[056] Figure 8 illustrates a system diagram for an enhanced conversational voice user interface according to one embodiment of the invention.

DETAILED DESCRIPTION

[057] The following detailed description refers to the accompanying drawings, and describes exemplary embodiments of the invention. Other embodiments are contemplated and modifications may be made to the exemplary embodiments without departing from the spirit, functionality and scope of the invention. Therefore, the following detailed descriptions are not meant to limit the invention.

[058] According to one embodiment of the invention, a telematic natural language speech interface and non-speech interface are provided for use in mobile environments and telematic applications. The system, or portions thereof, may be used in vehicles, while on foot or at a fixed location such as an office or home, or at other locations. An overall block diagram of one embodiment of the invention is illustrated in Figure 1.

[059] A speech unit 128 and/or a keypad 14 may be integrally coupled to a mobile structure 10 or may be part of mobile devices 36, fixed home or office computer systems 44, or other devices. Mobile devices 36 may include mobile telephones, personal digital assistants, digital radios, compact disk players, navigation systems, or other mobile devices. The mobile devices 36 may be configured to integrate with set-top boxes, alarm clocks, radios, or other electronic components. The speech unit 128 and/or keypad 14 may be interfaced to a Telematics Control Unit (TCU) 28 through one or more data interfaces 26. According to some embodiments, the main speech-processing unit 98 may be embedded in one or more TCU 28. In some embodiments, the components of the speech unit 128 may also be distributed between one or more TCUs.

[060] A speech-processing unit may be built into mobile devices 36 and may be coupled with the data interfaces 26 though a wireless, or wired, handheld interface 20. Other user interface peripherals may be connected to the TCU through the data interfaces and may include displays 18, such as touch screen displays for displaying text, graphics and video; keypads 14 for receiving textual data input; video cameras 16 for receiving multimedia communications or conferences; a pointing device or stylus, or other user interface peripherals. Other devices may be connected to the TCU through the data interfaces including wide-area RF transceivers 24, navigation

system components 22, or other devices. The navigation system may include several components such as, Global Positioning System (GPS) receivers or other radiolocation system receivers, gyroscopes or other inertial measurement equipment, distance measurement sensors, such as odometers, or other components. Radiolocation equipment may receive coded signals from one or more satellites or terrestrial sources 40. The one or more location service servers 48 may assist the navigation system. Other systems that can connect to the TCU through the data interfaces may include automotive control computers, digital control interfaces for devices such as media players or other electronic systems, measurement sensors, or other specialized electronic equipment.

[061] The control and device interfaces 30 may connect the TCU 28 to various devices 32 within the mobile structure 10. The control and device interfaces 30 may be used to execute local or remote commands from users of the natural language interface. In some cases, the control and device interfaces 30 may include specialized hardware for interacting with different types of devices. The hardware interfaces may include analog or digital signal interfaces for device control, along with analog or digital interfaces for measurements that may control devices 32. These interfaces may also include specialized software that encapsulates or abstracts specific behavior of the devices 32. The interface software may include one or more drivers that are specific to the hardware interface and to one or more agents. The domain agents may include the specialized software behavior and data required to control a particular device or class of devices. New or updated behavior may be added to the system by updating data managers that are associated with specific devices or class of devices. The devices 32 may include manual controls or manual overrides 34. For safety reasons, the control and device interfaces 30 may incorporate fail-safe systems that, for example, may verify operating limits before changing settings, ensuring that commands do not conflict with settings from manual controls, and will not, in some combination with other commands or device settings, create an unsafe situation. The software behavior and data that ensures safe operations may be included within the domain agent that is specific to the device or class of devices. Examples of devices and system that can be controlled through the control and device interfaces 30 include, power management systems, measurement sensors, door locks, window controls, interior temperature controls, shifting of the transmission, turn signals,

lights, safety equipment, engine ignition, cruise control, fuel tank switches, seat adjustments, specialize equipment such as winches, lifting systems or loading system, and other systems.

[062] The wide-area RF transceiver 24 may communicate with one or more wide-area wireless networks 38, which may be connected to data networks 42, including the Internet, the Public Switched Telephone Network (PSTN) 42 or other data networks. The wide-area wireless networks can be of any suitable terrestrial or satellite based type. Mobile devices 36 may communicate with one or more local or wide-area wireless networks. Home or office systems 44, equipped with wired or wireless network interfaces, may communicate through the data networks or PSTN.

[063] According to one embodiment of the invention, data and agents may be stored and synchronized in mobile structures 10, mobile devices 36 and/or fixed systems 44 having one or more main speech-processing units 98. The synchronization between these different systems can occur on the wide area wireless network 38, the data network 42, through the handheld interface 20, or other local data connections. The synchronization may be performed automatically when any two or more of the computers are connected to these networks. Alternatively, the synchronization may be performed on demand under user control. The synchronization process attempts to determine which version of a data element or an agent are the newest or most up-to-date and propagates that element. Thus, synchronization is an incremental change processes. In some cases, a complete replacement of a database, a portion of the database or of one or more agents may be performed rather than performing a series of incremental updates.

[064] The wide-area wireless networks 38, the data networks 42 or PSTN, may connect mobile structures 10, mobile devices 36 and fixed computers 44 to one or more servers that provide one of more services. According to one embodiment of the invention, an interactive natural language user interface may be provided that supports a transfer of data or transmission of speech, text, video and other formats. For data centric applications, a standardized data transfer format may be used including, for example, Hypertext Markup Language over Hypertext Transfer Protocol (HTTP), Extensible Markup Language (XML), and other data formats or schemas, over HTTP or other transfer protocol, Electronic Data Interchange formats over a variety of transport protocols, etc. It will be understood that the exact configuration of the

servers may be determined on a case-by-case basis with consideration being given to the exact combinations of services being offered, the service providers providing the services, the contractual relationships between the service provider, and other factors. The invention supports any suitable configuration. In each case these servers may themselves be distributed over one or more public or private networks. Some examples of servers that may be used to deliver these services, are given below:

1. One or more payment service providers 56 that supply payment capabilities to users. These payment services may include electronic wallet capabilities for one or more payment accounts and may include payment security information, payment account information, transaction histories, account balance information, and other information. Exemplary payment types include, stored value accounts, promotional accounts, credit accounts, telecommunications billing accounts, debit accounts, or other payment types that are accessed using online or offline methods. Payments can be computed in any manner including payment for a specific good or service, subscription payment, metered payment, or other payments. The payment services can be distributed in a number of ways. Examples of computers and servers used to store and process payment transaction information include, smart cards, main speech processing units 128, mobile devices 36, TCUs 28, fixed personal computers 44, payment gateways, payment servers 56, or other systems.

2. One or more Customer Relationship Management (CRM) systems 52 may supply any number of consumer and business customer services as discussed above. The CRM system can supply automated services or services that are partly or completely manual. For manual services one or more customer service representatives may use one or more service representative workstations 54. The CRM system and the service representative workstation may be connected to one or more data networks 42 or the PSTN. Any other servers may also have connections to one or more service representative workstations, which may be in common or independent of each other.

3. One or more specialized service servers 50 may support specialized consumer and business services.

4. One or more location service servers 48 may supply location information and location based services. The location data is used as a data

input to the location services that may then be distributed in any suitable manner including, on main speech processing units 128, mobile devices 36, TCUs 28, fixed personal computers 44, other servers (i.e. 46, 50, 52, 56) or the one or more location services servers 48.

5. One or more emergency services servers 46 may supply both public and private emergency services to the users.

[065] Main speech processing unit 98, speech unit 128 and keyboard 14 may be distributed in mobile devices 36 in a number of ways. For example, these units may be attached to the mobile devices 36 as independent components or as a single integrated component. In another embodiment of the invention, some or all of the main speech processing unit 98 and speech unit 128 may be embedded in one or more of the TCUs 28, mobile devices 36, fixed computer systems 44, or other devices.

[066] Figure 2 illustrates an alternative block diagram of a second embodiment of the invention. In this embodiment, the main speech processing unit 98 and speech unit 128 are external to the TCU 28. These components may be housed in one or more packages or may be included in a single integrated package.

[067] In all other respects, the second embodiment resembles the first embodiment illustrated in Figure 1. It will be understood that the exact distribution and packaging of the main speech processing unit 98 and speech unit 128 may be determined by the details of the deployment situation and should not change the functionality, capabilities or spirit of the invention in any way.

[068] Figure 3 illustrates a block diagram of one embodiment of the mobile devices 36 that include main speech processing unit 98 and speech unit 128 embedded in the mobile device 36, mobile telephone or other mobile device. The main speech processing unit may interface to one or more processing units 70 located in the mobile device 36. The processing units 70 may include one or more central processing units, one or more data and address busses, data interfaces, volatile memory, or other components. The processing units 70 may use one of more types of nonvolatile memory 80 for software and data storage. Suitable types of nonvolatile memory 80 may include flash memory and hard disk drives. In some embodiments, the main speech processing unit 98 may be integrated with the one or more processing units 70.

[069] According to one embodiment, users may interact with the mobile device 36 through the speech unit 128, the keypad 74 or keyboard, a display 72 that displays

text, graphics, video, or other peripheral. In some embodiments the display may be a touch screen type. Alternatively, a pointing device (not shown) may be used, among other devices.

[070] The mobile device 36 may connect to one or more wired or wireless wide-area or local-area networks through one or more interfaces. A wide-area network transceiver 78 may connect to the wide-area wireless network 38 or the data network 42 using a wireless or wired connection, including an IP connection, a dial-up PSTN network connection or other connections. The local-area network transceiver 76 may connect to wired or wireless local area networks. These networks may include the handheld interface 20 or connections to fixed computer systems 44. In both mobile device environments and fixed computer environments, communications between humans and machines may not provide accurate results at least because natural language commands may be incomplete in their definition. The occurrence of inaccurate results may be reduced by leveraging context of previous utterances, knowledge of the domain and/or the user's history of interests and preferences or other factors.

[071] According to one embodiment of the invention, processing may be performed at the mobile devices 36. Alternatively, the processing may be performed on the server side. In yet another embodiment, the processing may occur on both the mobile devices 36 and the server side. The commands may be processed on-board the mobile devices 36 to enable the mobile devices 36 to control themselves and/or to control other mobile devices 36, fixed computers 44, mobile telephones, or other devices. Additionally, mobile devices 36 may track context. According to another embodiment of the invention illustrated in FIG. 7, a context manager 702 may be centrally maintained to allow input to and output from multiple mobile devices 36. Each of the mobile devices may communicate with the context manager 702 to register via registration module 712, wherein registration may indicate events that the mobile devices 36 may subscribe to. The context manager 702 may receive input in Context XML form, for example. The other registered mobile devices 36 may be informed of context changes through context tracking module 714, thereby synchronizing the context across the registered mobile devices 36. According to one embodiment of the invention, registered mobile devices 36 may be added or removed.

The registered mobile devices 36 may be dynamic link libraries (DLLs), object code, or other data that may be specific to the mobile device 36.

[072] According to yet another alternative embodiment of the invention, context information may be determined from a command or request that is presented as a text message and/or a command or request that is presented as a verbal utterance and processed using a multi-pass automatic speech recognition module that transcribes the verbal utterance to a text message. The multi-pass automatic speech recognition module may use a dictation grammar or a large vocabulary grammar, among other resources, to transcribe the verbal utterance into a text message. On platforms where a dictation grammar is not available, the multi-pass automatic speech recognition may use a virtual dictation grammar that uses decoy words for out-of-vocabulary words. Decoy words may include utility words, nonsense words, isolated syllables, isolated distinct sounds and other decoy words.

[073] According to one embodiment of the invention, the text message may be searched for a particular character, group of characters, word, group of words, and other text combinations. The text combination may be compared against entries in a context description grammar that is associated with each agent 106. If a match is identified between an active grammar in the context description grammar and the command and/or request, then the match may be scored. The agents 106 may be ranked based on the determined score. In generating an aggregate response from the one or more responses received from the agents 106, the ordering of the responses from the individual agents may be determined based on the rank of agents 106. The aggregate response may be generated by a response generator module. Agents 106 may update a context stack, that includes an ordered list of command contexts, to enable follow-up requests.

[074] According to another embodiment of the invention, if a match is not found, or only a partial match is found, between the text message and active grammars, then a knowledge-enhanced speech recognition system may be used to semantically broaden the search. The knowledge-enhanced speech recognition system may be used to determine the intent of the request and/or to correct false recognitions. The knowledge-enhanced speech recognition may access a set of expected contexts that are stored in a context stack to determine a most likely context. The knowledge-enhanced speech recognition may use context specific matchers that are able to

identify context such as time, location, numbers, dates, categories (e.g., music, movies, television, addresses, etc.) and other context. The matching may be performed by comparing a character, group of characters, a word, group of words, and other text combinations. Alternatively, or in addition to text based matching, the matching may be performed using phonetic matching, among other techniques. The results of any match may be used to generate a command and/or request that is communicated to agents 106 for additional processing. According to one embodiment of the invention, non-speech interface 114 may show system, state and history information in a more concise manner than is possible through the speech interface. Non-speech interface 114 may be accessed to create or extend capabilities of agents 106. These operations may include scripting of agents, adding data to the agent or databases 102 used by the agent, adding links to information sources, among other operations.

[075] According to another embodiment of the invention, mobile devices 36 may be configured to allow speech annotations of objects stored thereon. The objects may include photographs, calendar entries, email messages, instant messages, phonebook entries, voice mail entries, digital movies or other objects. Mobile devices 36 may transcribe the speech annotations to textual annotations and store the textual annotations with the object. Alternatively, mobile devices 36 may be configured to enable users to input non-speech annotations, such as textual descriptions, that are stored along with the objects.

[076] According to one embodiment of the invention, the annotated objects may be stored on a server side, a client side, a combination of server side and client side, or according to other configurations. The invention further contemplates collaboratively exchanging and sharing the annotated objects among distributed workgroups that may include centralized servers having shared workspaces for providing common object storage and retrieval facilities. The shared workspaces may be implemented on the centralized servers and may be accessed from different platforms using the mobile devices 36. The system may include a peer-to-peer system for accessing the annotated objects.

[077] According to one embodiment of the invention, the non-speech annotations may be classified and searched. In an alternative embodiment, the speech annotations may be classified and searched. However, classifying and searching speech

annotations may be considerably more difficult than classifying and searching non-speech annotations.

[078] According to another embodiment of the invention, the textual annotations and/or textual descriptions that are associated with the objects may be stored as metadata, thereby enabling searching for the objects using the metadata. The metadata may include GPS information, environmental information, geographic information, or other information. For example, proximity to famous landmarks may be determined using GPS information, environmental information, geographic information, or other information, and this information may be integrated into the metadata associated with the objects. According to one embodiment of the invention, GPS coordinates may be stored in the metadata associated with the objects and users may search for selected objects based on the GPS coordinates. A user may provide a speech command such as "show me all the photos of Greece." In this case, the system would limit the type of object to photographs and would determine the GPS coordinates of Greece. The system would then search the metadata for objects that correspond to photographs and that also satisfy GPS coordinates for Greece. According to another embodiment of the invention, including GPS coordinates in the metadata of objects enables post-processing of the objects based on the GPS coordinates. For example, the objects may be subjected to an initial sort based on generalized GPS coordinates that are stored in the metadata and may be subjected to additional sorting based on more particular criteria for the GPS coordinates. Thus, a user may initially search for object metadata that corresponds to a location near a famous landmark and may use image matching to label the objects with the searchable metadata (i.e., textual descriptions). For example, using the GPS coordinates stored in the metadata, users may first determine that photographs were taken at the Jefferson Memorial and may use this information to label the photographs with searchable metadata (i.e., textual descriptions) including "photo of the Jefferson Memorial." One of ordinary skill in the art will readily appreciate that metadata may include various types of information and may be searched using the various types of information.

[079] According to one embodiment of the invention, the textual annotations may be communicated using a short message service on the mobile telephone or other device. Short message service is a text message service that enables sending and receiving of

short textual messages. The textual messages may be stored at data centers for forwarding to intended recipients. Other configurations may be used.

[080] According to another embodiment of the invention, the mobile devices 36 may support multi-modal communications that enable displaying of non-speech search results on a graphical interface and receipt of speech commands to provide a follow-up search, among other configurations. For example, the user may be presented with textual search results corresponding to a name of a famous person and the user may provide a speech command to find a biography of the famous person. The system may maintain the context of the textual search results to find the biography associated with the famous person. By contrast, known systems may perform a follow-up search on the term "biography" and may present a dictionary definition of the term "biography."

[081] Figure 4 illustrates an embodiment that uses a fixed computer 44. In some embodiments, the main speech processing unit 98 and speech unit 128 may be embedded into the fixed computer 44. The main speech processing unit may interface with one or more processing units 84 that are associated with the fixed computer. The processing units may include one or more central processing units, one or more data and address busses, data interfaces, volatile memory, or other components. The processing unit may use one of more types of nonvolatile memory 94 for software and data storage. Suitable types of nonvolatile memory include, for example, flash memory and hard disk drives. In some embodiments, the main speech processing unit 98 may be integrated with the one or more processing units 84. According to one embodiment of the invention, processing may be performed at the fixed computers 44. The commands may be processed on-board to enable the fixed computers 44 to control themselves and/or to control other fixed computers 44, mobile devices 36, mobile telephones, or other devices. Additionally, fixed computers 44 may track context as described above.

[082] According to another embodiment of the invention, fixed computers 44 may be configured to allow verbal annotations of objects stored thereon. Fixed computers 44 may transcribe the verbal annotation to text and store the textual annotation with the object. Alternatively, fixed computers 44 may be configured to enable users to manually input textual descriptions that are stored along with the objects. According to one embodiment of the invention, the textual annotations and/or textual

descriptions may be classified and searched. In an alternative embodiment, fixed computers 44 may classify and search the verbal annotations rather than textual annotations. However, classifying and searching verbal annotations may be considerably more difficult than classifying and searching textual annotations and/or textual descriptions.

[083] In another embodiment of the invention, users may interact with fixed computers 44 using speech unit 128, the keyboard 88 or keypad, a display 86 for displaying text, graphics, video, or other peripherals. According to some embodiments of the invention, the display may be a touch screen type. Alternatively, a pointing device (not shown) may be used, along with other devices. Fixed computers 44 may be coupled to one or more wired or wireless wide-area or local-area networks through one or more interfaces. A wide-area transceiver 92 may connect to the wide-area wireless network 38 or the data network 42, using a wireless or wired connection, including an IP network, a dial-up PSTN network connection, or other connections. The local-area network transceiver 90 may connect to wired or wireless local area networks. These networks may include connections to mobile devices 36.

[084] In order for devices to properly respond to requests and/or commands that are submitted in a natural language form, machine processable requests and/or algorithms may be formulated after the natural form questions or commands have been parsed and interpreted. Algorithms describe how the machines should gather data to respond to the questions or commands. Depending on the nature of the requests or commands, there may not be a simple set of requests and/or algorithms that will return an adequate response. Several requests and algorithms may need to be initiated and even these requests and algorithms may need to be chained or concatenated to achieve a complete response. Further, no single available source may contain the entire set of results needed to generate a complete response. Thus, multiple requests and/or algorithms, perhaps with several parts, may be generated to access multiple data sources that are located both locally or remotely. Not all of the data sources, requests and/or algorithms may return useful results or any results at all. Useful results that are returned are often embedded in other information and may need to be extracted from the other information. For example, a few key words or numbers may need to be "scraped" from a larger amount of other information in a text string, table, list, page,

video stream or other information. At the same time, extraneous information including graphics or pictures may be removed to process the response. In any case, the multiple results must be evaluated and combined to form the best possible response, even in cases where some requests do not return useful results or fail to produce results entirely. In cases where the command is determined to be ambiguous or the result is inherently subjective, determining the results to present in the response is a complex process. Finally, to maintain a natural interaction, responses should be returned to the user rapidly. Managing and evaluating complex and uncertain requests, while maintaining real-time performance, is a significant challenge.

[085] The invention provides a complete speech-based command generation, information query, retrieval, processing and presentation environment or a combination of speech-based and non-speech-based command generation, information query, retrieval, processing and presentation environment for telematic applications. In addition, the invention may be useful for controlling the system itself and/or external devices. This integrated environment makes maximum use of context, prior information and domain and user specific profile data to achieve a natural environment for one or more users submitting requests or commands in multiple domains. Through this integrated approach, a complete speech-based natural language command, algorithm and response environment or a combination of speech-based and non-speech-based command, algorithm and response environment may be created.

[086] The telematic natural language interface may be deployed as part of, or a peripheral to a TCU or other mobile devices 36, as part of a mobile device interfaced to vehicle computers and other mobile system through wired, wireless, optical, or other types of connections or fixed computers interfaced to the vehicle computers or other systems through a combination of wired, wireless, optical and/or other types of connections. Alternatively, the components of the interactive natural language telematic interface may be distributed in any suitable manner between these multiple computing platforms.

[087] Figure 5 illustrates one exemplary schematic diagram view of a system 90 according to an embodiment of the invention. The system 90 may include a main unit 98, speech units 128, and multi-modal devices 155. In an alternative embodiment, the system 98 may include completely separate systems for main unit 98, speech units

128 and multi-modal devices 155. The event manager 100 may mediate interactions between components of main unit 98. The event manager 100 provides a multi-threaded environment allowing the system 98 to operate on multiple commands or questions from multiple user sessions, without conflict, and in an efficient manner to maintain real-time response capabilities.

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

[089] According to one embodiment of the invention, personalized cognitive model 810 is a model derived from a user's interaction pattern with the system and may be used to predict what actions the user may take next in time, thus assisting with speech recognition and/or question or command recognition. Personalized cognitive model 810 may track actions performed by the user. When the system is attempting to predict user behavior, the personalized cognitive model may be consulted first. The system may have multiple personalized cognitive models, wherein one may correspond to each user.

[090] According to another embodiment of the invention, general cognitive model 806 is a statistical abstract that corresponds to interaction patterns with the system for multiple users. Data stored within general cognitive model 806 may be used to predict a user's next action, thus assisting with speech recognition and/or question or command recognition. The general cognitive model 806 may also track what actions

a particular user has performed and may be used when the user interacts with the system in a way that is not handled in the personalized cognitive model.

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

[092] The conversational speech analyzer 804 may also access the general cognitive model 806 and/or the personalized cognitive model 810 to further refine a context, domain knowledge, preferences, and/or other cognitive qualities to enhance the interpretation of questions and/or commands. Based on information received from general cognitive model 806, environmental model 808 and/or the personalized cognitive model 810, the system may enhance responses to commands and questions by including a prediction of user behavior.

[093] Adaptive Misrecognition Analysis Engine 812 may analyze and store textual messages, including transcribed utterances, that are identified as being unrecognized or incorrectly recognized by conversational speech analyzer 804. Upon a determination that text is unrecognized, the system may generate an unrecognized event. For example, an unrecognized event may result from not finding a match to text and/or the transcribed utterance.

[094] According to one embodiment of the invention, the system may implement one or more techniques to determine that textual messages are incorrectly recognized. For example, the user may command the system to play a specific song title and the system may misrecognize the requested song title and provide a song having a different title, the system may misrecognize an incorrect song title and

provide a song having a different than requested title, the system may misrecognize a correct song title and provide a song having an invalid song title, among other variations. When the system misrecognizes a request, the user typically provides immediate feedback such as, overriding the command in a time shorter than the expected execution time of the command, repeating the original request, issuing a stop command, or taking other action, wherein the action may be presented verbally, non-verbally by pushing a selected button on a cell phone or remote control, or both, among other configurations. According to one embodiment of the invention, the system may detect the user action and may prompt the user to re-phrase the request to enable the system to modify words in the query. The user's actions may be analyzed in a statistical model to determine a frequency occurrence of misrecognitions for particular commands, with the results being used to update the corresponding personalized cognitive model 810.

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

[096] The event manager 100 may mediate interactions between other components of the invention. The event manager can provide a multi-threaded environment allowing the system to operate on multiple commands or questions from multiple user sessions without conflict and in an efficient manner, maintaining real-time response capabilities.

[097] Agents 106 may include a collection of grammars, criteria handlers, and algorithms that are accessible to respond to a set of requests and/or commands. Agents 106 may further contain packages of both generic and domain specific behavior for the system 98. Agents 106 may use nonvolatile storage for data, parameters, history information, and locally stored content provided in the system databases 102 or other local sources. One or more user profiles 110 may be provided

that include user specific data, parameters, and session and history information for determining the behavior of agents 106. One or more personality modules 108 may be provided in a data determining system that include personality characteristics for agents. The update manager 104 manages the automatic and manual loading and updating of agents 106 and their associated data from the Internet 146 or other network through the network interface 116.

[098] According to one embodiment of the invention, the speech-based interface for the system 90 may include one or more speech units 128. Speech units 128 may include one or more microphones, for example array microphone 134, to receive the utterances from the user. The speech received at the microphone 134 may be processed by filter 132 and passed to the speech coder 138 for encoding and compression. In one embodiment, a transceiver module 130 may transmit the coded speech to the main unit 98. Coded speech received from the main unit 98 is detected by the transceiver 130, then decoded and decompressed by the speech coder 138 and annunciated by the speaker 136.

[099] According to one embodiment of the invention, the non-speech-based interface for the system 90 may include one or more multi-modal devices 155 that may include mobile devices, stand alone or networked computers, personal digital assistances (PDAs), portable computer devices, or other multi-modal devices.

[0100] The speech units 128, multi-modal devices 155 and the main unit 98 may communicate over a communication link. The communication link may include a wired or wireless link. According to one embodiment, the communication link may comprise an RF link. The transceiver 130 on the speech unit may communicate coded speech data bi-directionally over the communication link with the transceiver 126 on the main unit 98. According to another embodiment, the RF link may use any standard local area wireless data protocols including the IEEE 802.11, Bluetooth or other standards. Alternatively, an infrared data link conforming to any suitable standard such as IrDA or other infrared standards can be used. In an alternative embodiment, wires may connect the speech unit 128 and the main unit 98, eliminating the need for one speech coder 138. Other wired or wireless analog or digital transmission techniques can be used.

[0101] According to one embodiment of the invention, coded speech received at the transceiver 126 on the main unit 98 is passed to the speech coder 122 for decoding

and decompression. The decoded speech may be processed by the speech recognition engine 120 using the context description grammar module 112, among other information. Any recognized information may be processed by the parser 118, which transforms information into complete algorithms and questions using data supplied by knowledge agents. Knowledge agents may be selected from the plurality of agents 106 using a grammar stack, wherein the knowledge agents provide information for generating a response to the question or command. The knowledge agents may then process the commands or questions by creating requests that are submitted to local databases 102 or submitted through the network interface 116 to external data sources over the Internet 146 or external other networks. Algorithms typically result in actions taken by the system 90 itself (i.e., pause or stop), or to a remote device or data source (i.e., download data or program, or control a remote device), through the network interface to the Internet or other data interface.

[0102] According to one embodiment of the invention, knowledge agents may return results of questions as responses to users. The responses may be created using the results of information requests, the system personality 108, the user preferences, other data in the user profile 110, and/or other information. Agents 106 may present these results using the speech unit 128. The agents 106 may create a response string, which is sent to the text to speech engine 124. The text to speech engine 124 may generate the utterances, which may be encoded and compressed by the speech coder 122. Once coded, the utterances are transmitted from the main unit 98 by the transceiver 126 to the transceiver 130 on the speech unit 128. The utterance is then decoded and decompressed by the speech coder 138 and output by the speaker 136. Alternatively, agents 106 may present the results using multi-modal devices 155.

[0103] The non-speech interface 114 may be part of, or separate from, the multi-modal devices 155 and may be used as a substitute or complement to the speech interface. For example, non-speech interface 114 may be used to present and interact with non-speech (e.g., graphical or tabular) information in a manner more easily understood by the user. According to one embodiment of the invention, multimodal support may be provided to maintain the context during both voice interaction and interaction through the non-speech interface 114. In one exemplary embodiment, a communication channel may be opened between multimodal devices 155 and the main user interface system 90 to allow multimodal devices 155 to input text

commands and questions. According to another embodiment of the invention, multimodal devices 155 may send a string of text or keywords for a command or question. The main interface system 90 may synchronize the context between multimodal device 155 and the speech units 128. In order to send a response to the corresponding device, the main user interface system 90 may track where the input came from so that the response may be sent to a TTS or multi-modal device 155.

[0104] According to one embodiment of the invention, non-speech interface 114 may show system, state and history information in a more concise manner than is possible through the speech interface. Non-speech interface 114 may be accessed to create or extend capabilities of agents 106. These operations may include scripting of agents, adding data to the agent or databases 102 used by the agent, adding links to information sources, among other operations.

[0105] According to another embodiment of the invention, system 90 may include different types of agents 106. In some embodiments of the invention, generic and domain specific behavior and information may be organized into domain agents 156. The system agent, on the other hand, may provide default functionality and basic services. The domain specific agents may provide complete, convenient and re-distributable packages or modules for each application area. In other words, the domain agents may include data that is needed to extend or modify the functionality of the system 90 in a current or new domain. Further, domain agents and their associated data can be updated remotely over a network as new behavior is added or new information becomes available. Domain agents may access a plurality of sources that may provide various services. Domain agents may use the services of other, typically more specialized, data managers and system agent. Agents are distributed and redistributed in a number of ways including on removable storage media, transfer over networks or attached to emails and other messages. The invention may provide license management capability allowing the sale of data managers by third parties to one or more users on a one time or subscription basis. In addition, users with particular expertise can create data managers, update existing data managers by adding new behaviors and information and making these data managers available to other users. A block diagram of the agent architecture according to an embodiment of the invention is shown in Figure 6.

[0106] Agents 106 may receive and return events to the event manager 100. Both system agents 150 and domain agents 156 may receive questions and commands from the parser 118. Based on keywords in the questions and commands and the structures of the questions and commands, the parser may invoke selected agents. Agents use the nonvolatile storage for data, parameters, history information and local content provided in the system databases 102.

[0107] According to one embodiment of the invention, when the system starts-up or boots-up, the agent manager 154 may load and initialize the system agent 150 and the one or more domain agents 156. Agent manager 154 includes knowledge of agents 106 and maps agents 106 to the agent library 158. At shutdown, the agent manager may unload the agents 106. The agent manager 154 also performs license management functions for the domain agents 156 and content in the databases 102.

[0108] The system agent 150 manages the criteria handlers 152 that handle specific parameters or values (criteria) used to determine context for questions and commands. According to one embodiment of the invention, criteria handlers 152 include parsing routines that are specialized to recognize particular parts of speech, such as times, locations, movie titles, and other parts of speech. The criteria handlers 152 may identify matching phrases and extract semantic attributes from the phrases. Both the system agent 150 and the domain agents 156 may use the criteria handlers 152. The various domain agents 156 may use the services of the system agent 150 and of other, typically more specialized, domain agents 156. The system agent 150 and the domain agents 156 may use the services of the agent library 158, which contains utilities for commonly used functions. According to one embodiment of the invention, the agent library may be a dynamic link library that implements one or more agents. The agent library may include utilities for text and string handling, network communications, database lookup and management, fuzzy and probabilistic evaluation, text to speech formats, and other utilities.

[0109] Domain agents 156 may be data-driven, scripted or created with compiled code. A base of generic agent may be used as the starting point for data-driven or scripted agents. Agents created with compiled code are typically built into dynamically linkable or loadable libraries. Developers of agents can add new functionality to the agent library 158 as required. Details of agent distribution and update, and agent creation or modification are discussed in sections below.

[0110] According to another embodiment of the invention, capabilities are provided to distribute and update system agents 150, domain agents 156, agent library 158 components, databases 102, and context description grammar 112 over wireless or wired networks 136, including IP networks and dial-up networks using the update manager 104. The network interface 116 may provide connections to one or more networks. The update manager 104 may also manage the downloading and installation of core system updates. The agent manager 154 may perform license management functions for the domain agents and the databases. The update manager 104 and agent manager 154 may perform these functions for all agents and database content including, agents and content available to all users or agents and/or content available to selected users. Examples of agent and database components added or updated on a periodic basis include:

- agents for new domains;
- additional domain knowledge for agents;
- new keywords for a domain, which can include names of politicians, athletes, entertainers, names of new movies or songs, etc. who have achieved recent prominence;
- links to a preferred set of information sources for the domains covered including links for, entertainment, news, sports, weather, etc.;
- updates to domain information based on, for example, changes to tax laws, company mergers, changing political boundaries;
- updates to content, including dictionaries, encyclopedias and almanacs; and
- other content and database components.

[0111] When a user requires or selects a new domain agent 156 or database element 102, the update manager 104 may connect to their source on the network 146 though the network interface 116, download and install the agent or data. To save system resources and to comply with any license conditions, the update manger 104 may uninstall agents that are no longer in use. In one embodiment of the invention, the update manager 104 may periodically query one or more sources of the licensed agents and database components to locate and download updates to agent executables, scripts or data as they become available. Alternatively, the agent sources may initiate

the downloading of agent updates of the registered or licensed agents to the update manager as they become available.

[0112] The agent manager 154 may provide license management clients that are capable of executing most any license terms and conditions. When a particular agent 106 and/or database element 102 is selected based on a submitted command, the agent manager 154 verifies that the use of the agent or data element is within the allowed terms and conditions, and if so, invokes the agent or allows access to the data element. License management schemes that can be implemented through the agent manager 154 include outright purchase, subscription for updates, one time or limited time use. Use of shared agents and data elements (such as those down-loaded from web sites maintained by groups of domain experts) may also be managed by the agent manager 154.

[0113] If questions or commands do not match an agent that is currently loaded on the system, the agent manager 154 may search the network 146 through the network interface 116 to find a source for a suitable agent. This process may be triggered, for example, when a query is made in a domain for which an agent is not available, when a new device is added to a mobile structure, or when the behavior of a mobile device is updated. Once located, the agent can be loaded under the control of the update manager 104, within the terms and conditions of the license agreement, as enforced by the agent manager.

[0114] New commands, keywords, information, or information sources can be added to any domain agent 156 by changing agent data or scripting. These configuration capabilities may allow users and content developers to extend and modify the behavior of existing domain agents 156 or to create new domain agents 156 from a generic agent without the need to create new compiled code. Thus, the modification of the domain agents 156 may range from minor data-driven updates by even the most casual users, such as specifying the spelling of words, to development of complex behavior using the scripting language as would typically be done by a domain expert. The user can create and manage modifications to domain agents 156 through speech interface commands or using non-speech interface 114. User-specific modifications of domain agents 156 are stored in conjunction with the user's profile 110 and accessed by the domain agent 156 at run-time.

[0115] The data used to configure data driven agents 156 may be structured in a manner to facilitate efficient evaluation and to help developers with organization. This data is used not only by the agent, but also in the speech recognition engine 120, the text to speech engine 124, and the parser 118. Examples of some major categories of data include:

1. Content packages may include questions or commands. Each command or question or group of commands or questions may include contexts used for creation of one or more requests. The domain agent 156 may pass a context description grammar expression to the parser 118 for evaluation of a context or question. An initial or default context is typically supplied for each command or question. The command or question includes a grammar for the management and evaluation of the context stack.
2. Parameters and other operating data on devices that are under control of the natural language speech interface. The domain agent 156 may use this data and parameters to determine how to execute a command, how to formulate the command string for the parser 118, determine if the command is feasible, and determine if the command can be executed within safety and operating limits.
3. Page lists or pointers to other local or network content sources. For each page or content source, there may be a pointer (e.g. URL, URI, or other pointer) to the page or source. Each page may have specific scraping information used to extract the data of interest. The scraping information may include, for example, matching patterns, HTML or other format parsing information.
4. A response list, determining the response of the domain agent 156 to a particular command or question given the context, the user profile and the information retrieved. Responses may include diagnostic error messages or requests for more information if the question or command cannot yet be resolved from the known information. Responses can be based on or be dependent on thresholds or probabilistic or fuzzy weights for the variables.
5. Substitution lists that include variable substitutions and transformations, often applied by the agents 150, 156 in the formatting of requests and results. For example, a stock domain specific agent 156 would

use a substitution list of company trading symbols, company names and commonly used abbreviations. Substitutions and transformations can be performed on commands and questions to create precise queries, which can be applied against one or more information sources or to results for creating more meaningful output to the user. Substitution lists also include information for optimally dealing with structured information, such as HTML formatted page parsing and evaluation.

6. Personalities used for responses. Personalities can be constructed by combining multiple traits in a weighted manner. Weights can be specified for each agent's domain area to create one or more specific personalities. Examples of personality traits include, sarcasm, humor, irritation, and sympathy, and other traits.

7. Public and user specific parameters for sources, substitutions, transformations, variables or criteria. The public parameter lists are part of the agent package 156. The user specific parameters are included in the user profile 110.

[0116] Commands and questions are interpreted, requests formulated, responses created and results presented can be based on the user's personal or user profile 110 values. Personal profiles may include information specific to the individual, their interests, their special use of terminology, the history of their interactions with the system, domains of interest, or other factors. The personal profile data can be used by the agents 106, the speech recognition engine 120, the text to speech engine 124, and the parser 118. Preferences can include, special (modified) commands, past behavior or history, questions, information sources, formats, reports, alerts or other preferences. User profile data can be manually entered by the user and/or can be learned by the system based on user behavior. User profile values can include: 1) spelling preferences; 2) date of birth for user, family and friends; 3) income level; 4) gender; 5) occupation; 6)location information such as, home address, neighborhood, and business address, paths traveled, locations visited; 7) vehicle type or types; 8) vehicle operator certifications, permits or special certificates; 9) history of commands and queries; 10) telecommunications and other service providers and services; 11) financial and investment information; 12) synonyms (i.e., a nick name for someone,

different terms for the same item); 13) special spelling; 14) keywords; 15) transformation or substitution variables; 16) domains of interest; or, 17) other values.

[0117] End users may use the data driven agent 156 extension and modification facilities and values stored in user profiles 110 to create special reports, packages of queries, alerts and output formats. A single alert or report can be configured to use multiple data sources and other variables (i.e., time, location, measured value) value to determine when alerts should be sent. For example, an alert can be generated by sampling a stock price every 15 minutes and sending an alert if the price drops below some value. In another example, an alert may be generated when a particular condition or combination of conditions occur for the vehicles. Alerts and reports can be directed to a local or remote output.

[0118] To create a report, the user may first specify a set of commands or questions. Next, the user can create or select a format for the report. Finally the user may name the report. A report can have variable parameters. For example, a user may create a company stock report and execute the report by stating its name and the company name, which gives the user selected information in a specified format for that company. In another example, a user can create a "morning" report, which presents selected multimedia information from different sources (news, sports, traffic, weather) in the order and formats desired. In yet another example, the user can create a report on the status of one or more vehicle systems. Alerts and reports can be created using only voice commands and responses, commands and responses through the graphical user interface 114, or a combination of the two. Reports can be run locally or remotely with respect to the vehicle. To create a report, alert, or other specialized behavior, the user performs a number of steps including: 1) specifying the command to run a report or alert; 2) specifying the question or questions, including keywords, used for a query; 3) setting the criteria for running the report such as on command or when a particular condition is met; 4) defining preferred information sources; 5)defining preferences for order of result evaluation by source, value, and other parameters; 6) specifying the presentation medium for a report or alert, such as an email, the text to speech engine, a message to a pager, or a text and graphics display; and, 7) specifying the preferred format for the report, such as information to be presented, order of information to be presented, preferred abbreviations or other variable substitutions.

[0119] Filtering and noise elimination may be important in facilitating the various functionalities of system 90 and may improve operation in noisy mobile environments. Recognition and parsing of the user's speech is implemented with good signal to noise ratio at the input to the speech recognition engine 120. To provide acceptable results, a set of acoustic models, an array microphone 134, a filter 132, or other components, may be employed. If a good signal to noise ratio cannot be achieved, a noise identification algorithm may be used and the appropriate acoustic model, for example, one that has been trained in conditions similar to the identified noise, may be selected. According to one embodiment of the invention, the microphone array, filters and speech coder 138 are physically separated from the main unit 98 into a speech unit 128, and connected using a wireless link. Since bandwidth on a wireless connection is at a premium, the speech coder dynamically adapts the digitization rate and compression of the captured speech.

[0120] Some embodiments of the invention may use one or more arrays of microphones 134 to provide better directional signal capture and noise elimination than can be achieved with a single microphone. The microphone array can be one-dimensional (a linear array) or two-dimensional (a circle, square, triangle or other suitable shape). The beam pattern of the array can be fixed or made adaptive though use of analog or digital phase shifting circuitry. The pattern of the active array is steered to point in the direction of the one or more users speaking. At the same time, nulls can be added to the pattern to notch out point or limited area noise sources. The use of the array microphone also helps reduce the cross talk between output from the text to speech engine 124 through the speaker 136 or from another user talking and detection of the user's speech.

[0121] The invention may use an analog or digital filter 132 between the array microphone or conventional microphone 134 and the speech coder 138. The pass band of the filter can be set to optimize the signal to noise ratio at the input to the speech recognition engine 120. In some embodiments, the filter is adaptive, using band shaping combined with notch filtering to reject narrow-band noise. One embodiment employs adaptive echo cancellation in the filter. The echo cancellation helps prevent cross talk between output from the text to speech engine and detection of the user's speech as well as suppression of environmentally caused echoes.

Algorithms comparing the background noise to the signal received from the users speech may be used to optimize the band-shaping parameters of the adaptive filter.

[0122] The speech received by the array microphone 134 and passed through the filter 132 may be sent to the speech digitizer or coder 138. The speech coder may use adaptive lossy audio compression to optimize bandwidth requirements for the transmission of the coded speech to the speech recognition engine 120 over a wireless link. The lossy coding is optimized to preserve only the components of the speech signal required for optimal recognition. Further, the lossy compression algorithms that may be used are designed to prevent even momentary gaps in the signal stream, which can cause errors in the speech recognition engine. The digitized speech may be buffered in the coder and the coder may adapt the output data rate to optimize the use of the available bandwidth. The use of the adaptive speech coder is particularly advantageous when a band-limited wireless link is used between the coder and the speech recognition engine.

[0123] The microphone can be complemented with an analog or digital (i.e., Voice over IP) speech interface. This interface allows a remote user to connect to the system and interact with it in the same manner possible if they were physically present.

[0124] In an alternative embodiment, the array microphone can be replaced by a set of physically distributed microphones or headsets worn by the users. The distributed microphones can be placed in different parts of a vehicle, different parts of a room or in different rooms of a building. The distributed microphones can create a three-dimensional array to improve signal to noise ratio. The headset can use a wireless or wired connection.

[0125] While the invention is intended to be able to accept most any natural language question or command, ambiguity may be a problem. To assist users formulate concise questions and commands, the system can support a voice query language. The language may be structured to allow a variety of queries and commands with minimal ambiguity. Thus, the voice query language helps users clearly specify the keywords or contexts of the question or command along with the parameters or criteria. The language can provide a grammar to clearly specify the keyword used to determine the context and present a set of one or criteria or parameters. A user asking a question or

stating a command in the voice query language may nearly always be guaranteed to receive a response.

[0126] The voice query language may be sensitive to the contents of the context stack, wherein a context defines a set of questions that can be activated or deactivated during a conversation. According to one embodiment, each agent may designate one context to be the root context that defines base algorithms that the agent implements. Thus, follow-on questions can be asked using an abbreviated grammar, since key words and criteria can be inherited from the context stack. For example, the user can simply ask about another keyword if the criteria of the question remain constant.

[0127] The system 90, may provide built in training capabilities to help users learn the best methods to formulate their questions and commands. The interactive training allows users to audibly or visibly see the machine interpretation of their queries and provides suggestions on how to better structure a query. Using the interactive training, users can quickly become comfortable with the voice query language and, at the same time, learn how to optimize the amount of information required with each step of a dialog.

[0128] The output of the speech coder 122 may be fed to the speech recognition engine 120. The speech recognition engine 120 recognizes words and phrases, using information in the context description grammar 112 and passes these to the parser 118 for interpretation. The speech recognition engine 120 may determine the user's identity by voice and name for each utterance. Recognized words and phrases may be tagged with this identity in all further processing. Thus, as multiple users engage in overlapping sessions, the tags added by the speech recognition engine 120 to each utterance allows other components of the system 90 to tie that utterance to the correct user and dialog. The user recognition capability may further be used as a security measure for applications, such as auctions or online shopping, where this is required. Voice characteristics of each user may be contained in the user profile 110.

[0129] According to one embodiment of the invention, users may start a dialog with the system 90 when they first address it. This can be done by speaking a generic word ("computer") or addressing a specific name ("Fred"), which may be generally tied to a system personality 108. Once the user starts the dialog, it may be recognized by the speech recognition engine 120, using unique characteristics of the user's speech. At

the end of a dialog or to interrupt a dialog, the user may utter a dismissal word ("good bye").

[0130] According to another embodiment of the invention, the system 90 may employ a speech recognition engine 124 that gains improved word recognition accuracy using data from context description grammar 112, user profiles 110, and the agents 106, among other components. At the same time, the fuzzy set possibilities or prior probabilities for the words in the context description grammar may be dynamically updated to maximize the probability of correct recognition at each stage of the dialog. The probabilities or possibilities may be dynamically updated based on a number of criteria including the application domain, the questions or commands, contexts, the user profile and preferences, user dialog history, the recognizer dictionary and phrase tables, and word spellings, among other criteria.

[0131] For uncommon words or new vocabulary words, the user may be given the option to spell the words. The spelling can be done by saying the names or the letters or using a phonetic alphabet. The phonetic alphabet can be a default one or one of the user's choosing.

[0132] Alternatively, when a user submits a word that is not recognized at all or is not correctly recognized by the speech recognition engine 120, then the user may be asked to spell the word. The speech recognition engine 120 determines this condition based on confidence level for the scoring process. The word may be looked up in a dictionary and the pronunciation for the word is added to either the dictionary, the agent 106, or the user's profile 110. The word pronunciation can then be associated with the domain, the question, the context and the user. Through this process, the speech recognition engine learns with time and improves in accuracy. To assist users in spelling words, an individualized phonetic alphabet can be used. Each user can modify the standard phonetic alphabets with words, which they can remember more easily.

[0133] Once the words and phrases have been recognized by the speech recognition engine 120, the tokens and user identification may be passed to the parser 118. The parser examines the tokens for the questions or commands, context and criteria. The parser may determine a context for an utterance by applying prior probabilities or fuzzy possibilities to keyword matching, user profile 110, dialog history, and context stack contents. The context of a question or command may determine the domain

and, thereby, the domain agent 156, if any, to be invoked. For example, a question with the keywords "temperature" implies a context value of weather for the question. Within a different dialog, the keyword "temperature" can imply a context for a measurement. The parser dynamically receives keyword and associated prior probability or fuzzy possibility updates from the system agent 150 or an already active domain agent 156. Based on these probabilities or possibilities the possible contexts are scored and the top one or few are used for further processing.

[0134] The parser 118 uses a scoring system to determine the most likely context or domain for a user's question and/or command. The score is determined from weighing a number of factors including, the user profile 110, the domain agent's data content and previous context. Based on this scoring, the system 90 invokes the correct agent. If the confidence level of the score is not high enough to ensure a reliable response, the system 90 may ask the user to verify whether the question and/or command is correctly understood.

[0135] In general, the question that is asked by the system 90 may be phrased to indicate the context of the question including all criteria or parameters. For example, the question can be in the form of: "Did I understand that you want such-and-such" If the user confirms that the question is correct the system proceeds to produce a response. Otherwise, the user can rephrase the original question, perhaps adding additional information to remove ambiguity, or the system can ask one or more questions to attempt to resolve the ambiguity.

[0136] Once the context for the question or command has been determined, the parser 118 can invoke the correct agent 156, 150. To formulate a question or command in the regular grammar used by agents, the parser will preferably determine required and optional values for the criteria or parameters. These criteria may have been explicitly supplied by the user or may need to be inferred. The parser may make use of the criteria handlers 152 supplied by the system agent. The criteria handlers can provide context sensitive procedures for extracting the criteria or parameters from the user's question or command. Some criteria may be determined by executing algorithms in the agent, while others may be determined by applying probabilistic or fuzzy reasoning to tables of possible values. Prior probabilities or fuzzy possibilities and associated values may be received from a number of sources including, for example, the history of the dialog, the user profile 110, and the agent. Based on user

responses, the prior probabilities or fuzzy possibilities may be updated as the system learns the desired behavior. For a weather context, examples of criteria include, location, date and time. Other criteria can include command criteria (i.e., yes/no, on/off, pause, stop), and spelling. Special criteria handlers are available from the system agent for processing lists, tables, barge-in commands, long strings of text and system commands.

[0137] The criteria handlers 152 can operate iteratively or recursively on the criteria extracted to eliminate ambiguity. This processing may help reduce the ambiguity in the user's question or command. For example, if the user has a place name (or other proper noun) in their utterance the parser 118 can use services of the domain agent 156 to look up tables in the databases 102 for place names or can attempt to determine which word is the proper noun from the syntax of the utterance. In another example, the user asks, "what about flight one hundred and twenty too?" The parser and domain agent use flight information in the database and network information along with context to determine the most plausible interpretation among; flight 100 and flight 20 also, flight 100 and flight 22, flight 122, etc.

[0138] Once the context and the criteria are determined, the parser 118 may form the question or command in a standard format or hierarchical data structure used for processing by the agents 150, 156. The parser 118 may fill in all required and some optional tokens for the grammar of the context. Often the tokens must be transformed to values and forms acceptable to the agents. The parser obtains the required transformations from the agents, dialog history or user profile 110. Examples of transformations or substitutions performed by the parser on tokens include: 1) substituting a stock symbol for a company name or abbreviation; 2) substituting a numerical value for a word or words; 3) adding a zip code to an address; and, 4) changing a place or other name to a commonly used standard abbreviation.

[0139] The agents 150, 156 may receive a command or question once the parser 118 has placed it in the required standard format. Based on the context, the parser can evoke the correct agent to process the question or command.

[0140] Commands can be directed to the system or to an external entity. System commands are generally directed to the system agent 150. Commands for external entities are generally processed by a domain agent 156, which includes the command context and behavior for the external entity.

[0141] Specific questions may be generally directed to one of the domain agents 156. The real-time selection of the correct agent allows the invention to dynamically switch contexts. Based on the question, command or context and the parameters or criteria, the domain agent may create one or more queries to one or more local or external information sources. Questions can be objective or subjective in nature. Results for objective questions can often be obtained by structured queries to one or more local or network information sources. Even for objective questions, the system may need to apply probabilistic or fuzzy set analysis to deal with cases of conflicting information or incomplete information. Information to answer subjective questions is generally obtained by one or more ad-hoc queries to local or network data sources, followed by probabilistic or fuzzy set evaluation of the one results to determine a best answer.

[0142] Once the domain agent 156 has formulated the one or more queries, they may be sent to local and/or network information sources. The queries may be performed in an asynchronous manner to account for the fact that sources respond at different speeds or may fail to respond at all. Duplicate queries can be sent to different information sources to ensure that at least one source responds with a useful result in a timely manner. Further, if multiple results are received in a timely manner, they can be scored by the system to determine which data is most reliable or appropriate. Examples of data sources accommodated include, HTTP data sources, sources with meta-data in various formats including XML, measurement data from sensors using various formats, device 32 setting parameters, entertainment audio, video and game files including MP3, databases using query languages and structured responses such as SQL, and other data sources.

[0143] The local information sources can be stored in one or more system databases 102 or can be on any local data storage such as a set of CDs or DVDs in a player or other local data storage. In other cases, local information can be obtained from vehicle system settings or measurement devices. Network information sources can be connected to, the control and device interfaces 30, the data interfaces 26, the Internet 42 or other network and accessed through a series of plug-ins or adaptors, known a pluggable sources, in the network interface 116. The pluggable sources are capable of executing the protocols and interpreting the data formats for the data sources of interest. The pluggable source provides information scrapping forms and procedures

for each source to the domain agents 156. If a new type of data source is to be used a new plug-in or adaptor can be added to the appropriate interface.

[0144] The domain agent 156 can evaluate the results of the one or more queries as they arrive. The domain agent may score the relevance of the results based on results already received, the context, the criteria, the history of the dialog, the user profile 110 and domain specific information using probabilistic or fuzzy scoring techniques. Part of the dialog history is maintained in a context stack. The weight of each context for the scoring may be based on the relevance of one context to another and the age of the contexts. Other scoring variables can be associated through the context stack. Contexts can also be exclusive, so that previous contexts have no weight in the scoring.

[0145] Based on the on-going scoring processes, the domain agent 156 may determine if a single best answer can be extracted. For most questions, the desired result may include a set of tokens that may be found to formulate an answer. Once a value has been found for each of these tokens, the results are ready for presentation to the user. For example, for a question on weather, the tokens can include the date, day of week, predicted high temperature, predicted low temperature, chance of precipitation, expected cloud cover, expected type of precipitation and other tokens. Results processed in this manner may include error messages. For subjective questions, this determination is made by determining a most likely answer or answers, extracted by matching of the results received. If no satisfactory answer can be inferred from the results of the query, the agent can do one of the following:

1. Ask the user for more information, typically through the speech interface, and based on the results obtained formulate new queries. This approach is applied when an irresolvable ambiguity arises in the formulation of a response.
2. Formulate new queries based on the results received from the first set of queries. This approach is typically applied in cases where the responses received do not include all the required information. Information sources to queries can be inferred from the results already obtained (i.e., links in an HTML document or measurements or settings from other devices 32) or from other sources. Using this approach one or more sets of queries and responses can be chained without the need for action by the user.

3. Wait for additional queries to return results.

[0146] In any case, the domain agent 156 may continue to make queries and evaluate results until a satisfactory response is constructed. In doing so, the agent can start several overlapping query paths or threads of inquiry, typically mediated by the event manager 100. This technique, combined with the use of asynchronous queries from multiple data sources provides the real-time response performance required for a natural interaction with the user.

[0147] The domain agent 156 may apply conditional scraping operations to each query response as it is received. The conditional scraping actions may depend on the context, the criteria, user profile 110, and domain agent coding and data. For each token to be extracted a scraping criteria 152 can be created using the services of the system agent 150. The scraping criteria may use format specific scraping methods including, tables, lists, text, and other methods. One or more scraping criteria can be applied to a page or results set. Once additional results are received, the domain agent can create new scraping criteria to apply to results already acquired. The conditional scarping process removes extraneous information, such as graphics, which need not be further processed or stored, improving system performance.

[0148] Specific commands are generally directed to one of the domain agents 156. The real-time selection of the correct agent allows the invention to dynamically switch contexts. Command oriented domain agents 156 evaluate the command and the state of vehicle systems, system capabilities, and measurements to determine if the command can be executed at all or if doing so will exceed operating or safety limits. If the command is ambiguous or cannot be executed for some other reason, the system may ask the user for more information or may suggest what the problem is and a likely approach to the solution. The domain agent may format the command for the specific device 32 and control and device interface 30. This formatting may involve variable substitution, inference of missing values and other formatting. Variable substitution and inference depends on the command context, the user profile 110, command history, state of vehicle systems and measured values, and other factors. A complex command can result in more atomic commands being sent to multiple devices, perhaps in a sequence. The sequence and nature of subsequent commands may depend on the previous commands, results of pervious commands, device settings and other measurements. As a command is executed, measurements are made

and results collected to determine if the execution was correct and the desired state or states were reached.

[0149] Once the domain agent 156 has created a satisfactory response to a question, or to a command, the agent may format that response for presentation. Typically, the domain agent can format the response into the markup format used by the text to speech engine 124. The domain agent may format the result presentation using available format templates and based on the context, the criteria, and the user profile 110. The domain agent may perform variable substitutions and transformations to produce a response best understood and most natural to the user. The domain agent may vary the order of presentation of tokens and the exact terminology used to create a more natural response to the user. The domain agent may also select the presentation personality 108 to be used.

[0150] For both command and query responses, the domain agent 156 may select the presentation template, determine order of presentation for tokens and determine variable substitutions and transformations using probabilistic or fuzzy set decision methods. The template used to form the presentation can be from the domain agent itself or from the user profile 110. The user profile can completely specify the presentation format or can be used to select and then modify an existing presentation format. Selection and formatting of presentation template can also depend on the presentation personality 108. At the same time, the characteristics of the personality used for the response are dynamically determined using probabilities or fuzzy possibilities derived from the context, the criteria, the domain agent itself and the user profile 110.

[0151] The domain agent 156 may apply a number of transformations to the tokens before presentation to the user. These variable substitutions and transformations may be derived from a number of sources including, domain information carried by the agent, the context, the token values, the criteria, the personality 108 to be used, and the user profile 110. Examples of variable substitutions and transformations include: 1) substitution of words for numbers; 2) substitution of names for acronyms or symbols (i.e., trading symbols); 3) use of formatting information derived from the information sources (i.e., HTML tags); 4) nature of the response including, text, long text, list, table; 5) possible missing information or errors; 6) units for measurement

(i.e., English or metric); and, 7) preferred terminology from the user profile or presentation personality 108.

[0152] The invention may provide special purpose presentation capabilities for long text strings, tables, lists and other large results sets. Domain agents 156 may use special formatting templates for such results. The system agent 150 can provide special criteria handlers 152 for presentation and user commands for large results sets. The presentation templates used by the domain agents for large results sets typically include methods for summarizing the results and then allowing the user to query the result in more detail. For example, initially only short summaries, such as headlines or key numbers, are presented. The user can then query the results set further. The criteria handlers provide users with the capability to browse large results sets. Commands provided by the criteria handlers for large results sets include, stop, pause, skip, rewind, start, and forward.

[0153] Some information, in formats such as video, pictures and graphics, may be best presented in a displayed format. The domain agents 156 apply suitable presentation templates in these cases and present the information through the non-speech interface 114. The system agent 150 provides special criteria handlers 152 for presentation and user commands for display presentation and control.

[0154] Although particular embodiments of the invention have been shown and described, it will be understood that it is not intended to limit the invention to the embodiments that are disclosed and it will be obvious to those skilled in the art that various changes and modifications may be made without departing from the spirit and scope of the invention. Thus, the invention is intended to cover alternatives, modifications, and equivalents, which may be included within the spirit and scope of the invention as defined by the claims.

We claim:

1. A mobile device for processing speech and non-speech communications, comprising:

an identification module that identifies a user that communicates with the mobile device;

a matching module that matches the identified user to one or more profiles;

a transcription module that transcribes the speech and the non-speech communications to create a speech-based textual message and a non-speech-based textual message, wherein the transcription module accesses the one or more profiles;

a merging module that merges the speech-based textual message and the non-speech-based textual message to generate a query;

a search module that searches the query for text combinations;

a comparison module that compares the text combinations to entries in a context description grammar;

a plurality of domain agents that are associated with the context description grammar;

a scoring module that provides relevance scores based on results from the comparison module;

a domain agent selector that selects domain agents based on results from the scoring module; and

a response generating module that communicates with the selected domain agents to obtain content that is gathered by the selected domain agents and that generates a response from the content, wherein the content is arranged in a selected order based on results from the scoring module.

2. The mobile device according to claim 1, wherein the response generating module generates an aggregate response that includes the content gathered by the selected domain agents.

3. The mobile device according to claim 1, wherein the identification module supports multiple users that access the mobile device during a same session.

4. The mobile device according to claim 3, wherein the session may include an overlapping session or an interleaved session.

5. The mobile device according to claim 3, wherein the identification module verifies an identify of the multiple users based on voiceprint matching or password matching.

6. The mobile device according to claim 4, wherein, during the interleaved session, the mobile device processes one or more queries in an order of receipt.

7. The mobile device according to claim 1, wherein, during the interleaved session, the mobile device processes one or more queries based on an order determined by a length of the queries.

8. The mobile device according to claim 1, wherein the mobile device includes (i) a personal digital assistant, (ii) a cellular telephone, (iii) a portable computer, (iv) a desktop computer, or any combination of (i) to (iv).

9. The mobile device according to claim 1, wherein the mobile device receives follow-up speech and non-speech communications and wherein the transcription module transcribes the follow-up speech and non-speech communications to create a follow-up speech-based textual message and a follow-up non-speech-based textual message.

10. The mobile device according to claim 9, wherein the merging module merges the follow-up speech-based textual message and the follow-up non-speech-based textual message to generate a follow-up query.

11. The mobile device according to claim 1, further comprising a personality module that facilitates formatting the response.

12. The mobile device according to claim 1, wherein the mobile device is configured to communicate with a context stack that includes one or more contexts that are selected based on the query.

13. The mobile device according to claim 12, wherein the scoring module determines the one or more contexts based on at least applying prior probabilities or fuzzy possibilities to (i) keyword matching, (ii) user profiles, (iii) a dialog history, or any combination of (i) to (iii).

14. The mobile device according to claim 1, wherein at least one of the domain agents creates and directs a request to a local information source or a network information source.

15. The mobile device according to claim 1, wherein at least one of the domain agents creates and directs a command to a remote or local device.

16. The mobile device according to claim 10, wherein the follow-up query is associated with a same context as the query.

17. The mobile device according to claim 1, wherein the mobile device supports Internet browsing and wherein the generated query is used to conduct the Internet browsing.

18. The mobile device according to claim 1, wherein the mobile device supports multimodal communications.

19. The mobile device, comprising:

an object that is accessible by the mobile device;

an input device that receives speech and non-speech annotations corresponding to the object;

a transcription module that transcribes the speech and the non-speech annotations to create a speech-based textual message and a non-speech-based textual message; and

an association module that associates the object with the speech-based textual message and the non-speech-based textual message.

20. The mobile device according to claim 19, further comprising:

a classification module that classifies the speech-based textual message and the non-speech-based textual message; and

a searching module that searches the speech-based textual message and the non-speech-based textual message based on a search request.

21. The mobile device according to claim 19, wherein the association module associates the object with the speech and non-speech annotations.

22. The mobile device according to claim 21, further comprising:

a classification module that classifies the speech and non-speech annotations; and

a searching module that searches the speech and non-speech annotations based on a search request.

23. The mobile device according to claim 19, wherein the object includes digital media.

24. The mobile device according to claim 23, wherein the object includes a digital photograph, a calendar entry, an email message, an instant message, a phonebook entry, a voicemail entry or a digital movie.

25. The mobile device according to claim 19, wherein the association module associates the object to the speech and non-speech annotations, the speech-based textual message, and the non-speech-based textual message.

26. The mobile device according to claim 25, wherein the object is retrievable using (i) a voice command, (i) a text command, or a combination of (i) and (ii).

27. The mobile device according to claim 19, wherein the object is stored (i) in the mobile device, (ii) at a server, or a combination of (i) and (ii).

28. The mobile device according to claim 19, wherein the object includes metadata.

29. The mobile device according to claim 28, wherein the metadata includes global positioning system information, environmental information, or geographic information.

30. The mobile device according to claim 28, further comprising a metadata searching module that enables searching of the metadata.

31. A system, comprising:

a plurality of mobile devices that support multimodal communications, wherein two or more mobile devices are associated with a user; and

a context module that is coupled to the plurality of mobile devices, wherein a same context is maintained at the two or more mobile devices that are associated with the user.

32. The system according to claim 31, wherein the plurality of mobile devices comprise:

an identification module that identifies a user that communicates with the mobile device;

a matching module that matches the identified user to one or more profiles;

a transcription module that transcribes the speech and the non-speech communications to create a speech-based textual message and a non-speech-based textual message, wherein the transcription module accesses the one or more profiles;

a merging module that merges the speech-based textual message and the non-speech-based textual message to generate a query;

a search module that searches the query for text combinations;

a comparison module that compares the text combinations to entries in a context description grammar;

a plurality of domain agents that are associated with the context description grammar;

a scoring module that provides relevance scores based on results from the comparison module;

a domain agent selector that selects domain agents based on results from the scoring module; and

a response generating module that communicates with the selected domain agents to obtain content that is gathered by the selected domain agents and that generates a response from the content, wherein the content is arranged in a selected order based on results from the scoring module.

33. The system according to claim 31, wherein the context module is configured to broadcast the same information to two or more mobile devices and to receive a response from one of the mobile devices.

34. The system according to claim 31, wherein the context module broadcasts a first type of information and receives a second type of information.

35. The system according to claim 34, wherein first type of information and the second type of information include speech information or non-speech information.

36. The system according to claim 35, wherein the first type of information includes non-speech information that is presented on a graphical user interface and the second type of information includes speech information that is applied to search a network.

37. The system according to claim 36, wherein the first type of information and the second type of information correspond to the same context.

38. A method of processing speech and non-speech communications in a multimodal environment, comprising:

receiving the speech and non-speech communications;

identifying a user that provides the speech and non-speech communications;

matching the identified user to one or more profiles;

transcribing the speech and non-speech communications to create a speech-based textual message and a non-speech-based textual message, wherein the transcribing includes accessing the one or more profiles;

merging the speech-based textual message and the non-speech-based textual message to generate a query;

searching the query for text combinations;

comparing the text combinations to entries in a context description grammar;

accessing a plurality of domain agents that are associated with the context description grammar;

generating a relevance score based on results from comparing the text combinations to entries in the context description grammar;

selecting one or more domain agents based on results from the relevance score;

obtaining content that is gathered by the selected domain agents; and

generating a response from the content, wherein the content is arranged in a selected order based on results from the relevance score.

39. The method according to claim 38, further comprising generating an aggregate response that includes the content that is gathered by the selected domain agents.

40. The method according to claim 38, further comprising:

receiving a follow-up speech and non-speech communications;

transcribing the follow-up speech and non-speech communications to create a follow-up speech-based textual message and a follow-up non-speech-based textual message; and

merging the follow-up speech-based textual message and the follow-up non-speech-based textual message to generate a follow-up query.

41. The method according to claim 38, further comprising a personality module that communicates the response to a user.

42. The method according to claim 38, wherein the system supports multiple users during a same session.

43. The method according to claim 42, wherein the session may include an overlapping session or an interleaved session.

44. The method according to claim 42, wherein the multiple users are identified based on voiceprint matching or password matching.

45. The method according to claim 43, further comprising processing one or more queries in an order of receipt.

46. The method according to claim 43, further comprising processing one or more queries based on an order determined by a length of the queries.

47. The method according to claim 38, further comprising generating a context stack that includes one or more contexts that are selected based on the query.

48. The method according to claim 47, wherein the one or more contexts are generated based on applying prior probabilities or fuzzy possibilities to (i) keyword matching, (ii) user profiles, (iii) a dialog history, or any combination of (i) to (iii).

49. The method according to claim 38, further comprising conducting Internet browsing with the generated query.

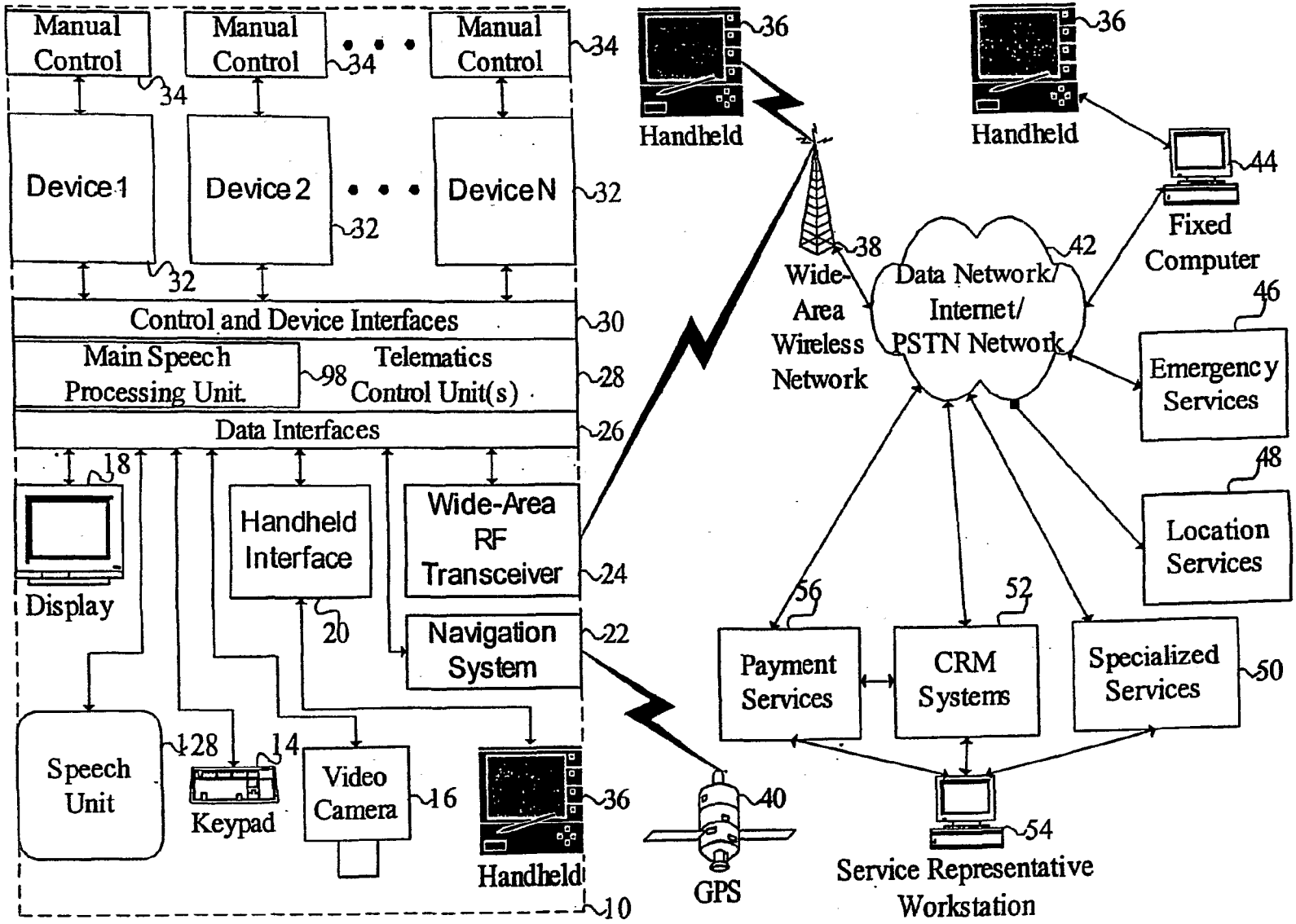

Manual Control
Manual Control
Manual Control
34
Device 1
Device 2
Device N
32
Control and Device Interfaces
30
Main Speech Processing Unit
98
Telematics Control Unit(s)
28
Data Interfaces
26
18
Display
Handheld Interface
20
Wide-Area RF Transceiver
24
Navigation System
22
Speech Unit
128
Keypad
14
Video Camera
16
Handheld
36
10
Handheld
36
Handheld
36
Wide-Area Wireless Network
38
Data Network/ Internet/ PSTN Network
42
Fixed Computer
44
Emergency Services
46
Location Services
48
Payment Services
56
CRM Systems
52
Specialized Services
50
GPS
40
Service Representative Workstation
54

Figure 1

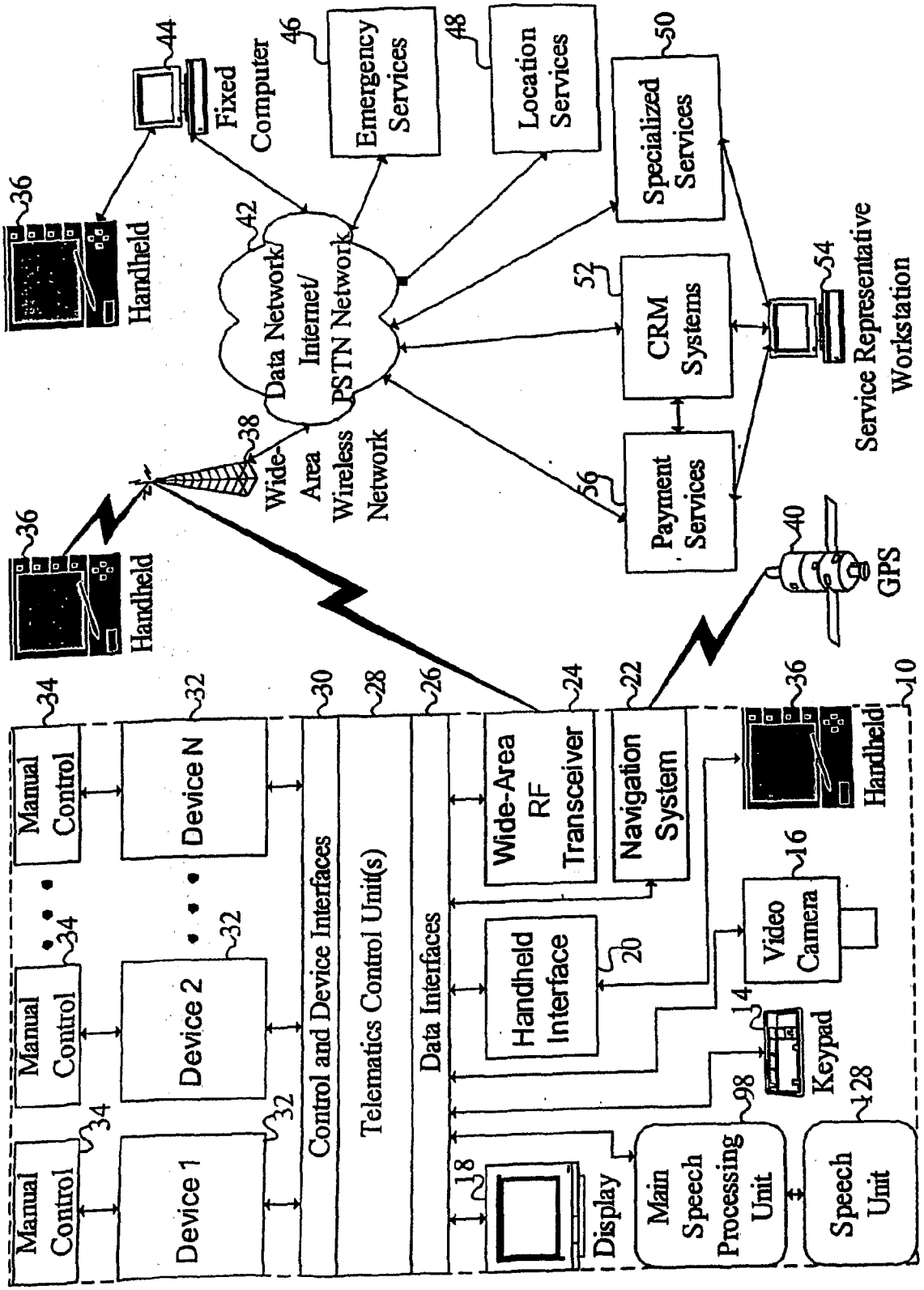

Manual Control
Device 1
Manual Control
Device 2
Manual Control
Device N
Control and Device Interfaces
Telematics Control Unit(s)
Data Interfaces
Handheld Interface
Wide-Area RF Transceiver
Navigation System
Video Camera
Keypad
Display
Main Speech Processing Unit
Speech Unit
Handheld
GPS
Payment Services
CRM Systems
Service Representative Workstation
Specialized Services
Location Services
Emergency Services
Fixed Computer
Data Network/ Internet/ PSTN Network
Wide-Area Wireless Network
Handheld
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
52
54
56
98

Figure 2

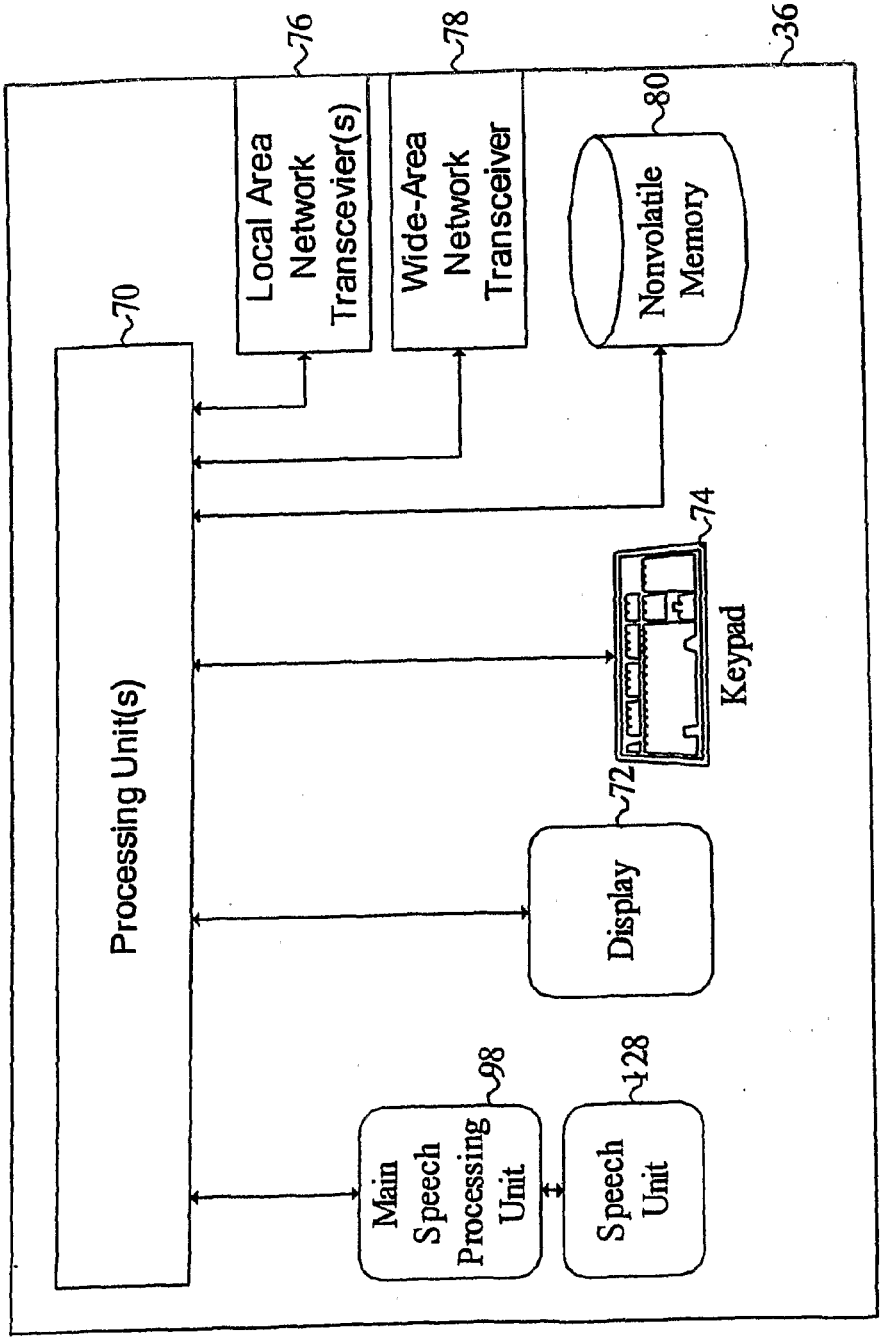

Processing Unit(s)
70
Local Area Network Transcevier(s)
76
Wide-Area Network Transceiver
78
Nonvolatile Memory
80
36
74
Keypad
72
Display
Main Speech Processing Unit
98
Speech Unit
128

Figure 3.

