US 20100023320A1

(19) United States

(12) Patent Application Publication

Di Cristo et al.

(10) Pub. No.: US 2010/0023320 A1

(43) Pub. Date: Jan. 28, 2010

(54) SYSTEM AND METHOD OF SUPPORTING ADAPTIVE MISRECOGNITION IN CONVERSATIONAL SPEECH

(75) Inventors: **Philippe Di Cristo**, Bellevue, WA (US); **Chris Weider**, Everett, WA (US); **Robert A. Kennewick**, Seattle, WA (US)

Correspondence Address:
PILLSBURY WINTHROP SHAW PITTMAN, LLP
P.O. BOX 10500
MCLEAN, VA 22102 (US)

(73) Assignee: **VoiceBox Technologies, Inc.**, Kirkland, WA (US)

(21) Appl. No.: **12/571,795**

(22) Filed: **Oct. 1, 2009**

Related U.S. Application Data

(63) Continuation of application No. 11/200,164, filed on Aug. 10, 2005, now Pat. No. 7,620,549.

Publication Classification

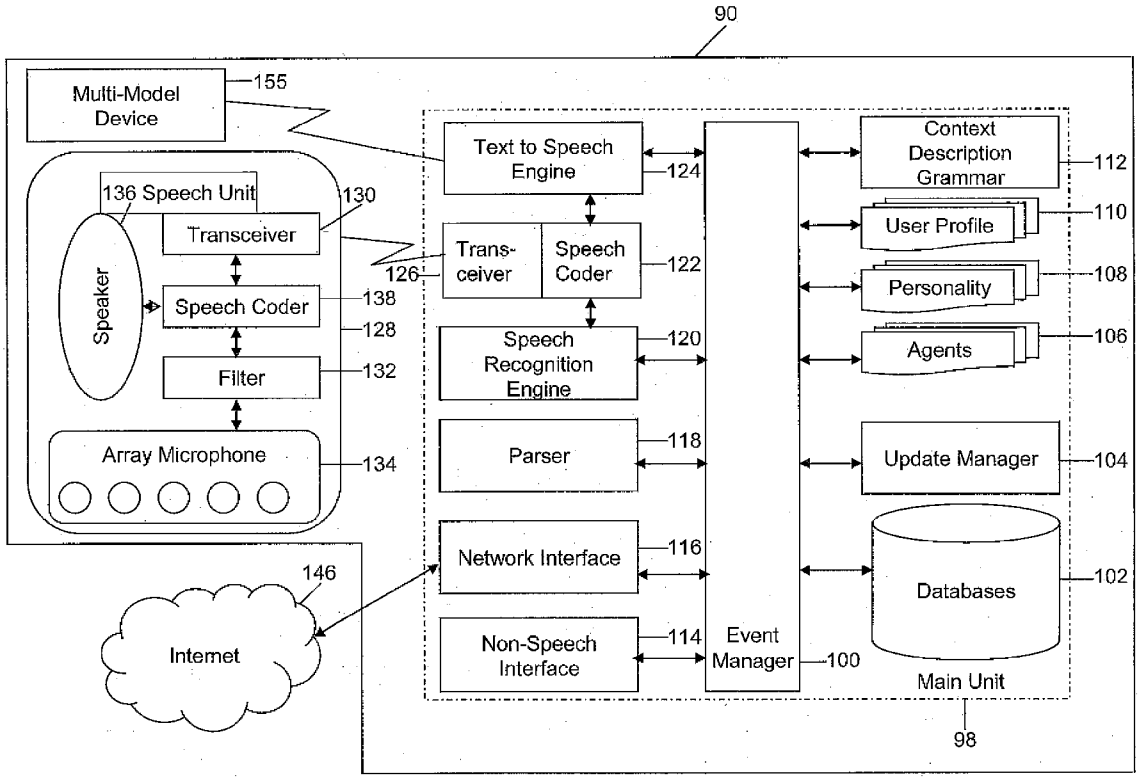
(51) **Int. Cl.**

G06F 17/27	(2006.01)
G10L 15/18	(2006.01)
G10L 15/00	(2006.01)

(52) **U.S. Cl.** **704/9**; 704/257; 704/240; 704/E15.001; 704/E15.018

(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.



MICROSOFT CORP.
EXHIBIT 1007

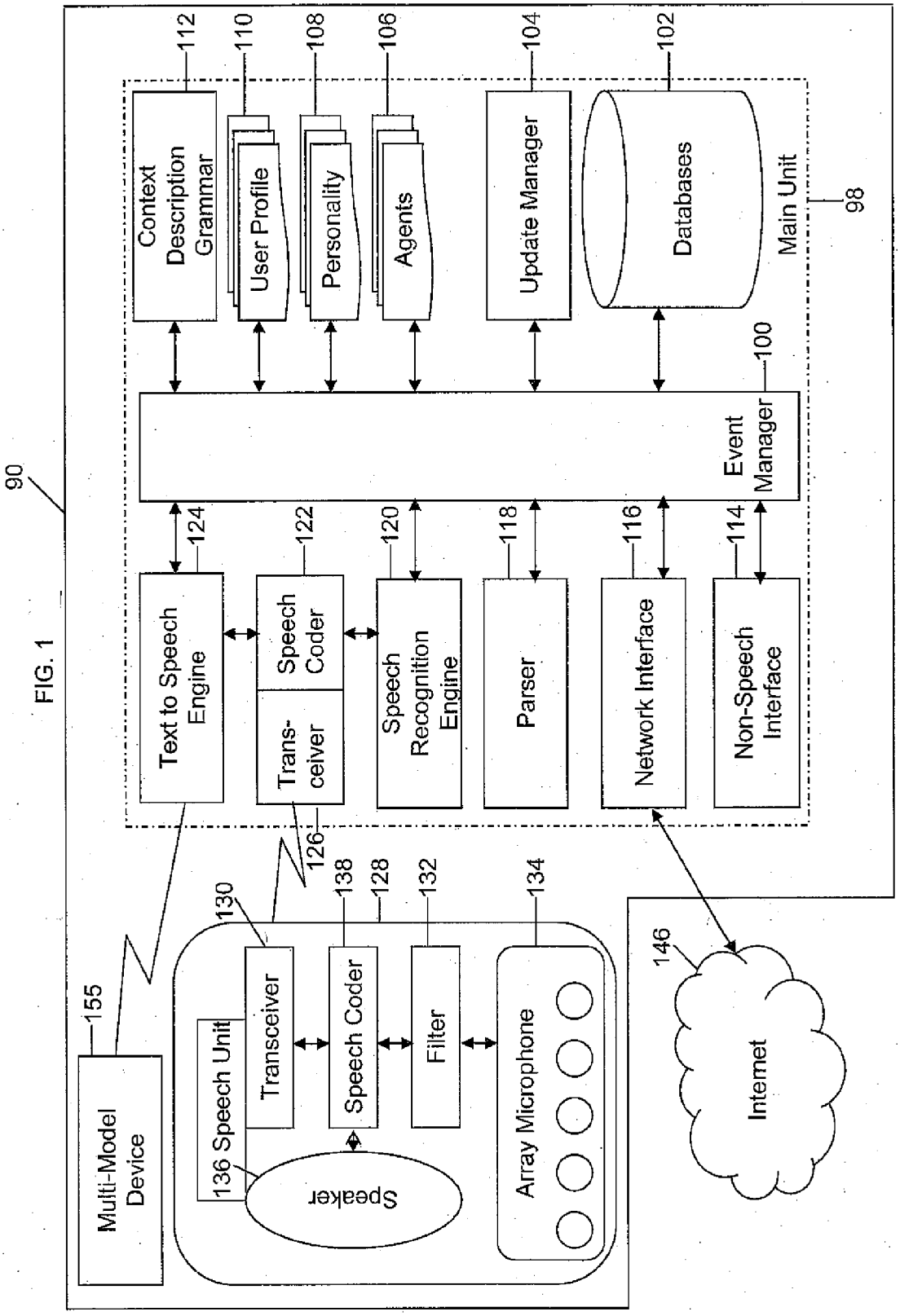

90
Context Description Grammar
User Profile
Personality
Agents
Update Manager
Databases
Main Unit
98
112
110
108
106
104
102
Event Manager
100
Text to Speech Engine
Speech Coder
Trans-ceiver
Speech Recognition Engine
Parser
Network Interface
Non-Speech Interface
124
122
120
118
116
114
126
130
138
128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

