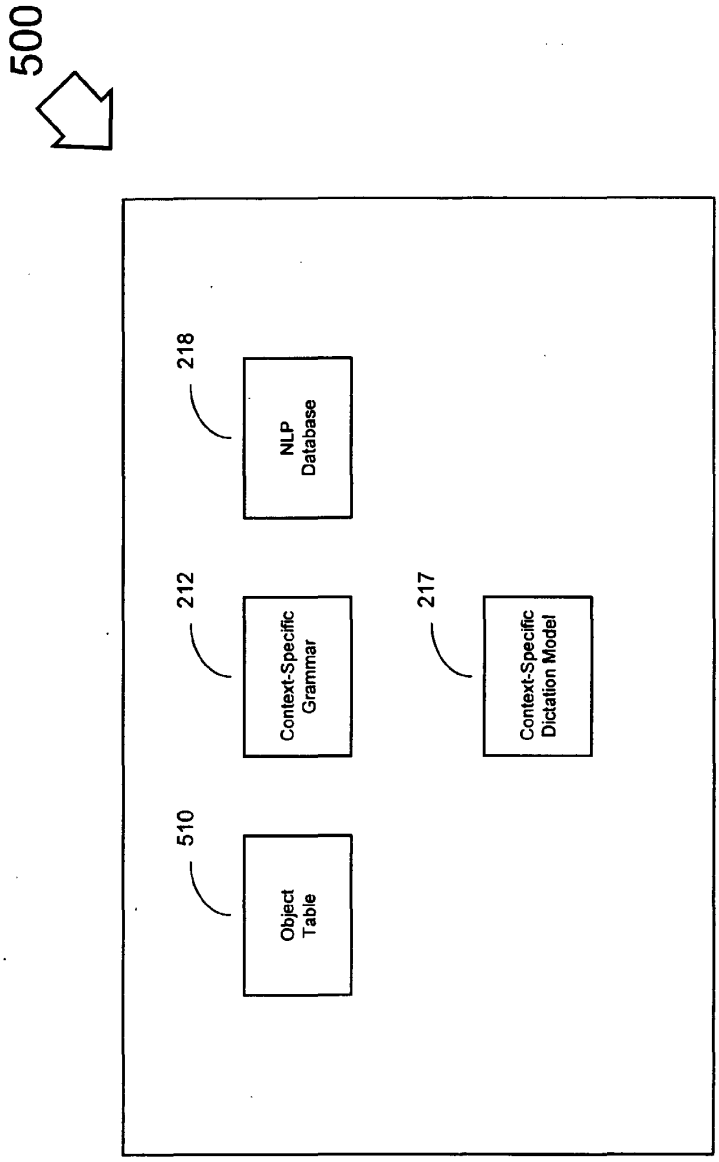

500
218
NLP Database
212
Context-Specific Grammar
217
Context-Specific Dictation Model
510
Object Table

FIG. 5

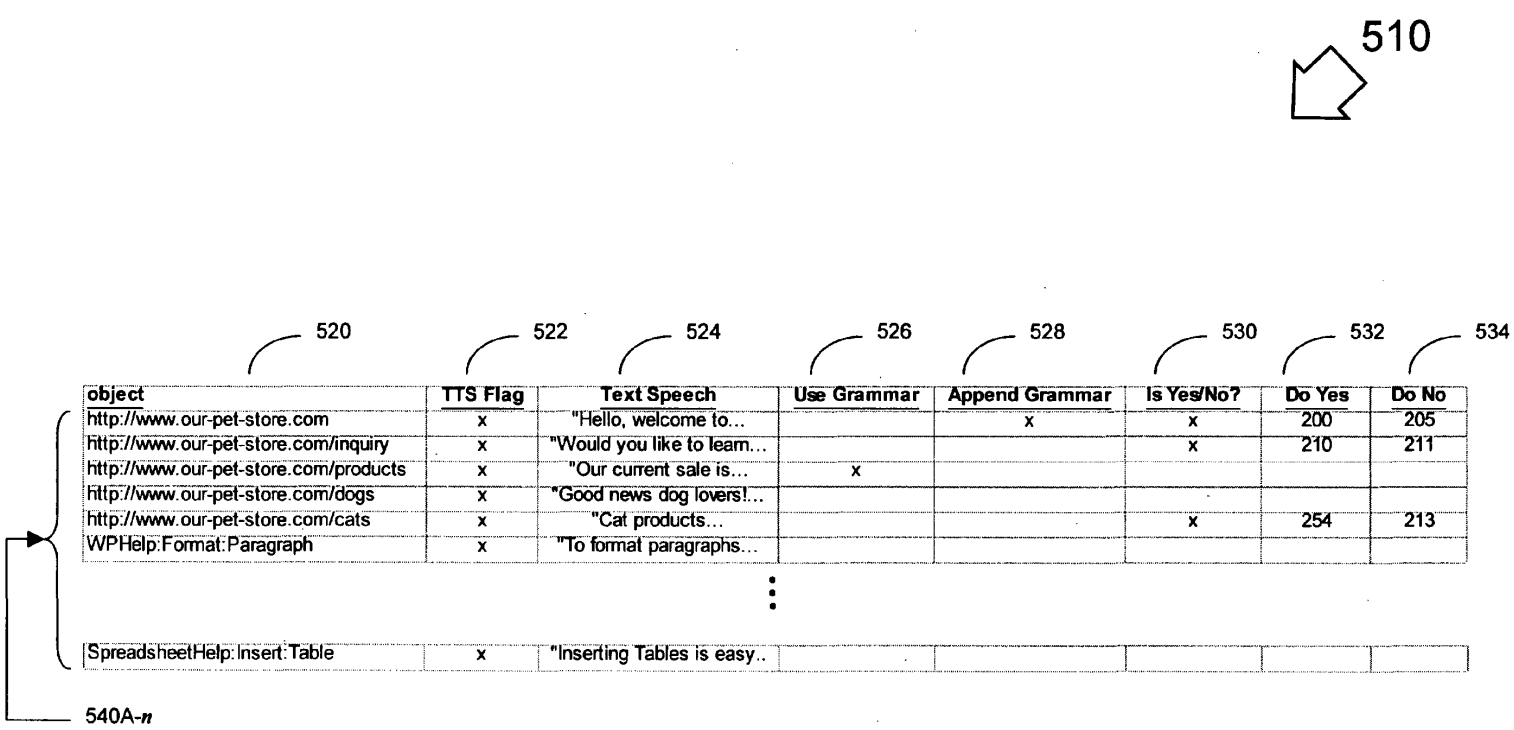

510
520
522
524
526
528
530
532
534
object
TTS Flag
Text Speech
Use Grammar
Append Grammar
Is Yes/No?
Do Yes
Do No
http://www.our-pet-store.com
x
"Hello, welcome to...
x
x
200
205
http://www.our-pet-store.com/inquiry
x
"Would you like to learn...
x
210
211
http://www.our-pet-store.com/products
x
"Our 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

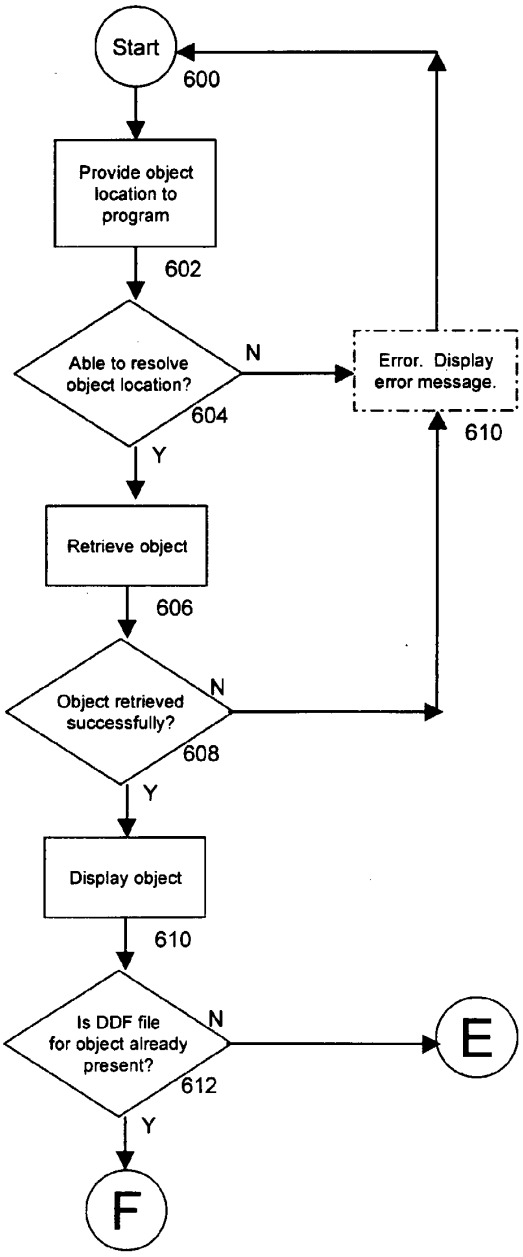

Start
600
Provide object location to program
602
Able to resolve object location?
604
N
Error. Display error message.
610
Y
Retrieve object
606
Object retrieved successfully?
608
N
Y
Display object
610
Is DDF file for object already present?
612
N
E
Y
F

FIG. 7A

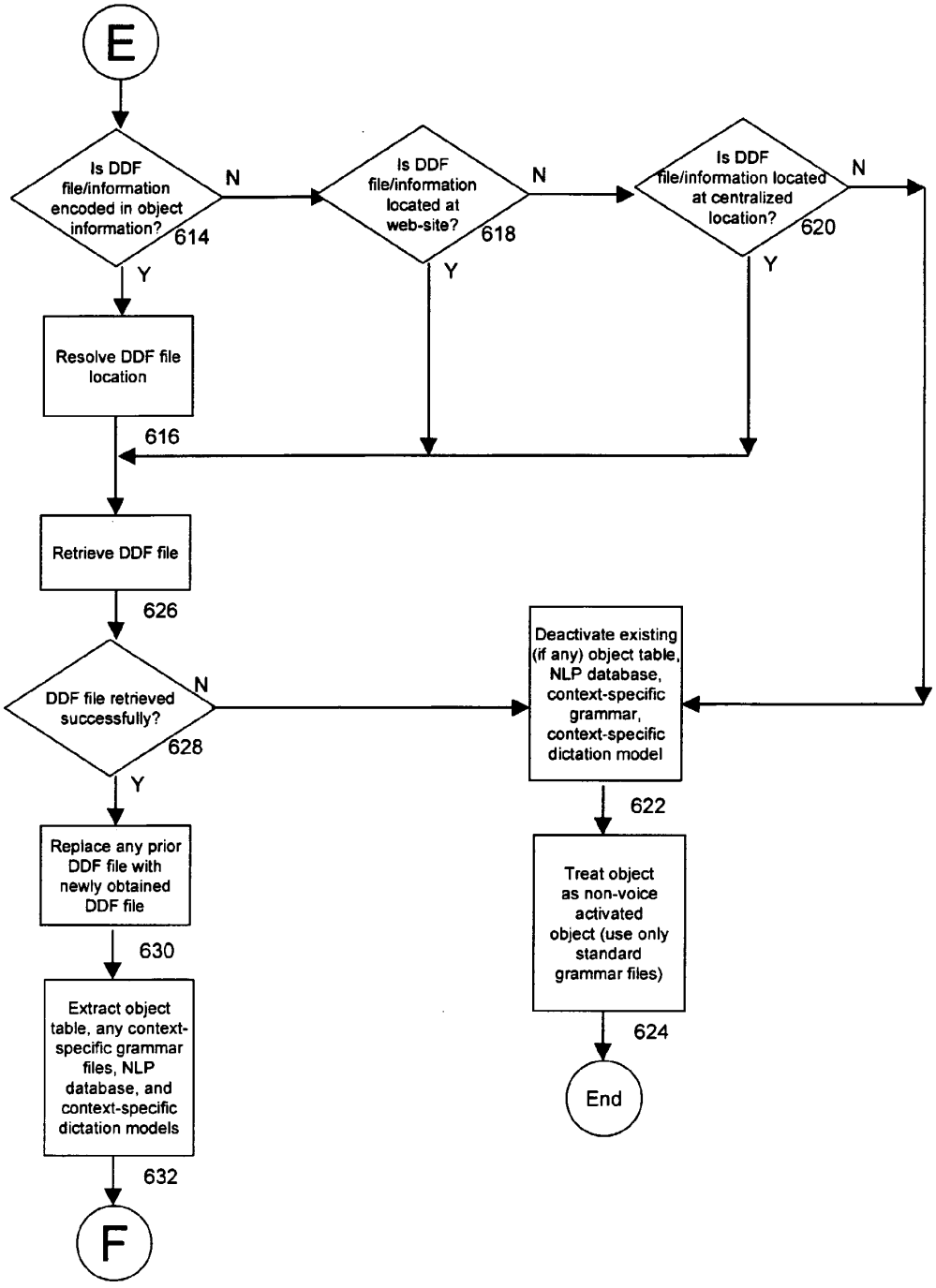

E
Is DDF file/information encoded in object information?
614
N
Y
Is DDF file/information located at web-site?
618
N
Y
Is DDF file/information located at centralized location?
620
N
Y
Resolve DDF file location
616
Retrieve DDF file
626
DDF file retrieved successfully?
628
N
Y
Deactivate existing (if any) 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