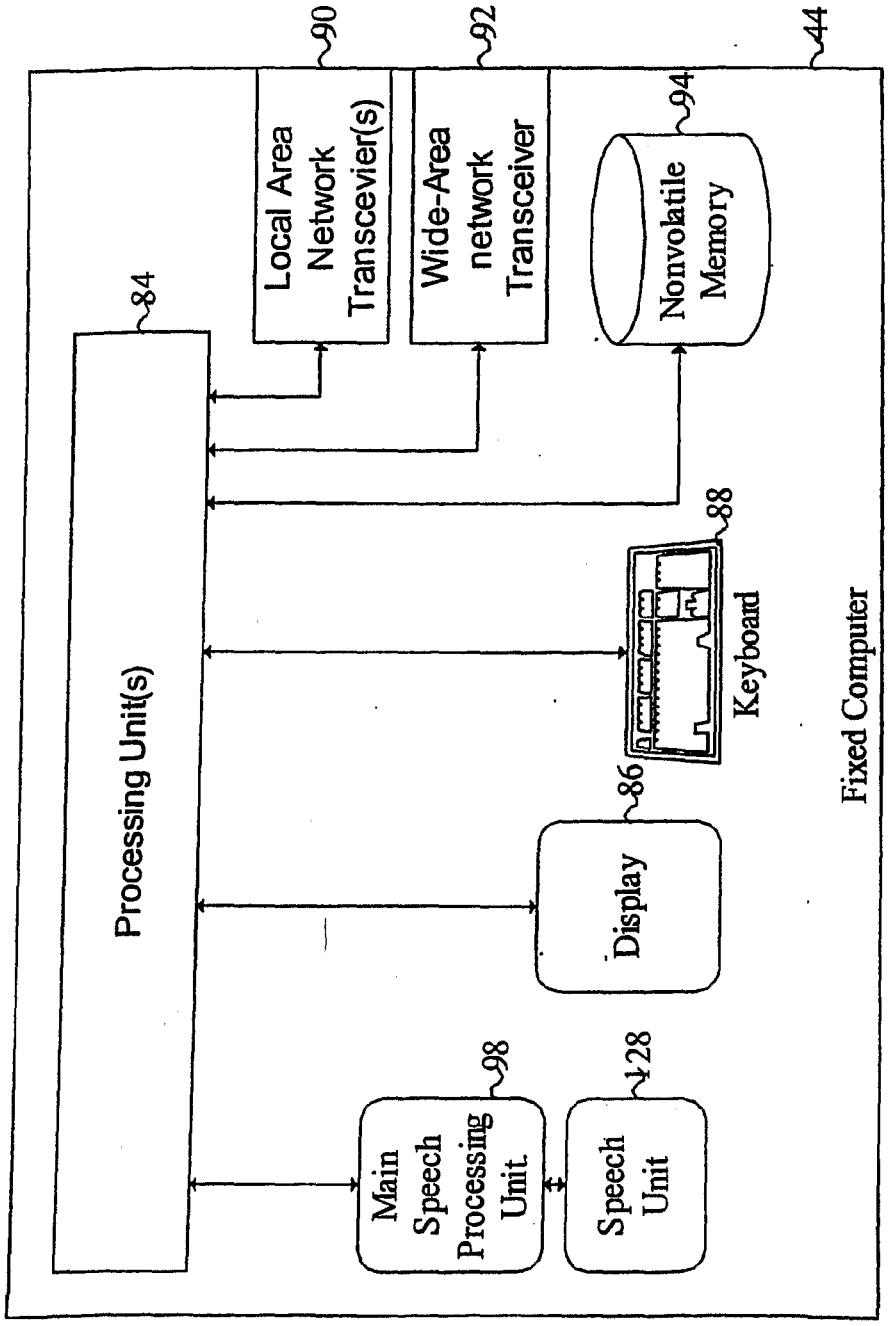

Fixed Computer
44
Local Area Network Transcevier(s)
90
Wide-Area network Transceiver
92
Nonvolatile Memory
94
Processing Unit(s)
84
Keyboard
88
Display
86
Main Speech Processing Unit
98
Speech Unit
128

Figure 4

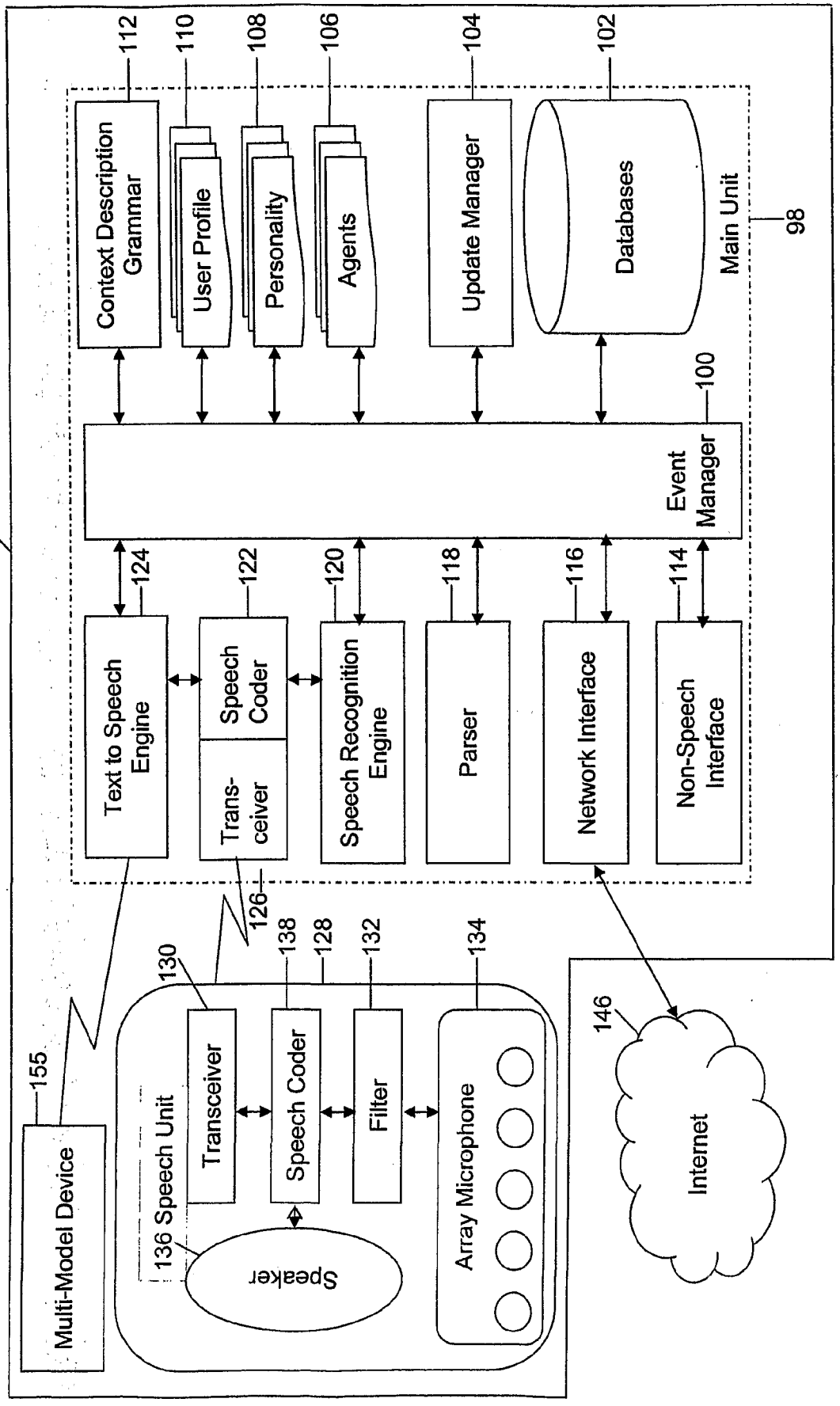

90
Context Description Grammar
112
User Profile
110
Personality
108
Agents
106
Update Manager
104
Databases
102
Main Unit
98
Event Manager
100
Text to Speech Engine
124
Speech Coder
Trans-ceiver
122
126
Speech Recognition Engine
120
Parser
118
Network Interface
116
Non-Speech Interface
114
Multi-Model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

FIG. 5

FIG. 6

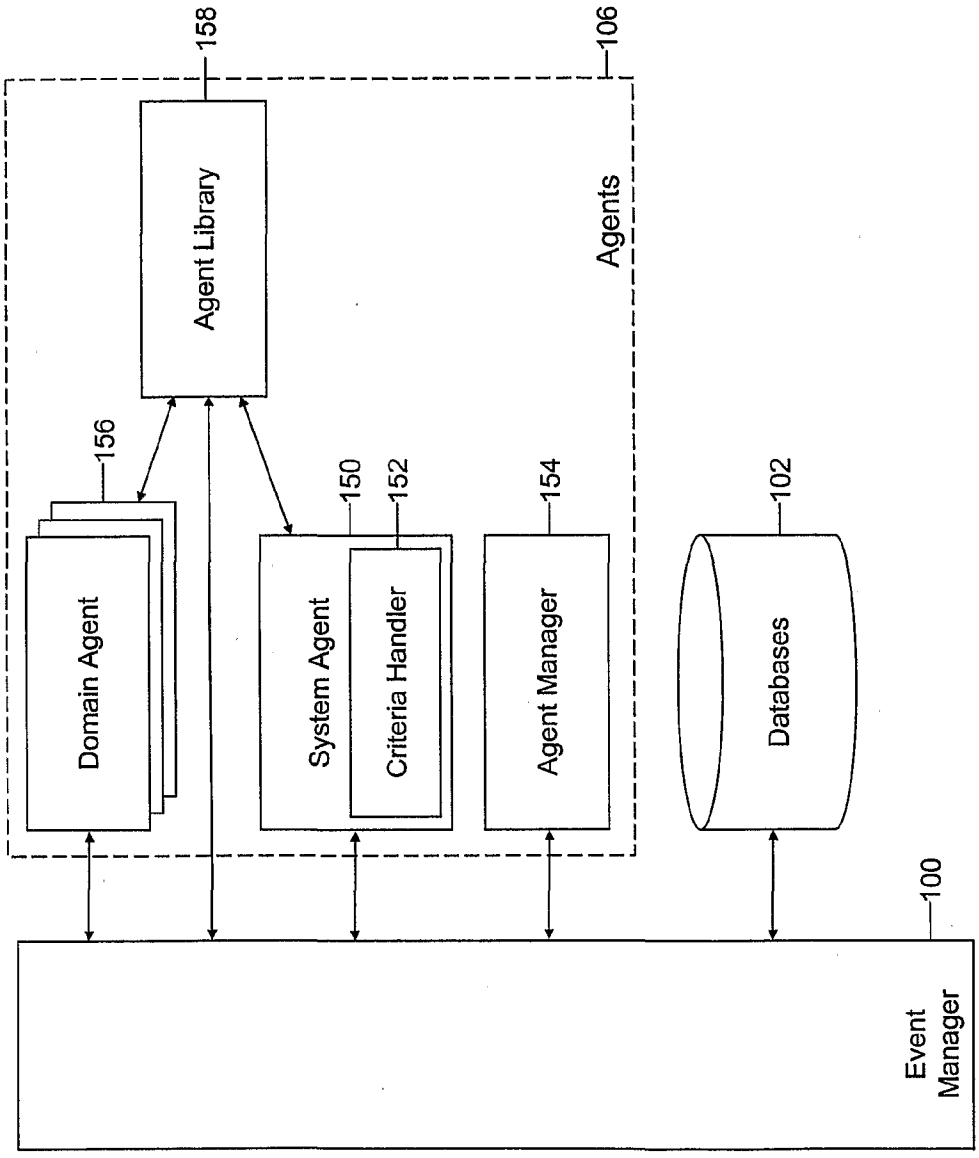
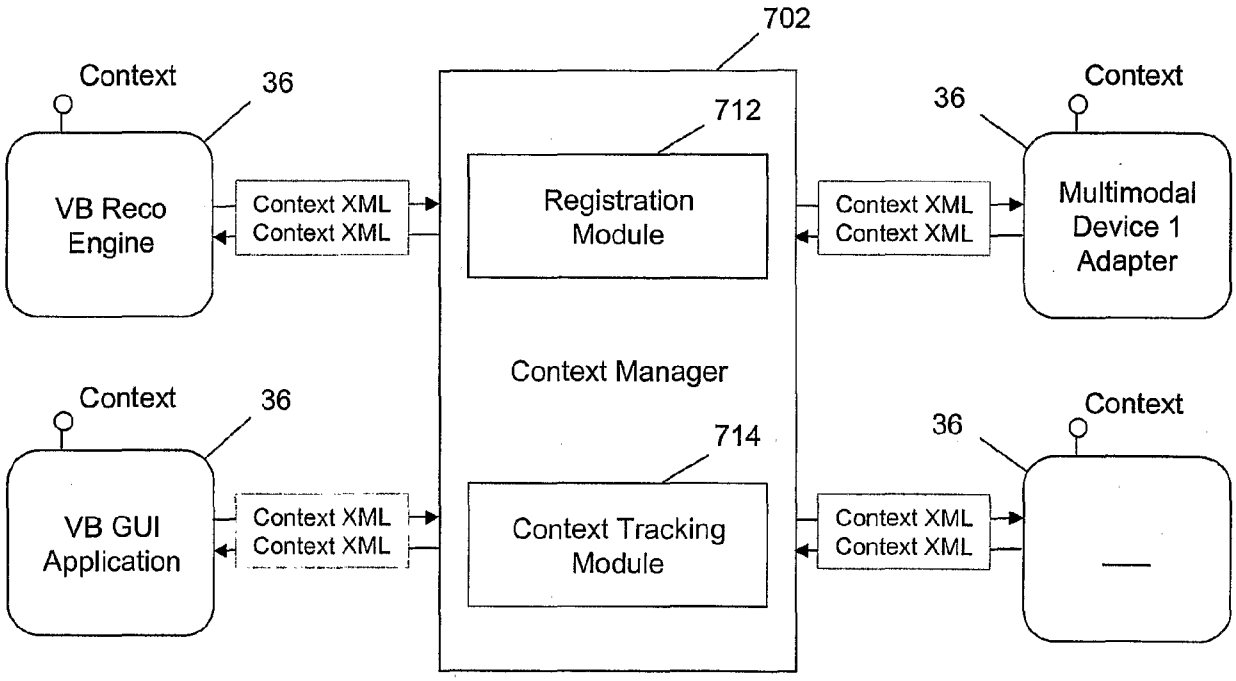

158
106
Agent Library
Agents
156
150
152
154
102
Domain Agent
System Agent
Criteria Handler
Agent Manager
Databases
100
Event Manager

FIG. 7


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

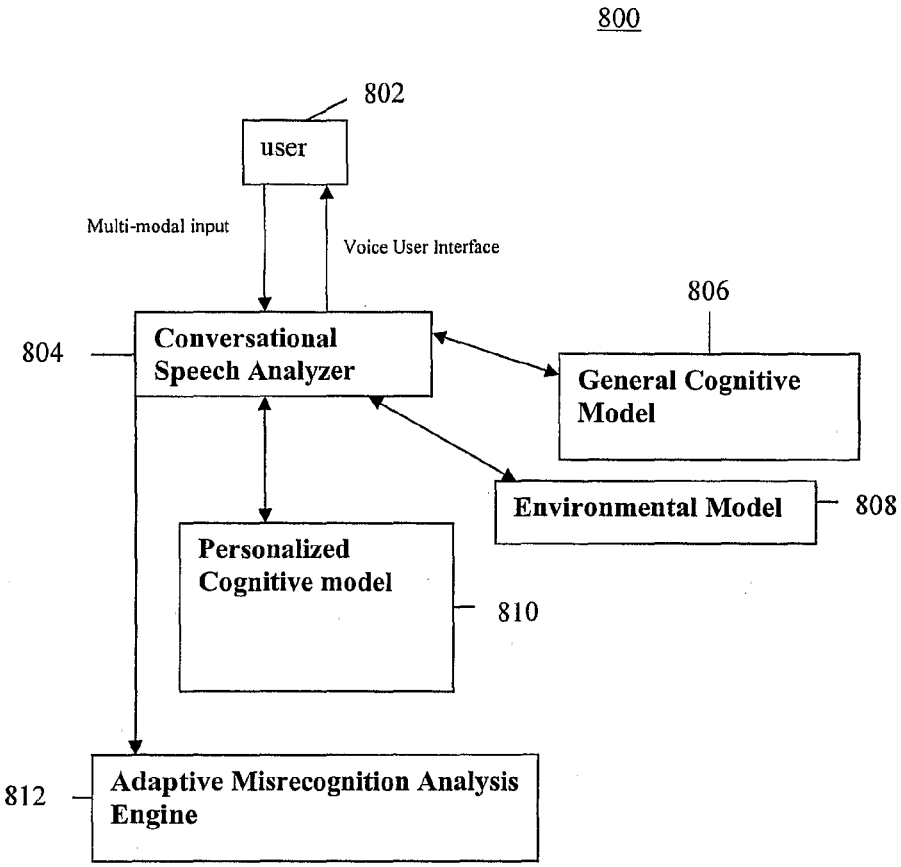

800
802
user
Multi-modal input
Voice User Interface
806
804
Conversational Speech Analyzer
General Cognitive Model
Environmental Model
808
Personalized Cognitive model
810
812
Adaptive Misrecognition Analysis Engine

FIGURE 8

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 March 2007 (08.03.2007)

PCT

(10) International Publication Number
WO 2007/027989 A2

(51) International Patent Classification:
G10L 15/18 (2006.01)

(21) International Application Number:
PCT/US2006/034184

(22) International Filing Date: 31 August 2006 (31.08.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/712,412 31 August 2005 (31.08.2005) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [US/US]; 10505 NE 38th Place, Building 9, Kirkland, WA 98033 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **DI CRISTO, Philippe** [FR/US]; 177 107th Avenue, N.E., #1007, Bellevue, WA 98004 (US). **KE, Min** [US/US]; 3221 201st Place, S.E., Bothell, WA 98012 (US). **KENNEWICK, Robert, A.** [US/US]; 2204 S. Eddy St., Seattle, WA 98108 (US). **TJALVE, Michael** [DK/US]; 14904 SE 47th Court, Bellevue, WA 98006 (US).

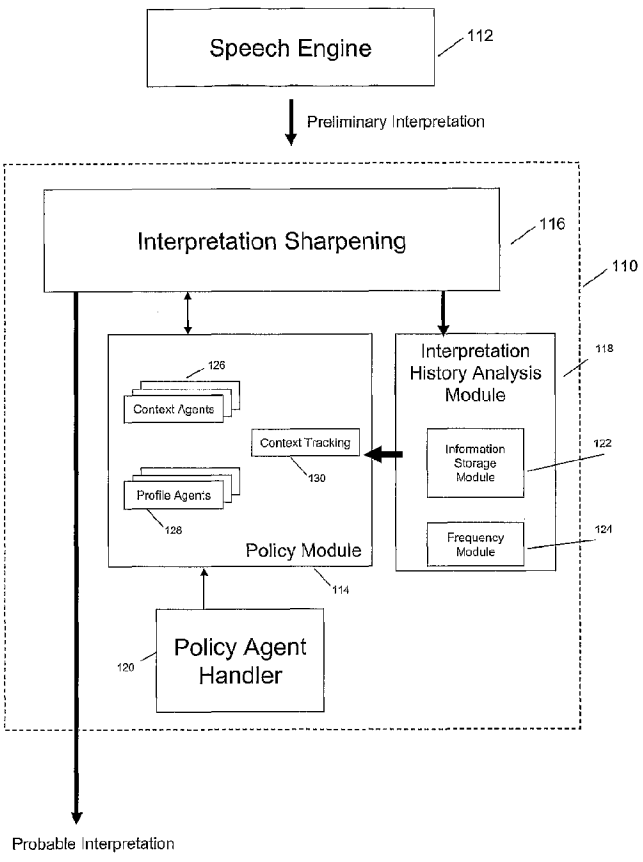
(74) Agents: **GATTO, James, G.** et al.; PILLSBURY WINTHROP SHAW PITTMAN LLP, P.O. Box 10500, Mclean, VA 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

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

[Continued on next page]

(54) Title: DYNAMIC SPEECH SHARPENING



(57) Abstract: An enhanced system for speech interpretation is provided. The system may include receiving a user verbalization and generating one or more preliminary interpretations of the verbalization by identifying one or more phonemes in the verbalization. An acoustic grammar may be used to map the phonemes to syllables or words, and the acoustic grammar may include one or more linking elements to reduce a search space associated with the grammar. The preliminary interpretations may be subject to various post-processing techniques to sharpen accuracy of the preliminary interpretation. A heuristic model may assign weights to various parameters based on a context, a user profile, or other domain knowledge. A probable interpretation may be identified based on a confidence score for each of a set of candidate interpretations generated by the heuristic model. The model may be augmented or updated based on various information associated with the interpretation of the verbalization.

WO 2007/027989 A2

FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

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

DYNAMIC SPEECH SHARPENING

CROSS REFERENCE TO RELATED APPLICATIONS

[001] This application claims the benefit of provisional U.S. Patent Application Serial No. 60/712,412, entitled "Dynamic Speech Sharpening," filed August 31, 2005, and which is hereby incorporated by reference in its entirety.

[002] This application is related to the following co-pending applications, which are hereby incorporated by reference in their entirety: "Systems and Methods for Responding to Natural Language Speech Utterance," Ser. No. 10/452,147, filed June 3, 2003, and published as US Patent Application Publication No. US 2004/0044516; "Mobile Systems and Methods for Responding to Natural Language Speech Utterance," Ser. No. 10/618,633, filed July 15, 2003, and published as US Patent Application Publication No. US 2004/0193420; "Systems and Methods for Responding to Natural Language Speech Utterance," Ser. No. 11/197,504, filed August 5, 2005; "System and Method of Supporting Adaptive Misrecognition in Conversational Speech," Ser. No. 11/200,164, filed August 10, 2005; and "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," Ser. No. 11/212,693, filed August 29, 2005.

FIELD OF THE INVENTION

[003] The invention is related generally to automated speech interpretation, and in particular, to enhancing the accuracy and performance of speech interpretation engines.

BACKGROUND OF THE INVENTION

[004] The field of automated speech interpretation is in increasingly higher demand. One use of automated speech interpretation is to provide voice requests to electronic devices. This may enable a user to simply speak to an electronic device rather than manually inputting requests, or other information, through pressing buttons, uploading information, or by other request input methods. Controlling various electronic devices through speech may enable the user to use the electronic devices more efficiently.

[005] However, existing technology in the field of automated speech interpretation, such as standard speech engines, automatic speech recognition (ASR), and other systems for interpreting speech, are unable to process a speech signal in an efficient manner, often constructing large grammars that include a large number of items,

nodes, and transitions, which is a concern particularly for large-list recognition for embedded applications. If the grammar for an embedded application grows too much, it may not fit within the constrained space of an embedded application. With limited CPU power, response time and performance is easily affected due to the significant time needed to compile and load the grammar. Response time is further degraded because the speech engine has to parse through a large number of transition states to come up with a recognition result. Even when the speech engine is able recognize a word, the results are often unreliable because large grammars introduce greater risk of confusion between items as the size of the grammar increases. Existing techniques focus on reducing the size of a grammar tree by removing command variants or criteria items, but this approach strips functionality from the application.

[006] In addition to the performance problems associated with speech recognition engines that employ large word grammars, existing speech processing engines are unable to interpret natural human speech with a suitable accuracy to sufficiently control some electronic devices. In particular, speech interpretation engines still have substantial problems with accuracy and interpreting words that are not defined in a predetermined vocabulary or grammar context. Poor quality microphones, extraneous noises, unclear or grammatically incorrect speech by the user, or an accent of the user may also cause shortcomings in accuracy, such as when a particular sound cannot be mapped to a word in the grammar.

[007] In light of these and other problems, there is a need for enhanced automated speech interpretation that may interpret natural human speech with an augmented accuracy.

SUMMARY OF THE INVENTION

[008] According to one aspect of the invention, a system for enhancing automated speech interpretation is provided. The system may include a set of techniques for use in a speech-to-text engine to enhance accuracy and performance, for example, by reducing the search space of the speech engine. The problems with large-list recognition for embedded applications may also be improved by using phonetic dictation, which may recognize a phoneme string by disregarding the notion of words. The system may also use one or more post-processing techniques to sharpen an output of a preliminary speech interpretation made by a speech engine. The system may be modeled at least partially after one or more speech pattern

recognition techniques used by humans, such as interpreting speech using words, word sequences, word combinations, word positions, context, phonetic similarities between two or more words, parts of speech, or other techniques.

[009] In one implementation of the invention, the system may receive a verbalization made by a user, where a speech engine may receive the verbalization. The speech engine may output information relating to a plurality of preliminary interpretations of the verbalization, where the plurality of preliminary interpretations represent a set of best guesses at the user verbalization. According to one aspect of the invention, the performance of the speech engine may be improved by using phoneme recognition. Phoneme recognition may disregard the notion of words, instead interpreting a verbalization as a series of phonemes, which may provide out-of-vocabulary (OOV) capabilities, such as when a user misspeaks or an electronic capture devices drops part of a speech signal, or for large-list applications, such as city and street names or song titles, for example. Phoneme recognition may be based on any suitable acoustic grammar that maps a speech signal into a phonemic representation. For example, the English language may be broken down into a detailed grammar of the phonotactic rules of the English language. Portions of a word may be represented by a syllable, which may be further broken down into core components of an onset, a nucleus, and a coda, which may be further broken down into sub-categories. Various different acoustic grammars may be formed as trees with various branches representing many different syllables forming a speech signal.

[010] According to another aspect of the invention, the performance of the speech engine and the phonemic recognition may be improved by pruning the search space used by the speech engine using a common phonetic marker. In one implementation, the acoustic grammar may be represented entirely by a loop of phonemes. In another implementation, the speech engine may reduce the search space by reducing the number of transitions in a grammar tree, thereby speeding up the process of compiling, loading, and executing the speech engine. For example, the phoneme loop may include a linking element between transitions. This may reduce the number of grammar transitions, such that grammar paths merge after a first transition and diverge after the linking element. In one implementation of the invention, a common acoustic element that is part of a speech signal may be used as the linking element. In one implementation of the invention, the acoustic element may be one that is very likely to be triggered even if it is unpronounced. For

example, a *schwa* in the English language may be used as the linking element because *schwa* represents an unstressed, central vowel that is likely to be spoken even if unintended. Those skilled in the art will appreciate that acoustic models for different languages may use other frequently elided phonemes as linking elements to reduce the search space used by the speech engine.

[011] The speech engine may generate a plurality of preliminary interpretations representing a set of best guesses at the user verbalization. The preliminary interpretations may be stored in a matrix, array, or another form, and may be provided to an interpretation sharpening module to determine a probable interpretation of a verbalization made by a user by applying heuristic policies against the preliminary interpretation to identify dominant words and/or phrases. According to various aspects of the invention, the interpretation sharpening module may include a policy module that may manage and/or provide one or more policies that enable the sharpening module to generate a plurality of probable interpretations of the verbalization made by the user. For example, according to one aspect of the invention, the plurality of preliminary interpretations may be applied against one or more policies to generate a set of hypotheses as to a candidate recognition. Each hypothesis may be reanalyzed to generate an interpretation score that may relate to a likelihood of the probable interpretation being a correct interpretation of the verbalization, and the preliminary interpretation corresponding to the highest (or lowest) interpretation score may then be designated as a probable interpretation of the verbalization. The designated probable interpretation may be stored and used for augmenting the policies to improve accuracy.

[012] According to one aspect of the invention, the policy module may include one or more agents that represent domains of knowledge. The agents may compete using a weighted model to revise a preliminary interpretation by determining context and intent. Relevant substitution of suspect words and phrases may be based on phonetic similarities or domain appropriateness. A domain agent may include one or more domain parameters for determining a probable interpretation from a preliminary interpretation. For example, domain parameters may include a policy vocabulary, a word position in the verbalization, a word combination, a sentence structure, or other parameters. A domain agent may include a parameter weighting scheme that may weight individual parameters according to one or more weighting factors, such as, a frequency of use, a difficulty to understand, or other factors.

[013] According to one aspect of the invention, the domain agents may revise a preliminary interpretation into a probable interpretation using phonetic fuzzy matching (PFM). In one implementation of the invention, the speech engine may output a phoneme stream that is applied against a model of phoneme feature similarities, drawn from domain agents, to identify a closest phonetic match using a multi-pass method. Domain agents may be loaded and prioritized into an M-Tree, which accounts for the possibility of the speech engine dropping or adding phonemes. An M-Tree may be an index structure that resolves similarity queries between phonemes using a closest-distance metric based on relative weightings of phoneme misrecognition, phoneme addition, and phoneme deletion. The M-Tree may be updated using an adaptive misrecognition model. For example, information about a verbalization and its components, as well as a probability that the probable interpretation was correct, may be stored and used for adapting the policy module for the user.

[014] In one implementation of the invention, the domain agents in the policy module may include one or more profile agents that may manage and/or provide one or more profile policies for revising a preliminary interpretation of a phoneme stream. For example, a profile agent may correspond to a user and may include one or more profile parameters tailored to the user. The profile agent may be used as a base policy to interpret any verbalizations made by the user. In other implementations, a profile agent may correspond to a particular language, a regional accent, or other profiles for interpreting a user verbalization. The profile agents may be augmented to enable the system to provide more accurate interpretations of verbalizations made by the user. The augmentation may include a user augmentation, such as providing additional vocabulary (e.g., names in an address book), one or more personalized pronunciations or other pronunciation information, or other user provided augmentations. The augmentation may also include a non-user provided augmentation, such as updates generated by a third party (e.g., a commercial administration and/or maintenance entity), or other non-user provided augmentations. The augmentation may be automated, such as adjusting a profile parameter-weighting scheme through an adaptive misrecognition model, as discussed above.

[015] In another implementation of the invention, the domain agents in the policy module may include one or more context agents that may manage and/or provide

one or more context policies for revising a preliminary interpretation of a phoneme stream. For example, a context agent may correspond to a context, such as song titles, city and street names, movie titles, finance, or other contexts. A context agent may include one or more context parameters that may be tailored to a verbalization context. The context policy may enhance an ability of the system related to interpreting verbalizations made by the user in the verbalization context corresponding to the context agent. The context agents may be augmented to enable the system to provide more accurate interpretations of verbalizations made in a verbalization context corresponding to the context agent. The augmentation may include a user provided augmentation, a non-user provided augmentation, an automated augmentation, or other augmentations. The augmentation may be automated, such as adjusting a profile parameter-weighting scheme through an adaptive misrecognition model, as discussed above.

[016] According to various implementations of the invention, the policy module may determine which profile agents and/or which context agents to use through a set of heuristics provided in a context-tracking module. In one implementation, the context-tracking module may use phonetic fuzzy matching to track a series of verbalizations by the user to identify a verbalization context. The context-tracking module may utilize one or more M-Trees to track the series of verbalizations and determine a closest phonetic match. The context-tracking module may track one or more past verbalization contexts for the series of verbalizations, one or more current verbalization contexts for the series of verbalizations, and/or make predictions regarding one or more future verbalization contexts for the series of verbalizations. The policy module may utilize information about the verbalization context of the series of verbalizations generated by the context tracking module to manage and/or provide one or more profile and/or context agents.

[017] According to one aspect of the invention, the system may include an interpretation history analysis module that may enable the system to augment one or more domain agents based on an analysis of past interpretations related to previously interpreted verbalizations. The augmentations enabled by the interpretation history analysis module may include a user augmentation, a third-party augmentation, an automated augmentation, or other augmentations. The interpretation history analysis module may include an information storage module that may store interpretation information related to past verbalizations, such as one

or more preliminary interpretations associated with a past verbalization, one or more interpretation scores associated with a past verbalization, one or more probable interpretations associated with a past verbalization, whether or not a past verbalization was interpreted correctly, or other information. A frequency module may be included in the interpretation history module, and the frequency module may use some or all of the information stored in the information storage module to generate one or more frequencies related to one or more past verbalizations. For example, the frequency module may calculate a frequency of word usage, word combinations, phonetic homonyms, interpretation errors for a particular verbalization, or other frequencies.

[018] The Information generated and/or stored by the interpretation history analysis module may be used to augment the profile and/or context agents in order to enhance the accuracy of subsequent interpretations. In some implementations, an adaptive misrecognition model may use one or more generated frequencies to augment one or more agents. For example, one or more parameters and/or weighting schemes of an agent or phonetic model may be augmented based on a frequency generated by the interpretation history analysis module. Other augmentations using information stored and/or generated by the interpretation history analysis module may be made, and the system may include a policy agent handler that may augment, update, remove, and/or provide one or more domain agents to the system. A domain agent may comprise a profile or context agent, and the policy agent handler may be controlled, directly or indirectly by a third party (e.g. a commercial entity, etc.). The policy agent handler may augment, update, remove, and/or provide domain agents to the system as part of a commercial agreement, such as a licensing agreement, a subscription agreement, a maintenance agreement, or other agreements.

[019] Other objects and advantages of the invention will be apparent to those skilled in the art based on the following detailed description and accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[020] Fig. 1 illustrates an exemplary system for enhancing automated speech interpretation according to one implementation of the invention.

[021] Fig. 2 illustrates an exemplary grammar tree for enhancing the performance of a speech engine according to one implementation of the invention.

[022] Fig. 3 illustrates an exemplary flow chart of a method for enhancing automated speech interpretation according to one implementation of the invention.

DETAILED DESCRIPTION

[023] Fig. 1 illustrates an exemplary system for enhancing automated speech interpretation according to one implementation of the invention. A speech-to-text processing engine 112 may receive a user verbalization, and speech engine 112 may generate one or more preliminary interpretations of the user verbalization. The preliminary interpretations may represent a set of best guesses as to the user verbalization arranged in any predetermined form or data structure, such as an array, a matrix, or other forms. In one implementation of the invention, speech engine 112 may generate the preliminary interpretations by performing phonetic dictation to recognize a stream of phonemes, instead of a stream of words. Phonemic recognition provides several benefits, particularly in the embedded space, such as offering out-of-vocabulary (OOV) capabilities, improving processing performance by reducing the size of a grammar, and eliminating the need to train Statistic Language Models (SLMs). Those skilled in the art will recognize other advantages of phonemic recognition.

[024] Speech engine 112 may apply the phoneme stream against one or more acoustic grammars that reliably map a speech signal to a phonemic representation in order to generate the plurality of preliminary interpretations. Characteristics of a speech signal may be mapped to a phonemic representation to construct a suitable acoustic grammar, and various acoustic grammars may be included in speech engine 112 to generate one or more preliminary interpretations according to the various acoustic grammars. For example, the English language may be mapped into a detailed acoustic grammar representing the phonotactic rules of English, where words may be divided into syllables, which may further be divided into core components of an onset, a nucleus, and a coda, which may be further broken down into one or more sub-categories.

[025] Once the phonotactic rules of a speech signal have been identified, a detailed acoustic grammar tree may be constructed that accounts for the nuances of the speech signal. The acoustic grammar may include a loop of phonemes, or the phoneme loop may include a linking element to reduce a size of a search space associated with the grammar. Using the English language as an example, the grammar tree may include various branches representing English language

syllables. The speech engine may traverse one or more grammar trees to generate one or more preliminary interpretations of a phoneme stream as a series of syllables that map to a word or phrase. By using phonemic recognition rather than word recognition, the size of the grammar can be reduced, which reduces the amount of time required to compile, load, and execute speech interpretation. Moreover, because the grammar maintains a high level of phonotactic constraints and therefore a large number of syllables, speech engine 112 may be very precise in generating phonemic representations of human verbalizations.

[026] An acoustic grammar used by speech engine 112 may be further optimized to reduce compile time, load time, and execution time by reducing the size of a search space associated with the acoustic grammar. Referring now to Fig. 2, a traditional grammar tree 120 is compared to an exemplary grammar tree according to one aspect of the invention to demonstrate the performance enhancements of speech engine 112. In traditional speech processing engines, nodes in a grammar tree 210 tend to represent words, or large-list applications may be supported provided through a grammar tree 210 where the nodes represent items in the large-list. This requires the speech engine to parse through a large number of transition states to come up with a recognition result, which degrades response time. An example of this is seen in the following grammar structure:

[027] "<street name> <city name>" → e.g., "NE 24th Street Bellevue"

[028] In the above example, a large list of street names is followed by a large list of city names. Assuming three elements in the list of street names, and three elements in the list of city names, this results in twenty-one transitions, which may be represented by traditional grammar tree 210. Every end-node of the first list is followed by all entries in the second list, potentially leading to very large grammars because most real-world large-list applications are likely to include much more than three list items. For example, a city may have hundreds or thousands of street names, and there may be hundreds or thousands of city names. Moreover, every element in the second segment of traditional grammar tree 210 is repeated, once for each first segment, which introduces redundancy.

[029] According to an aspect of the invention, the problems with traditional grammar trees may be resolved by using phonemic acoustic grammars instead of large-lists. The grammar may further be improved by including linking elements to reduce the number of transition states in the grammar. Thus, a grammar tree with a linking

element 220 will merge after a first segment and then spread out again at a second segment, where the segments may represent a phoneme in an acoustic grammar, as discussed above. For example, assume a two-syllable word in an acoustic grammar consisting of three phonemes, which is able to reduce the number of transitions from twenty-one in a traditional grammar tree 210 to twelve in a grammar tree with a linking element 220. Two syllables and three phonemes are chosen to show the reduction in search space in a grammar tree with a linking element 220 as opposed to a corresponding traditional grammar tree 210, although a real-world acoustic grammar modeled after a language is likely to have a maximum of roughly fifty phonemes. Moreover, the search space may be further reduced by restricting available transitions based on phonotactic constraints for an acoustic model.

[030] Using the approach described in Fig. 2, adding a linking element to an acoustic grammar may reduce both grammar size and response time. Part of a speech signal may be mapped to the linking element in order to maintain the phonotactic rules of the acoustic grammar. The linking element may be an acoustic element that is likely to be triggered even if unpronounced. For example, a *schwa* represents an unstressed, central vowel in the English language (e.g., the first and last sound in the word "arena" is *schwa*). The phoneme *schwa* is an ideal linking element because of how it is represented in a frequency spectrum. That is, *schwa* is a brief sound and when a person opens their mouth to speak, there is a strong likelihood of passing through the frequencies of *schwa* even if unintended. Those skilled in the art will recognize that this approach may be extended to acoustic models of speech signals for other languages by using frequently elided phonemes as linking elements to reduce the search space of an acoustic grammar.

[031] Referring again to Fig. 1, speech engine 112 may receive a user verbalization and process the verbalization into a plurality of preliminary interpretations using the techniques described above. That is, the verbalization may be interpreted as a series of phonemes, and the series of phonemes may be mapped to one or more preliminary interpretations by traversing one or more acoustic grammars that are modeled after grammar 220 of Fig. 2. The plurality of preliminary interpretations may take the form of words, parts of words, phrases, utterances, or a combination thereof, and the plurality of preliminary interpretations may be arranged as a matrix, an array, or in another form. The plurality of preliminary interpretations are then

passed to a speech sharpening engine 110 for deducing a most probable interpretation of the verbalization.

[032] According to various aspects of the invention, speech sharpening engine 110 may include an interpretation sharpening module 116, a policy module 114, an interpretation history analysis module 118, and a policy agent handler 120. The plurality of preliminary interpretations may be received by interpretation sharpening module 116, which forwards the preliminary interpretations to policy module 114 for further processing. Policy module 114 may include one or more context agents 126, one or more profile agents 128, and a context tracking module 130 that collectively revise the plurality of preliminary interpretations into a set of hypotheses that represent candidate recognitions of the verbalization. Policy module 114 may assign each hypothesis an interpretation score, and interpretation sharpening module 116 may designate the hypothesis with the highest (or lowest) interpretation score as a probable interpretation.

[033] According to one aspect of the invention, policy module 114 may include one or more context agents 126. Context agents 126 may represent domains of knowledge corresponding to a given context, such as song titles, city and street names, finance, movies, or other contexts. Context agents 126 may use context objects and associated dynamic languages to represent a corresponding context. Policy module 114 may also include one or more profile agents 128. Profile agents 128 may represent domains of knowledge corresponding to a given profile, such as a specific user, language, accent, or other profiles. Profile agents 128 may use profile objects and dynamic languages to represent a corresponding profile. Dynamic languages for context agents 126 or profile agents 128 may specify vocabularies, word combinations, phrases, sentence structures, criteria, and priority weightings for any given context or profile, respectively. The priority weightings may weight individual parameters according to one or more weighting factors, such as assigning a weight according to a frequency of use, a difficulty to understand, or other factors. Policy module 114 may also include a context-tracking module 130. Context tracking module 130 may track a verbalization context of a consecutive series of verbalizations. Context tracking module 130 may utilize one or more conversation trees to track the series of verbalizations. Context tracking sub-module 214 may track one or more past or current verbalization contexts of the series of verbalizations, and/or may make predictions regarding one or more future

verbalization contexts of the series of verbalizations. Policy module 114 may utilize information about the verbalization context, generated by context tracking module 130, to generate one or more sharpened interpretations and corresponding interpretation scores.

[034] In some implementations, policy module 114 may use context tracking module 130 to apply objects from one or more context agents 126 and/or profile agents 128 to the preliminary interpretations provided by speech engine 112. The various agents may compete with each other using a set of heuristics in a phonetic fuzzy matcher, where an intent or context of the user may be identified based on the set of heuristics about how a request may be phrased in a given domain. A closest phonetic match may be identified for suspect words and/or phrases among the plurality of preliminary interpretations.

[035] The phonetic fuzzy matcher may include an M-Tree that is populated with context objects, profile objects, and/or dynamic language data from one or more of context agents 126 and/or profile agents 128. M-Trees are known to those skilled in the art. The M-Tree may assign relative priority weights to the context objects, profile objects, and/or dynamic language data in order to account for the possibility of misrecognized phonemes, extraneous phonemes, or erroneously deleted phonemes. A closest distance metric associated with the M-Tree may be used given the relative weightings of phoneme misrecognition, phoneme addition, and phoneme deletion for various contexts and/or profiles.

[036] According to one aspect of the invention, one or more passes may be taken over the plurality of preliminary interpretations to identify dominant words and/or phrases among the plurality of preliminary interpretations. Using the M-Tree weighted model, one or more candidate interpretations may be made based on relevant substitution of suspect words and/or phrases based on phonetic similarities and/or domain appropriateness. For example, if a set of dominant words appear to be a movie name, a candidate interpretation will substitute the relevant words and/or phrases to generate a candidate interpretation about movies. After a set of candidate interpretations have been generated, the candidate interpretations are analyzed using the M-Tree weighted model. With the relevant domains constrained by the candidate interpretations, a confidence or interpretation score may be assigned to each candidate interpretation, with the interpretation score representing a likelihood that a particular candidate interpretation is a correct interpretation of the

verbalization. The candidate interpretations may then be returned to interpretation sharpening module 116, and interpretation sharpening module 116 may select a candidate interpretation with a highest (or lowest) interpretation score as a probable interpretation of the verbalization.

[037] According to various implementations of the invention, speech sharpening engine 110 may include an interpretation history analysis module 118. Interpretation history analysis module 118 may include an information storage module 122 a frequency module 124. Information storage module 122 may store information related to verbalizations, including components of verbalizations, preliminary interpretations, dominant words and/or phrases, candidate interpretations, probable interpretations, and/or interpretation scores associated with verbalizations, as well as whether or not a verbalization was interpreted correctly, or other information. Interpretation history analysis module 118 may also include a frequency module 124. Frequency module 124 may use some or all of the information stored in information storage module 122 to generate one or more frequencies related to one or more past verbalizations. For example, frequency module 124 may calculate a word usage frequency, a word combination frequency, a frequency related to a set of verbalizations that are phonetically similar but have distinct meanings, an interpretation error frequency for a particular verbalization, or other frequencies.

[038] Information stored and/or generated by interpretation history analysis module 118 may be used to augment speech sharpening engine 110. In some implementations, the information may be used to adjust various weights used in phonetic models, such as context agents 126 or profile agents 128, as well as adapting the relative weights in the M-Tree in context tracking module 130 to enhance accuracy for subsequent verbalizations. In another implementation, the stored information may be sent to a third-party or commercial entity for analyzing the data and developing new domain agents or further improving the accuracy of speech sharpening engine 110. For example, one or more parameters and/or weighting schemes of an agent may be augmented based on a frequency generated by interpretation history analysis module 118. Other augmentations related to information stored on and/or generated by interpretation history analysis module 118 may be made. Speech sharpening engine 110 may also include a policy agent handler 120 that may augment, update, remove, and/or provide one or more domain agents to policy module 114. A domain agent may include one or more new,

modified, or updated context agents 126 and/or profile agents 128. Policy agent handler 120 may also augment or update the M-Tree in context tracking module 130 to adjustments in priority weighting schemes or phonetic models. Policy agent handler 120 may be controlled, directly or indirectly, by a third party, such as a commercial entity, and domain agents may be augmented, updated, removed, and/or provided by policy agent handler 120 as part of a commercial agreement, licensing agreement, subscription agreement, maintenance agreement, or other agreement.

[039] Referring to Fig. 3, a flow chart demonstrating an exemplary method for enhancing the performance and accuracy of speech interpretation is provided. The method may begin by receiving a user verbalization at an operation 312. The received user verbalization may be electronically captured at operation 312, such as by a microphone or other electronic audio capture device. The electronically captured verbalization may be provided to a speech interpretation engine, such as speech engine 112 in Fig. 1.

[040] The speech interpretation may then generate one or more preliminary interpretations of the received verbalization at an operation 314. According to one implementation of the invention, the plurality of preliminary interpretations may be generated using phonetic dictation, grammar trees with linking elements, or any combination thereof to improve performance and enhance accuracy. Phonetic dictation and reducing a search space of a grammar tree by including linking elements is discussed in greater detail above. The preliminary interpretations may be arranged in any predetermined form, such as an array, a matrix, or other forms.

[041] In an operation 320, the preliminary interpretations may be provided to a speech sharpening engine. The speech sharpening engine may take one or more passes over the plurality of preliminary interpretations to identify dominant words and/or phrases in operation 320. This information may then be used to generate one or more candidate interpretations. The candidate interpretations may be based on various domain agents, such as context agents and/or profile agents, which may be organized as a weighted domain model, such as an M-Tree. For example, if a set of dominant words sound like a movie name, apply policies operation 320 may generate a candidate interpretation that substitutes relevant words and/or phrases based on a domain agent populated with movie titles. Additional passes may be made over the candidate interpretations, which may be constrained by domain

information associated with the candidate interpretations, to thereby generate a confidence score or interpretation score for each candidate interpretation. The interpretation score may represent a likelihood that a particular candidate interpretation is a correct interpretation of the verbalization received in operation 312. The operation of apply policies 320 is described in greater detail above in reference to Fig. 1.

[042] The candidate interpretations and corresponding interpretation scores may then be analyzed to determine a probable interpretation in an operation 322. In one implementation of the invention, a candidate interpretation with a highest (or lowest) score may be designated as a probable interpretation. The probable interpretation may then be output in an operation 324, such as for use in a voice-activated vehicular navigation system, a voice-controlled server or desktop computer, or other electronic device that can be controlled using voice commands.

[043] Information relating to the verbalization and the interpretations of the verbalization may be provided in a store interpretation operation 325. Store interpretation operation 324 may store interpretation information related to verbalizations, such as components of verbalizations, preliminary interpretations, dominant words and/or phrases, candidate interpretations, probable interpretations, and/or interpretation scores associated with verbalizations, as well as whether or not a verbalization was interpreted correctly, or other information. In some implementations of the invention, some or all of the interpretation information stored at store interpretation operation 324 may be used to determine one or more frequencies at a determine frequencies operation 326. The frequencies calculated at determine frequencies operation 326 may include one or more frequencies related to past verbalizations, such as, a word usage frequency, a word combination frequency, a frequency related to a set of verbalizations that are phonetically similar but have distinct meanings, an interpretation error frequency for a particular verbalization, or other frequencies. Determine frequencies operation 326 may be performed by interpretation history analysis module 118.

[044] In various implementations, a decision may be made whether to augment a speech sharpening engine in an augmentation decision operation 328. The decision concerning system augmentation may be based at least in part on information generated at determine frequencies block 326, such as one or more frequencies, or other information. If it is decided that no augmentation is needed, no further action is

taken until another verbalization is captured, and the method ends. In some instances, decision operation 328 may determine that augmentation may be made and control passes to an augment system operation 330. Augment system operation 330 may include making an augmentation to a speech sharpening engine. For example, one or more domain agents may be augmented to reflect probabilities of an interpretation being a correct interpretation of a verbalization, to update a user profile, or other augmentation. Dynamic languages associated with context agents and/or profile agents may be augmented, or parameters weights may be augmented to enhance accuracy when interpreting subsequent verbalizations. For example, an adaptive misrecognition technique may adjust the various weights in a phonetic model or update similarity weights for regional accents, or other augmentations may be made. In parallel to augment system operation 330, new agent policies may be received in an operation 332. For example, a third party or commercial entity may redesign or modify various domain agents, new domain agents may be developed and installed as plug-ins, domain agents that are unreliable may be removed, or other augmentations or modifications may be made. Thus, the method continually refines the domain agents and the weighting of various parameters in order to refine the accuracy of the speech sharpening engine for subsequent verbalizations.

[045] The above disclosure has been described in terms of specific exemplary aspects, implementations, and embodiments of the invention. However, those skilled in the art will recognize various changes and modifications that may be made without departing from the scope and spirit of the invention. For example, references throughout the specification to "one implementation," "one aspect," "an implementation," or "an aspect" may indicate that a particular feature, structure, or characteristic is included in at least one implementation. However, the particular features, structures, or characteristics may be combined in any suitable manner in one or more implementations. Therefore, the specification and drawings are to be regarded as exemplary only, and the scope of the invention is to be determined solely by the appended claims.

CLAIMS

What is claimed is:

1. A computer-implemented method for interpreting speech, the method comprising:

receiving a user verbalization;

parsing the verbalization into a plurality of phonemes; and

generating at least one interpretation of the verbalization based on the plurality of phonemes.

2. The method of claim 1, wherein generating the at least one preliminary interpretation includes traversing an acoustic grammar based on the plurality of phonemes.

3. The method of claim 2, wherein the acoustic grammar is a phonemic representation of a speech signal.

4. The method of claim 3, wherein the acoustic grammar includes a linking element between segments, the linking element being an acoustic element in the speech signal.

5. The method of claim 4, wherein the acoustic element is a phoneme representing an unstressed vowel.

6. The method of claim 1, further comprising:

revising the at least one interpretation, wherein the revising includes generating a plurality of candidate interpretations of the verbalization;

assigning an interpretation score to each candidate interpretation; and

selecting a candidate interpretation with a highest interpretation score as a probable interpretation of the verbalization.

7. The method of claim 6, wherein revising the at least one interpretation includes:

applying the at least one interpretation against at least one domain agent;

identifying suspect words and/or phrases in the at least one interpretation;

identifying a closest phonetic match for the suspect words and/or phrases based on a relative weighting of parameters in at least one domain associated with the at least one domain agent; and

substituting the closest phonetic match for the suspect words and/or phrases to generate a candidate interpretation.

8. The method of claim 7, wherein the closest phonetic match is stored in an M-Tree, and wherein identifying the closest phonetic match includes using a closest-distance metric associated with the M-Tree.

9. The method of claim 7, wherein the parameters include at least one of a context of the verbalization, a verbalization criteria, a history of user verbalizations, a user profile, or domain specific information.

10. The method of claim 7, further comprising adjusting the relative weighting of parameters in response to the selection of a probable interpretation of the verbalization.

11. A method for sharpening a preliminary interpretation of a user verbalization, the preliminary interpretation generated by a speech interpretation engine, the method comprising:

revising the preliminary interpretation, wherein the revising includes generating a plurality of candidate interpretations of the verbalization;

assigning an interpretation score to each candidate interpretation; and

selecting a candidate interpretation with a highest interpretation score as a probable interpretation of the verbalization.

12. The method of claim 11, wherein revising the preliminary interpretation includes:

applying the preliminary interpretation against at least one domain agent;

identifying suspect words and/or phrases in the preliminary interpretation;

identifying a closest phonetic match for the suspect words and/or phrases based on a relative weighting of parameters in at least one domain associated with the at least one domain agent; and

substituting the closest phonetic match for the suspect words and/or phrases to generate a candidate interpretation.

13. The method of claim 12, wherein the closest phonetic match is stored in an M-Tree, and wherein identifying the closest phonetic match includes using a closest-distance metric associated with the M-Tree.

14. The method of claim 12, wherein the parameters include at least one of a context of the verbalization, a verbalization criteria, a history of user verbalizations, a user profile, or domain specific information.

15. The method of claim 12, further comprising adjusting the relative weighting of parameters in response to the selection of a probable interpretation of the verbalization.

16. A system for interpreting speech, the system comprising:

a device that receives a user verbalization and generates an electronic signal corresponding to the user verbalization; and

a speech interpretation engine that receives the electronic signal corresponding to the user verbalization, wherein the speech interpretation engine parses the signal into a plurality of phonemes and generates at least one interpretation of the verbalization based on the plurality of phonemes.

17. The system of claim 16, wherein generating the at least one preliminary interpretation includes traversing an acoustic grammar based on the plurality of phonemes.

18. The system of claim 17, wherein the acoustic grammar is a phonemic representation of a speech signal.

19. The system of claim 18, wherein the acoustic grammar includes a linking element between segments, the linking element being an acoustic element in the speech signal.

20. The system of claim 19, wherein the acoustic element is a phoneme representing an unstressed vowel.

21. The system of claim 16, further comprising a speech sharpening engine configured to:

revise the at least one interpretation by generating a plurality of candidate interpretations of the verbalization;

assign an interpretation score to each candidate interpretation; and

select a candidate interpretation with a highest interpretation score as a probable interpretation of the verbalization.

22. The system of claim 21, wherein the speech sharpening engine is configured to revise the at least one interpretation by:

applying the at least one interpretation against at least one domain agent;

identifying suspect words and/or phrases in the at least one interpretation;

identifying a closest phonetic match for the suspect words and/or phrases based on a relative weighting of parameters in at least one domain associated with the at least one domain agent; and

substituting the closest phonetic match for the suspect words and/or phrases to generate a candidate interpretation.

23. The system of claim 22, wherein the closest phonetic match is stored in an M-Tree, and wherein identifying the closest phonetic match includes using a closest-distance metric associated with the M-Tree.

24. The system of claim 22, wherein the parameters include at least one of a context of the verbalization, a verbalization criteria, a history of user verbalizations, a user profile, or domain specific information.

25. The system of claim 22, wherein the speech sharpening engine is further configured to adjust the relative weighting of parameters in response to the selection of a probable interpretation of the verbalization.

26. A system for sharpening a preliminary interpretation of a user verbalization received by an electronic capture device, the preliminary interpretation generated by a speech interpretation engine, the system comprising a speech sharpening engine containing at least one processor configured to:

revise the preliminary interpretation by generating a plurality of candidate interpretations of the verbalization;

assign an interpretation score to each candidate interpretation; and

select a candidate interpretation with a highest interpretation score as a probable interpretation of the verbalization.

27. The system of claim 26, wherein the speech sharpening engine is further configured to revise the preliminary interpretation by:

applying the preliminary interpretation against at least one domain agent;

identifying suspect words and/or phrases in the preliminary interpretation;

identifying a closest phonetic match for the suspect words and/or phrases based on a relative weighting of parameters in at least one domain associated with the at least one domain agent; and

substituting the closest phonetic match for the suspect words and/or phrases to generate a candidate interpretation.

28. The system of claim 27, wherein the closest phonetic match is stored in an M-Tree, and wherein identifying the closest phonetic match includes using a closest-distance metric associated with the M-Tree.

29. The system of claim 27, wherein the parameters include at least one of a context of the verbalization, a verbalization criteria, a history of user verbalizations, a user profile, or domain specific information.

30. The system of claim 27, wherein the speech sharpening engine is further configured to adjust the relative weighting of parameters in response to the selection of a probable interpretation of the verbalization.

31. A computer-readable medium containing computer-executable instructions for interpreting speech, the computer-executable instructions configured to:

receive a user verbalization;

parse the verbalization into a plurality of phonemes; and

generate at least one interpretation of the verbalization based on the plurality of phonemes.

32. The computer-readable medium of claim 31, wherein the computer-executable instructions are configured to generate the at least one preliminary interpretation by traversing an acoustic grammar based on the plurality of phonemes.

33. The computer-readable medium of claim 32, wherein the acoustic grammar is a phonemic representation of a speech signal.

34. The computer-readable medium of claim 33, wherein the acoustic grammar includes a linking element between segments, the linking element being an acoustic element in the speech signal.

35. The computer-readable medium of claim 34, wherein the acoustic element is a phoneme representing an unstressed vowel.

36. The computer-readable medium of claim 31, wherein the computer-executable instructions are further configured to:

revise the at least one interpretation by generating a plurality of candidate interpretations of the verbalization;

assign an interpretation score to each candidate interpretation; and

select a candidate interpretation with a highest interpretation score as a probable interpretation of the verbalization.

37. The computer-readable medium of claim 36, wherein the computer-executable instructions are configured to revise the at least one interpretation by:

applying the at least one interpretation against at least one domain agent;

identifying suspect words and/or phrases in the at least one interpretation;

identifying a closest phonetic match for the suspect words and/or phrases based on a relative weighting of parameters in at least one domain associated with the at least one domain agent; and

substituting the closest phonetic match for the suspect words and/or phrases to generate a candidate interpretation.

38. The computer-readable medium of claim 37, wherein the closest phonetic match is stored in an M-Tree, and wherein identifying the closest phonetic match includes using a closest-distance metric associated with the M-Tree.

39. The computer-readable medium of claim 37, wherein the parameters include at least one of a context of the verbalization, a verbalization criteria, a history of user verbalizations, a user profile, or domain specific information.

40. The computer-readable medium of claim 37, wherein the computer-executable instructions are further configured to adjust the relative weighting of parameters in response to the selection of a probable interpretation of the verbalization.

41. A computer-readable medium containing computer-executable instructions for sharpening a preliminary interpretation of a user verbalization, the preliminary interpretation generated by a speech interpretation engine, the computer-executable instructions configured to:

revise the preliminary interpretation by generating a plurality of candidate interpretations of the verbalization;

assign an interpretation score to each candidate interpretation; and

select a candidate interpretation with a highest interpretation score as a probable interpretation of the verbalization.

42. The computer-readable medium of claim 41, wherein the computer-executable instructions are configured to revise the preliminary interpretation by:

applying the preliminary interpretation against at least one domain agent;

identifying suspect words and/or phrases in the preliminary interpretation;

identifying a closest phonetic match for the suspect words and/or phrases based on a relative weighting of parameters in at least one domain associated with the at least one domain agent; and

substituting the closest phonetic match for the suspect words and/or phrases to generate a candidate interpretation.

43. The computer-readable medium of claim 42, wherein the closest phonetic match is stored in an M-Tree, and wherein identifying the closest phonetic match includes using a closest-distance metric associated with the M-Tree.

44. The computer-readable medium of claim 42, wherein the parameters include at least one of a context of the verbalization, a verbalization criteria, a history of user verbalizations, a user profile, or domain specific information.

45. The computer-readable medium of claim 42, wherein the computer-executable instructions are further configured to adjust the relative weighting of parameters in response to the selection of a probable interpretation of the verbalization.

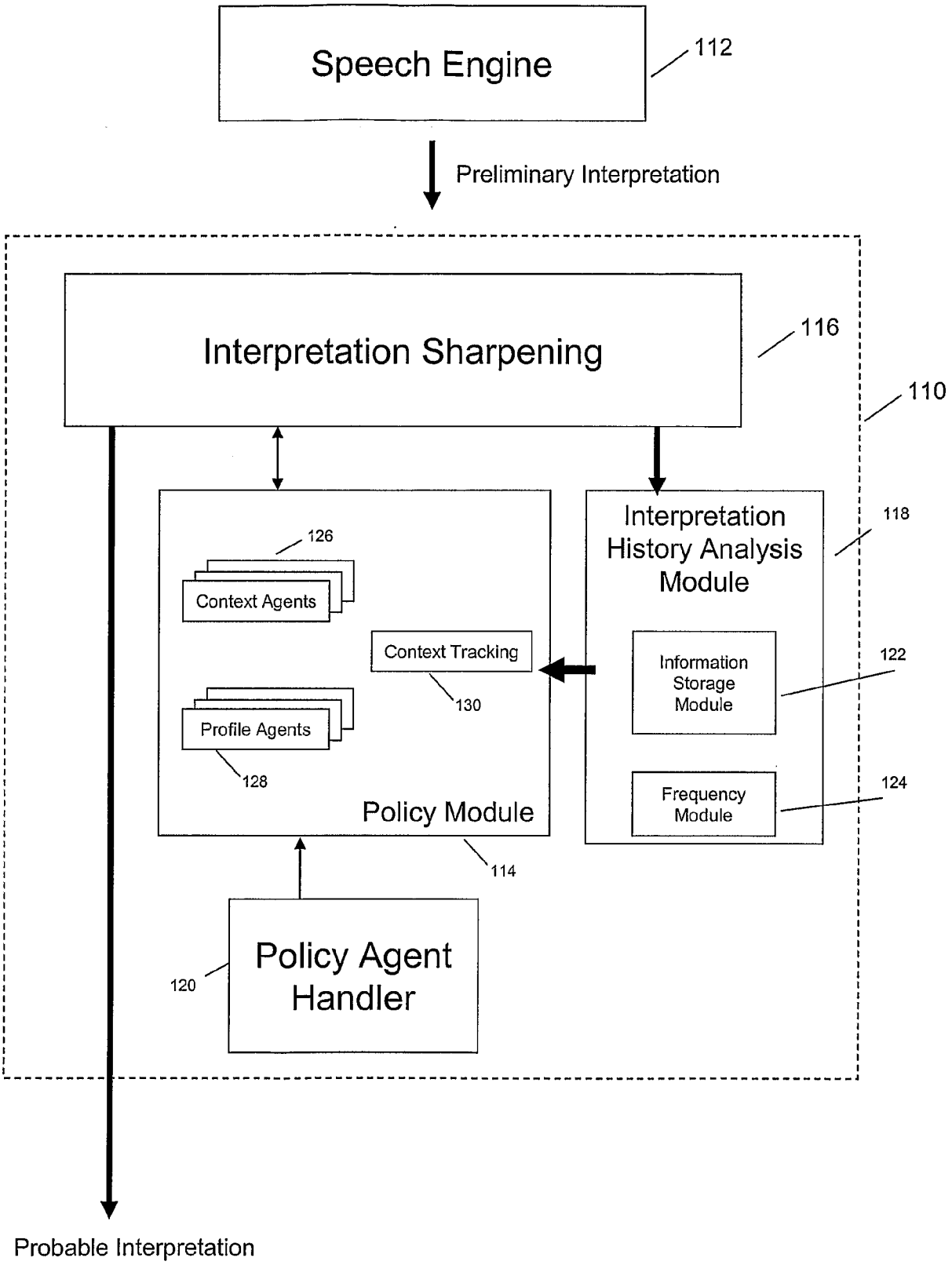
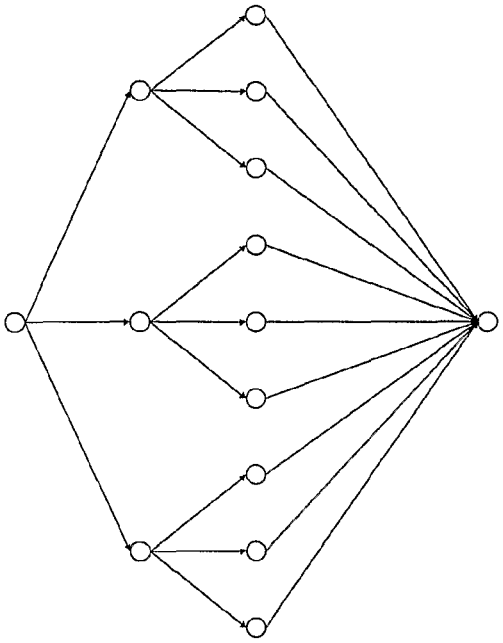

Speech Engine
112
Preliminary Interpretation
Interpretation Sharpening
116
110
Interpretation History Analysis Module
118
126
Context Agents
Context Tracking
130
Information Storage Module
122
Profile Agents
128
Frequency Module
124
Policy Module
114
Policy Agent Handler
120
Probable Interpretation

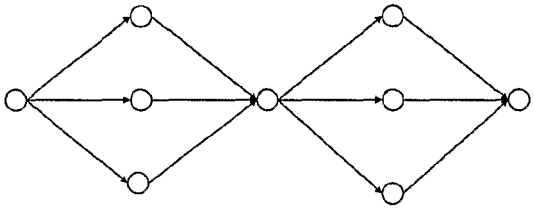
Fig. 1

Traditional Grammar Tree



210

Grammar Tree with Linking Element



220

Figure 2

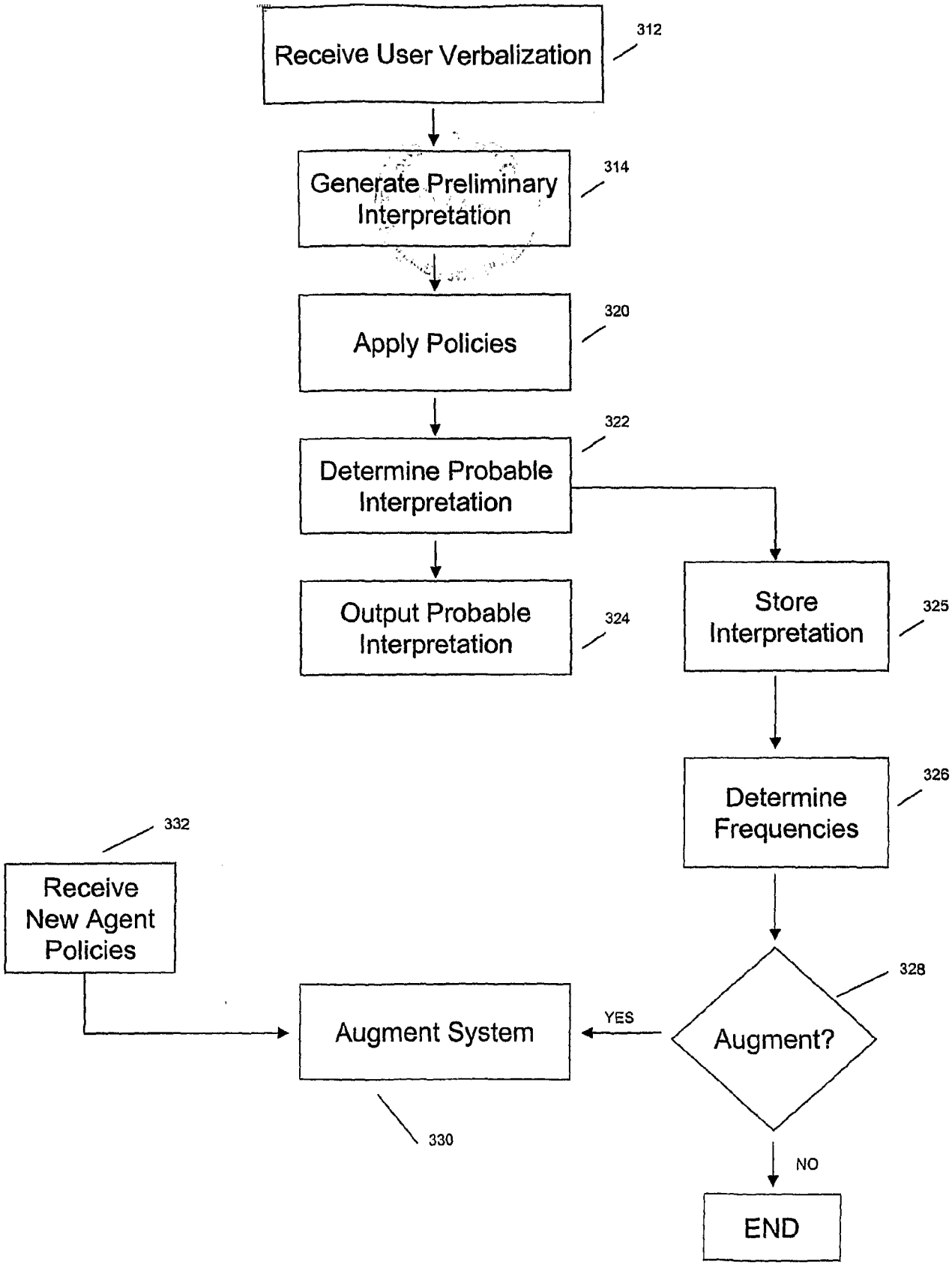

Receive User Verbalization
312
Generate Preliminary Interpretation
314
Apply Policies
320
Determine Probable Interpretation
322
Output Probable Interpretation
324
Store Interpretation
325
Determine Frequencies
326
Augment?
328
YES
NO
Augment System
330
Receive New Agent Policies
332
END

Fig. 3

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
14 August 2008 (14.08.2008)

PCT

(10) International Publication Number
WO 2008/098039 A1

(51) International Patent Classification:
G06Q 30/00 (2006.01)

(21) International Application Number:
PCT/US2008/053157

(22) International Filing Date: 6 February 2008 (06.02.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/671,526 6 February 2007 (06.02.2007) US

(71) Applicant *(for all designated States except US)*: **VOICE-BOX TECHNOLOGIES, INC.** [US/US]; 11980 NE 24th Street, Bellevue, WA 98005 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **FREEMAN, Tom** [US/US]; 6735 83rd Avenue, S.E., Mercer Island, WA 98040 (US). **KENNEWICK, Mike** [US/US]; 13407 Ne 37th Place, Bellevue, WA 98005 (US).

(74) Agents: **GATTO, James, G.** et al.; Pillsbury Winthrop Shaw Pittman LLP, P.O. Box 10500, McLean, VA 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

(54) **Title:** SYSTEM AND METHOD FOR SELECTING AND PRESENTING ADVERTISEMENTS BASED ON NATURAL LANGUAGE PROCESSING OF VOICE-BASED INPUT

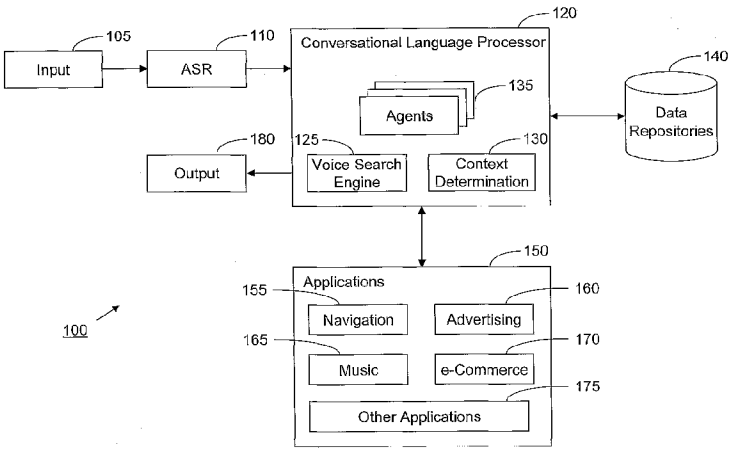


Fig. 1

(57) **Abstract:** A system and method for selecting and presenting advertisements based on natural language processing of voice-based inputs is provided. A user utterance may be received at an input device, and a conversational, natural language processor may identify a request from the utterance. At least one advertisement may be selected and presented to the user based on the identified request. The advertisement may be presented as a natural language response, thereby creating a conversational feel to the presentation of advertisements. The request and the user's subsequent interaction with the advertisement may be tracked to build user statistical profiles, thus enhancing subsequent selection and presentation of advertisements.

WO 2008/098039 A1

SYSTEM AND METHOD FOR SELECTING AND PRESENTING ADVERTISEMENTS BASED ON NATURAL LANGUAGE PROCESSING OF VOICE-BASED INPUT

FIELD OF THE INVENTION

[001] The present invention relates to selecting and presenting advertisements based on natural language processing of voice-based input.