FIG. 1

FIG. 2

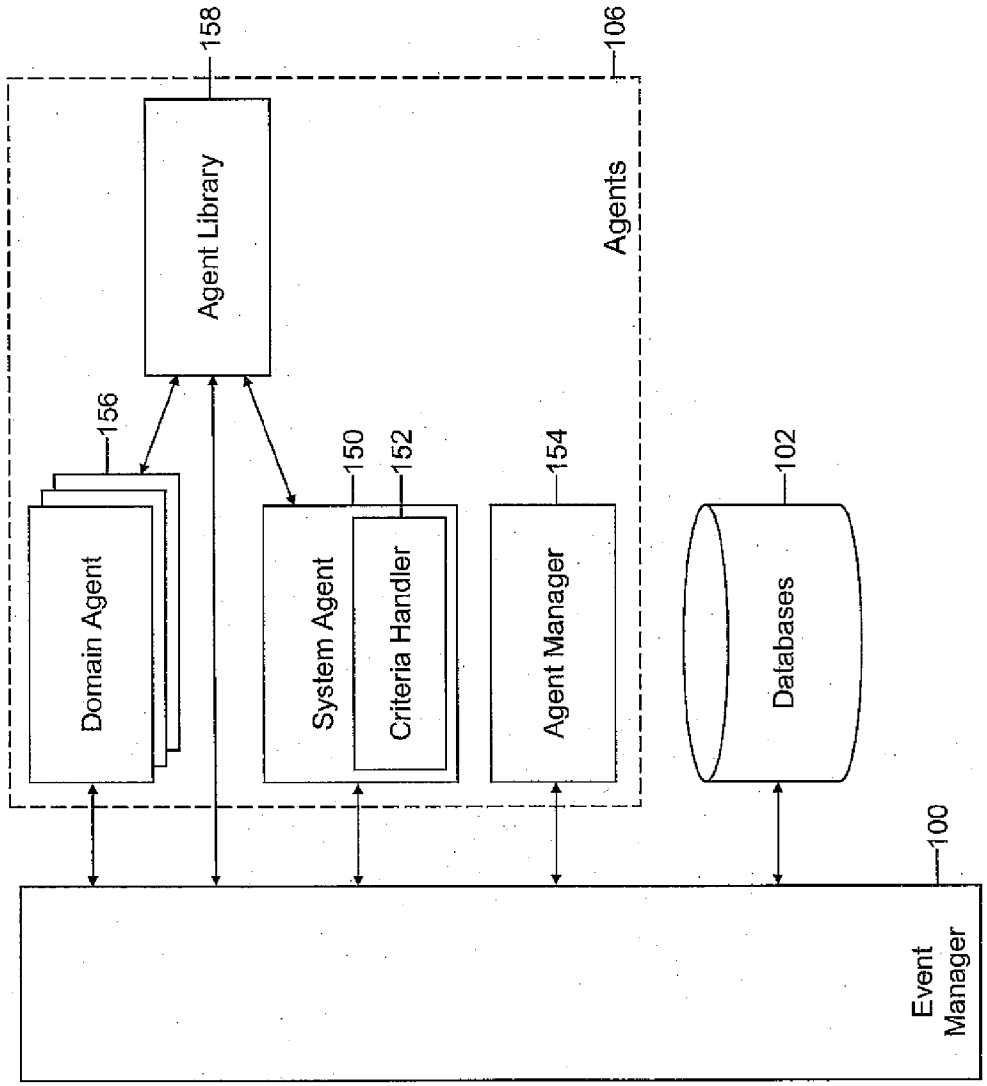

158
106
Agent Library
Agents
156
150
152
154
102
Domain Agent
System Agent
Criteria Handler
Agent Manager
Databases
100
Event Manager

FIG. 3

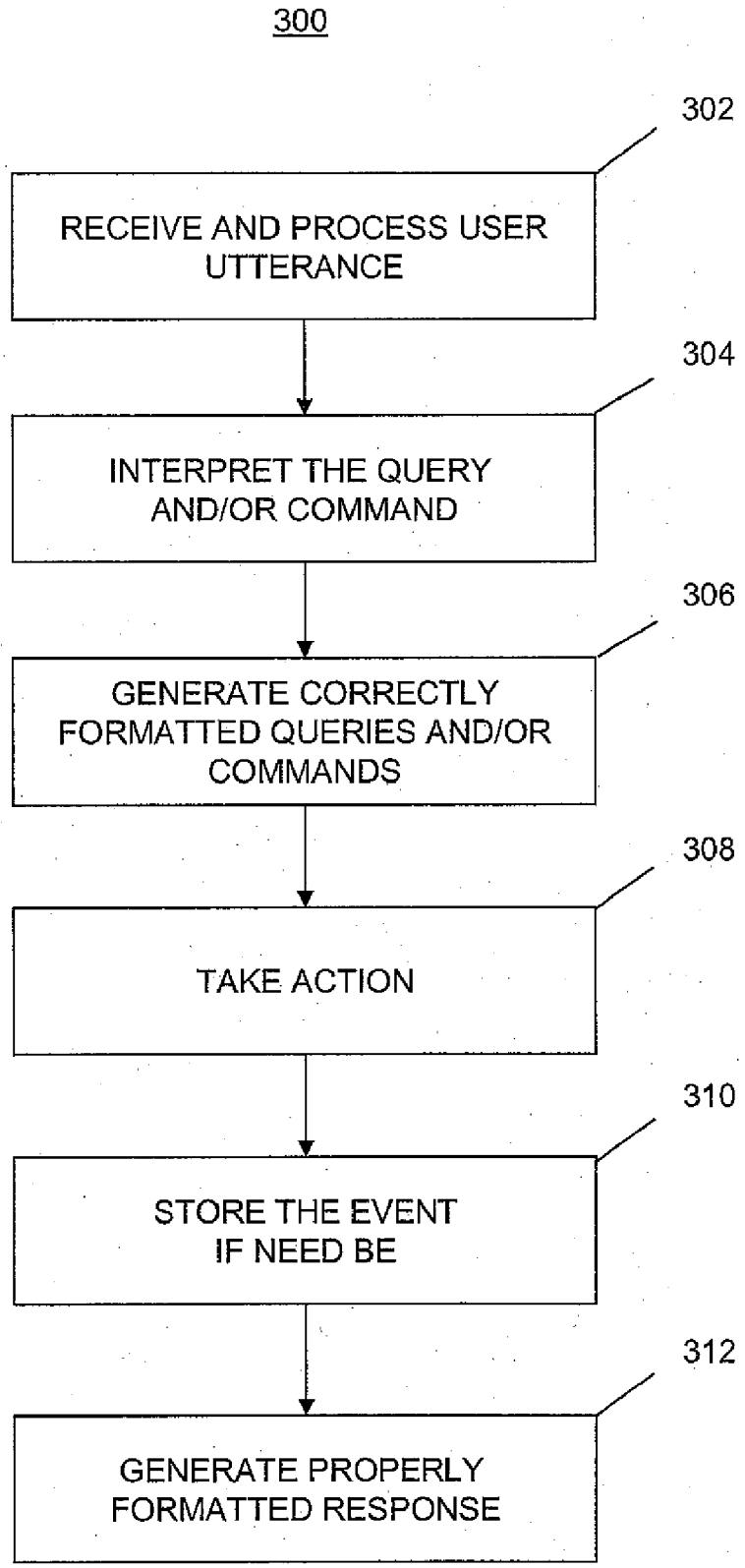

300
302
RECEIVE AND PROCESS USER UTTERANCE
304
INTERPRET THE QUERY AND/OR COMMAND
306
GENERATE CORRECTLY FORMATTED QUERIES AND/OR COMMANDS
308
TAKE ACTION
310
STORE THE EVENT IF NEED BE
312
GENERATE PROPERLY FORMATTED RESPONSE