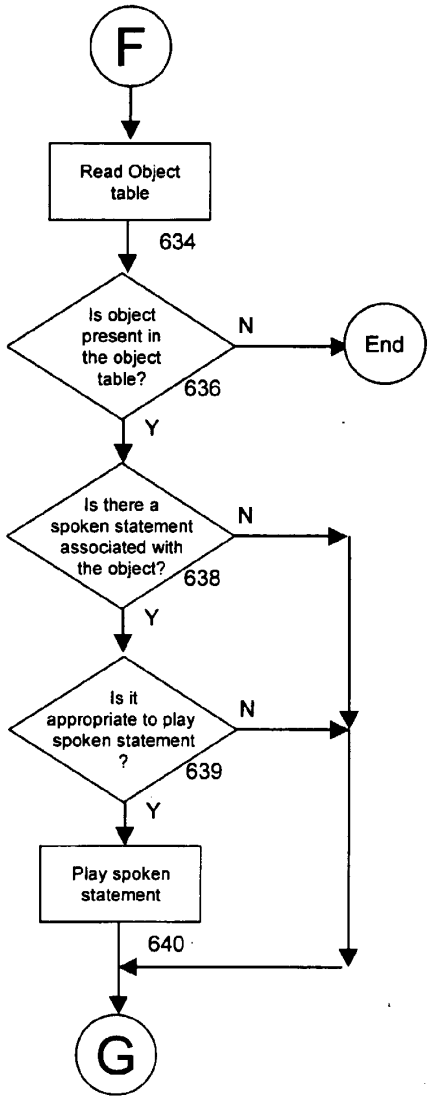

F
Read Object table
634
Is object present in the object table?
636
N
End
Y
Is there a spoken statement associated with the object?
638
N
Y
Is it appropriate to play spoken statement ?
639
N
Y
Play spoken statement
640
G

FIG. 7C

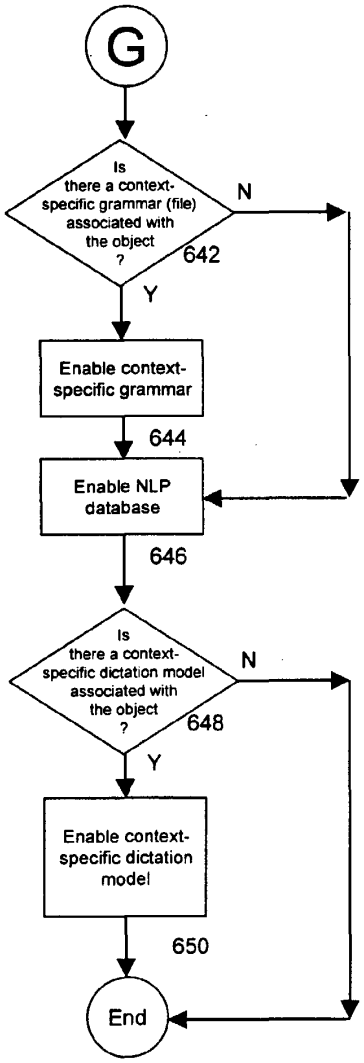

G
Is there a context-specific grammar (file) associated with the object ?
642
N
Y
Enable context-specific grammar
644
Enable NLP database
646
Is there a context-specific dictation model associated with the object ?
648
N
Y
Enable context-specific dictation model
650
End

FIG. 7D

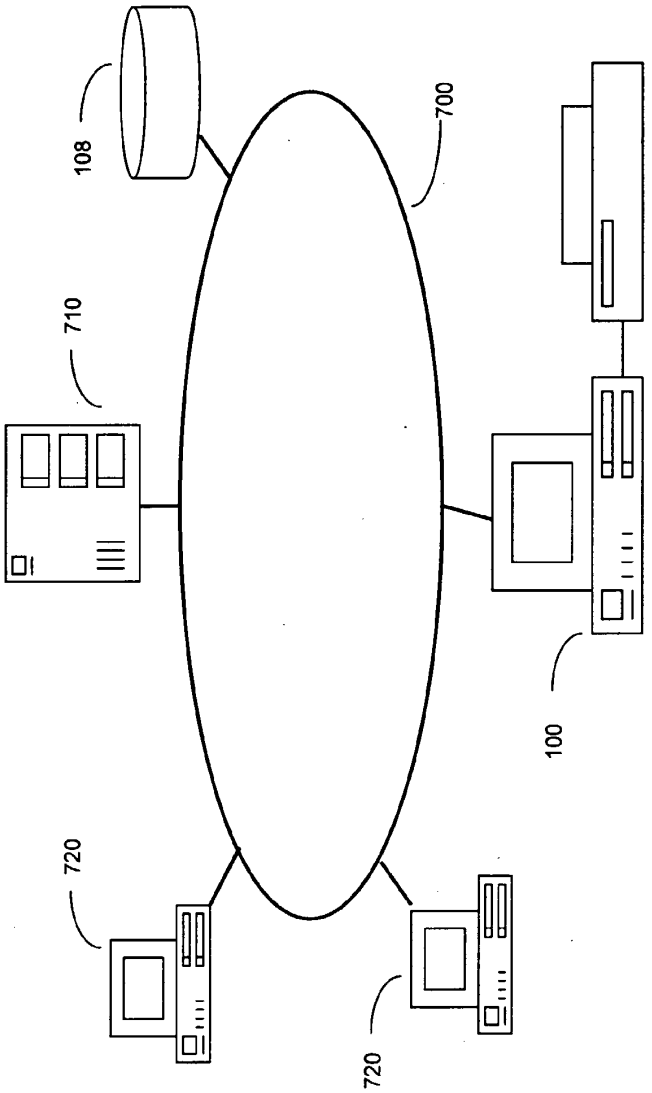


FIG. 8

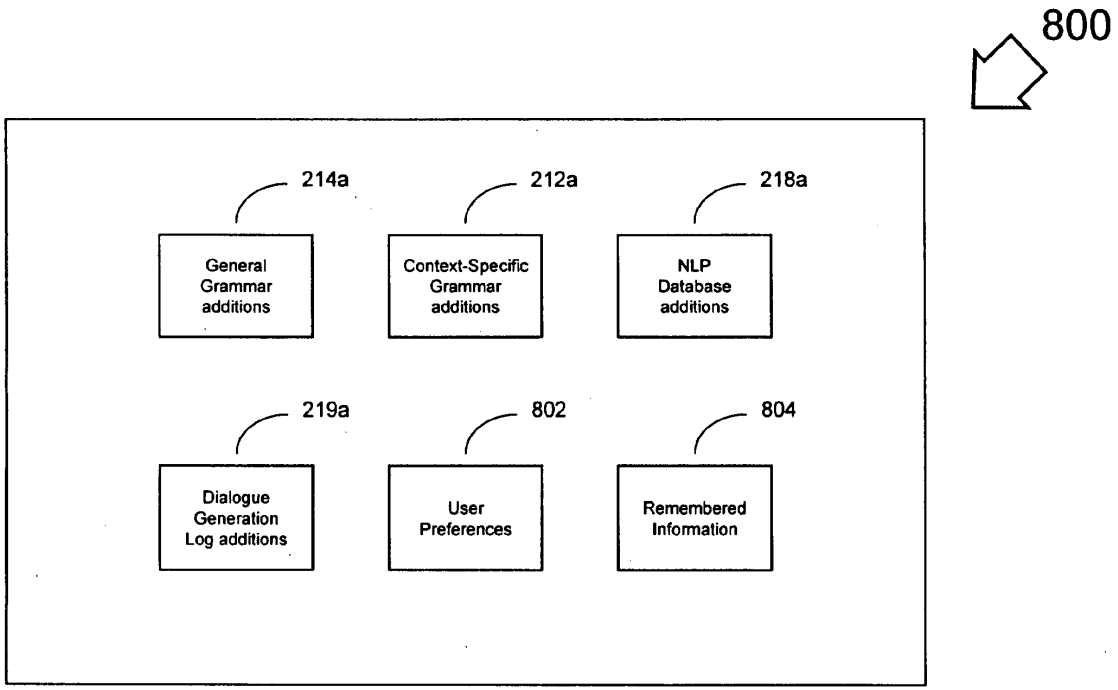


FIG. 9

Amazon/VB Assets
Exhibit 1003
Page 1048

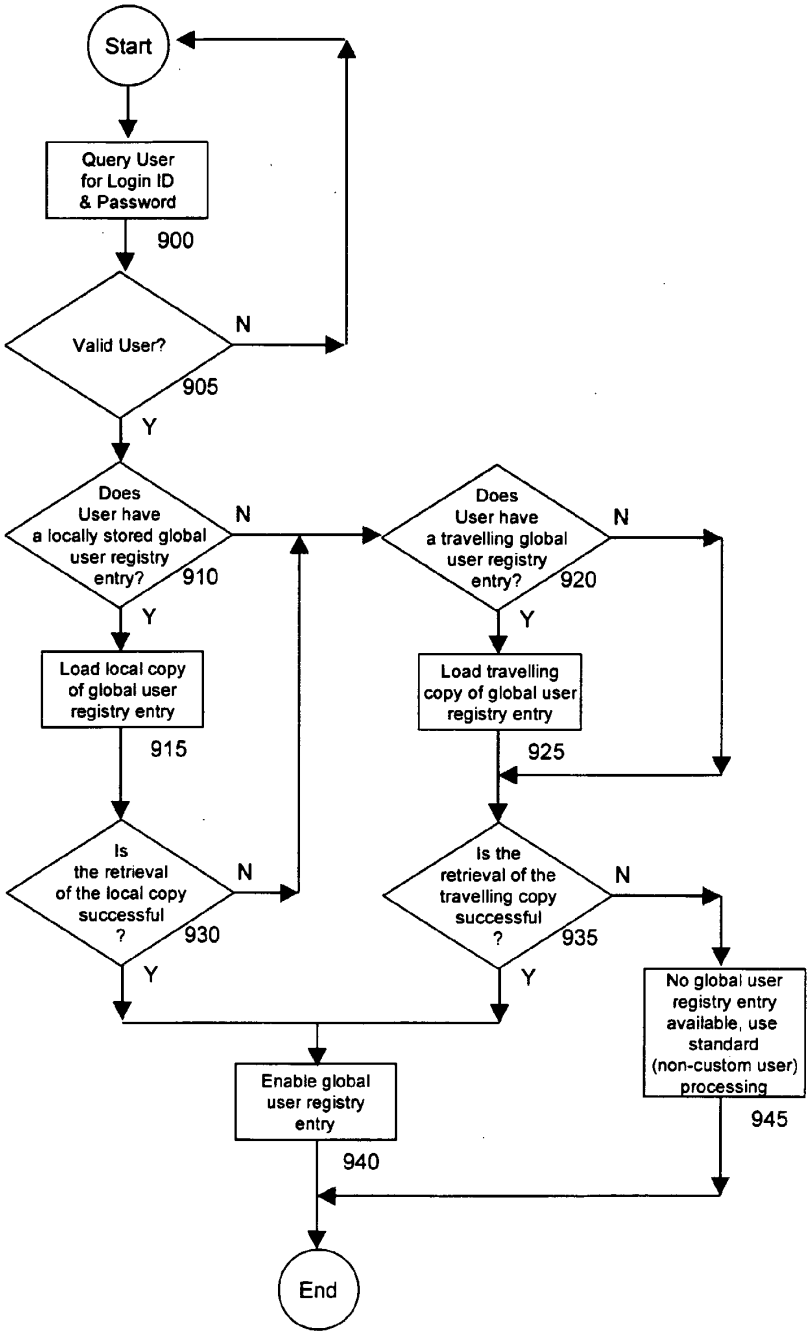

Start
Query User for Login ID & Password
900
Valid User?
905
N
Y
Does User have a locally stored global user registry entry?
910
N
Y
Load local copy of global user registry entry
915
Is the retrieval of the local copy successful ?
930
N
Y
Does User have a travelling global user registry entry?
920
N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

FIG. 10A

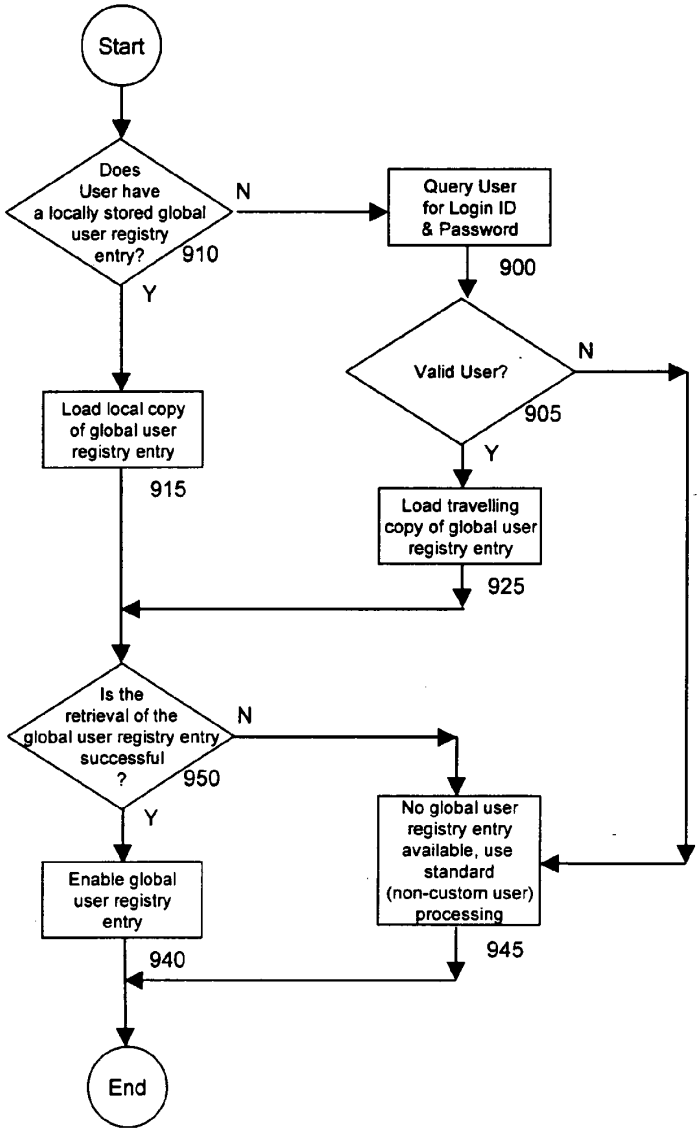

Start
Does User have a locally stored global user registry entry?
910
N
Y
Query User for Login ID & Password
900
Valid User?
905
N
Y
Load local copy of global user registry entry
915
Load travelling copy of global user registry entry
925
Is the retrieval of the global user registry entry successful ?
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