BACKGROUND OF THE INVENTION

[002] As technology advances, consumer electronics devices tend to play larger roles due to increased functionality and mobility. For example, mobile phones, navigation devices, embedded devices, and other such devices provide a wealth of functionality beyond core applications. However, increased functionality adds difficulty to the learning curve associated with using electronic devices, and increased mobility intensifies the demand for simple mechanisms to interact with devices on the go. For example, existing systems tend to have complex human to machine interfaces, which may inhibit mass-market adoption for various technologies. For example, when a user wishes to perform a relatively simple task on a mobile phone, such as purchasing a ring tone, the user often is forced to navigate through a series of menus and press a series of buttons. In some instances, this may result in the transaction not necessarily occurring, as the user may prefer to avoid the hassles altogether. As such, there is ever-growing demand for ways to exploit technology in intuitive ways.

[003] Voice recognition software may enable a user to exploit applications and features of a device that may otherwise be unfamiliar, unknown, or difficult to use. However, many existing voice user interfaces (when they actually work) still require significant learning on the part of the user. For example, users often cannot directly issue a request for a system to retrieve information or perform an action without having to memorize specific syntaxes, words, phrases, concepts, semantic indicators, or other keywords/qualifiers. Similarly, when users are uncertain of particular needs, many existing systems do not engage the user in a productive, cooperative dialogue to resolve requests and advance a conversation. Instead, many existing speech interfaces force users to use a fixed set commands or

keywords to communicate requests in ways that systems can understand. Using existing voice user interfaces, there is virtually no option for dialogue between the user and the system to satisfy mutual goals.

[004] The lack of adequate voice user interfaces results in missed opportunities for providing valuable and relevant information to users. Not only does this potentially leave user requests unresolved, in certain instances, providers of goods and services may lose out on potential business. In an increasingly global marketplace, where marketers are continually looking for new and effective ways to reach consumers, the problems with existing voice user interfaces leaves a large segment of consumer demand unfulfilled. Furthermore, existing techniques for marketing, advertising, or otherwise calling consumers to action fail to effectively utilize voice-based information, which is one of the most natural, intuitive methods of human interaction.

[005] Existing systems suffer from these and other problems.

SUMMARY OF THE INVENTION

[006] According to various aspects of the invention, a system and method for selecting and presenting advertisements based on natural language processing of voice-based inputs is provided. A natural language voice-based input may be received by a voice user interface. The voice-based input may include a user utterance, and a request may be identified from the utterance. Appropriate action may be taken to service the request, while one or more advertisements may be selected and presented to the user. Advertisements may be selected based on various criteria, including content of the input (e.g., concepts, semantic indicators, etc.), an activity related to the input (e.g., a relation to a request, a requested application, etc.), user profiles (e.g., demographics, preferences, location, etc.), or in other ways. A user may subsequently interact with the advertisement (e.g., via a voice-based input), and action may be taken in response to the interaction. Furthermore, the interaction may be tracked to build statistical profiles of user behavior based on affinities or clusters among advertisements, user profiles, contexts, topics, semantic indicators, concepts, or other criteria.

[007] According to various aspects of the invention, advertisers may create advertisements, which may be stored in an advertisement repository. For example,

advertisements may include sponsored messages, calls to action, purchase opportunities, trial downloads, or any other marketing communication, as would be apparent to those skilled in the art. Advertisers may specify various parameters to associate with the advertisements, such as various contexts or topic concepts (e.g., semantic indicators for a "music" concept may include words such as "music," "tunes," "songs," etc.), target demographics (e.g., a preferred audience), marketing criteria or prices for insertion (e.g., dynamic or static pricing based on various marketing criteria), or other information, as would be apparent. The advertisement repository may be associated with a server, where in response to a voice-based input from a user (e.g., at a voice-enabled device), a communications link may be established with the server. Information may be extracted from the voice-based input (e.g., words in the input, applications requested by the input, etc.), and the extracted information may be correlated with user profiles, advertisement parameters, or other information to determine which advertisements to select in relation to the voice-based input. The server may subsequently communicate the selected advertisements to the user, and the server may track the user's subsequent interaction with the selected advertisements.

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

BRIEF DESCRIPTION OF THE DRAWINGS

[009] Fig. 1 illustrates a block diagram of an exemplary system for implementing a voice user interface according to various aspects of the invention.

[010] Fig. 2 illustrates a block diagram of an exemplary advertising system according to various aspects of the invention.

[011] Fig. 3 illustrates a flow diagram of an exemplary method for selecting and presenting advertisements based on voice-based inputs according to various aspects of the invention.

DETAILED DESCRIPTION

[012] Referring to Fig. 1, an exemplary system 100 for implementing a voice user interface is illustrated according to various aspects of the invention. System 100 may enable users to perform various tasks on a voice-enabled device. For example, users may control navigation

devices, media devices, personal computers, personal digital assistants, or any other device supporting voice-based inputs. System 100 may enable users to request voice-enabled devices to retrieve information or perform various tasks, among other things, using natural language voice-based inputs. For example, system 100 may interpret natural language voice-based inputs and generate responses using, among other things, techniques described in co-pending U.S. Patent Application Ser. No. 10/452,147, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed June 3, 2003; and U.S. Patent Application Ser. No. 10,618,633, entitled "Mobile Systems and Methods for Responding to Natural Language Speech Utterance," filed June 15, 2003, both of which are hereby incorporated by reference in their entirety.

[013] According to various aspects of the invention, system 100 may receive a user input, including at least a voice-based user utterance, at an input device 105. Input device 105 may include any suitable device, or combination of devices, for receiving a voice-based input (e.g., a microphone). In various implementations, input device 105 may include a multi-modal input, such as a touch-screen interface, keypad, or other input. The received utterance may be processed by an Automatic Speech Recognizer 110. Automatic Speech Recognizer 110 may generate one or more preliminary interpretations of the utterance using various techniques. For example, Automatic Speech Recognizer 110 may interpret the utterance using techniques of phonetic dictation to recognize a stream of phonemes. Further, Automatic Speech Recognizer 110 may perform post-processing to enhance the preliminary interpretations. For example, Automatic Speech Recognizer 110 may vary interpretations of an utterance, or components of an utterance, from one context to another. Other techniques for enhancing an interpretation of a user utterance may be used, such as those described in co-pending U.S. Patent Application Ser. No. 11/513,269, entitled "Dynamic Speech Sharpening," filed August 31, which is hereby incorporated by reference in its entirety.

[014] The one or more preliminary interpretations may be provided to a conversational language processor 120. Conversational language processor 120 may include a voice search engine 125, a context determination module 130, and one or more agents 135, among other things, to enable cooperative, conversational interaction between the user and system 100. Conversational language processor 120 may be communicatively coupled to one or more data repositories 140 and one or more applications 150. Conversational language processor

120 may generate a domain-specific conversational response, which may be returned to the user as an output 180. Output 180 may include a multi-modal output (e.g., by simultaneously returning a voice-based response and displaying information on a display device).

[015] System 100 may further include an interaction with one or more applications 150 to service one or more requests in the utterance. For example, the utterance may include one or more requests for performing an action, retrieving information, or various combinations thereof. Output 180 may include a conversational response to advance a conversation to service requests by invoking one or more applications 150, as appropriate. For example, applications 150 may include a navigation application 155, an advertising application 160, a music application, an electronic commerce application 170, and/or other applications 175. Furthermore, Automatic Speech Recognizer 110, conversational language processor 120, data repositories 140, and/or applications 150 may reside locally (e.g., on a user device), remotely (e.g., on a server), and/or hybrid local/remote processing models may be used (e.g., lightweight applications may be processed locally while computationally intensive applications may be processed remotely).

[016] Conversational language processor 120 may build long-term and/or short-term shared knowledge in one or more knowledge source. For example, shared knowledge sources may include information about previous utterances, requests, and other user interactions to inform generating an appropriate response to a current utterance. The shared knowledge may include public/non-private (i.e., environmental) knowledge, as well as personal/private (i.e., historical) knowledge. For example, conversational language processor 120 may use context determination module 130 to establish a context for a current utterance by having domain agents 135 competitively generate a context-based interpretation of the utterance (e.g., by scoring possible interpretations and selecting a highest scoring interpretation). As such, agents 135 may model various domains (e.g., navigation, music, a specific user, global users, advertising, e-commerce, etc.), and conversational language processor 120 may interpret and/or respond to a voice-based input accordingly. For example, context-based interpretations and responses to a voice-based input may be generated using techniques described in co-pending U.S. Patent Application Ser. No. 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed August 5, 2005; and U.S. Patent Application Ser. No. 11/212,693,

entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," filed August 29, 2005, both of which are hereby incorporated by reference in their entirety.

[017] Furthermore, conversational language processor 120 may support adaptive misrecognition to reinterpret a current utterance and/or one or more previous utterances. For example, information contained in a current utterance may indicate that interpretations for one or more previous utterances were incorrect, and therefore, the previous utterances may be reinterpreted to improve subsequent interpretations. Accordingly, conversational language processor 120 may use the techniques described herein, along with various other techniques, to interpret and respond to conversational, natural language utterances. Conversational language processor 120 may use various other techniques as will be apparent, such as those described in co-pending U.S. Patent Application Ser. No. 11/200,164, entitled "System and Method of Supporting Adaptive Misrecognition in Conversational Speech," filed August 10, 2005; and U.S. Patent Application Ser. No. 11/580,926, entitled "System and Method for a Cooperative Conversational Voice User Interface," filed October 16, 2006, both of which are hereby incorporated by reference in their entirety.

[018] Referring to Fig. 2, an exemplary advertising system 200 is illustrated according to various aspects of the invention. System 200 may include a server 230 for receiving one or more advertisements from an advertiser 220, wherein the advertisements may be stored in a data repository 260 associated with server 230. For example, advertisements may include sponsored messages or marketing communications, calls to action, purchase opportunities, trial downloads, coupons, or any other suitable marketing, advertising, campaign, or other information, as would be apparent to those skilled in the art. A voice-enabled device 210 may receive a voice-based input and establish communications with advertising server 230. Subsequently, advertising server 230 may select one or more advertisements from among the advertisements stored in data repository 260, and the selected advertisements may be provided to the voice-enabled device for presentation to a user.

[019] Advertiser 220 may access advertising server 230 via an advertiser interface 245. Advertisers 220 may upload targeted advertisements to server 230 via advertiser interface 245, and server 230 may store the advertisements in data repository 260. The advertisements may include graphically-based advertisements that include banners, images,

audio, video, or any suitable combination thereof. Furthermore, the advertisements may include interactive or embedded information, such as links, metadata, or computer-executable instructions, or any suitable combination thereof. Advertisers may specify criteria for a campaign or targeting information for an advertisement (e.g., a start date, an end date, budget information, geo-targeting information, conceptual or contextual information, or any other suitable criteria), which may be used to facilitate selecting an advertisement in relation to a particular voice-based input.

[020] In addition to providing interface 245 for advertisers, server 230 may include a content/action identification module 235, a user profile module 240, an advertisement selection module 250, and a tracking module 255. Users may submit voice-based requests to voice-enabled device 210, and voice-enabled device 210 may communicate information about the voice-based input to server 230. Server 230 may invoke advertisement selection module 250 to extract relevant information from about the voice-based input, where advertisement selection module 250 may select one or more advertisements relevant to the voice-based input based on information extracted using content/action identification module 235 and/or user profile module 240.

[021] For example, content/action identification module 235 may identify content of the voice-based input (e.g., words in the input), requested information (e.g., search results, a web page, music, video, graphics, or other information), requested actions (e.g., calculating a navigation route, placing a telephone call, playing a song, etc.), a category or topic related to the input (e.g., music, business, stocks, sports, navigation, movies, etc.), or other criteria to use in selecting an advertisement. Further, user profile module 240 may identify characteristics of a specific user (e.g., demographics, personal preferences, location-based information, etc.), global user profiles (e.g., demographic profiles, click-through rates, etc.), or other criteria to use in selecting an advertisement. Moreover, advertisement selection module 250 may account for where a request originates from. For example, advertisements may be selected based on a default user location (e.g., identified from a user profile), current geolocation information (e.g., identified from a navigation device), whether an affiliate or partner of server 230 initiated the request, or other criteria.

[022] For instance, a user may request airline reservations via voice-enabled device 210, and content/action identification module 235 may identify specific words used in the request, a category related to the request (e.g., travel, airlines, hotels, etc.), or other

information. Furthermore, user profile module 240 may identify relevant characteristics of the user (e.g., user-specific demographics, location information, preferred airlines or hotels, etc.), as well as global user characteristics (e.g., most popular airlines). In various implementations, advertisements may be selected by assigning a score to each advertisement (e.g., based on click-through rates, relevance metrics, target audiences, etc.). As such, advertisement selection module 250 may correlate the information about the request to select advertisements stored in data repository 260, and server 230 may communicate the selected advertisements to voice-enabled device 210. Furthermore, selected advertisements may be presented according to a predetermined ordering or ranking (e.g., based on a ranking of relevance to an advertisement).

[023] In various implementations, advertisement selection module 250 may retrieve a predetermined number of advertisements for any given request. Furthermore, the selected advertisements may depend upon a presentation format. For example, advertisements may be selected based on an amount of available space on a display of voice-enabled device 210 and/or a size/shape of the selected advertisements. In another example, voice-based advertisements may be selected and presented to the user audibly (e.g., a "hands-free" advertisement may be preferred when voice-enabled device 210 is a telematics device).

[024] Furthermore, the user's subsequent interaction with an advertisement may be tracked using tracking module 255. For example, tracking module 255 may determine whether a conversion or click-through occurs for each advertisement presented to users. Further, tracking module 255 may maintain accounting and/or billing information associated with advertisers 220. For example, advertisers 220 may specify a maximum insertion cost, a cost-per-click-through, an average insertion cost, or other criteria specifying a budget constraint for an advertisement. As such, tracking module 255 may track which advertisements are selected and/or presented, which advertisements result in a conversion or click-through, whether a click-through or conversion results in a transaction or sale, associations between advertisements and users, requests, concepts, semantic indicators, and/or other criteria. For example, tracking user interaction with advertisements may be used to build user-specific and/or global statistical profiles that map or cluster advertisements to topics, semantic indicators, contexts, concepts, etc. based on user behavior, demographics, targeting constraints, content of advertisements, content of requests, actions associated with requests, or other statistically relevant information.

Accordingly, the tracking information may be used to bill or invoice advertisers 220, as well as to improve subsequent performance and relevance of advertisements selected using advertisement selection module 250. Other techniques and features of selecting and presenting advertisements based on voice-based inputs may suitably be employed, as would be apparent.

[025] Referring to Fig. 3, an exemplary method for selecting and presenting advertisements based on a voice-based input is illustrated according to various aspects of the invention. The method may begin in an operation 305, where a voice-based input, including at least a user utterance, may be received at a voice user interface. The voice user interface may include any suitable mechanism for receiving the utterance (e.g., a microphone), and may interface with any suitable voice-enabled device, as would be apparent, including personal navigation devices, personal digital assistants, media devices, telematics devices, personal computers, mobile phones, or others.

[026] Subsequently, one or more requests included in the voice-based input may be identified in an operation 310. For example, the requests may include requests to retrieve information, perform tasks, explore or gather information, or otherwise interact with a system or device. For example, a voice-based input to a navigation device may include a request to calculate a route or retrieve location-based information. In another example, a voice-based input to a mobile phone may include a request to place a telephone call, purchase a ringtone, or record a voice-memo. Furthermore, in various implementations, voice-based inputs may include multiple requests, multi-modal requests, cross-device requests, cross-application requests, or other types of requests. For example, an utterance received in operation 305 may be: "Get me a route to Chang's Restaurant, and call them so I can make a reservation." The utterance may thus include multiple requests, including cross-device requests (e.g., calculate a route using a navigation device, and make a telephone call using a mobile phone), as well as cross-application requests (e.g., search for an address and/or phone number using a voice search engine, and calculate a route using a navigation application).

[027] The requests may be part of a conversational interaction between a user and a system or device, whereby an interpretation of requests in a current utterance may be based upon previous utterances in a current conversation, utterances in previous conversations, context-based information, local and/or global user profiles, or other

information. For example, a previous request may be reinterpreted based on information included in subsequent requests, a current request may be interpreted based on information included in previous requests, etc. Furthermore, the conversational interaction may take various forms, including query-based conversations, didactic conversations, exploratory conversations, or other types of conversations. For example, the conversational language processor may identify a type of conversation, and information may be extracted from the utterance accordingly to identify the one or more requests in operation 310. Moreover, the conversational language processor may determine whether any of the requests are incomplete or ambiguous, and action may be taken accordingly (e.g., a system response may prompt a user to clarify an incomplete and/or ambiguous request). The conversational language processor may therefore use various techniques to identify a conversation type, interpret utterances, identify requests, or perform other tasks, such as those described in the aforementioned co-pending patent applications, which are hereby incorporated by reference in their entirety.

[028] Upon identifying the one or more requests, action may be taken based on the identified requests in an operation 315, while one or more advertisements may be selected in an operation 320 (described in greater detail below). For example, one or more context-appropriate applications may be invoked to service the requests in operation 315 (e.g., a voice search engine, a navigation application, an electronic commerce application, or other application may be invoked depending upon the request). Furthermore, in operation 320, information may be communicated to an advertising server to select one or more advertisements related to the request. Thus, as shown in Fig. 3, taking action in operation 315 and selecting advertisements in operation 320 may be related operations (e.g., advertisements may be selected to help in interpreting incomplete and/or ambiguous requests).

[029] Upon taking action in operation 315 (e.g., to service the request) and selecting one or more advertisements in operation 320 (e.g., in relation to the request), an output may be presented to the user in operation 325. The output may indicate a result of the action associated with operation 315. For example, the output may include requested information, an indication of whether a requested task was successfully completed, whether additional information is needed to service the request (e.g., including a prompt for the information), or other information relating to an action based on the request. Furthermore,

the output may include advertisements, as selected in operation 320. For example, the output may include text-based, graphic-based, video-based, audio-based, or other types of advertisements, as would be apparent to those skilled in the art. Further, the output may include other types of advertisements, including calls to action (e.g., a location-based coupon or purchase opportunity, trial downloads, or other actionable advertising or marketing).

[030] Advertisements may be selected in relation to a request based on various criteria. For example, an advertisement may be selected based on words or other content of the request, relevant words or content related to the words or content of the request, etc. In another example, the advertisement may be selected based on requested tasks/information (e.g., a request for movie showtimes may result in an advertisement being selected for a particular theater). In yet another example, the advertisement may be selected based on a topic or category associated with the requested tasks/information (e.g., a request to purchase airline tickets may result in an advertisement being selected for a hotel in a destination associated with a reserved flight). In still other examples, the advertisement may be selected based on location information, (e.g., advertisements may be selected based on a proximity to a user geolocation identified using a navigation device), user-specific and/or global user profiles (e.g., advertisements may be selected based on user-specific and/or global preferences, advertiser campaign criteria, etc.).

[031] Content of a voice-based input may be determined based on various criteria, including contextual or conceptual information (e.g., semantic indicators, qualifiers, or other information). For example, a given concept may include various semantically equivalent indicators having an identical meaning. Thus, for instance, a voice-based input may be "Play some tunes!" or "Play some music!" or other variants thereof, each of which may be interpreted as relating to a specific idea (or concept) of "Music." Thus, concept or content information in a request may be used to select an advertisement. For example, a user may request to calculate a route in Seattle, Washington (e.g., "How do I get to the Space Needle?"). Based on a context of the requested task (e.g., "Navigation," "Seattle," etc.), a voice search engine may retrieve an address of the Space Needle and a navigation application may calculate the route. Furthermore, user profile information may indicate that the user is visiting Seattle from out-of-town (e.g., the profile may indicate that the user's home is Sacramento), and therefore, an advertisement for popular points-of-interest

in Seattle may be selected. In another example, the user may request information about a sporting event (e.g., "Get me the kickoff time for the Eagles game on Sunday"). Based on a context of the requested information (e.g., "Search," "Sports," "Philadelphia," etc.), the requested information may be retrieved, while an advertisement for Eagles apparel or memorabilia may be selected.

[032] In various instances, concepts, semantic indicators, qualifiers, or other information included in, or inferred from, a request may indicate an exploratory nature for the request. In other words, the exploratory request may identify a goal for a conversation, instead of a particular task to perform or information to retrieve. As such, in various implementations, an advertisement may be selected in operation 320 in an effort to advance the conversation towards the goal. For example, an exploratory request may include a request for a navigation route (e.g., "I feel like going to a museum, find me something interesting"). Based on a context of the requested task (e.g., "Navigation," "Points of Interest," etc.), the goal of the conversation may be identified, and the request may be serviced in operation 315 (e.g., a voice search engine may locate nearby points of interest based on user preferred topics). Further, the advertising application may select an appropriate advertisement in operation 320, where the advertisement may be selected in an attempt to advance the conversation towards the goal. For example, statistical profiles (e.g., user profiles, global profiles, topic-based profiles, etc.) may reflect an affinity between an advertisement for a particular museum and other users sharing similar demographics or other characteristics with the requesting user. Thus, in addition to retrieving information about museums in operation 315, an advertisement for a museum likely to be of interest to the user may be selected in operation 320.

[033] In various instances, a request may include incomplete, ambiguous, unrecognized, or otherwise insufficient semantic indicators, context, qualifiers, or other information needed to identify the request. In other words, the request may include inadequate information to identify or infer a task to perform, information to retrieve, or a goal for a conversation. Thus, as much information as possible may be extracted and/or inferred from the request based on shared knowledge such as context, user or global profile information, previous utterances, previous conversations, etc. As such, servicing the request may include generating a response and/or communicating with an advertising application to advance a conversation toward a serviceable request. For example, servicing the request in operation

315 and selecting an advertisement in operation 320 may include generating a response and/or selecting an advertisement to frame a subsequent user input, thereby advancing the conversation.

[034] For example, the request may include incomplete, ambiguous, or unrecognized information (e.g., "Do you know [*mumbled words*] Seattle?"). A context of the requested task may be identified (e.g., "Seattle"), yet the identified context may be insufficient to adequately take action to service the request. Additional information may be inferred based on previous utterances in the conversation, profile information, or other information. However, when the additional information fails to provide adequate information to infer a reasonable hypothesis, servicing the request in operation 315 may include generating a response to frame a subsequent user input and advance the conversation (e.g., information about various topics may be retrieved based on a user's preferred topics). Further, the advertising application may select an advertisement in operation 320 to advance the conversation (e.g., advertisements may be selected based on user and/or global profiles reflecting an affinity between certain advertisements associated with Seattle and user preferences, profiles, etc.). Thus, by selecting an advertisement, indicating dissatisfaction with an advertisement, or otherwise interacting with an advertisement, the interaction may be used to build context and shared knowledge for a subsequent course of the conversation. For example, a user may select an advertisement, and an interpretation of a subsequent voice-based input (e.g., "Call them," "What's the price range?" etc.) may be interpreted with shared knowledge of the advertisement that the voice-based input relates to. Thus, advertisements may be used in a way that enables advertisers to market to consumers, while also improving the consumers' interaction with a device. Other advantages will be apparent to those skilled in the art.

[035] It will be apparent that operation 320 may use various techniques to select advertisements based on voice-based inputs and/or requests included therein. For example, an advertiser may specify a target audience, marketing criteria, campaign strategies, budget constraints, concepts, semantic indicators, related topics, categories, and/or any other suitable information to associate with an advertisement. For instance, advertisers may pay a premium to prioritize an advertisement in relation to similar advertisements (e.g., advertisements associated with competitors). In another example, various statistical profiles may define affinities between advertisements, topics, users, etc.

(e.g., based on click-through or conversion rates, or other tracking information, as described in more detail below). Thus, advertisements may be selected in operation 320 using various techniques, including content of the request, an activity/action associated with the request, user profiles, user preferences, statistical metrics, advertiser-specified criteria, to advance a conversation, to resolve ambiguous requests, or in various other ways, as will be apparent.

[036] The output presented to the user in operation 325 may be provided to the user in various ways. For example, in various implementations, the output may include a voice-based or otherwise audible response. In another example, when an associated device includes a display mechanism, the output may be displayed on the display device. It will be apparent that many combinations or variants thereof may be used, such as augmenting a voice-based response with information on a display device. For example, a user may request information about restaurants, and an advertisement may be selected based on a user preference indicating a favorite type of restaurant (e.g., a Chinese restaurant may be selected based on a user profile indicating a preference for Chinese). Therefore, in one example, the output presented in operation 325 may display information about various restaurants matching the requested information, while a voice-based advertisement for the Chinese restaurant may be played to the user (e.g., via a speaker or other suitable mechanism for playing voice back to the user). Many other variations will be apparent (e.g., a graphical advertisement may be displayed on a display device, while a corresponding or different voice-based advertisement may be played audibly).

[037] Subsequent interaction between the user and the presented advertisements may be monitored in a decisional operation 330. For instance, when the user elects to interact with the advertisement, action may be taken based on the interaction in an operation 335. The interaction may take various forms, including additional voice-based inputs or other suitable mechanisms for interacting with advertisements (e.g., clicking on an advertisement displayed on a personal digital assistant using an associated stylus). For example, a user may initially request information from a voice-enabled media device (e.g., a satellite radio player) about a song currently playing (e.g., "What is this song?"). In addition to outputting the requested information about the song (e.g., "This song is Double Barrel by Dave and Ansel Collins."), a selected advertisement may enable the user to purchase a ringtone for a mobile phone that corresponds to the song. In this example, the interaction may include a request to purchase the ringtone (e.g., "Yeah, I'll buy that"), and action taken in operation

335 may include completing a transaction for the ringtone and/or downloading the ringtone to the mobile phone. Furthermore, additional advertisements may be selected in an operation 340 based on the interaction, using similar techniques as described in connection with operation 320 (e.g., advertisements for additional ringtones, similar musicians, etc. may be selected). Processing may subsequently return to operation 325 to present output resulting from the interaction.

[038] User advertisement interaction may be tracked in an operation 345. For example, operation 345 may track historical data about users, conversations, topics, contexts, or other criteria to associate information with the selected advertisement. The tracking information may therefore be used to build statistical profiles defining affinities, click-through or conversion rates, or other information about various advertisements, topics, or other criteria on a user-specific and/or a global-user level. Thus, clusters or mappings may be created between advertisements, topics, concepts, demographics, or other criteria based on user behavior with the advertisements (e.g., whether a user interacts with the advertisement in operation 330).

[039] For instance, certain advertisements may experience high click-through rates in relation to a first context and/or topic, but low click-through rates in relation to a second context and/or topic, and therefore, when requests relate to the first context and/or topic, the advertisement may be more likely to be selected in subsequent operations 320/340. In another example, global statistical profiles may indicate that an advertisement experiences more click-throughs by users of a particular demographic, and therefore, the advertisement may be more likely to be selected for users falling within the demographic. Many different techniques for tracking and building statistical profiles will be apparent.

[040] Implementations of the invention may be made in hardware, firmware, software, or any combination thereof. The invention may also be implemented as instructions stored on a machine-readable medium, which may be read and executed by one or more processors. A machine-readable medium may include any mechanism for storing or transmitting information in a form readable by a machine (e.g., a computing device). For example, a machine-readable storage medium may include read only memory, random access memory, magnetic disk storage media, optical storage media, flash memory devices, and others, and a machine-readable transmission media may include forms of propagated signals, such as carrier waves, infrared signals, digital signals, and others. Further, firmware, software,

routines, or instructions may be described in the above disclosure in terms of specific exemplary aspects and implementations of the invention, and performing certain actions. However, it will be apparent that such descriptions are merely for convenience and that such actions in fact result from computing devices, processors, controllers, or other devices executing the firmware, software, routines, or instructions.

[041] Aspects and implementations may be described as including a particular feature, structure, or characteristic, but every aspect or implementation may not necessarily include the particular feature, structure, or characteristic. Further, when a particular feature, structure, or characteristic is described in connection with an aspect or implementation, it will be apparent to effect such feature, structure, or characteristic in connection with other aspects or implementations whether or not explicitly described. Thus, various changes and modifications may be made, without departing from the scope and spirit of the invention. The specification and drawings are to be regarded as exemplary only, and the scope of the invention is to be determined solely by the appended claims.

CLAIMS

What is claimed is:

1. A method for selecting and presenting advertisements in response to voice-based inputs, the method comprising:

receiving a voice-based input, the received input including a request;

retrieving information from at least one knowledge source based on the received request;

selecting an advertisement based on the received request and the retrieved information; and

presenting the selected advertisement to a user.

2. The method of claim 1, wherein the retrieved information is related to one or more of content of the voice-based input, an action associated with the request, a location of the user, a user-specific profile, or a global user profile.

3. The method of claim 1, wherein selecting the advertisement includes identifying an advertisement related to resolving the request when the received input includes insufficient information to resolve the request.

4. The method of claim 1, further comprising tracking an interaction between the user and the presented advertisement.

5. The method of claim 4, wherein the tracking is used to build statistical profiles of one or more users.

6. The method of claim 5, wherein the statistical profiles identify affinities among one or more advertisements, concepts, categories, and/or user profiles.

7. The method of claim 4, wherein the tracking is used to build a context for interpreting subsequent voice-based inputs.

8. The method of claim 4, further comprising taking an action in response to the interaction.

9. The method of claim 8, wherein the action includes performing a task and/or retrieving information based on the interaction.

10. The method of claim 1, wherein the at least one knowledge source includes public information generic to the user, and private information specific to the user.

11. A system for selecting and presenting advertisements in response to voice-based inputs, the system comprising:

at least one input device that receives a voice-based input, the received input including a request; and

at least one processor operable to:

retrieve information from at least one knowledge source based on the received request;

select an advertisement based on the received request and the retrieved information; and

present the selected advertisement to a user.

12. The system of claim 11, wherein the retrieved information is related to one or more of content of the voice-based input, an action associated with the request, a location of the user, a user-specific profile, or a global user profile.

13. The system of claim 11, wherein selecting the advertisement includes identifying an advertisement related to resolving the request when the received input includes insufficient information to resolve the request.

14. The system of claim 11, further comprising tracking an interaction between the user and the presented advertisement.

15. The system of claim 14, wherein the tracking is used to build statistical profiles of one or more users.

16. The system of claim 15, wherein the statistical profiles identify affinities among one or more advertisements, concepts, categories, and/or user profiles.

17. The system of claim 14, wherein the tracking is used to build a context for interpreting subsequent voice-based inputs.

18. The system of claim 14, further comprising taking an action in response to the interaction.

19. The system of claim 18, wherein the action includes performing a task and/or retrieving information based on the interaction.

20. The system of claim 11, wherein the at least one knowledge source includes public information generic to the user, and private information specific to the user.

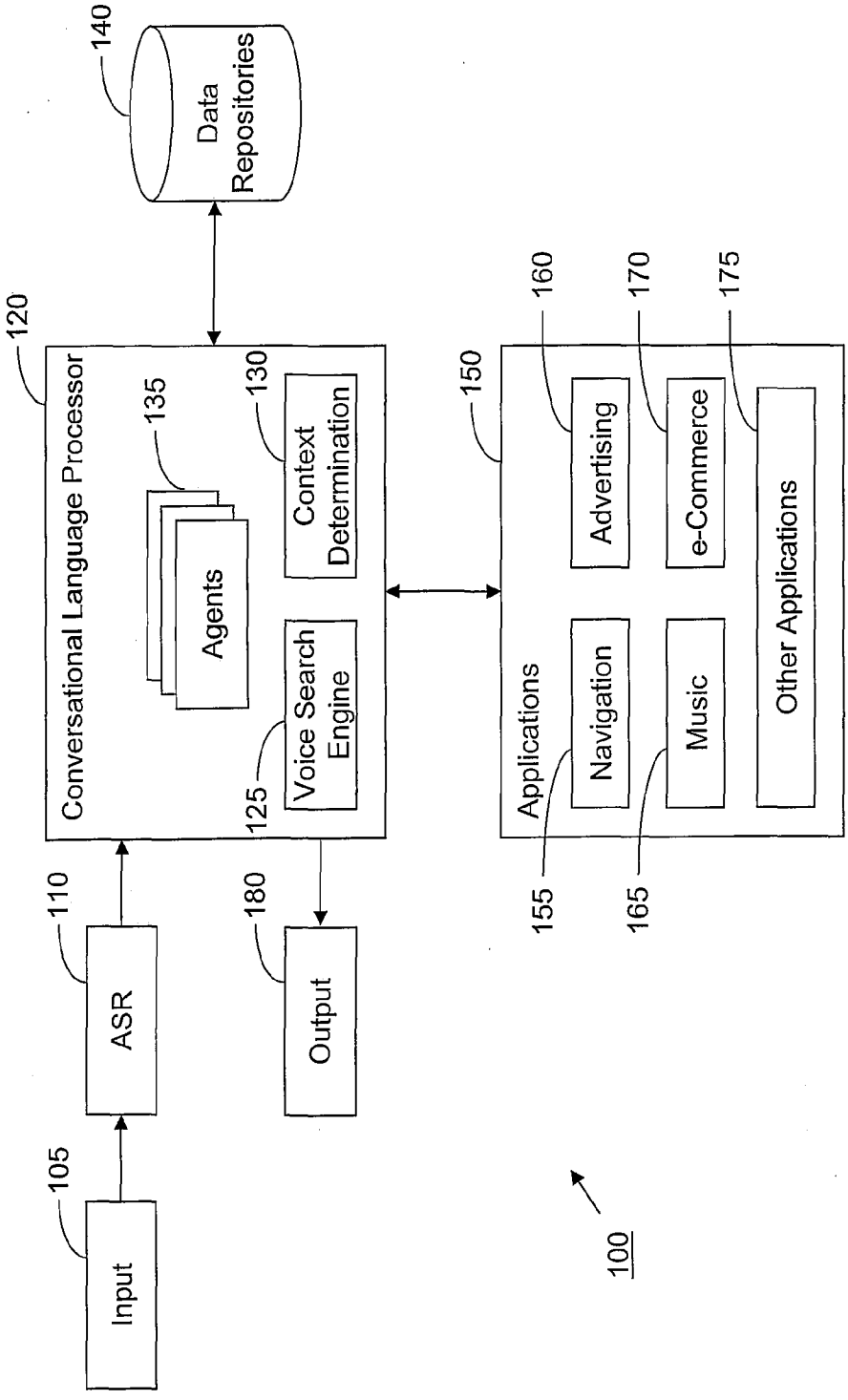

140
Data Repositories
120
Conversational Language Processor
135
Agents
130
Context Determination
125
Voice Search Engine
150
Applications
160
Advertising
170
e-Commerce
175
Other Applications
Navigation
Music
155
165
110
ASR
180
Output
105
Input
100

Fig. 1

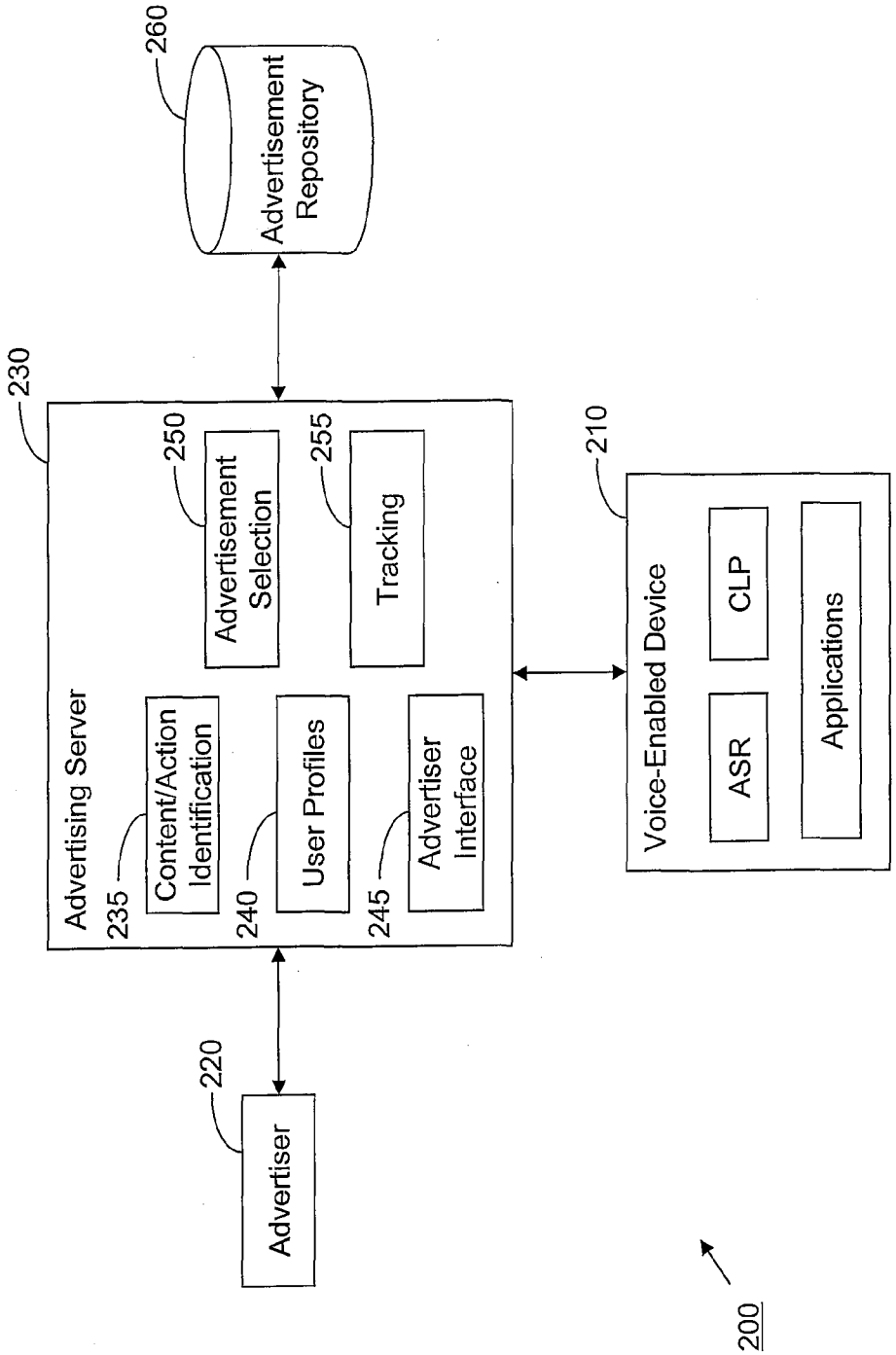

260
Advertisement Repository
230
Advertising Server
250
Advertisement Selection
255
Tracking
235
Content/Action Identification
240
User Profiles
245
Advertiser Interface
210
Voice-Enabled Device
ASR
CLP
Applications
220
Advertiser
200

Fig. 2

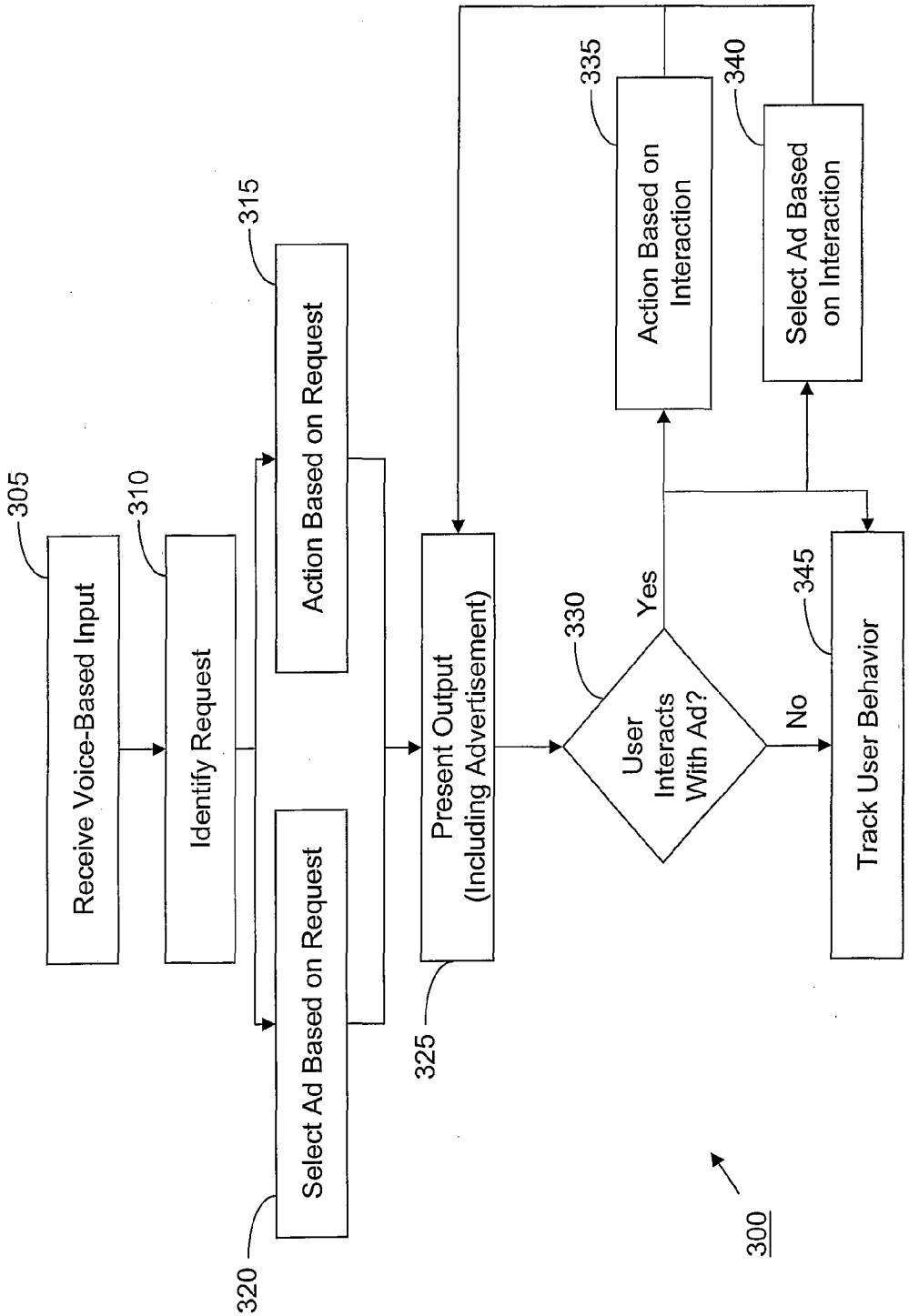

305
Receive Voice-Based Input
310
Identify Request
315
Action Based on Request
320
Select Ad Based on Request
325
Present Output (Including Advertisement)
330
User Interacts With Ad?
Yes
No
335
Action Based on Interaction
340
Select Ad Based on Interaction
345
Track User Behavior
300

Fig. 3

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 08/53157

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

B. FIELDS SEARCHED

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

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

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

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

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

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

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

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

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
2 October 2008 (02.10.2008)

PCT

(10) International Publication Number
WO 2008/118195 A2

(51) International Patent Classification:
G06F 17/28 (2006.01)

(21) International Application Number:
PCT/US2007/081481

(22) International Filing Date: 16 October 2007 (16.10.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/580,926 16 October 2006 (16.10.2006) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [US/US]; 11980 Ne 24th Street, Bellevue, WA 98005 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **BALDWIN, Larry** [US/US]; 25498 S.e. 274th Place, Maple Valley, WA 98038 (US). **FREEMAN, Tom** [US/US]; 6735 83rd Avenue S.e., Mercer Island, WA 98040 (US). **TJALVE, Michael** [DK/US]; 14904 S.e. 47th Court, Bellevue, WA 98006 (US). **EBERSOLD, Blane** [US/US]; 4721 Admiral Way, S.w., Seattle, WA 98116 (US). **WEIDER, Chris** [US/US]; 2618 Grand Avenue, #b-508, Everett, WA 98201 (US).

(74) Agents: **BARUFKA, Jack, S.** et al.; Pillsbury Winthrop Shaw Pittman Llp, P.O. Box 10500, Mclean, VA 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

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

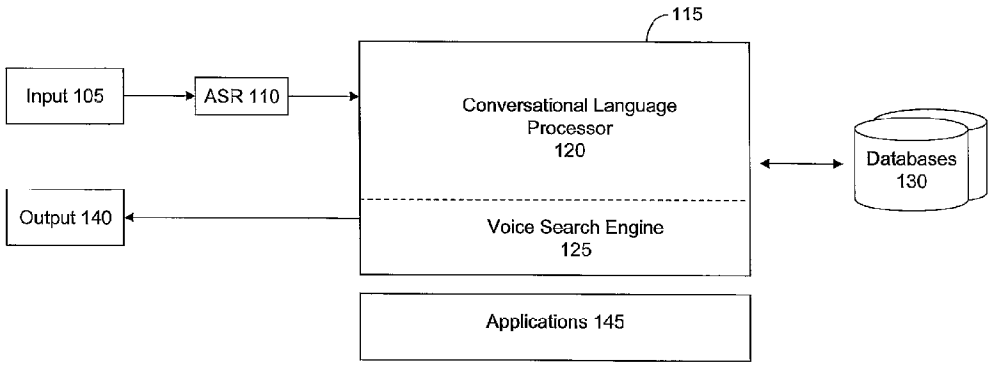


FIG. 1

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

WO 2008/118195 A2

SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

FIELD OF THE INVENTION

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

BACKGROUND OF THE INVENTION

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

[003] In contrast, complex Human-to-Machine interfaces do not allow users to exploit technology in an intuitive way, which inhibits mass-market adoption for various technologies. Incorporating a speech interface helps to alleviate this burden by making interaction easier and faster, but existing speech interfaces (when they actually work) still require significant learning on the part of the user. That is, existing speech interfaces are unable to bridge the gap between archaic Human-to-Machine interfaces and conversational speech that would make interaction with systems feel normal. Users should be able to directly request what they want from a system in a normal, conversational fashion, without having to memorize exact words or phrases. Alternatively, when users are uncertain of particular needs, they should be able to engage the system in a productive, cooperative dialogue to resolve their

requests. Instead, existing speech interfaces force users to dumb down their requests to match simple sets of instructions in simple languages in order to communicate requests in ways that systems can understand. Using existing speech interfaces, there is virtually no option for dialogue between the user and the system to satisfy mutual goals.

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

SUMMARY OF THE INVENTION

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

[006] According to an aspect of the invention, an exemplary system architecture for implementing a cooperative conversational voice user interface is provided. The system may receive an input, which may include a human utterance received by an input device, where the utterance may include one or more requests. As used herein, an "utterance" may be words, syllables, phonemes, or any other audible sound made by a human being. As used herein, a "request" may be a command, directive, or other instruction for a device, computer, or other machine to retrieve information, perform a task, or take some other action. In one implementation, the input may be a multi-modal input, where at least part of the multi-modal input is an utterance. The utterance component of the input may be processed by an Automatic Speech Recognizer to generate one or more preliminary interpretations of the utterance. The one or more preliminary interpretations may then be provided to a conversational speech engine for further processing, where the conversational speech engine may communicate with one or more databases to generate an adaptive conversational response, which may be returned to the user as an output. In one implementation, the output may be a multi-modal output. For example, the utterance may include a request to perform an action, and the output may include a conversational response reporting success or failure, as well as an execution of the action.

[007] According to another aspect of the invention, an exemplary conversational

speech engine may generate an adaptive conversational response to a request or series of requests. The conversational speech engine may include a free form voice search module that may understand an utterance made using typical, day-to-day language (i.e., in free form), and may account for variations in how humans normally speak, the vocabulary they use, and the conditions in which they speak. To account for intangible variables of human speech, the free form search module may include models of casual human speech. For example, in one implementation, the free form search module may understand specialized jargon and/or slang, tolerate variations in word order, and tolerate verbalized pauses or stuttered speech. For example, formalized English requests, where a verb precedes a noun, may be treated in an equivalent manner to requests where the noun precedes the verb. In another implementation, compound requests and/or compound tasks with multiple variables may be identified in a single utterance. By identifying all relevant information for completing one or more tasks from a single utterance, advantages may be provided over existing voice user interfaces, such as Command and Control systems that use verbal menus to restrict information that a person can provide at a given point. In another implementation, inferring intended requests from incomplete or ambiguous requests may provide a conversational feel. By modeling what contextual signifiers, qualifiers, or other information may be required to perform a task in an identified context, an adaptive response may be generated, such as prompting a user for missing contextual signifiers, qualifiers, or other information. In one implementation, the response may ask for missing information in a way that most restricts possible interpretations, and the response may be framed to establish a domain for a subsequent user utterance. In another implementation, common alternatives for nouns and verbs may be recognized to reflect variations in usage patterns according to various criteria. Thus, variations in expression may be supported because word order is unimportant or unanticipated, and nouns and/or verbs may be represented in different ways to give simplistic, yet representative, examples. In another implementation, requests may be inferred from contradictory or otherwise inaccurate information, such as when an utterance includes starts and stops, restarts, stutters, run-on sentences, or other imperfect speech. For example, a user may sometimes change their mind, and thus alter the request in mid-utterance, and the imperfect speech feature may nonetheless be able to infer a request based on models of human speech. For example, various models may indicate that a last criterion is most likely to be correct, or intonation, emphasis, stress, use of the word "not," or other

models may indicate which criterion is most likely to be correct.

[008] According to another aspect of the invention, the conversational speech engine may include a noise tolerance module that may discard words or noise which has no meaning in a given context to reduce a likelihood of confusion. Moreover, the noise tolerance module may filter out environmental and non-human noise to further reduce a likelihood of confusion. In one implementation, the noise tolerance module may cooperate with other modules and features to filter out words that do not fit into an identified context. For example, the noise tolerance module may filter other human conversations and/or utterances within a range of one or more microphones. For example, a single device may include multiple microphones, or multiple devices may each include one or more microphones, and the noise tolerance module may collate inputs and cooperatively filter out sound by comparing a speech signal from the various microphones. The noise tolerance module may also filter out non-human environmental noise within range of the microphones, out-of-vocabulary words caused by speaker ambiguity or malapropisms, or other noise that may be unrelated to a target request. Performance benchmarks for the noise tolerance module may be defined by noise models based on human criteria. For example, if a driver of a car is 92% likely to be understood by a passenger when traveling at 65 miles-per-hour with windows cracked, then performance benchmarks for the noise tolerance module may have a similar performance under such conditions.

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

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

[010] The context determination process may provide advantages over existing voice user interfaces by continually updating one or more models of an existing context and establishing context as a by-product of a conversation, which cannot be established a priori. Rather, the context determination process may track conversation topics and attempt to fit a current utterance into recent contexts, including switching between contexts as tasks are completed, partially completed, requested, etc. The context determination process may identify one or more context domains for an utterance by defining a collection of related functions that may be useful for users in various context domains. Moreover, each context domain may have relevant vocabularies and thought collections to model word groupings, which when evaluated together, may disambiguate one context domain from another. Thus, eliminating out-of-context words and noise words when searching for relevant combinations may enhance accuracy of inferences. This provides advantages over existing systems that attempt to assign meaning to every component of an utterance (i.e., including out-of-context words and noise words), which results in nearly infinite possible combinations and greater likelihood of confusion. The context determination process may also be self-aware, assigning degrees of certainty to one or more generated hypotheses, where a hypothesis may be developed to account for variations in environmental conditions, speaker ambiguity, accents, or other factors. By identifying a context, capabilities within the context, vocabularies within the context, what tasks are done most often historically in the context, what task was just completed, etc., the context determination process may establish intent from rather meager phonetic clues. Moreover, just as in human-to-human conversation, users may switch contexts at any time without confusion, enabling various context domains to be rapidly selected, without menu-driven dead ends, when an utterance is unambiguous.

[011] According to another aspect of the invention, an exemplary cooperative conversational model may build upon free form voice search, noise tolerance, and context determination to implement a conversational Human-to-Machine interface that reflects human interaction and normal conversational behavior. That is, the cooperative

conversational model enables humans and machines to participant in a conversation with an accepted purpose or direction, with each participant contributing to the conversation for the benefit of the conversation. By taking advantage of human presumptions about utterances that humans rely upon, both as speakers and listeners, a Human-to-Machine interface may be analogous to everyday human-to-human conversation. In one implementation, the exemplary cooperative conversation model may take incoming data (shared knowledge) to inform a decision (intelligent hypothesis building), and then may refine the decision and generate a response (adaptive response building).

[012] According to another aspect of the invention, shared knowledge may include both short-term and long-term knowledge. Short-term knowledge may accumulate during a single conversation, where input received during a single conversation may be retained. The shared knowledge may include cross-modality awareness, where in addition to accumulating input relating to user utterances, requests, locations, etc., the shared knowledge may accumulate a current user interface state relating to other modal inputs to further build shared knowledge models. The shared knowledge may be used to build one or more intelligent hypotheses using current and relevant information, build long-term shared knowledge by identifying information with long-term significance, and generate adaptive responses with relevant state and word usage information. Moreover, because cooperative conversations model human conversations, short-term session data may be expired after a psychologically appropriate amount of time, thereby humanizing system behavior, reducing a likelihood of contextual confusion based on stale data, while also adding relevant information from an expired session context to long-term knowledge models. Long-term shared knowledge may generally be user-centric, rather than session-based, where inputs may be accumulated over time to build user, environmental, cognitive, historical, or other long-term knowledge models. Long-term and short-term shared knowledge may be used simultaneously anytime a user engages in a cooperative conversation. Long-term shared knowledge may include explicit and/or implicit user preferences, a history of recent contexts, requests, tasks, etc., user-specific jargon related to vocabularies and/or capabilities of a context, most often used word choices, or other information. The long-term shared knowledge may be used to build one or more intelligent hypotheses using current and relevant information, generate adaptive responses with appropriate word choices when unavailable via short-term shared knowledge, refine long-term shared knowledge models, identify a frequency of

specific tasks, identify tasks a user frequently has difficulty with, or provide other information and/or analysis to generate more accurate conversational responses. Shared knowledge may also be used to adapt a level of unprompted support (e.g., for novices versus experienced users, users who are frequently misrecognized, etc.) Thus, shared knowledge may enable a user and a voice user interface to share assumptions and expectations such as topic knowledge, conversation history, word usage, jargon, tone, or other assumptions and/or expectations that facilitate a cooperative conversation between human users and a system.

[013] According to another aspect of the invention, a conversation type may be identified for any given utterance. Categorizing and developing conceptual models for various types of exchanges may consistently align user expectations and domain capabilities. One or more intelligent hypotheses may be generated as to a conversation type by considering conversational goals, participant roles, and/or an allocation of information among the participants. Based on the conversational goals, participant roles, and allocation of information, the intelligent hypotheses may consider various factors to classify a conversation (or utterance) into general types of conversations that can interact with one another to form many more variations and permutations of conversation types (e.g., a conversation type may change dynamically as information is reallocated from one participant to another, or as conversational goals change based on the reallocation of information).

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

[015] According to another aspect of the invention, syntactically, grammatically, and contextually sensitive "intelligent responses" may be generated from the intelligent hypotheses that can be used to generate a conversational experience for a user, while

also guiding the user to reply in a manner favorable for recognition. The intelligent responses may create a conversational feel by adapting to a user's manner of speaking, framing responses appropriately, and having natural variation and/or personality (e.g., by varying tone, pace, timing, inflection, word use, jargon, and other variables in a verbal or audible response).

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

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

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

convenient.

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

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

BRIEF DESCRIPTION OF THE DRAWINGS

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

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

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

DETAILED DESCRIPTION

[024] Referring to Fig. 1, an exemplary system architecture for implementing a cooperative conversational voice user interface is illustrated according to one aspect of the invention. The system may receive an input 105 from a user, where in one implementation, input 105 may be an utterance received by an input device (e.g., a microphone), where the utterance may include one or more requests. Input 105 may also be a multi-modal input, where at least part of the multi-modal input is an utterance. For example, the input device may include a combination of a microphone and a touch-screen device, and input 105 may include an utterance that includes a request relating to a portion of a display on the touch-screen device that the user is touching. For instance, the touch-screen device may be a navigation device, and input 105 may include an utterance

of "Give me directions to here," where the user may be requesting directions to a desired destination on the display of the navigation device.

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

[026] Referring to Fig. 2, an exemplary block diagram is provided illustrating a conversational speech engine 215 according to one aspect of the invention. Conversational speech engine 215 may include a conversational language processor 220 that generates an adaptive conversational response to a request or series of requests using a free form voice search module 245, a noise tolerance module 250, and/or a context determination process 255. According to one aspect of the invention, modules 245-255 may communicate with a voice search engine 225 that includes one or more context

domain agents 230 and/or one or more vocabularies 235 to aid in interpreting utterances and generating responses, as described in "Enhancing the VUE™ (Voce-User-Experience) Through Conversational Speech," by Tom Freeman and Larry Baldwin, which is herein incorporated by reference in its entirety. Conversational speech engine 215 may generate an adaptive conversational response to one or more requests, where the requests may depend on unspoken assumptions, incomplete information, context established by previous utterances, user profiles, historical profiles, environmental profiles, or other information. Moreover, conversational speech engine 215 may track which requests have been completed, which requests are being processed, and/or which requests cannot be processed due to incomplete or inaccurate information, and the response may be generated accordingly.

[027] According to one aspect of the invention, free form voice search module 245 may understand an utterance made using typical, day-to-day language (i.e., in free form), and may account for variations in how humans normally speak, the vocabulary they use, and the conditions in which they speak. Because variables such as stress, distraction, and serendipity are always different and infinitely varied, free form search module 245 may be designed with a goal of understanding that no human will come to the same Human-to-Machine interface situation in the same way twice. Thus, free form search module 245 may implement one or more features that model casual human speech. In various implementations, free form search module 245 may include, among other things, a free form utterance feature, a one-step access feature, an inferencing intended operations feature, an alternative expression feature, and/or an imperfect speech feature.

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

cooperating with context determination process 255, or other features, and identifying a relevant context domain agent 230 and/or vocabulary 235 to interpret the utterance).

[029] The one-step access feature may understand utterances that include compound requests with multiple variables. For example, a user utterance may be "What is the forecast for Boston this weekend?" The one-step access feature may identify "weather" as a context (e.g., by cooperating with context determination process 255, or other features, and identifying "forecast" as a synonym of "weather"), and search for a city equal to "Boston" and a time equal to "weekend." By identifying all relevant information for completing a task from a single utterance, the one-step access feature may overcome drawbacks of existing voice user interfaces, such as Command and Control systems that use verbal menus to restrict information that a person can provide at a given point (e.g., a Command and Control system for a phone directory service may say: "State please," . . . "City please," . . . "What listing," etc.). Moreover, some utterances may include compound requests, and the one-step access feature may decompose the compound requests into sub-tasks. For example, a user utterance of "I need to be at a meeting tomorrow in San Francisco at 8:00 am" may be decomposed into a set of sub-tasks such as (1) checking availability and reserving a flight on an evening before the meeting, (2) checking availability and reserving a hotel, (3) checking availability and reserving a car, etc., where users may further designate preferences for various tasks (e.g., first check availability on an airline for which the user is a frequent flyer). Depending on a level of shared knowledge about a user's preferences and/or historical patterns, the one-step access feature may infer additional tasks from a request. For example, in the above example, the one-step access feature may also check a weather forecast, and if the weather is "nice" (as defined by the user preferences and/or as inferred from historical patterns), the one-step access feature may schedule a tee-time at a preferred golf course in San Francisco.

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

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

[031] The alternative expression feature may recognize common alternatives for nouns and verbs to reflect variations in usage patterns according to various criteria. For example, users may vary expression based on age, socio-economics, ethnicity, user whims, or other factors. Thus, the alternative expression feature may support variations in expression where word order is unimportant or unanticipated. Alternatives in expression based on various criteria or demographics may be loaded into context domain agents 230 and/or vocabularies 235, and the alternative expression feature may update context domain agents 230 and/or vocabularies 235 based on inferred or newly discovered variations. In one implementation, conversational speech engine 215 may include a subscription interface to update changes to context domain agents 230 and/or vocabularies 235 (e.g., a repository may aggregate various user utterances and deploy updates system wide). In operation, the alternative expression feature may allow nouns and/or verbs to be represented in different ways to give simplistic, yet representative, examples. For example, a user interested in a weather forecast for Washington, D.C. may provide any of the following utterances, each of which are interpreted equivalently: "What's the weather like in DC," "Is it raining inside the Beltway," Gimme the forecast for the capital," etc. Similarly, utterances of "Go to my home," "Go home," "Show route to home," and "I would like to know my way home" may all be interpreted equivalently, where a user profile may include the user's home address and a navigation route to the home address may be calculated.

[032] The imperfect speech feature may be able to infer requests from contradictory or

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

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

those conditions.

[034] According to one aspect of the invention, conversational speech engine 215 may include a context determination process 255 that determines one or more contexts for a request to establish meaning within a conversation. The one or more contexts may be determined by having one or more context domain agents compete to determine a most appropriate domain for a given utterance, as described in copending U.S. Patent Application Ser. No. 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," and copending U.S. Patent Application Ser. No. 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," both of which are hereby incorporated by reference in their entirety. Once a given context domain agent "wins" the competition, the winning agent may be responsible for establishing or inferring further contexts and updating short-term and long-term shared knowledge. If there is a deadlock between context domain agents, an adaptive conversational response may prompt the user to assist in disambiguating between the deadlocked agents. For example, a user utterance of "What about traffic?" may have a distinct meaning in various contexts. That is, "traffic" may have a first meaning when the user is querying a system's media player (i.e., "traffic" would be a Rock and Roll band led by singer/songwriter Steve Winwood), a second meaning when the user is querying a search interface regarding Michael Douglas films (i.e., "traffic" would be a film directed by Steven Soderbergh), a third meaning when the user is querying a navigation device for directions to an airport (i.e., "traffic" would be related to conditions on roads along a route to the airport).

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

[036] Context determination process 255 may overcome drawbacks of existing systems by continually updating one or more models of an existing context, where establishing

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

[037] The profiles may be particularly meaningful when attempting to disambiguate between contexts where a word has different meanings in different contexts. For example, a user may utter "What's the weather in Seattle?" and context determination process 255 may establish Weather as a context, as well as establishing Seattle as an environmental context. The user may then utter "and Portland?" and context determination process 255 may return a weather report for Portland, Oregon based on the Weather and an environmental proximity between Portland, Oregon and Seattle, Washington. The user may then ask "What time does the game start?"" and a search for sports events with teams from Seattle and/or Portland may occur, with results presented conversationally according to methods described in greater detail below in Fig. 3. Correlatively, had user originally uttered "What's the weather in Portsmouth, New Hampshire," in the second utterance, context determination process 255 may instead retrieve a weather report for Portland, Maine based on an environmental proximity to New Hampshire. Moreover, when environmental profiles, contextual shared knowledge, and/or other short-term and/or long-term shared knowledge does not provide enough information to

disambiguate between possibilities, responses may prompt the user with a request for further information (e.g., "Did you mean Portland, Maine, or Portland, Oregon?").

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

[039] Context domain agents 230 may also be self-aware, assigning degrees of certainty to one or more generated hypotheses, where a hypothesis may be developed to account for variations in environmental conditions, speaker ambiguity, accents, or other factors. Conceptually, context domain agents 230 may be designed to model utterances as a hard-of-hearing person would at a noisy party. By identifying a context, capabilities within the context, vocabularies within the context, what tasks are done most often historically in the context, what task was just completed, etc., a context domain agent 230 may establish intent from rather meager phonetic clues. Moreover, the context stack may be one of a plurality of components for establishing context, and thus not a constraint upon the user. All context domains may be accessible, allowing the user to switch contexts at any time without confusion. Thus, just as in human- to-human conversation, context domains

may be rapidly selected, without menu-driven dead ends, when an utterance is unambiguous. For example, a user may utter, "Please call Rich Kennewick on his cell phone," and a system response of "Do you wish me to call Rich Kennewick on his cell?" may be generated. The user may decide at that point to call Rich Kennewick later, and instead, listen to some music. Thus, the user may then utter, "No, play the Louis Armstrong version of Body and Soul from my iPod," and a system response of "Playing Body and Soul by Louis Armstrong" may be generated as Body and Soul is played through a media player. In this example, the later utterance has no contextual connection to the first utterance, yet because request criteria in the utterances are unambiguous, contexts can be switched easily without relying on the context stack.

[040] Referring to Fig. 3, an exemplary cooperative conversational model 300 is illustrated according to an aspect of the invention. Cooperative conversational model 300 may build upon free form voice search 245, noise tolerance 250, and context determination 255 to implement a conversational Human-to-Machine interface that reflects how humans interact with each other and their normal behavior in conversation. Simply put, cooperative conversational model 300 enables humans and machines to participate in a conversation with an accepted purpose or direction, with each participant contributing to the conversation for the benefit of the conversation. That is, cooperative conversational model 300 incorporates technology and process-flow that takes advantage of human presumptions about utterances that humans rely upon, both as speakers and listeners, thereby creating a Human-to-Machine interface that is analogous to everyday human-to-human conversation. In one implementation, a cooperative conversation may take incoming data (shared knowledge) 305 to inform a decision (intelligent hypothesis building) 310, and then may refine the decision and generate a response (adaptive response building) 315.

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

[042] Input received during a single conversation may be retained in a Session Input Accumulator. The Session Input Accumulator may include cross-modality

awareness, where in addition to accumulating input relating to user utterances, requests, locations, etc., the Session Input Accumulator may accumulate a current user interface state relating to other modal inputs to further build shared knowledge models and more accurate adaptive responses (e.g., when a user utters a request relating to a portion of a touch-screen device, as described above). For example, the Session Input Accumulator may accumulate inputs including recognition text for each utterance, a recorded speech file for each utterance, a list-item selection history, a graphical user interface manipulation history, or other input data. Thus, the Session Input Accumulator may populate Intelligent Hypothesis Builder 310 with current and relevant information, build long-term shared knowledge by identifying information with long-term significance, provide Adaptive Response Builder 315 with relevant state and word usage information, retain recent contexts for use with Intelligent Hypothesis Builder 310, and/or retain utterances for reprocessing during multi-pass evaluations. Moreover, because cooperative conversations 300 model human conversations, short-term session data may be expired after a psychologically appropriate amount of time, thereby humanizing system behavior. For example, a human is unlikely to recall a context of a conversation from two years ago, but because the context would be identifiable by a machine, session context is expired after a predetermined amount of time to reduce a likelihood of contextual confusion based on stale data. However, relevant information from an expired session context may nonetheless be added to user, historical, environmental, cognitive, or other long-term knowledge models.

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

knowledge models, identify a frequency of specific tasks, identify tasks a user frequently has difficulty with, or provide other information and/or analysis to generate more accurate conversational responses.

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

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

[046] For example, in one implementation, a query conversation may include a conversational goal of getting a discrete piece of information or performing a particular task, where a leader of the query conversation may have a specific goal in

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

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

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

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

[047] Intelligent Hypothesis Builder 310 may use an identified conversation type to assist in generating a set of hypotheses as to a user's intent in an utterance. In addition, Intelligent Hypothesis Builder 310 may use short-term shared knowledge from the Session Input Accumulator to proactively build and evaluate interaction with a user as a conversation progresses, as well as long-term shared knowledge to proactively build and evaluate interaction with the user over time. Intelligent Hypothesis Builder 310 may thus adaptively arrive at a set of n-best hypotheses about user intent, and the n-best hypotheses may be provided to an Adaptive Response Builder 315. In addition, Intelligent Hypothesis Builder 310 may model human-to-human interaction by calculating a degree of certainty for each of the hypotheses. That is, just as humans rely on knowledge shared by participants to examine how much and what kind of information was available, Intelligent Hypothesis Builder 310 may leverage the identified conversation type and short-term

and long-term shared knowledge to generate a degree of certainty for each hypothesis.

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

[049] According to another aspect of the invention, Intelligent Hypothesis Builder 310 may use short-term knowledge to generate one or more implicit hypotheses of a user's intent when an utterance may be missing required qualifiers or other information needed to complete a request or task. Each hypothesis may have a corresponding degree of certainty. For instance, when a conversation begins, short-term knowledge stored in the Session Input Accumulator may be empty, and as the conversation progresses, the Session Input Accumulator may build a history of the conversation. Intelligent Hypothesis Builder 310 may use data in the Session Input Accumulator to supplement or infer additional information about a current utterance. For example, Intelligent Hypothesis Builder 310 may evaluate a degree of certainty based on a number of previous requests relevant to the current utterance. In another example, when the current utterance contains insufficient information to complete a request or task, data in the Session Input Accumulator may be used to infer missing information so that a hypothesis can be generated. In still another example, Intelligent Hypothesis Builder 310 may identify syntax and/or grammar to be used by Adaptive Response Builder 315 to formulate personalized and conversational response. In yet another example, when the current utterance contains a threshold amount of information needed to complete a request or task, data in the Session Input Accumulator may be relied upon to tune a degree of certainty.

[050] According to another aspect of the invention, Intelligent Hypothesis Builder 310 may use long-term shared knowledge to generate one or more implicit hypotheses of a user's intent when an utterance is missing qualifiers or other information needed to complete a request or task. Each hypothesis may have a corresponding degree of certainty. Using long-term knowledge may be substantially similar to using short-term shared knowledge, except that information may be unconstrained by a current session,

and an input mechanism may include information from additional sources other than conversational sessions. For example, Intelligent Hypothesis Builder 310 may use information from long-term shared knowledge at any time, even when a new conversation is initiated, whereas short-term shared knowledge may be limited to an existing conversation (where no short-term shared knowledge would be available when a new conversation is initiated). Long-term shared knowledge may come from several sources, including user preferences or a plug-in data source (e.g., a subscription interface to a remote database), expertise of a user (e.g., based on a frequency of errors, types of tasks requested, etc., the user may be identified as a novice, intermediate, experienced, or other type of user), agent-specific information and/or language that may also apply to other agents (e.g., by decoupling information from an agent to incorporate the information into other agents), frequently used topics passed in from the Session Input Accumulator, frequently used verbs, nouns, or other parts of speech, and/or other syntax information passed in from the Session Input Accumulator, or other sources of long-term shared knowledge may be used.

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

grammar, context domain agent specific information and/or preferences (e.g., a travel context domain agent may know a user frequently requests information about France, which may be shared with a movie context domain agent so that responses may occasionally include French movies).

[052] According to another aspect of the invention, Adaptive Response Builder 315 may build syntactically, grammatically, and contextually sensitive "intelligent responses" that can be used with one or more agents to generate a conversational experience for a user, while also guiding the user to reply in a manner favorable for recognition. In one implementation, the intelligent responses may include a verbal or audible reply played through an output device (e.g., a speaker), and/or an action performed by a device, computer, or machine (e.g., downloading a web page, showing a list, executing an application, etc.). In one implementation, an appropriate response may not require conversational adaptation, and default replies and/or randomly selected response sets for a given task may be used.

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

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

Accumulator) and long-term (from one or more profiles) shared knowledge about how a user utters a request, the responses may be modeled using techniques used to recognize requests. Adaptive Response Builder 315 may rate possible responses statistically and/or randomize responses, which creates an opportunity to build an exchange with natural variation and conversational feel. This may be a significant advantage over existing voice user interfaces with incongruous input and output, where the input is "conversational" and the output is "computerese." The following examples may demonstrate how a response may adapt to a user's input word choices and manner of speaking:

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

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

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

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

User	What's the weather like in Dallas?

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

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

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

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

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

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

CLAIMS

What is claimed is:

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

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

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

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

generating a response based on the determined intended meaning.

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

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

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

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

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

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

utterance.

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

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

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

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

12. The method of claim 1, further comprising:

learning that the determined intended meaning is incorrect;

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

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

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

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

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

instructions operable when executed to:

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

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

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

generate a response based on the determined intended meaning.

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

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

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

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

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

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

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

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

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

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

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

learn that the determined intended meaning is incorrect;

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

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

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

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

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

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

one or more processors collectively operable to:

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

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

generate a response based on the determined intended meaning.

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

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

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

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

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

determine the intended meaning by generating at least one hypothesis of an intended meaning for the at least one utterance, the at least one hypothesis including a degree of certainty.