FIG. 4A

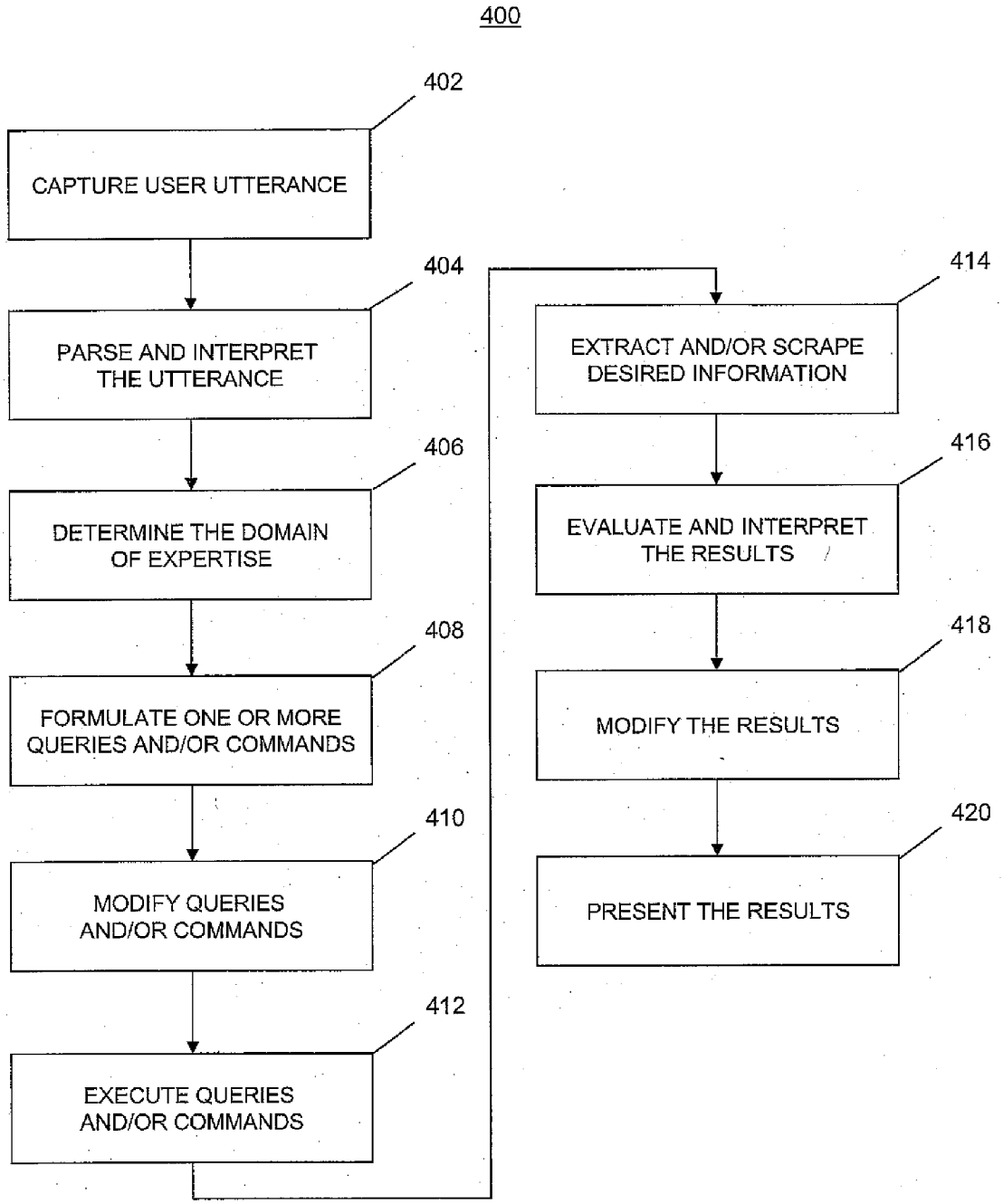

400
402
CAPTURE USER UTTERANCE
404
PARSE AND INTERPRET THE UTTERANCE
406
DETERMINE THE DOMAIN OF EXPERTISE
408
FORMULATE ONE OR MORE QUERIES AND/OR COMMANDS
410
MODIFY QUERIES AND/OR COMMANDS
412
EXECUTE QUERIES AND/OR COMMANDS
414
EXTRACT AND/OR SCRAPE DESIRED INFORMATION
416
EVALUATE AND INTERPRET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

FIG. 4B

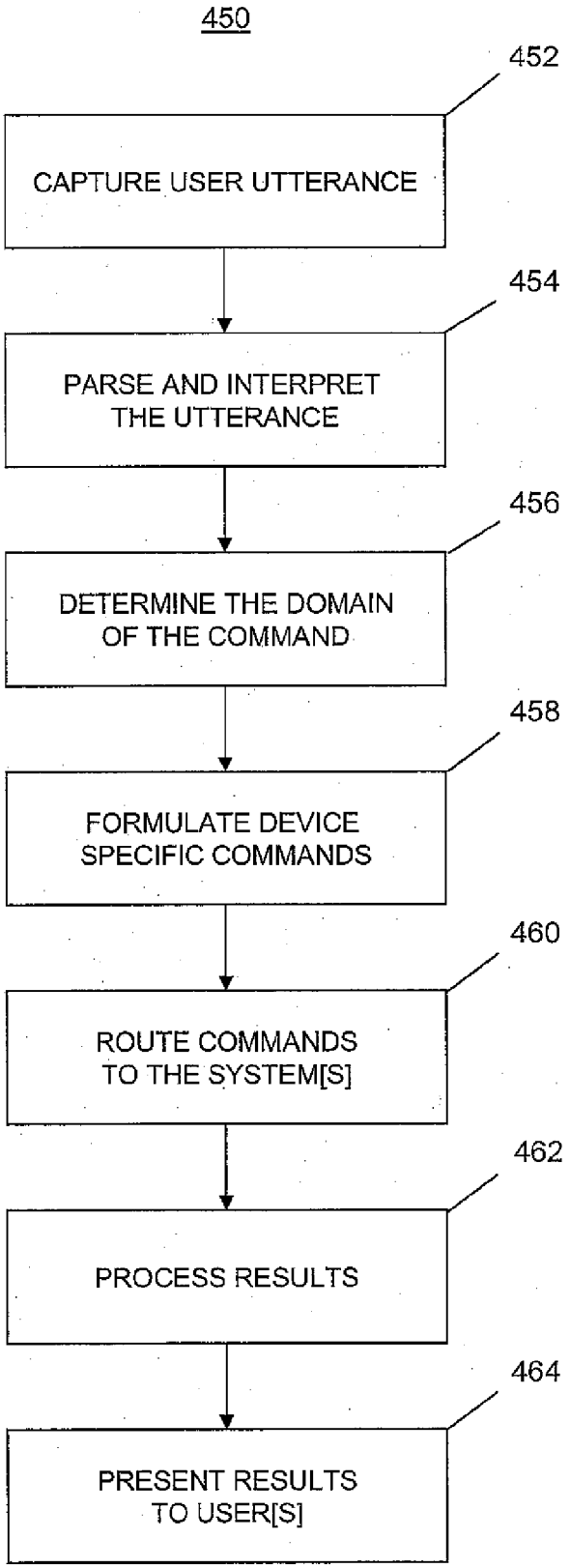

450
452
CAPTURE USER UTTERANCE
454
PARSE AND INTERPRET THE UTTERANCE
456
DETERMINE THE DOMAIN OF THE COMMAND
458
FORMULATE DEVICE SPECIFIC COMMANDS
460
ROUTE COMMANDS TO THE SYSTEM[S]
462
PROCESS RESULTS
464
PRESENT RESULTS TO USER[S]