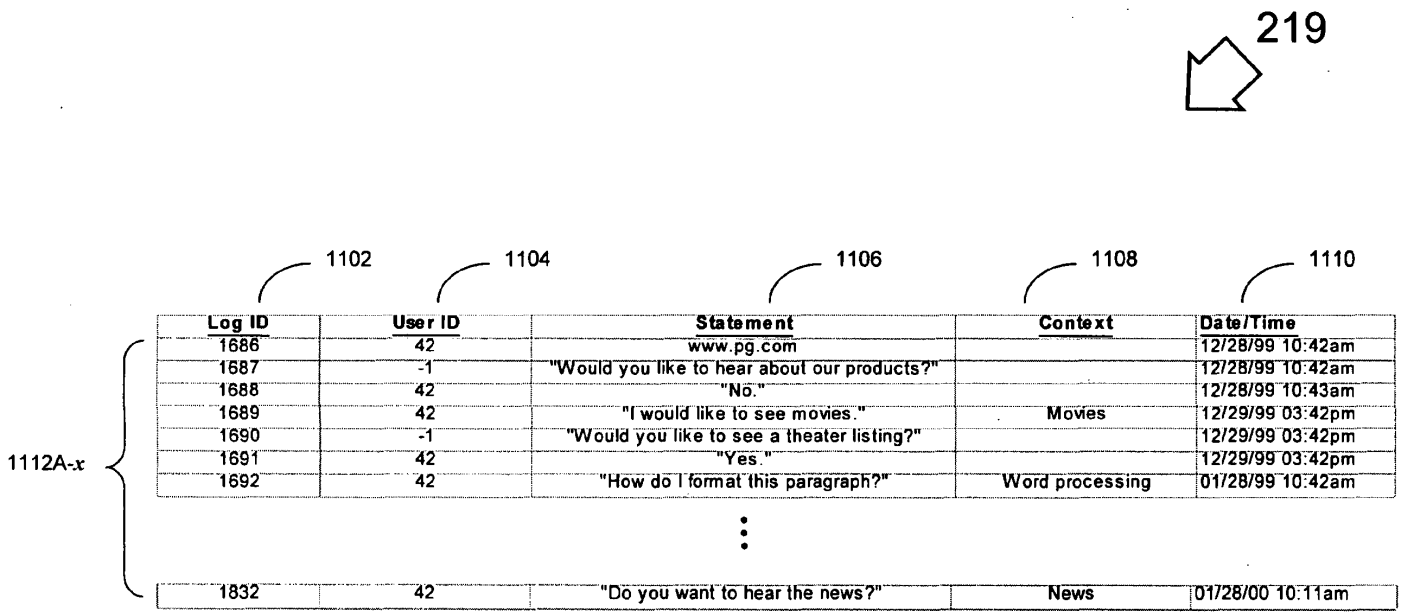

219
1102
1104
1106
1108
1110
1112A-x
Log ID
User ID
Statement
Context
Date/Time
1686
42
www.pg.com
12/28/99 10:42am
1687
-1
"Would you like to hear about our products?"
12/28/99 10:42am
1688
42
"No."
12/28/99 10:43am
1689
42
"I would like to see movies."
Movies
12/29/99 03:42pm
1690
-1
"Would you like 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

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)

INTERNATIONAL SEARCH REPORT

International Application No
PCT/US 01/11138

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

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

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

Amazon/VB Assets
Exhibit 1003
Page 1054

(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).

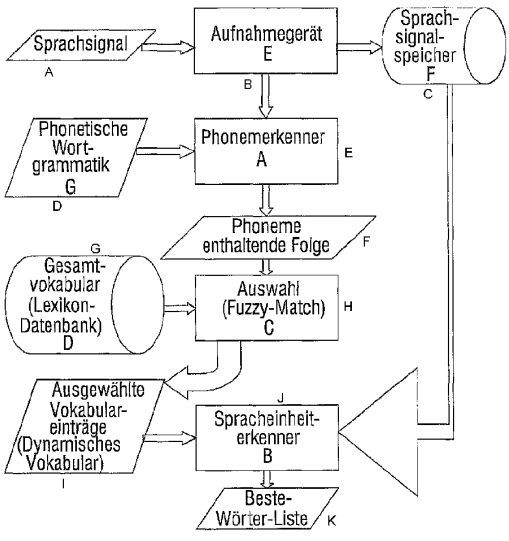
(74) Gemeinsamer Vertreter: **SIEMENS AKTIENGESELLSCHAFT**; Postfach 22 16 34, 80506 München (DE).

(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.

Amazon/VB Assets
Exhibit 1003
Page 1065

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.

Amazon/VB Assets
Exhibit 1003
Page 1069

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

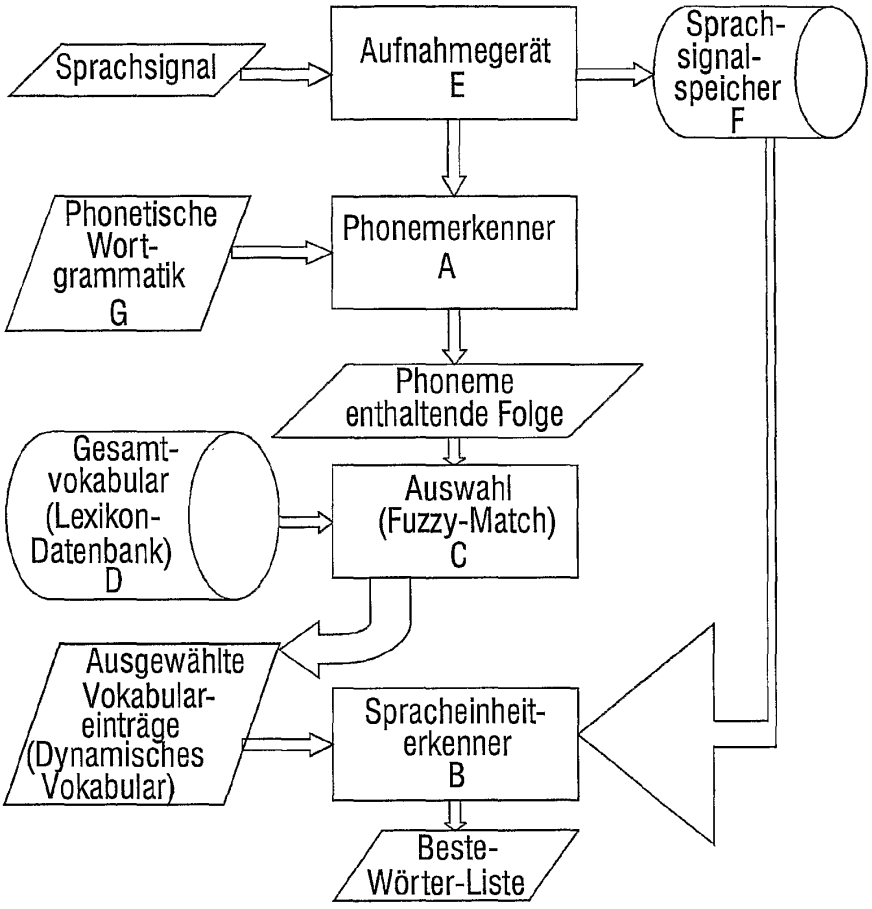

Sprachsignal
Aufnahmegerät E
Sprach-signal-speicher F
Phonetische Wort-grammatik G
Phonemerkenner A
Phoneme enthaltende Folge
Gesamt-vokabular (Lexikon-Datenbank) D
Auswahl (Fuzzy-Match) C
Ausgewählte Vokabular-einträge (Dynamisches Vokabular)
Spracheinheit-erkenner B
Beste-Wörter-Liste

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

Amazon/VB Assets
Exhibit 1003
Page 1077

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

Amazon/VB Assets
Exhibit 1003
Page 1078

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

Amazon/VB Assets
Exhibit 1003
Page 1079

(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 non-speech communications and presenting the expected results for a particular question or command.

Amazon/VB Assets
Exhibit 1003
Page 1080

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.

Amazon/VB Assets
Exhibit 1003
Page 1081

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

Amazon/VB Assets
Exhibit 1003
Page 1082

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

Amazon/VB Assets
Exhibit 1003
Page 1083

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

Amazon/VB Assets
Exhibit 1003
Page 1084

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

Amazon/VB Assets
Exhibit 1003
Page 1085

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,

Amazon/VB Assets
Exhibit 1003
Page 1086

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

Amazon/VB Assets
Exhibit 1003
Page 1087

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

Amazon/VB Assets
Exhibit 1003
Page 1088

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.

Amazon/VB Assets
Exhibit 1003
Page 1089

[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

Amazon/VB Assets
Exhibit 1003
Page 1090

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;

Amazon/VB Assets
Exhibit 1003
Page 1091

- 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;

Amazon/VB Assets
Exhibit 1003
Page 1092

- 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

Amazon/VB Assets
Exhibit 1003
Page 1093

[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,

Amazon/VB Assets
Exhibit 1003
Page 1094

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

Amazon/VB Assets
Exhibit 1003
Page 1095

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

Amazon/VB Assets
Exhibit 1003
Page 1096

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.

Amazon/VB Assets
Exhibit 1003
Page 1097

[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

Amazon/VB Assets
Exhibit 1003
Page 1098

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.

Amazon/VB Assets
Exhibit 1003
Page 1099

[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,

Amazon/VB Assets
Exhibit 1003
Page 1100

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

Amazon/VB Assets
Exhibit 1003
Page 1101

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

Amazon/VB Assets
Exhibit 1003
Page 1102

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:

Amazon/VB Assets
Exhibit 1003
Page 1103

- 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;

Amazon/VB Assets
Exhibit 1003
Page 1104

- 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.

Amazon/VB Assets
Exhibit 1003
Page 1105

[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

Amazon/VB Assets
Exhibit 1003
Page 1106

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

Amazon/VB Assets
Exhibit 1003
Page 1107

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

Amazon/VB Assets
Exhibit 1003
Page 1108

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

Amazon/VB Assets
Exhibit 1003
Page 1109

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

Amazon/VB Assets
Exhibit 1003
Page 1110

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

Amazon/VB Assets
Exhibit 1003
Page 1111

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

Amazon/VB Assets
Exhibit 1003
Page 1112

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

Amazon/VB Assets
Exhibit 1003
Page 1113

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

Amazon/VB Assets
Exhibit 1003
Page 1114

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

Amazon/VB Assets
Exhibit 1003
Page 1115

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.

Amazon/VB Assets
Exhibit 1003
Page 1116

[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

Amazon/VB Assets
Exhibit 1003
Page 1117

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

Amazon/VB Assets
Exhibit 1003
Page 1118

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.

Amazon/VB Assets
Exhibit 1003
Page 1119

[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.

Amazon/VB Assets
Exhibit 1003
Page 1120

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.

Amazon/VB Assets
Exhibit 1003
Page 1121

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;

Amazon/VB Assets
Exhibit 1003
Page 1122

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).

Amazon/VB Assets
Exhibit 1003
Page 1123

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).

Amazon/VB Assets
Exhibit 1003
Page 1124

FIG. 1

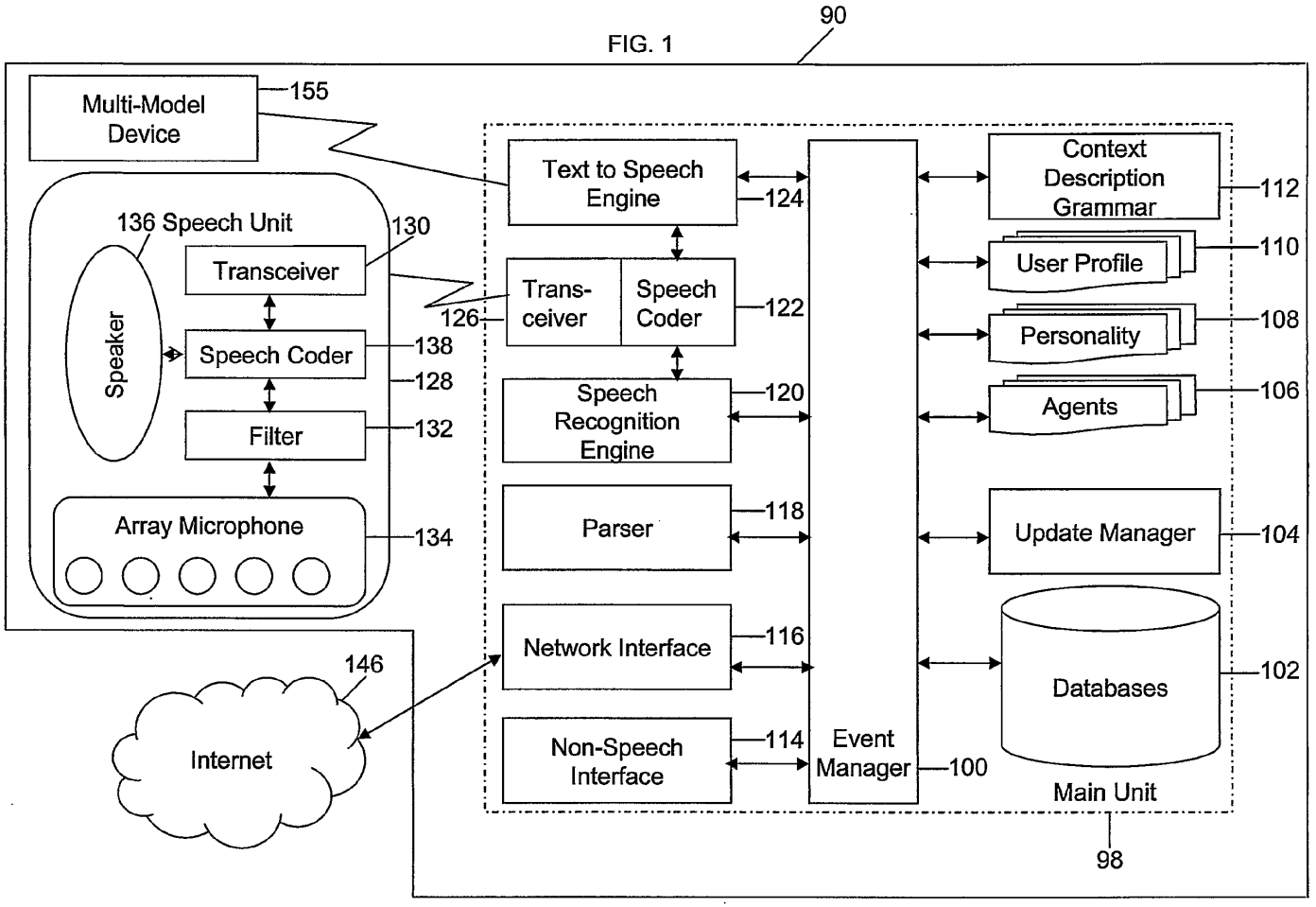

90
Multi-Model Device
155
136 Speech Unit
130
Transceiver
Speaker
Speech Coder
138
128
Filter
132
Array Microphone
134
126
Trans-ceiver
Speech Coder
122
Text to Speech Engine
124
Speech Recognition Engine
120
Parser
118
Network Interface
116
Non-Speech Interface
114
Event Manager
100
Main Unit
98
Context Description Grammar
112
User Profile
110
Personality
108
Agents
106
Update Manager
104
Databases
102
146
Internet