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

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

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

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

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

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

learn that the determined intended meaning is incorrect;

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

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

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

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

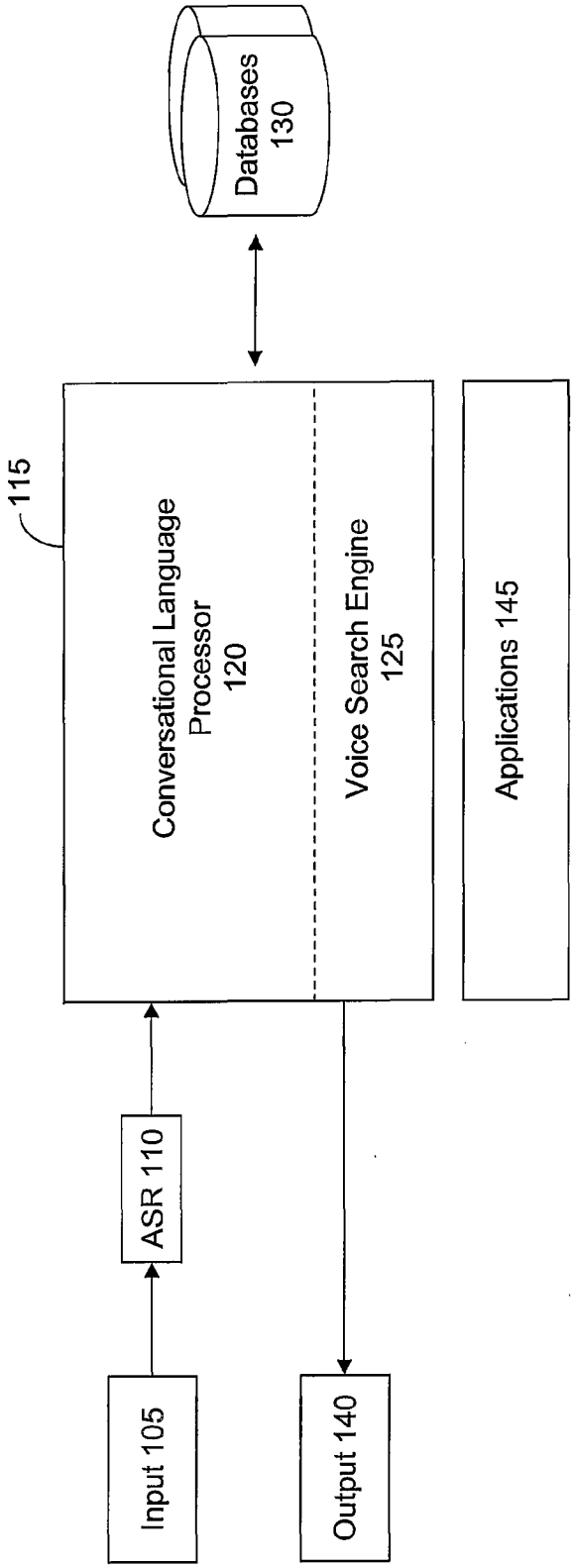

Databases
130
115
Conversational Language
Processor
120
Voice Search Engine
125
Applications 145
ASR 110
Input 105
Output 140

FIG. 1

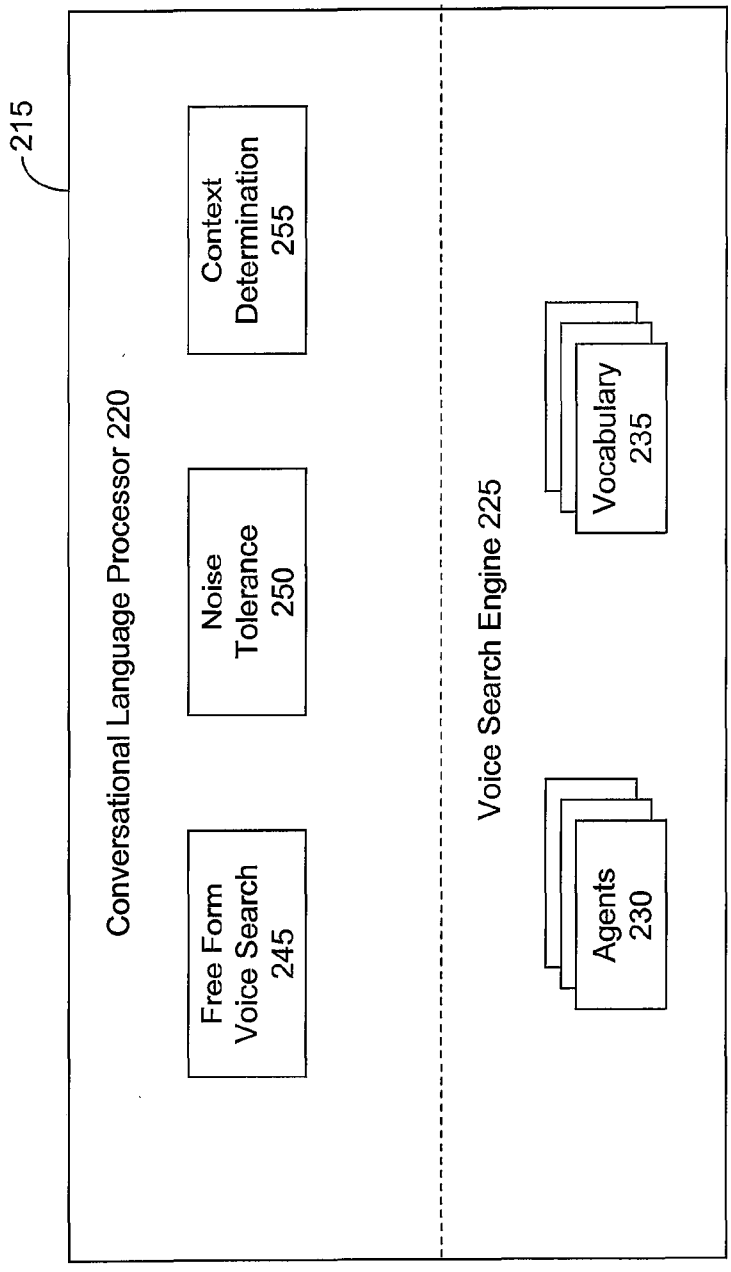

215
Conversational Language Processor 220
Free Form Voice Search 245
Noise Tolerance 250
Context Determination 255
Voice Search Engine 225
Agents 230
Vocabulary 235

FIG. 2

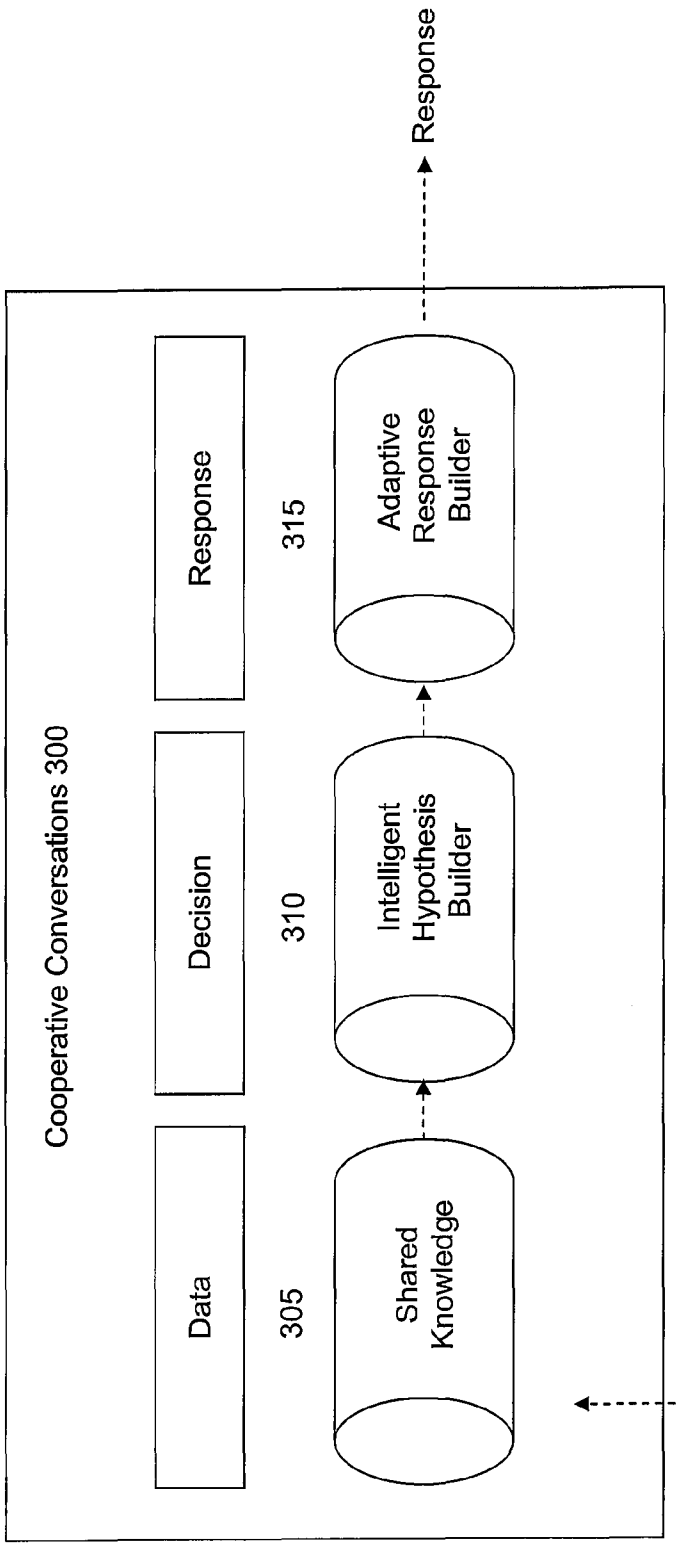

Cooperative Conversations 300
Data
Decision
Response
305
310
315
Shared Knowledge
Intelligent Hypothesis Builder
Adaptive Response Builder
Input
Response

FIG. 3

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 June 2009 (18.06.2009)

PCT

(10) International Publication Number
WO 2009/075912 A1

(51) International Patent Classification:
G06F 17/27 (2006.01)

(21) International Application Number:
PCT/US2008/069508

(22) International Filing Date: 9 July 2008 (09.07.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/954,064 11 December 2007 (11.12.2007) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [US/US]; 11980 NE 24th Street, Bellevue, Washington 98005 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **KENNEWICK, Michael** [US/US]; 12196 NE 24th Street, Bellevue, Washington 98005 (US). **CHEUNG, Catherine** [US/US]; 1639 207th Avenue NE, Sammamish, Washington 98074 (US). **BALDWIN, Larry** [US/US]; 23125 SE 243rd Place, Maple Valley, Washington 98038 (US). **SALOMON, Ari** [US/US]; 4133 96th Avenue SE, Mercer Island, Washington 98040 (US). **TJALVE, Michael** [DK/US]; 14904 SE 47th Court, Bellevue, Washington 98006 (US). **GUTTIGOLI, Sheetal** [IN/US]; 11314 Ohde Circle, Kirkland, Washington 98033 (US). **ARMSTRONG, Lynn** [US/US]; 19807 183rd Place NE, Woodinville, Washington 98077 (US). **DI CRISTO, Philippe** [FR/US]; 177 107th Avenue NE, #1007, Bellevue, Washington 98004 (US). **ZIMMERMAN, Bernie** [US/US]; 135 S. 107th Street, Seattle, Washington 98168 (US). **MENAKER, Sam** [US/US]; 480 150th Place NE, Bellevue, Washington 98007 (US).

(74) Agents: **GATTO, James G.** et al.; Pillsbury Winthrop Shaw Pittman LLP, P.O. Box 10500, McLean, Virginia 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),

[Continued on next page]

(54) Title: PROVIDING A NATURAL LANGUAGE VOICE USER INTERFACE IN AN INTEGRATED VOICE NAVIGATION SERVICES ENVIRONMENT

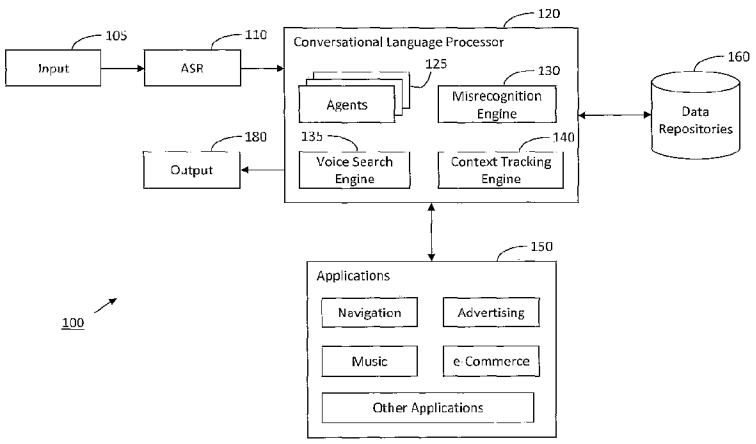


Figure 1

(57) Abstract: A conversational, natural language voice user interface may provide an integrated voice navigation services environment. The voice user interface may enable a user to make natural language requests relating to various navigation services, and further, may interact with the user in a cooperative, conversational dialogue to resolve the requests. Through dynamic awareness of context, available sources of information, domain knowledge, user behavior and preferences, and external systems and devices, among other things, the voice user interface may provide an integrated environment in which the user can speak conversationally, using natural language, to issue queries, commands, or other requests relating to the navigation services provided in the environment.

WO 2009/075912 A1

European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

PROVIDING A NATURAL LANGUAGE VOICE USER INTERFACE IN AN INTEGRATED VOICE NAVIGATION SERVICES ENVIRONMENT

FIELD OF THE INVENTION

[001] The present invention relates to a natural language voice user interface that facilitates cooperative, conversational interactions in an integrated voice navigation services environment, and in particular, to a natural language voice user interface in which users can request navigation services using conversational, natural language queries or commands.

BACKGROUND OF THE INVENTION

[002] As technology advances, consumer electronics tend to play increasingly significant roles in everyday life. As a result, users tend to expect greater functionality, mobility, and convenience from their electronic devices, as exemplified by modern mobile phones, navigation devices, personal digital assistants, portable media players, and other devices often providing a wealth of functionality beyond core applications. However, the greater functionality often tends to be accompanied by significant learning curves and other barriers that prevent users from fully exploiting device capabilities (e.g., features may often be buried within difficult to navigate menus or interfaces). Moreover, although increasing demand for mobility magnifies the need for simple on-the-go device interaction mechanisms, existing systems often have complex human to machine interfaces. For example, existing human to machine interfaces tend to primarily utilize various combinations of keyboards, keypads, point and click techniques, touch screen displays, or other interface mechanisms. However, these interfaces may often be unsuitable for mobile or vehicular devices (e.g., navigation devices), as they tend to be cumbersome in environments where speed of interaction and dangers of distraction pose significant issues. As such, existing systems often fall short in providing simple and intuitive interaction mechanisms, potentially inhibiting mass-market adoption for certain technologies. As such, there is an ever-growing demand for ways to exploit technology in intuitive ways.

[003] In response to these and other problems, various existing systems have turned to voice recognition software to simplify human to machine interactions. For example, voice recognition software can enable a user to exploit applications and features of a device that may otherwise be unfamiliar, unknown, or difficult to use. However, existing voice user interfaces, when they

actually work, still require significant learning on the part of the user. For example, existing voice user interfaces (e.g., command and control systems) often require users to memorize syntaxes, words, phrases, or other keywords or qualifiers in order to issue queries or commands. Similarly, when users may be uncertain of exactly what to request, or what a device may be capable of, existing systems cannot engage with the user in a productive, cooperative, natural language dialogue to resolve requests and advance conversations. Instead, many existing speech interfaces force users to use predetermined commands or keywords to communicate requests in ways that systems can understand. By contrast, cognitive research on human interaction demonstrates that a person asking a question or giving a command typically relies heavily on context and shared knowledge of an answering person. Similarly, the answering person also tends to rely on the context and shared knowledge to inform what may be an appropriate response. However, existing voice user interfaces do not adequately utilize context, shared knowledge, or other similar information to provide an environment in which users and devices can cooperate to satisfy mutual goals through conversational, natural language interaction.

[004] Furthermore, demand for global positional systems and other navigation-enabled devices has grown significantly in recent years. Navigation devices often tend to be used while a user may be driving, on-the-go, or in other environments where having a hands-free interface provides critical advantages. For example, a user may want to avoid being distracted by looking away from the road, yet the user may also want to interact with a navigation device, for example, to calculate a route to a destination, recalculate the route in response to traffic, find a local restaurant, gas station, or other point of interest, or perform another navigation related task. In these and other instances, efficiently processing a natural language voice-based input could enable the user to interact with the navigation device in a safer, simpler, and more effective way. However, existing systems often fall short in providing an integrated, conversational, natural language voice user interface that can provide such advantages in navigation and other mobile environments.

[005] Existing systems suffer from these and other problems.

SUMMARY OF THE INVENTION

[006] According to various aspects of the invention, various problems associated with existing systems may be addressed by a conversational, natural language voice user interface that provides an integrated voice navigation services environment.

[007] According to various aspects of the invention, the natural language voice user interface may resolve voice requests relating to navigation (e.g., calculating routes, identifying locations, displaying maps, etc.). The navigation application can provide a user with interactive, data-driven directions to a destination or waypoint, wherein the user can specify the destination or waypoint using free-form natural language (e.g., the user can identify full or partial destinations, including a specific address, a general vicinity, a city, a name or type of a place, a name or type of a business, a name of a person, etc.). As free form voice destination inputs may be provided in many different forms, post-processing may be performed on full or partial voice destination inputs to identify a suitable destination address for calculating a route (e.g., a closest address that makes "sense"). For example, an utterance containing a full or partial destination may be analyzed to identify one or more probable destinations (e.g., an N-best list of destinations). The N-best list may be post-processed to assign weights or rankings to the probable destinations (e.g., based on a degree of certainty that a given probable destination corresponds to an intended destination). Thus, a route may be calculated from a user's current location to a most heavily weighted one of the probable destinations in the N-best list. Further, when a voice destination entry includes a partial destination, a final destination may be successively refined over one or more subsequent voice destination entries. The navigation application may also provide dynamic, data-driven directions or routing to a destination. For instance, the navigation application may access data associated with various user-specific and environmental data sources to provide personalized data-driven directions along a route, which can be recalculated or modified based on information taken from the data sources. As such, data may be obtained dynamically to identify alternate routes, recalculate routes, or otherwise provide routing services. Further, possible answers or responses to a given utterance may be filtered according to a current route.

[008] According to various aspects of the invention, the natural language voice user interface may dynamically generate and/or load recognition grammars for interpreting what was said in an utterance (e.g., content of the utterance). Information contained in the dynamic recognition grammars may be used by a navigation agent, an Automatic Speech Recognizer, a context stack,

or various other components in the voice user interface that use grammar information. By efficiently generating, updating, loading, extending, or otherwise building dynamic grammars based on various factors, processing bottlenecks can be avoided, conflicts can be reduced, and other aspects of interpreting an utterance using a recognition grammar can be optimized. For example, a size of a generated grammar may be constrained by an amount of resources available in a system (e.g., in embedded devices or other devices having low amounts of dynamic memory, the constrained grammar size may limit a quantity of resources to be occupied). In another example, the size of the dynamic grammar can be reduced by eliminating redundant keywords, criteria, or other information available in the context stack, the shared knowledge, or other local sources. Thus, favorability of correct interpretations may be improved by reducing perplexity in the grammar (e.g., when two or more elements may likely be confused, one or more of the elements may be eliminated to reduce confusion).

[009] According to various aspects of the invention, the natural language voice user interface may generate dynamic recognition grammars using techniques of geographical chunking. A user's location can be determined at a given moment to determine one or more geographic proximities, which can be used to form an appropriate topological domain for the grammar. For example, the topological domains may reflect physical proximities (e.g., a distance from a current location), civil organization proximities (e.g., regions, states, cities, neighborhoods, subdivisions, localities, etc.), temporal proximities (e.g., amounts of travel time from the current location), directional proximities (e.g., based on directional travel vectors), or various combinations thereof. As a result, by mapping the user's geographic proximities to one or more topological domains, dynamic grammars may be pruned, extended, swapped in or out of memory, or otherwise generated and/or loaded to provide optimal recognition based on location, time, travel, or other factors (e.g., information may be swapped in and out of a grammar as a user moves from one area to another, ensuring that system resources utilize information presently relevant to a given location).

[010] According to various aspects of the invention, the natural language voice user interface may include dynamic grammars formed from one or more topological domains, which may be subdivided into a plurality of tiles, which may be further subdivided into a plurality of subtiles. Thus, information used to build the dynamic grammar can be subdivided or weighed in various ways to determine what information should be included in the grammar. Moreover,

geographical chunks based on physical, civil organization, temporal, directional, or other proximities may be extended into other domains in which a topological taxonomy can be placed. As a result, in addition to having relevance in navigation or other location-dependent systems, the geographical chunking techniques can be applied in other contexts or domains in which geography or location may be relevant. Further, a server operably coupled to the voice user interface may analyze various forms of information to build or refine a source of grammar information. For example, when various devices communicate with the server, information may be communicated to the server may be used to update proximities, topological domains, tiles, subtiles, peer-to-peer affinities, or other grammar information.

[011] According to various aspects of the invention, the natural language voice user interface may calculate routes, provide dynamic data-driven directions to a destination, provide dynamic routing to a destination, perform post-processing of full or partial destination entries, or otherwise provide various voice navigation services. Further, destinations and/or routes may be identified using techniques of successive refinement for voice destination entries, wherein context, agent adaptation, and shared knowledge, among other things, can help a user to narrow down a final destination using voice commands, multi-modal commands, or various combinations thereof. However, it will be apparent that the successive refinement techniques can be applied to various tasks in which generalized approximations can be successively refined through voice or multi-modal commands to narrow down information sought by a user, including various other domains, contexts, applications, devices, or other components that employ the techniques described herein.

[012] According to various aspects of the invention, the natural language voice user interface may enable successive refinement of a final destination by progressively narrowing the final destination. For example, successively refining the destination may be modeled after patterns of human interaction in which a route or a destination may be narrowed down or otherwise refined over a course of interaction. For example, a user may generally approximate a destination, which may result in a route being calculated along a preferred route to the approximated destination. While en route to the approximated destination, the user and the voice user interface may cooperatively refine the final destination through one or more subsequent interactions. Thus, a user may provide a full or partial destination input using free form natural language, for example, including voice commands and/or multi-modal commands. One or more

interpretations of a possible destination corresponding to the voice destination input may be organized in an N-best list. The list of possible destinations may be post-processed to assign weights or ranks to one or more of the entries therein, thus determining a most likely intended destination from a full or partial voice destination input. Thus, the post-processing operation may rank or weigh possible destinations according to shared knowledge about the user, domain-specific knowledge, dialogue history, or other factors. As a result, the full or partial destination input may be analyzed to identify an address to which a route can be calculated, for example, by resolving a closest address that makes "sense" relative to the input destination. Subsequent inputs may provide additional information relating to the destination, and the weighted N-best list may be iteratively refined until the final destination can be identified through successive refinement. As a result, when a suitable final destination has been identified, the route to the final destination may be completed.

[013] According to various aspects of the invention, the natural language voice user interface may include one or more advertising models for generating and/or detecting events relating to location dependent advertisements for navigation systems (e.g., as generated by a local or remote advertising engine, or via a data channel, or in other ways). For example, navigation systems typically include various mechanisms for determining a current location (e.g., a global positioning system, a radio frequency identification system, a system that determines location based on a distance to an identifiable wireless tower or access point, etc.). The location detection system may thus detect information associated with a radio frequency identifier over a data channel used by a marketer to provide advertisements. The marketer may broadcast the advertisement via the data channel, such that the navigation system triggers an event when within a suitable proximity of the RFIDs. Thus, information associated with the event may be filtered according to the current routing information or other contextual parameters to determine what action should be taken in response thereto. In other instances, advertisements may be uploaded to a server by one or more advertising partners, wherein the uploaded advertisements may be associated with metadata or other descriptive information that identifies a target audience, location-dependent information, or other criteria. In another example, a plurality of advertisements may be stored locally at the voice user interface, and an inferencing engine may determine appropriate circumstances in which an event should be generated to deliver one or more of the advertisements to a user. As a result, it will be apparent that advertising events may

be generated in a number of ways, and may be generated and/or detected locally, remotely, by detecting RFIDs, or in other ways.

[014] According to various aspects of the invention, the natural language voice user interface may track user interactions with delivered advertisements. In this way, affinity based models may be generated, for example, to ensure that promotions or advertisements will be delivered to a likely target audience. Thus, an event relating to a given advertisement may be generated and/or detected when shared knowledge about a user's behavior, preferences, or other characteristics match one or more criteria associated with peer-to-peer affinities associated with the advertisement. In other examples, an advertising model may include mobile pay-per-use systems, peer-to-peer local guides or recommendations, or other forms of advertising. Additionally, various aspects of the advertising model, such as the local guides and recommendations, may be generated according to a mapping applied to various topological domains. For example, certain types of advertisements may be dependent on geographic or topological characteristics, and such advertisements may be associated with a topological taxonomy based on geographical chunks. As a result, various advertising events may be generated and/or detected according to physical proximities, temporal proximities, directional proximities, civil organization proximities, or various combinations thereof.

[015] According to various aspects of the invention, the natural language voice user interface may enable a user to provide requests (e.g., queries, commands, or other requests) to the navigation device using natural language. As such, the user and the navigation device may engage in a cooperative, conversational dialogue to resolve the request. For example, the voice user interface may use prior context, dialogue histories, domain knowledge, short and long-term shared knowledge relating to user behavior and preferences, noise tolerance, and cognitive models, among various other things, to provide an integrated environment in which users can speak conversationally, using natural language, to issue queries, commands, or other requests that can be understood and processed by a machine. Accordingly, the voice user interface may understand free form human utterances, freeing the user from restrictions relating to how commands, queries, or other types of requests should be formulated. Instead, the user can use a natural or casual manner of speaking to request various voice services in an integrated environment, in which various devices can be controlled in a conversational manner, using natural language. For example, the voice user interface may be aware of data and services

associated with the navigation device, a media device, a personal computer, a personal digital assistant, a mobile phone, or various other computing devices or systems available in the environment.

[016] According to various aspects of the invention, the natural language voice user interface may include an input mechanism that receives a voice-based input, which includes at least an utterance or verbalization spoken by a user. The input mechanism may include a suitable device or combination of devices that can receive voice-based inputs (e.g., a directional microphone, an array of microphones, or other devices that encode speech). The input mechanism can be optimized to maximize gain in a direction of a user, cancel echoes, null point noise sources, perform variable rate sampling, filter out background conversations or environmental noise, or otherwise optimize fidelity of encoded speech. As such, the input mechanism may generate encoded speech generated in a manner that tolerates noise or other factors that could otherwise interfere with interpreting speech. Further, in various implementations, the input mechanism may include one or more other (non-voice) input modalities, which can be processed and/or correlated with one or more previous, current, or subsequent utterances or other voice-based inputs. As such, a user can provide other forms of input using a touch-screen interface, a stylus/tablet interface, a keypad or keyboard, or other input interfaces, for example, to clarify utterances or provide additional information about the utterances using other input modalities. For instance, the user could touch a stylus or other pointing device to a portion of a map displayed on a touch-screen interface, while also providing an utterance relating to the touched portion (e.g., "Show me restaurants around here."). In this example, the inputs can be correlated to interpret "around here" as likely referring to the touched portion of the map, as distinct from the user's current location or some other meaning.

[017] According to various aspects of the invention, the natural language voice user interface may include an Automatic Speech Recognizer that processes encoded speech to generate one or more preliminary interpretations of what was said in an utterance (e.g., content of the utterance). For example, the Automatic Speech Recognizer may generate the preliminary interpretations using phonetic dictation to recognize a stream of phonemes based on a dynamically adaptable recognition grammar. The dynamically adaptable recognition grammar may be based on dictionaries or phrases from various input domains (e.g., domains for languages, navigation, music, movies, weather, various temporal or geographic proximities, or various other domains).

Thus, the Automatic Speech Recognizer may generate one or more interpretations of an utterance, which may be represented as a series of phonemes or syllables. The one or more interpretations can be analyzed (e.g., utilizing phonotactic rules or models of human speech) to generate an N-best list of preliminary interpretations as to what was spoken by the user. The preliminary interpretations may then be provided to a conversational language processor, which utilizes shared knowledge, contextual information, and various other sources of information to generate an intelligent hypothesis as to an actual meaning, the user's intent, or other aspects of the utterance. By formulating the hypothesis using various features and components that model everyday human-to-human conversations, the conversational language processor may generate a hypothesis as to the meaning or intent of the utterance, which can inform a process of resolving one or more requests contained in the utterance.

[018] According to various aspects of the invention, the natural language voice user interface may include, among other things, a context tracking engine that establishes meaning for a given utterance. For example, the context tracking engine can manage competitions among one or more context-specific domain agents that establish the meaning (e.g., redistributable, adaptable engines or modules that provide functionality for a given context, domain, system, or application). The domain agents may analyze preliminary interpretations of an utterance to generate a domain-specific probable interpretation. For example, one or more of the agents may include adaptable vocabularies, concepts, available tasks, or other forms of information specific to the respective domain or context. In addition, the agents can use a voice search engine to search a network for information that may not be available within the system. The probable interpretation can be assigned a weighted ranking or score, which can be used to select a "winning" one of the agents. Thus, the winning one of the agents may be designated responsible for establishing or inferring further information (e.g., based on domain or context-specific information), updating the shared knowledge, or resolving requests in the utterance, among other things. The context tracking engine may also maintain a context stack to track conversation topics, track previously invoked agents, evaluate criteria, weigh parameters, or otherwise maintain information relating to a conversational context (e.g., the context stack may be traversed in light of recent contexts, frequently used contexts, or other information included therein to determine a most likely intent of the user). By identifying a context, or correlatively, by identifying capabilities, tasks, vocabularies, or other information within the context, the

context tracking engine can provide relevant information for establishing intent in addition to phonetic clues associated with the utterance (e.g., a word with multiple possible meanings may be disambiguated based on its meaning in a given context, previous usage in a dialogue, etc.).

[019] According to various aspects of the invention, the natural language voice user interface may utilize various forms of information to enable sharing of assumptions and expectations relating to a given utterance, conversation, or other human to machine interaction. For example, to inform decision making in the voice user interface, the voice user interface may include information sources containing short-term and long-term shared knowledge relating to user behavior, preferences, or other characteristics (e.g., short-term and long-term profiles of a specific user, global users, peer users, etc.). The short-term shared knowledge may accumulate information during a current conversation to dynamically establish awareness of a state of the voice user interface (e.g., recognition text for previous utterances, a cross-modal user interface manipulation history, a list of previously selected tasks or invoked queries, or other information). Storage of the short-term knowledge may be modeled after human interaction, such that certain information may be expired after a psychologically appropriate amount of time (e.g., to expunge stale data), whereas information with long-term significance can be added to the long-term shared knowledge (e.g., to establish persistent awareness of data likely to remain static over time). As such, the long-term shared knowledge may profile or otherwise model various characteristics, preferences, behavior, or other information relating to a user based on information accumulated over time (e.g., user-specific jargon, demographics, cognitive patterns, frequently requested tasks, favorite topics or concepts, etc.). Accordingly, the voice user interface may utilize various forms of information available via a context tracking engine, domain agents, a voice search engine, shared knowledge, internal or external databases, data associated with other devices, or other knowledge sources. As a result, a conversational type or goal associated with utterance can be identified. Based on the available information and the type or goal of the conversation, the voice user interface may attempt to resolve the utterance (e.g., by invoking an agent, which utilizes one or more applications to perform a requested task, retrieve requested information, etc.).

[020] According to various aspects of the invention, the natural language voice user interface may generate cross-modal intelligent responses sensitive to syntax, grammar, and context, which may provide a conversational feel to system-generated responses. When available, a generated

intelligent response may present results of a resolved request (e.g., feedback relating to a task, information retrieved as a result of a query, etc.). Further, the intelligent responses may be provided across modalities, for example, using various combinations of verbal and/or non-verbal outputs (e.g., information may be presented on a display device, via an audible alert, a verbalized output, another output mechanism, or various combinations thereof). Further, verbalized components of the cross-modal response may be adapted to the user's manner of speaking (e.g., by varying tone, pace, timing, or other variables), thus creating verbal responses having natural variation and personality. The intelligent responses may also be formulated to provide an output that guides the user towards a subsequent response (e.g., a subsequent utterance) that may be more likely to be recognized. For example, when an utterance cannot be resolved because of an ambiguous context, unrecognizable words or phrases, or other factors that can result in an ambiguous or unrecognized interpretation, the intelligent response can be framed to disambiguate context or request additional information from the user to clarify a meaning of the utterance. Further, when subsequent information indicates that a given interpretation or hypothesis was incorrect, one or more previous utterances may be reinterpreted to refine context and update short-term or long-term conversational models (e.g., an utterance of "No, I meant . . ." may indicate that a previous utterance was interpreted incorrectly, in which case various previous utterances may be reinterpreted based on a correct interpretation, thus building a more accurate conversational context). As a result, the voice user interface may tolerate full or partial failure through adaptive mechanisms.

[021] According to various aspects of the invention, the natural language voice user interface may provide voice navigation services within an agent-based architecture. The architecture may include a plurality of adaptable agents (e.g., specialized software, data, content, or other information that provide functionality, behavior, services, data, and other information in a plurality of respective contextual domains), at least one of which provides navigation services (e.g., route calculation, map control, location-sensitive information, data-driven directions, etc.). As the agents process requests, the agents may autonomously react, adapt, and otherwise reconfigure to provide optimal voice services in respective domains. For example, by building context over time (e.g., by generating short and long-term profiles of a user, conversations with the user, frequent topics or preferences, etc.), the agents may automatically combine knowledge, adapt preferences, remove conflicts, or perform other adaptations to refine or otherwise optimize

an operational framework thereof. Adaptation of the agents, which include at least a navigation agent, may occur across the plurality of agents, for example, in response to various ones of the agents resolving voice-based requests. As such, the adaptation may occur autonomously as a by-product of the agents providing voice services, generating inferences, identifying affinities (e.g., among users, peers, communities, etc.), receiving updates from external sources (e.g., using an update manager), or in other ways, as will be apparent.

[022] According to various aspects of the invention, the natural language voice user interface may include an agent-based architecture for providing voice navigation services. For example, the agent-based architecture may include one or more domain or context-specific agents, which include at least a navigation agent. The navigation agent may include, among other things, various navigation-specific content packages (e.g., dictionaries, available queries, tasks, commands, dynamic grammars, etc.), response lists (e.g., appropriate responses to commands, queries, or other requests), personality profiles (e.g., for creating a natural feel to system-generated speech), substitution lists (e.g., for substituting or transforming data into a structured form that can be understood by a target information source), or various other forms of navigation-specific information. Further, the navigation agent may be associated with pointers to local or remote data sources, parameters and operating data provided associated with other services in the architecture, or various other forms of information. For example, the data sources used by the navigation agent may include data relating to navigation, points-of-interest, traffic, events, parking, personal data, peer affinities, or various other sources of information. Further, the data sources may be populated, extended, pruned, or otherwise constructed through adaptation, analysis of various models, communication with a data service, or in other ways, as will be apparent.

[023] According to various aspects of the invention, the natural language voice user interface may include a navigation agent, which may be coupled with various sources of information, and may make use of context, communicating with various other adaptable agents and other system components to provide voice navigation services. For example, the navigation agent may use contextual information relating to a navigation domain, including tracked topics, user locations, routes traveled, previous requests, user interface states, user behaviors, preferences, demographics, or other characteristics, or various other types of contextual information. As a result, the navigation agent may have various sources of knowledge and resources available to

resolve voice navigation requests. For example, the navigation agent may generate inferences (e.g., using an inference engine) using the available knowledge and resources to apply various rules, policies, or other inferencing techniques to generate interpretations of an utterance (e.g., phonctic fuzzy matching, inductive logic, Bayesian probability analysis, monotonic or non-monotonic reasoning, etc.). As such, the navigation agent can infer keywords or criteria not explicitly provided in the utterance, determine suitable responses to subjective or indeterminate utterances, generate events, identify peer affinities, or otherwise generate inferences for resolving navigation-related requests.

[024] According to various aspects of the invention, the natural language voice user interface may include one or more inference engines, which can generate various inferences through awareness of previous context, short-term or long-term shared knowledge, command histories, states of vehicular systems, user interface states, and various other data sources. In various implementations, one or more of the agents may be associated with a respective inference engine that can generate inferences using domain-specific knowledge, rules, policies, or other criteria. For instance, the inference engines may identify keywords or criteria missing in an utterance, infer intended meanings, autonomously suggest available tasks, or otherwise assist an associated agent in identifying queries, commands, or other requests contained in an utterance. Moreover, when information cannot be suitably resolved using information sources associated with the navigation agent, or through generating inferences, the information may be requested from one or more other agents, other devices, network information sources (e.g., via a voice search engine), or in other ways, as will be apparent. Upon identifying the information through one or more of the other sources, the requesting agent may be adapted to make the information subsequently available. Thus, various devices, applications, systems, and other components of an architecture may cooperatively share available information and services (e.g., context, dialogue histories, shared knowledge, maps, points of interest, contact lists, user or peer affinities, dynamic grammars, available applications, command histories, etc.). Accordingly, the architecture may provide an integrated voice navigation services environment in which users can speak natural language requests relating to various available contexts, domains, applications, devices, information sources, or various combinations thereof.

[025] According to various aspects of the invention, the natural language voice user interface may accept natural language voice-based inputs to control an electronic device that can provide

navigational information, in addition to various other devices associated with an environment in which the voice user interface operates. Furthermore, various functional aspects of the voice user interface may reside at a client device, at a server, or various combinations thereof.

[026] According to various aspects of the invention, the natural language voice user interface may support multi-modal voice inputs. Thus, a given multi-modal voice input may include at least a voice component (e.g., an utterance) in addition to one or more non-voice input components (e.g., inputs provided via a keypad, a touch-screen, a stylus/tablet combination, a mouse, a keyboard, or other input modalities). As such, the non-voice input component can provide additional information or clarification relating to the utterance, adding to an amount of input information available when processing voice. For example, the user can use other input modalities to clarify a meaning of the utterance, provide additional information about the utterance, reduce a number of device interactions needed to make a given request, or otherwise provide additional in relation to a given utterance.

[027] According to various aspects of the invention, the natural language voice user interface may utilize various of cognitive models, contextual models, user-specific models, or other models to identify queries, commands, or other requests in a voice input. For example, a given input may include information relating to one or more contextual domains, one or more of which may be invoked to interpret and/or infer keywords, concepts, or other information contained in the input. Moreover, short-term and long-term shared knowledge about a user's behavior and preferences may be used in a hybrid recognition model that also considers semantic analysis and contextual inferences. For example, certain syllables, words, phrases, requests, queries, commands, or other information may be more likely to occur in a given context. Thus, the hybrid recognition model may analyze semantic patterns to resolve what was said by an utterance, and may further rely on contextual history or other information to resolve what was meant by the utterance. The hybrid recognition model may be used in conjunction with, or independently of, a peer to peer recognition model. For example, recognition models may include awareness of global usage patterns, preferences, or other characteristics of peer users, where certain keywords, concepts, queries, commands, or other aspects of a contextual framework may employed by peer users within the context.

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

BRIEF DESCRIPTION OF THE DRAWINGS

[029] Figure 1 illustrates a block diagram of an exemplary system for providing a navigation device with a conversational, natural language voice user interface according to various aspects of the invention.

[030] Figure 2 illustrates a block diagram of exemplary agent-based architecture for providing a navigation device with a conversational, natural language voice user interface according to various aspects of the invention.

[031] Figure 3 illustrates a flow diagram of an exemplary method for dynamically generating recognition grammars for a navigation device having a conversational, natural language voice user interface according to various aspects of the invention.

[032] Figure 4 illustrates a flow diagram of an exemplary method for processing multi-modal voice-based inputs provided to a navigation device having a conversational, natural language voice user interface according to various aspects of the invention.

[033] Figure 5 illustrates a flow diagram of an exemplary method for calculating routes based on voice-based inputs provided to a navigation device having a conversational, natural language voice user interface according to various aspects of the invention.

[034] Figure 6 illustrates a flow diagram of an exemplary method for providing voice services based on a current navigation route according to various aspects of the invention.

DETAILED DESCRIPTION

[035] According to various aspects of the invention, Figure 1 illustrates a block diagram of an exemplary system 100 for providing a navigation device with a conversational, natural language voice user interface. As illustrated in Figure 1, the system 100 may enable a user to engage in a natural language, cooperative, conversational dialogue with a voice-enabled navigation device. For example, the system 100 may understand free form human utterances, freeing the user from being restricted in how commands, queries, or other types of requests should be formulated. Instead, the user can use a casual or natural manner of speaking to request various voice navigation services, among various other voice services (e.g., services relating to telematics, communications, media, messaging, external systems, marketing, information retrieval, or various other computational services). As such, the user can use system 100 to control

navigation devices, media devices, personal computers, personal digital assistants, wireless phones, or other computing devices or systems in a conversational manner, using natural language. By making use of context, prior information, domain knowledge, short and long-term shared knowledge, noise tolerance, and cognitive models, among other things, system 100 can provide an integrated environment in which users can speak conversationally, using natural language, to issue queries, commands, or other requests that can be understood and processed by a machine. For example, in various implementations, system 100 may utilize one or more techniques as described in co-pending U.S. Patent Application 10/452,147, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed June 3, 2003, and co-pending U.S. Patent Application 10/618,633, entitled "Mobile Systems and Methods for Responding to Natural Language Speech Utterance," filed June 15, 2003, the disclosure of which are hereby incorporated by reference in their entirety.

[036] According to various aspects of the invention, system 100 may include an input mechanism 105 that receives a voice-based input, including at least one utterance or verbalization spoken by a user. The input mechanism 105 may include a device or combination of devices suitable for receiving a voice-based input (e.g., a directional microphone, an array of microphones, or other devices that encode speech). In various implementations, the input mechanism 105 can be optimized to receive the voice-based input, for example, by maximizing gain in a direction of a user, cancelling echoes, nulling point noise sources, performing variable rate sampling, filtering out background conversations or environmental noise, or performing various other techniques for maximizing fidelity of encoded speech. As such, the encoded speech generated by input mechanism 105 may be highly tolerant of noise or other factors that may potentially interfere with interpreting human speech.

[037] Further, in various implementations, the input mechanism 105 may be coupled to other input modalities, in that various forms of input other than voice can be processed and/or correlated with one or more previous, current, or subsequent voice-based inputs. For example, the input mechanism 105 may be coupled to a touch-screen interface, a stylus/tablet interface, a keypad or keyboard, or other devices or system interfaces, as will be apparent. As a result, an amount of input information potentially available for system 100 to process voice may be maximized, as the user can clarify utterances or provide additional information about utterances using other input modalities. For instance, the user could touch a stylus or other pointing device

to a portion of a map displayed on a touch-screen interface, while also providing an utterance relating to the touched portion (e.g., "Show me restaurants around here."). In this example, system 100 may correlate the inputs to interpret "around here" as likely referring to the touched portion of the map, as distinct from the user's current location or another meaning.

[038] System 100 may also include an Automatic Speech Recognizer 110 that receives the encoded voice input and generates one or more preliminary interpretations thereof. For example, the Automatic Speech Recognizer 110 may recognize the voice-based input using phonetic dictation to recognize a stream of phonemes based on a dynamically adaptable recognition grammar. As a result, the Automatic Speech Recognizer 110 may provide out-of-vocabulary capabilities, which may be tolerant of a user misspeaking, portions of a speech signal being dropped, or other factors that could interfere with interpreting an utterance. The dynamically adaptable recognition grammar may be based on dictionaries or phrases from various input domains (e.g., domains for different languages, navigation, music, movies, weather, various temporal or geographic proximities, or various other domains). Further, performance of the Automatic Speech Recognizer 110 may be improved, for example, by pruning a search space associated with the recognition grammar (e.g., the grammar can include a linking element, such as *schwa*, to represent an unstressed, central vowel that tends to be spoken frequently, even without the user's conscious awareness). Thus, using these and other techniques, Automatic Speech Recognizer 110 may analyze an incoming encoded utterance to represent portions of the utterance as a series of phonemes or syllables, which can be further broken down into core components of an onset, a nucleus, and a coda, among other sub-categories. The series of phonemes or syllables can then be analyzed (e.g., utilizing phonotactic rules that model human speech) to identify a plurality of preliminary interpretations or best guesses (e.g., an N-best list) as to what was actually spoken by the user. It will be apparent, however, that the Automatic Speech Recognizer 110 may use various techniques to generate the preliminary interpretations of the encoded utterance, including those described, for example, in co-pending U.S. Patent Application Ser. No. 11/513,269, entitled "Dynamic Speech Sharpening," filed August 31, 2006, the disclosure of which is hereby incorporated by reference in its entirety.

[039] The plurality of preliminary interpretations generated by the Automatic Speech Recognizer 110 may be provided to a conversational language processor 120, which utilizes shared knowledge to generate an intelligent hypothesis of an actual meaning, a user's intent, or

other aspect of the voice-based input. The conversational language processor 120 may formulate the hypothesis using various features and components that collectively operate to model everyday human-to-human conversations.

[040] For example, the conversational language processor 120 may include a context tracking engine 140 that establishes meaning for a given utterance by, among other things, managing a competition among one or more context-specific domain agents 125 (e.g., redistributable, adaptable packages or modules that provide functionality for a given context, domain, system, or application). For example, the preliminary interpretations may be ranked by the Automatic Speech Recognizer 110, and the agents 125 may further analyze the preliminary interpretations to generate a weighted ranking or score, which can be used to select a "winning" one of the agents 125. The winning one of the agents 125 may then be responsible for establishing or inferring further information, updating the shared knowledge, or performing other tasks to aid in generating the intelligent hypothesis. Moreover, the context tracking engine 140 may use a context stack to track conversation topics, evaluate criteria, weigh parameters, or otherwise maintain contextual information for generating the hypothesis based on a conversational context (e.g., a context stack may be traversed in light of recent contexts, frequently used contexts, or other information included therein to determine a most likely intent of the user). By identifying a context, or correlatively, by identifying capabilities, tasks, vocabularies, or other information within the context, the context tracking engine 140 can provide highly relevant information for establishing intent, apart from meager phonetic clues.

[041] Furthermore, the conversational language processor 120 may utilize various other forms of knowledge to inform the generation of the intelligent hypothesis. For example, various agents 125 may adaptively include domain-specific or context-specific vocabularies, concepts, available tasks, or other forms of information relevant to the respective domain or context. In addition, the various components associated with the conversational language processor 120 can invoke a voice search engine 135 (e.g., an engine that searches a network for information) to resolve information that may not be internally available (e.g., when an external knowledge source may be useful in clarifying an intent behind a particular word, command, query subject, or other aspect of an utterance).

[042] System 100 may also enable the user and the system 100 to share assumptions and expectations relating to a given utterance, conversation, or other interaction. For example, the

conversational language processor 120 may be coupled to one or more data repositories 160 that store short-term and long-term shared knowledge that inform decision making in the conversational language processor 120. The short-term shared knowledge may accumulate information during a current conversation (e.g., recognition text for previous utterances, a cross-modal user interface manipulation history, a list of previously selected tasks or invoked queries, or other information), thus dynamically establishing awareness of a cross-modal state of the voice user interface. Storage of the short-term knowledge may be modeled after human interaction, thus certain data may be expired after a psychologically appropriate amount of time (e.g., to expunge stale data), and information with long-term significance can be added to long-term shared knowledge (e.g., a new address of a long-term contact of the user). As such, the long-term shared knowledge may profile or otherwise model environmental, cognitive, historical, demographic, or other aspects of a user based on information accumulated over time (e.g., user-specific jargon, frequently requested tasks, favorite topics or concepts, etc.).

[043] Accordingly, the conversational language processor 120 includes various features that can be used to generate intelligent hypotheses as to a user's intent in a given utterance. For example, the hypothesis may be based on information provided via the context tracking engine 140, the agents 125, the voice search engine 135, the shared knowledge, or other knowledge sources. As a result, the conversational language processor 120 may attempt to identify a conversational type or goal of the utterance (e.g., a query for retrieving a discrete piece of information, a didactic interaction for clarifying information provided by the voice user interface, or an exploratory interaction in which conversational goals may be improvised as the conversation progresses, etc.). Based on the information available and the type or goal of the conversation, the generated hypothesis can be assigned a degree of certainty, which can inform how conversational language processor 120 resolves the utterance. For example, when the degree of certainty indicates that sufficient information has been identified, the conversational language processor may invoke one or more of the agents 125, which in turn can utilize one or more applications 150 to perform a requested task (e.g., a task relating to a navigation application, an advertising application, a music application, an electronic commerce application, or other suitable applications or tasks). In another example, one or more of the agents 125 may query the data repositories 160 or the voice search engine 135 to retrieve requested information, or otherwise take action to resolve a request for information contained in the utterance.

[044] Additionally, the conversational language processor 120 may generate cross-modal intelligent responses, which may be syntactically, grammatically, and contextually sensitive, thus providing a conversational feel to system-generated responses. When available, the intelligent response may present results of a performed task or an executed query to the user, and the response may be provided across modalities. For example, verbal and/or non-verbal outputs 180 may be used separately or in concert (e.g., by presenting information using a display device, an audible alert, a verbalized output, another output mechanism, or various combinations thereof). Further, a verbalized component of the cross-modal output 180 may be adapted to the user's manner of speaking (e.g., by varying tone, pace, timing, or other variables), thus creating verbal responses having natural variation and personality.

[045] The intelligent responses may also be formulated to provide an output 180 that guides the user towards a subsequent response likely to be favorable for recognition. For example, when the degree of certainty reflects an ambiguous context (e.g., when the competition results in a deadlock among the agents 125), an adaptive misrecognition engine 130 may identify the utterance as ambiguous or unrecognized, and the intelligent response can be framed to disambiguate context, or request a subsequent response from the user to clarify the meaning of the utterance, for example. Further, when the conversational language processor 120 or the misrecognition engine 130 determines that a given interpretation or hypothesis was incorrect, one or more previous utterances may be reinterpreted to refine context and build more accurate short-term or long-term conversational models (e.g., an utterance of "No, I meant . . ." may indicate that a previous utterance was interpreted incorrectly, in which case various previous utterances may be reinterpreted based on a correct interpretation, thus building a more accurate conversational context). As a result, the conversational language processor 120 may recover from full or partial failure. Additional techniques for adaptively responding to misrecognition or ambiguity may be included, such as those described in co-pending U.S. Patent Application Ser. No. 11/200,164, entitled "System and Method of Supporting Adaptive Misrecognition in Conversational Speech," filed August 10, 2005, the disclosure of which is hereby incorporated by reference in its entirety.

[046] Additional information relating to the various techniques described herein, as well as other techniques that system 100 uses to provide a conversational, natural language interaction may be provided, for example, in co-pending U.S. Patent Application Ser. No. 11/197,504,

entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed August 5, 2005, co-pending U.S. Patent Application Ser. No. 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," filed August 29, 2005, and co-pending U.S. Patent Application Ser. No. 11/580,926, entitled "System and Method for a Cooperative Conversational Voice User Interface," filed October 16, 2006, the disclosures of which are hereby incorporated by reference in their entirety.

[047] As such, system 100 may provide an environment where conversational, natural language interactions can occur between a user and the system 100. Further, as will be described in greater below, system 100 can be implemented to provide the conversational, natural language interactions, for example, as a voice user interface to one or more electronic devices or applications, including a navigation device. As will also be apparent from the further descriptions provided herein, the system 100 has capabilities for cross-device and cross-modal awareness, such that system 100 provides an environment in which the user can engage in a cooperative, natural language dialogue to request voice services relating to many different devices or applications.

[048] According to various aspects of the invention, Figure 2 illustrates a block diagram of exemplary agent-based architecture 200, which can provide a conversational, natural language voice user interface to a navigation device. The architecture 200 may include a plurality of agents 225a-n, which include specialized software, data, content, or other information that provide functionality, behavior, services, data, and other information in a plurality of respective contextual domains. The architecture 200 may be an integrated and dynamically adaptable environment, in which the agents 225 autonomously react, adapt, and reconfigure to provide optimal service in the environment. For example, by building context over time (e.g., by generating short and long-term profiles of a user, conversations with the user, frequent topics or preferences, etc.), the navigation agent 225a may automatically combine knowledge, adapt preferences, remove conflicts, or perform other adaptations to refine or otherwise optimize an operational framework thereof.

[049] For example, the agents 225, which include at least a navigation agent 225a, may adapt through ongoing use of short-term shared knowledge 270 and long-term shared knowledge 275 about a user's behavior, preferences, or other characteristics. Further, agent adaptation may occur across the plurality of agents 225, for example, in response to various ones of the agents

225 resolving voice-based requests (e.g., by invoking one or more applications 250, querying one or more data sources 260, cooperating with one or more other agents 225, or learning through a misrecognition engine 230, a voice search engine 235, a context tracking engine 24, etc.). In various implementations, adaptation may occur autonomously as a by-product of the agents 225 providing voice services, generating inferences, identifying affinities (e.g., among users, peers, communities, etc.), receiving updates from external sources (e.g., an update manager can update content, behavior, data, commands, domain knowledge, keywords, concepts, dictionaries, or other information for one or more of the agents 225), or in other ways, as will be apparent.

[050] As illustrated in Figure 2, the architecture 200 provides the conversational, natural language voice user interface to the navigation device by way of at least a navigation agent 225a. The navigation agent 225a can include, among other things, various navigation-specific content packages (e.g., dictionaries, available queries, tasks, commands, dynamic grammars, etc.), response lists (e.g., appropriate responses to commands, queries, or other requests), personality profiles (e.g., for creating a natural feel to system-generated speech), substitution lists (e.g., for substituting or transforming data into a structured form that can be understood by a target information source), or various other forms of navigation-specific information. Further, the navigation agent 225a may be associated with parameters and operating data provided by other services in the architecture 200 (e.g., available devices, applications 250, agents 225, etc.), pointers to local or remote data sources (e.g., short-term shared knowledge 270, long-term shared knowledge 275, data sources 260, voice search engine 235, etc.), among various other forms of information.

[051] The data sources 260 used by the navigation agent 225a may include, among other things, data relating to navigation (e.g., maps, addresses, street names, directories, alternate routes, etc.), points-of-interest (e.g., restaurants, hotels, museums, tourist attractions, gas stations, etc.), traffic (e.g., dynamic road conditions, traffic, detours, or other traffic-related information), events (e.g., sporting events, concerts, protests, etc.), parking (e.g., parking garages, restricted areas or times, street lots, street parking, etc.), personal data (e.g., telephone numbers and addresses associated with a user's contact lists), peer affinities (e.g., recommendations based on similarities among users having similar preferences, demographic profiles, or other characteristics). The data

sources 260 may be populated in various ways, such as being based on one or models, received via a data service, extended or refined through adaptation, or in other ways, as will be apparent.

[052] Further, contextual information relating to a navigation domain may be maintained via the context tracking engine 240, short-term shared knowledge 270, and long-term shared knowledge 275, among other things. For example, the contextual information may relate to tracked topics, user location, routes traveled, previous requests, user interface state, user behavior, preferences, demographics, or other characteristics, among other types of contextual information. For example, context may be built and utilized using techniques described in greater detail above with reference to Figure 1, or in the above-referenced co-pending U.S. Patent Applications, or various combinations thereof. Moreover, when navigation agent 225a cannot find information locally within architecture 200, and the information cannot be inferred, the information may be requested from one or more other agents 225b-n (e.g., to request domain-specific information), other devices, the voice search engine 235 (e.g., to search a network), external sources (e.g., the update manager may be invoked to download available updates, which may have the information), or other sources, as will be apparent. If the information can be located through one or more of the other sources, the navigation agent 225a may be adapted, extended, or otherwise updated to make the information available subsequently.

[053] Accordingly, the navigation agent 225a may be coupled with various sources of information, and may make use of context, while communicating with various other adaptable agents 225b-n and other system components to provide voice navigation services. Thus, context, shared knowledge, maps, points of interest, contact lists, user or peer affinities, dynamic grammars, available applications, or various other aspects of the navigation architecture 200 can be shared or otherwise made available to various components, devices, or aspects of the architecture 200. As a result, an inference engine 255 included with the navigation agent 225a may various knowledge sources and other resources available to provide an integrated voice navigation services environment. For example, the inference engine 255 may generate inferences from the available knowledge and resources by applying one or more rules, policies, or other inferencing mechanisms to generate probable interpretations of utterances in the environment (e.g., phonetic fuzzy matching, inductive logic, Bayesian probability analysis, monotonic or non-monotonic reasoning, etc.).

[054] For instance, as described in greater detail above, the navigation agent 225a and one or more other agents 225b-n may compete in analyzing one or more preliminary interpretations of an utterance to generate one or more respective probable interpretations of the utterance. The navigation agent 225a may then be invoked to process the probable interpretation (e.g., by resolving information contained in the utterance) upon being identified as a "winning" one of the competing agents. For example, which of the competing agents wins may depend on context or other information contained in the utterance, whereby the navigation agent 225a may win and recalculate a route in response to an utterance of "This traffic is terrible, maybe back roads," while a music agent may win and change a radio channel in response to an utterance of "Traffic is terrible, maybe some Oldies," while a movie agent may win and search for movie show times in response to an utterance of "Traffic was terrible, maybe a comedy." Thus, when the navigation agent 225a generates a probable interpretation that results in being a winning one of the agents 225, the navigation agent 225a may manage processed for searching, inferring, or otherwise identifying information in light of context to provide voice navigation services using one or more applications 250 (e.g., navigation, grammar generation, location determination, response generation, advertising, device control, or other applications available within architecture 200).

[055] For example, the navigation agent 225a may utilize the inference engine 255 to infer keywords or criteria not explicitly provided in the utterance, determine suitable responses to subjective or indeterminate utterances (e.g., selecting a most likely answer to a query, asking a user for more information, etc.), generate events, identify peer affinities, or otherwise generate inferences for resolving navigation-related requests. The inference engine 255 may generate such inferences through awareness of previous context (e.g., through information provided by the context tracking engine 240), short-term or long-term shared knowledge 270, command histories, states of vehicular systems, user interfaces, or other devices, data sources 260, or other available information. For example, in an exemplary illustration, previous context, shared knowledge, and other information sources may be utilized in conjunction with the inference engine 255 and/or various other components of architecture 200 to enable a human to machine interaction that may occur as follows:

User:	"Show me restaurants around here."

Voice User Interface:	"All I can find is fast food, which you don't like."
User:	"Okay, how about closer to the hotel?"
Voice User Interface:	"Found an Indian restaurant with good reviews. Want me to call them?"
User:	"Yeah go ahead, that sounds good."

[056] In the above illustrated example, the first utterance establishes a context of "restaurants," and the voice user interface responds to the utterance in light of a shared knowledge inference indicating that the user does not like fast food. Furthermore, while the second utterance does not explicitly reference a restaurant, the context established in the first utterance can be used to infer the unreferenced information, forming an unambiguous request out of what would otherwise be ambiguous. Further still, long-term shared knowledge 275 about the user (e.g., the user may frequently go to Indian restaurants) and short-term knowledge 270 about the user (e.g., a hotel associated with a current route) can narrow a search space for resolving the second utterance. Moreover, the voice user interface may invoke the response generation application to advance the conversation towards a resolution (e.g., by prompting the user to call the restaurant). Because the integrated environment has cross-device awareness, the user's subsequent response can result in the device control application formulating an appropriate command structure to dial the user's mobile phone, which may or may not be distinct from the navigation device. Accordingly, as illustrated in this example, robust awareness of context, shared knowledge, available applications, and available systems or devices, among other things, has enabled the user to use natural language to locate a restaurant along a current route matching the user's preferences, while also enabling the user to dial a phone to call the restaurant, without the user having to manually interact with devices or otherwise interact with cumbersome human to machine interfaces.

[057] It will be understood, however, that the above illustrated dialogue provides but one example of a cooperative conversation that can occur between a user and a system that includes the above-described architecture 200. Because the architecture 200 supports free form interactions, and can tolerate serendipitous variations in human speech, vocabulary usage, or other factors, it will be understood that no two conversational interactions between the user and the system will necessarily be alike. As such, the user can seamlessly switch among contexts,

allowing subsequent utterances to occur in various available domains (e.g., a navigation-related utterance of "Take me to the Seahawks game" may be followed up with subsequent cross-domain utterances, such as "Are there tickets left?" or "When does it start?"). Thus, although various descriptions illustrate exemplary ways in which the agent-based architecture 200 can be used for voice navigation services, it will be apparent that, in a given utterance, the user can provide natural language voice-based requests relating to various available contexts, domains, applications, devices, information sources, or various combinations thereof.

[058] The navigation application available within architecture 200 may resolve natural language, voice-based requests relating to navigation (e.g., calculating routes, identifying locations, displaying maps, etc.). The navigation application can provide a user with interactive, data-driven directions to a destination or waypoint, wherein the user can specify the destination or waypoint using free-form natural language (e.g., the user can identify full or partial destinations, including a specific address, a general vicinity, a city, a name or type of a place, a name or type of a business, a name of a person, etc.). As a result, in various implementations, the navigation application may perform post-processing on the full or partial destinations to identify a suitable address for calculating a route (e.g., a closest address that makes "sense"). For example, an utterance containing a full or partial destination may be analyzed to identify one or more probable destinations (e.g., an N-best list of destinations), which may be subject to post-processing that weighs the identified probable destinations based on various factors. Thus, the probable destination most likely to provide a suitable (preliminary or complete) route may be identified from the N-best list.

[059] For example, in an illustrative conversation between a user and the voice user interface, the user may utter a voice-based input of "Take me to Seattle." Based on a current location of the user (e.g., as determined using one or more navigation sensors, radio frequency identifiers, local or remote map databases, etc.), the navigation application may select an address in Seattle that provides a route from the current location to Seattle (e.g., a central point in Seattle may be selected for long distance travel, whereas an address in northern Seattle may be selected for a current location being close to north Seattle). Further, when the requested destination does not specify a final destination (e.g., as in the above example), the user may successively refine the final destination in subsequent requests, and the post-processing may continue to select additional addresses based on the current location, a current route, or other factors. For instance,

continuing the above-provided example, a subsequent input of "Take me to Pike Street" may result in the post-processing analyzing the current route to determine an appropriate address on Pike Street, and possibly recalculating the route, as necessary. Thus, the current route may have the user driving north-bound on Interstate-5, such that an address may be selected on Pike Street on a same side as north-bound lanes of Interstate-5 (e.g., preserving current routing information). The final destination may continue to be successively refined, in combination with the post-processing. As a result, a user knowing only that a destination falls along a 500-block of Pike Street can specify, "501 Pike Street," and even if such an address does not exist, the post-processing may find a closest address that makes sense, thus routing the user to the 500-block of Pike Street.

[060] In addition, the navigation application may provide dynamic, data-driven directions to the destination or dynamically provide routing information, among other things. Furthermore, the response generation application may work together with the navigation application and/or one or more other applications to generate system responses personalized for the user. For instance, the navigation application may access data associated with various user-specific and environmental data sources to provide personalized data-driven directions along a route, which can be recalculated or modified based on information taken from the data sources. Thus, in one example, the response generation application may generate personalized data-driven directions as follows:

Voice User Interface:	"Go to Bob's House."
Voice User Interface:	"Take a left on 14th Street."
Voice User Interface:	"Make a right here to get on the highway, and prepare for heavy traffic."

[061] In the above-illustrated example, the response has been personalized by framing directions in terms of one of the user's contacts (i.e., Bob), while an available source of traffic data has been used to provide data-driven directions. In another example, data that may affect routing can be identified and used to recalculate a route or identify alternate routes (e.g., weather data may be used to avoid snowy roads, or event data may be used to avoid a route along which a protest may be occurring). As such, data may be obtained dynamically to identify alternate

routes, recalculate routes, or otherwise provide optimal routing service. In various implementations, the source of data may be based on persistent models, dynamically received data (e.g., a given highway may regularly have high volume during rush hour or sporting events), or other information. Routes may be recalculated automatically (e.g., user preferences may dictate that heavy traffic be avoided, or sporting event traffic may not be avoided when the user has tickets to the event), or the user can request recalculation through additional utterances (e.g., "I don't want to sit in traffic, can we go a different way?"). Further, possible answers or responses to a given utterance may be filtered according to a current route (e.g., possible results provided for an utterance of "Find me a gas station" may be filtered according to those gas stations within a given proximity of the route). Moreover, continuing the above-provided example, the user may later return to a point of origin, and personalized responses may include return directions, such as "Go back the way you came on 14th Street." Thus, the personalized responses may also create a natural feel to the directions, and may be based on context that builds over time.

[062] The navigation architecture 200 may be used in conjunction with an electronic device capable of accepting voice-based inputs and providing navigation-related information. For example, architecture 200 may be implemented within a handheld device (e.g., personal navigation device, mobile phone, etc.), a telematics devices (e.g., a vehicular navigation system), or various other suitable devices. Furthermore, various functional aspects of the architecture 200 may reside at a client device, at a server, or various combinations thereof. For example, a voice-based input may be received by the client device (e.g., a personal navigation device), which may communicate with the server to retrieve information, perform a task, process the input, or for other reasons. The server may subsequently return information to the client device in response to the request. In various implementations, however, the voice-based input may be processed within the system (e.g., processing delays may be reduced by minimizing communication with the server). Other suitable arrangements may be used, as will be apparent, and specific arrangements as to how information can be processed, retrieved, or returned to the user, among other things, may be highly dependent on context or specific implementations (e.g., what information can be obtained locally, remotely, or otherwise, what services a user subscribes to, etc.).

[063] According to various aspects of the invention, Figure 3 illustrates a flow diagram of an exemplary method 300 for dynamically generating and/or loading recognition grammars in a conversational, natural language voice user interface. For example, a grammar generator application (e.g., as illustrated in Figure 2) may dynamically generate and/or load recognition grammars for use by a navigation agent, an Automatic Speech Recognizer, a context stack, or other components in the voice user interface that incorporate a recognition grammar. While grammar generating method 300 may be used by various components of the natural language voice user interface, the method 300 may be particularly useful in reducing processing overhead and reducing perplexity or confusion. For example, by efficiently generating, updating, loading, extending, or otherwise building dynamic grammars based on various factors, as will described in greater detail herein, bottlenecks can be avoided, conflicts can be reduced, or other aspects of interpreting utterances using a grammar can be optimized.

[064] For example, in various implementations, grammar generation may be keyed to an amount of available resources in a system that includes the voice user interface. Thus, at an operation 310, the amount of system available resources may be calculated, and a subsequent operation 320 may constrain a size of the dynamic grammar based on the calculated resources. For instance, in embedded devices or other devices having low amounts of dynamic memory, the constrained grammar size may ensure that only as many resources as necessary will be occupied. Thus, throughput in the low-memory device will not be degraded by a large-list grammar over-occupying processing capacity. In another example, the size of the dynamic grammar can be reduced by eliminating redundant keywords, criteria, or other information available in the context stack, the shared knowledge, or other local sources. However, it will be apparent that various other techniques may be used to constrain grammar size, and that effectively considering available system resources can provide processing improvements in various environments.

[065] In a navigational context, in which a user's location can change from one moment to another, grammar generation may be further optimized using techniques of geographical chunking. For instance, a user's location can be determined at a given moment (e.g., via navigation sensors, global positioning systems, etc.), and an operation 330 can use the determined location to determine one or more geographic proximities associated with the location. Subsequently, at an operation 340, the determined proximities may be used to identify one or more topological domains appropriate for the determined proximities.

[066] As such, the proximities determined in operation 330 may be weighed or analyzed in various ways to determine how to utilize the topological domains identified in operation 340 in generating the dynamic grammar. For example, the topological domains may reflect physical proximities (e.g., distances from a current location, such as everything within a five-mile radius), civil organization proximities (e.g., regions, states, cities, neighborhoods, subdivisions, localities, etc.), temporal proximities (e.g., amounts of travel time from the current location, where road classes or other information may be used to estimate fastest/safest routes, such as highways and interstates being faster than city or surface roads), directional proximities (e.g., based on directional travel vectors, where a point ten miles ahead may be considered "closer" than a point one mile behind based on a direction of travel, an amount of time to turn around, etc.), or various combinations thereof. As a result, by mapping the user's geographic proximities to one or more topological domains, dynamic grammars may be pruned, extended, swapped in or out of memory, or otherwise generated and/or loaded to provide optimal recognition based on location, time, travel, or other factors (e.g., information may be swapped in and out of a grammar as a user drives from one area to another, optimizing the information along the drive to ensure that the system resources utilize information presently relevant for a given location).

[067] Furthermore, the topological domains may be subdivided into a plurality of tiles, which may in turn be subdivided into a plurality of subtiles. For example, a civil organization topological domain may include a tile representing grammar information for a state, and the tile may include one or more subtiles for counties within the state. Further, the county subtiles may include subtiles for various cities, neighborhoods, or other organizational boundaries within respective counties. Thus, the geographic proximities and topological domains used to build the dynamic grammar may be based on a variety of factors, and can be subdivided or weighed in various ways to determine what information to include in the grammar. Moreover, geographical chunks based on physical, civil organization, temporal, directional, or other proximities may be extended into various other domains in which a topological taxonomy can be placed. As a result, the geographical chunking techniques may have particular relevance in navigation or other location-dependent voice recognition systems, yet the techniques can be suitably applied in various contexts or domains in which geography or location may have relevance.

[068] As various implementations generate the recognition grammars dynamically, for example, to preserve system resources, information available to extend, update, or otherwise

update the grammar may be stored remotely (e.g., on a server). Thus, various implementations may include a system having a network connection, a data service, or another communication mechanism for establishing a link to the remote source. Based on the proximities and/or the topological domains identified for a given location, context, utterance, or otherwise, one or more grammar tiles and/or subtiles may be downloaded from the remote source at an operation 350. Further, the remote source may be adapted to store, analyze, or otherwise process certain information to refine the grammar information stored therein. For example, when a given device communicates with the remote source at operation 350, the remote source may receive identifying information relating to a user of the device (e.g., user preferences, user characteristics, request characteristics, etc.), requested information, or other information. Based on the received information, the remote source may dynamically update various tiles and/or subtiles, build affinities (e.g., among users, peers, communities, etc.), or perform other actions to refine a process by which relevant information can be identified (e.g., the remote source may determine that various different users, who share certain common demographics, may be interested in particular events, locations, etc.).

[069] Thus, the grammar information downloaded at operation 350 may be based on a variety of factors, including information in a current voice-based input, current geographical proximities of the user, topological domains or geographical chunks related to the current geographical proximities, or various kinds of affinities, among other things. Based on these and other factors, appropriate information may be retrieved from various local or remote data sources (e.g., a local repository may contain a great deal of grammar information, a portion of which may be used by the dynamic grammar to interpret an utterance). The retrieved information may then be used to generate the dynamic grammars at an operation 360. As a result, the dynamic grammars may be adapted to include information favorable for correctly interpreting a given utterance. Furthermore, at an operation 370, the favorability of correct interpretations may further be improved by post-processing the generated grammars using one or more optimization techniques, such as reducing perplexity in the grammar. For example, when the grammar generated at operation 360 includes two or more elements likely to be confused (e.g., two or more streets having a same name), the optimization operation 370 may eliminate one or more of the elements to reduce confusion (e.g., selecting the element closest to a user's current location, or other criteria). In another example, when an interpretation turns out to be incorrect (e.g., a

subsequent utterance may be "No, 14th Street Northwest, not Northeast"), the dynamic grammar may be pruned to eliminate information associated with the incorrect interpretation (e.g., information associated with Northeast may be removed). As a result, the optimization operation 370 can reduce overhead, assure that mistakes will not be repeated, and/or improve accuracy of interpretations, among other things.

[070] Although various exemplary operations have been described for dynamically extending, pruning, loading, or otherwise generating recognition grammars, it will be apparent that in operation, the natural language voice user interface may use various forms of information or techniques to determine an optimal grammar. For instance, in addition to being based on available system resources, likelihood of perplexity, geographical proximities, or user affinities, the grammar may be generated in response to conversational contexts, shared knowledge about a user, and/or adaptable agent behavior, among other factors. Accordingly, the dynamic grammar can be structured to recognize queries, commands, or other information in an utterance based on information most likely to be relevant for the utterance, with minimal ambiguity. Further, when the utterance cannot be resolved using the information currently in the grammar, various information sources or techniques may be available to update the grammar in a way that will be favorable for recognition. Thus, the above-described aspects and features of the grammar generation method 300 should be regarded as exemplary only, as many variations will be possible without departing from the inventive concepts described.

[071] According to various aspects of the invention, Figure 4 illustrates a flow diagram of an exemplary method 400 for processing multi-modal voice inputs using a conversational, natural language voice user interface. As described in connection with Figure 1 above, method 400 may be performed in a voice user interface having a mechanism for receiving voice inputs (e.g., a directional microphone, an array of microphones, or other devices that encode speech), wherein the voice inputs include at least a voice component (e.g., an utterance). The utterance, which may be received in an operation 410, may be provided using natural language or other free form manner of expression.

[072] Furthermore, to enable processing of multi-modal voice inputs, the voice user interface may have one or more input mechanisms for receiving non-voice input components (e.g., inputs provided via a keypad, a touch-screen, a stylus/tablet combination, a mouse, a keyboard, or other input modalities). As such, in operation 410, the received voice input may include a non-voice

input component in addition to the utterance, such that an amount of input information may be maximized through usage of multi-modal inputs. For example, the user can use other input modalities to clarify a meaning of the utterance, provide additional information about the utterance, reduce a number of device interactions needed to make a given request, or otherwise maximize an amount of information provided in connection with a given utterance. For example, a multi-modal voice input may be used to efficiently interact with a map displayed on a navigation device (e.g., an utterance of "I only need to see this area" may be coupled with a stylus input that circles a given portion of a map, and the combination of inputs may result in a display of the circled portion of the map).

[073] Thus, for a given voice input, a decisional operation 420 may determine whether a non-voice component accompanies an utterance in the voice input. For example, in various implementations, the various input mechanisms associated with the voice user interface may be coupled to an Alternating Voice and Data (AVD) system. As a result, input data may alternate between voice and data, or the input data may be multiplexed together in a single stream, for example, to allocate processing resources where needed at a given moment.

[074] When decisional operation 420 determines that a non-voice input component has been received, the voice and non-voice components may be correlated in an operation 430. For example, the voice component may be processed using accurate speech recognition techniques, and the voice component may further be parsed and interpreted in light of context, shared knowledge, and available data sources, among other things, as described in greater detail above. As such, correlation operation 430 may generate additional information that can inform context, shared knowledge, or other information for processing the voice input. For instance, as discussed above, short-term shared knowledge may include awareness of various user interface states, which may be affected by non-voice components. Thus, correlation operation 430 may assist the voice user interface in developing one or more preliminary interpretations of the utterance, for example, by relating information in the utterance to information in the non-voice inputs, or vice versa. For example, a multi-modal input could include a user touching a stylus to one of a number of restaurant listings displayed on a touch-screen interface, while also providing an utterance relating to the touched listing (e.g., "Call this one."). In this example, correlation operation 430 can interpret "this one" to refer to the touched telephone listing, providing additional information for processing the voice input.

[075] However, it will be apparent that voice inputs need not necessarily include non-voice components. As such, when the voice input includes only utterances, verbalizations, or other spoken components, decisional operation 420 may advance method 400 directly to an operation 440. For instance, as will be apparent from the further descriptions provided elsewhere in this specification, the natural language voice user interface may efficiently resolve many requests that include only voice. As such, a received voice input can include stand-alone utterances, such that a navigation device, vehicular system, mobile phone, or other device can be controlled in one step using voice. For example, using existing navigation systems or other navigation interfaces, a user must often take multiple steps to configure a displayed map, even for relatively simple tasks (e.g., multiple button and/or stylus inputs may be required to display a desired portion of a map). By contrast, using the natural language voice user interface, the user can simplify controlling a device using one-step voice control (e.g., a map may be controlled using utterances of "Move the map up," "Move the map Northeast," "Show me downtown Seattle," or other utterances that provide all information necessary to control a device in one-step), which may substantially reduce a number of device interactions needed to control the device.

[076] In various instances, (e.g., whether the input received in operation 410 includes an utterance, a non-voice component, or various combinations thereof), one or more queries and/or commands contained in the voice input may be identified in operation 440. Once the one or more preliminary interpretations of the multi-modal input have been generated, appropriate contextual information, shared knowledge, external knowledge sources, or other information may be used to identify the queries and/or the commands contained in the utterance. The information sources relied upon may be associated with various devices, systems, or other components of the integrated voice user interface environment. Thus, information associated with a navigation device (e.g., maps, points-of-interest, etc.) may be distributed for sharing with other devices in the environment (e.g., a personal digital assistant capable of rendering maps may display a map associated with a navigation device, or a mobile phone may dial a telephone number of a point-of-interest associated with the navigation device, etc.). Moreover, the information may be shared in a bi-directional manner, such that information sources associated various devices can be shared with other devices in the integrated environment.

[077] Information may be associated persistently (e.g., by storing all information on a server, an online data backup system, or other data repositories), or may be built dynamically as a

function of processing multi-modal voice inputs. For example, a user may maintain an address book of contacts, e-mail addresses, street addresses, telephone numbers, or other information using a commercial service provider (e.g., GMail, Yahoo Mail, or another service provider capable of storing address books). When a user makes a request (e.g., via voice input, non-voice input, or multi-modal input) in which the service provider manages information needed to resolve the request, the information may be pulled and shared with a device that will be used to resolve the request. For example, a mobile phone may be integrated with the navigation voice user interface (e.g., a mobile phone having a GPS receiver), and the mobile phone may initially (e.g., at a first-time use) have no local contacts, addresses, telephone numbers, or other information. Thus, a user utterance of "I'm running late for dinner at Jane's house, call her to let her know I'm on the way" may result in a query being formulated to the service provider in order to retrieve an address and telephone number associated with Jane's house for the mobile phone to process accordingly. Moreover, information relating to being "late" may be processed to generate a route that avoids traffic, uses highways, or otherwise reduces an amount of travel time.

[078] A variety of cognitive models, contextual models, user-specific models, or other models may be used to identify the query and/or command in operation 440. For example, a given input may include information relating to one or more contextual domains, one or more of which may be invoked to interpret and/or infer certain keywords, concepts, or other information contained in the input. Moreover, short-term and long-term shared knowledge about a user's behavior and preferences may be used in a hybrid recognition model, for example, relating to navigation queries and/or commands. The hybrid recognition model may be used to generate probable or likely interpretations of an utterance using a combination of semantic analysis and contextual inferences. For example, certain syllables, words, phrases, requests, queries, commands, or other identifiable aspects of an utterance may be identified as being more likely to occur based on a given contextual history. The contextual history can be based on the short-term or the long-term shared knowledge about the user, previous utterances in a current conversation, common requests in a given environment (e.g., a navigation device may be used most often for displaying maps, calculating routes, and identifying locations), or other information. Thus, the hybrid recognition model for operation 440 may include various processes for analyzing semantic

patterns to resolve what was said by an utterance, in addition to various processes for relying on contextual history to resolve what was meant by the utterance.