FIG. 5

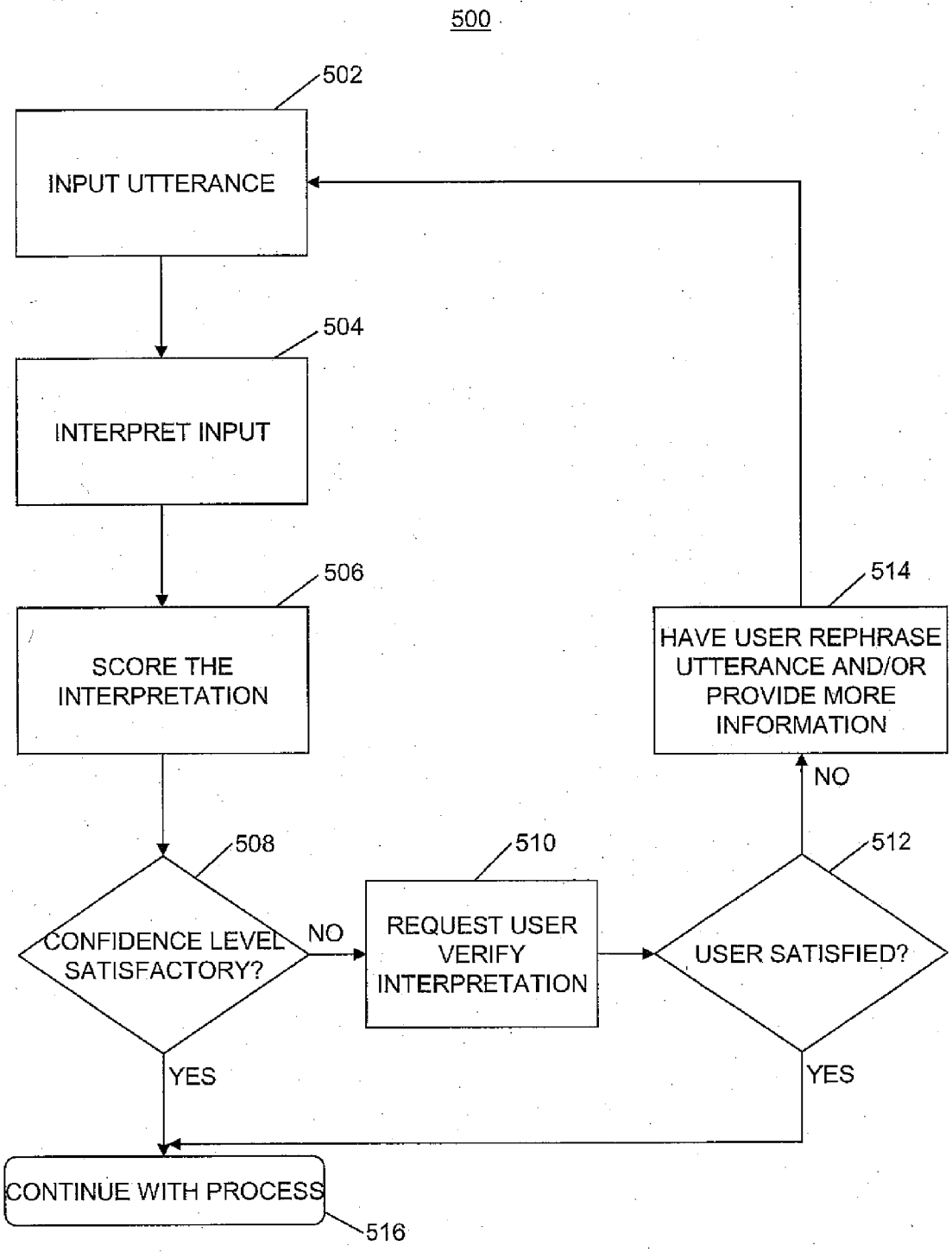

500
502
INPUT UTTERANCE
504
INTERPRET INPUT
506
SCORE THE INTERPRETATION
508
CONFIDENCE LEVEL SATISFACTORY?
NO
YES
510
REQUEST USER VERIFY INTERPRETATION
512
USER SATISFIED?
NO
YES
514
HAVE USER REPHRASE UTTERANCE AND/OR PROVIDE MORE INFORMATION
CONTINUE WITH PROCESS
516

FIG. 6

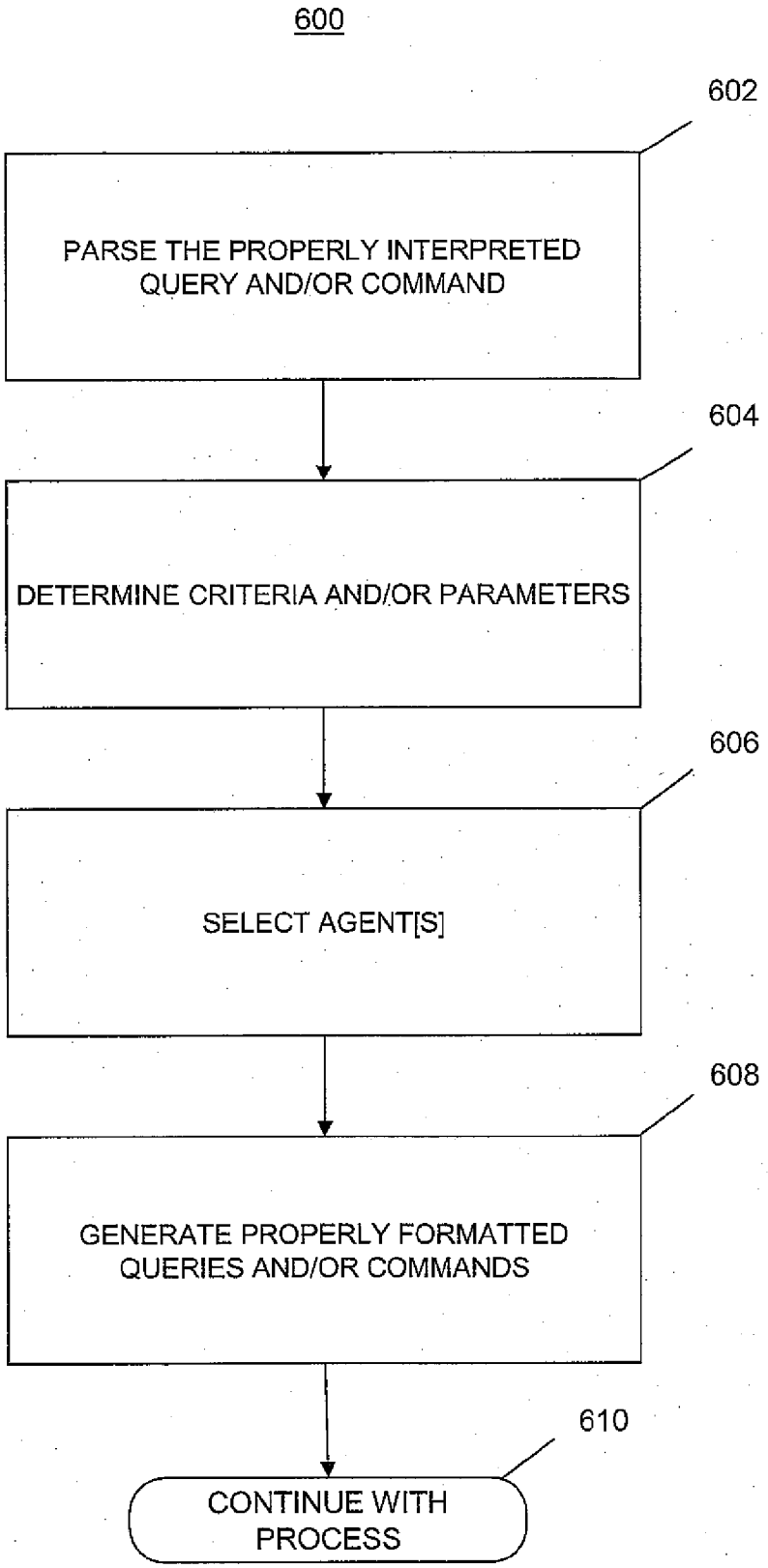
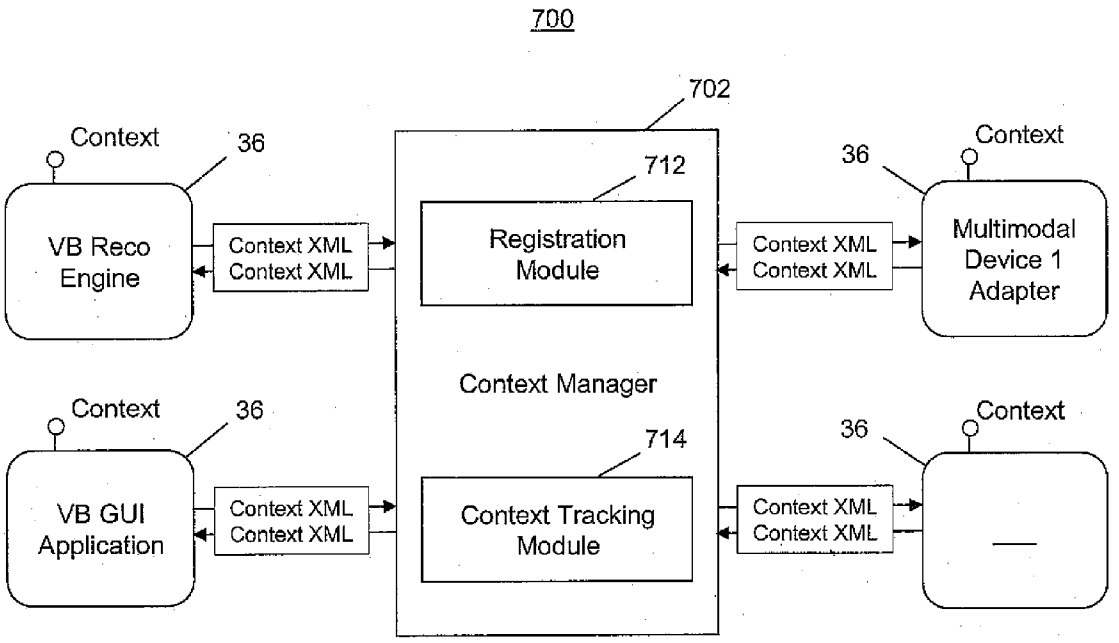

600
602
PARSE THE PROPERLY INTERPRETED QUERY AND/OR COMMAND
604
DETERMINE CRITERIA AND/OR PARAMETERS
606
SELECT AGENT[S]
608
GENERATE PROPERLY FORMATTED QUERIES AND/OR COMMANDS
610
CONTINUE WITH PROCESS