FIG. 2

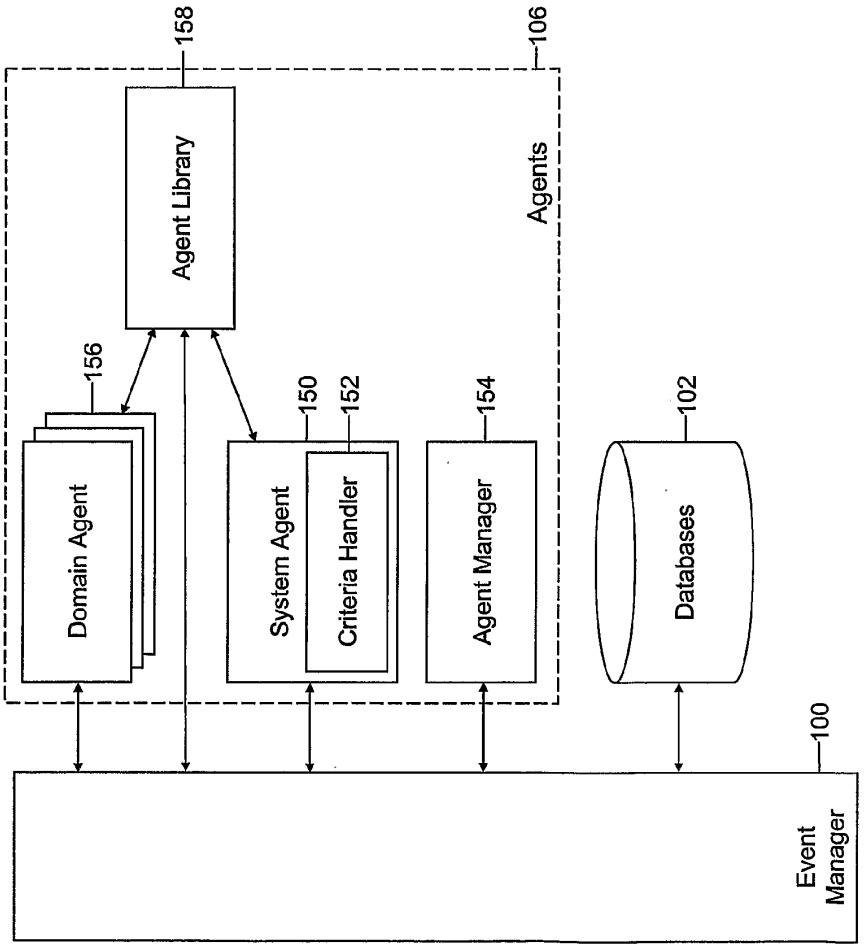

158
106
Agent Library
Agents
156
150
152
154
102
Domain Agent
System Agent
Criteria Handler
Agent Manager
Databases
100
Event Manager

FIG. 3

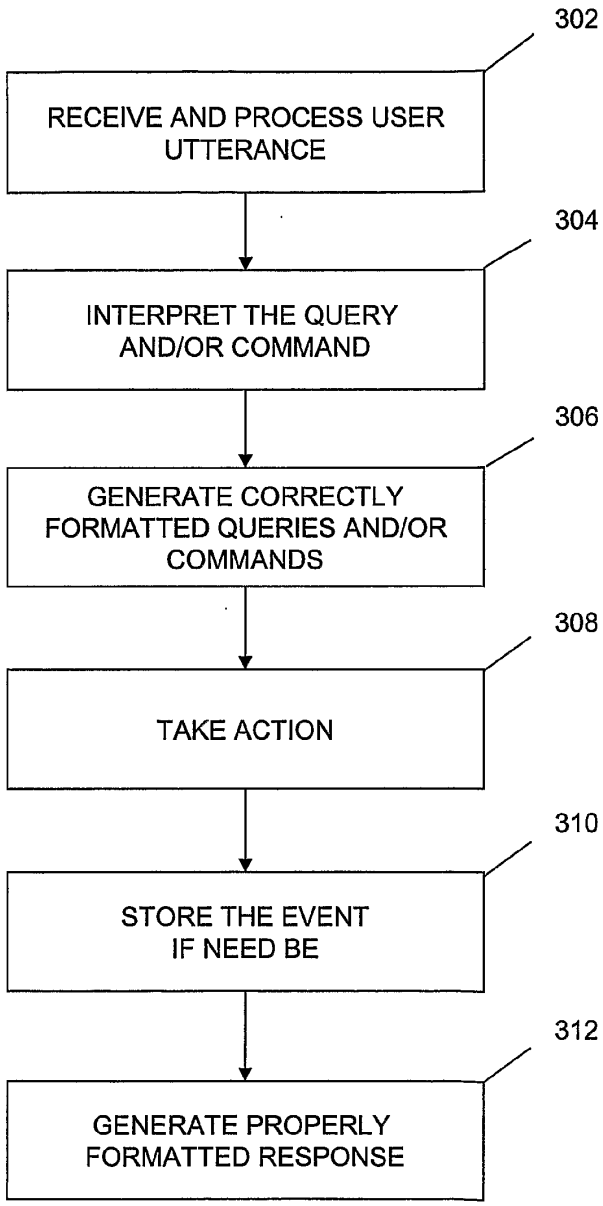

300
302
RECEIVE AND PROCESS USER UTTERANCE
304
INTERPRET THE QUERY AND/OR COMMAND
306
GENERATE CORRECTLY FORMATTED QUERIES AND/OR COMMANDS
308
TAKE ACTION
310
STORE THE EVENT IF NEED BE
312
GENERATE PROPERLY FORMATTED RESPONSE

FIG. 4A

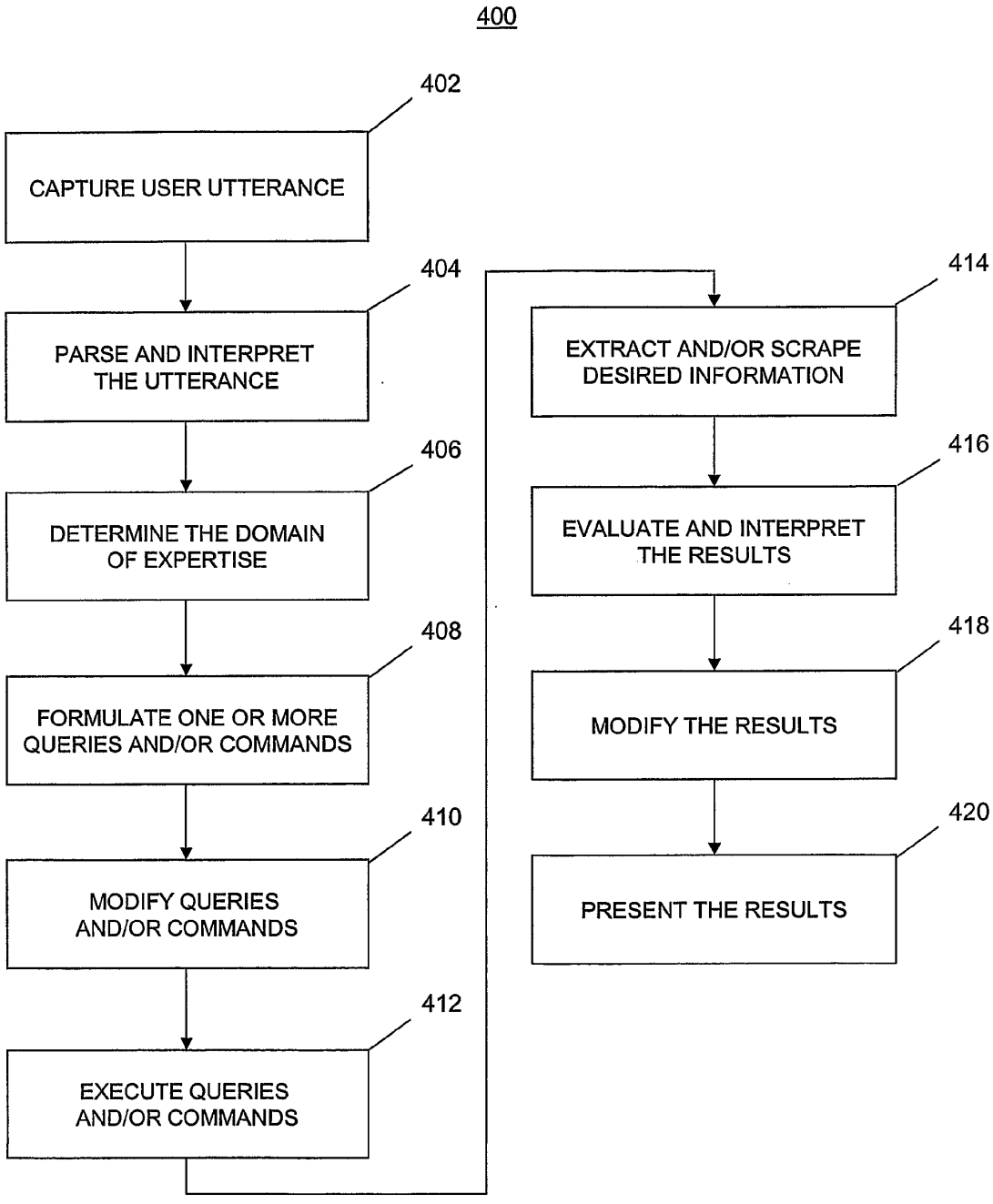
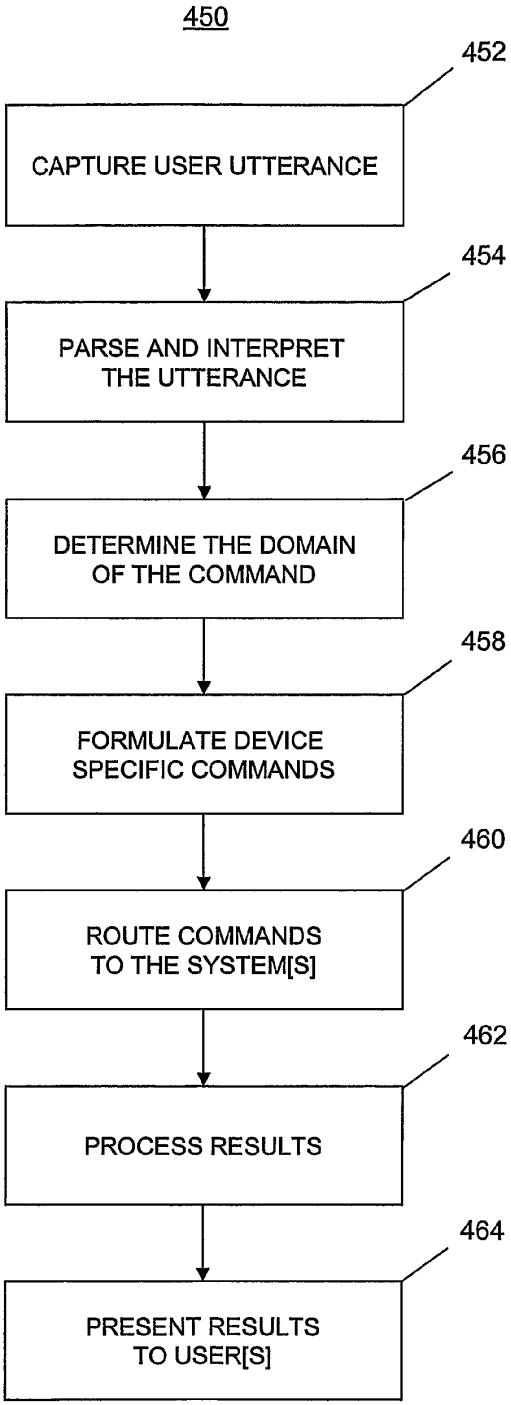

400
402
CAPTURE USER UTTERANCE
404
PARSE AND INTERPRET THE UTTERANCE
406
DETERMINE THE DOMAIN OF EXPERTISE
408
FORMULATE ONE OR MORE QUERIES AND/OR COMMANDS
410
MODIFY QUERIES AND/OR COMMANDS
412
EXECUTE QUERIES AND/OR COMMANDS
414
EXTRACT AND/OR SCRAPE DESIRED INFORMATION
416
EVALUATE AND INTERPRET THE RESULTS
418
MODIFY THE RESULTS
420
PRESENT THE RESULTS