[079] In addition, the hybrid recognition model may be used in conjunction with, or independently of, a peer to peer recognition model. For example, contextual histories may include various preferences or user characteristics, in addition to providing a basis for inferring information about a user based on patterns of usage or behavior, among others. As a result, the recognition models may include additional awareness relating to global usage patterns, preferences, or other characteristics of peer users on a global basis. For example, certain keywords, concepts, queries, commands, or other aspects of a contextual framework may be commonly employed by all users within a context. In another example, users of certain demographics may share common jargon, slang, or other semantic speech patterns. As a result, operation 440 may utilize various recognition models, which consider context and semantics in various dimensions, to identify queries or command. For example, in addition to information generally available within a given environment or context, a voice input may be recognized using context and semantics based on short-term or long-term behavior or preferences for a specific user, global users, peer users, or other meaningful user abstractions.

[080] Accordingly, upon identifying a suitable query and/or command contained in the utterance, a suitable response may be generated, or a task may be performed in an operation 450. As will be apparent from the descriptions already provided, the generated response and/or the performed task may be highly dependent on context and information contained in a given voice input. As a result, when certain voice inputs may be unclear, ambiguous, or otherwise unclear, the generated response may prompt a user for an additional input, and the response may be framed in a way favorable for recognition (e.g., an utterance of "Where is Joe's Pizza" provided while a user is in New York City may return too many results to be meaningful, such that results may be weighed and displayed for a user to select a correct one of the results).

[081] Furthermore, voice inputs may be used to perform compound requests, which could otherwise be impossible to perform using a single manual input. For example, a single voice-based input may include a compound map control request, such as "Show me downtown Seattle." As such, operation 450 may perform tasks of retrieving a map of Seattle and automatically zooming in on a downtown area of the Seattle map. Further, one or more responses or other outputs may be generated, such as suggesting a route to a frequent destination,

identifying possible points-of-interest to the user, or searching for traffic or event notifications, among other things. Many other variations will be apparent, including characteristics of the received inputs (e.g., utterances, non-voice inputs, etc.), the requested queries or commands, the generated responses, and the performed tasks, among other things. As such, it will be apparent that the method 400 illustrated herein may enable users to request many different navigation related tasks verbally, non-verbally, or various combinations thereof, such that the users can request various kinds of information or tasks available in the environment (e.g., relating to various devices, information sources, etc.). Thus, method 400 may operate in an environment in which the natural language voice user interface has been associated with one or more devices that provide navigation services.

[082] According to various aspects of the invention, Figure 5 illustrates a flow diagram of an exemplary method for calculating navigation routes from voice destination entries provided to a navigation device having a conversational, natural language voice user interface. As previously described, the voice user interface may calculate routes, provide dynamic data-driven directions to a destination, provide dynamic routing to a destination, perform post-processing of full or partial destination entries, or otherwise provide various voice navigation services. Thus, as illustrated in Figure 5, the method for calculating routes may support successive refinement of a voice destination entry, wherein context, agent adaptation, and shared knowledge, among other things, can help a user to narrow down a final destination using voice commands, multi-modal commands, or various combinations thereof. However, though Figure 5 specifically illustrates an exemplary implementation of successive refinement for voice destination entry, it will be apparent that the techniques described herein can be applied in performing various tasks in which generalized approximations can suitably be refined through successive voice or multi-modal commands to narrow down information sought by a user in various domains, contexts, applications, devices, or other components that employ the techniques described herein. As a result, successive refinement can be implemented in various domains that enable a user to "drill down" to a specific piece of information or data through specifying more and more specific criteria about the information sought.

[083] As illustrated in Figure 5, for example, successive refinement of a final destination may includes instances in which a user specifies the destination in a single utterance, over a plurality of utterances (e.g., which sequentially narrow down approximations of the destination), as well

as instances in which a system identifies the destination (e.g., through generating inferences, using context or shared knowledge, etc.). For example, a given approximation of a final destination may be associated with a finite number of possible destinations, and the system may be able to unambiguously identify the final destination from among the finite possible destinations by generating inferences, by relying on context, shared knowledge, and other information, or by other ways, as will be apparent. Moreover, in another example, successively refining the destination may be modeled after patterns of human interaction between passengers in a taxicab, or other similar situations in which a route or a destination may be narrowed down or otherwise refined over a course of interaction. For example, passengers in taxicabs sometimes specify a general approximation of a destination (e.g., a passenger picked up in Arlington, Virginia may state, "I'm going to the Foggy Bottom area of D.C."), which may result in a cab driver beginning to drive along a preferred route to the approximated destination (e.g., the driver may know to head towards Rock Creek Parkway when driving from Arlington to Foggy Bottom). While en route to the approximated destination, the passenger and/or the driver may cooperate to refine the final destination through one or more subsequent interactions (e.g., the cab driver may ask, "Where in Foggy Bottom are you headed?", and the passenger may specify "McFadden's," which may provide sufficient information for the driver to determine a route to the destination).

[084] However, it will be apparent that no two sequences of interactions will necessarily be alike. For example, the passenger may provide additional information relating to the final destination without being prompted by the driver, or the driver may not begin driving until the final destination has been identified (e.g., when several routes may be available, and the final destination has an impact on which route will be selected). As such, cooperative, conversational interactions may have a general goal of calculating a route to a final destination, yet the interactions may occur in many different ways, where a given interaction will likely be highly dependent on context, shared knowledge, agent adaptation, points of origin and/or presence, user preferences, dynamic data (e.g., relating to traffic, events, external systems, etc.), among many other factors.

[085] Accordingly, as further described herein, the method illustrated in Figure 5 may enable voice destination entry through successive refinement using the aforementioned factors, among other things, to inform route calculation based on full or partial destination inputs. For instance,

in similar fashion as described in the above cab driver example, the voice user interface (e.g., the cab driver) may have knowledge that certain routes may be preferable to reach a certain destination from a current point of presence (e.g., Arlington to Foggy Bottom via Rock Creek Parkway). Further, as will be described in greater detail in Figure 6, a search space associated with subsequent interactions may be constrained according to a current route, context, or other knowledge (e.g., locations named "McFadden's" that fall outside a certain proximity of Foggy Bottom may be excluded).

[086] Thus, in an operation 510, a user may provide a full or partial destination input using free form natural language, for example, including voice commands and/or multi-modal commands (e.g., an input may include an utterance of "I'm going here," coupled with a touched point on a display). The full or partial destination may be specified in various ways, including by specific address, place name, person's name, business name, neighborhood, city, region, or various other ways (e.g., a voice destination entry may be provided in an exploratory manner, such as when a user wants to visit a museum, but has yet to decide which one to visit). Furthermore, because successive refinement tolerates incomplete or partial addresses, voice destination inputs may be provided in a way that uses the successive refinement in reverse, for example, by narrowly specifying a desired destination in order to calculate a route to a broader vicinity of the desired destination (e.g., a route to South Philadelphia could be identified through an utterance of "Pat's Steaks in Philly," even though Pat's Steaks may not be a user's actual destination).

[087] The voice destination input may be parsed or otherwise analyzed using one or more dynamically adaptable recognition grammars, for example, as described above in reference to Figure 3. For example, recognition grammars may be loaded, generated, extended, pruned, or otherwise adapted based on various factors, including a proximity to a user's point of presence (e.g., as the user moves from one area to another, the recognition grammar may be optimized based on a current location, a direction of travel, temporal constraints, etc.), a contextual history (e.g., as the user interacts with the voice user interface, the grammar may adapt based on dictionaries, keywords, concepts, or other information associated with other contexts, domains, devices, applications, etc.), or other factors, as will be apparent. As such, an operation 520 may include generating one or more interpretations of the voice destination input, which may be analyzed using various data sources in order to generate an N-best list of possible destinations (e.g., a navigation agent may query a directory, a voice search engine, or other components to

identify one or more destinations that at least partially match criteria contained in the voice destination input).

[088] The generated list of possible destinations may be post-processed in an operation 530 in order to assign weights or ranks to one or more of the entries in the N-best list. The post-processing may include analyzing the destination list generated in operation 520 according to various factors in order to determine a most likely intended destination from a full or partial voice destination input. For example, post-processing operation 530 may rank or weigh possible destinations according to shared knowledge about the user, domain-specific knowledge, dialogue history, or other factors. Furthermore, the post-processing operation 530 may analyze the full or partial destination input in order to identify an address to which a route can be calculated, for example, by resolving a closest address that makes "sense" relative to the input destination. For example, a user may specify a partial destination that identifies a broad and approximated area (e.g., "Take me to Massachusetts"), and depending on a user's current location, direction of travel, preferences, or other information, post-processing operation 530 may select an address makes sense for calculating a route (e.g., an address in Cape Cod may be selected for a user having relatives that live in Cape Cod, whereas an address in Boston may be selected for a user who may be traveling to various popular sightseeing areas, etc.).

[089] As a result, the weighed list of N-best destinations may be evaluated in an operation 540 to determine a suitably identifiable destination exists in the list. For example, a full or partial voice destination entry may be ambiguous, or certain criteria or keywords in a voice destination entry may be unrecognizable, such that a highest ranked destination in the weighted list does not exceed a minimal confidence level needed to identify a destination. For instance, a user located in Oklahoma may utter a partial destination, as in "I'm heading to Washington," and the decisional operation 540 may return a negative indication when the voice user interface cannot disambiguate between Washington state, Washington, D.C., Washington University in Saint Louis, and a town of Washington located slightly south of Oklahoma City. In another example, a user originating in Saint Louis may provide a voice destination entry of "Take me to Springfield," which could result in an unidentifiable destination even though multiple destinations may satisfy the minimal confidence level. In this example, "Springfield" may be unidentifiable as a destination because Springfield, Illinois and Springfield, Missouri both reside within reasonable distances of Saint Louis, yet directions of travel to either destination may be

entirely opposite (e.g., a route to the Illinois Springfield includes traveling north on Interstate-55, whereas a route to the Missouri Springfield includes traveling southwest on Interstate-44). Thus, to avoid routing the user in a direction opposite from an intended destination, processing may instead branch to an operation 550 to generate a prompt for resolving the destination.

[090] For example, the prompt may ask the user a question or provide information that can clarify an intended meaning of the received voice destination input. For instance, in the Washington example given above, a multi-modal user prompt may display the four possible destinations on a display device (i.e., Washington state, Washington, D.C., Washington University, and Washington, Oklahoma). The multi-modal prompt may further include system-generated speech, for example, stating, "I found several possible Washingtons, did you mean one of these or something else?" As a result, processing then proceeds back to operation 510, in which the user can disambiguate the intended destination through another input (e.g., by touching one of the displayed destinations when the display device includes a touch-screen, or by verbalizing one of the destinations, or by providing additional information indicating that none of the displayed destinations were intended). Thereafter, subsequent N-best lists may be more likely to result in an identifiable destination, as an original voice destination input can be interpreted in view of a dialogue history.

[091] Alternatively, when at least one of the possible destinations can be identified unambiguously, while meeting the minimal confidence level, a positive indication may result in processing proceeding to another decisional operation 560, which controls how a route will be calculated to the identified destination (e.g., a highest ranking entry in the weighted N-best list of destinations). The decisional operation 560 may determine whether the identified destination provides a full or otherwise complete destination. For example, through successive refinement, a partial voice destination entry may result in an identifiable, yet incomplete, destination. As such, an operation 570 may calculate a preliminary route to the identified partial destination. Thereafter, subsequent processing returns to operation 510, where the voice user interface may await an additional voice destination input that refines the partial destination. Further, in various implementations, one or more system prompts may be generated to request the additional voice destination inputs (e.g., as the user approaches the partial destination and additional information will be needed to provide further routing, or as the user approaches a point in the preliminary

route where distinct routes may be taken to different topographical subtiles or points within the partial destination, etc.).

[092] As such, the processing operations 510-570 may be iteratively repeated until the final destination can be identified through successive refinement using one or more multi-modal voice destination entries. As a result, when a suitable final destination has been identified, operation 560 may be followed by an operation 580 that completes the route to the final destination. For example, the route may be completed by extrapolating the route to the partial destination into a complete route to the final destination. Further, it will be apparent that voice destination entries may be successively refined into a final destination en route (e.g., the final destination may be successively refined as a user proceeds along a preliminary route), in advance (e.g., the user may choose to drive upon a complete route to a final destination being identified), or in other ways.

[093] Moreover, as described in greater detail herein, the calculated route may be dynamically adjusted or rerouted based on subsequent inputs, generated inferences, dynamic data, or various other sources of information (e.g., inferences may be generated, which may result in dynamic routing, in response to dynamic data relating to traffic conditions, detours, events, weather conditions, or user preferences, or other factors). Thus, Figure 6 illustrates a flow diagram of an exemplary method for providing voice services based on a current navigation route according to various aspects of the invention. The voice user interface may use the method illustrated in Figure 6 to provide the voice services in a dynamic, data-driven manner, wherein the voice services may be provided in a customized manner, for example, based on context, shared knowledge about a user, or other sources of information.

[094] Accordingly, providing voice navigation services based on a current route includes identifying the current navigation route in an operation 610. The route may be identified from a preliminary route or a complete route, as calculated according to the techniques described above in reference to Figure 5. Thus, current routing information may be made available to various system components in order to process events or voice inputs, generate responses, formulate queries, resolve commands, or perform other processing services available in a voice navigation services environment. As such, the current routing information may provide further context relevant to interpreting voice inputs, taking action towards resolving the inputs, or responding to events, among other things. Thus, though not illustrated in Figure 6, various other sources of contextual information may be utilized to provide voice navigation services in view of the

current route. For example, user preferences may limit places to which a vehicle can travel, or for how long the vehicle can travel, among other things (e.g., for keeping a delivery or passenger route on a regular route, or for placing parental controls on teenage drivers, etc.). In another example, the voice user interface may have awareness of one or more external systems or devices, sources of dynamic data, or other information, which can be utilized to determine how to best provide the voice navigation services.

[095] Thus, upon identifying the current route in operation 610, along with other appropriate information relating to context, shared knowledge, or otherwise, a decisional operation 620 may determine whether the identified information results in an event being generated and/or detected. When no event can be generated and/or detected, processing proceeds to a decisional operation 630, which determines whether a voice input has been received from the user. As such, the method illustrated herein may loop through decisional operations 620 and 630 until one or more of an event has been detected and/or a voice input has been received.

[096] One type of event that can be detected and/or generated in decisional operation 620 may relate to location dependent advertisements for navigation systems (e.g., as generated by a local or remote advertising engine, or via a data channel, or in other ways). For example, in various implementations, the advertising events may be generated and/or detected using techniques described in co-pending U.S. Patent Application 11/671,526, entitled "System and Method for Selecting and Presenting Advertisements Based on Natural Language Processing of Voice-Based Input," filed February 6, 2007, the disclosure of which is hereby incorporated by reference in its entirety. For instance, as will be apparent, a navigation system will necessarily include one or more systems or mechanisms for determining the navigation system's current location (e.g., a global positioning system, a radio frequency identification system, a system that determines location based on a distance to an identifiable wireless tower or access point, etc.). Further, the natural language voice user interface may be able to determine location in a number of other ways, in addition to or apart from the location detection system (e.g., an adaptable prompt may affirmatively request the location information from a user, for example, upon failure of the location detection system, or other techniques utilizing shared knowledge or context may be used, as will be apparent).

[097] Thus, in one example, operation 620 may include detecting and/or generating events relating to a location dependent advertisement. A marketer or other entity seeking to reach a

target audience may simply scatter radio frequency identifiers (RFIDs) along a roadside, or another place in which a system may detect the RFIDs. The marketer or other entity may then broadcast one or more advertisements via a data channel associated with the RFIDs, such that the navigation system may trigger an event when within a suitable proximity of the RFIDs. Thus, processing would proceed to an operation 570, in which information associated with the event may be filtered according to the current routing information or other contextual parameters. As applied to this example, the filtering operation 650 may include monitoring the data channel to analyze information associated with the detected RFID, thus receiving the advertisements broadcasted over the data channel. Thus, an advertisement including speech, for example, may be analyzed in a similar manner to how user speech would be analyzed (e.g., maintaining awareness that the speech relates to an advertisement).

[098] Accordingly, the advertisement event data may be filtered in view of the current route, personal information (e.g., previous queries, preferences, history, etc.), or other factors to determine what action should be taken in an operation 660. For example, the advertisement may be associated with a local restaurant seeking to reach personal navigation devices that may not necessarily have a complete business database. Therefore, in one example, filtering operation 650 could discard the advertisement when shared knowledge indicates that the user prefers to reach destinations as fast as possible, without making unnecessary stops. In this case, operation 660 may effectively be a null operation, or may simply include displaying an image associated with the advertisement to allow the user to make an independent choice as to whether or not to stop. However, in a distinct example, the user may have provided a previous voice input indicating a desire to stop driving and eat lunch at some point along the route. Thus, in this example, operation 660 may include a spoken output that brings the restaurant to the user's attention, possibly asking the user whether the route should be temporarily diverted to stop for lunch at the restaurant.

[099] However, it should be appreciated that the above described RFID-based advertisement provides but one example of a possible advertising event that may be detected in operation 620. For instance, advertisements may be uploaded to a server by one or more advertising partners, wherein the uploaded advertisements may be associated with metadata or other descriptive information that identifies a target audience, location-dependent information, or other criteria. In another example, a plurality of advertisements may be stored locally at the voice user interface,

and an inferencing engine may determine appropriate circumstances in which an event should be generated to deliver one or more of the advertisements to a user. As a result, it will be apparent that advertising events may be generated in a number of ways, and may be generated and/or detected locally, remotely, by detecting RFIDs, or in other ways.

[0100] Further, as advertisements get delivered to users (e.g., in operation 660), the users' subsequent interaction with the delivered advertisements may be tracked. In this way, affinity based models may be generated, for example, to ensure that promotions or advertisements will be delivered to a likely target audience. Thus, an event relating to a given advertisement may be generated and/or detected in operation 620 when shared knowledge about a user's behavior, preferences, or other characteristics match one or more criteria associated with peer-to-peer affinities associated with the advertisement. For example, a marketer may be offering promotional coupons intended for a specific demographic audience, and the peer-to-peer affinities may be used to ensure that the coupons will only be provided to users likely to be in the intended audience (e.g., to avoid providing an entire market, including unlikely customers, with the coupons). Thus, when the user falls within the specified demographic audience, one or more events may be generated on a server, locally by an advertising engine, or in other ways to deliver one or more of the promotional coupons to the user.

[0101] In other examples, the advertising model may include mobile pay-per-use systems (e.g., network access may be provided through a wireless access point on a pay-per-use basis), peer-to-peer local guides or recommendations (e.g., affinities may indicate that various categories of peer users may be interested in similar points-of-interests, events, local restaurants, stores, activities, tourist attractions, or other aspects of a local area). Additionally, as described in greater detail above in reference to Figure 3, various aspects of the advertising model, such as the local guides and recommendations, may be generated according to a mapping applied to various topological domains. For example, local guides or recommendations may be dependent on topological characteristics (e.g., different guides for various cities, communities, localities, etc.). As such, the local guides, recommendations, or other aspects of the advertising model may be associated with a topological taxonomy based on geographical chunking. As a result, various advertising events may be generated and/or detected according to physical proximities, temporal proximities, directional proximities, civil organization proximities, or various combinations

thereof (e.g., peer-to-peer recommendations may be varied depending on a user's location, direction of travel, temporal constraints, etc.).

[0102] In addition to the advertising related events that can be generated and/or detected in operation 620, the natural language voice user interface may generate additional events through awareness of context, shared knowledge about a user, external systems and devices, or other information. For example, as discussed above in reference to Figure 2, the voice user interface may implemented within an agent-based architecture, in which one or more of a plurality of agents may include an inference engine. However, it will be apparent that the inference engine may be arranged within the voice user interface in various other ways. For example, the voice user interface may include the inference engine within a conversational language processor, where in such implementations, the inference engine may generate, detect, and distribute events to various components of the voice user interface. In another example, a managing inference engine may coordinate event generation and/or detection among the inference engines associated with the agents. As such, it will be apparent that the voice user interface may include various suitable arrangements of one or more inference engines, such that events and inferences can be detected, generated, and distributed to various system components as they may arise.

[0103] Thus, the one or more inference engines may utilize awareness of context, shared knowledge about a user, dynamic data sources, data and services associated with external or devices, among other information to generate and/or detect the events or other inferences identified in operation 620. When the events or other inferences occur when a current navigation route exists, the events or other inferences may be filtered in operation 650 prior to being a response being generated and/or a task being performed in operation 660.

[0104] For example, the inference engine may utilize personal information, such as a user frequently visiting a particular person in the user's address book, to generate an event that suggests rerouting the current route. Thus, in this example, the inference engine may generate the event (which may be detected in operation 620) when the current route or current location information reflects a certain proximity to an address associated with the person in the address book. Operation 650 could then filter the event by determining possible responses or courses of action for handling the event in a manner consistent with the current route. For example, information associated with the current route may indicate that the user was out running errands and may now be heading home. Further, a calendar in the user's address book may indicate that

the user has an appointment later in the day, but that no calendar entries exist for a few hours. Thus, an exemplary response generated in operation 660 may include a speech-based output of, "You are close to Jane's house, would you like to stop by?"

[0105] Although the above-provided example illustrates an event generated using personal information, it will be apparent that various available information sources may be utilized to generate and/or detect events. For example, events may be based on transient or dynamic data relating to communities, traffic, weather, and/or many other sources of data (e.g., a generated event may result in an output of "Tonight's soccer game is being played at a field near your house, do you want to go?", in which the soccer game may be a local community event).

[0106] In yet another example, various types of external system awareness may trigger events and resulting voice responses from the voice user interface. For example, the voice user interface may be coupled to various vehicular or telematics systems, providing awareness over a state of such systems. Thus, an event may be generated when the vehicular system indicates that the vehicle will soon run out of gas, resulting in a voice response prompting a user to stop for gas. Further, the external system awareness may be filtered according to the current routing information (e.g., in conjunction with the inference engine), wherein a calculation may be made indicating that a level of gas will be sufficient to get to a next gas station along the route, but that the level of gas will not be sufficient to get to a second gas station along the route. As a result, the route-based filtering operation 650 may result in a response being generated in operation 660 that provides a voice warning that the user must stop for gas at the next gas station.

[0107] Returning to operation 630, user-provided voice inputs may also be processed in view of current routing information. As previously discussed, interpretations of what was said in a voice input (e.g., content of an utterance) may be based on various phonetic models, dictionaries, context histories, dialogue histories, and other information that can form a dynamic recognition grammar. Moreover, context, shared knowledge, and available data and services, among other information, may be used to interpret what meant by the voice input (e.g., an intent of the utterance). As a result, using the various techniques described herein and in the incorporated co-pending U.S. Patent Applications, a query, command, or other request contained in the utterance may be identified in an operation 640. When a current route exists, a domain of possible tasks, commands, query answers, responses, and/or other system actions may be filtered according to the current route. For example, a voice input of "Where's the closest bathroom" may be

analyzed in view of the current route. Various proximities (e.g., a direction of travel, physical distances, temporal distances, etc.) may then be utilized to determine what bathrooms may be appropriate in view of the current route. Thus, when a route has the user traveling a long distance, for example, a bathroom at a rest area twenty miles ahead may be favored over a bathroom at a restaurant ten miles off a highway (e.g., to avoid diverting the user from the current route).

[0108] Furthermore, in many instances, an agent or other processing component may need to formulate a plurality of queries to various different data sources in order to resolve a request. As such, operation 660 for performing a task may determine what inferences, metadata, or other information may be needed to formulate queries to the data sources in order to resolve the request. As a result, voice requests relating to navigation may include information that may not necessarily be linked to navigation, or that may not necessarily be in a database or other source of navigation information. Even so, by being coupled with various available sources of information, complex or compound queries may be formulated to retrieve information from various data sources in order to resolve a given request. Furthermore, a given utterance may include multiple requests, queries, or commands, such that a given response or task performed in response to the utterance may include interactions with various different system components. Thus, in an illustrative example, an utterance of "Find the 7-11 closest to the Woodland Park Zoo" may include a first query to identify a location of Woodland Park Zoo, a second query to find one or more 7-11's within a certain radius of the Woodland Park Zoo location, a third query to determine a current location of the user, and a fourth query that invokes an inference engine to determine which 7-11 may be "closest" based on various user proximities. Thus, supposing that two 7-11's exist equidistantly from the Woodland Park Zoo, one in a direction east of the Zoo and one in a direction west of the Zoo, the user's current location may be used to determine which 7-11 to select; by contrast, a selection based purely on physical proximity to the Zoo could result in ambiguity.

[0109] Although the above-provided descriptions highlight various specific examples and illustrations of how the natural language voice user interface may provide voice navigation services, it will be understood that the descriptions, examples, and modes of operation provided above should be regarded as exemplary only. The natural language voice user interface may be highly tolerant of variations in how humans interact verbally, and may utilize context, shared

knowledge, and available data and systems to effectively interact with any given user in a conversational manner that advances mutual goals. As a result, interactions between the user and the voice user interface may be regarded as unique occurrences, as users may speak free form requests relating to various available data sources, applications, or devices in a given interaction. Thus, the voice user interface may dynamically adapt, learn, and otherwise refine an operational framework to provide optimal services in the voice navigation services environment.

[0110] Implementations of the invention may be made in hardware, firmware, software, or various combinations thereof. The invention may also be implemented as instructions stored on a machine-readable medium, which may be read and executed by one or more processors. A machine-readable medium may include a mechanism for storing or transmitting information in a form readable by a machine (e.g., a computing device). For example, a machine-readable storage medium may include read only memory, random access memory, magnetic disk storage media, optical storage media, flash memory devices, and others, and a machine-readable transmission media may include forms of propagated signals, such as carrier waves, infrared signals, digital signals, and others. Further, firmware, software, routines, or instructions may be described in the above disclosure in terms of specific exemplary aspects and implementations of the invention, and performing certain actions. However, those skilled in the art will recognize that such descriptions are merely for convenience and that such actions in fact result from computing devices, processors, controllers, or other devices executing the firmware, software, routines, or instructions.

[0111] Aspects and implementations may be described as including a particular feature, structure, or characteristic, but every aspect or implementation may not necessarily include the particular feature, structure, or characteristic. Further, when a particular feature, structure, or characteristic is described in connection with an aspect or implementation, it will be understood that one skilled in the art may be able to effect such feature, structure, or characteristic in connection with other aspects or implementations, whether or not explicitly described. Thus, various changes and modifications may be made to the provided description without departing from the scope or spirit of the invention. As such, the specification and drawings should be regarded as exemplary only, and the scope of the invention to be determined solely by the appended claims.

CLAIMS

What is claimed is:

1. A computer-implemented method for providing a natural language voice user interface for a navigation device, comprising:

receiving a navigation-related voice-based input at an input mechanism associated with the navigation device, the voice-based input including at least a natural language utterance provided by a user;

generating one or more preliminary interpretations of the natural language utterance using a dynamic recognition grammar associated with a speech recognition engine;

analyzing the preliminary interpretations using a conversational language processor, the conversational language processor using shared knowledge and information associated with a navigation context to determine a probable interpretation of the natural language utterance in the navigation context; and

communicating the probable interpretation of the natural language utterance to a navigation agent associated with the navigation context, the navigation agent operable to:

identify one or more navigation-related requests contained in the voice-based input based on the probable interpretation of the natural language utterance; and

resolve the requests using information associated with one or more of a plurality of information sources, the plurality of information sources including at least a navigation-specific information source.

2. The method of claim 1, the navigation agent further operable to determine that at least one of the requests includes an approximation of requested information, wherein the conversational language processor manages a dialogue with the user to resolve the approximated information, the managed dialogue including one or more interactions with the user that successively refine the approximated information.

3. The method of claim 2, each of the interactions including one or more of prompting the user to refine the approximated information or receiving a subsequent input from the user that refines the approximated information.

4. The method of claim 1, the navigation-related requests including a request for a route to a full or partial address, the navigation agent operable to resolve the request for the route by:

selecting a destination for the route, the selected destination having an address that best corresponds to the full or partial address;

calculating a route from a current location of the user to the selected destination; and

generating directions to guide the user on the calculated route to the selected destination, the generated directions dynamically driven by data from a plurality of information sources.

5. The method of claim 4, the dynamic data-driven directions including information about destinations, points of interest, traffic, parking, weather, events, or other information relevant to the calculated route.

6. The method of claim 4, further comprising:

receiving a subsequent voice-based input at the input mechanism associated with the navigation device, the subsequent voice-based input including at least one request provided by the user; and

invoking at least one of a plurality of domain agents to resolve the request included in the subsequent voice-based input, the at least one domain agent operable to filter results of the request according to the calculated route.

7. The method of claim 1, the dynamic recognition grammar including information associated with one or more topological domains.

8. The method of claim 7, the topological domains including physical, temporal, directional, and civil organizational proximities with respect to a current location of the user.

9. The method of claim 1, the conversational language processor coupled to an inferencing engine operable to generate one or more inferences, wherein the conversational language processor uses the generated inferences to determine the probable interpretation of the natural language utterance.

10. The method of claim 9, wherein the conversational language processor further uses the generated inferences to suggest one or more services available in the navigation context to the user.

11. The method of claim 1, the navigation-related requests including a multi-modal request to control a map display, the navigation agent operable to resolve the request to control the map display by:

associating a non-voice component of the voice-based input with the probable interpretation of the utterance, the non-voice component identifying a portion of the map display; and

issuing a command to control the identified portion of the map display in accordance with the probable interpretation of the utterance.

12. The method of claim 1, wherein the plurality of information sources include the navigation-specific information source, the shared knowledge, the information associated with the navigation context, and one or more information sources relating to maps, destinations, directories, points of interest, traffic, parking, weather, events, user address books, user devices and systems, a search engine, and a plurality of domain agents.

13. The method of claim 12, the shared knowledge including dialogue history information, request history information, user interface state information, short-term user profile information, long-term user profile information, peer user profile information, and user location information.

14. A system for providing a natural language voice user interface for a navigation device, comprising:

an input mechanism that receives a navigation-related voice-based input, the voice-based input including at least a natural language utterance provided by a user;

a speech recognition engine having a dynamic recognition grammar associated therewith, the speech recognition engine operable to generate one or more preliminary interpretations of the natural language utterance using the dynamic recognition grammar;

a plurality of domain agents associated with respective contexts, the plurality of domain agents including at least a navigation agent associated with a navigation context; and

a conversational language processor operable to:

analyze the preliminary interpretations using shared knowledge and information associated with the navigation context to determine a probable interpretation of the natural language utterance in the navigation context; and

communicate the probable interpretation of the natural language utterance to the navigation agent associated with the navigation context, the navigation agent operable to identify one or more navigation-related requests contained in the voice-based input based on the probable interpretation of the natural language utterance and resolve the requests using information associated with one or more of a plurality of information sources, the plurality of information sources including at least a navigation-specific information source.

15. The system of claim 14, the navigation agent further operable to determine that at least one of the requests includes an approximation of requested information; and

the conversational language processor further operable to manage a dialogue with the user to resolve the approximated information, the managed dialogue including one or more interactions with the user that successively refine the approximated information.

16. The system of claim 15, each of the interactions including one or more of prompting the user to refine the approximated information or receiving a subsequent input from the user that refines the approximated information.

17. The system of claim 14, the navigation-related requests including a request for a route to a full or partial address, the navigation agent operable to resolve the request for the route by:

selecting a destination for the route, the selected destination having an address that best corresponds to the full or partial address;

calculating a route from a current location of the user to the selected destination; and

generating directions to guide the user on the calculated route to the selected destination, the generated directions dynamically driven by data from a plurality of information sources.

18. The system of claim 17, the dynamic data-driven directions including information about destinations, points of interest, traffic, parking, weather, events, or other information relevant to the calculated route.

19. The system of claim 17, the one or more input mechanisms further receiving a subsequent voice-based input, the subsequent voice-based input including at least one request provided by the user, and the conversational language processor further operable to:

invoking at least one of a plurality of domain agents to resolve the request included in the subsequent voice-based input, the at least one domain agent operable to filter results of the request according to the calculated route.

20. The system of claim 14, the dynamic recognition grammar including information associated with one or more topological domains.

21. The system of claim 20, the topological domains including physical, temporal, directional, and civil organizational proximities with respect to a current location of the user.

22. The system of claim 14, further comprising an inferencing engine coupled to the conversational language processor that generates one or more inferences, wherein the conversational language processor uses the generated inferences to determine the probable interpretation of the natural language utterance.

23. The system of claim 22, further comprising an inferencing engine coupled to the conversational language processor, wherein the conversational language processor further uses the generated inferences to suggest one or more services available in the navigation context to the user.

24. The system of claim 14, the navigation-related requests including a multi-modal request to control a map display, the navigation agent operable to resolve the request to control the map display by:

associating a non-voice component of the voice-based input with the probable interpretation of the utterance, the non-voice component identifying a portion of the map display; and

issuing a command to control the identified portion of the map display in accordance with the intended meaning of the utterance.

25. The system of claim 14, wherein the plurality of information sources include the navigation-specific information source, the shared knowledge, the information associated with the navigation context, and one or more information sources relating to maps, destinations, directories, points of interest, traffic, parking, weather, events, user address books, user devices and systems, a search engine, and a plurality of domain agents.

26. The system of claim 25, the shared knowledge including dialogue history information, request history information, user interface state information, short-term user profile information, long-term user profile information, peer user profile information, and user location information.

27. A method for providing a natural language voice user interface in a voice navigation services environment, comprising:

receiving a voice-based destination input provided by a user, the voice-based destination input including at least a natural language utterance;

generating one or more preliminary interpretations of the natural language utterance using a dynamic recognition grammar associated with a speech recognition engine;

analyzing the preliminary interpretations using a conversational language processor, the conversational language processor using shared knowledge and information associated with a navigation context to determine an intended destination provided in the natural language utterance, the intended destination including an approximation of a destination; and

communicating the intended destination to a navigation agent, the navigation agent operable to provide a route to the intended destination by:

selecting a preliminary destination for the route, the selected preliminary destination having an address that best corresponds to the approximated destination;

calculating a route from a current location of the user to the selected preliminary destination; and

participating in a dialogue with the user to resolve a final destination for the calculated route, the dialogue includes one or more interactions with the user that successively refine the approximated destination until the final destination has been resolved.

28. The method of claim 27, each of the interactions including one or more of prompting the user to refine the approximated destination or receiving a subsequent input from the user that refines the approximated information.

29. The method of claim 27, the navigation agent further operable to dynamically recalculate the route in response to the successive refinements of the approximated destination.

30. The method of claim 27, the final destination being resolvable substantially later in time with respect to an initial calculation of the route.

31. The method of claim 27, the navigation agent operable to select the preliminary destination by identifying one or more addresses possibly corresponding to the approximated destination and selecting a highest ranked one of the identified addresses as the preliminary destination.

32. The method of claim 27, the identified addresses ranked according to a proximity to the current location of the user or the approximated destination.

33. The method of claim 32, the proximity to the based on one or more topological domains, including one or more of physical, temporal, directional, and civil organizational proximities with respect to the current location of the user or the approximated destination.

34. The method of claim 27, the navigation agent further operable to generate directions to guide the user on the calculated route, the generated directions dynamically driven by data from a plurality of information sources.

35. The method of claim 34, the plurality of information sources including information relating to destinations, points of interest, traffic, parking, weather, events, or other information relevant to the calculated route.

36. The method of claim 27, the shared knowledge including dialogue history information, request history information, user interface state information, short-term user profile information, long-term user profile information, peer user profile information, and user location information.

37. A method for providing a natural language voice user interface in a voice navigation services environment, comprising:

receiving a navigation-related voice-based input provided by a user, the voice-based input including at least a natural language utterance;

identifying a current location of a user using a location detection system;

determining one or more topological domains associated with the current location of the user;

generating a dynamic recognition grammar that includes grammar information associated with the determined topological domains; and

using the dynamic recognition grammar to generate one or more interpretations of the natural language utterance.

38. The method of claim 37, the dynamic recognition grammar organizing the grammar information according to geographic chunks associated with the topological domains.

39. The method of claim 38, the geographic chunks including physical proximities defined by a distance from the current location of the user.

40. The method of claim 38, the geographic chunks including temporal proximities defined by a time of travel from the current location of the user.

41. The method of claim 38, the geographic chunks including directional proximities defined by a directional vector of travel by the user.

42. The method of claim 38, the geographic chunks including civil organizational proximities defined by continents, countries, regions, states, cities, localities, neighborhoods, and communities.

43. The method of claim 38, further comprising dividing one or more of the geographic chunks into a plurality of tiles, the dynamic recognition grammar further organizing the grammar information according to the plurality of tiles.

44. The method of claim 43, further comprising dividing one or more of the plurality of tiles into a plurality of subtiles, the dynamic recognition grammar further organizing the grammar information according to the plurality of subtiles.

45. The method of claim 37, the dynamic recognition grammar having a size constrained according to memory or resource availability in the voice navigation services environment.

46. The method of claim 37, further comprising modifying the grammar information included in the dynamic recognition grammar in response to changes in the topological domains associated with the current location of the user.

47. The method of claim 37, further comprising determining affinities between the user and one or more peers of the user, the dynamic recognition grammar further organizing the grammar information according to the determined affinities.

48. A method for providing advertisements in a voice navigation services environment, comprising:

identifying a current location of a user using a location detection system;

retrieving shared knowledge and information associated with a navigation context, the retrieved information used to determine probable interpretations of natural language utterances received in the voice navigation services environment;

identifying one or more advertisements to provide to the user, the one or more advertisements identified based on an affinity to one or more of the current location of the user, the shared knowledge, or the information associated with the navigation context; and

generating a multi-modal output to provide the one or more identified advertisements to the user.

49. The method of claim 48, the multi-modal output including a system-generated speech utterance.

50. The method of claim 48, the location detection system receiving information relating to the one or more advertisements over a data channel.

51. The method of claim 50, the data channel associated with a radio frequency identifier detected by the location detection system.

52. The method of claim 48, the shared knowledge including dialogue history information, request history information, user interface state information, short-term user profile information, long-term user profile information, peer user profile information, and user location information.

53. The method of claim 48, further comprising:

receiving a voice-based input that includes at least a natural language utterance;

using the shared knowledge and the information associated with the navigation context to determine a probable interpretation of the received natural language utterance, the advertisements identified based on the probable interpretation of the natural language utterance.

54. The method of claim 48, the advertisement including a guide to a community local to the current location of the user.

55. The method of claim 54, the local community guide including recommendations for one or more points of interest, events, restaurants, stores, activities, or tourist attractions.

56. The method of claim 54, the local community guide including information associated with one or more of maps, destinations, directories, traffic, parking, or weather.

57. The method of claim 54, the local community guide based on affinities between the user and one or more of the current location of the user, one or more peers of the user, or one or more communities associated with the user.

58. The method of claim 48, the advertisement including one or more of an image, a banner, an audio message, a video message, a promotional offer, a coupon, and a data stream.

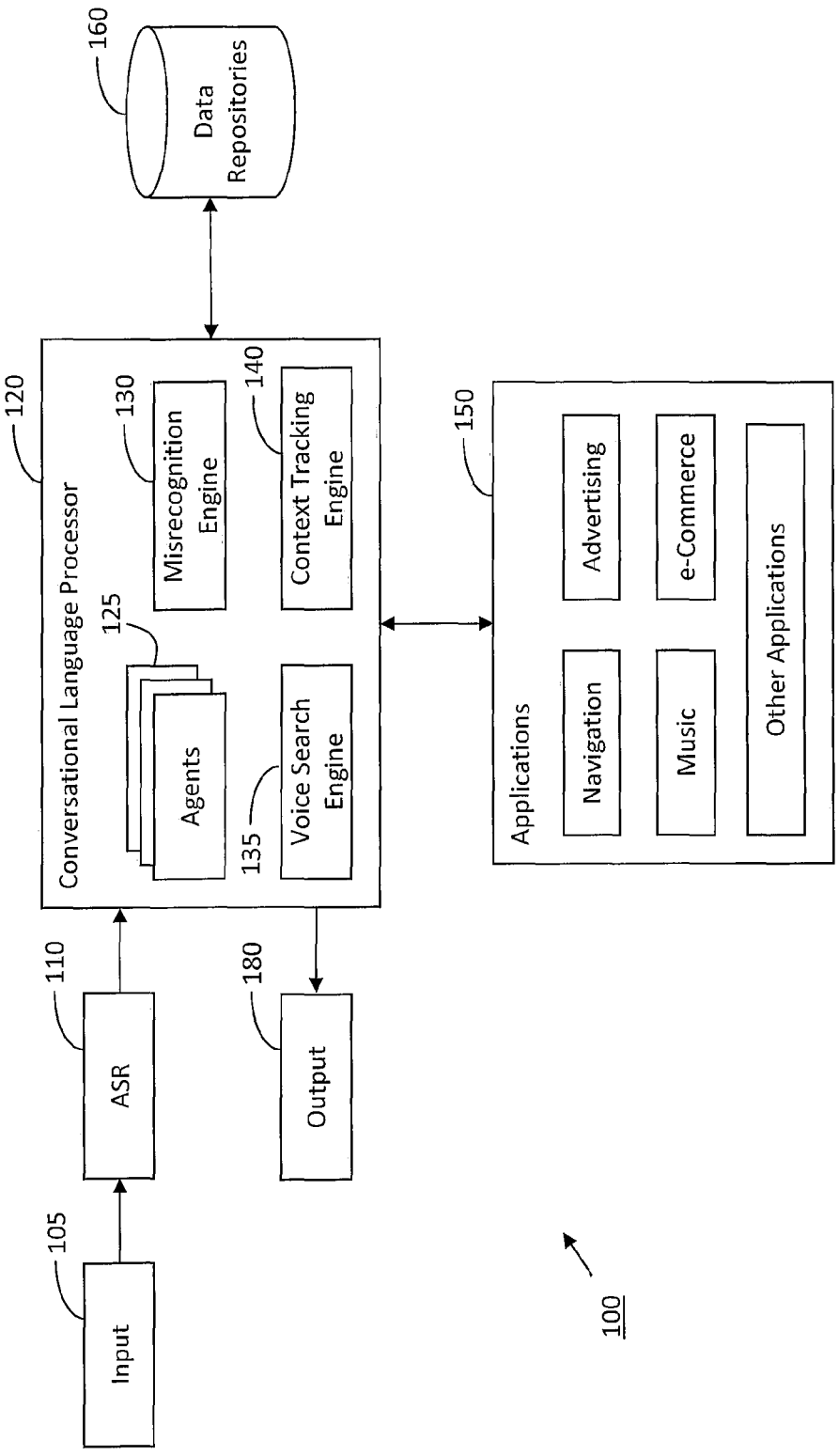

160
Data Repositories
120
Conversational Language Processor
130
Misrecognition Engine
140
Context Tracking Engine
125
Agents
135
Voice Search Engine
150
Applications
Advertising
Navigation
e-Commerce
Music
Other Applications
110
ASR
180
Output
105
Input
100

Figure 1

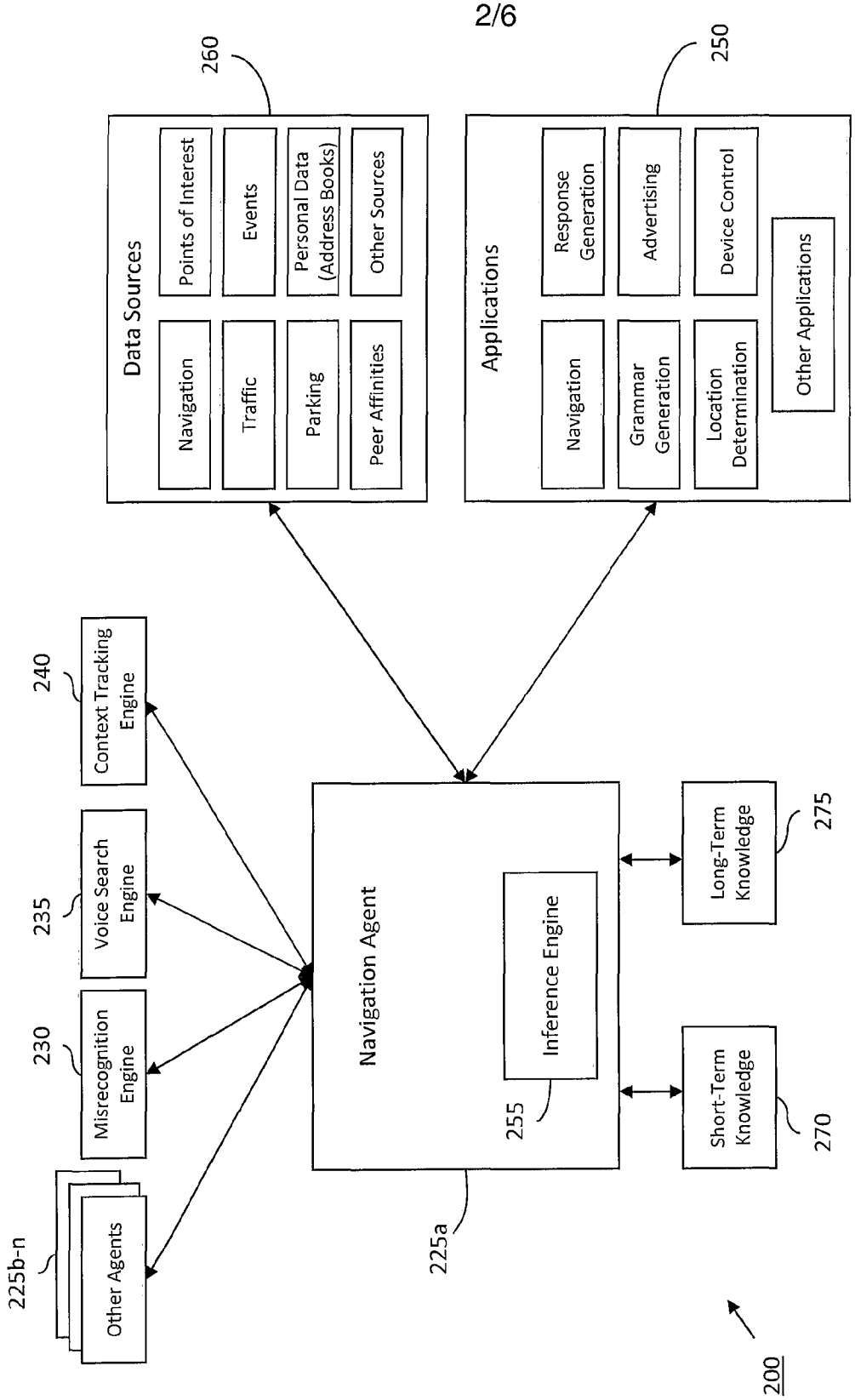

260
Data Sources
Points of Interest
Events
Personal Data (Address Books)
Other Sources
Navigation
Traffic
Parking
Peer Affinities
250
Applications
Response Generation
Advertising
Device Control
Navigation
Grammar Generation
Location Determination
Other Applications
240
Context Tracking Engine
235
Voice Search Engine
230
Misrecognition Engine
225b-n
Other Agents
Navigation Agent
Inference Engine
255
225a
Long-Term Knowledge
275
Short-Term Knowledge
270
200

Figure 2

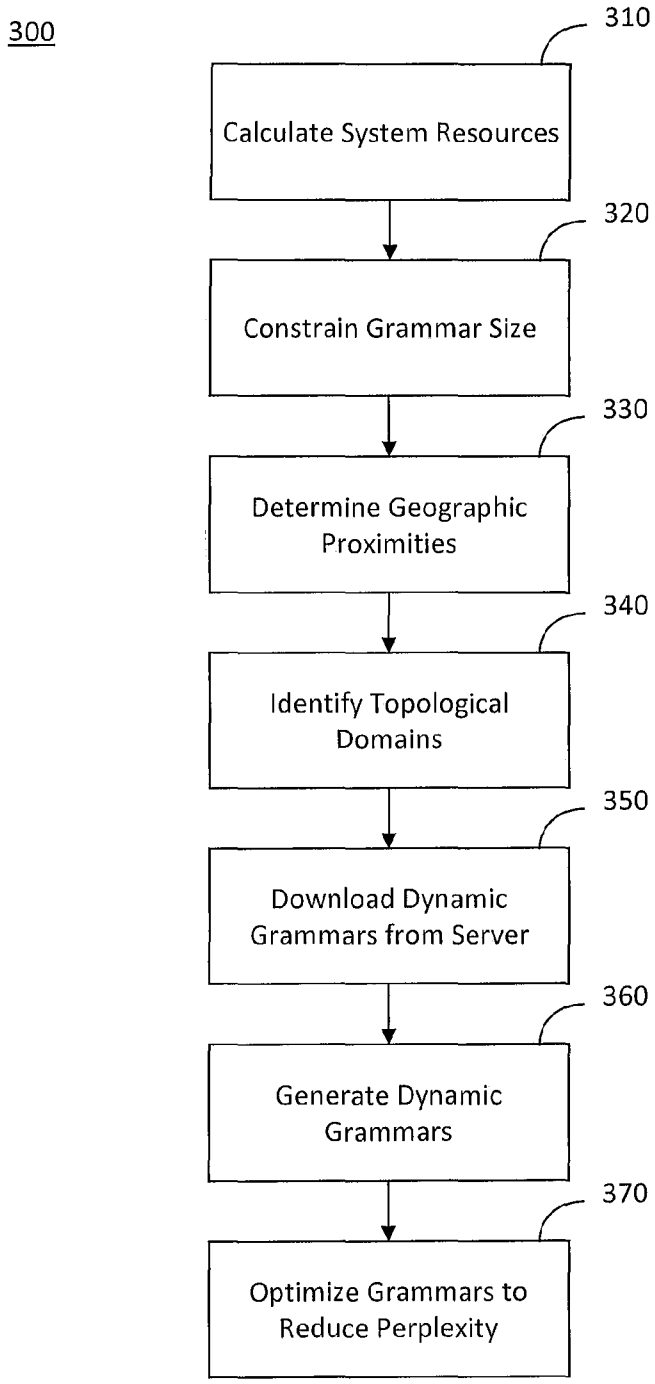

300
310
Calculate System Resources
320
Constrain Grammar Size
330
Determine Geographic Proximities
340
Identify Topological Domains
350
Download Dynamic Grammars from Server
360
Generate Dynamic Grammars
370
Optimize Grammars to Reduce Perplexity

Figure 3

400

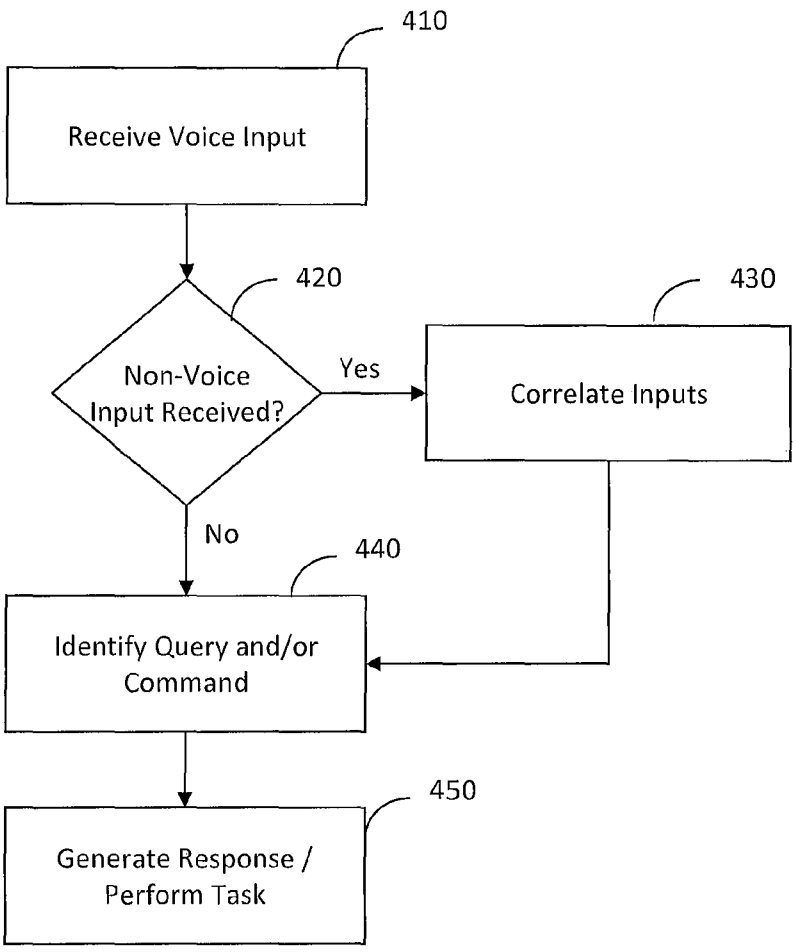

410
Receive Voice Input
420
Non-Voice Input Received?
Yes
430
Correlate Inputs
No
440
Identify Query and/or Command
450
Generate Response / Perform Task

Figure 4

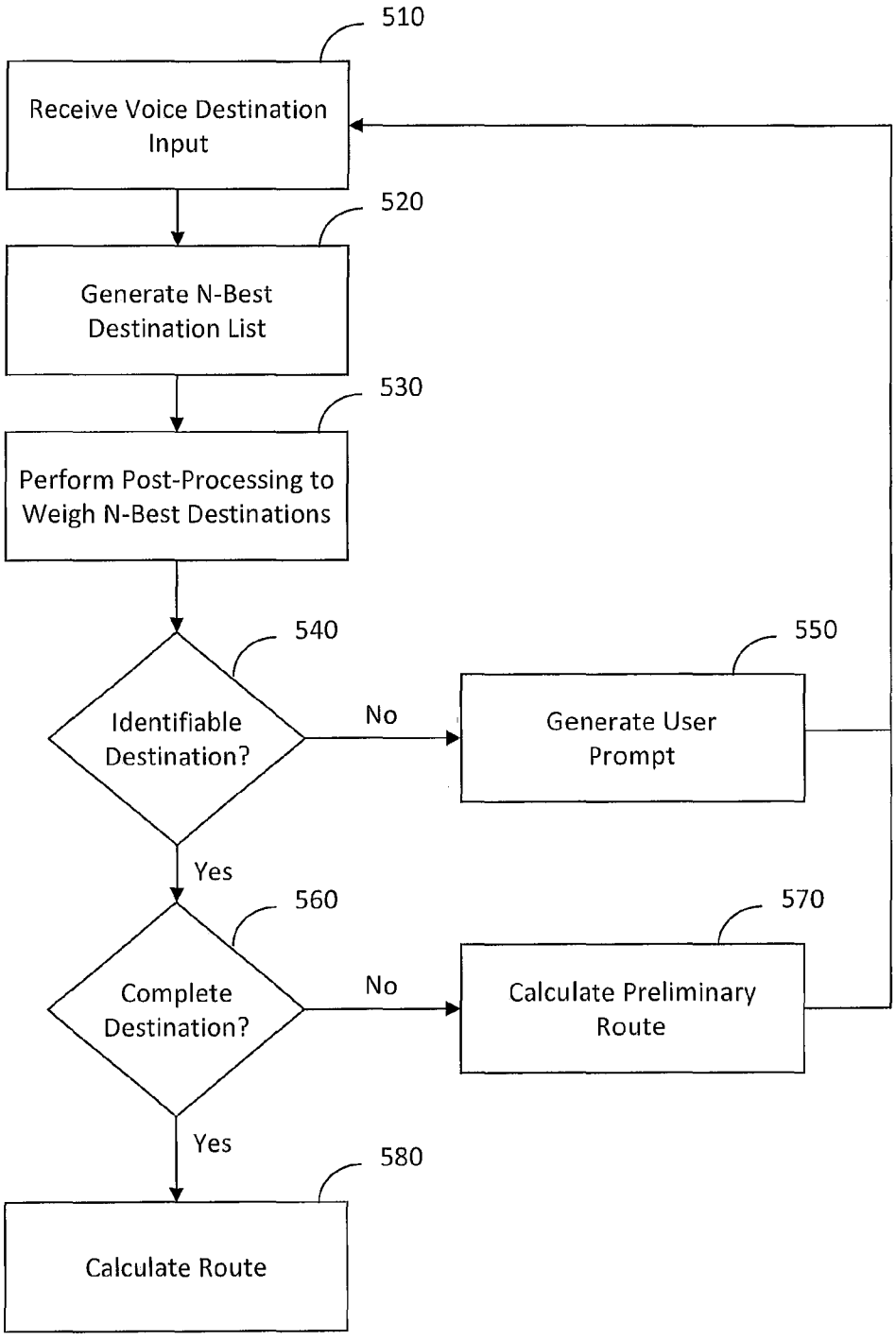

510
Receive Voice Destination Input
520
Generate N-Best Destination List
530
Perform Post-Processing to Weigh N-Best Destinations
540
Identifiable Destination?
No
550
Generate User Prompt
Yes
560
Complete Destination?
No
570
Calculate Preliminary Route
Yes
580
Calculate Route

Figure 5

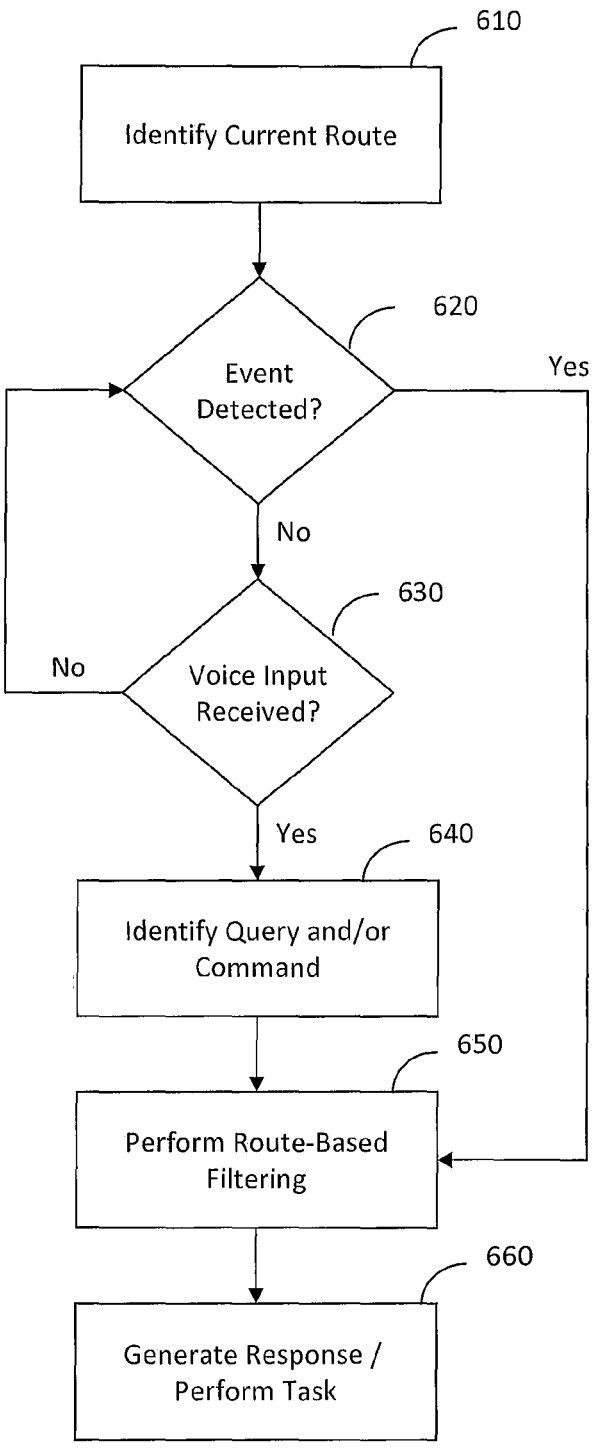

610
Identify Current Route
620
Event Detected?
Yes
No
630
Voice Input Received?
No
Yes
640
Identify Query and/or Command
650
Perform Route-Based Filtering
660
Generate Response / Perform Task

Figure 6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 08/69508

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 17/27 (2008.04)
USPC - 704/9
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC - 704/9

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC - 704/1, 2, 8, 9, 200, 206, 221, E15.045

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PubWEST (PGPB, USPT, EPAB, JPAB); Google Scholar
Search Terms Used: map, gps, speech, voice, talk, navigation, natural, conversational, grammar, interpret, software, computer, input, interface, user interface, refine, zoom, route, directions

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X -- Y	US 2007/0050191 A1 (WEIDER et al.) 01 March 2007 (01.03.2007), entire document, especially paras. [0061]-[0073], [0100], [0121], [0123], [0126], [0133], [0140], [0142], [0147], [0153], [0167], [0189]-[0194], [0210]-[0211], [0215]	1-42, 45-58 ---------- 43-44
Y	US 5,902,347 A (BACKMAN et al.) 11 May 1999 (11.05.1999), entire document, especially col 1 ln 32-34, col 4 ln 33-44	43-44
A	US 5,274,560 A (LARUE) 28 December 1993 (28.12.1993), entire document	1-58
A	US 6,598,018 B1 (JUNQUA) 22 July 2003 (22.07.2003), entire document	1-58

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

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

Date of the actual completion of the international search	Date of mailing of the international search report
25 September 2008 (25.09.2008)	02 OCT 2008
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

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

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
3 December 2009 (03.12.2009)

PCT

(10) International Publication Number
WO 2009/145796 A1

(51) International Patent Classification:
G06F 13/12 (2006.01)

(21) International Application Number:
PCT/US2008/069524

(22) International Filing Date:
9 July 2008 (09.07.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
12/127,343 27 May 2008 (27.05.2008) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [US/US]; 11980 NE 24th Street, Bellevue, Washington 98005 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **KENNEWICK, Robert** [US/US]; 13660 NE 42nd St., Bellevue, WA 98005 (US). **WEIDER, Chris** [US/US]; 10757 Greenwood Avenue N. #220, Seattle, Washington 98133 (US).

(74) Agents: **GATTO, James, G.** et al.; Pillsbury Winthrop Shaw Pittman LLP, P.O. Box 10500, McLean, Virginia 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) **Title**: SYSTEM AND METHOD FOR AN INTEGRATED, MULTI-MODAL, MULTI-DEVICE NATURAL LANGUAGE VOICE SERVICES ENVIRONMENT

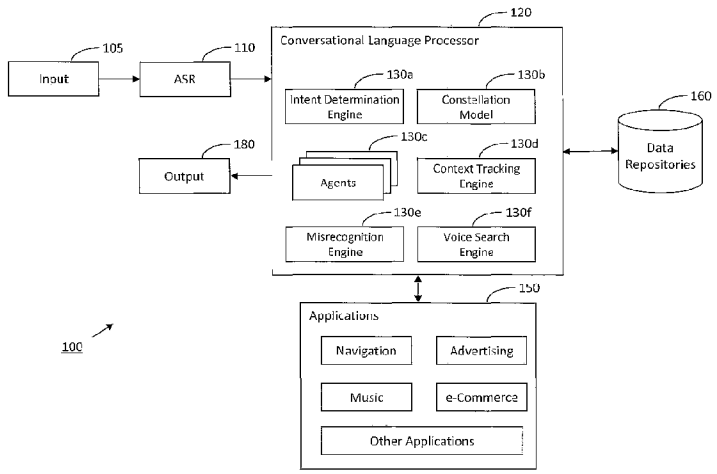


Figure 1

(57) **Abstract**: A system and method for an integrated, multi-modal, multi-device natural language voice services environment may be provided. In particular, the environment may include a plurality of voice-enabled devices each having intent determination capabilities for processing multi-modal natural language inputs in addition to knowledge of the intent determination capabilities of other devices in the environment. Further, the environment may be arranged in a centralized manner, a distributed peer-to-peer manner, or various combinations thereof. As such, the various devices may cooperate to determine intent of multi-modal natural language inputs, and commands, queries, or other requests may be routed to one or more of the devices best suited to take action in response thereto.

WO 2009/145796 A1

SYSTEM AND METHOD FOR AN INTEGRATED, MULTI-MODAL, MULTI-DEVICE NATURAL LANGUAGE VOICE SERVICES ENVIRONMENT

FIELD OF THE INVENTION

[001] The invention relates to an integrated voice services environment in which a plurality of devices can provide various voice services by cooperatively processing free form, multi-modal, natural language inputs, thereby facilitating conversational interactions between a user and one or more of the devices in the integrated environment.

BACKGROUND OF THE INVENTION

[002] As technology has progressed in recent years, consumer electronic devices have emerged to become nearly ubiquitous in the everyday lives of many people. To meet the increasing demand that has resulted from growth in the functionality and mobility of mobile phones, navigation devices, embedded devices, and other such devices, a wealth of features and functions are often provided therein in addition to core applications. Greater functionality also introduces the trade-offs, however, including learning curves that often inhibit users from fully exploiting all of the capabilities of their electronic devices. For example, many existing electronic devices include complex human to machine interfaces that may not be particularly user-friendly, which inhibits mass-market adoption for many technologies. Moreover, cumbersome interfaces often result in otherwise desirable features being buried (e.g., within menus that may be tedious to navigate), which has the tendency of causing many users to not use, or even know about, the potential capabilities of their devices.

[003] As such, the increased functionality provided by many electronic devices often tends to be wasted, as market research suggests that many users only use only a fraction of the features or applications available on a given device. Moreover, in a society where wireless networking and broadband access are increasingly prevalent, consumers tend to naturally desire seamless mobile capabilities from their electronic devices. Thus, as consumer demand intensifies for simpler mechanisms to interact with electronic devices, cumbersome interfaces that prevent quick and focused interaction can become an important concern. Accordingly, the ever-growing demand for mechanisms to use technology in intuitive ways remains largely unfulfilled.

[004] One approach towards simplifying human to machine interactions in electronic devices includes the use of voice recognition software, which can enable users to exploit features that

could otherwise be unfamiliar, unknown, or difficult to use. For example, a recent survey conducted by the Navteq Corporation, which provides data used in a variety of applications such as automotive navigation and web-based applications, demonstrates that voice recognition often ranks among the features most desired by consumers of electronic devices. Even so, existing voice user interfaces, when they actually work, still tend to require significant learning on the part of the user.

[005] For example, many existing voice user interface only support requests formulated according to specific command-and-control sequences or syntaxes. Furthermore, many existing voice user interfaces cause user frustration or dissatisfaction because of inaccurate speech recognition. Similarly, by forcing a user to provide pre-established commands or keywords to communicate requests in ways that a system can understand, existing voice user interfaces do not effectively engage the user in a productive, cooperative dialogue to resolve requests and advance a conversation towards a mutually satisfactory goal (e.g., when users may be uncertain of particular needs, available information, or device capabilities, among other things). As such, existing voice user interfaces tend to suffer from various drawbacks, including significant limitations on engaging users in a dialogue in a cooperative and conversational manner.

[006] Additionally, many existing voice user interfaces fall short in utilizing information distributed across various different domains or devices in order to resolve natural language voice-based inputs. Thus, existing voice user interfaces suffer from being constrained to a finite set of applications for which they have been designed, or to devices on which they reside. Although technological advancement has resulted in users often having several devices to suit their various needs, existing voice user interfaces do not adequately free users from device constraints. For example, users may be interested in services associated with different applications and devices, but existing voice user interfaces tend to restrict users from accessing the applications and devices as they see fit. Moreover, users typically can only practicably carry a finite number of devices at any given time, yet content or services associated with users' other devices that currently being used may be desired in various circumstances. Accordingly, although users tend to have varying needs, where content or services associated with different devices may be desired in various contexts or environments, existing voice technologies tend to fall short in providing an integrated environment in which users can request content or services associated with virtually any device or network. As such, constraints on information availability

and device interaction mechanisms in existing voice services environments tend to prevent users from experiencing technology in an intuitive, natural, and efficient way.

[007] Existing systems suffer from these and other problems.

SUMMARY OF THE INVENTION

[008] According to various aspects of the invention, a system and method for an integrated, multi-modal, multi-device natural language voice services environment may include a plurality of voice-enabled devices each having intent determination capabilities for processing multi-modal natural language inputs in addition to knowledge of the intent determination capabilities of other devices in the environment. Further, the environment may be arranged in a centralized manner, a distributed peer-to-peer manner, or various combinations thereof. As such, the various devices may cooperate to determine intent of multi-modal natural language inputs, and commands, queries, or other requests may be routed to one or more of the devices best suited to take action in response thereto.

[009] According to various aspects of the invention, the integrated natural language voice services environment arranged in the centralized manner includes an input device that receives a multi-modal natural language input, a central device communicatively coupled to the input device, and one or more secondary devices communicatively coupled to the central device. Each of the input device, the central device, and the one or more secondary devices may have intent determination capabilities for processing multi-modal natural language inputs. As such, an intent of a given multi-modal natural language input may be determined in the centralized manner by communicating the multi-modal natural language input from the input device to the central device. Thereafter, the central device may aggregate the intent determination capabilities of the input device and the one or more secondary devices and determine an intent of the multi-modal natural language input using the aggregated intent determination capabilities. The input device may then receive the determined intent from the central device and invoke at least one action at one or more of the input device, the central device, or the secondary devices based on the determined intent.

[010] According to various aspects of the invention, the integrated natural language voice services environment arranged in the distributed manner includes an input device that receives a multi-modal natural language input, a central device communicatively coupled to the input

device and one or more secondary devices communicatively coupled to the input device, wherein each of the input device and the one or more secondary devices may have intent determination capabilities for processing multi-modal natural language inputs, as in the centralized implementation. However, the distributed implementation may be distinct from the centralized implementation in that a preliminary intent of the multi-modal natural language input may be determined at the input device using local intent determination capabilities. The multi-modal natural language input may then be communicated to one or more of the secondary devices (e.g., when a confidence level of the intent determination at the input device falls below a given threshold). In such cases, each of the secondary devices determine an intent of the multi-modal natural language input using local intent determination capabilities. The input device collates the preliminary intent determination and the intent determinations of the secondary devices, and may arbitrate among the collated intent determinations to determine an actionable intent of the multi-modal natural input.

[011] According to various aspects of the invention, the integrated natural language voice services environment arranged in a manner that dynamically selects between a centralized model and a distributed model. For example, the environment includes an input device that receives a multi-modal natural language input one or more secondary devices communicatively coupled to the input device, each of which have intent determination capabilities for processing multi-modal natural language inputs. A constellation model may be accessible to each of the input device and the one or more secondary devices, wherein the constellation model describes the intent determination capabilities of the input device and the one or more secondary devices. The multi-modal natural language input can be routed for processing at one or more of the input device or the secondary devices to determine an intent thereof based on the intent determination capabilities described in the constellation model. For example, when the constellation model arranges the input device and the secondary devices in the centralized manner, one of the secondary devices may be designated the central device and the natural language input may be processed as described above. However, when the multi-modal natural language cannot be communicated to the central device, the constellation model may be dynamically rearranged in the distributed manner, whereby the input device and the secondary devices share knowledge relating to respective local intent determination capabilities and operate as cooperative nodes to

determine the intent of the multi-modal natural language input using the shared knowledge relating to local intent determination capabilities.

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

BRIEF DESCRIPTION OF THE DRAWINGS

[013] Figure 1 illustrates a block diagram of an exemplary multi-modal electronic device that may be provided in an integrated, multi-device natural language voice services environment, according to various aspects of the invention.

[014] Figure 2 illustrates a block diagram of an exemplary centralized implementation of the integrated, multi-modal, multi-device natural language voice service environment, according to various aspects of the invention.

[015] Figure 3 illustrates a flow diagram of an exemplary method for processing multi-modal, natural language inputs at an input device in the centralized implementation of the integrated, multi-modal, multi-device natural language voice service environment, according to various aspects of the invention.

[016] Figure 4 illustrates a flow diagram of an exemplary method for processing multi-modal, natural language inputs at a central device in the centralized implementation of the integrated, multi-modal, multi-device natural language voice service environment, according to various aspects of the invention.

[017] Figure 5 illustrates a flow diagram of an exemplary method for processing multi-modal, natural language inputs at a secondary device in the centralized implementation of the integrated, multi-modal, multi-device natural language voice service environment, according to various aspects of the invention.

[018] Figure 6 illustrates a block diagram of an exemplary distributed implementation of the integrated, multi-modal, multi-device natural language voice service environment, according to various aspects of the invention.

[019] Figure 7 illustrates a flow diagram of an exemplary method for processing multi-modal, natural language inputs at an input device in the distributed implementation of the integrated, multi-modal, multi-device natural language voice service environment, according to various aspects of the invention.

DETAILED DESCRIPTION

[020] According to various aspects of the invention, Figure 1 illustrates a block diagram of an exemplary multi-modal electronic device 100 that may be provided in a natural language voice services environment that includes one or more additional multi-modal devices (e.g., as illustrated in Figures 2 and 6). As will be apparent, the electronic device 100 illustrated in Figure 1 may be any suitable voice-enabled electronic device (e.g., a telematics device, a personal navigation device, a mobile phone, a VoIP node, a personal computer, a media device, an embedded device, a server, or another electronic device). The device 100 may include various components that collectively provide a capability to process conversational, multi-modal natural language inputs. As such, a user of the device 100 may engage in multi-modal conversational dialogues with the voice-enabled electronic device 100 to resolve requests in a free form, cooperative manner.

[021] For example, the natural language processing components may support free form natural language utterances to liberate the user from restrictions relating to how commands, queries, or other requests should be formulated. Rather, the user may employ any manner of speaking that feels natural in order to request content or services available through the device 100 (e.g., content or services relating to telematics, communications, media, messaging, navigation, marketing, information retrieval, etc.). For instance, in various implementations, the device 100 may process natural language utterances utilizing techniques described in co-pending U.S. Patent Application 10/452,147, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed June 3, 2003, and U.S. Patent Application 10/618,633, entitled "Mobile Systems and Methods for Responding to Natural Language Speech Utterance," filed June 15, 2003, the disclosures of which are hereby incorporated by reference in their entirety.

[022] Moreover, because the device 100 may be deployed in an integrated multi-device environment, the user may further request content or services available through other devices deployed in the environment. In particular, the integrated voice services environment may include a plurality of multi-modal devices, each of which include natural language components generally similar to those illustrated in Figure 1. The various devices in the environment may serve distinct purposes, however, such that available content, services, applications, or other capabilities may vary among the devices in the environment (e.g., core functions of a media

device may vary from those of a personal navigation device). Thus, each device in the environment, including device 100, may have knowledge of content, services, applications, intent determination capabilities, and other features available through the other devices by way of a constellation model 130b. Accordingly, as will be described in greater detail below, the electronic device 100 may cooperate with other devices in the integrated environment to resolve requests by sharing context, prior information, domain knowledge, short-term knowledge, long-term knowledge, and cognitive models, among other things.

[023] According to various aspects of the invention, the electronic device 100 may include an input mechanism 105 that can receive multi-modal natural language inputs, which include at least an utterance spoken by the user. As will be apparent, the input mechanism 105 may include any appropriate device or combination of devices capable of receiving a spoken input (e.g., a directional microphone, an array of microphones, or any other device that can generate encoded speech). Further, in various implementations, the input mechanism 105 can be configured to maximize fidelity of encoded speech, for example, by maximizing gain in a direction of the user, cancelling echoes, nulling point noise sources, performing variable rate sampling, or filtering environmental noise (e.g., background conversations). As such, the input mechanism 105 may generate encoded speech in a manner that can tolerate noise or other factors that could otherwise interfere with accurate interpretation of the utterance.