FIG. 7


700
702
Context
36
VB Reco Engine
Context XML
Context XML
712
Registration Module
Context Manager
714
Context Tracking Module
Context
36
VB GUI Application
Context XML
Context XML
Context XML
Context XML
36
Context
Multimodal Device 1 Adapter
Context XML
Context XML
36
Context

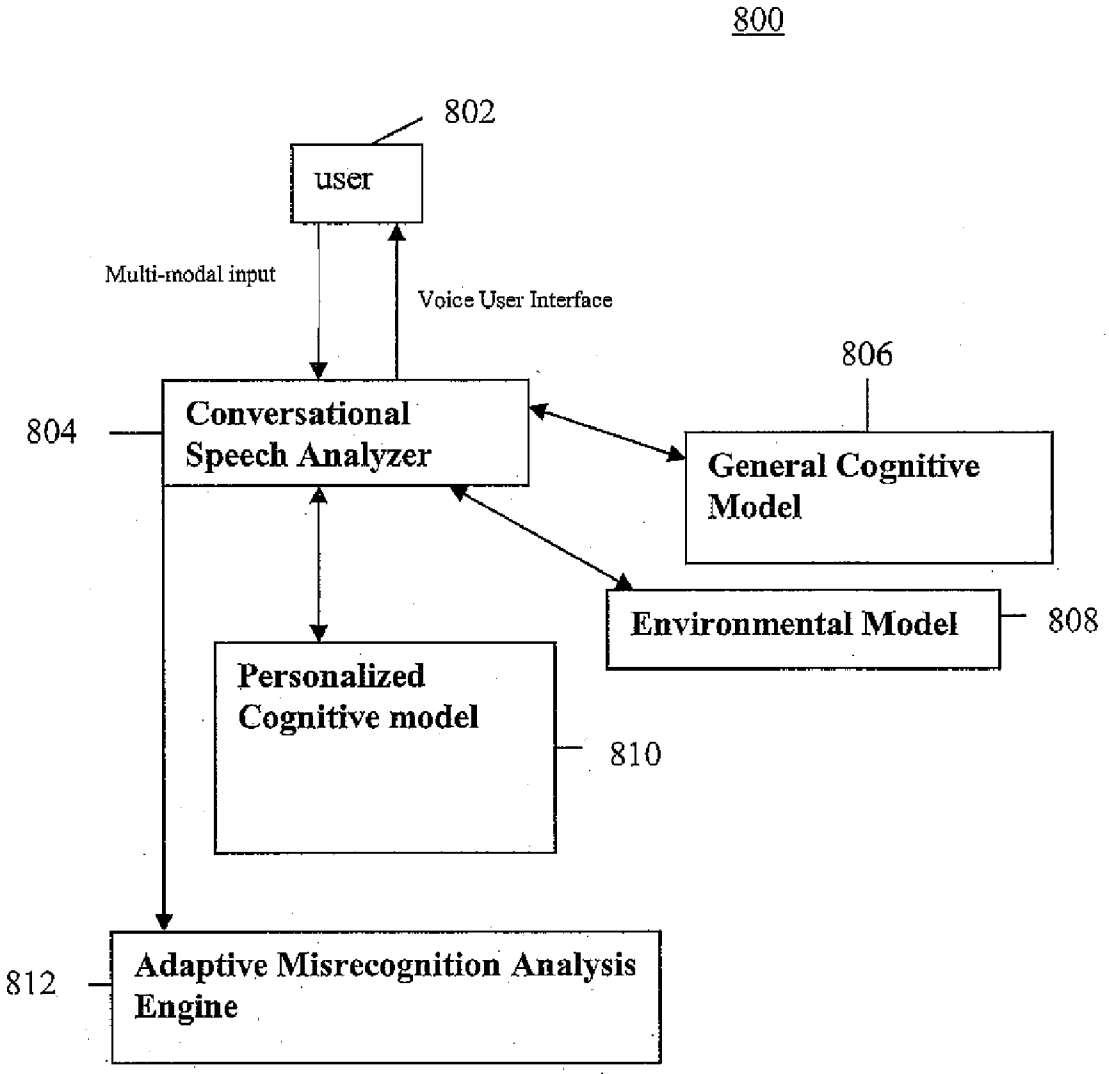

800
802
user
Multi-modal input
Voice User Interface
806
804
Conversational Speech Analyzer
General Cognitive Model
Environmental Model
808
Personalized Cognitive model
810
812
Adaptive Misrecognition Analysis Engine

FIGURE 8

SYSTEM AND METHOD OF SUPPORTING ADAPTIVE MISRECOGNITION IN CONVERSATIONAL SPEECH

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application is a continuation of co-pending U.S. patent application Ser. No. 11/200,164, entitled "System and Method of Supporting Adaptive Misrecognition in Conversational Speech," filed Aug. 10, 2005, the contents of which are hereby incorporated by reference in their entirety.

BACKGROUND OF THE INVENTION

[0002] 1. Field of the Invention

[0003] 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.

[0004] 2. Background of the Related Art

[0005] 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.

[0006] 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

[0007] According to an aspect of the invention, one object of the invention is to overcome these and other drawbacks of prior systems.

[0008] According to one aspect of the invention, users may provide the system with spoken utterances and/or other communications in a natural language format, including imperfect information such as, incomplete thoughts, incomplete sentences, incomplete phrases, slang terminology, repeated words, word variations, synonyms, or other imperfect information. In order for machines (e.g., computer devices) to properly respond to questions and/or commands that are presented in the natural language format, the questions and/or commands may be parsed and interpreted to formulate machine processable queries and algorithms.

[0009] 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 fill 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.

[0010] 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 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.

[0011] According to one embodiment of the invention, multi-modal input may be received from users at a conversational speech analyzer that is coupled to several components, including a general cognitive model, an environmental model, a personalized cognitive model, and an adaptive misrecognition analysis engine, or other components. According to one embodiment of the invention, the conversational speech analyzer may include one or more speech recognition engines that transcribe utterances to text.

[0012] 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.

[0013] 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.

[0014] 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.

[0015] According to one embodiment of the invention, infrastructure may be provided to maintain context information during multimodal interactions, such as speech and/or non-speech interactions. According to one exemplary embodiment of the invention, context information may be maintained in a multimodal environment by providing communication channels between multimodal devices, or mobile devices, and the system. The communication channels allow the system to receive multimodal input such as text-based commands and questions and/or voice-based commands and questions. According to another embodiment of the invention, the multimodal input may include a string of text, such as keywords, that are received as commands or questions. According to yet another embodiment of the invention, the system may synchronize the context between the multimodal devices and the speech-based units. In order to send a response to the corresponding device, the system may track the source and send the response to the corresponding speech interface or the non-speech interface.

[0016] 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.

[0017] 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.

[0018] According to one embodiment of the invention, the text message may be searched for a particular character, group of characters, word, group of words, and other text combinations. The text combination may be compared against entries in a context description grammar that is associated with each agent. If a match is identified between an active grammar in the context description grammar and the command and/or request, then the match may be scored. The agents may be ranked based on the determined score. In generating an aggregate response from the one or more responses received from the agents, the ordering of the responses from the individual agents may be determined based on the rank of agents. The aggregate response may be generated by a response generator module. Agents may update a context stack, that includes an ordered list of command contexts, to enable follow-up requests.

[0019] 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.

[0020] 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.

[0021] According to another aspect of the invention domain specific behavior and information may be organized into data managers. Data managers are autonomous executables that receive, process and respond to user questions, queries and commands. The data managers provide complete, convenient

and re-distributable packages or modules of functionality, typically for a specific domain of application. Data managers may be complete packages of executable code, scripts, links to information, and other forms of communication data that provide a specific package of functionality, usually in a specific domain. In other words, data managers include components for extending the functionality to a new domain. Further, data managers and their associated data may be updated remotely over a network as new behavior is added or new information becomes available. Data managers may use system resources and the services of other, typically more specialized, data managers. Data managers may be distributed and redistributed in a number of ways including on removable storage media, transfer over networks or attached to emails and other messages. An update manager may be used to add new data managers to the system or update existing data managers.

[0022] 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.

[0023] 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.

[0024] 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.

[0025] 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 tine for the users favorite TV program. A command must be executed to set a video recorder to capture this program.

[0026] For utterances including questions and/or requests, the system may perform multiple steps that may include one or more of:

- **[0027]** capturing the user's questions and/or requests through speech recognition components that operate in a variety of real-world environments;
- **[0028]** parsing and interpreting the question and/or request;
- **[0029]** determining the domain of expertise and context, invoking the proper resources, including agents;
- **[0030]** 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;
- **[0031]** performing presentation formatting, variable substitutions and transformations to modify the requests to a form that yields desired results from the available sources;
- **[0032]** a executing the multiple requests or commands in an asynchronous manner and dealing gracefully with failures;
- **[0033]** a 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;
- **[0034]** 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;
- **[0035]** performing formatting, variable substitutions and transformations to modify the results to a form most easily understood by the user; and
- **[0036]** presenting the compound result, through a text to speech engine or a multimodal interface, to the user in a useful and expected manner.