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

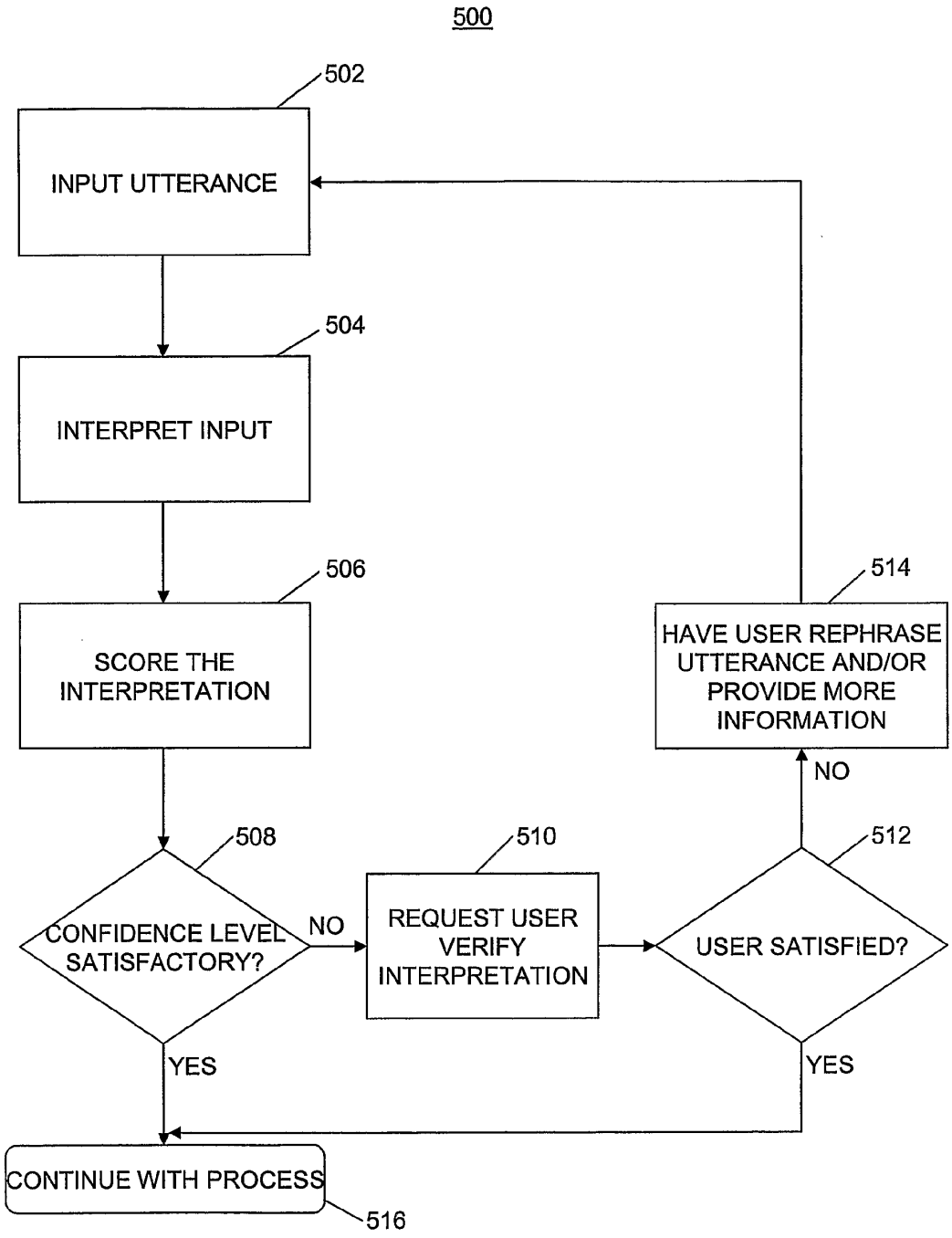

500
502
INPUT UTTERANCE
504
INTERPRET INPUT
506
SCORE THE INTERPRETATION
508
CONFIDENCE LEVEL SATISFACTORY?
NO
510
REQUEST USER VERIFY INTERPRETATION
512
USER SATISFIED?
NO
514
HAVE USER REPHRASE UTTERANCE AND/OR PROVIDE MORE INFORMATION
YES
YES
CONTINUE WITH PROCESS
516

FIG. 6

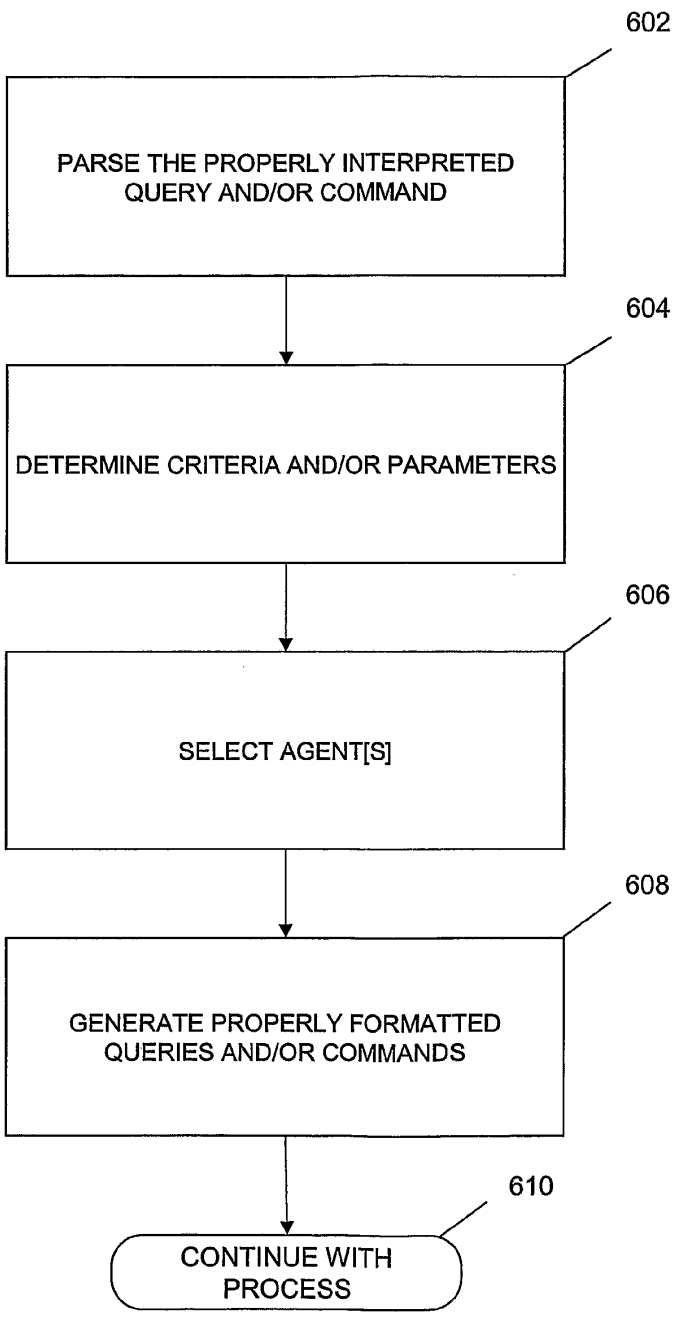
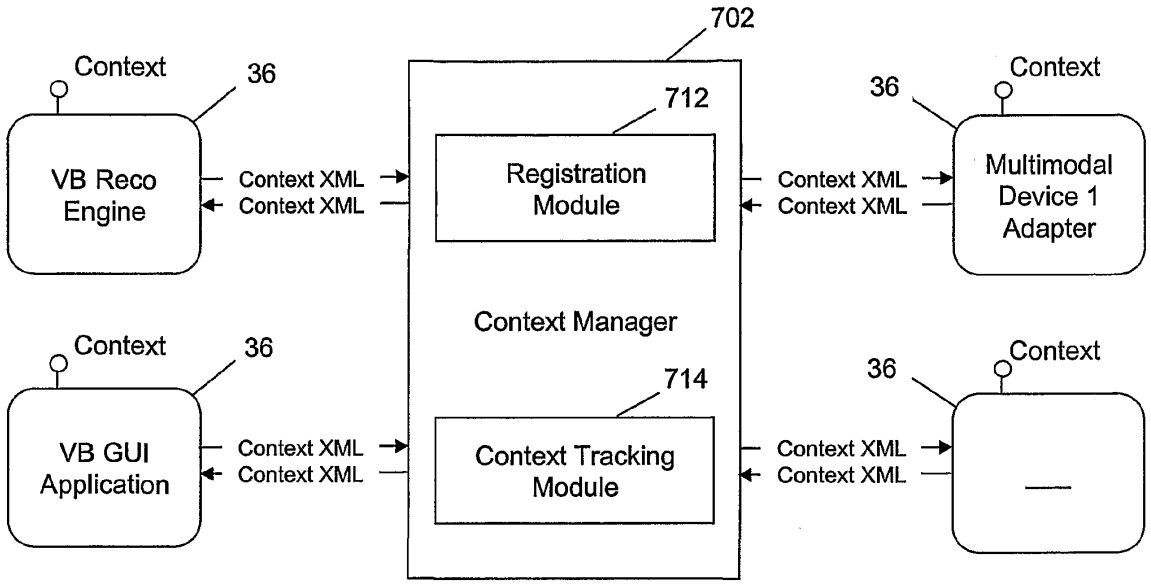

600
602
PARSE THE PROPERLY INTERPRETED QUERY AND/OR COMMAND
604
DETERMINE CRITERIA AND/OR PARAMETERS
606
SELECT AGENT[S]
608
GENERATE PROPERLY FORMATTED QUERIES AND/OR COMMANDS
610
CONTINUE WITH PROCESS

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.

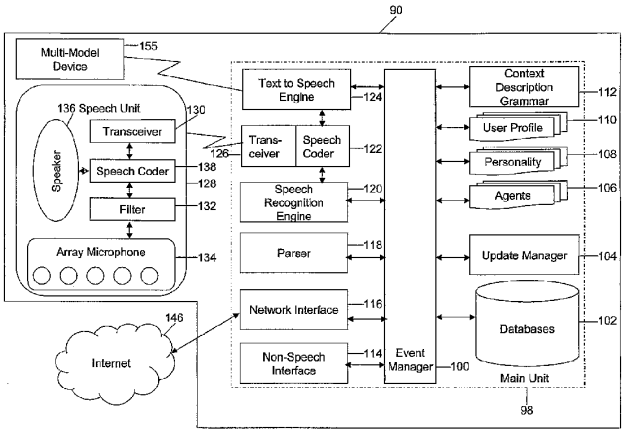
(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, 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

Amazon/VB Assets
Exhibit 1003
Page 1136

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

Amazon/VB Assets
Exhibit 1003
Page 1137

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

Amazon/VB Assets
Exhibit 1003
Page 1138

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

Amazon/VB Assets
Exhibit 1003
Page 1139

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

Amazon/VB Assets
Exhibit 1003
Page 1140

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

Amazon/VB Assets
Exhibit 1003
Page 1141

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.

Amazon/VB Assets
Exhibit 1003
Page 1142

[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

Amazon/VB Assets
Exhibit 1003
Page 1143

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.

Amazon/VB Assets
Exhibit 1003
Page 1144

[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;

Amazon/VB Assets
Exhibit 1003
Page 1145

- 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

Amazon/VB Assets
Exhibit 1003
Page 1146

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:

Amazon/VB Assets
Exhibit 1003
Page 1147

- 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;

Amazon/VB Assets
Exhibit 1003
Page 1148

- 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;

Amazon/VB Assets
Exhibit 1003
Page 1149

[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

Amazon/VB Assets
Exhibit 1003
Page 1150

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

Amazon/VB Assets
Exhibit 1003
Page 1151

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

Amazon/VB Assets
Exhibit 1003
Page 1152

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

Amazon/VB Assets
Exhibit 1003
Page 1153

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

Amazon/VB Assets
Exhibit 1003
Page 1154

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

Amazon/VB Assets
Exhibit 1003
Page 1155

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

Amazon/VB Assets
Exhibit 1003
Page 1156

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

Amazon/VB Assets
Exhibit 1003
Page 1157

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.

Amazon/VB Assets
Exhibit 1003
Page 1158

[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

Amazon/VB Assets
Exhibit 1003
Page 1159

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:

Amazon/VB Assets
Exhibit 1003
Page 1160

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.

Amazon/VB Assets
Exhibit 1003
Page 1161

[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

Amazon/VB Assets
Exhibit 1003
Page 1162

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-

Amazon/VB Assets
Exhibit 1003
Page 1163

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.

Amazon/VB Assets
Exhibit 1003
Page 1164

[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

Amazon/VB Assets
Exhibit 1003
Page 1165

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

Amazon/VB Assets
Exhibit 1003
Page 1166

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

Amazon/VB Assets
Exhibit 1003
Page 1167

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

Amazon/VB Assets
Exhibit 1003
Page 1168

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.

Amazon/VB Assets
Exhibit 1003
Page 1169

[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

Amazon/VB Assets
Exhibit 1003
Page 1170

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;

Amazon/VB Assets
Exhibit 1003
Page 1171

- 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.

Amazon/VB Assets
Exhibit 1003
Page 1172

[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

Amazon/VB Assets
Exhibit 1003
Page 1173

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

Amazon/VB Assets
Exhibit 1003
Page 1174

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.

Amazon/VB Assets
Exhibit 1003
Page 1175

[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

Amazon/VB Assets
Exhibit 1003
Page 1176

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,

Amazon/VB Assets
Exhibit 1003
Page 1177

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.

Amazon/VB Assets
Exhibit 1003
Page 1178

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).

Amazon/VB Assets
Exhibit 1003
Page 1179

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.

Amazon/VB Assets
Exhibit 1003
Page 1180

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.