[024] Furthermore, in various implementations, the input mechanism 105 may include various other input modalities (i.e., the input mechanism 105 may be arranged in a multi-modal environment), in that non-voice inputs can be correlated and/or processed in connection with one or more previous, contemporaneous, or subsequent multi-modal natural language inputs. For example, the input mechanism 105 may be coupled to a touch-screen interface, a stylus and tablet interface, a keypad or keyboard, or any other suitable input mechanism, as will be apparent. As a result, an amount of information potentially available when processing the multi-modal inputs may be maximized, as the user can clarify utterances or otherwise provide additional information in a given multi-modal natural language input using various input modalities. For instance, in an exemplary illustration, the user could touch a stylus or other pointing device to a portion of a touch-screen interface of the device 100, while also providing an utterance relating to the touched portion of the interface (e.g., "Show me restaurants around here"). In this example, the natural language utterance may be correlated with the input received

via the touch-screen interface, resulting in "around here" being interpreted in relation to the touched portion of the interface (e.g., as opposed to the user's current location or some other meaning).

[025] According to various aspects of the invention, the device 100 may include an Automatic Speech Recognizer 110 that generates one or more preliminary interpretations of the encoded speech, which may be received from the input mechanism 105. For example, the Automatic Speech Recognizer 110 may recognize syllables, words, or phrases contained in an utterance using one or more dynamically adaptable recognition grammars. The dynamic recognition grammars may be used to recognize a stream of phonemes through phonetic dictation based on one or more acoustic models. Furthermore, as described in co-pending U.S. Patent Application 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed August 5, 2005, the disclosure of which is hereby incorporated by reference in its entirety, the Automatic Speech Recognizer 110 may be capable of multi-pass analysis, where a primary speech recognition engine may generate a primary interpretation of an utterance (e.g., using a large list dictation grammar) and request secondary transcription from one or more secondary speech recognition engines (e.g., using a virtual dictation grammar having decoy words for out-of-vocabulary words).

[026] Thus, the Automatic Speech Recognizer 110 may generate preliminary interpretations of an utterance in various ways, including exclusive use of a dictation grammar or virtual dictation grammar, or use of various combinations thereof (e.g., when the device 100 supports multi-pass analysis). In any event, the Automatic Speech Recognizer 110 may provide out-of-vocabulary capabilities and may tolerate portions of a speech signal being dropped, the user misspeaking, or other variables that may occur in natural language speech (e.g., stops and starts, stutters, etc.). Furthermore, the recognition grammars employed by the Automatic Speech Recognizer 110 may include vocabularies, dictionaries, syllables, words, phrases, or other information optimized according to various contextual or application-specific domains (e.g., navigation, music, movies, weather, shopping, news, languages, temporal or geographic proximities, or other suitable domains). Moreover, environmental knowledge (e.g., peer-to-peer affinities, capabilities of devices in the environment, etc.), historical knowledge (e.g., frequent requests, prior context, etc.), or other types of knowledge can be used to continually optimize the information contained in the recognition grammars on a dynamic basis.

[027] For example, information contained in the recognition grammars may be dynamically optimized to improve a likelihood of a given utterance being recognized accurately (e.g., following an incorrect interpretation of a word, the incorrect interpretation may be removed from the grammar to reduce a likelihood of the incorrect interpretation being repeated). Accordingly, the Automatic Speech Recognizer 110 may use a number of techniques to generate preliminary interpretations of natural language utterances, including those described, for example, in co-pending U.S. Patent Application 11/513,269, entitled "Dynamic Speech Sharpening," filed August 31, 2006, the disclosure of which is hereby incorporated by reference in its entirety. Furthermore, the techniques used by the Automatic Speech Recognizer 110 associated with the device 100 may be considered in defining intent determination capabilities of the device 100, and such capabilities may be shared with other devices in the environment to enable convergence of speech recognition throughout the environment (e.g., because various devices may employ distinct speech recognition techniques or have distinct grammars or vocabularies, the devices may share vocabulary translation mechanisms to enhance system-wide recognition).

[028] According to various aspects of the invention, the Automatic Speech Recognizer 110 may provide one or more preliminary interpretations of a multi-modal input, including an utterance contained therein, to a conversational language processor 120. The conversational language processor 120 may include various components that collectively operate to model everyday human-to-human conversations in order to engage in cooperative conversations with the user to resolve requests based on the user's intent. For example, the conversational language processor 120 may include, among other things, an intent determination engine 130a, a constellation model 130b, one or more domain agents 130c, a context tracking engine 130d, a misrecognition engine 130e, and a voice search engine 130f. Furthermore, the conversational language processor 120 may be coupled to one or more data repositories 160 and applications associated with one or more domains. Thus, the intent determination capabilities of the device 100 may be defined based on the data and processing capabilities of the Automatic Speech Recognizer 110 and the conversational language processor 120.

[029] More particularly, the intent determination engine 130a may establish meaning for a given multi-modal natural language input based on a consideration of the intent determination capabilities of the device 100 as well as the intent determination capabilities of other devices in the integrated voice services environment. For example, the intent determination capabilities of

the device 100 may be defined as a function of processing resources, storage for grammars, context, agents, or other data, and content or services associated with the device 100 (e.g., a media device with a small amount of memory may have a smaller list of recognizable songs than a device with a large amount of memory). Thus, the intent determination engine 130a may determine whether to process a given input locally (e.g., when the device 100 has intent determination capabilities that suggest favorable conditions for recognition), or whether to route information associated with the input to other devices, which may assist in determining the intent of the input.

[030] As such, to determine which device or combination of devices should process an input, the intent determination engine 130a may evaluate the constellation model 130b, which provides a model of the intent determination capabilities for each of the devices in the integrated voice services environment. For instance, the constellation model 130b may contain, among other things, knowledge of processing and storage resources available to each of the devices in the environment, as well as the nature and scope of domain agents, context, content, services, and other information available to each of the devices in the environment. As such, using the constellation model 130b, the intent determination engine 130a may be able to determine whether any of the other devices have intent determination capabilities that can be invoked to augment or otherwise enhance the intent determination capabilities of the device 100 (e.g., by routing information associated with a multi-modal natural language input to the device or devices that appear best suited to analyze the information and therefore determine an intent of the input). Accordingly, the intent determination engine 130a may establish the meaning of a given utterance by utilizing the comprehensive constellation model 130b that describes capabilities within the device 100 and across the integrated environment. The intent determination engine 130a may therefore optimize processing of a given natural language input based on capabilities throughout the environment (e.g., utterances may be processed locally to the device 100, routed to a specific device based on information in the constellation model 130b, or flooded to all of the devices in the environment in which case an arbitration may occur to select a best guess at an intent determination).

[031] Although the following discussion will generally focus on various techniques that can be used to determine the intent of multi-modal natural language inputs in the integrated multi-device environment, it will be apparent that the natural language processing capabilities of any

one of the devices may extend beyond the specific discussion that has been provided herein. As such, in addition to the co-pending U.S. Patent Applications referenced above, further natural language processing capabilities that may be employed include those described in co-pending U.S. Patent Application 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed August 5, 2005, U.S. Patent Application 11/200,164, entitled "System and Method of Supporting Adaptive Misrecognition in Conversational Speech," filed August 10, 2005, U.S. Patent Application 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," filed August 29, 2005, U.S. Patent Application 11/580,926, entitled "System and Method for a Cooperative Conversational Voice User Interface," filed October 16, 2006, U.S. Patent Application 11/671,526, entitled "System and Method for Selecting and Presenting Advertisements Based on Natural Language Processing of Voice-Based Input," filed February 6, 2007, and U.S. Patent Application 11/954,064, entitled "System and Method for Providing a Natural Language Voice User Interface in an Integrated Voice Navigation Services Environment," filed December 11, 2007, the disclosures of which are hereby incorporated by reference in their entirety.

[032] According to various aspects of the invention, Figure 2 illustrates a block diagram of an exemplary centralized implementation of the integrated, multi-modal, multi-device natural language voice service environment. As will be apparent from the further description to be provided herein, the centralized implementation of the integrated, multi-device voice services environment may enable a user to engage in conversational, multi-modal natural language interactions with any one of voice-enabled devices 210a-n or central voice-enabled device 220. As such, the multi-device voice services environment may collectively determine intent for any given multi-modal natural language input, whereby the user may request content or voice services relating to any device or application in the environment, without restraint.

[033] As illustrated in Figure 2, the centralized implementation of the multi-device voice service environment may include a plurality of voice-enabled devices 210a-n, each of which include various components capable of determining intent of natural language utterances, as described above in reference to Figure 1. Furthermore, as will be apparent, the centralized implementation includes a central device 220, which contains information relating to intent determination capabilities for each of the other voice-enabled devices 210a-n. For example, in various exemplary implementations, the central device 220 may be designated as such by virtue

of being a device most capable of determining the intent of an utterance (e.g., a server, home data center, or other device having significant processing power, memory resources, and communication capabilities making the device suitable to manage intent determination across the environment). In another exemplary implementation, the central device 220 may be dynamically selected based on one or more characteristics of a given multi-modal natural language input, dialogue, or interaction (e.g., a device may be designated as the central device 220 when a current utterance relates to a specific domain).

[034] In the centralized implementation illustrated in Figure 2, a multi-modal natural language input may be received at one of the voice-enabled devices 210a-n. Therefore, the receiving one of the devices 210a-n may be designated as an input device for that input, while the remaining devices 210a-n may be designated as secondary devices for that input. In other words, for any given multi-modal natural language input, the multi-device environment may include an input device that collects the input, a central device 220 that aggregates intent determination, inferencing, and processing capabilities for all of the devices 210a-n in the environment, and one or more secondary devices that may also be used in the intent determination process. As such, each device 210 in the environment may be provided with a constellation model that identifies all of the devices 210 having incoming and outgoing communication capabilities, thus indicating an extent to which other devices may be capable of determining intent for a given multi-modal natural language input. The constellation model may further define a location of the central device 220, which aggregates context, vocabularies, content, recognition grammars, misrecognitions, shared knowledge, intent determination capabilities, inferencing capabilities, and other information from the various devices 210a-n in the environment.

[035] Accordingly, as communication and processing capabilities permit, the central device 220 may be used as a recognizer of first or last resort. For example, because the central device 220 converges intent determination capabilities across the environment (e.g., by aggregating context, vocabularies, device capabilities, and other information from the devices 210a-n in the environment), inputs may be automatically routed to the central device 220 when used as a recognizer of first resort, or as a recognizer of last resort when local processing at the input device 210 cannot determine the intent of the input with a satisfactory level of confidence. However, it will also be apparent that in certain instances the input device 210 may be unable to make contact with the central device 220 for various reasons (e.g., a network connection may be

unavailable, or a processing bottleneck at the central device 220 may cause communication delays). In such cases, the input device 210 that has initiated contact with the central device 220 may shift into decentralized processing (e.g., as described in reference to Figure 6) and communicate capabilities with one or more of the other devices 210a-n in the constellation model. Thus, when the central device 220 cannot be invoked for various reasons, the remaining devices 210a-n may operate as cooperative nodes to determine intent in a decentralized manner.

[036] Additionally, in the multi-device voice services environment, the central device 220 and the various other devices 210a-n may cooperate to create a converged model of capabilities throughout the environment. For example, as indicated above, in addition to having intent determination capabilities based on processing resources, memory resources, and device capabilities, each of the devices 210a-n and the central device 220 may include various other natural language processing components. The voice services environment may therefore operate in an integrated manner by maintaining not only a complete model of data, content, and services associated with the various devices 210a-n, but also of other natural language processing capabilities and dynamic states associated with the various devices 210a-n. As such, the various devices 210a-n may operate with a goal of converging capabilities, data, states, and other information across the device, either on one device (e.g., the central device 220) or distributed among the various devices 210a-n (e.g., as in the decentralized implementation to be described in Figure 6).

[037] For example, as discussed above, each device 210 includes an Automatic Speech Recognizer, one or more dynamically adaptable recognition grammars, and vocabulary lists used to generate phonemic interpretations of natural language utterances. Moreover, each device 210 includes locally established context, which can range from information contained in a context stack, context and namespace variables, vocabulary translation mechanisms, short-term shared knowledge relating to a current dialogue or conversational interaction, long-term shared knowledge relating to a user's learned preferences over time, or other contextual information. Furthermore, each device 210 may have various services or applications associated therewith, and may perform various aspects of natural language processing locally. Thus, additional information to be converged throughout the environment may include partial or preliminary utterance recognitions, misrecognitions or ambiguous recognitions, inferencing capabilities, and

overall device state information (e.g., songs playing in the environment, alarms set in the environment, etc.).

[038] Thus, various data synchronization and referential integrity algorithms may be employed in concert by the various devices 210a-n and the central device 220 to provide a consistent worldview of the environment. For example, information may be described and transmitted throughout the environment for synchronization and convergence purposes using the Universal Plug and Play protocol designed for computer ancillary devices, although the environment can also operate in a peer-to-peer disconnected mode (e.g., when the central device 220 cannot be reached). However, in various implementations, the environment may also operate in a peer-to-peer mode regardless of the disconnected status, as illustrated in Figure 6, for example, when the devices 210a-n have sufficient commensurate resources and capabilities for natural language processing.

[039] In general, the algorithms for convergence in the environment can be executed at various intervals, although it may be desirable to limit data transmission so as to avoid processing bottlenecks. For example, because the convergence and synchronization techniques relate to natural language processing, in which any given utterance will typically be expressed over a course of several seconds, information relating to context and vocabulary need not be updated on a time frame of less than a few seconds. However, as communication capabilities permit, context and vocabulary could be updated more frequently to provide real-time recognition or the appearance of real-time recognition. In another implementation, the convergence and synchronization may be permitted to run until completion (e.g., when no requests are currently pending), or the convergence and synchronization may be suspended or terminated when a predetermined time or resource consumption limit has been reached (e.g., when the convergence relates to a pending request, an intent determination having a highest confidence level at the time of cut-off may be used).

[040] By establishing a consistent view of capabilities, data, states, and other information throughout the environment, an input device 210 may cooperate with the central device 220 and one or more secondary devices (i.e., one or more of devices 210a-n, other than the input device) in processing any given multi-modal natural language input. Furthermore, by providing each device 210 and the central device 220 with a constellation model that describes a synchronized state of the environment, the environment may be tolerant of failure by one or more of the

devices 210a-n, or of the central device 220. For example, if the input device 210 cannot communicate with the central device 220 (e.g., because of a server crash), the input device 210 may enter a disconnected peer-to-peer mode, whereby capabilities can be exchanged with one or more devices 210a-n with which communications remain available. To that end, when a device 210 establishes new information relating to vocabulary, context, misrecognitions, agent adaptation, intent determination capabilities, inferencing capabilities, or otherwise, the device 210 may transmit the information to the central device 220 for convergence purposes, as discussed above, in addition to consulting the constellation model to determine whether the information should be transmitted to one or more of the other devices 210a-n.

[041] For example, suppose the environment includes a voice-enabled mobile phone that has nominal functionality relating to playing music or other media, and which further has a limited amount of local storage space, while the environment further includes a voice-enabled home media system that includes a mass storage medium that provides dedicated media functionality. If the mobile phone were to establish new vocabulary, context, or other information relating to a song (e.g., a user downloads the song or a ringtone to the mobile phone while on the road), the mobile phone may transmit the newly established information to the home media system in addition to the central device 220. As such, by having a model of all of the devices 210a-n in the environment and transmitting new information to the devices where it will most likely be useful, the various devices may handle disconnected modes of operation when the central device 220 may be unavailable for any reason, while resources may be allocated efficiently throughout the environment.

[042] Thus, based on the foregoing discussion, it will be apparent that a centralized implementation of an integrated multi-device voice services environment may generally include a central device 220 operable to aggregate or converge knowledge relating to content, services, capabilities, and other information associated with various voice-enabled devices 210a-n deployed within the environment. In such centralized implementations, the central device 220 may be invoked as a recognizer of first or last resort, as will be described in greater detail with reference to Figures 3-5, and furthermore, the other devices 210a-n in the environment may be configured to automatically enter a disconnected or peer-to-peer mode of operation when the central device 220 cannot be invoked for any reason (i.e., devices may enter a decentralized or distributed mode, as will be described in greater detail with reference to Figures 6-7).

Knowledge and capabilities of each of the devices 210a-n may therefore be made available throughout the voice services environment in a centralized manner, a distributed manner, or various combinations thereof, thus optimizing an amount of natural language processing resources used to determine an intent of any given multi-modal natural language input.

[043] According to various aspects of the invention, Figure 3 illustrates a flow diagram of an exemplary method for processing multi-modal, natural language inputs at an input device in the centralized implementation of the integrated, multi-modal, multi-device natural language voice service environment. Similarly, Figures 4 and 5 illustrate corresponding methods associated with a central device and one or more secondary devices, respectively, in the centralized voice service environment. Furthermore, it will be apparent that the processing techniques described in relation to Figures 3-5 may generally be based on the centralized implementation illustrated in Figure 2 and described above, whereby the input device may be assumed to be distinct from the central device, and the one or more secondary devices may be assumed to be distinct from the central device and the input device. However, it will be apparent that various instances may involve a natural language input being received at the central device or at another device, in which case the techniques described in Figures 3-5 may be vary depending on circumstances of the environment (e.g., decisions relating to routing utterances to a specific device or devices may be made locally, collaboratively, or in other ways depending on various factors, such as overall system state, communication capabilities, intent determination capabilities, or otherwise).

[044] As illustrated in Figure 3, a multi-modal natural language input may be received at an input device in an operation 310. The multi-modal input may include at least a natural language utterance provided by a user, and may further include other input modalities such as audio, text, button presses, gestures, or other non-voice inputs. It will also be apparent that prior to receiving the natural language input in operation 310, the input device may be configured to establish natural language processing capabilities. For example, establishing natural language processing capabilities may include, among other things, loading an Automatic Speech Recognizer and any associated recognition grammars, launching a conversational language processor to handle dialogues with the user, and installing one or more domain agents that provide functionality for respective application domains or contextual domains (e.g., navigation, music, movies, weather, information retrieval, device control, etc.).

[045] The input device may also be configured to coordinate synchronization of intent determination capabilities, shared knowledge, and other information with the central device and the secondary devices in the environment prior to receiving the input at operation 310. For example, when the input device installs a domain agent, the installed domain agent may bootstrap context variables, semantics, namespace variables, criteria values, and other context related to that agent from other devices in the system. Similarly, misrecognitions may be received from the central device and the secondary devices in order to enable correction of agents that use information relevant to the received misrecognitions, and vocabularies and associated translation mechanisms may be synchronized among the devices to account for potential variations between the Automatic Speech Recognizers used by the various devices (e.g., each device in the environment cannot be guaranteed to use the same Automatic Speech Recognizer or recognition grammars, necessitating vocabulary and translation mechanisms to be shared among the devices that share intent determination capabilities).

[046] Upon establishing and synchronizing natural language processing capabilities and subsequently receiving a multi-modal natural language input in operation 310, the input device may determine whether the environment has been set up to automatically transmit the input to the central device in a decisional operation 320. In such a case, processing proceeds to an operation 360 for transmitting the input to the central device, which may then process the input according to techniques to be described in relation to Figure 4. If the environment has not been set up to automatically communicate the input to the central device, however, processing proceeds to an operation 330, where the input device performs transcription of the natural language utterance contained in the multi-modal input. For example, the input device may transcribe the utterance using the Automatic Speech Recognizer and recognition grammars associated therewith according to techniques described above and in the above-referenced U.S. Patent Applications.

[047] Subsequently, in an operation 340, an intent of the multi-modal natural language input may be determined at the input device using local natural language processing capabilities and resources. For example, any non-voice input modalities included in the input may be merged with the utterance transcription and a conversational language processor associated with the input device may utilize local information relating to context, domain knowledge, shared knowledge, context variables, criteria values, or other information useful in natural language

processing. As such, the input device may attempt to determine a best guess as to an intent of the user that provided the input, such as identifying a conversation type (e.g., query, didactic, or exploratory) or request that may be contained in the input (e.g., a command or query relating to one or more domain agents or application domains).

[048] The intent determination of the input device may be assigned a confidence level (e.g., a device having an Automatic Speech Recognizer that implements multi-pass analysis may assign comparatively higher confidence levels to utterance transcriptions created thereby, which may result in a higher confidence level for the intent determination). The confidence level may be assigned based on various factors, as described in the above-referenced U.S. Patent Applications. As such, a decisional operation 350 may include determining whether the intent determination of the input device meets an acceptable level of confidence. When the intent determination meets the acceptable level confidence, processing may proceed directly to an operation 380 where action may be taken in response thereto. For example, when the intent determination indicates that the user has requested certain information, one or more queries may be formulated to retrieve the information from appropriate information sources, which may include one or more of the other devices. In another example, when the intent determination indicates that the user has requested a given command (e.g., to control a specific device), the command may be routed to the appropriate device for execution.

[049] Thus, in cases where the input device can determine the intent of a natural language input without assistance from the central device or the secondary devices, communications and processing resources may be conserved by taking immediate action as may be appropriate. On the other hand, when the intent determination of the input device does not meet the acceptable level of confidence, decisional operation 350 may result in the input device requesting assistance from the central device in operation 360. In such a case, the multi-modal natural language input may be communicated to the central device in its entirety, whereby the central device processes the input according to techniques described in Figure 4. However, should transmission to the central device fail for some reason, the input device may shift into a disconnected peer-to-peer mode where one or more secondary devices may be utilized, as will be described below in relation to Figure 7. When transmission to the central device occurs without incident, however, the input device may receive an intent determination from the central device in an operation 370, and may further receive results of one or more requests that the central device was able to

resolve, or requests that the central device has formulated for further processing on the input device. As such, the input device may take action in operation 380 based on the information received from the central device in operation 370. For example, the input device may route queries or commands to local or remote information sources or devices based on the intent determination, or may present results of the requests processed by the central device to the user.

[050] Referring to Figure 4, the central device may receive the multi-modal natural language input from the input device in an operation 410. The central device, having aggregated context and other knowledge from throughout the environment, may thus transcribe the utterance in an operation 420 and determine an intent of the input from the transcribed utterance in an operation 430. As such, the central device may consider information relating to context, domain agents, applications, and device capabilities throughout the environment in determining the intent of the utterance, including identification of one or more domains relevant to the input. However, it will be apparent that utilizing information aggregated from throughout the environment may cause ambiguity or uncertainty in various instances (e.g., an utterance containing the word "traffic" may have a different intent in domains relating to movies, music, and navigation).

[051] As such, once the central device has attempted to determine the intent of the natural language input, a determination may be made in an operation 440 as to whether one or more secondary devices (i.e., other devices in the constellation besides the input device) may also be capable of intent determination in the identified domain or domains. When no such secondary devices can be identified, decisional operation 440 may branch directly to an operation 480 to send to the input device the determined intent and any commands, queries, or other requests identified from the determined intent.

[052] On the other hand, when one or more secondary devices in the environment have intent determination capabilities in the identified domain or domains, the natural language input may be sent to such secondary devices in an operation 450. The secondary devices may then determine an intent as illustrated in Figure 5, which may include techniques generally similar to those described above in relation to the input device and central device (i.e., the natural language input may be received in an operation 510, an utterance contained therein may be transcribed in an operation 520, and an intent determination made in an operation 530 may be returned to the central device in an operation 540).

[053] Returning to Figure 4, the central device may collate intent determination responses received from the secondary devices in an operation 460. For example, as indicated above, the central device may identify one or more secondary devices capable of determining intent in a domain that the central device has identified as being relevant to the natural language utterance. As will be apparent, the secondary devices invoked in operation 450 may often include a plurality of devices, and intent determination responses may be received from the secondary devices in an interleaved manner, depending on processing resources, communications throughput, or other factors (e.g., the secondary devices may include a telematics device having a large amount of processing power and a broadband network connection and an embedded mobile phone having less processing power and only a cellular connection, in which case the telematics device may be highly likely to provide results to the central device before the embedded mobile phone). Thus, based on potential variations in response time of secondary devices, the central device may be configured to place constraints on collating operation 460. For example, the collating operation 460 may be terminated as soon as an intent determination has been received from one of the secondary devices that meets an acceptable level of confidence, or the operation 460 may be cut off when a predetermined amount of time has lapsed or a predetermined amount of resources have been consumed. In other implementations, however, it will be apparent that collating operation 460 may be configured to run to completion, regardless of whether delays have occurred or suitable intent determinations have been received. Further, it will be apparent that various criteria may be used to determine whether or when to end the collating operation 460, including the nature of a given natural language input, dialogue, or other interaction, or system or user preferences, among other criteria, as will be apparent.

[054] In any event, when the collating operation 460 has completed, a subsequent operation 470 may include the central device arbitrating among the intent determination responses received from one or more of the secondary devices previously invoked in operation 450. For example, each of the invoked secondary devices that generate an intent determination may also assign a confidence level to that intent determination, and the central device may consider the confidence levels in arbitrating among the responses. Moreover, the central device may associate other criteria with the secondary devices or the intent determinations received from the secondary devices to further enhance a likelihood that the best intent determination will be used. For example, various ones of the secondary devices may only be invoked for partial recognition in

distinct domains, and the central device may aggregate and arbitrate the partial recognitions to create a complete transcription. In another example, a plurality of secondary devices may be invoked to perform overlapping intent determination, and the central device may consider capabilities of the secondary devices to weigh the respective confidence levels (e.g., when one of two otherwise identical secondary devices employs multi-pass speech recognition analysis, the secondary device employing the multi-pass speech recognition analysis may be weighed as having a higher likelihood of success). It will be apparent that the central device may be configured to arbitrate and select one intent determination from among all of the intent hypotheses, which may include the intent determination hypothesis generated by the central device in operation 430. Upon selecting the best intent determination hypothesis, the central device may then provide that intent determination to the input device in operation 480, as well as any commands, queries, or other requests that may be relevant thereto. The input device may then take appropriate action as described above in relation to Figure 3.

[055] According to various aspects of the invention, Figure 6 illustrates a block diagram of an exemplary distributed implementation of the integrated, multi-modal, multi-device natural language voice service environment. As described above, the distributed implementation may also be categorized as a disconnected or peer-to-peer mode that may be employed when a central device in a centralized implementation cannot be reached or otherwise does not meet the needs of the environment. The distributed implementation illustrated in Figure 6 may be generally operate with similar purposes as described above in relation to the centralized implementation (i.e., to ensure that the environment includes a comprehensive model of aggregate knowledge and capabilities of a plurality of devices 610a-n in the environment). Nonetheless, the distributed implementation may operate in a somewhat different manner, in that one or more of the devices 610a-n may be provided with the entire constellation model, or various aspects of the model may be distributed among the plurality of devices 610a-n, or various combinations thereof.

[056] Generally speaking, the plurality of voice-enabled devices 610a- may be coupled to one another by a voice services interface 630, which may include any suitable real or virtual interface (e.g., a common message bus or network interface, a service-oriented abstraction layer, etc.). The various devices 610a-n may therefore operate as cooperative nodes in determining intent for multi-modal natural language utterances received by any one of the devices 610. Furthermore,

the devices 610a-n may share knowledge of vocabularies, context, capabilities, and other information, while certain forms of data may be synchronized to ensure consistent processing among the devices 610a-n. For example, because natural language processing components used in the devices 610a-n may vary (e.g., different recognition grammars or speech recognition techniques may exist), vocabulary translation mechanisms, misrecognitions, context variables, criteria values, criteria handlers, and other such information used in the intent determination process should be synchronized to the extent that communication capabilities permit.

[057] By sharing intent determination capabilities, device capabilities, inferencing capabilities, domain knowledge, and other information, decisions as to routing an utterance to a specific one of the devices 610a-n may be made locally (e.g., at an input device), collaboratively (e.g., a device having particular capabilities relevant to the utterance may communicate a request to process the utterance), or various combinations thereof (e.g., the input device may consider routing to secondary devices only when an intent of the utterance cannot be determined). Similarly, partial recognition performed at one or more of the devices 610a-n may be used to determine routing strategies for further intent determination of the utterance. For example, an utterance that contains a plurality of requests relating to a plurality of different domains may be received at an input device that can only determine intent in one of the domains. In this example, the input device may perform partial recognition for the domain associated with the input device, and the partial recognition may also identify the other domains relevant to the utterance for which the input device does not have sufficient recognition information. Thus, the partial recognition performed by the input device may result in identification of other potentially relevant domains and a strategy may be formulated to route the utterance to other devices in the environment that include recognition information for those domains.

[058] As a result, multi-modal natural language inputs, including natural language utterances, may be routed among the various devices 610a-n in order to perform intent determination in a distributed manner. However, as the capabilities and knowledge held by any one of the devices 610a-n may vary, each of the devices 610a-n may be associated with a reliability factor for intent determinations generated by the respective devices 610a-n. As such, to ensure that final intent determinations can be relied upon with a sufficient level of confidence, knowledge may be distributed among the devices 610a-n to ensure that reliability metrics for intent determinations provided by each of the devices 610a-n are commensurable throughout the environment. For

example, additional knowledge may be provided to a device having a low intent determination reliability, even when such knowledge results in redundancy in the environment, to ensure commensurate reliability of intent determination environment-wide.

[059] Therefore, in distributed implementations of the integrated voice services environment, utterances may be processed in various ways, which may depend on circumstances at a given time (e.g., system states, system or user preferences, etc.). For example, an utterance may be processed locally at an input device and only routed to secondary devices when an intent determination confidence level falls below a given threshold. In another example, utterances may be routed to a specific device based on the modeling of knowledge and capabilities discussed above. In yet another example, utterances may be flooded among all of the devices in the environment, and arbitration may occur whereby intent determinations may be collated and arbitrated to determine a best guess at intent determination.

[060] Thus, utterances may be processed in various ways, including through local techniques, centralized techniques, distributed techniques, and various combinations thereof. Although many variations will be apparent, Figure 7 illustrates an exemplary method for combined local and distributed processing of multi-modal, natural language inputs in a distributed implementation of the voice service environment, according to various aspects of the invention. In particular, the distributed processing may begin in an operation 710, where a multi-modal natural language input may be received at an input device. The input device may then utilize various natural language processing capabilities associated therewith in an operation 720 to transcribe an utterance contained in the multi-modal input (e.g., using an Automatic Speech Recognizer and associated recognition grammars), and may subsequently determine a preliminary intent of the multi-modal natural language input in an operation 730. It will be apparent that operations 710 through 730 may generally be performed using local intent determination capabilities associated with the input device.

[061] Thereafter, the input device may invoke intent determination capabilities of one or more secondary devices in an operation 740. More particularly, the input device may provide information associated with the multi-modal natural language input to one or more of the secondary devices, which may utilize local intent determination capabilities to attempt to determine intent of the input using techniques as described in relation to Figure 5. It will also be apparent that, in various implementations, the secondary devices invoked in operation 740 may

include only devices having intent determination capabilities associated with a specific domain identified as being associated with the input. In any event, the input device may receive intent determinations from the invoked secondary devices in an operation 750, and the input device may then collate the intent determinations received from the secondary devices. The input device may then arbitrate among the various intent determinations, or may combine various ones of the intent determinations (e.g., when distinct secondary devices determine intent in distinct domains), or otherwise arbitrate among the intent determinations to determine a best guess at the intent of the multi-modal natural language input (e.g., based on confidence levels associated with the various intent determinations). Based on the determined intent, the input device may then take appropriate action in an operation 770, such as issuing one or more commands, queries, or other requests to be executed at one or more of the input device or the secondary devices.

[062] Furthermore, in addition to the exemplary implementations described above, various implementations may include a continuous listening mode of operation where a plurality of devices may continuously listen for multi-modal voice-based inputs. In the continuous listening mode, each of the devices in the environment may be triggered to accept a multi-modal input when one or more predetermined events occur. For example, the devices may each be associated with one or more attention words, such as "Phone, <multi-modal request>" for a mobile phone, or "Computer, <multi-modal request>" for a personal computer. When one or more of the devices in the environment recognize the associated attention word, keyword activation may result, where the associated devices trigger to accept the subsequent multi-modal request. Further, where a plurality of devices in a constellation may be listening, the constellation may use all available inputs to increase recognition rates.

[063] Moreover, it will be apparent that the continuous listening mode may be applied in centralized voice service environments, distributed centralized voice service environments, or various combinations thereof. For example, when each device in the constellation has a different attention word, any given device that recognizes an attention word may consult a constellation model to determine a target device or functionality associated with the attention word. In another example, when a plurality of devices in the constellation share one or more attention words, the plurality of devices may coordinate with one another to synchronize information for processing the multi-modal input, such as a start time for an utterance contained therein.

[064] Implementations of the invention may be made in hardware, firmware, software, or various combinations thereof. The invention may also be implemented as instructions stored on a machine-readable medium, which may be read and executed by one or more processors. A machine-readable medium may include various mechanisms for storing or transmitting information in a form readable by a machine (e.g., a computing device). For example, a machine-readable storage medium may include read only memory, random access memory, magnetic disk storage media, optical storage media, flash memory devices, and others, and a machine-readable transmission media may include forms of propagated signals, such as carrier waves, infrared signals, digital signals, and others. Further, firmware, software, routines, or instructions may be described in the above disclosure in terms of specific exemplary aspects and implementations of the invention, and performing certain actions. However, it will be apparent that such descriptions are merely for convenience and that such actions in fact result from computing devices, processors, controllers, or other devices executing the firmware, software, routines, or instructions.

[065] Aspects and implementations may be described as including a particular feature, structure, or characteristic, but every aspect or implementation may not necessarily include the particular feature, structure, or characteristic. Further, when a particular feature, structure, or characteristic has been described in connection with an aspect or implementation, it will be understood that such feature, structure, or characteristic may be included in connection with other aspects or implementations, whether or not explicitly described. Thus, various changes and modifications may be made to the preceding description without departing from the scope or spirit of the invention, and the specification and drawings should therefore be regarded as exemplary only, and the scope of the invention determined solely by the appended claims.

CLAIMS

What is claimed is:

1. A method for providing an integrated natural language voice services environment, the integrated voice services environment including an input device that receives a multi-modal natural language input, a central device communicatively coupled to the input device, and one or more secondary devices communicatively coupled to the central device, each of the input device, the central device, and the one or more secondary devices having intent determination capabilities for processing multi-modal natural language inputs, the method comprising:

communicating the multi-modal natural language input from the input device to the central device, the central device operable to:

aggregate the intent determination capabilities of the input device and the one or more secondary devices; and

determine an intent of the multi-modal natural language input received at the input device using the aggregated intent determination capabilities;

receiving the determined intent of the multi-modal natural language input at the input device; and

invoking at least one action at one or more of the input device, the central device, or the secondary devices based on the determined intent.

2. The method of claim 1, the intent determination capabilities for a given one of the input device, the central device, or the secondary devices based on at least one of processing power, storage resources, natural language processing capabilities, or local knowledge.

3. The method of claim 1, the central device further operable to:

communicate the multi-modal natural language input to each of the secondary devices, wherein each of the secondary devices determine an intent of the multi-modal natural language input received at the input device using local intent determination capabilities;

receive the intent determined by each of the secondary devices; and

arbitrate among the intent determinations of the secondary devices to determine the intent of the multi-modal natural input.

4. The method of claim 3, wherein the central device arbitrates among the intent determinations of the secondary devices based on respective confidence levels associated therewith.

5. The method of claim 4, the central device further operable to weigh the confidence level of the intent determinations of the secondary devices, wherein intent determinations of secondary devices that use multi-pass speech recognition techniques receive a higher weight.

6. The method of claim 1, further comprising sharing knowledge among the input device, the central device, and the secondary devices to update the intent determination capabilities thereof.

7. The method of claim 6, wherein the shared knowledge includes information relating to at least one of vocabularies, translation mechanisms, misrecognitions, context, or states of the devices in the environment.

8. The method of claim 1, the input device operable to:

determine an intent of the multi-modal natural language input received at the input device using local intent determination capabilities; and

invoke the at least one action without communicating the multi-modal natural language input to the central device when the intent determination of the input device meets a predetermined level of confidence.

9. The method of claim 1, wherein the voice services environment operates in a continuous listening mode.

10. A method for providing an integrated natural language voice services environment, the integrated voice services environment including an input device that receives a multi-modal natural language input and one or more secondary devices communicatively coupled to the input device, each of the input device and the one or more secondary devices having intent

determination capabilities for processing multi-modal natural language inputs, the method comprising:

determining a preliminary intent of the multi-modal natural language input, wherein the input device determines the preliminary intent using local intent determination capabilities;

communicating the multi-modal natural language input to one or more of the secondary devices, each of which determine an intent of the multi-modal natural language input using local intent determination capabilities;

collating the preliminary intent determination and the intent determinations of the secondary devices;

arbitrating among the collated intent determinations to determine an actionable intent of the multi-modal natural input; and

invoking at least one action at one or more of the input device or the secondary devices based on the determined actionable intent.

11. The method of claim 10, further comprising sharing knowledge among the input device and the secondary devices relating to respective local intent determination capabilities.

12. The method of claim 11, further comprising associating a reliability factor with the local intent determination capabilities of each of the input device and the one or more secondary devices, wherein an aggregation of the reliability factors for the local intent determination capabilities provide commensurable intent determination capabilities throughout the environment.

13. The method of claim 11, wherein the shared knowledge includes information relating to at least one of processing power, storage resources, natural language processing capabilities, or knowledge associated with the input device and the secondary devices.

14. The method of claim 10, further comprising sharing knowledge among the input device and the secondary devices to update the intent determination capabilities thereof.

15. The method of claim 14, wherein sharing the knowledge includes transmitting knowledge established on the input device to the one or more secondary devices.

16. The method of claim 14, wherein sharing the knowledge includes transmitting knowledge established on one of the secondary devices to the input device and to the secondary devices other than the device where the knowledge was established.

17. The method of claim 14, wherein the shared knowledge includes information relating to at least one of vocabularies, translation mechanisms, misrecognitions, context, or states of the devices in the environment.

18. The method of claim 10, wherein the voice services environment operates in a continuous listening mode.

19. A system for providing an integrated natural language voice services environment, the system comprising:

an input device that receives a multi-modal natural language input;

one or more secondary devices communicatively coupled to the input device, each of the input device and the one or more secondary devices having intent determination capabilities for processing multi-modal natural language inputs; and

a constellation model accessible to each of the input device and the one or more secondary devices, the constellation model describing the intent determination capabilities of the input device and the one or more secondary devices, the multi-modal natural language input routed for processing at one or more of the input device or the secondary devices to determine an intent thereof based on the intent determination capabilities described in the constellation model.

20. The system of claim 19, wherein the constellation model describes, for each of the input device and the one or more secondary devices, one or more of processing power, storage resources, natural language processing capabilities, or knowledge associated therewith.

21. The system of claim 19, wherein the constellation model dynamically selecting one of a centralized model or a distributed model for operating the environment.

22. The system of claim 21, the centralized model designating one of the secondary devices as a central device, the central device operable to:

aggregate the intent determination capabilities of the input device and the one or more secondary devices; and

determine the intent of the multi-modal natural language input using the aggregated intent determination capabilities.

23. The system of claim 22, wherein the constellation model selects the distributed model when the multi-modal natural language cannot be communicated to the central device.

24. The system of claim 23, wherein selecting the distributed model causes the input device and the one or more secondary devices to:

share knowledge relating to respective local intent determination capabilities; and

operate as cooperative nodes to determine the intent of the multi-modal natural language input using the shared knowledge relating to local intent determination capabilities.

25. The system of claim 24, wherein operating as cooperative nodes includes determining whether the intent of the multi-modal natural language input should be determined by the input device, by one or more of the secondary devices, or by each of the input device and the one or more secondary devices.

26. The system of claim 19, further comprising at least one domain agent associated with the input device, wherein the at least one domain agent bootstraps information relating to a context associated with the domain agent from one or more of the secondary devices.

27. The system of claim 19, wherein the voice services environment operates in a continuous listening mode.

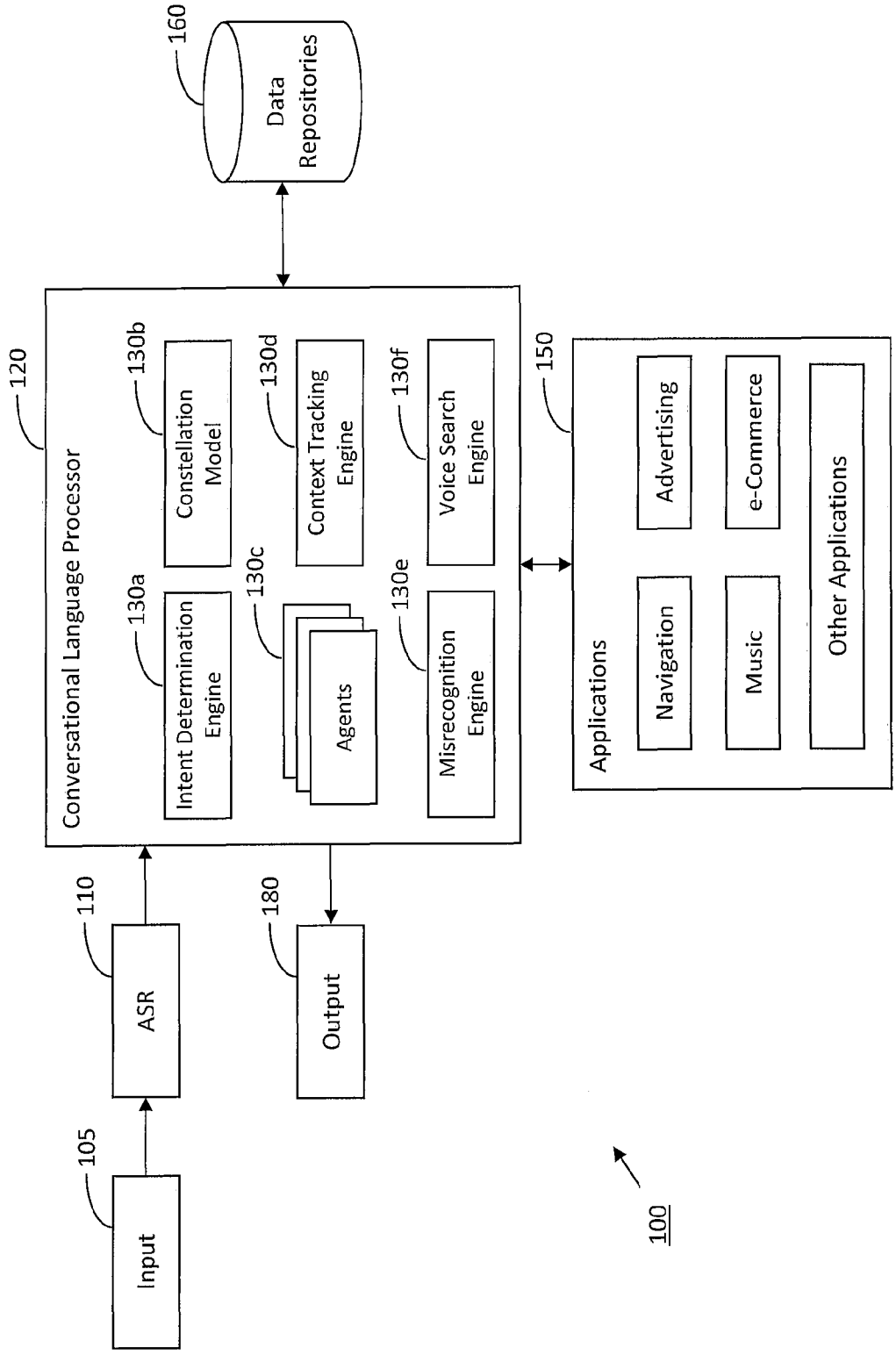

160
Data Repositories
120
Conversational Language Processor
130b
Constellation Model
130d
Context Tracking Engine
130f
Voice Search Engine
150
130a
Intent Determination Engine
130c
Agents
130e
Misrecognition Engine
Applications
Advertising
Navigation
e-Commerce
Music
Other Applications
110
ASR
180
Output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

Figure 1

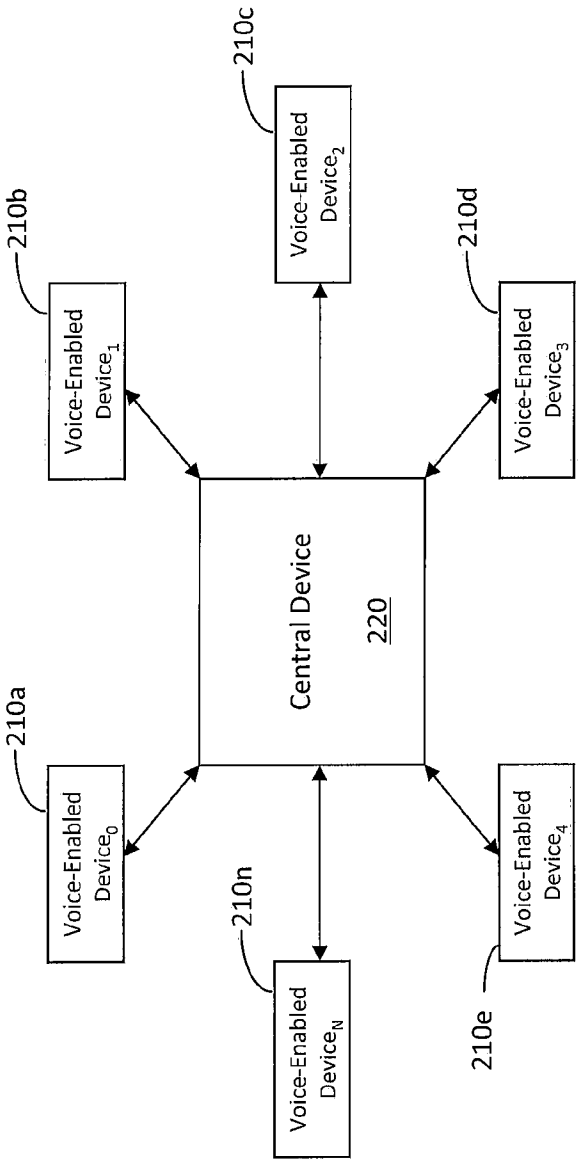

210c
Voice-Enabled Device$_2$
210b
Voice-Enabled Device$_1$
210d
Voice-Enabled Device$_3$
Central Device
220
210a
Voice-Enabled Device$_0$
210n
Voice-Enabled Device$_N$
Voice-Enabled Device$_4$
210e

Figure 2

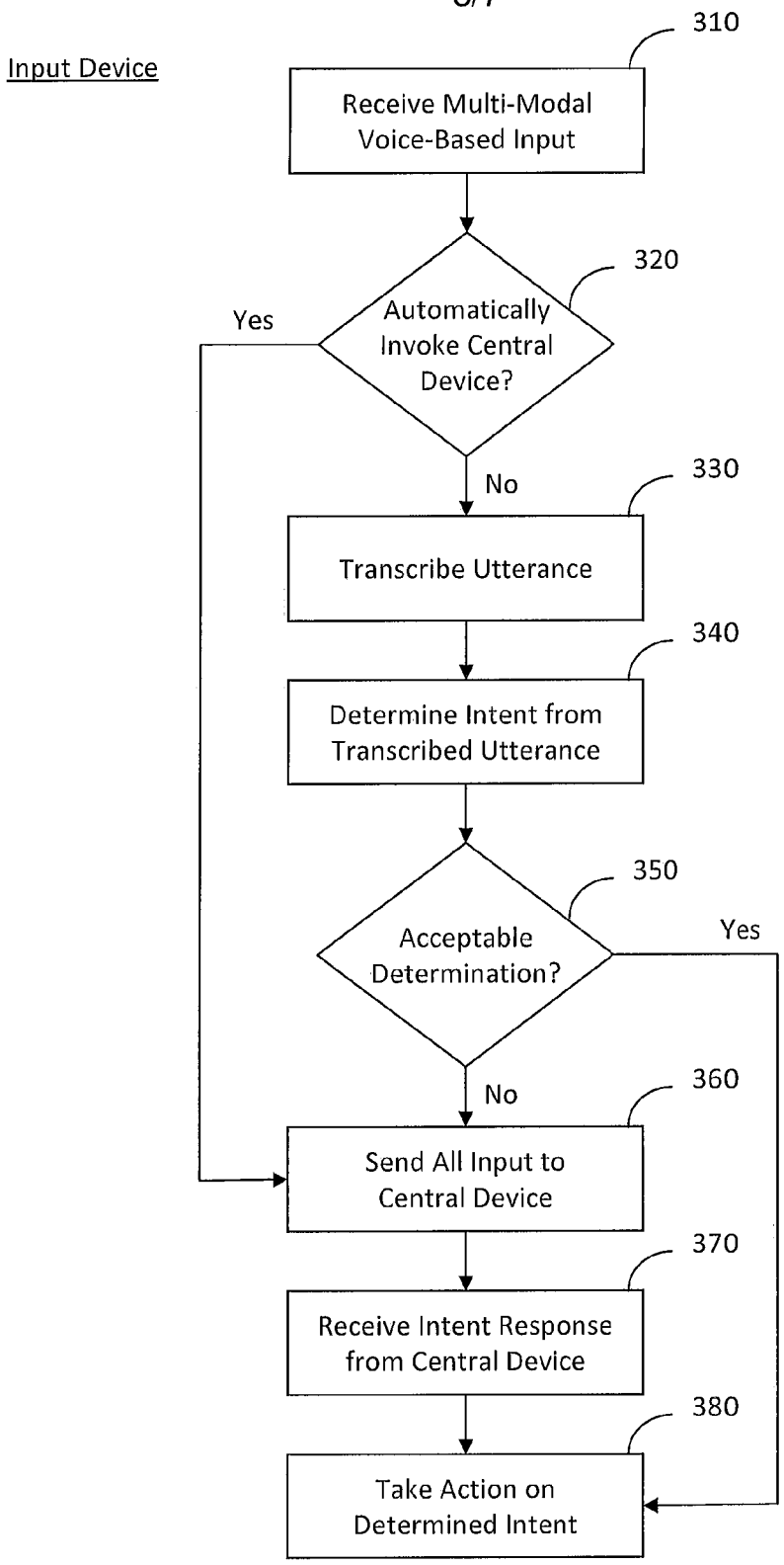

Input Device
310
Receive Multi-Modal Voice-Based Input
320
Automatically Invoke Central Device?
Yes
No
330
Transcribe Utterance
340
Determine Intent from Transcribed Utterance
350
Acceptable Determination?
Yes
No
360
Send All Input to Central Device
370
Receive Intent Response from Central Device
380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

Figure 3

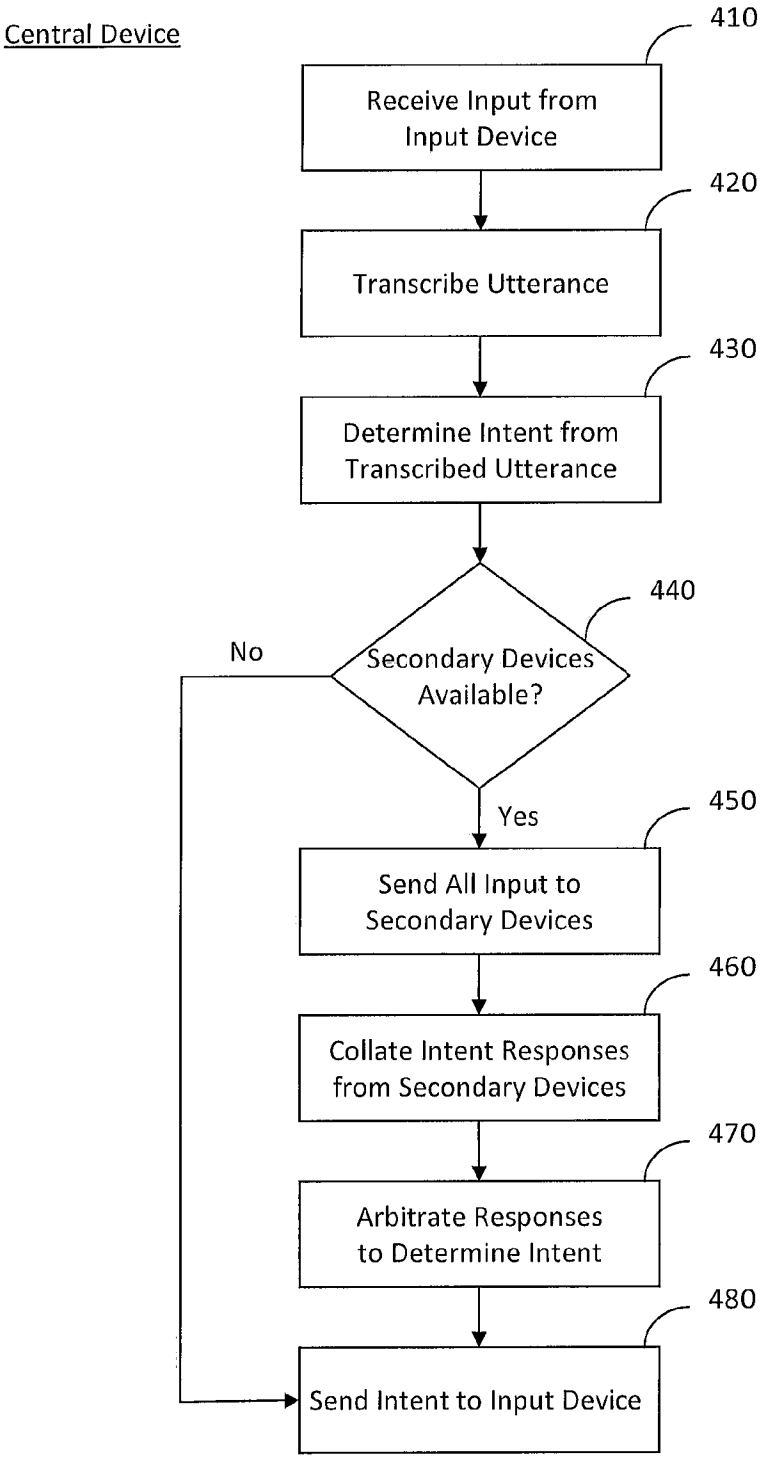

Central Device
410
Receive Input from Input Device
420
Transcribe Utterance
430
Determine Intent from Transcribed Utterance
440
Secondary Devices Available?
No
Yes
450
Send All Input to Secondary Devices
460
Collate Intent Responses from Secondary Devices
470
Arbitrate Responses to Determine Intent
480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

Figure 4

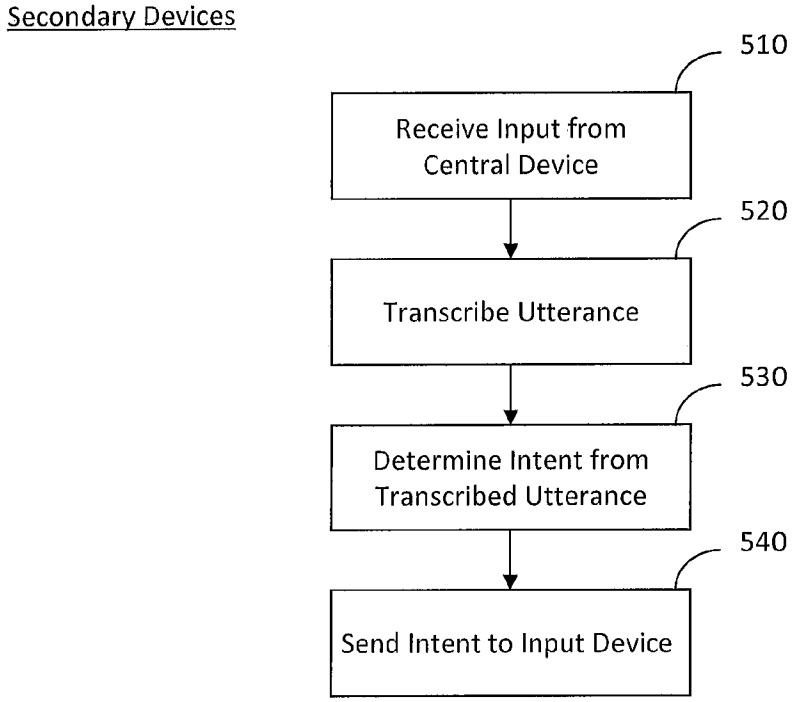

Secondary Devices
510
Receive Input from Central Device
520
Transcribe Utterance
530
Determine Intent from Transcribed Utterance
540
Send Intent to Input Device

Figure 5

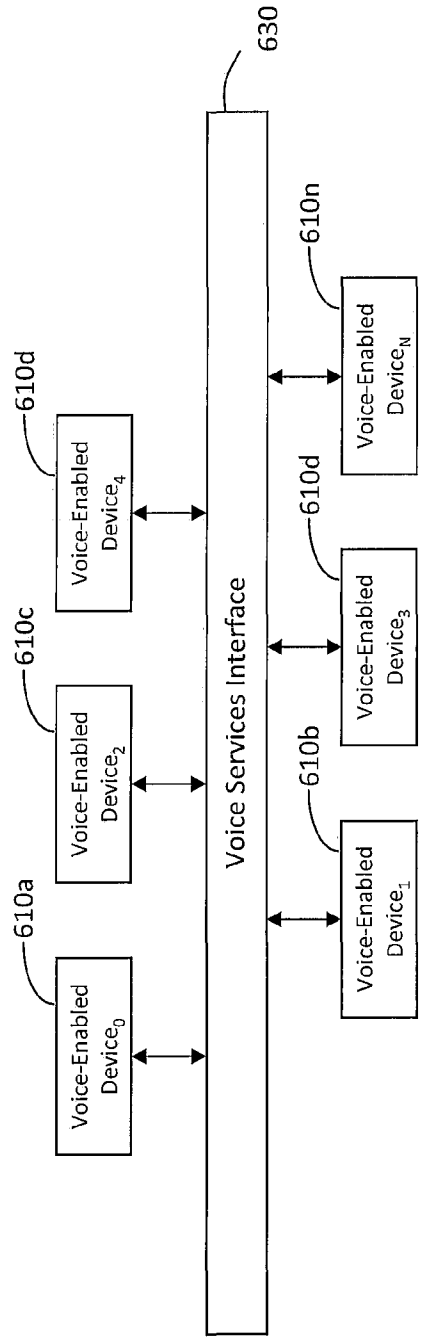

630
Voice Services Interface
610a
Voice-Enabled Device0
610b
Voice-Enabled Device1
610c
Voice-Enabled Device2
610d
Voice-Enabled Device3
610d
Voice-Enabled Device4
610n
Voice-Enabled DeviceN

Figure 6

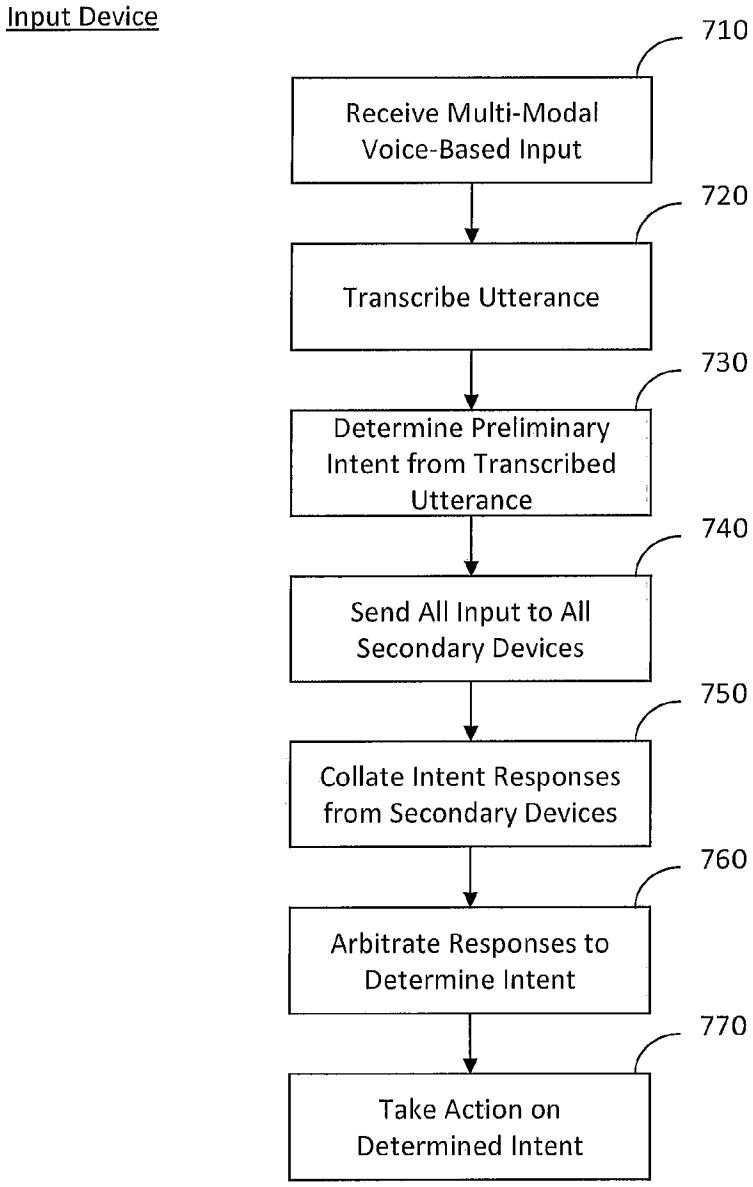

Input Device
710
Receive Multi-Modal Voice-Based Input
720
Transcribe Utterance
730
Determine Preliminary Intent from Transcribed Utterance
740
Send All Input to All Secondary Devices
750
Collate Intent Responses from Secondary Devices
760
Arbitrate Responses to Determine Intent
770
Take Action on Determined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

Figure 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 08/69524

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 13/12 (2008.04)
USPC - 710/65
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC: 710/65

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 710/11, 14, 37, 50, 73; 370/353, 458, 448; 379/907-908; 704/200, 246, 251-252 (see text terms below)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PubWEST (PGPB, USPT, USOC, EPAB, JPAB); DialogPRO; Google (incl. Patents, Scholar) - natural language, multi-modal, central, input, device, client, server, process, distributed, local, distributed system, local, context, speech

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 7,376,645 B2 (BERNARD) 20 May 2008 (20.05.2008), Abstract, FIGS.1-3; and Col 1, ln 19-30; Col 1, ln 35 to Col 2, ln 2; Col 2, ln 21 to Col 4, ln 16; Col 4, ln 21-56, 60; Col 5, ln 10; Col 6, ln 1-5, 7, 22-23; Col 6, ln 37 to Col 7, ln 4; Col 7, ln 13-26; Col 7, ln 52 to Col 10, ln 37	1-27
A	US 2004/0025115 A1 (SIENEL et al.) 5 February 2004 (05.02.2004)	1-27
A	US 2007/0050191 A1 (WEIDER et al.) 1 March 2007 (01.03.2007)	1-27
A	US 2004/0166832 A1 (PORTMAN et al.) 26 August 2004 (26.08.2004)	1-27
A	US 6,615,172 B1 (BENNETT et al.) 2 September 2003 (02.09.2003)	1-27

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

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

Date of the actual completion of the international search	Date of mailing of the international search report
04 October 2008 (04.10.2008)	09 OCT 2008
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

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

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 August 2010 (26.08.2010)

PCT

(10) International Publication Number
WO 2010/096752 A1

(51) International Patent Classification:
G10L 15/00 (2006.01)

(21) International Application Number:
PCT/US2010/024895

(22) International Filing Date:
22 February 2010 (22.02.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
12/389,678 20 February 2009 (20.02.2009) US

(71) Applicant *(for all designated States except US)*: **VOICEBOX TECHNOLOGIES, INC.** [US/US]; 11980 NE 24th Street, Suite 100, Bellevue, WA 98005 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **BALDWIN, Larry** [US/US]; 23125 SE 243rd Place, Maple Valley, WA 98038 (US). **WEIDER, Chris** [US/US]; 10757 Greenwood Avenue North #220, Seattle, WA 98133 (US).

(74) Agents: **ALI, Syed, Jafar** et al.; Pillsbury Winthrop Shaw Pittman LLP, P.O. Box 10500, McLean, VA 22102 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

Published:

— *with international search report (Art. 21(3))*

(54) **Title**: SYSTEM AND METHOD FOR PROCESSING MULTI-MODAL DEVICE INTERACTIONS IN A NATURAL LANGUAGE VOICE SERVICES ENVIRONMENT

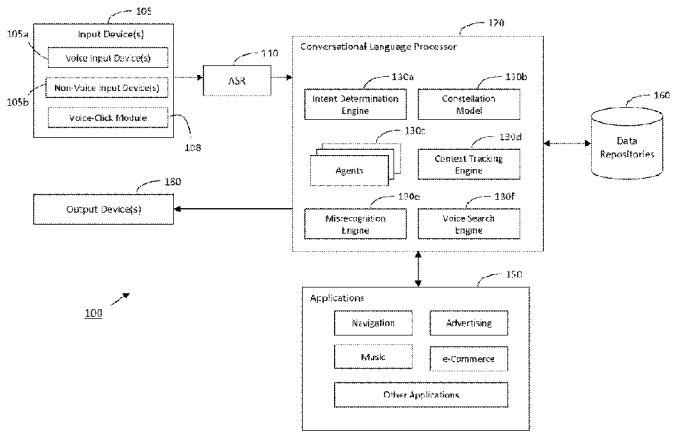


Figure 1

(57) **Abstract**: A system and method for processing multi-modal device interactions in a natural language voice services environment may be provided. In particular, one or more multi-modal device interactions may be received in a natural language voice services environment that includes one or more electronic devices. The multi-modal device interactions may include a non-voice interaction with at least one of the electronic devices or an application associated therewith, and may further include a natural language utterance relating to the non-voice interaction. Context relating to the non-voice interaction and the natural language utterance may be extracted and combined to determine an intent of the multi-modal device interaction, and a request may then be routed to one or more of the electronic devices based on the determined intent of the multi-modal device interaction.

WO 2010/096752 A1

SYSTEM AND METHOD FOR PROCESSING MULTI-MODAL DEVICE INTERACTIONS IN A NATURAL LANGUAGE VOICE SERVICES ENVIRONMENT

CROSS-REFERENCE TO RELATED APPLICATIONS

[001] This application claims the benefit of U.S. Patent Application Serial No. 12/389,678, entitled "SYSTEM AND METHOD FOR PROCESSING MULTI-MODAL DEVICE INTERACTIONS IN A NATURAL LANGUAGE VOICE SERVICES ENVIRONMENT", filed February 20, 2009, which is hereby incorporated by reference in its entirety.

FIELD OF THE INVENTION

[002] The invention relates to an integrated natural language voice services environment for processing multi-modal interactions with or one or more devices and/or applications, wherein the multi-modal interactions may provide additional context for cooperatively interpreting and otherwise processing a natural language utterance accompanying the multi-modal interactions.

BACKGROUND OF THE INVENTION

[003] As technology has progressed in recent years, consumer electronic devices have emerged to become nearly ubiquitous in the everyday lives of many people. To meet the increasing demand that has resulted from growth in the functionality and mobility of mobile phones, navigation devices, embedded devices, and other such devices, many devices offer a wealth of features and functions in addition to core applications. Greater functionality also introduces trade-offs, however, including learning curves that often inhibit users from fully exploiting all of the capabilities of their electronic devices. For example, many existing electronic devices include complex human to machine interfaces that may not be particularly user-friendly, which can inhibit mass-market adoption for many technologies. Moreover, cumbersome interfaces often result in otherwise desirable features being difficult to find or use (e.g., because of menus that are complex or otherwise tedious to navigate). As such, many users tend not to use, or even know about, many of the potential capabilities of their devices.

[004] As such, the increased functionality of electronic devices often tends to be wasted, as market research suggests that many users only use only a fraction of the features or applications available on a given device. Moreover, in a society where wireless networking and broadband access are increasingly prevalent, consumers tend to naturally desire seamless

mobile capabilities from their electronic devices. Thus, as consumer demand intensifies for simpler mechanisms to interact with electronic devices, cumbersome interfaces that prevent quick and focused interaction become an important concern. Nevertheless, the ever-growing demand for mechanisms to use technology in intuitive ways remains largely unfulfilled.

[005] One approach towards simplifying human to machine interactions in electronic devices has included the use of voice recognition software, which has the potential to enable users to exploit features that would otherwise be unfamiliar, unknown, or difficult to use. For example, a recent survey conducted by the Navteq Corporation, which provides data used in a variety of applications such as automotive navigation and web-based applications, demonstrates that voice recognition often ranks among the features most desired by consumers of electronic devices. Even so, existing voice user interfaces, when they actually work, still require significant learning on the part of the user.

[006] For example, many existing voice user interface only support requests formulated according to specific command-and-control sequences or syntaxes. Furthermore, many existing voice user interfaces cause user frustration or dissatisfaction because of inaccurate speech recognition. Similarly, by forcing a user to provide pre-established commands or keywords to communicate requests in ways that a system can understand, existing voice user interfaces do not effectively engage the user in a productive, cooperative dialogue to resolve requests and advance a conversation towards a satisfactory goal (e.g., when users may be uncertain of particular needs, available information, device capabilities, etc.). As such, existing voice user interfaces tend to suffer from various drawbacks, including significant limitations on engaging users in a dialogue in a cooperative and conversational manner.

[007] Additionally, many existing voice user interfaces fall short in utilizing information distributed across different domains, devices, and applications in order to resolve natural language voice-based inputs. Thus, existing voice user interfaces suffer from being constrained to a finite set of applications for which they have been designed, or to devices on which they reside. Although technological advancement has resulted in users often having several devices to suit their various needs, existing voice user interfaces do not adequately free users from device constraints. For example, users may be interested in services associated with different applications and devices, but existing voice user interfaces tend to restrict users from accessing the applications and devices as they see fit. Moreover, users typically can only practicably carry

a finite number of devices at any given time, yet content or services associated with users' other devices currently being used may be desired in various circumstances.

[008] Accordingly, although users tend to have varying needs, where content or services associated with different devices may be desired in various contexts or environments, existing voice technologies tend to fall short in providing an integrated environment in which users can request content or services associated with virtually any device or network. As such, constraints on information availability and device interaction mechanisms in existing voice services environments tend to prevent users from experiencing technology in an intuitive, natural, and efficient way. For instance, when a user wishes to perform a given function using a given electronic device but does not necessarily know how to go about performing the function, the user typically cannot engage in a multi-modal interaction with the device to simply utter words in natural language to request the function.

[009] Furthermore, relatively simple functions can often be tedious to perform using electronic devices that do not have voice recognition capabilities. For example, purchasing new ring-tones for a mobile phone tends to be a relatively straightforward process, but users must typically navigate several menus and press many different buttons in order to complete the process. As such, it becomes apparent that interaction with electronic devices could be far more efficient if users were able to use natural language to exploit buried or otherwise difficult to use functionality. Existing systems suffer from these and other problems.

SUMMARY OF THE INVENTION

[010] According to one aspect of the invention, a system and method for processing multi-modal device interactions in a natural language voice services environment may be provided. In particular, one or more multi-modal interactions may be received in a natural language voice services environment that includes one or more electronic devices. The multi-modal device interactions may include a user engaging in a non-voice interaction with one or more of the electronic devices or applications associated with the devices, while also providing a natural language utterance in relation to the non-voice interaction. For example, the non-voice device interaction may comprise the user selecting a particular segment, item, data, point of focus, or attention focus, or otherwise engaging in one or more unique and distinguishable interactions with the electronic devices or applications associated therewith. As such, context may be

extracted from the natural language utterance, and the non-voice device interaction may provide additional context for the natural language utterance. The context of the utterance and the non-voice device interaction may then be combined to determine an intent of the multi-modal device interaction, wherein one or more of the electronic devices may process a request based on the intent of the multi-modal device interaction.

[011] According to one aspect of the invention, at least one of the electronic devices may include an input device configured to receive voice-based inputs. In one implementation, the voice-based input device may be signaled to capture the natural language utterance in response to detecting the non-voice interaction with the one or more electronic devices or applications. Furthermore, the natural language voice services environment may include one or more listeners established for the electronic devices and the associated applications, wherein the listeners may be configured to detect the non-voice interaction with the electronic devices or applications. As such, information relating to the non-voice interaction and the accompanying natural language utterance may be aligned to enable cooperative processing of the utterance and the non-voice device interaction.

[012] According to one aspect of the invention, at least one transaction lead may be generated based on the intent of the multi-modal device interaction. For example, a further multi-modal device interaction may be received, wherein the further multi-modal device interaction may relate to the transaction lead generated for the first multi-modal device interaction. At least one request may then be routed to one or more of the electronic devices based on an intent determined for the further multi-modal device interaction, whereby a transaction click-through may be processed in response to receiving the device interaction relating to the generated transaction lead. For example, the transaction lead may include an advertisement or a recommendation that is selected based on the intent of the original multi-modal device interaction, while the further multi-modal device interaction may include a user selecting the advertisement or recommendation. Thus, the selection of the advertisement or recommendation may be considered a transaction click-through, which may generate revenue for a particular entity (e.g., a provider of the natural language voice services environment).

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

BRIEF DESCRIPTION OF THE DRAWINGS

[014] Figure 1 illustrates a block diagram of an exemplary system for processing multi-modal device interactions in a natural language voice services environment, according to various aspects of the invention.

[015] Figure 2 illustrates a block diagram of an exemplary method for synchronizing multi-modal devices in a natural language voice service environment, according to various aspects of the invention.

[016] Figure 3 illustrates a flow diagram of an exemplary method for processing multi-modal device interactions in a natural language voice service environment, according to various aspects of the invention.

[017] Figure 4 illustrates a flow diagram of an exemplary method for processing multi-modal device interactions to generate one or more transaction leads in a natural language voice service environment, according to various aspects of the invention.

DETAILED DESCRIPTION

[018] According to various aspects of the invention, Figure 1 illustrates a block diagram of an exemplary system 100 for processing multi-modal device interactions in a natural language voice services environment. As will be apparent from the further description to be provided herein, the system 100 illustrated in Figure 1 may include an input device 105, or a combination of input devices 105, which enable users to interact with the system 100 in a multi-modal manner. In particular, the system 100 may include various natural language processing components, including at least a voice-click module 108, which can collectively process a user's multi-modal interactions with one or more input devices 105. For example, in one implementation, the input devices 105 may include any suitable combination of at least one voice input device 105a (e.g., a microphone) and at least one non-voice input device 105b (e.g., a mouse, touch-screen display, wheel selector, etc.). As such, the input devices 105 may include any suitable combination of electronic devices having mechanisms for receiving both voice-based and non-voice-based inputs (e.g., a microphone coupled to one or more of a telematics device, personal navigation device, mobile phone, VoIP node, personal computer, media device, embedded device, server, or other electronic device). As such, the system 100 may enable users to engage in multi-modal conversational interactions with the one or more

electronic input devices 105 or applications associated with the electronic devices 105, wherein the system 100 may process the device interactions in a free-form and cooperative manner suitable for routing tasks or otherwise resolving requests.

[019] As indicated above, in one implementation, the system may include various natural language processing components that can support free-form utterances and/or other forms of device interactions, which may liberate users from restrictions relating to the manner of formulating commands, queries, or other requests. As such, users may interact with the input devices 105 using any manner of speaking into the voice input device 105a or otherwise interacting with the non-voice input devices 105b in order to request content or services available in the system 100. For instance, a user may request any content or services that may be available in the system 100 by providing a natural language utterance to the voice input device 105a. In one implementation, the utterance may then be processed using techniques described in U.S. Patent No. 7,398,209, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," issued July 8, 2008, and U.S. Patent Application 10/618,633, entitled "Mobile Systems and Methods for Responding to Natural Language Speech Utterance," filed June 15, 2003, the disclosures of which are hereby incorporated by reference in their entirety. In addition, the user may interact with one or more of the non-voice input devices 105b to provide further context or other information relating to the utterance and/or the requested content or services.