[0037] 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.

[0038] 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.

[0039] 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 identity 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.

[0040] 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.

[0041] According to one aspect of the invention, the system may attempt to provide rapid responses in order to provide a natural question and response environment. The rapid responses may be provided without obtaining additional information. The system may determine agent composition, context and/or domain for a user's question or command, for example, by using a real-time scoring system or other technique. Based on this determination, the system may trigger one or more agents to respond to the user's question or command. The agents may make one or more requests and rapidly return a formatted response. Thus, users may receive direct responses to a set of questions, each with a different response or context. In some cases, the available information, including the request results, may not adequately answer the questions presented. In such situations, the user may be asked one or more follow-up questions to resolve the ambiguity. Additional requests may then be made-before an adequate response is provided. In these cases, the system may use context information, user profile information and/or domain specific information to minimize the interaction with the user required to deliver a response.

[0042] 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.

[0043] 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.

[0044] 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.

[0045] 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.

[0046] 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 EP conferencing), operating integrated messaging environments and other control operations.

[0047] 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:

[0048] capturing the user's command through speech recognition that operates in a variety of real-world environments;

[0049] parsing and interpreting the command;

[0050] determining the domain for the command and context, invoking the proper resources, including agents;

[0051] formulating device specific commands for the system or external devices;

[0052] routing commands to the system, or external devices, including external devices connected to data networks;

[0053] receiving and processing results of commands, including errors; and

[0054] optionally, providing responses to users indicating the success or failure of the command, and possibly including state information.

[0055] 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.

[0056] 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.

[0057] 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 voice-print matching, password or pass-phrase matching, or other security safeguards.

[0058] When multiple users are engaged in interleaved sessions, the system may gracefully resolve conflicts using a probabilistic or fizzy 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.

[0059] 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.

[0060] 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.

[0061] 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.

[0062] Some example applications for the invention includes but are not limited to the following:

[0063] 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;

[0064] automatic telephone dialing, reading and sending emails and pages by voice and other communications control functions;

[0065] map, location and direction applications;

[0066] movie or other entertainment locator, review information and ticket purchasing;

[0067] television, radio or other home entertainment schedule, review information and device control from a local or remote user;

[0068] weather information for the local area or other locations;

[0069] 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.;

[0070] flight or other scheduled transportation information and ticketing;

[0071] reservations for hotels, rental cars and other travel services;

[0072] 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;

[0073] sports scores, news stories, schedules, alerts, statistics, back ground and history information, etc.;

[0074] 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;

[0075] rights management for information or content used or published;

[0076] horoscopes, daily jokes and comics, crossword puzzle retrieval and display and related entertainment or diversions;

[0077] recipes, meal planning, nutrition information and planning, shopping lists and other home organization related activities;

[0078] as an interface to auctions and online shopping, and where the system can manage payment or an electronic wallet;

[0079] 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;

[0080] 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;

[0081] organization and calendar management for families, businesses and other groups of users including the management of, meetings, appointments, and events; and

[0082] 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.

[0083] 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

[0084] The invention will be described by reference to the preferred and alternative embodiments thereof in conjunction with the drawings in which:

[0085] FIG. **1** is an overall diagrammatic view according to one embodiment of the invention;

[0086] FIG. **2** is a schematic block diagram showing the agent architecture according to one embodiment of the invention;

[0087] 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;

[0088] FIG. **4**A is a process for receiving natural language speech-based queries and/or commands and generating a response according to one embodiment of the invention;

[0089] FIG. **4**B 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;

[0090] FIG. **5** is a process for correctly interpreting a user's utterance according to one embodiment of the invention; and

[0091] 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.

[0092] FIG. **7** illustrates a diagram for centrally maintaining context; according to one embodiment of the invention.

[0093] FIG. **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

[0094] 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.

[0095] 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.

[0096] Useful results that are returned are often embedded in other information and may need to be extracted from the other information. For example, a few key words or numbers may need to be "scraped" from a larger amount of other information in a text string, table, list, page, video stream or other information. At the same time, extraneous information including graphics or pictures may be removed to process the response. In any case, the multiple results must be evaluated and combined to form the best possible response, even in cases where some requests do not return useful results or fail to produce results entirely. In cases where the question is determined to be ambiguous or the result is inherently subjective, determining the results to present in the response is a complex process. Finally, to maintain a natural interaction, responses should be returned to the user rapidly. Managing and evaluating complex and uncertain requests, while maintaining real-time performance, is a significant challenge.

[0097] 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.

[0098] 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.

[0099] 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.

[0100] FIG. **8** illustrates one exemplary embodiment of an enhanced conversational user interface **800** that receives multi-modal input from users **802** and communicates with a conversational speech analyzer **804**. The conversational speech analyzer **804** may be coupled to several components, including a general cognitive model **806**, an environmental model **808**, a personalized cognitive model **810**, and an adaptive misrecognition analysis engine **812**, or other components. According to one embodiment of the invention, the conversational speech analyzer **804** may include one or more speech recognition engines that transcribe utterances to textual messages. The transcription may be performed using one or more conversational grammars, a context description grammar **112**, or other transcription methods. According to one embodiment of the invention, data that is entered in a text format may be merge with data that is transcribed to a textual format from the utterance. Conversational speech analyzer **804** also may include a semantic knowledge-based model that analyzes the text message and detects command components.

[0101] 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.

[0102] 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.

[0103] According to one embodiment of the invention, the environmental model **808** may include information associated with the user's environment and surroundings. The information may include the type of environment that a user is in (e.g., quiet or noisy); details of a microphone and/or speaker system; the user's current global position and movement, as may be determined by GPS; current system status, such as what song/movie is playing, is the system in the midst of retrieving something, or other system status; details on all voice-enabled devices in the immediate vicinity, such as a presence of a voice-enabled TV, stereo, and DVD player in the same room; a user's credit card information, such as numbers and current balances, wherein the user may ask a mobile telephone to download and pay for a video and the system may respond that there are insufficient funds; or other information. The information may be accessed to invoke a context, domain knowledge, preferences, and/or other cognitive quality that enhances the interpretation of questions and/or commands.

[0104] 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.

[0105] 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.

[0106] 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 rephrase 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 misrecogni-

tions for particular commands, with the results being used to update the corresponding personalized cognitive model **810**.

[0107] According to another embodiment of the invention, the conversational speech analyzer **804** may access the personalized cognitive model **810** to proactively select a next best (or nth best) match for the received text. A match may be confirmed by user action that include the user not immediately canceling the command. The misrecognitions may also be analyzed to potentially determine personalized tuning parameters for the speech recognition components of the system. For example, the system may, over time, tune the speech recognition engine to increase recognition by analyzing how the speech recognition engine misrecognizes the verbal utterance.

[0108] 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**.

[0109] 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**.

[0110] 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.

[0111] The speech units **128**, multi-modal devices **155** and the main unit **98** may communicate over a communication link. The communication link may include a wired or wireless link. According to one embodiment, the communication link may comprise an RF link. The transceiver **130** on the speech unit may communicate coded speech data bi-directionally over the communication link with the transceiver **126** on the main unit **98**. According to another embodiment, RF link may use any standard local area wireless data protocols including the IEEE 802.11, Bluetooth or other standards. Alternatively, an infrared data link conforming to any suitable standard such as IrDA or other infrared standards can be used. In an alternative embodiment, wires may connect the speech unit **128** and the main unit **98**, eliminating the need for one speech coder **138**. Other wired or wireless analog or digital transmission techniques can be used.

[0112] 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.

[0113] 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**.

[0114] The non-speech interface **114** may be integral with, or separate from, the multi-modal devices **155** and may be used as a substitute or may complement the speech interface. For example, non-speech interface **114** may be used to present and to interact with non-speech (e.g., graphical or tabular) information in a manner more easily understood by the user. According to one embodiment of the invention, multimodal support may be provided to maintain the context during both voice interaction and interaction through the non-speech interface **114**. In one exemplary embodiment, a communication channel may be opened between multimodal devices **155** and the main user interface system **90** to allow multimodal devices **155** to input text commands and questions. The multimodal devices **155** may received and/or provide one or more types of communications, including speech, text, digital audio files, dual-tone-multi-frequency (DTMF) or other communication types. According to another embodiment of the invention, multimodal devices **155** may send a string of text or keywords for a command or question. The main interface system **90** may synchronize the context between multimodal device **155** and the speech units **128**. In order to send a response to the corresponding device, the main user interface system **90** may track where the input came from so that the response may be sent to a TTS or multi-modal device **155**.