Amazon/VB Assets
Exhibit 1003
Page 1181

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;

Amazon/VB Assets
Exhibit 1003
Page 1182

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.

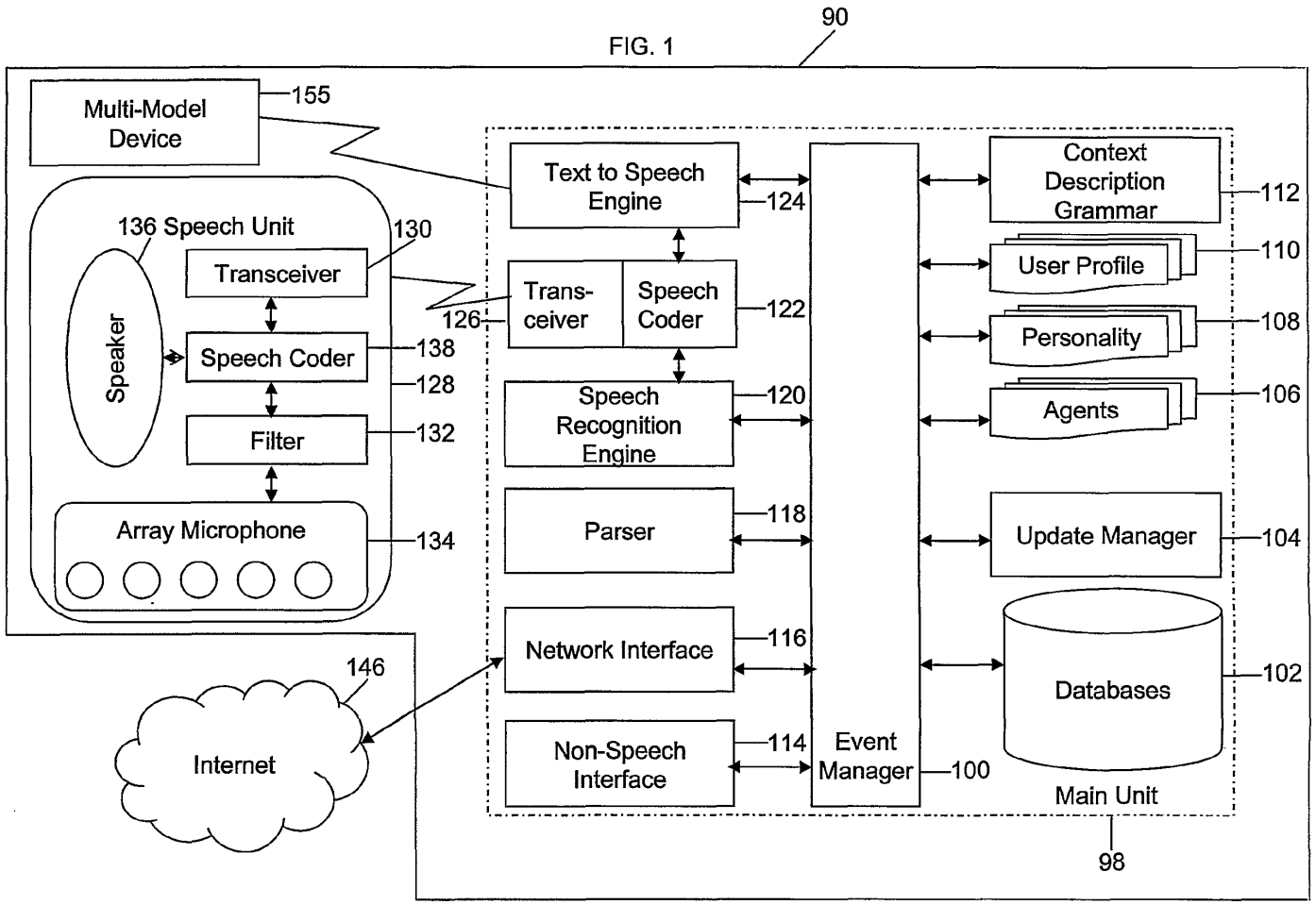
42. The method according to claim 37, further comprising a personality module that communicates the response to a user.

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).

Amazon/VB Assets
Exhibit 1003
Page 1183

FIG. 1



Amazon/VB Assets
Exhibit 1003
Page 1184

Amazon/VB Assets
Exhibit 1003
Page 1185

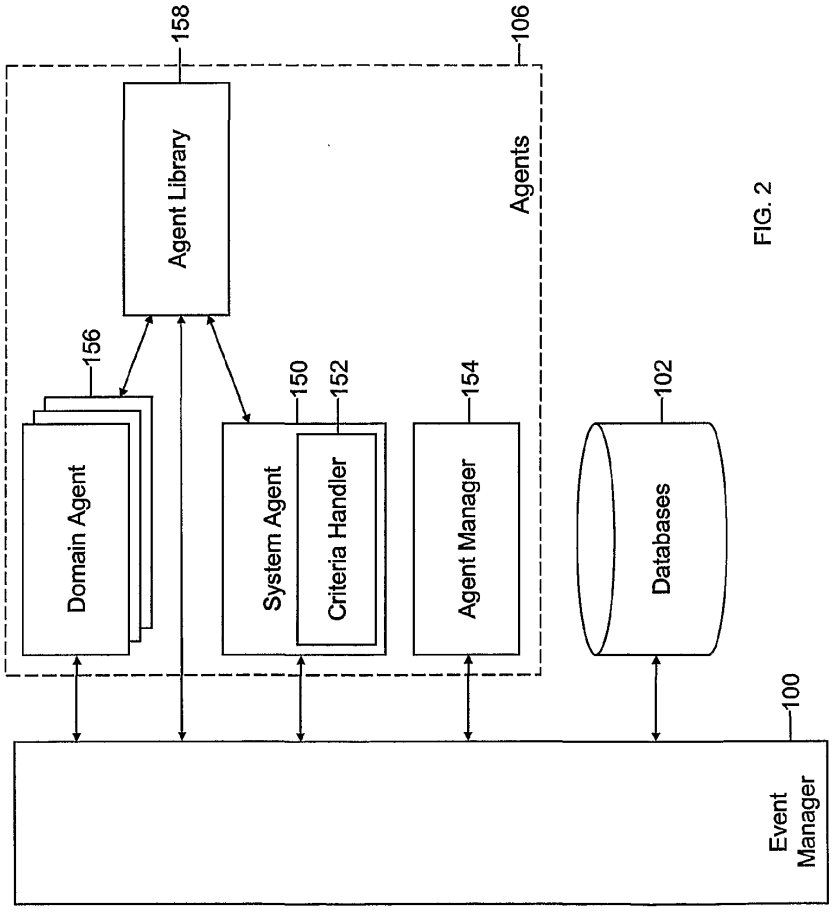


FIG. 2

FIG. 3

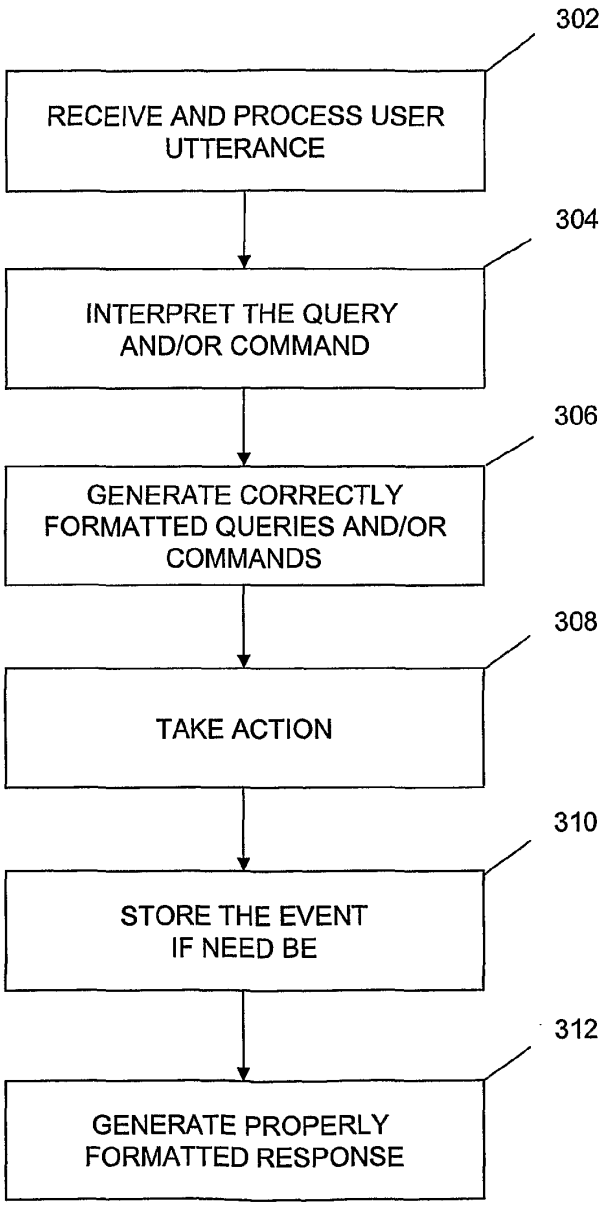

300
302
RECEIVE AND PROCESS USER UTTERANCE
304
INTERPRET THE QUERY AND/OR COMMAND
306
GENERATE CORRECTLY FORMATTED QUERIES AND/OR COMMANDS
308
TAKE ACTION
310
STORE THE EVENT IF NEED BE
312
GENERATE PROPERLY FORMATTED RESPONSE

FIG. 4A

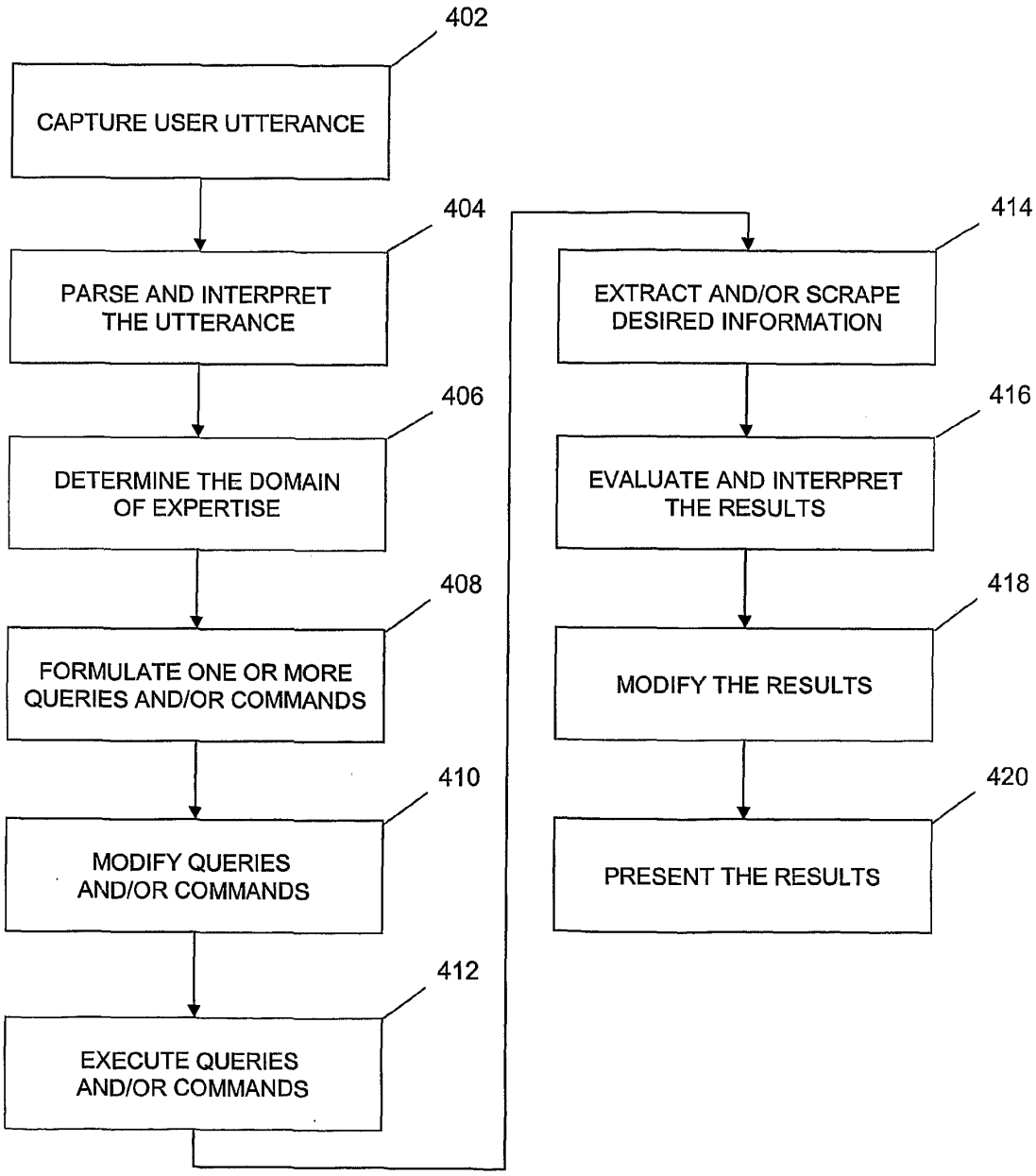

400
402
CAPTURE USER UTTERANCE
404
PARSE AND INTERPRET THE UTTERANCE
406
DETERMINE THE DOMAIN OF EXPERTISE
408
FORMULATE ONE OR MORE QUERIES AND/OR COMMANDS
410
MODIFY QUERIES AND/OR COMMANDS
412
EXECUTE QUERIES AND/OR COMMANDS
414
EXTRACT AND/OR SCRAPE DESIRED INFORMATION
416
EVALUATE AND INTERPRET THE RESULTS
418
MODIFY THE RESULTS
420
PRESENT THE RESULTS