[020] In one implementation, the system 100 may be coupled to various other systems that include additional multi-modal devices, with the other systems having natural language processing capabilities similar to those shown in Figure 1. The system 100 may therefore provide an interface to a multi-device environment in which users may request content or services available through the various additional devices in the environment. For example, in one implementation, the system 100 may include a constellation model 130b that provides knowledge relating to content, services, applications, intent determination capabilities, and other features available through the other systems and devices in the environment. For example, in one implementation, the system 100 may interact with devices, applications, or other systems in the environment to cooperatively resolve requests, as described in co-pending U.S. Patent Application 12/127,343, entitled "System and Method for an Integrated, Multi-Modal, Multi-Device Natural Language Voice Services Environment," filed May 27, 2008, the

disclosure of which is hereby incorporated by reference in its entirety. For example, the multi-device environment may share information among various systems and devices to provide a cooperative environment for resolving requests, wherein the shared information may relate to aspects such as device capabilities, context, prior interactions, domain knowledge, short-term knowledge, long-term knowledge, and cognitive models, among other things.

[021] As indicated above, the system 100 illustrated in Figure 1 may include, among other things, one or more electronic input devices 105 that collectively provide an interface (or combination of interfaces) for receiving one or more multi-modal device interactions from a user, wherein the device interactions include at least a user-spoken utterance. Although the implementation illustrated in Figure 1 includes a distinct voice input device 105a and non-voice input device 105b, it will be apparent that in one or more implementations the voice input device 105a and the non-voice input device 105b may be components of the same or separate devices. For example, the input devices 105 may include a microphone coupled to a mobile phone (i.e., the voice input device 105a), and may further include one or more buttons, selectable displays, wheel selectors, or other components coupled to the mobile phone (i.e., the non-voice input devices 105b). In another example, the input devices 105 may include a combination of a microphone coupled to a telematics device (i.e., the voice input device 105a), and may further include buttons, a touch-screen display, a track wheel, or other non-voice input devices 105b coupled to a media player that is communicatively coupled to, yet distinct from, the telematics device. Thus, the input devices 105 may include any suitable combination of communicatively coupled electronic devices that includes at least one input device for receiving natural language utterance inputs and at least one input device for receiving multi-modal non-voice inputs.

[022] In one implementation, a voice-click module 108 communicatively coupled to the one or more input devices 105 may enable cooperative processing of multi-modal device interactions received at the voice input device 105a and one or more of the non-voice input devices 105b. For example, the voice-click module 108 may provide the system 100 with information that can be used to process a natural language utterance received via the voice input device 105a in view of one or more non-voice device interactions received via the non-voice input devices 105b. The voice-click module 108 therefore enables the user to interact with the various input devices 105 in an intuitive and free-form manner, whereby the user may provide various types

of information to the system 100 when seeking to initiate action, retrieve information, or otherwise request content or services available in the system 100.

[023] The voice input device 105a may comprise any appropriate device, or combination of devices, which have capabilities for receiving natural language utterances or other forms of spoken input. For example, in one implementation, the voice input device 105a may include a directional microphone, an array of microphones, or other devices capable of creating encoded speech. In one implementation, the voice input device 105a may be configured to maximize fidelity of the encoded speech. For example, the voice input device 105a may be configured to maximize gain in a direction of the user, cancel echoes, null point noise sources, perform variable rate sampling, filter environmental noise or background conversations, or use other techniques to maximize the fidelity of the encoded speech. As such, the voice input device 105a may create the encoded speech in a manner tolerant of noise or other factors that could otherwise interfere with the system 100 accurately interpreting natural language utterances.

[024] The non-voice input device 105b may comprise any appropriate device, or combination of devices, which have capabilities for supporting non-voice device interactions. For example, in one implementation, the non-voice input device 105b may include a combination stylus and touch-screen or tablet interface, a BlackBerry® wheel selector, an iPod® click wheel, a mouse, a keypad, buttons, or any other devices that supports distinguishable non-voice device interactions. The user may therefore use the non-voice input devices 105b to make data selections or identify a point of focus (or attention focus) to be processed in connection with a related natural language utterance provided via the voice input device 105a. For example, the user may point a stylus at a specific segment of a touch-screen display, highlight text using a mouse, click a button, interact with an application, or otherwise engage in any suitable device interaction for selecting data or otherwise identifying a point of focus (i.e., voice-activating or "voice-clicking" the selected data and/or the identified point of focus).

[025] Furthermore, in addition to being usable for making data selections, identifying points of focus, or otherwise activating data to be interpreted in association with one or more utterances, the user may further use the non-voice input devices 105b to engage in specialized device interactions that have meaning in the system 100. For example, the specialized device interactions (which may be referred to as "clicks" or "voice-clicks") may include clicks lasting a given duration, clicks continuously held for a given duration, clicks made in a predetermined

sequence, or any other interaction or sequence of interactions that the input devices 105 and/or the voice-click module 108 can identify, detect, or otherwise distinguish.

[026] In one implementation, the specialized device interactions may be associated with one or more actions, queries, commands, tasks, or other requests associated with applications or services available in the system 100. In one implementation, the specialized device interactions may further include one or more actions, queries, commands, tasks, or other requests associated with any of various devices deployed in the multi-device environment, as described in the above-referenced co-pending U.S. Patent Application 12/127,343, entitled "System and Method for an Integrated, Multi-Modal, Multi-Device Natural Language Voice Services Environment," filed May 27, 2008. For example, a distinct sequence of clicking a stylus on a particular segment or item displayed on a touch-screen display may be defined as a specialized device interaction or voice-click for initiating a telephone call on a mobile phone, calculating a route on a navigation device, purchasing a song to a media player, or another type of request.

[027] Thus, the voice-click module 108 coupled to the input device 105 may continually monitor the user's interactions with the non-voice input device 105b to detect occurrences of at least one non-voice device interaction, which may be referred to herein as a "voice-click." The detected voice-click may therefore provide further context for processing a multi-modal device interaction, which may include the at least one voice-click and one or more natural language utterances, each of which may provide context for task specification. Thus, the voice-click may generally signal the system 100 that a current utterance or other voice-based input is to be processed together with a current interaction with one or more of the devices 105. For example, in one implementation, the current device interaction may include the user selecting, highlighting, or otherwise identifying a particular point of attention, object, or other item associated with one or more of the devices 105. As such, the current device interaction may provide context for sharpening recognition, interpretation, and understanding of the accompanying utterance, and moreover, the current utterance may provide information to enhance the context provided by the accompanying device interaction.

[028] In one implementation, the voice-click module 108 may determine the various voice-click interactions to be detected based on particular characteristics of the non-voice input devices 105b (e.g., the voice-click interactions may include distinguishable interactions that the non-voice input devices 105b support). For example, a multi-touch display typically includes a

touch-screen display device configured to support various distinguishable gestures for interacting with information displayed therein (e.g., a user may zoom in, zoom out, rotate, or otherwise control graphical information displayed on a multi-touch screen using specific gestures or other interaction techniques). Thus, in one example, the non-voice input devices 105b may include a multi-touch display, in which case the voice-click module 108 may be configured to detect an occurrence of a voice-click when a user engages in one or more of the distinguishable gestures supported by the non-voice multi-touch display 105b.

[029] In one implementation, a user may customize or otherwise modify the voice-click interactions to be detected by the voice-click module 108. In particular, the specific device interactions detected by the voice-click module 108 may be removed or modified, or new device interactions may be added. As such, the voice-click device interactions detected by the voice-click module 108 may include any appropriate interaction or combination of interactions that the non-voice input devices 105b and/or the voice-click module 108 can distinguish.

[030] When the voice-click module 108 detects a user's engaging in a voice-click device interaction, the voice-click module 108 may extract context information associated with the voice-click device interaction for voice-activation. In particular, the voice-click module 108 may identify information relating to a segment, item, point of focus, attention focus, or other data selected by the user, or otherwise identify information relating to a particular device interaction or sequence of device interactions engaged in by the user. The voice-click module 108 thus extracts the identified information relating to the detected voice-click, which may be used as context information to be associated with one or more prior, contemporaneous, or subsequent natural language utterances.

[031] Thus, in response to the voice-click module 108 detecting a voice-click (e.g., selection of an icon, a section of text, specific coordinates on a map display, or other information), the voice-click module 108 may signal the system 100 to use a natural language utterance voice input (which may be received via the voice input device 105a) as further context for determining an action, query, command, task, or other request to perform in order to service the detected voice-click. As such, the various natural language processing components in the system 100 may use the combined context of the voice-click and the accompanying natural language utterance to determine the intent of the voice-click device interaction, and to

appropriately route one or more actions, queries, commands, tasks, or other requests to any of the various devices deployed in the multi-device environment.

[032] For instance, in one implementation, the multi-device environment may include a voice-enabled navigation device. Thus, an exemplary voice-click device interaction may include a user touching a stylus to a specific intersection displayed on a touch-screen display 105b associated with the voice-enabled navigation device, while also providing an utterance such as "What restaurants are around here?" into a microphone 105a. In this example, the voice-click module 108 may extract information relating to the voice-clicked intersection, which may be used as context for processing the accompanying utterance (i.e., the selected intersection may provide context for interpreting "around here," as opposed to the user's current location or some other meaning). Moreover, as indicated above, the voice input may be used as additional context for determining task specification. Thus, the utterance may be further processed for recognition and conversational interpretation using the various natural language processing components of the system 100, as described in greater detail below.

[033] In one implementation, an Automatic Speech Recognizer (ASR) 110 may generate one or more preliminary interpretations of the utterance received via the voice input device 105a. For example, the ASR 110 may recognize syllables, words, phrases, or other acoustic characteristics of the utterance using one or more dynamically adaptable recognition grammars. In one implementation, the dynamic recognition grammars may be used to recognize a stream of phonemes using phonetic dictation based on one or more acoustic models (e.g., as described in co-pending U.S. Patent Application 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed August 5, 2005, the disclosure of which is hereby incorporated by reference in its entirety).

[034] In one implementation, the ASR 110 may be configured to perform multi-pass speech recognition, where a primary speech recognition engine may generate a primary transcription of the utterance (e.g., using a large list dictation grammar), and may subsequently request one or more secondary transcriptions from one or more secondary speech recognition engines (e.g., using a virtual dictation grammar having decoy words for out-of-vocabulary words). In one implementation, the primary speech recognition engine may request the secondary transcriptions based on a confidence level for the primary transcription.

[035] The recognition grammars employed in the ASR 110 may include various vocabularies, dictionaries, syllables, words, phrases, or other information for recognizing utterances. In one implementation, information contained in the recognition grammars may be dynamically optimized to improve a likelihood of accurate recognition for a given utterance (e.g., following an incorrect interpretation of a word or phrase, the incorrect interpretation may be removed from the grammar to reduce a likelihood of repeating the incorrect interpretation). Additionally, various forms of knowledge can be used to continually optimize the information contained in the recognition grammars on a dynamic basis. For example, the system 100 may have knowledge that includes environmental knowledge (e.g., peer-to-peer affinities, capabilities of the various devices in the environment, etc.), historical knowledge (e.g., frequent requests, prior context, etc.), or short term shared knowledge relating to a current conversational dialogue or interaction, among other types of knowledge.

[036] In one implementation, the information in the recognition grammars may be further optimized according to context or application-specific domains. In particular, similar utterances may be interpreted differently depending on a context to which the utterance relates, including navigation, music, movies, weather, shopping, news, languages, temporal or geographic proximities, or other contexts or domains. For example, an utterance containing the word "traffic" may be subject to different interpretations depending on whether the context relates to navigation (i.e., road conditions), music (i.e., the 1960's rock band), or movies (i.e., the film directed by Steven Soderbergh). Accordingly, the ASR 110 may use various techniques to generate a preliminary interpretation of the natural language utterance, such as those described in the above-referenced co-pending U.S. Patent Applications and/or co-pending U.S. Patent Application 11/513,269, entitled "Dynamic Speech Sharpening," filed August 31, 2006, the disclosure of which is hereby incorporated by reference in its entirety.

[037] As such, the ASR 110 may provide one or more preliminary interpretations of the natural language utterance included in the voice-click to a conversational language processor 120. The conversational language processor 120 may include various natural language processing components collectively configured to model human-to-human conversations or interactions. For example, the conversational language processor 120 may include, among other things, an intent determination engine 130a, a constellation model 130b, one or more domain agents 130c, a context tracking engine 130d, a misrecognition engine 130e, and a voice

search engine 130f. Furthermore, the conversational language processor 120 may be coupled to one or more data repositories 160 and one or more applications 150 associated with various contexts or domains.

[038] Thus, the system 100 may use the various natural language processing components associated with the conversational language processor 120 in order to engage the user in a cooperative conversation and resolve voice-click device interactions based on the user's intent in initiating the voice-click. More particularly, the intent determination engine 130a may establish meaning for a given multi-modal device interaction based on capabilities of the system 100 as well as the capabilities of any other devices in the multi-device environment. For instance, referring to the example above where the user voice-clicked a particular intersection to determine "What restaurants are around here," the conversational language processor 120 may determine a conversational goal of the voice-click (e.g., "What" may indicate that the utterance relates to a query requesting data retrieval). In addition, the conversational language processor 120 may invoke the context tracking engine 130d to determine context for the voice-click. For example, to determine the voice-click context, the context tracking engine 130d may combine the context associated with the identified point of focus (i.e., the selected intersection) with context associated with the utterance (i.e., restaurants).

[039] As a result, the combined context of the voice-click (which includes both the device interaction and the accompanying utterance) may provide sufficient information for routing a specific query. For example, the query may include various parameters or criteria relating to restaurants and the identified intersection. The conversational language processor 120 may then select a particular device, application, or other component to which the query may be routed for processing. For example, in one implementation, the conversational language processor 120 may evaluate the constellation model 130b, which contains a model of capabilities for each device in the multi-device environment. In one implementation, the constellation model 130b may contain, among other things, knowledge of processing and storage resources available to each of the devices in the environment, as well as the nature and scope of domain agents, context, capabilities, content, services, or other information for each of the devices.

[040] As such, using the constellation model 130b and/or other information, the conversational language processor 120 may determine which device or combination of devices

has suitable capabilities that can be invoked to process a given voice-click device interaction. For instance, again referring to the example given above, the conversational language processor 120 may determine that the context of the voice-click relates to an interaction with a navigation device and therefore route the query for processing using a navigation application 150. Results of the query may then be processed (e.g., to weigh the results based on knowledge of the user, such as a preference for vegetarian restaurants) and returned to the user via an output device 180.

[041] According to various aspects of the invention, Figure 2 illustrates a block diagram of an exemplary method for synchronizing multi-modal devices in a natural language voice service environment. As described above, multi-modal device interactions (or "voice-clicks") may generally occur when a user engages in one or more interactions with one or more multi-modal devices while also providing one or more natural language utterances that relate to the interactions with the multi-modal devices. In one implementation, context information relating to the interactions with the multi-modal devices may be combined with context information relating to the natural language utterances to determine an intent of the voice-click (e.g., to initiate a particular action, query, command, task, or other request).

[042] In one implementation, various natural language processing components may be configured to continually listen or otherwise monitor the multi-modal devices to determine when voice-clicks occur. As such, the method illustrated in Figure 2 may be used to calibrate or otherwise configure the components responsible for continually listening or otherwise monitoring the multi-modal devices. For example, in one implementation, the natural language voice service environment may include a plurality of multi-modal devices that provide different capabilities or services, and the user may engage in one or more voice-clicks to request services relating to any of the various devices or capabilities in any given device interaction.

[043] To enable continuous listening for multi-modal device interactions or voice-clicks, each of the plurality of devices in the environment may be configured to receive information relating to a voice-click. Thus, in one implementation, an operation 210 may include establishing device listeners for each of the plurality of devices in the environment. Additionally, operation 210 may be performed in response to one or more new devices being added to the environment. The device listeners established in operation 210 may include any suitable combination of instructions, firmware, or other routines configured to be executed on one or more processing

devices or other hardware components. For each device in the environment, the associated device listener may communicate with the device to determine capabilities, features, supported domains, or other information associated with the device. In one implementation, the device listeners may be configured to communicate with the device using the Universal Plug and Play protocol designed for ancillary computer devices. It will be apparent, however, that any suitable mechanism for communicating with the multi-modal devices may be used.

[044] When device listeners have been established for each of the devices in the environment (or when device listeners have been established for new devices added to the environment), the various device listeners may be synchronized in an operation 220. In particular, each of the various devices may have different internal clocks or other timing mechanisms, wherein operation 220 may include synchronizing the various device listeners according to the respective internal clocks or timing mechanisms of the devices. In one implementation, synchronizing the device listeners may include each of the respective device listeners publishing information relating to the internal clock or timing of the associated device.

[045] Thus, when one or more multi-modal interactions or voice-clicks subsequently occur for one or more of the devices, the associated device listener may detect information associated with the voice-click in an operation 230. For example, in one implementation, the various device listeners established in operation 210 may be associated with the voice-click module described above and illustrated in Figure 1. Operation 230 may therefore include one or more of the device listeners or the voice-click module detecting an occurrence of a user interacting with one or more of the devices (e.g., selecting data associated with the device, identifying a point of focus or attention focus associated with the device, or otherwise engaging in one or more interactions or sequences of interactions with the device). Furthermore, an operation 240 may then comprise capturing an utterance from the user that relates to the device interaction detected in operation 230.

[046] For example, a user viewing a web page presented on a display device may see a product name on the web page and desire more information about purchasing the product. The user may select text from the web page that contains the product name (e.g., using a mouse or keyboard to highlight the text), and then initiate a voice-click to ask "Is this available on Amazon.com?" In this example, operation 230 may include a device listener associated with the display device detecting the selection of the text associated with the product name, while

operation 240 may include capturing the utterance inquiring about the availability of the product on Amazon.com.

[047] As described above, each device that receives an input from the user may have an internal clock or timing mechanism. In an operation 250, each device may therefore determine when the input was received from a local perspective, and notify the voice-click module that the input was received. In particular, a given voice-click may include at least a natural language utterance received via a voice input device in addition to one or more further interactions with one or more other devices. The utterance may be received prior to, contemporaneously with, or subsequent to the device interactions, whereby operation 250 includes determining the timing of the device interaction for correlation with the associated utterance. In particular, using the device listener signals synchronized as described in reference to operation 220, an operation 260 may include aligning the signals for the device interactions and the utterance. In matching the device interaction and utterance signals, a voice-click input may be generated that includes aligned voice and non-voice components. The voice-click input may then be subject to further natural language processing, as described in greater detail below.

[048] According to various aspects of the invention, Figure 3 illustrates a flow diagram of an exemplary method for processing multi-modal device interactions in a natural language voice service environment. As described above, multi-modal device interactions (or "voice-clicks") may generally occur when a user interacts with one or more multi-modal devices while also providing one or more natural language utterances related to the device interactions. As such, in one implementation, the method illustrated in Figure 3 may be performed when one or more natural language processing components continually listen or otherwise monitor the one or more multi-modal devices to determine when one or more voice-clicks occur.

[049] In one implementation, one or more device interactions may be defined as initiating a voice-click. For example, any given electronic device may generally support various distinguishable interactions, which may result in a given action, command, query, or other request being executed. Thus, in one implementation, any appropriate combination of device interactions that a given electronic device can uniquely recognize or otherwise use to generate a uniquely recognizable signal may be defined as a voice-click, wherein the voice-click may provide a signal indicating that a natural language utterance is to be processed together with context associated with the relevant device interactions. For example, devices that have a four-

way or a five-way navigation button may support specific distinguishable interactions, wherein pressing the navigation button in a particular way may cause a specific task or other action to be performed, such as controlling a map display or calculating a route. In another example, a BlackBerry® device having a wheel selector may support interactions such as wheeling a cursor over a particular point of focus or attention focus, pressing the wheel to select specific data or a given application, or various other interactions. Various other device interactions may be used to indicate when a natural language utterance is to be processed together with context associated with the device interactions, without limitation, wherein the specific device interactions may vary in any given implementation. For example, the relevant device interactions may further include one or more of gesturing with a pointing or drawing instrument on a touch-sensitive screen (e.g., drawing an ear-shaped squiggle), a unique interaction method such as a long touch or double-tap, and/or if the system is operating in the continuous listening mode described above, a predefined contextual command word may indicate that a current device context is to be processed with a portion of the voice-based input that follows the contextual command word (e.g., the command word be "OK," "Please," "Computer," or another suitable word, wherein a user may select a specific point on a map and say "Please zoom in," or say "OK read it" when an e-mail is displayed, etc.).

[050] As such, an operation 310 may include processing multi-modal device interactions in a natural language voice service environment to detect an occurrence of one or more device interaction that signal initiation of a voice-click. In particular, the device interactions detected in operation 310 may include any suitable interactions that cause an electronic device to generate a unique, recognizable, or otherwise distinguishable signal that relates to a user selecting data, identifying a point of focus or attention focus, invoking an application or task, or interacting with the device in another way, depending on the specific capabilities of the device.

[051] In addition to the specific signal that the device generates in response to the user interaction, the interaction detected in operation 310 may signal initiation of a voice-click, whereby a prior, contemporaneous, or subsequent natural language voice input will provide further context for interpreting the device interaction detected in operation 310. For example, a natural language processing system may generally be configured to accept a voice input when particular device interactions occur (e.g., pressing a button to turn on a microphone). Thus, in the method illustrated in Figure 3, the device interactions that signal an incoming voice input

may further include any suitable interaction or combination of interactions with an electronic device, including interactions associated with a user selecting data, identifying a point of focus or attention focus, invoking an application or task, or interacting with the device in another way, depending on the specific capabilities of the device.

[052] As such, when a voice-click device interaction has been detected in operation 310, a voice-click signal may be generated in an operation 320 to indicate that a natural language voice input should be associated with the interaction detected in operation 320. Subsequently, an operation 330 may include capturing a user utterance to be associated with the interaction detected in operation 310. In one implementation, the interaction detected in operation 310 may signal that a subsequent voice input will be provided, but it will be apparent that in one or more implementations the utterance captured in operation 330 may be provided prior to or contemporaneously with the interaction detected in operation 310 (e.g., a user may provide an utterance such as "Look up this artist on iTunes®" and subsequently voice-click the artist's name on a media player, or the user may provide the utterance while voice-clicking the artist's name, or the user may voice-click the artist's name and then provide the utterance).

[053] When information relating to the voice-click device interactions and associated natural language utterance have been received, an operation 340 may include extracting and combining context information for the device interactions and the associated utterance. In particular, the context information extracted from the voice-click device interactions may include information relating to a segment, item, point of focus, attention focus, or data selected by the user, or to a particular device interaction or sequence of device interactions engaged in by the user. The extracted context for the device interaction may then be combined with context extracted for the natural language utterance captured in operation 330, wherein the combined context information may be used to determine an intent of the voice-click in an operation 350.

[054] For example, in an exemplary voice-click device interaction, a user may be selectively copying a music collection from a media player to a backup storage device. While browsing the music on the media player, the user may encounter a particular song and voice-click the song while saying "Copy this whole album" (e.g., by pressing a particular button on the media player for an extended period of time while highlighting the song). In this example, operation 310 may include detecting the interaction of the extended button press, which causes the voice-click

signal to be generated in operation 320. The utterance of "Copy this whole album" may then be captured in operation 330, and context information associated with the voice-click device interaction and the utterance may be combined in operation 340. In particular, the context of the device interaction may include information relating to the selected song, among other things (e.g., the context may further include information contained in metadata associated with the song, such as an ID3 tag for music files). Furthermore, the context of the utterance may include information identifying a copy operation and an album that contains the selected song.

[055] As such, context information relating to voice-click interactions with the multi-modal devices may be combined with context information relating to the natural language utterances, whereby an operation 350 may determine the intent of the voice-click interaction. For instance, referring to the example above, the intent determined in operation 350 may include an intent to copy an album that contains the highlighted song from the media player to the backup storage device. Thus, in response to determining the intent of the voice-click in operation 350, one or more requests may be appropriately routed in an operation 360. In the example being discussed herein, operation 360 may include routing one or more requests to the media player to identify all of the data associated with the album that contains the highlighted song, and one or more requests to any appropriate combination of devices that can manage copying of the identified data from the media player to the backup storage device (e.g., a personal computer that interfaces with both the media player and the storage device).

[056] According to various aspects of the invention, Figure 4 illustrates a flow diagram of an exemplary method for processing multi-modal device interactions to generate a transaction lead or "click-through" in a natural language voice service environment. In particular, the method illustrated in Figure 4 may be used to generate a transaction lead or click-through in combination with one or more actions performed in response to one or more voice-click device interactions being detected.

[057] For example, an operation 410 may include detecting one or more voice-click device interactions received from a user, wherein the voice-click device interactions may comprise any suitable combination of one or more device interactions coupled with one or more related natural language utterances. The user's intent in engaging in the voice-click device interactions may then be determined in an operation 420, and a subsequent operation 430 may include routing one or more requests to one or more processing devices based on the determined

intent in order to resolve the voice-click interaction. In one implementation, operations 410, 420, and 430 may be performed in a manner similar to that described above in reference to Figure 2 and Figure 3, whereby signals for the device interactions may be aligned with signals for one or more natural language utterances and context information may be extracted from the signals to determine the intent of the voice-click device interaction.

[058] In addition to routing one or more requests based on the user's intent, the method illustrated in Figure 4 may further comprise generating one or more transaction leads that may result in one or more click-throughs. For example, a click-through may generally refer to an instance of a user clicking or otherwise selecting an electronic advertisement to access one or more services associated with the advertiser. In many electronic systems, click-throughs or click-through rates can provide mechanisms for measuring a user's interactions with an electronic advertisement, which may provide various measurements that advertisers can use to determine an amount to pay an entity that delivers the advertisement to users.

[059] As such, the method illustrated in Figure 4 may generate transaction leads, including advertisements or recommendations, whereby a user's voice-based input combined with a specific device interaction may provide further focus for generating transaction leads. In this manner, the advertisements or recommendations provided to the user may be more relevant to specific information with which the user may be interacting. Furthermore, using natural language cognitive models and shared knowledge relating to a user's preferences may provide further context for targeted transaction leads tailored to the specific user, which may thus be more likely to result in a click-through that can generate payment for a voice services provider.

[060] Thus, in addition to routing one or more requests based on the user's intent in engaging in the voice-click device interaction, an operation 440 may include generating one or more transaction leads based on the determined intent. In particular, based on the combined context of the device interaction and the associated natural language utterance, the transaction leads may be processed in a manner "closer" to the user, in that local voice and non-voice context can be used as state data in any appropriate system that performs targeted advertising. For instance, referring to the example given above where a user selects an intersection displayed on a navigation device while saying "Find restaurants are around here," the transaction leads generated in operation 440 may include one or more advertisements or recommendations for restaurants near the intersection, which may be targeted to the user

based on knowledge of the user's short-term and long-term preferences (e.g., a preferred type of restaurant, a preferred price range, etc.).

[061] The transaction leads may then be presented to the user (e.g., as selectable points on a map display). The user's subsequent multi-modal device interactions may then be monitored in an operation 450 to determine whether or when one or more further multi-modal device interactions occur. If no further interactions occur, a determination may be made that the user did not act on the transaction lead and the process may then end. If an additional multi-modal interaction does occur, however, the multi-modal input may be processed in an operation 480 to determine an intent of the input and route one or more requests accordingly. In addition, an operation 460 may include determining whether the multi-modal input relates to the transaction lead generated in operation 440. For example, the user may select one of the advertised or recommended restaurants by providing an utterance, a non-voice device interaction, or a voice-click device interaction requesting further action or information relating to the transaction lead. In such a case, an operation 470 may further include processing a transaction click-through in relation to the transaction lead generated in operation 440, wherein the transaction click-through may be used to determine payment or otherwise generate revenue for a particular entity (e.g., a provider of the voice services or another entity associated with the transaction lead or transaction click-through).

[062] Implementations of the invention may be made in hardware, firmware, software, or various combinations thereof. The invention may also be implemented as instructions stored on a machine-readable medium, which may be read and executed by one or more processors. A machine-readable medium may include various mechanisms for storing or transmitting information in a form readable by a machine (e.g., a computing device). For example, a machine-readable storage medium may include read only memory, random access memory, magnetic disk storage media, optical storage media, flash memory devices, or other storage media, and a machine-readable transmission media may include forms of propagated signals, such as carrier waves, infrared signals, digital signals, or other transmission media. Further, firmware, software, routines, or instructions may be described in the above disclosure in terms of specific exemplary aspects and implementations of the invention, and performing certain actions. However, it will be apparent that such descriptions are merely for convenience, and

that such actions in fact result from computing devices, processors, controllers, or other devices executing the firmware, software, routines, or instructions.

[063] Although the descriptions provided herein have generally focused on techniques for processing multi-modal device interactions in a natural language voice services environment, it will be apparent that various further natural language processing capabilities may be used in connection with, in addition to, or in lieu of those described in connection with the specific aspects and implementations discussed herein. For example, in addition to the techniques described in the co-pending U.S. Patent Applications referenced above, the systems and methods described herein may further utilize natural language processing capabilities as described in co-pending U.S. Patent Application 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed August 5, 2005, U.S. Patent Application 11/200,164, entitled "System and Method of Supporting Adaptive Misrecognition in Conversational Speech," filed August 10, 2005, U.S. Patent Application 11/212,693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," filed August 29, 2005, U.S. Patent Application 11/580,926, entitled "System and Method for a Cooperative Conversational Voice User Interface," filed October 16, 2006, U.S. Patent Application 11/671,526, entitled "System and Method for Selecting and Presenting Advertisements Based on Natural Language Processing of Voice-Based Input," filed February 6, 2007, and U.S. Patent Application 11/954,064, entitled "System and Method for Providing a Natural Language Voice User Interface in an Integrated Voice Navigation Services Environment," filed December 11, 2007, the disclosures of which are hereby incorporated by reference in their entirety.

[064] Accordingly, aspects and implementations of the invention may be described herein as including a particular feature, structure, or characteristic, but it will be apparent that every aspect or implementation may or may not necessarily include the particular feature, structure, or characteristic. In addition, when a particular feature, structure, or characteristic has been described in connection with a given aspect or implementation, it will be understood that such feature, structure, or characteristic may be included in connection with other aspects or implementations, whether or not explicitly described. Thus, various changes and modifications may be made to the preceding description without departing from the scope or spirit of the

invention, and the specification and drawings should therefore be regarded as exemplary only, with the scope of the invention determined solely by the appended claims.

CLAIMS

What is claimed is:

1. A method for processing one or more multi-modal device interactions in a natural language voice services environment that includes one or more electronic devices, comprising:

detecting at least one multi-modal device interaction, wherein the multi-modal device interaction includes a non-voice interaction with at least one of the electronic devices or an application associated with at least one of the electronic devices, and wherein the multi-modal device interaction further includes at least one natural language utterance relating to the non-voice interaction;

extracting context information relating to the multi-modal device interaction, wherein the extracted context information includes context relating to the non-voice interaction, and wherein the extracted context information further include context relating to the natural language utterance;

combining the context relating to the non-voice interaction and the context relating to the natural language utterance;

determining an intent of the multi-modal device interaction based on the combined context relating to the non-voice interaction and the natural language utterance; and

routing at least one request to one or more of the electronic devices based on the determined intent of the multi-modal device interaction.

2. The method of claim 1, wherein at least one of the electronic devices includes an input device configured to receive the natural language utterance.

3. The method of claim 2, further comprising signaling the input device to capture the natural language utterance in response to the non-voice interaction being detected.

4. The method of claim 3, further comprising:

establishing one or more device listeners in the natural language voice services environment, the device listeners configured to detect the non-voice interaction; and

aligning information relating to the non-voice interaction detected by the device listeners and the natural language utterance captured by the input device.

5. The method of claim 1, further comprising:

generating at least one transaction lead based on the determined intent of the multi-modal device interaction;

receiving at least one further multi-modal device interaction relating to the generated transaction lead; and

processing a transaction click-through in response to receiving the multi-modal device interaction relating to the generated transaction lead.

6. The method of claim 5, wherein the generated transaction lead includes at least one of an advertisement or a recommendation relating to the determined intent of the multi-modal device interaction.

7. The method of claim 1, wherein the non-voice interaction comprises a selection of a segment, item, data, or application associated with one or more of the electronic devices.

8. The method of claim 1, wherein the non-voice interaction comprises an identification of a point of focus or an attention focus associated with one or more of the electronic devices.

9. The method of claim 1, wherein the non-voice interaction comprises one or more unique and distinguishable interactions associated with one or more of the electronic devices.

10. A system for processing one or more multi-modal device interactions in a natural language voice services environment that includes one or more electronic devices, wherein the system comprises one or more processing devices configured to:

detect at least one multi-modal device interaction, wherein the multi-modal device interaction includes a non-voice interaction with at least one of the electronic devices or an application associated with at least one of the electronic devices, and wherein the multi-modal device interaction further includes at least one natural language utterance relating to the non-voice interaction;

extract context information relating to the multi-modal device interaction, wherein the extracted context information includes context relating to the non-voice interaction, and wherein the extracted context information further include context relating to the natural language utterance;

combine the context relating to the non-voice interaction and the context relating to the natural language utterance;

determine an intent of the multi-modal device interaction based on the combined context relating to the non-voice interaction and the natural language utterance; and

route at least one request to one or more of the electronic devices based on the determined intent of the multi-modal device interaction.

11. The system of claim 10, wherein at least one of the electronic devices includes an input device configured to receive the natural language utterance.

12. The system of claim 11, the processing devices further configured to signal the input device to capture the natural language utterance in response to the non-voice interaction being detected.

13. The system of claim 12, the processing devices further configured to:

establish one or more device listeners in the natural language voice services environment, the device listeners configured to detect the non-voice interaction; and

align information relating to the non-voice interaction detected by the device listeners and the natural language utterance captured by the input device.

14. The system of claim 10, the processing devices further configured to:

at least one transaction lead based on the determined intent of the multi-modal device interaction;

receive at least one further multi-modal device interaction relating to the generated transaction lead; and

process a transaction click-through in response to receiving the multi-modal device interaction relating to the generated transaction lead.

15. The system of claim 14, wherein the generated transaction lead includes at least one of an advertisement or a recommendation relating to the determined intent of the multi-modal device interaction.

16. The system of claim 10, wherein the non-voice interaction comprises a selection of a segment, item, data, or application associated with one or more of the electronic devices.

17. The system of claim 10, wherein the non-voice interaction comprises an identification of a point of focus or an attention focus associated with one or more of the electronic devices

18. The system of claim 10, wherein the non-voice interaction comprises one or more unique and distinguishable interactions associated with one or more of the electronic devices.

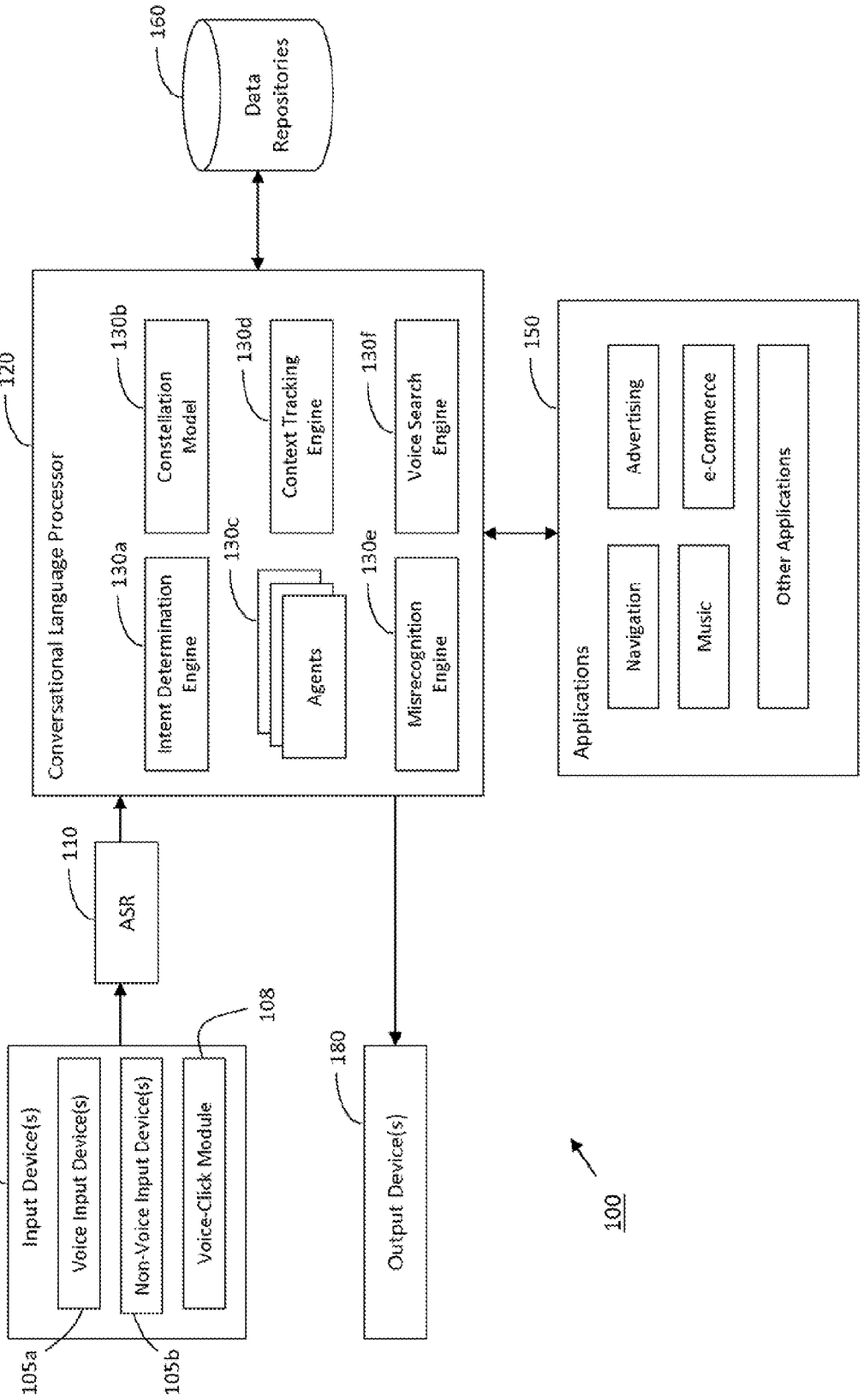

160
Data Repositories
120
Conversational Language Processor
130b
Constellation Model
130d
Context Tracking Engine
130f
Voice Search Engine
130a
Intent Determination Engine
130c
Agents
130e
Misrecognition Engine
150
Applications
Advertising
Navigation
Music
e-Commerce
Other Applications
110
ASR
105
Input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

Figure 1

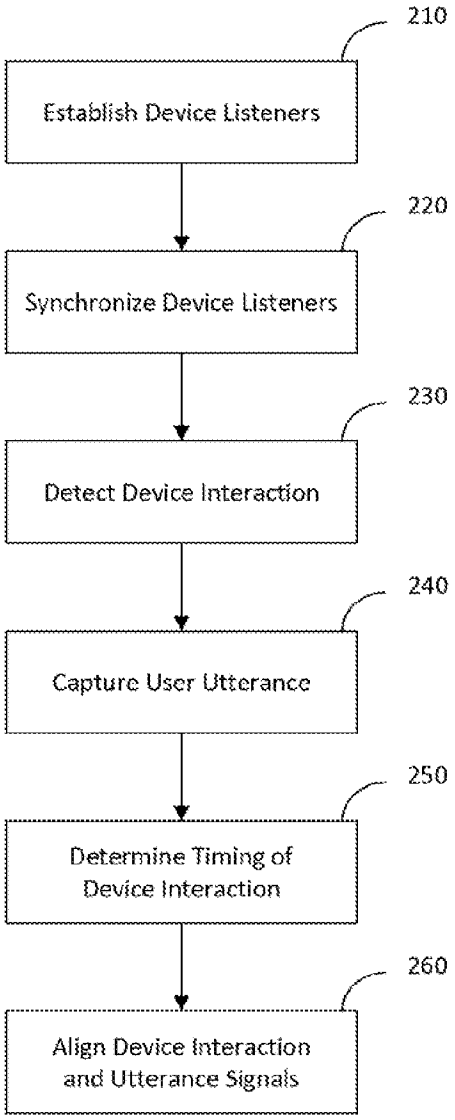

210
Establish Device Listeners
220
Synchronize Device Listeners
230
Detect Device Interaction
240
Capture User Utterance
250
Determine Timing of Device Interaction
260
Align Device Interaction and Utterance Signals

Figure 2

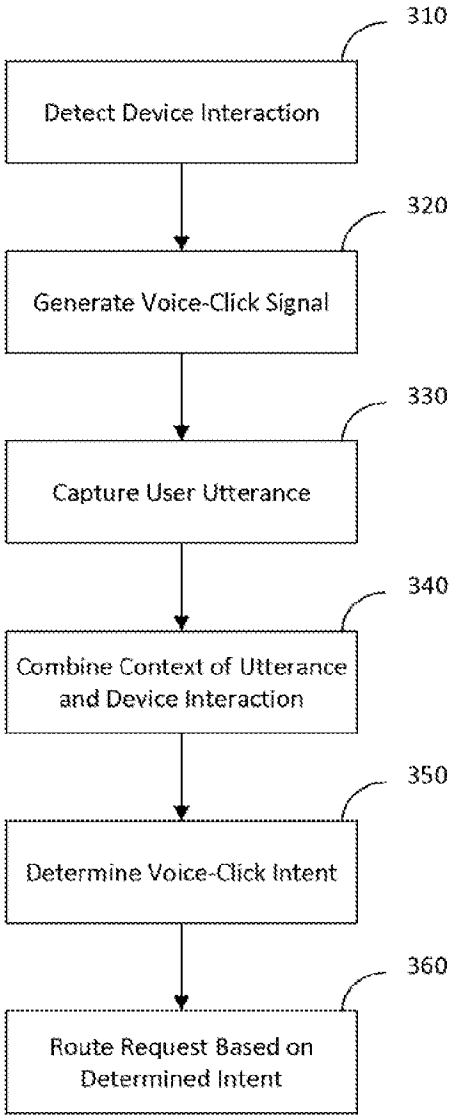

310
Detect Device Interaction
320
Generate Voice-Click Signal
330
Capture User Utterance
340
Combine Context of Utterance and Device Interaction
350
Determine Voice-Click Intent
360
Route Request Based on Determined Intent

Figure 3

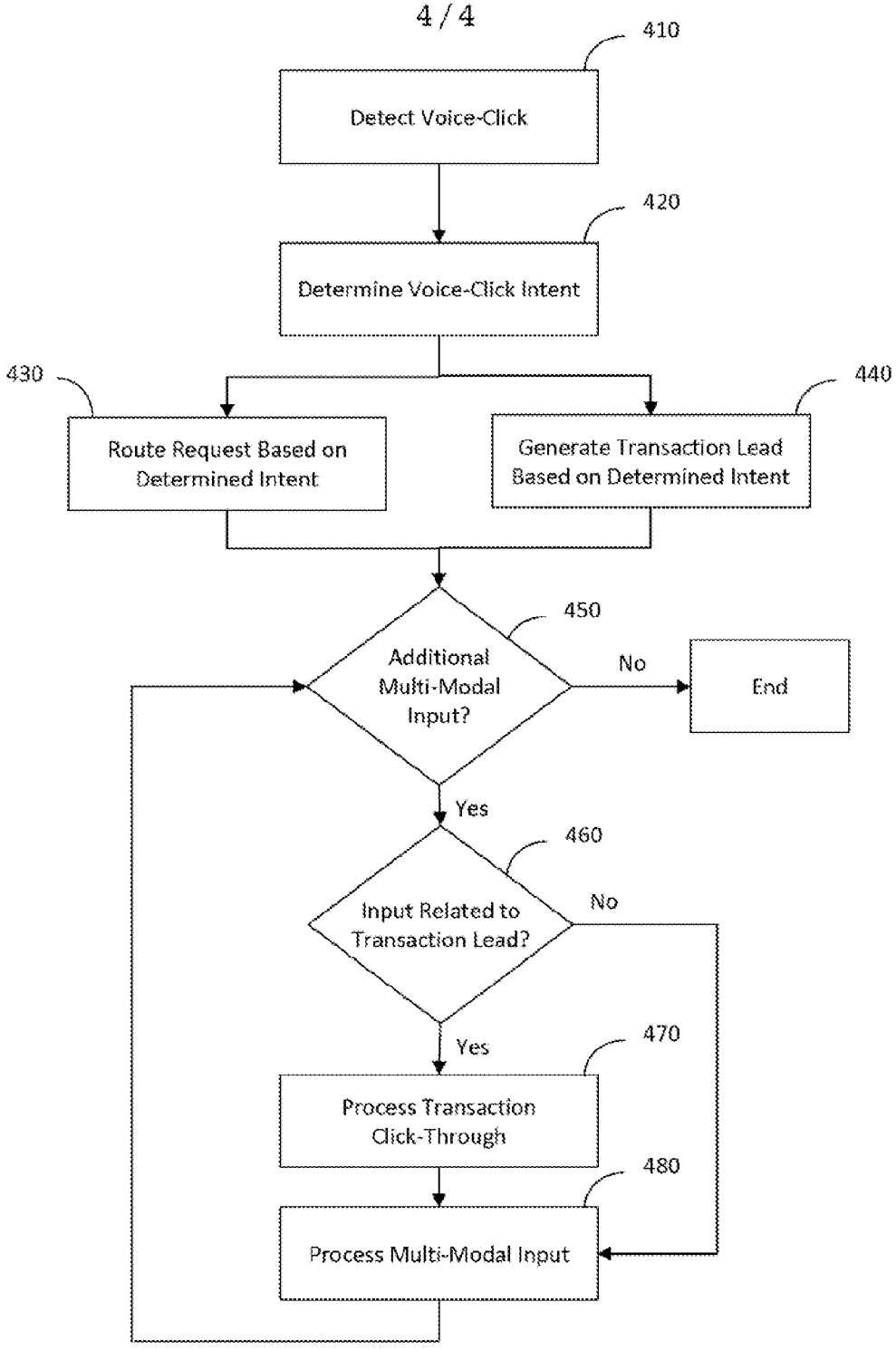

410
Detect Voice-Click
420
Determine Voice-Click Intent
430
Route Request Based on Determined Intent
440
Generate Transaction Lead Based on Determined Intent
450
Additional Multi-Modal Input?
No
End
Yes
460
Input Related to Transaction Lead?
No
Yes
470
Process Transaction Click-Through
480
Process Multi-Modal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

Figure 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 10/24895

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G10L 15/00 (2010.01)
USPC - 704/257
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC: 704/257

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 704/270.1; 704/200; 704/9 (search term limited)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Electronic Databases Searched: pubWEST(PGPB,USPT,USOC,EPAB,JPAB); googleScholar
Search Terms Used: natural language, multimodal, speech/voice identification/recognition, voice/keyboard/mouse click, context/intent, marketing/advertising

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US 2007/0050191 A1 (WEIDER et al.) 01 March 2007 (01.03.2007) Entire document, especially: para [0001], [0016], [0018], [0020], [0031], [0043], [0116], [0133]; Fig. 7, 8	1-2, 7-11 and 16-18 --- 3-6 and 12-15
Y	US 2007/0043574 A1 (COFFMAN et al.) 22 February 2007 (22.02.2007) Entire document, especially: Abstract, para [0154], [0227]	3, 4, 12 and 13
Y	US 2008/0189110 A1 (FREEMAN et al.) 07 August 2008 (07.08.2008) Entire document, especially: para [0006], [0013], [0024], [0028] and Fig. 3	5, 6, 14 and 15
A	US 2008/0235027 A1 (CROSS) 25 September 2008 (25.09.2008)	1-18
A	US 2008/0091406 A1 (BALDWIN et al.) 17 April 2008 (14.04.2008)	1-18

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

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

Date of the actual completion of the international search	Date of mailing of the international search report
28 March 2010 (28.03.2010)	13 APR 2010
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (July 2009)

Electronic Patent Application Fee Transmittal

Application Number:	12617506
Filing Date:	12-Nov-2009
Title of Invention:	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE
First Named Inventor/Applicant Name:	Philippe Di Cristo
Filer:	Syed Jafar Ali/Miriam E. Nelsen
Attorney Docket Number:	090611-0382843

Filed as Small Entity

Utility under 35 USC 111(a) Filing Fees

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Miscellaneous:				
Request for continued examination	2801	1	405	405
		Total in USD ($)		**405**

Electronic Acknowledgement Receipt

EFS ID:	8972878
Application Number:	12617506
International Application Number:	
Confirmation Number:	4703
Title of Invention:	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE
First Named Inventor/Applicant Name:	Philippe Di Cristo
Customer Number:	00909
Filer:	Syed Jafar Ali/Miriam E. Nelsen
Filer Authorized By:	Syed Jafar Ali
Attorney Docket Number:	090611-0382843
Receipt Date:	06-DEC-2010
Filing Date:	12-NOV-2009
Time Stamp:	15:08:45
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	Deposit Account
Payment was successfully received in RAM	$405
RAM confirmation Number	1423
Deposit Account	033975
Authorized User	

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:

Charge any Additional Fees required under 37 C.F.R. Section 1.21 (Miscellaneous fees and charges)

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		090611-0382843_RCE_IDS.pdf	848165 fa816630bb606a703ccd751d298005f034b4e3df	yes	15

Multipart Description/PDF files in .zip description

Document Description	Start	End
Request for Continued Examination (RCE)	1	1
Transmittal Letter	2	3
Information Disclosure Statement (IDS) Filed (SB/08)	4	15

Warnings:					
Information:					
2	Foreign Reference	WO1999046763.pdf	1247910 62eacd3541d472fdc4873f7f386c31e69fd9ea4e	no	28
Warnings:					
Information:					
3	Foreign Reference	WO2000021232.pdf	3963803 203ae5a44439b38f56ae96cdbaa5a64e9e3c570c	no	77
Warnings:					
Information:					
4	Foreign Reference	WO2000046792.pdf	1259047 705348c0646506c1bc55d53c27f1ae60013291b1	no	34
Warnings:					
Information:					
5	Foreign Reference	WO2001078065.pdf	2063508 477bf8d473bf86193d1f7e268aa8aa10625dfa37	no	63
Warnings:					
Information:					
6	Foreign Reference	WO2004072954.pdf	1105346 c4acb3accfd8c87c21d4bf9b2cc1c8dfe629a50c	no	25
Warnings:					
Information:					
7	NPL Documents	REUTERS.pdf	79394 7f52f8fb293a5d88b570212d1f5399bb61aacdb3	no	1

Warnings:					
Information:					
8	NPL Documents	LIN.pdf	546736 3dce1acf1fbd1bc2a4a0da820e559e124949d760	no	4
Warnings:					
Information:					
9	Foreign Reference	WO2007019318.pdf	3196832 ff1325b971941c7a88c2c7871f5785b0154e9307	no	55
Warnings:					
Information:					
10	Foreign Reference	WO2007021587.pdf	3525793 196b60abf639ab219e2472fadc79b2bf65c0ee3a	no	61
Warnings:					
Information:					
11	Foreign Reference	WO2007027546.pdf	3873530 5cc6e95751b76c1f016b92415f2860dd2cc39298	no	74
Warnings:					
Information:					
12	Foreign Reference	WO2007027989.pdf	1432119 6c33baceb03630bcc7fc958ccf493ab880e7e9e3	no	28
Warnings:					
Information:					
13	Foreign Reference	WO2008098039.pdf	1103202 fcab991006a05016a57e05548f6b00f78732c946	no	24
Warnings:					
Information:					
14	NPL Documents	KUHN.pdf	438455 5e927839fd4a238b3722e336346ee055e7fcae45	no	5
Warnings:					
Information:					
15	NPL Documents	BELVIN.pdf	468676 93c339f6044e2eb4cff4b1c9087be3a99e3767d4	no	5
Warnings:					
Information:					
16	Foreign Reference	WO2008118195.pdf	2058896 635503176929e0d666e3e8683aa40f3a9643658e	no	39

Warnings:					
Information:					
17	Foreign Reference	WO2009075912.pdf	3688456 02c65abf9ea5c876fb8e4a654567b258513f4e25	no	69
Warnings:					
Information:					
18	Foreign Reference	WO2009145796.pdf	1934285 e71f0a22dfc12ce94396146b078264c9e5df68d9	no	39
Warnings:					
Information:					
19	Foreign Reference	WO2010096752.pdf	1502381 f4c719245f739ab6223fbd8f94c7b17cd9acf7b7	no	33
Warnings:					
Information:					
20	NPL Documents	LIND.pdf	729916 67eb78f7014eeb103df2382cdcecf0f9807d755a	no	6
Warnings:					
Information:					
21	NPL Documents	ZHAO.pdf	1517578 2808fd320bc985b8850a5a6c3fa515213ffa320c	no	5
Warnings:					
Information:					
22	NPL Documents	CHAI.pdf	1024450 9bf84f1bd4b08cba3cd8e8f3c6d5bb73ea5f9e5b	no	10
Warnings:					
Information:					
23	NPL Documents	CHEYER.pdf	136148 ef2636c9907ae9ab23536f71fecb9453fed36f4a	no	11
Warnings:					
Information:					
24	NPL Documents	ELIO.pdf	934606 5312283c44d9e175a26f467c8014f5de541895b2	no	10
Warnings:					
Information:					
25	NPL Documents	TURUNEN.pdf	179154 01427e429808c0e3724a46047f4d0cfc9abf99d1	no	2

Warnings:					
Information:					
26	NPL Documents	MAO.pdf	87646 07b0d85ca1604fefcd7cd99772e25ff506ff1ef0	no	4
Warnings:					
Information:					
27	NPL Documents	VANHOUCKE.pdf	259233 84ea32d24072d520d3c972cb6af5e8dc4b7835f8	no	4
Warnings:					
Information:					
28	NPL Documents	WENG.pdf	346056 0b8aed950361e7f52843a57fd30ba770384b6363	no	4
Warnings:					
Information:					
29	NPL Documents	ELMELIANI.pdf	296007 f2fb2b6713c7266a4c98b7c64837ebb75106cc03	no	4
Warnings:					
Information:					
30	NPL Documents	ARRINGTON.pdf	627466 b532ea1dde8275e00c1d288d7e7f9a17df60711b	no	4
Warnings:					
Information:					
31	NPL Documents	BAZZI.pdf	361077 67a2e53341c20cd2f0fecfdfc9ca971d47945b8c	no	4
Warnings:					
Information:					
32	NPL Documents	OSHAUGHNESSY.pdf	4563513 93d11484104e08a3f70ac1884e29bccb188a88f0	no	34
Warnings:					
Information:					
33	Fee Worksheet (PTO-875)	fee-info.pdf	30431 2ac648938d1dda204d02d54423d92565a7aa0630	no	2
Warnings:					
Information:					
			Total Files Size (in bytes):	45429815	

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

New Applications Under 35 U.S.C. 111

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

National Stage of an International Application under 35 U.S.C. 371

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

New International Application Filed with the USPTO as a Receiving Office

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

UNITED STATES PATENT AND TRADEMARK OFFICE

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

NOTICE OF ALLOWANCE AND FEE(S) DUE

909 7590 12/20/2010

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

EXAMINER	
ALBERTALLI, BRIAN LOUIS	
ART UNIT	PAPER NUMBER
2626	

DATE MAILED: 12/20/2010

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
12/617,506	11/12/2009	Philippe Di Cristo	090611-0382843	4703

TITLE OF INVENTION: SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE

APPLN. TYPE	SMALL ENTITY	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	YES	$755	$300	$0	$1055	03/21/2011

THE APPLICATION IDENTIFIED ABOVE HAS BEEN EXAMINED AND IS ALLOWED FOR ISSUANCE AS A PATENT. PROSECUTION ON THE MERITS IS CLOSED. THIS NOTICE OF ALLOWANCE IS NOT A GRANT OF PATENT RIGHTS. THIS APPLICATION IS SUBJECT TO WITHDRAWAL FROM ISSUE AT THE INITIATIVE OF THE OFFICE OR UPON PETITION BY THE APPLICANT. SEE 37 CFR 1.313 AND MPEP 1308.

THE ISSUE FEE AND PUBLICATION FEE (IF REQUIRED) MUST BE PAID WITHIN THREE MONTHS FROM THE MAILING DATE OF THIS NOTICE OR THIS APPLICATION SHALL BE REGARDED AS ABANDONED. THIS STATUTORY PERIOD CANNOT BE EXTENDED. SEE 35 U.S.C. 151. THE ISSUE FEE DUE INDICATED ABOVE DOES NOT REFLECT A CREDIT FOR ANY PREVIOUSLY PAID ISSUE FEE IN THIS APPLICATION. IF AN ISSUE FEE HAS PREVIOUSLY BEEN PAID IN THIS APPLICATION (AS SHOWN ABOVE), THE RETURN OF PART B OF THIS FORM WILL BE CONSIDERED A REQUEST TO REAPPLY THE PREVIOUSLY PAID ISSUE FEE TOWARD THE ISSUE FEE NOW DUE.

HOW TO REPLY TO THIS NOTICE:

I. Review the SMALL ENTITY status shown above.

If the SMALL ENTITY is shown as YES, verify your current SMALL ENTITY status:

A. If the status is the same, pay the TOTAL FEE(S) DUE shown above.

B. If the status above is to be removed, check box 5b on Part B - Fee(s) Transmittal and pay the PUBLICATION FEE (if required) and twice the amount of the ISSUE FEE shown above, or

If the SMALL ENTITY is shown as NO:

A. Pay TOTAL FEE(S) DUE shown above, or

B. If applicant claimed SMALL ENTITY status before, or is now claiming SMALL ENTITY status, check box 5a on Part B - Fee(s) Transmittal and pay the PUBLICATION FEE (if required) and 1/2 the ISSUE FEE shown above.

II. PART B - FEE(S) TRANSMITTAL, or its equivalent, must be completed and returned to the United States Patent and Trademark Office (USPTO) with your ISSUE FEE and PUBLICATION FEE (if required). If you are charging the fee(s) to your deposit account, section "4b" of Part B - Fee(s) Transmittal should be completed and an extra copy of the form should be submitted. If an equivalent of Part B is filed, a request to reapply a previously paid issue fee must be clearly made, and delays in processing may occur due to the difficulty in recognizing the paper as an equivalent of Part B.

III. All communications regarding this application must give the application number. Please direct all communications prior to issuance to Mail Stop ISSUE FEE unless advised to the contrary.

IMPORTANT REMINDER: Utility patents issuing on applications filed on or after Dec. 12, 1980 may require payment of maintenance fees. It is patentee's responsibility to ensure timely payment of maintenance fees when due.

PART B - FEE(S) TRANSMITTAL

Complete and send this form, together with applicable fee(s), to: Mail **Mail Stop ISSUE FEE**
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450
or Fax (571)-273-2885

INSTRUCTIONS: This form should be used for transmitting the ISSUE FEE and PUBLICATION FEE (if required). Blocks 1 through 5 should be completed where appropriate. All further correspondence including the Patent, advance orders and notification of maintenance fees will be mailed to the current correspondence address as indicated unless corrected below or directed otherwise in Block 1, by (a) specifying a new correspondence address; and/or (b) indicating a separate "FEE ADDRESS" for maintenance fee notifications.

CURRENT CORRESPONDENCE ADDRESS (Note: Use Block 1 for any change of address)

909 7590 12/20/2010

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

Note: A certificate of mailing can only be used for domestic mailings of the Fee(s) Transmittal. This certificate cannot be used for any other accompanying papers. Each additional paper, such as an assignment or formal drawing, must have its own certificate of mailing or transmission.

Certificate of Mailing or Transmission

I hereby certify that this Fee(s) Transmittal is being deposited with the United States Postal Service with sufficient postage for first class mail in an envelope addressed to the Mail Stop ISSUE FEE address above, or being facsimile transmitted to the USPTO (571) 273-2885, on the date indicated below.

(Depositor's name)
(Signature)
(Date)

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
12/617,506	11/12/2009	Philippe Di Cristo	090611-0382843	4703

TITLE OF INVENTION: SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE

APPLN. TYPE	SMALL ENTITY	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	YES	$755	$300	$0	$1055	03/21/2011

EXAMINER	ART UNIT	CLASS-SUBCLASS
ALBERTALLI, BRIAN LOUIS	2626	704-270100

1. Change of correspondence address or indication of "Fee Address" (37 CFR 1.363).

❑ Change of correspondence address (or Change of Correspondence Address form PTO/SB/122) attached.

❑ "Fee Address" indication (or "Fee Address" Indication form PTO/SB/47; Rev 03-02 or more recent) attached. **Use of a Customer Number is required.**

2. For printing on the patent front page, list

(1) the names of up to 3 registered patent attorneys or agents OR, alternatively,

(2) the name of a single firm (having as a member a registered attorney or agent) and the names of up to 2 registered patent attorneys or agents. If no name is listed, no name will be printed.

1 _____

2 _____

3 _____

3. ASSIGNEE NAME AND RESIDENCE DATA TO BE PRINTED ON THE PATENT (print or type)

PLEASE NOTE: Unless an assignee is identified below, no assignee data will appear on the patent. If an assignee is identified below, the document has been filed for recordation as set forth in 37 CFR 3.11. Completion of this form is NOT a substitute for filing an assignment.

(A) NAME OF ASSIGNEE (B) RESIDENCE: (CITY and STATE OR COUNTRY)

Please check the appropriate assignee category or categories (will not be printed on the patent) : ❑ Individual ❑ Corporation or other private group entity ❑ Government

4a. The following fee(s) are submitted:

❑ Issue Fee

❑ Publication Fee (No small entity discount permitted)

❑ Advance Order - # of Copies _____

4b. Payment of Fee(s): **(Please first reapply any previously paid issue fee shown above)**

❑ A check is enclosed.

❑ Payment by credit card. Form PTO-2038 is attached.

❑ The Director is hereby authorized to charge the required fee(s), any deficiency, or credit any overpayment, to Deposit Account Number _____ (enclose an extra copy of this form).

5. **Change in Entity Status** (from status indicated above)

❑ a. Applicant claims SMALL ENTITY status. See 37 CFR 1.27.

❑ b. Applicant is no longer claiming SMALL ENTITY status. See 37 CFR 1.27(g)(2).

NOTE: The Issue Fee and Publication Fee (if required) will not be accepted from anyone other than the applicant; a registered attorney or agent; or the assignee or other party in interest as shown by the records of the United States Patent and Trademark Office.

Authorized Signature _____ Date _____

Typed or printed name _____ Registration No. _____

This collection of information is required by 37 CFR 1.311. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, Virginia 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, Virginia 22313-1450.

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

UNITED STATES PATENT AND TRADEMARK OFFICE

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

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
12/617,506	11/12/2009	Philippe Di Cristo	090611-0382843	4703

909 7590 12/20/2010

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

EXAMINER	
ALBERTALLI, BRIAN LOUIS	
ART UNIT	PAPER NUMBER
2626	

DATE MAILED: 12/20/2010

Determination of Patent Term Adjustment under 35 U.S.C. 154 (b)
(application filed on or after May 29, 2000)

The Patent Term Adjustment to date is 0 day(s). If the issue fee is paid on the date that is three months after the mailing date of this notice and the patent issues on the Tuesday before the date that is 28 weeks (six and a half months) after the mailing date of this notice, the Patent Term Adjustment will be 0 day(s).

If a Continued Prosecution Application (CPA) was filed in the above-identified application, the filing date that determines Patent Term Adjustment is the filing date of the most recent CPA.

Applicant will be able to obtain more detailed information by accessing the Patent Application Information Retrieval (PAIR) WEB site (http://pair.uspto.gov).

Any questions regarding the Patent Term Extension or Adjustment determination should be directed to the Office of Patent Legal Administration at (571)-272-7702. Questions relating to issue and publication fee payments should be directed to the Customer Service Center of the Office of Patent Publication at 1-(888)-786-0101 or (571)-272-4200.

Notice of Allowability	Application No.	Applicant(s)	
	12/617,506	DI CRISTO ET AL.	
	Examiner	Art Unit	
	BRIAN L. ALBERTALLI	2626	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address--

All claims being allowable, PROSECUTION ON THE MERITS IS (OR REMAINS) CLOSED in this application. If not included herewith (or previously mailed), a Notice of Allowance (PTOL-85) or other appropriate communication will be mailed in due course. **THIS NOTICE OF ALLOWABILITY IS NOT A GRANT OF PATENT RIGHTS.** This application is subject to withdrawal from issue at the initiative of the Office or upon petition by the applicant. See 37 CFR 1.313 and MPEP 1308.

1. ☒ This communication is responsive to *6 December 2010*.

2. ☒ The allowed claim(s) is/are *1-20*.

3. ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
   a) ☐ All b) ☐ Some* c) ☐ None of the:
      1. ☐ Certified copies of the priority documents have been received.
      2. ☐ Certified copies of the priority documents have been received in Application No. _____ .
      3. ☐ Copies of the certified copies of the priority documents have been received in this national stage application from the International Bureau (PCT Rule 17.2(a)).

   * Certified copies not received: _____.

Applicant has THREE MONTHS FROM THE "MAILING DATE" of this communication to file a reply complying with the requirements noted below. Failure to timely comply will result in ABANDONMENT of this application.
THIS THREE-MONTH PERIOD IS NOT EXTENDABLE.

4. ☐ A SUBSTITUTE OATH OR DECLARATION must be submitted. Note the attached EXAMINER'S AMENDMENT or NOTICE OF INFORMAL PATENT APPLICATION (PTO-152) which gives reason(s) why the oath or declaration is deficient.

5. ☐ CORRECTED DRAWINGS (as "replacement sheets") must be submitted.
   (a) ☐ including changes required by the Notice of Draftsperson's Patent Drawing Review (PTO-948) attached
      1) ☐ hereto or 2) ☐ to Paper No./Mail Date _____.
   (b) ☐ including changes required by the attached Examiner's Amendment / Comment or in the Office action of Paper No./Mail Date _____.

   Identifying indicia such as the application number (see 37 CFR 1.84(c)) should be written on the drawings in the front (not the back) of each sheet. Replacement sheet(s) should be labeled as such in the header according to 37 CFR 1.121(d).

6. ☐ DEPOSIT OF and/or INFORMATION about the deposit of BIOLOGICAL MATERIAL must be submitted. Note the attached Examiner's comment regarding REQUIREMENT FOR THE DEPOSIT OF BIOLOGICAL MATERIAL.

Attachment(s)

1. ☒ Notice of References Cited (PTO-892)
2. ☐ Notice of Draftperson's Patent Drawing Review (PTO-948)
3. ☐ Information Disclosure Statements (PTO/SB/08), Paper No./Mail Date _____
4. ☐ Examiner's Comment Regarding Requirement for Deposit of Biological Material
5. ☐ Notice of Informal Patent Application
6. ☐ Interview Summary (PTO-413), Paper No./Mail Date _____ .
7. ☐ Examiner's Amendment/Comment
8. ☒ Examiner's Statement of Reasons for Allowance
9. ☐ Other _____.

DETAILED ACTION

Allowable Subject Matter

Claims 1-20 are allowed.

The following is an examiner's statement of reasons for allowance:

Claims 1-20 are allowed for the reasons given in the previous Office Action. Additionally, the IDS submitted 6 December 2010 does not raise any new prior art issues or rejections.

Any comments considered necessary by applicant must be submitted no later than the payment of the issue fee and, to avoid processing delays, should preferably accompany the issue fee. Such submissions should be clearly labeled "Comments on Statement of Reasons for Allowance."

Conclusion

Any inquiry concerning this communication or earlier communications from the examiner should be directed to BRIAN L. ALBERTALLI whose telephone number is (571)272-7616. The examiner can normally be reached on Monday-Thursday, 8 AM to 6:30 PM.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, David Hudspeth can be reached on (571) 272-7843. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

BLA 12/14/10
/Brian L Albertalli/
Primary Examiner, Art Unit 2626

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	1A	4,430,669	02/07/1984	Cheung	358	122	
	1B	5,155,743	10/13/1992	Jacobs	375	28	
	1C	5,274,560	12/28/1993	LaRue	364	444	
	1D	5,377,350	12/27/1994	Skinner	395	600	
	1E	5,386,556	01/31/1995	Hedin et al.	395	600	
	1F	5,424,947	06/13/1995	Nagao et al.	364	419.08	
	1G	5,471,318	11/28/1995	Ahuja et al.	358	400	
	1H	5,475,733	12/12/1995	Eisdorfer et al.	379	52	
	1I	5,488,652	01/30/1996	Bielby *et al.*	379	88	
	1J	5,499,289	03/12/1996	Bruno et al.	379	220	
	1K	5,500,920	03/19/1996	Kupiec	395	2.79	
	1L	5,517,560	05/14/1996	Greenspan	379	114	
	1M	5,533,108	07/02/1996	Harris et al.	379	201	
	1N	5,537,436	07/16/1996	Bottoms et al.	375	222	
	1O	5,539,744	07/23/1996	Chu et al.	370	60	
	1P	5,557,667	09/17/1996	Bruno et al.	379	201	
	1Q	5,563,937	10/08/1996	Bruno et al.	379	201	
	1R	5,590,039	12/31/1996	Ikeda et al.	395	759	
	1S	5,617,407	04/01/1997	Bareis	369	275.3	
	1T	5,633,922	05/27/1997	August et al.	379	220	
	1U	5,675,629	10/07/1997	Raffel et al.	379	58	
	1V	5,696,965	12/09/1997	Dedrick	395	610	
	1W	5,708,422	01/13/1998	Blonder et al.	340	825.34	

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO
	1X	WO 99/46763	09/16/1999	International				
	1Y	WO 00/21232	04/13/2000	International				
	1Z	WO 00/46792	08/10/2000	International				
	1AA	WO 01/78065	10/18/2001	International				
	1BB	WO 2004/072954	08/26/2001	International				

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

	1CC	REUTERS, "IBM to Enable Honda Drivers to Talk to Cars", Charles Schwab & Co., Inc., July 28, 2002, 1 page.
	1DD	LIN, Bor-shen, *et al.*, "A Distributed Architecture for Cooperative Spoken Dialogue Agents with Coherent Dialogue State and History", ASRU'99, 1999, 4 pages.

EXAMINER	DATE CONSIDERED

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

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.A./

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

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

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO
	2X	WO 2007/019318	02/15/2007	International				
	2Y	WO 2007/021587	02/22/2007	International				
	2Z	WO 2007/027546	03/08/2007	International				
	2AA	WO 2007/027989	03/08/2007	International				
	2BB	WO 2008/098039	08/14/2008	International				

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

	2CC	KUHN, Thomas, *et al.*, "Hybrid In-Car Speech Recognition for Mobile Multimedia Applications", Vehicular Technology Conference, IEEE, July 1999, pages 2009-2013.
	2DD	BELVIN, Robert, *et al.*, "Development of the HRL Route Navigation Dialogue System", Proceedings of the First International Conference on Human Language Technology Research, San Diego, 2001, pages 1-5.

EXAMINER	DATE CONSIDERED

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

402527183v1

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.A./

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

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

FOREIGN PATENT DOCUMENTS

	REF	DOCUMENT NO.	DATE	COUNTRY	CLASS	SUBCLASS	Translations YES	Translations NO
	3X	WO 2008/118195	10/02/2008	International				
	3Y	WO 2009/075912	06/18/2009	International				
	3Z	WO 2009/145796	12/03/2009	International				
	3AA	WO 2010/096752	08/26/2010	International				

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

	3BB	LIND, R., *et al.*, "The Network Vehicle—A Glimpse into the Future of Mobile Multi-Media", IEEE Aerosp. Electron. Systems Magazine, Volume 14, No. 9, September 1999, pages 27-32.
	3CC	ZHAO, Yilin, "Telematics: Safe and Fun Driving", IEEE Intelligent Systems, Volume 17, Issue 1, 2002, pages 10-14.

EXAMINER	DATE CONSIDERED

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

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.A./

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

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

FOREIGN PATENT DOCUMENTS

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

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

	4X	CHAI *et al.*, "MIND: A Semantics-Based Multimodal Interpretation Framework for Conversational System", Proceedings of the International CLASS Workshop on Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems, June 2002, pages 37-46.
	4Y	CHEYER *et al.*, "Multimodal Maps: An Agent-Based Approach", International Conference on Cooperative Multimodal Communication (CMC/95), May 24-26, 1995, pages 111-121.

EXAMINER	DATE CONSIDERED

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

402527183v1

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.A./

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

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

FOREIGN PATENT DOCUMENTS

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

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

	5X	ELIO *et al.*, "On Abstract Task Models and Conversation Policies" in Workshop on Specifying and Implementing Conversation Policies, Autonomous Agents '99, Seattle, 1999, 10 pages.
	5Y	TURUNEN, "Adaptive Interaction Methods in Speech User Interfaces", Conference on Human Factors in Computing Systems, Seattle, Washington, 2001, pages 91-92.

EXAMINER	DATE CONSIDERED

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

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.A./

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

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

FOREIGN PATENT DOCUMENTS

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

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

	6X	MAO, Mark Z., "Automatic Training Set Segmentation for Multi-Pass Speech Recognition", Department of Electrical Engineering, Stanford University, CA, copyright 2005, IEEE, pages I-685 to I-688.
	6Y	VANHOUCKE, Vincent, "Confidence Scoring and Rejection Using Multi-Pass Speech Recognition", Nuance Communications, Menlo Park, CA, 2005, 4 pages.

EXAMINER	DATE CONSIDERED

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

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.A./

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

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

FOREIGN PATENT DOCUMENTS

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

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

	7X	WENG, Fuliang, *et al.*, "Efficient Lattice Representation and Generation", Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, 1998, 4 pages.
	7Y	EL MELIANI *et al.*, "A Syllabic-Filler-Based Continuous Speech Recognizer for Unlimited Vocabulary", Canadian Conference on Electrical and Computer Engineering, Volume 2, September 5-8, 1995, pages 1007-1010.

EXAMINER	DATE CONSIDERED

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

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.A./

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

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

FOREIGN PATENT DOCUMENTS

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

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

	8X	ARRINGTON, Michael, "Google Redefines GPS Navigation Landscape: Google Maps Navigation for Android 2.0", TechCrunch, printed from the Internet <http://www.techcrunch.com/2009/10/28/google-redefines-car-gps-navigation-google-maps-navigation-android/>, October 28, 2009, 4 pages.
	8Y	BAZZI, Issam *et al.*, "Heterogeneous Lexical Units for Automatic Speech Recognition: Preliminary Investigations", Processing of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Volume 3, June 5-9, 2000, XP010507574, pages 1257-1260.

EXAMINER	DATE CONSIDERED

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

402527183v1

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.A./

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	9A	7,340,040	03/04/2008	Saylor *et al.*	379	67.1	
	9B	7,376,645	05/20/2008	Bernard	707	3	
	9C	7,398,209	07/08/2008	Kennewick *et al.*	704	255	
	9D	7,424,431	09/09/2008	Greene *et al.*	704	270	
	9E	7,461,059	12/02/2008	Richardson *et al.*	707	5	
	9F	7,487,088	02/03/2009	Gorin *et al.*	704	240	
	9G	7,493,559	02/17/2009	Wolff *et al.*	715	727	
	9H	7,502,738	03/10/2009	Kennewick *et al.*	704	257	
	9I	7,620,549	11/17/2009	Di Cristo *et al.*	704	257	
	9J	7,634,409	12/15/2009	Kennewick *et al.*	704	257	
	9K	7,640,160	12/29/2009	Di Cristo *et al.*	704	257	
	9L	7,693,720	04/06/2010	Kennewick *et al.*	704	275	
	9M	7,809,570	10/05/2010	Kennewick *et al.*	704	257	
	9N	7,818,176	10/19/2010	Freeman *et al.*	704	270	
	9O	2001/0041980	11/15/2001	Howard *et al.*	704	270	
	9P	2002/0049805	04/25/2002	Yamada *et al.*	709	202	
	9Q	2002/0065568	05/30/2002	Silfvast *et al.*	700	94	
	9R	2002/0069059	06/06/2002	Smith	704	257	
	9S	2002/0082911	06/27/2002	Dunn *et al.*	705	14	
	9T	2002/0087525	07/04/2002	Abbott *et al.*	707	3	
	9U	2002/0120609	08/29/2002	Lang *et al.*	707	1	
	9V	2002/0124050	09/05/2002	Middeljans	709	203	
	9W	2002/0138248	09/26/2002	Corston-Oliver *et al.*	704	1	

FOREIGN PATENT DOCUMENTS

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

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

	9X	O'SHAUGHNESSY, Douglas, "Interacting with Computers by Voice: Automatic Speech Recognition and Synthesis", Proceedings of the IEEE, Volume 91, No. 9, September 1, 2003, XP011100665, pages 1272-1305.