[0115] 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.

[0116] 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.

[0117] 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 ran 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.

[0118] 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.

[0119] 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**.

[0120] 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**.

[0121] 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**.

[0122] 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.

[0123] 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.

[0124] 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:

[0125] agents for new domains;

[0126] additional domain knowledge for agents;

[0127] 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;

[0128] links to a preferred set of information sources for the domains covered including links; for, entertainment, news, sports, weather, etc.;

[0129] updates to domain information based on, for example, changes to tax laws, company mergers, changing political boundaries;

[0130] updates to content, including dictionaries, encyclopedias and almanacs; and

[0131] other content and database components.

[0132] When a user requires or selects a new domain agent **156** or database element **102**, the update manager **104** may connect to their source on the network **146** though the network interface **116**, download and install the agent or data. To save system resources and to comply with any license conditions, the update manger **104** may uninstall agents that are no longer in use. In one embodiment of the invention, the update manager **104** may periodically queries one or more sources of the licensed agents and database components to locate and download updates to agent executables, scripts or data as they become available. Alternatively, the agent sources may initiate the downloading of agent updates of the registered or licensed agents to the update manager as they become available.

[0133] 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 downloaded from web sites maintained by groups of domain experts) may also be managed by the agent manager **154**.

[0134] 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.

[0135] 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.

[0136] 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:

[0137] 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.

[0138] 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.

[0139] 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 infor-

mation. Responses can be based on or dependent on thresholds or probabilistic or fuzzy weights for the variables.

[0140] 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.

[0141] 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.

[0142] 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**.

[0143] 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:

[0144] spelling preferences;

[0145] date of birth for user, family and friends;

[0146] income level;

[0147] gender;

[0148] occupation;

[0149] location information such as, home address, neighborhood, and business address;

[0150] car type;

[0151] telecommunications and other service providers and services;

[0152] financial and investment information;

[0153] synonyms (i.e., a nick name for someone);

[0154] special spelling;

[0155] keywords;

[0156] transformation or substitution variables;

[0157] domains of interest; and

[0158] other values.

[0159] 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:

[0160] specify the command to run a report or alert;

[0161] specify the question or questions, including keywords, used for a query;

[0162] set the criteria for running the report such as on command or when a particular condition is met;

[0163] define preferred information sources;

[0164] define preferences for order of result evaluation by source, value, etc.;

[0165] 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

[0166] 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.

[0167] 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.

[0168] 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.

[0169] 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.

[0170] 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.

[0171] 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.

[0172] 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.

[0173] 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.

[0174] 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.

[0175] 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.

[0176] 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**.

[0177] 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").

[0178] 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 **10**, 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.

[0179] 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.

[0180] 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 pronun-

ciation 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.

[0181] Once the words and/or phrases have been recognized by the speech recognition engine **120**, the tokens and user identification is passed to the parser **118**. The parser **118** examines the tokens for the questions or commands, context and criteria. The parser **118** determines a context for an utterance by applying prior probabilities or fuzzy possibilities to keyword matching, user profile **110**, and dialog history. The context of a question or command determines the domain and thereby, the domain agent **156**, if any, to be evoked. For example, a question with the keywords "temperature" implies a context value of weather for the question. The parser dynamically receives keyword and associated prior probability or fuzzy possibility updates from the system agent **150** or an already active domain agent **156**. Based on these probabilities or possibilities the possible contexts are scored and the top one or few are used for farther processing.

[0182] 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.

[0183] 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.

[0184] 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.

[0185] 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.

[0186] 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:

- **[0187]** substituting a stock symbol for a company name or abbreviation;
- **[0188]** substituting a numerical value for a word or words;
- **[0189]** adding a zip code to an address; and
- **[0190]** changing a place or other name to a commonly used standard abbreviation.

[0191] 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.

[0192] 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.

[0193] Specific questions are generally directed to one of the domain agents **156**. Based on the question or context and the parameters or criteria, the domain agent creates one or more queries to one or more local or external information sources. Questions can be objective or subjective in nature. Results for objective questions can often be obtained by structured queries to one or more local or network information sources. Even for objective questions, the system **90** may need to apply probabilistic or fuzzy set analysis to deal with cases of conflicting information or incomplete information. Information to answer subjective questions is generally obtained by one or more ad-hoc queries to local or network data sources, followed by probabilistic or fuzzy set evaluation of the one results to determine a best answer.

[0194] 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.

[0195] 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**.

[0196] 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.

[0197] 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:

[0198] 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.

[0199] 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.

[0200] 3. Wait for additional queries to return results.

[0201] 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.

[0202] 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.

[0203] 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.

[0204] 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**.

[0205] 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:

[0206] substitution of words for numbers;

[0207] substitution of names for acronyms or symbols (i.e., trading symbols);

[0208] use of formatting information derived from the information sources (i.e., HTML tags);

[0209] nature of the response including, text, long text, list, table;

[0210] possible missing information or errors;

[0211] units for measurement (i.e., English or metric); and

[0212] preferred terminology from the user profile **110** or presentation personality **108**.

[0213] 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 her. 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.

[0214] 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.

[0215] 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.

[0216] 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**.

[0217] 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.

[0218] FIG. **4**A illustrates an exemplary process **400** for receiving natural language speech-based requests and/or commands and generating a response using the system **90** according to another embodiment of the invention. The process **400** may begin when a user's utterance (i.e., user query and/or command) is captured using speech recognition that operates in a variety of real-world environments at **402**. Once the utterance is captured, it is parsed and interpreted to determine the query and/or command that is contained in the utterance at **404**. According to one embodiment of the invention, the utterance may be transcribed to text and analyzed using the general cognitive model **806**, the environmental model **808**, and/or the personalized cognitive model **810**. Next, the request and/or command are reviewed to determine the domain of expertise required and the context of the query, invoking the proper resources including, for example, agents at **406**. The original query and/or command submitted by the user may require that multiple queries and/or commands be generated.

[0219] 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.

[0220] 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.

[0221] 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.

[0222] 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.

[0223] FIG. **4**B 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**.

[0224] 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**.

[0225] Some of the operations depicting in FIG. **4**A 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.

[0226] 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].

[0227] 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. **4**A. 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. **4**A and **4**B, is parsed at operation **602**.

[0228] 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 **10**, and the agent.

[0229] 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, Right **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**.

[0230] 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**.

[0231] 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.

[0232] The descriptions above are intended to be illustrative, not limiting. Thus, it will be apparent to one skilled in the art that modifications may be made to the invention as described without departing from the scope of the claims set out below.

We claim:

1. A system for processing natural language utterances, comprising:

a multimodal device configured to receive a natural language utterance;

a speech recognition engine configured to recognize one or more words from the natural language utterance;

a parser configured to generate an interpretation of the natural language utterance from the one or more recognized words, and further configured to generate a request based on the interpretation of the natural language utterance;

a domain agent configured to process the generated request; and

an adaptive misrecognition engine configured to:

monitor one or more actions associated with the domain agent processing the request; and

determine whether the interpretation of the natural language utterance is correct or incorrect based on the one or more monitored actions.

2. The system of claim **1**, wherein the adaptive misrecognition engine is further configured to generate an unrecognized event in response to determining that the interpretation of the natural language utterance is incorrect.

3. The system of claim **2**, farmer comprising an analyzer configured to:

track an interaction pattern with the system over time for a user that provided the natural language utterance;

generate a personalized cognitive model for the user based on the interaction pattern tracked for the user; and

use the personalized cognitive model to predict the one or more actions associated with the domain agent processing the request.

4. The system of claim **3**, wherein the analyzer is further configured to update the personalized cognitive model based on a frequency of incorrect interpretations for the request.

5. The system of claim **3**, wherein the parser is further configured to generate a plurality of interpretations of the natural language utterance, and wherein the analyzer is further configured to use the personalized cognitive model to select a next best interpretation from the plurality of interpretations in response to the adaptive misrecognition engine determining that the interpretation generated by the parser is incorrect.

6. The system of claim **3**, wherein the analyzer is further configured to track interaction patterns with the system over time for a plurality users.

7. The system of claim **6**, wherein the analyzer is further configured to generate a generalized cognitive model for the plurality of users based on the interaction patterns tracked for the plurality of users, wherein the generalized cognitive model includes a statistical abstract that corresponds to the tracked interaction patterns.

8. The system of claim **7**, wherein the analyzer is further configured to use the generalized cognitive model to predict the one or more actions associated with the domain agent processing the request.

9. The system of claim **7**, wherein the analyzer is further configured to update the generalized cognitive model based on a frequency of incorrect interpretations for the request.

10. The system of claim **3**, wherein the analyzer is further configured to generate an environmental model that includes information associated with at least one of environmental conditions or surroundings associated with the user.