FIG. 4B

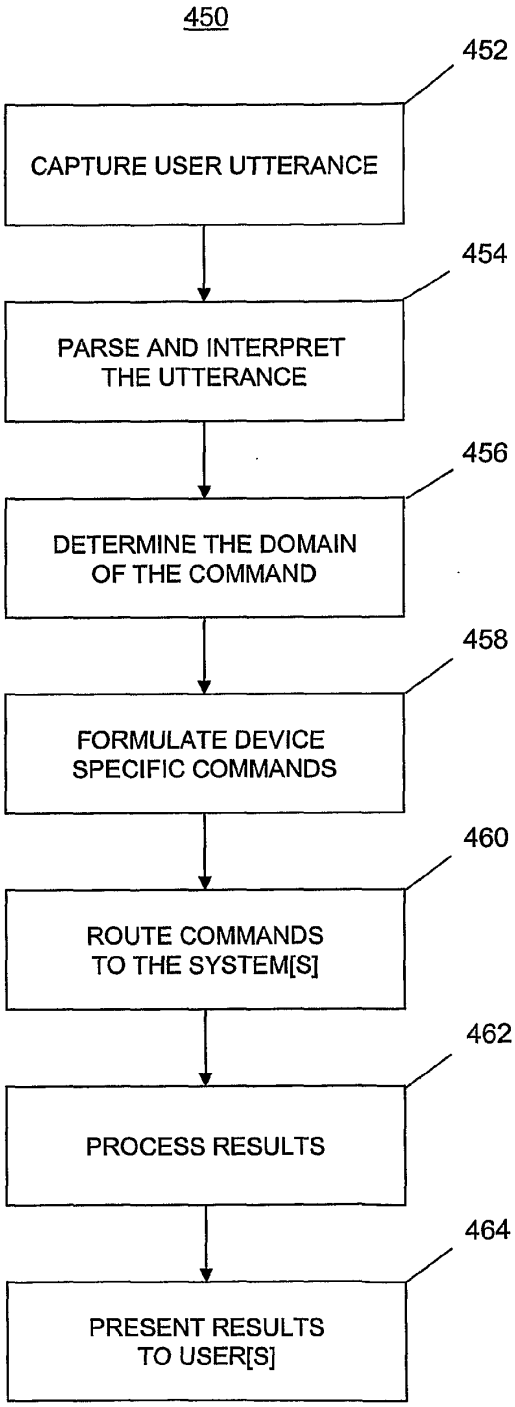

450
452
CAPTURE USER UTTERANCE
454
PARSE AND INTERPRET THE UTTERANCE
456
DETERMINE THE DOMAIN OF THE COMMAND
458
FORMULATE DEVICE SPECIFIC COMMANDS
460
ROUTE COMMANDS TO THE SYSTEM[S]
462
PROCESS RESULTS
464
PRESENT RESULTS TO USER[S]

FIG. 5

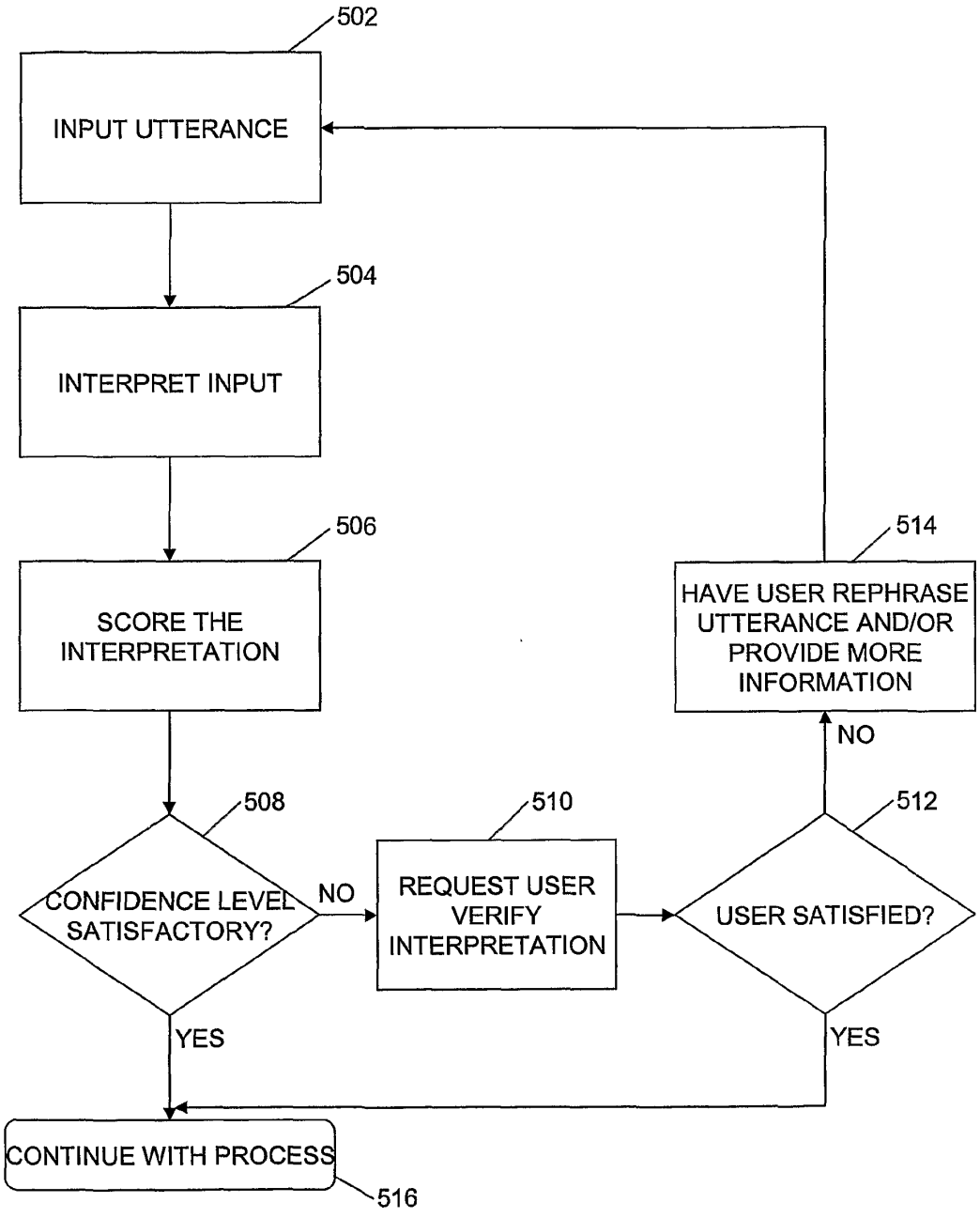

500
502
INPUT UTTERANCE
504
INTERPRET INPUT
506
SCORE THE INTERPRETATION
508
CONFIDENCE LEVEL SATISFACTORY?
NO
510
REQUEST USER VERIFY INTERPRETATION
512
USER SATISFIED?
NO
514
HAVE USER REPHRASE UTTERANCE AND/OR PROVIDE MORE INFORMATION
YES
YES
CONTINUE WITH PROCESS
516

FIG. 6

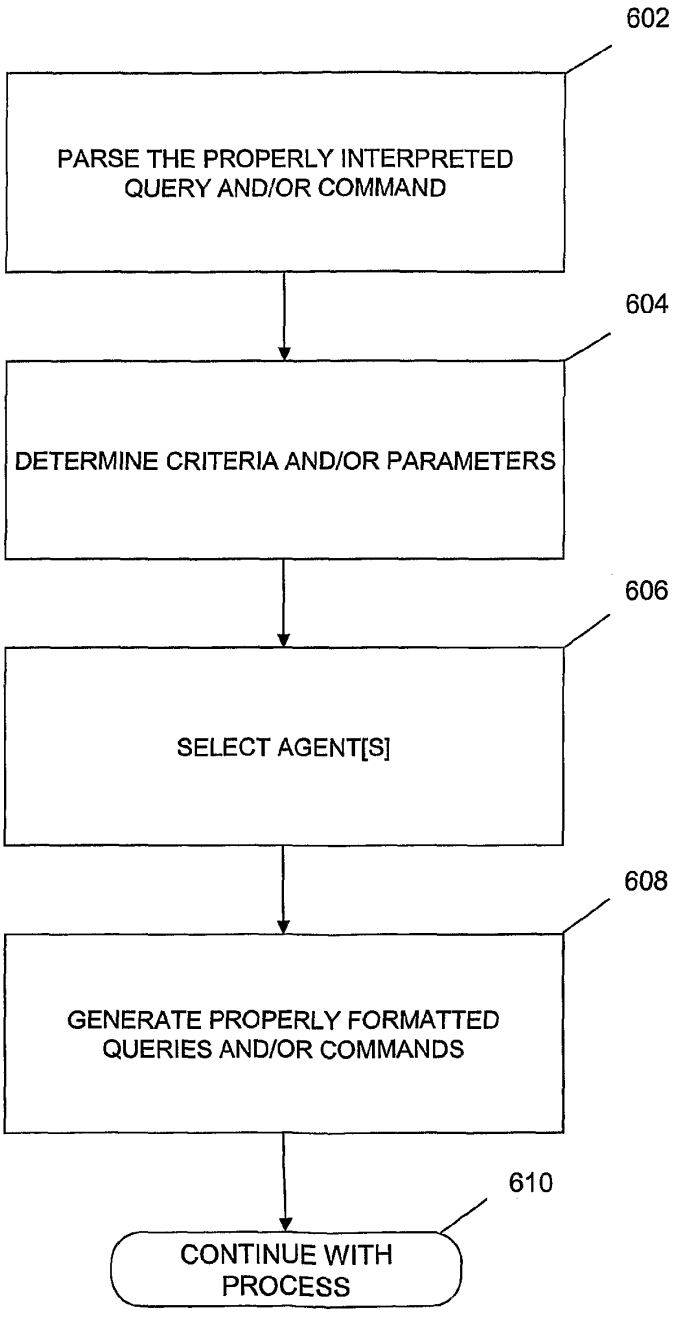
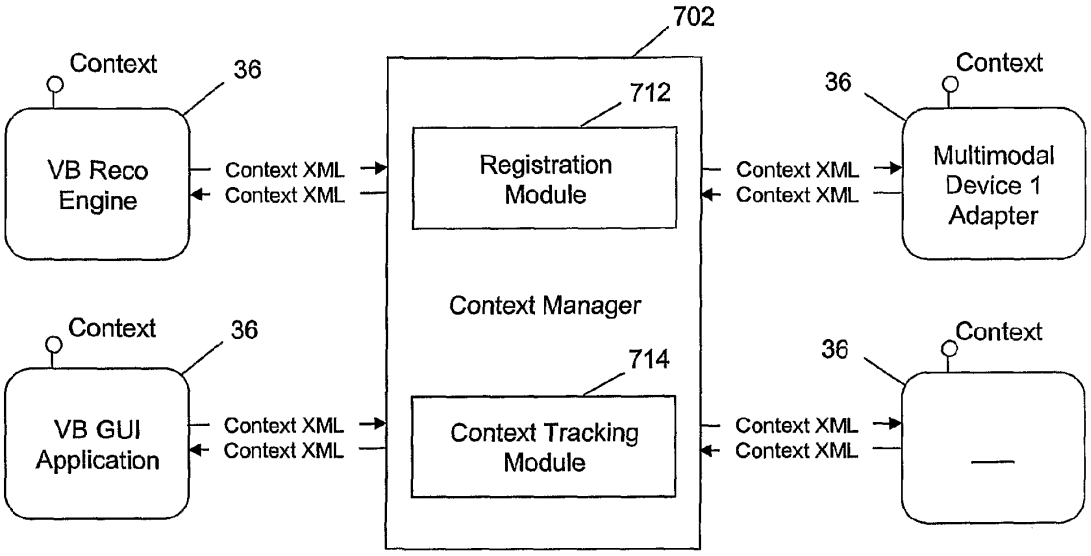

600
602
PARSE THE PROPERLY INTERPRETED QUERY AND/OR COMMAND
604
DETERMINE CRITERIA AND/OR PARAMETERS
606
SELECT AGENT[S]
608
GENERATE PROPERLY FORMATTED QUERIES AND/OR COMMANDS
610
CONTINUE WITH PROCESS

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

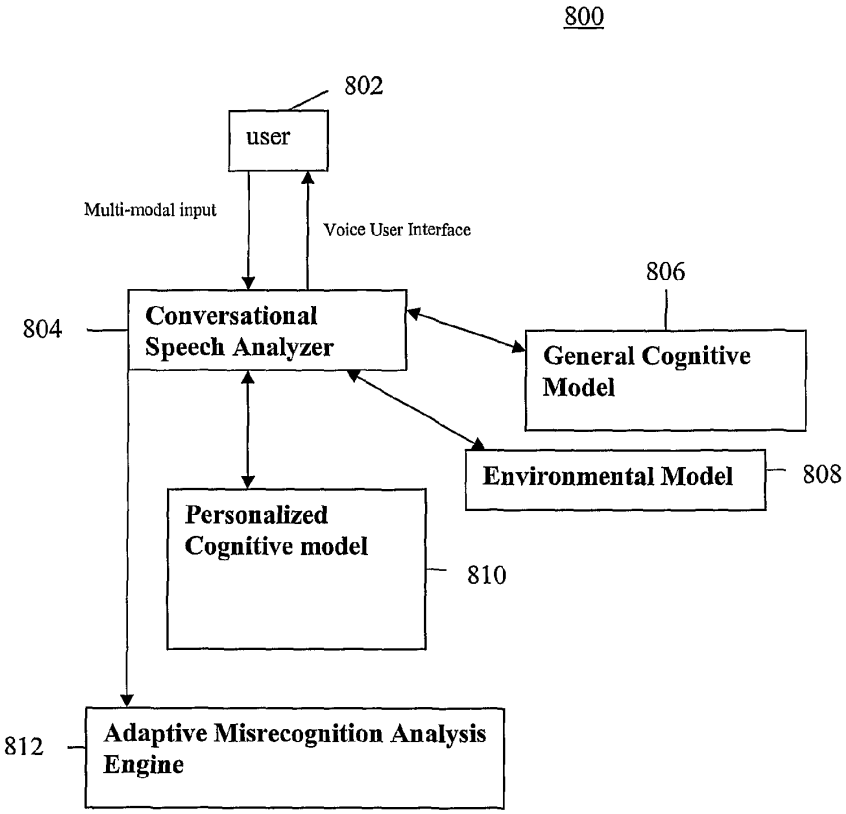

800
802
user
Multi-modal input
Voice User Interface
806
804
Conversational Speech Analyzer
General Cognitive Model
Environmental Model
808
Personalized Cognitive model
810
812
Adaptive Misrecognition Analysis Engine

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; 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).

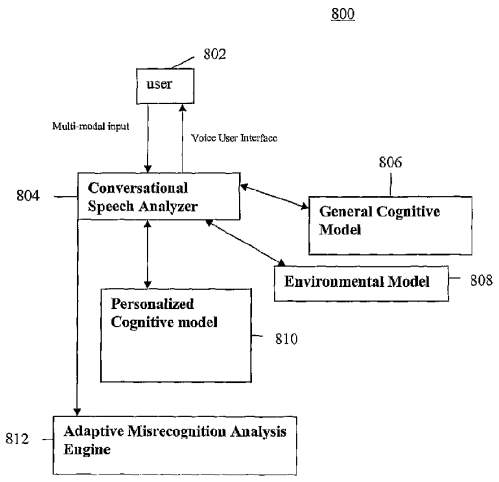
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, 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:
— *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)

Electronic Patent Application Fee Transmittal

Application Number:	
Filing Date:	
Title of Invention:	SYSTEM AND METHOD FOR PROVIDING A NATURAL LANGUAGE CONTENT DEDICATION SERVICE
First Named Inventor/Applicant Name:	Mike Kennewick
Filer:	Syed Jafar Ali/Miriam E. Nelsen
Attorney Docket Number:	090611-0384422

Filed as Small Entity

Utility under 35 USC 111(a) Filing Fees

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Utility filing Fee (Electronic filing)	4011	1	82	82
Utility Search Fee	2111	1	270	270
Utility Examination Fee	2311	1	110	110
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				
Miscellaneous:				
		Total in USD ($)		**462**

Electronic Acknowledgement Receipt	
EFS ID:	8812263
Application Number:	12943699
International Application Number:	
Confirmation Number:	4185
Title of Invention:	SYSTEM AND METHOD FOR PROVIDING A NATURAL LANGUAGE CONTENT DEDICATION SERVICE
First Named Inventor/Applicant Name:	Mike Kennewick
Customer Number:	00909
Filer:	Syed Jafar Ali/Miriam E. Nelsen
Filer Authorized By:	Syed Jafar Ali
Attorney Docket Number:	090611-0384422
Receipt Date:	10-NOV-2010
Filing Date:	
Time Stamp:	17:12:29
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	Deposit Account
Payment was successfully received in RAM	$462
RAM confirmation Number	3937
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		090611-0384422_ApplicationPapers.pdf	877863 7fca4f10f941f7540141e9ec52fd6a106856e7bc	yes	17

Multipart Description/PDF files in .zip description

Document Description	Start	End
Transmittal of New Application	1	1
Application Data Sheet	2	3
Transmittal Letter	4	5
Information Disclosure Statement (IDS) Filed (SB/08)	6	17

Warnings:

Information:

2		090611-0384422_Specification.pdf	3074418 a59b31ba6b62625c7b10f6542d11547ad93db17e	yes	59

Multipart Description/PDF files in .zip description

Document Description	Start	End
Specification	1	54
Claims	55	58
Abstract	59	59

Warnings:

Information:

3	Drawings-only black and white line drawings	090611-0384422_Drawings.pdf	105596 51fe25d07a1100d2c51321925e925ee9ebd57440	no	8

Warnings:

Information:

4	Foreign Reference	WO1999046763.pdf	1247910 62eacd3541d472fdc4873f7f386c31e69fd9ea4e	no	28

Warnings:

Information:

5	Foreign Reference	WO2000021232.pdf	3963803 203ae5a44439b38f56ae96cdbaa5a64e9e3c570c	no	77
Warnings:					
Information:					
6	Foreign Reference	WO2000046792.pdf	1259047 705348c0646506c1bc55d53c27f1ae60013291b1	no	34
Warnings:					
Information:					
7	Foreign Reference	WO2001078065.pdf	2063508 477bf8d473bf86193d1f7e268aa8aa10625dfa37	no	63
Warnings:					
Information:					
8	Foreign Reference	WO2004072954.pdf	1105346 c4acb3accfd8c87c21d4bf9b2cc1c8dfe629a50c	no	25
Warnings:					
Information:					
9	Foreign Reference	WO2007019318.pdf	3196832 ff1325b971941c7a88c2c7871f5785b0154e9307	no	55
Warnings:					
Information:					
10	Foreign Reference	WO2007021587.pdf	3525793 196b60abf639ab219e2472fadc79b2bf65c0ee3a	no	61
Warnings:					
Information:					
11	NPL Documents	REUTERS.pdf	79394 7f52f8fb293a5d88b570212d1f5399bb61aacdb3	no	1
Warnings:					
Information:					
12	NPL Documents	LIN.pdf	546736 3dce1acf1fbd1bc2a4a0da820e559e124949d760	no	4
Warnings:					
Information:					
13	NPL Documents	KUHN.pdf	438455 5e927839fd4a238b3722e336346ee055e7fcae45	no	5
Warnings:					
Information:					

14	NPL Documents	BELVIN.pdf	468676 93c339f6044e2eb4cff4b1c9087be3a99e3767d4	no	5
Warnings:					
Information:					
15	NPL Documents	LIND.pdf	729916 67eb78f7014eeb103df2382cdcecf0f9807d755a	no	6
Warnings:					
Information:					
16	NPL Documents	ZHAO.pdf	1517578 2808fd320bc985b8850a5a6c3fa515213ffa320c	no	5
Warnings:					
Information:					
17	NPL Documents	CHAI.pdf	1024450 9bf84f1bd4b08cba3cd8e8f3c6d5bb73ea5f9e5b	no	10
Warnings:					
Information:					
18	NPL Documents	CHEYER.pdf	136148 ef2636c9907ae9ab23536f71fecb9453fed36f4a	no	11
Warnings:					
Information:					
19	NPL Documents	ELIO.pdf	934606 5312283c44d9e175a26f467c8014f5de541895b2	no	10
Warnings:					
Information:					
20	NPL Documents	TURUNEN.pdf	179154 01427e429808c0e3724a46047f4d0cfc9abf99d1	no	2
Warnings:					
Information:					
21	NPL Documents	MAO.pdf	87646 07b0d85ca1604fefcd7cd99772e25ff506ff1ef0	no	4
Warnings:					
Information:					
22	NPL Documents	VANHOUCKE.pdf	259233 84ea32d24072d520d3c972cb6af5e8dc4b7835f8	no	4
Warnings:					
Information:					

23	NPL Documents	WENG.pdf	346056 0b8aed950361e7f52843a57fd30ba770384b6363	no	4
Warnings:					
Information:					
24	NPL Documents	ELMELIANI.pdf	296007 f2fb2b6713c7266a4c98b7c64837ebb75106cc03	no	4
Warnings:					
Information:					
25	NPL Documents	ARRINGTON.pdf	627466 b532ea1dde8275e00c1d288d7e7f9a17df60711b	no	4
Warnings:					
Information:					
26	NPL Documents	BAZZI.pdf	361077 67a2e53341c20cd2f0fecfdfc9ca971d47945b8c	no	4
Warnings:					
Information:					
27	NPL Documents	OSHAUGHNESSY.pdf	4563513 93d11484104e08a3f70ac1884e29bccb188a88f0	no	34
Warnings:					
Information:					
28	Fee Worksheet (PTO-875)	fee-info.pdf	33341 15b2f3e6d992ef5dda7491b0131527 5f4211535a	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	33049568		

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.

Electronic Acknowledgement Receipt	
EFS ID:	8812263
Application Number:	12943699
International Application Number:	
Confirmation Number:	4185
Title of Invention:	SYSTEM AND METHOD FOR PROVIDING A NATURAL LANGUAGE CONTENT DEDICATION SERVICE
First Named Inventor/Applicant Name:	Mike Kennewick
Customer Number:	00909
Filer:	Syed Jafar Ali/Miriam E. Nelsen
Filer Authorized By:	Syed Jafar Ali
Attorney Docket Number:	090611-0384422
Receipt Date:	10-NOV-2010
Filing Date:	
Time Stamp:	17:12:29
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	Deposit Account
Payment was successfully received in RAM	$462
RAM confirmation Number	3937
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		090611-0384422_ApplicationPapers.pdf	877863 7fca4f10f941f7540141e9ec52fd6a106856e7bc	yes	17

Multipart Description/PDF files in .zip description

Document Description	Start	End
Transmittal of New Application	1	1
Application Data Sheet	2	3
Transmittal Letter	4	5
Information Disclosure Statement (IDS) Filed (SB/08)	6	17

Warnings:

Information:

2		090611-0384422_Specification.pdf	3074418 a59b31ba6b62625c7b10f6542d11547ad93db17e	yes	59

Multipart Description/PDF files in .zip description

Document Description	Start	End
Specification	1	54
Claims	55	58
Abstract	59	59

Warnings:

Information:

3	Drawings-only black and white line drawings	090611-0384422_Drawings.pdf	105596 51fe25d07a1100d2c51321925e925ee9ebd57440	no	8

Warnings:

Information:

4	Foreign Reference	WO1999046763.pdf	1247910 62eacd3541d472fdc4873f7f386c31e69fd9ea4e	no	28

Warnings:

Information:

5	Foreign Reference	WO2000021232.pdf	3963803 203ae5a44439b38f56ae96cdbaa5a64e9e3c570c	no	77
Warnings:					
Information:					
6	Foreign Reference	WO2000046792.pdf	1259047 705348c0646506c1bc55d53c27f1ae60013291b1	no	34
Warnings:					
Information:					
7	Foreign Reference	WO2001078065.pdf	2063508 477bf8d473bf86193d1f7e268aa8aa10625dfa37	no	63
Warnings:					
Information:					
8	Foreign Reference	WO2004072954.pdf	1105346 c4acb3accfd8c87c21d4bf9b2cc1c8dfe629a50c	no	25
Warnings:					
Information:					
9	Foreign Reference	WO2007019318.pdf	3196832 ff1325b971941c7a88c2c7871f5785b0154e9307	no	55
Warnings:					
Information:					
10	Foreign Reference	WO2007021587.pdf	3525793 196b60abf639ab219e2472fadc79b2bf65c0ee3a	no	61
Warnings:					
Information:					
11	NPL Documents	REUTERS.pdf	79394 7f52f8fb293a5d88b570212d1f5399bb61aacdb3	no	1
Warnings:					
Information:					
12	NPL Documents	LIN.pdf	546736 3dce1acf1fbd1bc2a4a0da820e559e124949d760	no	4
Warnings:					
Information:					
13	NPL Documents	KUHN.pdf	438455 5e927839fd4a238b3722e336346ee055e7fcae45	no	5
Warnings:					
Information:					

14	NPL Documents	BELVIN.pdf	468676 93c339f6044e2eb4cff4b1c9087be3a99e3767d4	no	5
Warnings:					
Information:					
15	NPL Documents	LIND.pdf	729916 67eb78f7014eeb103df2382cdcecf0f9807d755a	no	6
Warnings:					
Information:					
16	NPL Documents	ZHAO.pdf	1517578 2808fd320bc985b8850a5a6c3fa515213ffa320c	no	5
Warnings:					
Information:					
17	NPL Documents	CHAI.pdf	1024450 9bf84f1bd4b08cba3cd8e8f3c6d5bb73ea5f9e5b	no	10
Warnings:					
Information:					
18	NPL Documents	CHEYER.pdf	136148 ef2636c9907ae9ab23536f71fecb9453fed36f4a	no	11
Warnings:					
Information:					
19	NPL Documents	ELIO.pdf	934606 5312283c44d9e175a26f467c8014f5de541895b2	no	10
Warnings:					
Information:					
20	NPL Documents	TURUNEN.pdf	179154 01427e429808c0e3724a46047f4d0cfc9abf99d1	no	2
Warnings:					
Information:					
21	NPL Documents	MAO.pdf	87646 07b0d85ca1604fefcd7cd99772e25ff506ff1ef0	no	4
Warnings:					
Information:					
22	NPL Documents	VANHOUCKE.pdf	259233 84ea32d24072d520d3c972cb6af5e8dc4b7835f8	no	4
Warnings:					
Information:					

23	NPL Documents	WENG.pdf	346056 0b8aed950361e7f52843a57fd30ba770384b6363	no	4
Warnings:					
Information:					
24	NPL Documents	ELMELIANI.pdf	296007 f2fb2b6713c7266a4c98b7c64837ebb75106cc03	no	4
Warnings:					
Information:					
25	NPL Documents	ARRINGTON.pdf	627466 b532ea1dde8275e00c1d288d7e7f9a17df60711b	no	4
Warnings:					
Information:					
26	NPL Documents	BAZZI.pdf	361077 67a2e53341c20cd2f0fecfdfc9ca971d47945b8c	no	4
Warnings:					
Information:					
27	NPL Documents	OSHAUGHNESSY.pdf	4563513 93d11484104e08a3f70ac1884e29bccb188a88f0	no	34
Warnings:					
Information:					
28	Fee Worksheet (PTO-875)	fee-info.pdf	33341 15b2f3e6d992ef5dda7491b0131527 5f4211535a	no	2
Warnings:					
Information:					
			Total Files Size (in bytes):	33049568	

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.