11. The system of claim **10**, wherein the environmental conditions or surroundings include one or more of a global position of the user, movement information associated with the user, quiet or noisy conditions associated with an environment of the user, or a vicinity to one or more voice-enabled devices.

12. The system of claim **10**, wherein the environmental model provides one or more of context, domain knowledge, preferences, or cognitive qualities to enhance the interpretation of the natural language utterance.

13. The system of claim **2**, further comprising an analyzer configured to:

analyze the unrecognized event to determine how the natural language utterance was incorrectly interpreted; and

determining one or more tuning parameters for at least one of the speech recognition engine or the parser based on how the natural language utterance was incorrectly interpreted, wherein the tuning parameters are used to improve interpretations of subsequent natural language utterances relating to the request.

14. The system of claim **1**, further comprising:

a knowledge-enhanced speech recognition engine configured to determine a most likely context for the natural language utterance, wherein the knowledge-enhanced speech recognition engine is further configured to:

identify one or more contexts that completely or partially match one or more text combinations, wherein identifying the one or more contexts includes comparing the text combinations against one or more grammar expression entries in a context description grammar;

provide a relevance score for each of the identified matching contexts; and

select the matching context having a highest score as the most likely context for the natural language utterance, wherein the domain agent configured to process the generated request is associated with the selected context; and

a response generating module configured to:

communicate the request to the domain agent associated with the selected context; and

generate a response to the natural language utterance using content gathered as a result of the domain agent processing the request, wherein the response arranges the content in an order based on the relevance scores for the identified matching contexts.

15. The system of claim **14**, wherein the response generated by the response generating module includes an aggregation of the content gathered as a result of the domain agent processing the request.

16. The system of claim **14**, further comprising a personality module configured to format the response.

17. The system of claim **14**, wherein the knowledge-enhanced speech recognition engine identifying the one or more contexts further includes comparing the text combinations against a context stack that stores one or more expected contexts.

18. The system of claim **14**, wherein the knowledge-enhanced speech recognition engine identifying the one or more contexts further includes applying prior probabilities or fuzzy possibilities to at least one of keyword matching, user profiles, or a dialog history.

19. The system of claim **14**, wherein the domain agent is further configured to direct a query to at least one of a local information source or a network information source to process the request.

20. The system of claim **19**, wherein the domain agent is further configured to evaluate a plurality of responses to the query to process the request.

21. The system of claim **14**, wherein the domain agent is further configured to direct a command to at least one of a local device or a remote device to process the request.

22. The system of claim **1**, wherein the multimodal device includes at least one of a personal digital assistant, a cellular telephone, a portable computer, or a desktop computer.

23. The system of claim **1**, wherein the multimodal device is further configured to subsequently receive one or more follow-up multimodal inputs.

24. The system of claim **23**, wherein the speech recognition engine is further configured to recognize one or more words from a natural language utterance provided in the follow-up multimodal input, and wherein the parser is further configured to generate an interpretation of the follow-up multimodal input from the one or more words recognized from the natural language utterance provided in the follow-up multimodal input.

25. The system of claim **24**, wherein the follow-up multimodal input includes a follow-up request associated with a same context as the request being processed by the domain agent.

26. The system of claim **1**, wherein the adaptive misrecognition engine determines that the interpretation of the natural language utterance was incorrect in response to a user providing a subsequent request to stop the request being processed by the domain agent.

27. The system of claim **1**, wherein the adaptive misrecognition engine determines that the interpretation of the natural language utterance was incorrect in response to a user repeating the natural language utterance.

28. The system of claim **1**, wherein the multimodal device is Other configured to receive a non-speech input relating to the natural language utterance, and wherein the system further comprises:

a transcription module configured to transcribe the non-speech input to create a non-speech-based transcription; and

a merging module configured to merge the recognized words and the non-speech-based transcription to create a merged transcription, wherein the parser is farther configured to generate the interpretation of the natural language utterance from the merged transcription.

29. A method for processing natural language utterances, comprising:

receiving a natural language utterance at a multimodal device;

recognizing one or more words from the natural language utterance using a speech recognition engine coupled to the multimodal device;

generating an interpretation of the natural language utterance from the one or more recognized words using a parser coupled to the multimodal device, wherein the parser further generates a request based on the interpretation of the natural language utterance;

invoking a domain agent configured to process the generated request;

monitoring one or more actions associated with the domain agent processing the request using an adaptive misrecognition engine; and

determining whether the interpretation of the natural language utterance is correct or incorrect based on the actions monitored using the adaptive misrecognition engine.

30. The method of claim **29**, further comprising generating an unrecognized event in response to the adaptive misrecognition engine determining that the interpretation of the natural language utterance is incorrect.

31. The method of claim **30**, further comprising:

tracking an interaction pattern over time for a user that provided the natural language utterance;

generating a personalized cognitive model for the user based on the interaction pattern tracked for the user; and

using the personalized cognitive model to predict the one or more actions associated with the domain agent processing the request.

32. The method of claim **31**, further comprising updating the personalized cognitive model based on a frequency of incorrect interpretations for the request.

33. The method of claim **31**, further comprising:

generating a plurality of interpretations of the natural language utterance; and

using the personalized cognitive model to select a next best interpretation from the plurality of interpretations in response to the adaptive misrecognition engine determining that the interpretation generated by the parser is incorrect.

34. The method of claim **31**, further comprising tracking interaction patterns over time for a plurality of users.

35. The method of claim **34**, further comprising generating a generalized cognitive model for the plurality of users based on the interaction patterns tracked for the plurality of users, wherein the generalized cognitive model includes a statistical abstract that corresponds to the tracked interaction patterns.

36. The method of claim **35**, further comprising using the generalized cognitive model to predict the one or more actions associated with the domain agent processing the request.

37. The method of claim **35**, further comprising updating the generalized cognitive model based on a frequency of incorrect interpretations for the request.

38. The method of claim **30**, hither comprising generating an environmental model that includes information associated with at least one of environmental conditions or surroundings associated with the user.

39. The method of claim **38**, wherein the environmental conditions or surroundings include one or more of a global position of the user, movement information associated with the user, quiet or noisy conditions associated with an environment of the user, or a vicinity to one or more voice-enabled devices.

40. The method of claim **38**, wherein the environmental model provides one or more of context, domain knowledge, preferences, or cognitive qualities to enhance the interpretation of the natural language utterance.

41. The method of claim **30**, hither comprising:

analyzing the unrecognized event to determine how the natural language utterance was incorrectly interpreted; and

determine one or more tuning parameters for at least one of the speech recognition engine or the parser based on how the natural language utterance was incorrectly interpreted, wherein the tuning parameters are used to improve interpretations of subsequent natural language utterances relating to the request.

42. The method of claim **29**, further comprising determining a most likely context for the natural language utterance using a knowledge-enhanced speech recognition engine, wherein determining the most likely context further includes:

identifying one or more contexts completely or partially match one or more text combinations, wherein identifying the one or more contexts includes comparing the text combinations against one or more grammar expression entries in a context description grammar;

providing a relevance score for each of identified matching contexts;

selecting the matching context having a highest score as the most likely context for the natural language utterance, wherein the domain agent configured to process the generated request is associated with the selected context;

communicating the request to the domain agent associated with the selected context; and

generating a response to the natural language utterance using content gathered as a result of the domain agent processing the request, wherein the response arranges the content in an order based on the relevance scores for the identified matching contexts.

43. The method of claim **42**, wherein the response includes an aggregation of the content gathered as a result of the domain agent processing the request.

44. The method of claim **42**, further comprising formatting the response using a personality module.

45. The method of claim **42**, wherein identifying the one or more contexts further includes comparing the text combinations against a context stack that stores one or more expected contexts.

46. The method of claim **45**, wherein identifying the one or more contexts further includes applying prior probabilities or fuzzy possibilities to at least one of keyword matching, user profiles, or a dialog history.

47. The method of claim **29**, further comprising receiving one or more follow-up multimodal inputs at the multimodal device.

48. The method of claim **29**, wherein the adaptive misrecognition engine determines that the interpretation of the natural language utterance was incorrect in response to the user providing a subsequent request to stop the request being processed by the domain agent.

49. The method of claim **29**, wherein the adaptive misrecognition engine determines that the interpretation of the natural language utterance was incorrect in response to the user repeating the natural language utterance.

50. The method of claim **29**, further comprising:

receiving a non-speech input relating to the natural language utterance at the multimodal device;

transcribing the non-speech input to create a non-speech-based transcription; and

merging the recognized words and the non-speech-based transcription to create a merged transcription wherein the parser is further configured to generate the interpretation of the natural language utterance from the merged transcription.

* * * * *