EXAMINER	DATE CONSIDERED

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

402527183v1

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.A./

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	10A	2002/0143535	10/03/2002	Kist *et al.*	704	251	
	10B	2002/0188602	12/12/2002	Stubler *et al.*	707	3	
	10C	2002/0198714	12/26/2002	Zhou	704	252	
	10D	2003/0014261	01/16/2003	Kageyama	704	275	
	10E	2003/0064709	04/03/2003	Gailey *et al.*	455	412	
	10F	2003/0088421	05/08/2003	Maes *et al.*	704	270.1	
	10G	2003/0097249	05/22/2003	Walker *et al.*	704	1	
	10H	2003/0112267	06/19/2003	Belrose	345	728	
	10I	2003/0135488	07/17/2003	Amir *et al.*	707	3	
	10J	2003/0182132	09/25/2003	Niemoeller	704	275	
	10K	2003/0236664	12/25/2003	Sharma	704	251	
	10L	2004/0006475	01/08/2004	Ehlen *et al.*	704	270.1	
	10M	2004/0025115	02/05/2004	Sienel *et al.*	715	513	
	10N	2004/0044516	03/04/2004	Kennewick *et al.*	704	5	
	10O	2004/0166832	08/26/2004	Portman *et al.*	455	412.1	
	10P	2004/0193408	09/30/2004	Hunt	704	209	
	10Q	2004/0193420	09/30/2004	Kennewick *et al.*	704	257	
	10R	2004/0199375	10/07/2004	Ehsani *et al.*	704	4	
	10S	2004/0243417	12/02/2004	Pitts, III *et al.*	704	276	
	10T	2005/0015256	01/20/2005	Kargman	704	272	
	10U	2005/0021334	01/27/2005	Iwahashi	704	240	
	10V	2005/0021826	01/27/2005	Kumar	709	232	
	10W	2005/0033574	02/10/2005	Kim *et al.*	704	251	

FOREIGN PATENT DOCUMENTS

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

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

EXAMINER	DATE CONSIDERED

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

402527183v1

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.A./

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	11A	2005/0043940	02/24/2005	Elder	704	9	
	11B	2005/0114116	05/26/2005	Fiedler	704	201	
	11C	2005/0246174	11/03/2005	DeGolia	704	270	
	11D	2006/0206310	09/14/2006	Ravikumar *et al.*	704	9	
	11E	2007/0033005	02/08/2007	Cristo *et al.*	704	9	
	11F	2007/0038436	02/15/2007	Cristo *et al.*	704	9	
	11G	2007/0043574	02/22/2007	Coffman *et al.*	704	275	
	11H	2007/0050191	03/01/2007	Weider *et al.*	704	275	
	11I	2007/0055525	03/08/2007	Kennewick *et al.*	704	257	
	11J	2007/0186165	08/09/2007	Maislos *et al.*	715	728	
	11K	2007/0214182	09/13/2007	Rosenberg	707	104.1	
	11L	2007/0250901	10/25/2007	McIntire *et al.*	725	146	
	11M	2007/0265850	11/15/2007	Kennewick *et al.*	704	257	
	11N	2007/0299824	12/27/2007	Pan *et al.*	707	3	
	11O	2008/0091406	04/17/2008	Baldwin *et al.*	704	4	
	11P	2008/0103761	05/01/2008	Printz *et al.*	704	9	
	11Q	2008/0115163	05/15/2008	Gilboa *et al.*	725	34	
	11R	2008/0133215	06/05/2008	Sarukkai	704	2	
	11S	2008/0189110	08/07/2008	Freeman *et al.*	704	251	
	11T	2008/0235023	09/25/2008	Kennewick *et al.*	704	257	
	11U	2008/0235027	09/25/2008	Cross	704	270.1	
	11V	2008/0319751	12/25/2008	Kennewick *et al.*	704	257	
	11W	2009/0117885	05/07/2009	Roth	455	414.3	

FOREIGN PATENT DOCUMENTS

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

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

EXAMINER	DATE CONSIDERED

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

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.A./

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	12A	2009/0144271	06/04/2009	Richardson *et al.*	707	5	
	12B	2009/0150156	06/11/2009	Kennewick *et al.*	704	257	
	12C	2009/0171664	07/02/2009	Kennewick *et al.*	704	257	
	12D	2009/0299745	12/03/2009	Kennewick *et al.*	704	257	
	12E	2010/0023320	01/28/2010	Di Cristo *et al.*	704	9	
	12F	2010/0049501	02/25/2010	Kennewick *et al.*	704	9	
	12G	2010/0049514	02/25/2010	Kennewick *et al.*	704	233	
	12H	2010/0063880	03/11/2010	Atsmon *et al.*	705	14.53	
	12I	2010/0145700	06/10/2010	Kennewick *et al.*	704	257	
	12J	2010/0204986	08/12/2010	Kennewick *et al.*	704	226	
	12K	2010/0204994	08/12/2010	Kennewick *et al.*	704	257	
	12L	2010/0217604	08/26/2010	Baldwin *et al.*	704	275	
	12M	2010/0286985	11/11/2010	Kennewick *et al.*	704	257	
	12N	2010/0299142	11/25/2010	Freeman *et al.*	704	9	

FOREIGN PATENT DOCUMENTS

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

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

EXAMINER	DATE CONSIDERED
/Brian Albertalli/	12/14/2010

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

402527183v1

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.A./

Issue Classification	Application/Control No.	Applicant(s)/Patent Under Reexamination
12617506	12617506	DI CRISTO ET AL.
	Examiner	**Art Unit**
	BRIAN L ALBERTALLI	2626

ORIGINAL		INTERNATIONAL CLASSIFICATION															
CLASS	SUBCLASS	CLAIMED										NON-CLAIMED					
704	270.1	G	1	0	L	21 / 00 (2006.0)											
CROSS REFERENCE(S)		G	0	6	F	17 / 27 (2006.0)											
		G	1	0	L	15 / 18 (2006.0)											

CLASS	SUBCLASS (ONE SUBCLASS PER BLOCK)				
704	9	257			

☒ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original	Final	Original

NONE		Total Claims Allowed:	
(Assistant Examiner)	(Date)	20	
/BRIAN L ALBERTALLI/ Primary Examiner.Art Unit 2626	12/14/10	O.G. Print Claim(s)	O.G. Print Figure
(Primary Examiner)	(Date)	1	7

Search Notes	Application/Control No.	Applicant(s)/Patent Under Reexamination
1261750 6	12617506	DI CRISTO ET AL.
	Examiner	**Art Unit**
	BRIAN L ALBERTALLI	2626

SEARCHED			
Class	**Subclass**	**Date**	**Examiner**

SEARCH NOTES		
Search Notes	**Date**	**Examiner**
Updated EAST Search from parent case 11/197,504	10/21/10	BA
IEEE Search	10/21/10	BA
INSPEC Search	10/21/10	BA
Google Scholar Search	10/21/10	BA
Updated EAST Search - see search history printout	12/14/10	BA

INTERFERENCE SEARCH			
Class	**Subclass**	**Date**	**Examiner**
	Interference Search	10/22/10	BA

Index of Claims	Application/Control No.	Applicant(s)/Patent Under Reexamination
12617506	12617506	DI CRISTO ET AL.
	Examiner	**Art Unit**
	BRIAN L ALBERTALLI	2626

✓	Rejected	-	Cancelled	N	Non-Elected	A	Appeal
=	Allowed	÷	Restricted	I	Interference	O	Objected

☒ Claims renumbered in the same order as presented by applicant ☐ CPA ☐ T.D. ☐ R.1.47

CLAIM		DATE								
Final	Original	10/22/2010	12/14/2010							
	1	=	=							
	2	=	=							
	3	=	=							
	4	=	=							
	5	=	=							
	6	=	=							
	7	=	=							
	8	=	=							
	9	=	=							
	10	=	=							
	11	=	=							
	12	=	=							
	13	=	=							
	14	=	=							
	15	=	=							
	16	=	=							
	17	=	=							
	18	=	=							
	19	=	=							
	20	=	=							

EAST Search History

EAST Search History (Prior Art)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
S1	822	context near2 xml	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/14 13:30
S2	30	S1 and "704"/$.ccls.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/14 13:31
S3	95	"704"/$.ccls. and (synchroni$10 with context$5)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/14 13:42
S6	902	"704"/$.ccls. and (mobile$5).ab.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/14 14:03
S7	119	"704"/$.ccls. and (synchroni$10 with mobile$5)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/15 11:00
S8	145	(synchroni$10 with context$5 with mobile$5)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/15 12:06
S9	83	S8 and (speech or voice)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/15 12:18

S10	578	"704"/$.ccls. and synchroni$10.ab.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/15 12:23
S11	902	"704"/$.ccls. and (mobile$5).ab.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/15 12:23
S12	17	S11 and S10	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/15 12:23
S13	1020	"704"/$.ccls. and synchroni$10.clm.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/15 12:34
S14	1477	"704"/$.ccls. and (mobile$5).clm.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/15 12:34
S15	64	S13 and S14	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/15 12:34
S16	69	"704"/$.ccls. and (synchroni$10 with ((language$5 near2 model$5) or grammar$5))	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/15 12:54
S17	1082	704/270.1.ccls.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/15 13:07

S18	255	S17 and synchroni $10	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/15 13:07
S19	207	S18 and ((language $5 near2 model$5) or grammar$5 or context$5 or (natural near2 language$5))	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/15 13:08
S20	181	S18 and context$10	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/15 13:10
S21	126	"704"/$.ccls. and (synchroni$10 with client$5)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/10/15 13:24
S25	46	US-4430669-$.DID. OR US-5155743-$.DID. OR US-5274560-$.DID. OR US-5377350-$.DID. OR US-5386556-$.DID. OR US-5424947-$.DID. OR US-5471318-$.DID. OR US-5475733-$.DID. OR US-5488652-$.DID. OR US-5499289-$.DID. OR US-5500920-$.DID. OR US-5517560-$.DID. OR US-5533108-$.DID. OR US-5537436-$.DID. OR US-5539744-$.DID. OR US-5557667-$.DID. OR US-5563937-$.DID. OR US-5590039-$.DID. OR US-5617407-$.DID. OR US-5633922-$.DID. OR US-	US-PGPUB; USPAT; USOCR; DERWENT	OR	OFF	2010/12/13 09:39

		5675629-$.DID. OR US-5696965-$.DID. OR US-5708422-$.DID.				
S26	44	US-5721938-$.DID. OR US-5722084-$.DID. OR US-5742763-$.DID. OR US-5748841-$.DID. OR US-5752052-$.DID. OR US-5754784-$.DID. OR US-5761631-$.DID. OR US-5774859-$.DID. OR US-5794196-$.DID. OR US-5797112-$.DID. OR US-5799276-$.DID. OR US-5802510-$.DID. OR US-5832221-$.DID. OR US-5867817-$.DID. OR US-5878385-$.DID. OR US-5878386-$.DID. OR US-5892813-$.DID. OR US-5895464-$.DID. OR US-5895466-$.DID. OR US-5902347-$.DID. OR US-5911120-$.DID. OR US-5918222-$.DID.	US-PGPUB; USPAT; USOCR; DERWENT	OR	OFF	2010/12/13 09:44
S27	46	US-5926784-$.DID. OR US-5933822-$.DID. OR US-5953393-$.DID. OR US-5963894-$.DID. OR US-5963940-$.DID. OR US-5987404-$.DID. OR US-5991721-$.DID. OR US-5995119-$.DID. OR US-6009382-$.DID. OR US-6014559-$.DID. OR US-6021384-$.DID. OR US-6044347-$.DID. OR US-6049602-$.DID. OR US-6058187-$.DID. OR US-	US-PGPUB; USPAT; USOCR; DERWENT	OR	OFF	2010/12/13 09:47

		6078886-$.DID. OR US-6081774-$.DID. OR US-6101241-$.DID. OR US-6119087-$.DID. OR US-6134235-$.DID. OR US-6144667-$.DID. OR US-6144938-$.DID. OR US-6154526-$.DID. OR US-6160883-$.DID.				
S28	46	US-6173279-$.DID. OR US-6175858-$.DID. OR US-6185535-$.DID. OR US-6192110-$.DID. OR US-6192338-$.DID. OR US-6195634-$.DID. OR US-6195651-$.DID. OR US-6208972-$.DID. OR US-6219346-$.DID. OR US-6219643-$.DID. OR US-6226612-$.DID. OR US-6233556-$.DID. OR US-6233559-$.DID. OR US-6246981-$.DID. OR US-6272455-$.DID. OR US-6288319-$.DID. OR US-6292767-$.DID. OR US-6308151-$.DID. OR US-6314402-$.DID. OR US-6366886-$.DID. OR US-6374214-$.DID. OR US-6381535-$.DID. OR US-6385596-$.DID.	US-PGPUB; USPAT; USOCR; DERWENT	OR	OFF	2010/12/13 09:48

S29	46	US-6385646-$.DID. OR US-6393428-$.DID. OR US-6404878-$.DID. OR US-6408272-$.DID. OR US-6411810-$.DID. OR US-6415257-$.DID. OR US-6418210-$.DID. OR US-6420975-$.DID. OR US-6430285-$.DID. OR US-6430531-$.DID. OR US-6434523-$.DID. OR US-6434524-$.DID. OR US-6442522-$.DID. OR US-6446114-$.DID. OR US-6453153-$.DID. OR US-6453292-$.DID. OR US-6456711-$.DID. OR US-6466654-$.DID. OR US-6466899-$.DID. OR US-6470315-$.DID. OR US-6498797-$.DID. OR US-6499013-$.DID. OR US-6501833-$.DID.	US-PGPUB; USPAT; USOCR; DERWENT	OR	OFF	2010/12/13 09:49
S30	46	US-6501834-$.DID. OR US-6513006-$.DID. OR US-6523061-$.DID. OR US-6532444-$.DID. OR US-6539348-$.DID. OR US-6553372-$.DID. OR US-6556970-$.DID. OR US-6556973-$.DID. OR US-6560576-$.DID. OR US-6567778-$.DID. OR US-6567797-$.DID. OR US-6570555-$.DID. OR US-6570964-$.DID. OR US-6574597-$.DID. OR US-6574624-$.DID. OR US-6581103-$.DID.	US-PGPUB; USPAT; USOCR; DERWENT	OR	OFF	2010/12/13 09:50

		OR US-6594257-$. DID. OR US-6598018-$.DID. OR US-6604075-$.DID. OR US-6604077-$. DID. OR US-6611692-$.DID. OR US-6614773-$.DID. OR US-6615172-$. DID.				
S31	46	US-6629066-$.DID. OR US-6631346-$. DID. OR US-6643620-$.DID. OR US-6650747-$.DID. OR US-6658388-$. DID. OR US-6678680-$.DID. OR US-6691151-$.DID. OR US-6708150-$. DID. OR US-6721001-$.DID. OR US-6721706-$.DID. OR US-6735592-$. DID. OR US-6741931-$.DID. OR US-6742021-$.DID. OR US-6751591-$. DID. OR US-6757718-$.DID. OR US-6795808-$.DID. OR US-6801604-$. DID. OR US-6801893-$.DID. OR US-6829603-$.DID. OR US-6833848-$. DID. OR US-6856990-$.DID. OR US-6865481-$.DID. OR US-6877134-$. DID.	US-PGPUB; USPAT; USOCR; DERWENT	OR	OFF	2010/12/13 09:50

S32	46	US-6901366-$.DID. OR US-6937977-$.DID. OR US-6944594-$.DID. OR US-6950821-$.DID. OR US-6973387-$.DID. OR US-6980092-$.DID. OR US-6990513-$.DID. OR US-6996531-$.DID. OR US-7016849-$.DID. OR US-7020609-$.DID. OR US-7027975-$.DID. OR US-7062488-$.DID. OR US-7082469-$.DID. OR US-7092928-$.DID. OR US-7127400-$.DID. OR US-7136875-$.DID. OR US-7137126-$.DID. OR US-7146319-$.DID. OR US-7203644-$.DID. OR US-7228276-$.DID. OR US-7231343-$.DID. OR US-7289606-$.DID. OR US-7301093-$.DID.	US-PGPUB; USPAT; USOCR; DERWENT	OR	OFF	2010/12/13 09:51
S33	46	US-7340040-$.DID. OR US-7376645-$.DID. OR US-7398209-$.DID. OR US-7424431-$.DID. OR US-7461059-$.DID. OR US-7487088-$.DID. OR US-7493559-$.DID. OR US-7502738-$.DID. OR US-7620549-$.DID. OR US-7634409-$.DID. OR US-7640160-$.DID. OR US-7693720-$.DID. OR US-7809570-$.DID. OR US-7818176-$.DID. OR US-20010041980-$.DID. OR US-	US-PGPUB; USPAT; USOCR; DERWENT	OR	OFF	2010/12/13 09:51

		20020049805-$. DID. OR US-20020065568-$. DID. OR US-20020069059-$. DID. OR US-20020082911-$. DID. OR US-20020087525-$. DID. OR US-20020120609-$. DID. OR US-20020124050-$. DID. OR US-20020138248-$. DID.				
S34	44	US-20020143535-$. DID. OR US-20020188602-$. DID. OR US-20020198714-$. DID. OR US-20030014261-$. DID. OR US-20030064709-$. DID. OR US-20030088421-$. DID. OR US-20030097249-$. DID. OR US-20030112267-$. DID. OR US-20030135488-$. DID. OR US-20030182132-$. DID. OR US-20030236664-$. DID. OR US-20040006475-$. DID. OR US-20040025115-$. DID. OR US-20040044516-$. DID. OR US-20040166832-$. DID. OR US-20040193408-$. DID. OR US-20040193420-$. DID. OR US-20040199375-$. DID. OR US-20040243417-$.	US-PGPUB; USPAT; USOCR; DERWENT	OR	OFF	2010/12/13 09:52

		DID. OR US-20050021334-$. DID. OR US-20050021826-$. DID. OR US-20050033574-$. DID.				
S35	46	US-20050043940-$. DID. OR US-20050114116-$. DID. OR US-20050246174-$. DID. OR US-20060206310-$. DID. OR US-20070033005-$. DID. OR US-20070038436-$. DID. OR US-20070043574-$. DID. OR US-20070050191-$. DID. OR US-20070055525-$. DID. OR US-20070186165-$. DID. OR US-20070214182-$. DID. OR US-20070250901-$. DID. OR US-20070265850-$. DID. OR US-20070299824-$. DID. OR US-20080091406-$. DID. OR US-20080103761-$. DID. OR US-20080115163-$. DID. OR US-20080133215-$. DID. OR US-20080189110-$. DID. OR US-20080235023-$. DID. OR US-20080235027-$. DID. OR US-20080319751-$. DID. OR US-20090117885-$. DID.	US-PGPUB; USPAT; USOCR; DERWENT	OR	OFF	2010/12/13 09:52

S36	6	US-20090144271-$. DID. OR US-20090150156-$. DID. OR US-20090299745-$. DID.	US-PGPUB; USPAT; USOCR; DERWENT	OR	OFF	2010/12/13 09:53
S37	491	(S25 S26 S27 S28 S29 S30 S31 S32 S33 S34 S35 S36)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/12/13 09:54
S38	208	S37 and (context$5)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/12/13 09:54
S39	70	S37 and (context$5).ab.	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/12/13 09:54
S40	7	S37 and ((sync$5 or synchroni$10) with context$5)	US-PGPUB; USPAT; USOCR; FPRS; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2010/12/13 09:56

EAST Search History (Interference)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
S22	273	(synchroni$10 with context$5).clm.	US-PGPUB; USPAT; UPAD	OR	OFF	2010/10/22 11:56
S23	24	(synchroni$10 with ((language$5 near2 model$5) or grammar$5)).clm.	US-PGPUB; USPAT; UPAD	OR	OFF	2010/10/22 11:57
S24	38	S22 and (speech or voice or language$5).clm.	US-PGPUB; USPAT; UPAD	OR	OFF	2010/10/22 11:57

12/14/10 10:52:04 AM
C:\Documents and Settings\balbertalli\My Documents\EAST\Workspaces\12617506.

w sp

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	A	6,622,119	09/16/2003	Ramaswamy *et al.*	704	9	
	B	7,415,414	08/19/2008	Azara *et al.*	704	270	

FOREIGN PATENT DOCUMENTS

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

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

EXAMINER	DATE CONSIDERED

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

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANTS :	Philippe DI CRISTO *et al.*	CONFIRMATION NO.:	4703
SERIAL NUMBER :	12/617,506	EXAMINER:	Brian Louis Albertalli
FILING DATE :	November 12, 2009	ART UNIT:	2626
FOR :	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE		

FURTHER SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT

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

Sir:

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

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

This Further Supplemental Information Disclosure Statement is being filed more than three months after the filing date of a national application other than a continued prosecution application under 37 C.F.R. § 1.53(d), or more than three months after the date of entry of the national stage as set forth in 37 C.F.R. § 1.491 in an international application, and after the mailing date of any of a Final Action under 37 C.F.R. § 1.113, a Notice of Allowance under 37

C.F.R. § 1.311, or an action that otherwise closes prosecution in the application, but on or before payment of the Issue Fee. 37 C.F.R. § 1.97(d).

I hereby certify that each item of information contained in this Further Supplemental Information Disclosure Statement was first cited in any communication from a foreign patent office in a counterpart foreign application not more than three months prior to the filing of this Further Supplemental Information Disclosure Statement. 37 C.F.R. § 1.97(e)(1).

These documents were cited in a search report by a foreign patent office in a counterpart foreign application. Submission of an English language version of the search report that indicates the degree of relevance found by the foreign office is provided.

The $180.00 fee [37 C.F.R. § 1.17(p)] is being paid with the Supplemental Information Disclosure Statement filed on even date herewith.

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

Respectfully submitted,

PILLSBURY WINTHROP SHAW PITTMAN LLP

By_____
Syed Jafar Ali
Registration No.: 58,780

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

Customer No.
00909

Dated: January 21, 2011

PATENT COOPERATION TREATY

PCT

INTERNATIONAL SEARCH REPORT

(PCT Article 18 and Rules 43 and 44)

Applicant's or agent's file reference 90611-390977	FOR FURTHER ACTION	see Form PCT/ISA/220 as well as, where applicable, item 5 below.
International application No. PCT/US 10/56109	International filing date *(day/month/year)* 10 November 2010 (10.11.2010)	(Earliest) Priority Date *(day/month/year)* 10 November 2009 (10.11.2009)
Applicant VOICEBOX TECHNOLOGIES, INC.		

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

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

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

1. Basis of the report

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

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

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

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

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

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

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

4. With regard to the title,

☒ the text is approved as submitted by the applicant.

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

5. With regard to the abstract,

☒ the text is approved as submitted by the applicant.

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

6. With regard to the drawings,

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

☒ as suggested by the applicant.

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

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

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

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 10/56109

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - G06F 17/27 (2010.01)
USPC - 704/9
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC: 704/9

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 704/257; 704/231; 704/E15.001 (see terms below)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PubWEST(PGPB,USPT,USOC,EPAB,JPAB), GOOGLE SCHOLAR
terms: purchase, paid subscription, content dedication, natural language, transaction, utterance, song, movie, video, buy to own, pay, annotate, tag, recipient, multimodel, interact, metadata.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2004/0193420 A1 (Kennewick et al.) 30 September 2004 (30.09.2004), fig. 1, 3, 5-6, para [0013]-[0020], [0082], [0095]-[0182].	1-20
A	US 2009/0216540 A1 (Tessel et al.) 27 August 2009 (27.08.2009), entire document.	1-20
A	US 2008/0140385 A1 (Mahajan et al.) 12 June 2008 (12.06.2008), entire document.	1-20

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

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

Date of the actual completion of the international search	Date of mailing of the international search report
18 December 2010 (18.12.2010)	07 JAN 2011
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (July 2009)

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	A	2008/0140385	06/12/2008	Mahajan *et al.*	704	9	
	B	2009/0216540	08/27/2009	Tessel *et al.*	704	275	

FOREIGN PATENT DOCUMENTS

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

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

EXAMINER	DATE CONSIDERED

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

Electronic Patent Application Fee Transmittal

Application Number:	12617506
Filing Date:	12-Nov-2009
Title of Invention:	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE
First Named Inventor/Applicant Name:	Philippe Di Cristo
Filer:	Syed Jafar Ali/Susan Trader
Attorney Docket Number:	090611-0382843

Filed as Large Entity

Utility under 35 USC 111(a) Filing Fees

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Extension-of-Time:				

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Miscellaneous:				
Submission- Information Disclosure Stmt	1806	1	180	180
		Total in USD ($)		**180**

Electronic Acknowledgement Receipt

EFS ID:	9277931
Application Number:	12617506
International Application Number:	
Confirmation Number:	4703
Title of Invention:	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE
First Named Inventor/Applicant Name:	Philippe Di Cristo
Customer Number:	00909
Filer:	Syed Jafar Ali/Susan Trader
Filer Authorized By:	Syed Jafar Ali
Attorney Docket Number:	090611-0382843
Receipt Date:	21-JAN-2011
Filing Date:	12-NOV-2009
Time Stamp:	12:01:35
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	Deposit Account
Payment was successfully received in RAM	$180
RAM confirmation Number	9514
Deposit Account	033975
Authorized User	

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)

1		0382843_SIDS.pdf	210524 9cdaa2830b8b46544715f27a50109729242e143c	yes	4
	Multipart Description/PDF files in .zip description				
	Document Description		**Start**	**End**	
	Transmittal Letter		1	3	
	Information Disclosure Statement (IDS) Filed (SB/08)		4	4	
Warnings:					
Information:					
2		0382843_FSIDS.pdf	337295 772fef43b9eb5b5d272283b68e58f185b35d0851	yes	5
	Multipart Description/PDF files in .zip description				
	Document Description		**Start**	**End**	
	Transmittal Letter		1	4	
	Information Disclosure Statement (IDS) Filed (SB/08)		5	5	
Warnings:					
Information:					
3	Fee Worksheet (PTO-875)	fee-info.pdf	30522 14ee2777c434e6a2e4d17d1f3f7824986a89e58a	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	578341		

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

New Applications Under 35 U.S.C. 111

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

National Stage of an International Application under 35 U.S.C. 371

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

New International Application Filed with the USPTO as a Receiving Office

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

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANTS :	Philippe DI CRISTO *et al.*	CONFIRMATION NO.:	4703
SERIAL NUMBER :	12/617,506	EXAMINER:	Brian Louis Albertalli
FILING DATE :	November 12, 2009	ART UNIT:	2626
FOR :	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE		

SUPPLEMENTAL INFORMATION DISCLOSURE STATEMENT

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

Sir:

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

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

This Supplemental Information Disclosure Statement is being filed more than three months after the filing date of a national application other than a continued prosecution

application under 37 C.F.R. § 1.53(d), or more than three months after the date of entry of the national stage as set forth in 37 C.F.R. § 1.491 in an international application, and after the mailing date of any of a Final Action under 37 C.F.R. § 1.113, a Notice of Allowance under 37 C.F.R. § 1.311, or an action that otherwise closes prosecution in the application, but on or before payment of the Issue Fee. 37 C.F.R. § 1.97(d).

These documents were cited by the U.S. Patent Office in co-pending Application Serial No. 12/571,795, filed October 1, 2009, in an Office Action dated December 21, 2010. I hereby certify that no item of information contained in this Supplemental Information Disclosure Statement was cited in a communication from a foreign patent office in a counterpart foreign application, and, to my knowledge after making reasonable inquiry, no item of information contained in this Supplemental Information Disclosure Statement was known to any individual designated in 37 C.F.R. § 1.56(c) more than three months prior to the filing of this Supplemental Information Disclosure Statement. 37 C.F.R. § 1.97(e)(2).

Please charge the $180.00 fee [37 C.F.R. § 1.17(p)] to Deposit Account No. 03-3975. In addition, please charge any fee deficiencies, or credit any overpayment, to Deposit Account No. 03-3975 as needed to ensure consideration of the Supplemental Information Disclosure Statement.

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

Respectfully submitted,

PILLSBURY WINTHROP SHAW PITTMAN LLP

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

Customer No.
00909

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

Dated: January 21, 2011

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICANTS :	Philippe DI CRISTO et al.	CONFIRMATION NO. :	4703
SERIAL NUMBER :	12/617,506	EXAMINER :	Brian Louis Albertalli
FILING DATE :	November 12, 2009	ART UNIT :	2626
FOR :	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE		

AMENDMENT UNDER 37 C.F.R. § 1.312

Mail Stop Issue Fee
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Dear Sir:

In response to the Notice of Allowance and Notice of Allowability mailed **December 20, 2010** ("Notice of Allowance"), and prior to payment of the Issue Fee, please amend the above-identified application as follows pursuant to 37 C.F.R. § 1.312 and MPEP § 714.16:

Amendments to the Specification begin on page **2** of this paper.

Amendments to the Claims begin on page **3** of this paper.

Remarks/Arguments begin on page **9** of this paper.

It is believed that no extensions of time or fees for net addition of claims are required beyond those that may otherwise be provided for in documents accompanying this paper. However, if additional extensions of time are necessary to prevent abandonment of this application, then such extensions of time are hereby petitioned for under 37 C.F.R. § 1.136(a), and any fees required therefore (including fees for net addition of claims) are hereby authorized to be charged to our Deposit Account No. 033975 (**Ref. No. 090611-0382843**).

AMENDMENT

Amendments to the Specification: (identifying insertions and ~~deletions~~).

- Please amend the section entitled "Cross-Reference to Related Applications," which begins on page 1, as follows:

This application is a continuation of ~~co-pending~~ U.S. Patent Application Serial No. 11/197,504, entitled "SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE," filed August 5, 2005, which issued as U.S. Patent No. 7,640,160 on December 29, 2009, the contents of which are hereby incorporated by reference in their entirety.

Amendments to the Claims: Please replace all prior versions and listings of claims with the following listing of claims.

LISTING OF CLAIMS:

1. (**Currently Amended**) A system for processing multi-modal natural language inputs, comprising:

a context manager communicatively coupled to a plurality of mobile devices, wherein the context manager includes one or more processors [[is]] configured to:

register the plurality of mobile devices with the context manager in response to a registration module associated with the context manager receiving a communication from the plurality of mobile devices;

subscribe the plurality of mobile devices registered with the context manager to one or more context events;

receive a context input from one or more of the plurality of mobile devices registered with the context manager, wherein the context input includes a context change event; and

inform the plurality of mobile devices registered with the context manager of the context change event, thereby synchronizing a context across the plurality of mobile devices.

2. (**Currently Amended**) The system of claim 1, wherein the one or more processors are ~~context manager is~~ configured to inform the plurality of mobile devices registered with the context manager of the context change event through a context tracking module associated with the context manager.

3. (**Original**) The system of claim 2, wherein the context tracking module is configured to track the context synchronized across the plurality of mobile devices.

4. (**Currently Amended**) The system of claim 1, wherein the one or more processors are configured to receive ~~context manager receives~~ the context input in a Context XML format.

5. (**Currently Amended**) The system of claim 1, wherein the ~~context manager is~~ one or more processors are further configured to remove one or more of the plurality of mobile devices from the plurality of mobile devices registered with the context manager.

6. (**Original**) The system of claim 1, wherein the context manager includes a plurality of dynamic link libraries that are respectively specific to the plurality of mobile devices registered with the context manager.

7. (**Original**) The system of claim 1, wherein at least one of the plurality of mobile devices registered with the context manager include:

a multi-modal voice user interface configured to receive a natural language utterance; and

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 contained in text transcribed from the natural language utterance, wherein identifying the matching contexts includes comparing the text combinations against one or more grammar expression entries in a context description grammar and one or more expected contexts stored in a context stack synchronized through the context manager;

score 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.

8. (**Currently Amended**) The system of claim 7, wherein the context change event received at the one or more processors ~~context manager~~ includes a communication from the at least one mobile device that indicates a change in context to the most likely context.

9. (**Original**) The system of claim 7, wherein the at least one mobile device further includes a multi-pass speech recognition module configured to:

use at least one of a dictation grammar or a virtual dictation grammar to transcribe the natural language utterance into the text, wherein the dictation grammar includes an unconstrained large vocabulary of words, and wherein the virtual dictation grammar includes a constrained vocabulary of words and a plurality of decoy words for out-of-vocabulary words;

dynamically generate the constrained vocabulary of words based on one or more previously successful transcriptions; and

transcribe the utterance using the multi-pass speech recognition module.

10. (**Original**) The system of claim 9, wherein the decoy words include utility words, nonsense words, isolated syllables, and isolated distinct sounds associated with a particular spoken language.

11. (**Original**) A method for processing multi-modal natural language inputs, comprising:

registering a plurality of mobile devices with a context manager in response to a registration module associated with the context manager receiving a communication from the plurality of mobile devices;

subscribing the plurality of mobile devices registered with the context manager to one or more context events;

receiving, at the context manager, a context input from one or more of the plurality of mobile devices registered with the context manager, wherein the context input includes a context change event; and

informing the plurality of mobile devices registered with the context manager of the context change event, wherein informing the plurality of mobile devices registered with the context manager of the context change event synchronizes a context across the plurality of mobile devices.

12. (**Original**) The method of claim 11, wherein a context tracking module associated with the context manager informs the plurality of mobile devices registered with the context manager of the context change event.

13. (**Original**) The method of claim 12, further comprising tracking, by the context tracking module, the context synchronized across the plurality of mobile devices.

14. (**Original**) The method of claim 11, wherein the context manager receives the context input in a Context XML format.

15. (**Original**) The method of claim 11, further comprising removing, by the context manager, one or more of the plurality of mobile devices from the plurality of mobile devices registered with the context manager.

16. (**Original**) The method of claim 11, wherein the context manager includes a plurality of dynamic link libraries that are respectively specific to the plurality of mobile devices registered with the context manager.

17. (**Original**) The method of claim 11, further comprising:

receiving a natural language utterance at a multi-modal voice user interface associated with at least one of the plurality of mobile devices registered with the context manager;

identifying, using a knowledge-enhanced speech recognition engine associated with the at least one mobile device, one or more contexts that completely or partially match one or more text combinations contained in text transcribed from the natural language utterance, wherein identifying the matching contexts includes comparing the text combinations against one or more grammar expression entries in a context description grammar and one or more expected contexts stored in a context stack synchronized through the context manager;

scoring each of the identified matching contexts; and

selecting the matching context having a highest score as a most likely context for the natural language utterance.

18. (**Original**) The method of claim 17, wherein the context change event received at the context manager includes a communication from the at least one mobile device that indicates a change in context to the most likely context.

19. (**Original**) The method of claim 17, further comprising:

using, by a multi-pass speech recognition module associated with the at least one mobile device, at least one of a dictation grammar or a virtual dictation grammar to transcribe the natural language utterance into the text, wherein the dictation grammar includes an unconstrained large vocabulary of words, and wherein the virtual dictation grammar includes a constrained vocabulary of words and a plurality of decoy words for out-of-vocabulary words;

dynamically generating the constrained vocabulary of words based on one or more previously successful transcriptions; and

transcribing the utterance using the multi-pass speech recognition module.

20. (**Original**) The method of claim 19, wherein the decoy words include utility words, nonsense words, isolated syllables, and isolated distinct sounds associated with a particular spoken language.

REMARKS

Upon entry of the foregoing Amendment, claims 1-20 are pending in the application. The Specification has been amended to update the data associated with the cross-referenced related application. Claims 1-2, 4-5, and 8 have been amended. No claims have been cancelled or newly added. Applicants believe that this Amendment does not add new matter and submit the following Remarks pursuant to 37 C.F.R. § 1.312 and MPEP § 714.16.

EXAMINER INTERVIEW

Applicants thank Examiner Albertalli for taking the time to discuss the issues addressed in the foregoing Amendment with Applicants' representative on February 2, 2011. During the discussion with the Examiner, Applicants' representative proposed making the foregoing minor claim amendments to improve the form and clarity associated with the claims, and the Examiner indicated that the foregoing Amendment would be entered. As such, Applicants respectfully request notice to that effect.

REASONS FOR AMENDMENT

Applicants submit that the foregoing Amendment is necessary to address minor issues relating to the form and clarity associated with system claims 1-10, and further to update the data associated with the cross-referenced related application in the Specification.

The amendments to claims 1-2, 4-5, and 8 and the amendments to the Specification were not presented earlier because the present application was allowed in the first Office Action, and therefore the need to address the minor claim issues and update the data associated with the cross-referenced related application was not discovered until after the Notice of Allowance was mailed. Additionally, because the foregoing Amendment does not alter the scope or interpretation associated with claims 1-2, 4-5, or 8, the claim amendments would not require any additional search or examination because the amendments merely address minor informalities in the claim language. The currently pending claims therefore remain patentable for at least the same reasons set forth in the Notice of Allowance.

CONCLUSION

Accordingly, for at least the reasons discussed above, and because this Amendment is being submitted prior to payment of the Issue Fee, the foregoing Amendment satisfies the requirements of 37 C.F.R. § 1.312 and the guidelines set forth in MPEP § 714.16. Applicants therefore respectfully request entry of the foregoing Amendment into the record.

If the Examiner believes, for any reason, that personal communication will expedite prosecution of this application, the Examiner is invited to telephone the undersigned at the number provided.

Date: **February 2, 2011**

Respectfully submitted,

By: _____

Syed Jafar Ali
Registration No. 58,780

PILLSBURY WINTHROP SHAW PITTMAN LLP
P.O. Box 10500
McLean, Virginia 22102
Main: 703-770-7900
Direct: 703-770-7540
Fax: 703-770-7901

Electronic Acknowledgement Receipt

EFS ID:	9363673
Application Number:	12617506
International Application Number:	
Confirmation Number:	4703
Title of Invention:	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE
First Named Inventor/Applicant Name:	Philippe Di Cristo
Customer Number:	00909
Filer:	Syed Jafar Ali/Miriam E. Nelsen
Filer Authorized By:	Syed Jafar Ali
Attorney Docket Number:	090611-0382843
Receipt Date:	02-FEB-2011
Filing Date:	12-NOV-2009
Time Stamp:	17:25:56
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1		090611-0382843_312Amendment.pdf	305089 b0a653e762221f1a391d8c17a3aa89e54f37a065	yes	10

	Multipart Description/PDF files in .zip description		
	Document Description	**Start**	**End**
	Amendment after Notice of Allowance (Rule 312)	1	1
	Specification	2	2
	Claims	3	8
	Applicant Arguments/Remarks Made in an Amendment	9	10
Warnings:			
Information:			
	Total Files Size (in bytes):	305089	

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

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

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

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

UNITED STATES PATENT AND TRADEMARK OFFICE

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

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
12/617,506	11/12/2009	Philippe Di Cristo	090611-0382843	4703

909 7590 02/22/2011

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

EXAMINER	
ALBERTALLI, BRIAN LOUIS	

ART UNIT	PAPER NUMBER
2626	

NOTIFICATION DATE	DELIVERY MODE
02/22/2011	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

docket_ip@pillsburylaw.com
margaret.drosos@pillsburylaw.com

Supplemental Notice of Allowability	Application No. 12/617,506	Applicant(s) DI CRISTO ET AL.	
	Examiner BRIAN L. ALBERTALLI	Art Unit 2626	

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address--

All claims being allowable, PROSECUTION ON THE MERITS IS (OR REMAINS) CLOSED in this application. If not included herewith (or previously mailed), a Notice of Allowance (PTOL-85) or other appropriate communication will be mailed in due course. **THIS NOTICE OF ALLOWABILITY IS NOT A GRANT OF PATENT RIGHTS.** This application is subject to withdrawal from issue at the initiative of the Office or upon petition by the applicant. See 37 CFR 1.313 and MPEP 1308.

1. ☒ This communication is responsive to *21 January 2011*.

2. ☐ The allowed claim(s) is/are _____.

3. ☐ Acknowledgment is made of a claim for foreign priority under 35 U.S.C. § 119(a)-(d) or (f).
   a) ☐ All b) ☐ Some* c) ☐ None of the:
      1. ☐ Certified copies of the priority documents have been received.
      2. ☐ Certified copies of the priority documents have been received in Application No. _____ .
      3. ☐ Copies of the certified copies of the priority documents have been received in this national stage application from the International Bureau (PCT Rule 17.2(a)).

   * Certified copies not received: _____.

Applicant has THREE MONTHS FROM THE "MAILING DATE" of this communication to file a reply complying with the requirements noted below. Failure to timely comply will result in ABANDONMENT of this application.
THIS THREE-MONTH PERIOD IS NOT EXTENDABLE.

4. ☐ A SUBSTITUTE OATH OR DECLARATION must be submitted. Note the attached EXAMINER'S AMENDMENT or NOTICE OF INFORMAL PATENT APPLICATION (PTO-152) which gives reason(s) why the oath or declaration is deficient.

5. ☐ CORRECTED DRAWINGS (as "replacement sheets") must be submitted.
   (a) ☐ including changes required by the Notice of Draftsperson's Patent Drawing Review (PTO-948) attached
      1) ☐ hereto or 2) ☐ to Paper No./Mail Date _____.
   (b) ☐ including changes required by the attached Examiner's Amendment / Comment or in the Office action of Paper No./Mail Date _____.

   Identifying indicia such as the application number (see 37 CFR 1.84(c)) should be written on the drawings in the front (not the back) of each sheet. Replacement sheet(s) should be labeled as such in the header according to 37 CFR 1.121(d).

6. ☐ DEPOSIT OF and/or INFORMATION about the deposit of BIOLOGICAL MATERIAL must be submitted. Note the attached Examiner's comment regarding REQUIREMENT FOR THE DEPOSIT OF BIOLOGICAL MATERIAL.

Attachment(s)

1. ☐ Notice of References Cited (PTO-892)
2. ☐ Notice of Draftperson's Patent Drawing Review (PTO-948)
3. ☒ Information Disclosure Statements (PTO/SB/08), Paper No./Mail Date _____
4. ☐ Examiner's Comment Regarding Requirement for Deposit of Biological Material
5. ☐ Notice of Informal Patent Application
6. ☐ Interview Summary (PTO-413), Paper No./Mail Date _____ .
7. ☐ Examiner's Amendment/Comment
8. ☐ Examiner's Statement of Reasons for Allowance
9. ☐ Other _____.

/Brian L Albertalli/
Primary Examiner, Art Unit 2626

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	A	6,622,119	09/16/2003	Ramaswamy *et al.*	704	9	
	B	7,415,414	08/19/2008	Azara *et al.*	704	270	

FOREIGN PATENT DOCUMENTS

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

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

EXAMINER	/Brian Albertalli/	DATE CONSIDERED	02/08/2011

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

402593666v1

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.A./

INFORMATION DISCLOSURE CITATION IN AN APPLICATION	Docket Number	Application Number
	090611-0382843	12/617,506
	Applicant(s)	
	Philippe DI CRISTO *et al.*	
	Filing Date	Group Art Unit
Form PTO-1449	November 12, 2009	2626

U.S. PATENT DOCUMENTS

*EXAMINER INITIAL	REF	DOCUMENT NO.	DATE	NAME	CLASS	SUBCLASS	FILING DATE IF APPROPRIATE
	A	2008/0140385	06/12/2008	Mahajan *et al.*	704	9	
	B	2009/0216540	08/27/2009	Tessel *et al.*	704	275	

FOREIGN PATENT DOCUMENTS

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

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

EXAMINER	DATE CONSIDERED
/Brian Albertalli/	02/08/2011

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

402593676v1

ALL REFERENCES CONSIDERED EXCEPT WHERE LINED THROUGH. /B.A./

UNITED STATES PATENT AND TRADEMARK OFFICE

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

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
12/617,506	11/12/2009	Philippe Di Cristo	090611-0382843	4703

909 7590 02/24/2011

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

EXAMINER	
ALBERTALLI, BRIAN LOUIS	
ART UNIT	**PAPER NUMBER**
2626	

NOTIFICATION DATE	DELIVERY MODE
02/24/2011	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

docket_ip@pillsburylaw.com
margaret.drosos@pillsburylaw.com

Response to Rule 312 Communication	Application No.	Applicant(s)
	12/617,506	DI CRISTO ET AL.
	Examiner	Art Unit
	BRIAN L. ALBERTALLI	2626

-- The MAILING DATE of this communication appears on the cover sheet with the correspondence address –

1. ☒ The amendment filed on *02 February 2011* under 37 CFR 1.312 has been considered, and has been:

   a) ☐ entered.

   b) ☒ entered as directed to matters of form not affecting the scope of the invention.

   c) ☐ disapproved because the amendment was filed after the payment of the issue fee.
   Any amendment filed after the date the issue fee is paid must be accompanied by a petition under 37 CFR 1.313(c)(1) and the required fee to withdraw the application from issue.

   d) ☐ disapproved. See explanation below.

   e) ☐ entered in part. See explanation below.

/Brian L Albertalli/
Primary Examiner, Art Unit 2626

PART B - FEE(S) TRANSMITTAL

Complete and send this form, together with applicable fee(s), to: Mail **Mail Stop ISSUE FEE**
Commissioner for Patents
P.O. Box 1450
Alexandria, Virginia 22313-1450
or Fax (571)-273-2885

INSTRUCTIONS: This form should be used for transmitting the ISSUE FEE and PUBLICATION FEE (if required). Blocks 1 through 5 should be completed where appropriate. All further correspondence including the Patent, advance orders and notification of maintenance fees will be mailed to the current correspondence address as indicated unless corrected below or directed otherwise in Block 1, by (a) specifying a new correspondence address; and/or (b) indicating a separate "FEE ADDRESS" for maintenance fee notifications.

CURRENT CORRESPONDENCE ADDRESS (Note: Use Block 1 for any change of address)

909 7590 12/20/2010

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

Note: A certificate of mailing can only be used for domestic mailings of the Fee(s) Transmittal. This certificate cannot be used for any other accompanying papers. Each additional paper, such as an assignment or formal drawing, must have its own certificate of mailing or transmission.

Certificate of Mailing or Transmission

I hereby certify that this Fee(s) Transmittal is being deposited with the United States Postal Service with sufficient postage for first class mail in an envelope addressed to the Mail Stop ISSUE FEE address above, or being facsimile transmitted to the USPTO (571) 273-2885, on the date indicated below.

(Depositor's name)

(Signature)

(Date)

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
12/617,506	11/12/2009	Philippe Di Cristo	090611-0382843	4703

TITLE OF INVENTION: SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE

APPLN. TYPE	SMALL ENTITY	ISSUE FEE DUE	PUBLICATION FEE DUE	PREV. PAID ISSUE FEE	TOTAL FEE(S) DUE	DATE DUE
nonprovisional	YES	$755	$300	$0	$1055	03/21/2011

EXAMINER	ART UNIT	CLASS-SUBCLASS
ALBERTALLI, BRIAN LOUIS	2626	704-270100

1. Change of correspondence address or indication of "Fee Address" (37 CFR 1.363).

☐ Change of correspondence address (or Change of Correspondence Address form PTO/SB/122) attached.

☐ "Fee Address" indication (or "Fee Address" Indication form PTO/SB/47; Rev 03-02 or more recent) attached. **Use of a Customer Number is required.**

2. For printing on the patent front page, list

(1) the names of up to 3 registered patent attorneys or agents OR, alternatively,

(2) the name of a single firm (having as a member a registered attorney or agent) and the names of up to 2 registered patent attorneys or agents. If no name is listed, no name will be printed.

1 Pillsbury Winthrop Shaw Pittman LLP

2

3

3. ASSIGNEE NAME AND RESIDENCE DATA TO BE PRINTED ON THE PATENT (print or type)

PLEASE NOTE: Unless an assignee is identified below, no assignee data will appear on the patent. If an assignee is identified below, the document has been filed for recordation as set forth in 37 CFR 3.11. Completion of this form is NOT a substitute for filing an assignment.

(A) NAME OF ASSIGNEE (B) RESIDENCE: (CITY and STATE OR COUNTRY)

VoiceBox Technologies, Inc. Bellevue, Washington

Please check the appropriate assignee category or categories (will not be printed on the patent) : ☐ Individual ☑ Corporation or other private group entity ☐ Government

4a. The following fee(s) are submitted:

☑ Issue Fee

☑ Publication Fee (No small entity discount permitted)

☐ Advance Order - # of Copies _____

4b. Payment of Fee(s): **(Please first reapply any previously paid issue fee shown above)**

☐ A check is enclosed.

☐ Payment by credit card. Form PTO-2038 is attached.

☑ The Director is hereby authorized to charge the required fee(s), any deficiency, or credit any overpayment, to Deposit Account Number 03-3975 (enclose an extra copy of this form).

5. **Change in Entity Status** (from status indicated above)

☐ a. Applicant claims SMALL ENTITY status. See 37 CFR 1.27.

☐ b. Applicant is no longer claiming SMALL ENTITY status. See 37 CFR 1.27(g)(2).

NOTE: The Issue Fee and Publication Fee (if required) will not be accepted from anyone other than the applicant; a registered attorney or agent; or the assignee or other party in interest as shown by the records of the United States Patent and Trademark Office.

Authorized Signature _____ Date February 24, 2011

Typed or printed name Syed Jafar Ali Registration No. 58,780

This collection of information is required by 37 CFR 1.311. The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, Virginia 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, Virginia 22313-1450.

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

Electronic Patent Application Fee Transmittal

Application Number:	12617506
Filing Date:	12-Nov-2009
Title of Invention:	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE
First Named Inventor/Applicant Name:	Philippe Di Cristo
Filer:	Syed Jafar Ali/Miriam E. Nelsen
Attorney Docket Number:	090611-0382843

Filed as Small Entity

Utility under 35 USC 111(a) Filing Fees

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Basic Filing:				
Pages:				
Claims:				
Miscellaneous-Filing:				
Petition:				
Patent-Appeals-and-Interference:				
Post-Allowance-and-Post-Issuance:				
Utility Appl issue fee	2501	1	755	755
Publ. Fee- early, voluntary, or normal	1504	1	300	300

Description	Fee Code	Quantity	Amount	Sub-Total in USD($)
Extension-of-Time:				
Miscellaneous:				
			Total in USD ($)	**1055**

Electronic Acknowledgement Receipt

EFS ID:	9524298
Application Number:	12617506
International Application Number:	
Confirmation Number:	4703
Title of Invention:	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE
First Named Inventor/Applicant Name:	Philippe Di Cristo
Customer Number:	00909
Filer:	Syed Jafar Ali/Miriam E. Nelsen
Filer Authorized By:	Syed Jafar Ali
Attorney Docket Number:	090611-0382843
Receipt Date:	24-FEB-2011
Filing Date:	12-NOV-2009
Time Stamp:	14:34:07
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	yes
Payment Type	Deposit Account
Payment was successfully received in RAM	$1055
RAM confirmation Number	848
Deposit Account	033975
Authorized User	

The Director of the USPTO is hereby authorized to charge indicated fees and credit any overpayment as follows:

Charge any Additional Fees required under 37 C.F.R. Section 1.21 (Miscellaneous fees and charges)

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Issue Fee Payment (PTO-85B)	090611-0382843_Issue_Pub_Transmittal.pdf	97733 05b26630aa5d785655b23af5ebe76ac2f36eea5e	no	1
Warnings:					
Information:					
2	Fee Worksheet (PTO-875)	fee-info.pdf	31711 36e0f13587315ce0342d304b1363b669d568a133	no	2
Warnings:					
Information:					
		Total Files Size (in bytes):	129444		

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

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

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

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

UNITED STATES PATENT AND TRADEMARK OFFICE

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

APPLICATION NO.	ISSUE DATE	PATENT NO.	ATTORNEY DOCKET NO.	CONFIRMATION NO.
12/617,506	03/29/2011	7917367	090611-0382843	4703

909 7590 03/09/2011

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

ISSUE NOTIFICATION

The projected patent number and issue date are specified above.

Determination of Patent Term Adjustment under 35 U.S.C. 154 (b)
(application filed on or after May 29, 2000)

The Patent Term Adjustment is 0 day(s). Any patent to issue from the above-identified application will include an indication of the adjustment on the front page.

If a Continued Prosecution Application (CPA) was filed in the above-identified application, the filing date that determines Patent Term Adjustment is the filing date of the most recent CPA.

Applicant will be able to obtain more detailed information by accessing the Patent Application Information Retrieval (PAIR) WEB site (http://pair.uspto.gov).

Any questions regarding the Patent Term Extension or Adjustment determination should be directed to the Office of Patent Legal Administration at (571)-272-7702. Questions relating to issue and publication fee payments should be directed to the Application Assistance Unit (AAU) of the Office of Data Management (ODM) at (571)-272-4200.

APPLICANT(s) (Please see PAIR WEB site http://pair.uspto.gov for additional applicants):

Philippe Di Cristo, Bellevue, WA;
Min Ke, Bothell, WA;
Robert A. Kennewick, Seattle, WA;
Lynn Elise Armstrong, Woodinville, WA;

PATENT
IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Current Date: September 26th, 2018

Application No.: 12/617506
Filed: 12.11.2009

Patent No.: 7917367
Issued: 29.03.2011

CHANGE OF ENTITY STATUS PURSUANT TO 37 C. F. R. § 1.27 (g)(2)

Commissioner for Patents
Mail Stop M Correspondence
P.O. Box 1450
Alexandria, VA 22313-1450

By Fax: 001-571-273-6500

Sir:

This communication hereby notifies the United States Patent and Trademark Office that **small entity status is no longer applicable** for the above-identified patent.

Respectfully submitted,

Signature
Printed Name: Scott Ouellette
TITLE: Attorney of Record Reg #38,573

CPA GLOBAL®

Liberation House,
Castle Street, St Helier,
Jersey JE1 1BL,
Channel Islands

t: +44 (0) 1534 888 711
e: enquiries@cpaglobal.com
f: +44 (0) 1534 888 747
cpaglobal.com

CPA Global Limited
Registered as a Company under the Companies (Jersey) Law 1991, number 93743. Registered office: Liberation House, Castle Street, St Helier, Jersey JE1 1BL, Channel Islands

FAX TO:

Pages 3

Commissioner of Patents and Trademarks
Attn. Maintenance Fee Branch
Office of Finance – Maintenance Fee Division
2051 Jamieson Ave – Suite 300
Alexandria, VA 22314

Fax no.: 001 571 273-6500

9 October 2018

URGENT

Dear Sirs,

Please see the attached entity status form to be amended to **SMALL** Entity status.

Please can you amend your records accordingly.

Many thanks for your assistance.

Kind regards,

Karen Greenham

Data Verification Specialist

CPA GLOBAL

Direct Dial: + 44(0)1534 811383
E-mail: kgreenham@cpaglobal.com

Certificate of Facsimile Transmission

I hereby certify that this document is being facsimile transmitted to the USPTO at 001 571-273-6500 on the date shown below.

Date
9-Oct-18

Signed

UNITED STATES PATENT AND TRADEMARK OFFICE

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

APPLICATION NO.	FILING DATE	FIRST NAMED INVENTOR	ATTORNEY DOCKET NO.	CONFIRMATION NO.
12/617,506	11/12/2009	Philippe Di Cristo	090611-0382843	4703

909 7590 06/13/2019

Pillsbury Winthrop Shaw Pittman, LLP
PO Box 10500
McLean, VA 22102

EXAMINER	
ALBERTALLI, BRIAN LOUIS	

ART UNIT	PAPER NUMBER
2626	

NOTIFICATION DATE	DELIVERY MODE
06/13/2019	ELECTRONIC

Please find below and/or attached an Office communication concerning this application or proceeding.

The time period for reply, if any, is set in the attached communication.

Notice of the Office communication was sent electronically on above-indicated "Notification Date" to the following e-mail address(es):

docket_ip@pillsburylaw.com

UNITED STATES PATENT AND TRADEMARK OFFICE

Commissioner for Patents
United States Patent and Trademark Office
P.O. Box 1450
Alexandria, VA 22313-1450
www.uspto.gov

In re Application of	:	
Di Cristo et al.	:	
Application No. 12/617,506	:	
Filed: November 12, 2009	:	NOTICE
For SYSTEMS AND METHODS FOR	:	
RESPONDING TO NATURAL	:	
LANGUAGE SPEECH UTTERANCE	:	

This is in response to the communication filed October 09, 2018, under 37 CFR 1.27(g)(2) seeking loss of entitlement to small entity status .

The statement claiming small entity status is no longer applicable has been made of record and large entity status has been accorded.

This patent is no longer entitled to small entity status. Accordingly, all future fees paid in this application must be paid at the large entity rate.

/JOANNE L BURKE/
Lead Paralegal Specialist, OPET

PATENT ASSIGNMENT COVER SHEET

Electronic Version v1.1
Stylesheet Version v1.2

EPAS ID: PAT6700444

SUBMISSION TYPE:	NEW ASSIGNMENT
NATURE OF CONVEYANCE:	ASSIGNMENT

CONVEYING PARTY DATA

Name	Execution Date
NUANCE COMMUNICATIONS, INC.	01/01/2021

RECEIVING PARTY DATA

Name:	VOICE INVENTIONS, LLC
Street Address:	2 EXECUTIVE DRIVE
Internal Address:	STE. 270
City:	FORT LEE
State/Country:	NEW JERSEY
Postal Code:	07024

PROPERTY NUMBERS Total: 29

Property Type	Number
Patent Number:	7398209
Patent Number:	7502738
Patent Number:	7809570
Patent Number:	8015006
Patent Number:	8112275
Patent Number:	8140327
Patent Number:	8155962
Patent Number:	8731929
Patent Number:	9734825
Patent Number:	7693720
Patent Number:	9031845
Patent Number:	7640160
Patent Number:	7917367
Patent Number:	8326634
Patent Number:	8849670
Patent Number:	9263039
Patent Number:	7620549
Patent Number:	8332224
Patent Number:	8620659

Property Type	Number
Patent Number:	9626959
Patent Number:	7949529
Patent Number:	8195468
Patent Number:	8447607
Patent Number:	8849652
Patent Number:	9495957
Patent Number:	7634409
Patent Number:	7983917
Patent Number:	8069046
Patent Number:	8150694

CORRESPONDENCE DATA

Fax Number:

Correspondence will be sent to the e-mail address first; if that is unsuccessful, it will be sent using a fax number, if provided; if that is unsuccessful, it will be sent via US Mail.

Email: eddie.weiss@voiceinventions.com
Correspondent Name: EDWARD WEISS
Address Line 1: 2 EXECUTIVE DRIVE
Address Line 2: STE. 270
Address Line 4: FORT LEE, NEW JERSEY 07024

NAME OF SUBMITTER:	EDWARD WEISS
SIGNATURE:	/Edward Weiss/
DATE SIGNED:	05/10/2021
	This document serves as an Oath/Declaration (37 CFR 1.63).

Total Attachments: 4
source=2020-12-24 - Assignment Nuance to Voice Inventions#page1.tif
source=2020-12-24 - Assignment Nuance to Voice Inventions#page2.tif
source=2020-12-24 - Assignment Nuance to Voice Inventions#page3.tif
source=2020-12-24 - Assignment Nuance to Voice Inventions#page4.tif

PATENT ASSIGNMENT AGREEMENT

This PATENT ASSIGNMENT AGREEMENT (the "Assignment Agreement") between Nuance Communications, Inc., a Delaware corporation, with an address at One Wayside Road, Burlington, MA 01803 ("Assignor"), and Voice Inventions, LLC, a Texas Limited Liability Company, with an address at 2 Executive Drive, Suite 270, Fort Lee, NJ 07024 ("Assignee").

RECITALS

WHEREAS the parties have entered into a certain Patent Purchase and Licensing Agreement (the "Purchase Agreement"), with an effective date of January 1, 2021;

WHEREAS Assignor possesses all legal rights, title, and interests in and to the patents listed in Exhibit A (the "Assigned Patent Portfolio") to this Assignment Agreement;

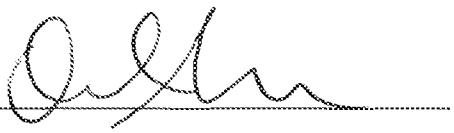
WHEREAS the parties desire to effectuate the assignment and transfer of the Assigned Patent Portfolio to Assignee such that Assignee shall acquire, and thus hereby acquires, all of Assignor's worldwide rights, title, and interests in and to the Assigned Patent Portfolio; and

WHEREAS the parties further desire that the assignment of the Assigned Patent Portfolio be undertaken in a such a manner as to allow Assignee to make such assignment a matter of record in the appropriate domestic and international patent offices;

NOW THEREFORE, for good and valuable consideration, the receipt of which by Assignor is hereby acknowledged, Assignor hereby sells, assigns, and transfers to Assignee, and Assignee hereby purchases and accepts from Assignor, all of Assignor's worldwide rights, title and interest, in and to the Assigned Patent Portfolio, including any and all of Assignor's rights to sue for past damages, subject to the terms and conditions of the Purchase Agreement.

IN WITNESS WHEREOF, intending to be legally bound, Assignor has caused this Assignment Agreement to by executed by its duly authorized representative and made effective, as of January 1, 2021.

Nuance Communications, Inc.
Assignor

By 

Name: David Greenbaum
Title: VP, Litigation & IP

Exhibit A

Assigned Patent Portfolio

	COUNTRY	TITLE	APPLICATION NUMBER	DATE FILED	PATENT NUMBER
1	US	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE	10/452,147	Jun 3, 2003	7,398,209
2	US	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE	11/747,547	May 11, 2007	7,502,738
3	US	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE	12/168,554	Jul 7, 2008	7,809,570
4	US	SYSTEMS AND METHODS FOR PROCESSING NATURAL LANGUAGE SPEECH UTTERANCES WITH CONTEXT-SPECIFIC DOMAIN AGENTS	12/130,397	May 30, 2008	8,015,006
5	US	SYSTEM AND METHOD FOR USER-SPECIFIC SPEECH RECOGNITION	12/765,733	Apr 22, 2010	8,112,275
6	US	SYSTEM AND METHOD FOR FILTERING AND ELIMINATING NOISE FROM NATURAL LANGUAGE UTTERANCES TO IMPROVE SPEECH RECOGNITION AND PARSING	12/765,753	Apr 22, 2010	8,140,327
7	US	METHOD AND SYSTEM FOR ASYNCHRONOUSLY PROCESSING NATURAL LANGUAGE UTTERANCES	12/838,982	Jul 19, 2010	8,155,962
8	US	AGENT ARCHITECTURE FOR DETERMINING MEANINGS OF NATURAL LANGUAGE UTTERANCES	12/365,516	Feb 4, 2009	8,731,929
9	US	SYSTEM AND METHOD FOR PROCESSING NATURAL LANGUAGE UTTERANCES	14/278,627	May 15, 2014	9,734,825
10	US	MOBILE SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCES	10/618,633	Jul 15, 2003	7,693,720
11	US	MOBILE SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE	12/705,311	Feb 12, 2010	9,031,845

12	US	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE	11/197,504	Aug 5, 2005	7,640,160
13	US	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE	12/617,506	Nov 12, 2009	7,917,367
14	US	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE	13/019,834	Feb 2, 2011	8,326,634
15	US	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE	13/690,895	Nov 30, 2012	8,849,670
16	US	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE	14/500,723	Sep 29, 2014	9,263,039

17	US	SYSTEM AND METHOD OF SUPPORTING ADAPTIVE MISRECOGNITION IN CONVERSATIONAL SPEECH	11/200,164	Aug 10, 2005	7,620,549
18	US	SYSTEM AND METHOD OF SUPPORTING ADAPTIVE MISRECOGNITION IN CONVERSATIONAL SPEECH	12/571,795	Oct 1, 2009	8,332,224
19	US	SYSTEM AND METHOD OF SUPPORTING ADAPTIVE MISRECOGNITION IN CONVERSATIONAL SPEECH	13/022,370	Feb 7, 2011	8,620,659
20	US	SYSTEM AND METHOD OF SUPPORTING ADAPTIVE MISRECOGNITION IN CONVERSATIONAL SPEECH	14/143,903	Dec 30, 2013	9,626,959

21	US	MOBILE SYSTEMS AND METHODS OF SUPPORTING NATURAL LANGUAGE HUMAN-MACHINE INTERACTIONS	11/212,693	Aug 29, 2005	7,949,529
22	US	MOBILE SYSTEMS AND METHODS OF SUPPORTING NATURAL LANGUAGE HUMAN-MACHINE INTERACTIONS	13/084,197	Apr 11, 2011	8,195,468
23	US	MOBILE SYSTEMS AND METHODS OF SUPPORTING NATURAL LANGUAGE HUMAN-MACHINE INTERACTIONS	13/488,299	Jun 4, 2012	8,447,607
24	US	MOBILE SYSTEMS AND METHODS OF SUPPORTING NATURAL LANGUAGE HUMAN-MACHINE INTERACTIONS	13/898,045	May 20, 2013	8,849,652

25	US	MOBILE SYSTEMS AND METHODS OF SUPPORTING NATURAL LANGUAGE HUMAN-MACHINE INTERACTIONS	14/467,641	Aug 25, 2014	9,495,957

26	US	DYNAMIC SPEECH SHARPENING	11/513,269	Aug 31, 2006	7,634,409
27	US	DYNAMIC SPEECH SHARPENING	12/608,544	Oct 29, 2009	7,983,917
28	US	DYNAMIC SPEECH SHARPENING	12/608,572	Oct 29, 2009	8,069,046
29	US	SYSTEM AND METHOD FOR PROVIDING AN ACOUSTIC GRAMMAR TO DYNAMICALLY SHARPEN SPEECH INTERPRETATION	13/150,977	Jun 1, 2011	8,150,694

30	CN	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE	2006800362006	Aug 4, 2006	ZL200680036200.6
31	CN	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE	2015102059396	Aug 4, 2006	ZL201510205939.6
32	CN	MOBILE SYSTEMS AND METHODS OF SUPPORTING NATURAL LANGUAGE HUMAN-MACHINE INTERACTIONS	200680038560	Aug 29, 2006	200680038560

PTO/AIA/81A (02-15)
Approved for use through 03/31/2021. OMB 0651-0035
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995 no persons are required to respond to a collection of information unless it displays a valid OMB control number

PATENT - POWER OF ATTORNEY OR REVOCATION OF POWER OF ATTORNEY WITH A NEW POWER OF ATTORNEY AND CHANGE OF CORRESPONDENCE ADDRESS

Patent Number	7917367
Issue Date	03/29/11
First Named Inventor	Di Cristo, Philippe
Title	Systems and methods for responding to natural language speech utterance
Attorney Docket No.	9032-0009

I hereby revoke all previous powers of attorney given in the above-identified patent.

☐ A Power of Attorney is submitted herewith.

OR

☒ I hereby appoint Practitioner(s) associated with the Customer Number identified in the box at right as my/our attorney(s) or agent(s) with respect to the patent identified above, and to transact all business in the United States Patent and Trademark Office connected therewith: 174889

OR

☐ I hereby appoint Practitioner(s) named below as my/our attorney(s) or agent(s) with respect to the patent identified above, and to transact all business in the United States Patent and Trademark Office connected therewith:

Practitioner(s) Name	Registration Number

Please recognize or change the correspondence address for the above-identified patent to:

☒ The address associated with the above-identified Customer Number.

OR

☐ The address associated with the Customer Number identified in the box at right:

OR

☐ Firm or Individual Name					
Address					
City		State		Zip	
Country					
Telephone		Email			

I am the:

☐ Applicant.

OR

☒ Patent owner.

Statement under 37 CFR 3.73(c) (Form PTO/AIA/96) submitted herewith or filed on 02/28/2023.

SIGNATURE of Applicant or Patent Owner

Signature	/Jeff Ronaldi/	Date	02/27/2023
Name	Jeff Ronaldi	Telephone	(571) 334-1112
Title and Company	Manager, Dialect, LLC		

NOTE: Signatures of all the applicants or patent owners of the entire interest or their representative(s) are required. If more than one signature is required, submit multiple forms, check the box below, and identify the total number of forms submitted in the blank below.

☐ A total of _____ forms are submitted.

This collection of information is required by 37 CFR 1.31, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public, which is to update (and by the USPTO to process) the file of a patent or reexamination proceeding. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

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

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

PTO/AIA/96 (08-12)
Approved for use through 01/31/2013. OMB 0651-0031
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

STATEMENT UNDER 37 CFR 3.73(c)

Applicant/Patent Owner: Dialect, LLC

Application No./Patent No.: 7917367 Filed/Issue Date: 03/29/11

Titled: Systems and methods for responding to natural language speech utterance

Dialect, LLC, a limited liability company

(Name of Assignee) (Type of Assignee, e.g., corporation, partnership, university, government agency, etc.)

states that, for the patent application/patent identified above, it is (choose **one** of options 1, 2, 3 or 4 below):

1. ☑ The assignee of the entire right, title, and interest.

2. ☐ An assignee of less than the entire right, title, and interest (check applicable box):

   ☐ The extent (by percentage) of its ownership interest is _____%. Additional Statement(s) by the owners holding the balance of the interest must be submitted to account for 100% of the ownership interest.

   ☐ There are unspecified percentages of ownership. The other parties, including inventors, who together own the entire right, title and interest are:

   Additional Statement(s) by the owner(s) holding the balance of the interest must be submitted to account for the entire right, title, and interest.

3. ☐ The assignee of an undivided interest in the entirety (a complete assignment from one of the joint inventors was made). The other parties, including inventors, who together own the entire right, title, and interest are:

   Additional Statement(s) by the owner(s) holding the balance of the interest must be submitted to account for the entire right, title, and interest.

4. ☐ The recipient, via a court proceeding or the like (*e.g.*, bankruptcy, probate), of an undivided interest in the entirety (a complete transfer of ownership interest was made). The certified document(s) showing the transfer is attached.

The interest identified in option 1, 2 or 3 above (not option 4) is evidenced by either (choose **one** of options A or B below):

A. ☐ An assignment from the inventor(s) of the patent application/patent identified above. The assignment was recorded in the United States Patent and Trademark Office at Reel _____, Frame _____, or for which a copy thereof is attached.

B. ☑ A chain of title from the inventor(s), of the patent application/patent identified above, to the current assignee as follows:

1. From: Di Cristo, Philippe et al. To: VoiceBox Technologies, Inc.

   The document was recorded in the United States Patent and Trademark Office at Reel 025674, Frame 0133, or for which a copy thereof is attached.

2. From: VoiceBox Technologies, Inc. To: VoiceBox Technologies Corporation

   The document was recorded in the United States Patent and Trademark Office at Reel 031715, Frame 0696, or for which a copy thereof is attached.

This collection of information is required by 37 CFR 3.73(b). The information is required to obtain or retain a benefit by the public which is to file (and by the USPTO to process) an application. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.11 and 1.14. This collection is estimated to take 12 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

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

Under the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless it displays a valid OMB control number.

STATEMENT UNDER 37 CFR 3.73(c)

3. From: VoiceBox Technologies Corporation To: Nuance Communications, Inc.

The document was recorded in the United States Patent and Trademark Office at
Reel 034703, Frame 0429, or for which a copy thereof is attached.

4. From: Nuance Communications, Inc. To: Voice Inventions, LLC

The document was recorded in the United States Patent and Trademark Office at
Reel 056185, Frame 0742, or for which a copy thereof is attached.

5. From: Voice Inventions, LLC To: Dialect, LLC

The document was recorded in the United States Patent and Trademark Office at
Reel 060820, Frame 0857, or for which a copy thereof is attached.

6. From: _____ To: _____

The document was recorded in the United States Patent and Trademark Office at
Reel _____, Frame _____, or for which a copy thereof is attached.

☐ Additional documents in the chain of title are listed on a supplemental sheet(s).

☑ As required by 37 CFR 3.73(c)(1)(i), the documentary evidence of the chain of title from the original owner to the assignee was, or concurrently is being, submitted for recordation pursuant to 37 CFR 3.11.

[NOTE: A separate copy (i.e., a true copy of the original assignment document(s)) must be submitted to Assignment Division in accordance with 37 CFR Part 3, to record the assignment in the records of the USPTO. See MPEP 302.08]

The undersigned (whose title is supplied below) is authorized to act on behalf of the assignee.

/Jeff Ronaldi/	02/27/2023
Signature	Date
Jeff Ronaldi	Manager, Dialect, LLC
Printed or Typed Name	Title or Registration Number

[Page 2 of 2]

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

Electronic Acknowledgement Receipt

EFS ID:	47601961
Application Number:	12617506
International Application Number:	
Confirmation Number:	4703
Title of Invention:	SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE
First Named Inventor/Applicant Name:	Philippe Di Cristo
Customer Number:	909
Filer:	David Alexander Gerasimow
Filer Authorized By:	
Attorney Docket Number:	090611-0382843
Receipt Date:	28-FEB-2023
Filing Date:	12-NOV-2009
Time Stamp:	11:04:03
Application Type:	Utility under 35 USC 111(a)

Payment information:

Submitted with Payment	no

File Listing:

Document Number	Document Description	File Name	File Size(Bytes)/ Message Digest	Multi Part /.zip	Pages (if appl.)
1	Power of Attorney	7917367.pdf	207201 83af0e953de044bec392d9e5063b00da29439756	no	2

Warnings:

Information:					
2	Assignee showing of ownership per 37 CFR 3.73	7917367-aia0096.pdf	202152 e346216a92c9dc5cab2c06ef7a2d9411fce08785	no	3
Warnings:					
Information:					
		Total Files Size (in bytes):	409353		

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

New Applications Under 35 U.S.C. 111

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

National Stage of an International Application under 35 U.S.C. 371

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

New International Application Filed with the USPTO as a Receiving Office

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

PTO/AIA/81A (02-15)
Approved for use through 03/31/2021. OMB 0651-0035
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE
Under the Paperwork Reduction Act of 1995 no persons are required to respond to a collection of information unless it displays a valid OMB control number

PATENT - POWER OF ATTORNEY OR REVOCATION OF POWER OF ATTORNEY WITH A NEW POWER OF ATTORNEY AND CHANGE OF CORRESPONDENCE ADDRESS

Patent Number	7917367
Issue Date	03/29/11
First Named Inventor	Di Cristo, Philippe
Title	Systems and methods for responding to natural language speech utterance
Attorney Docket No.	9032-0009

I hereby revoke all previous powers of attorney given in the above-identified patent.

☐ A Power of Attorney is submitted herewith.

OR

☒ I hereby appoint Practitioner(s) associated with the Customer Number identified in the box at right as my/our attorney(s) or agent(s) with respect to the patent identified above, and to transact all business in the United States Patent and Trademark Office connected therewith: 174899

OR

☐ I hereby appoint Practitioner(s) named below as my/our attorney(s) or agent(s) with respect to the patent identified above, and to transact all business in the United States Patent and Trademark Office connected therewith:

Practitioner(s) Name	Registration Number

Please recognize or change the correspondence address for the above-identified patent to:

☒ The address associated with the above-identified Customer Number.

OR

☐ The address associated with the Customer Number identified in the box at right:

OR

☐ Firm or Individual Name					
Address					
City		State		Zip	
Country					
Telephone		Email			

I am the:

☐ Applicant.

OR

☒ Patent owner.

Statement under 37 CFR 3.73(c) (Form PTO/AIA/96) submitted herewith or filed on 02/28/2023.

SIGNATURE of Applicant or Patent Owner

Signature	/Jeff Ronaldi/	Date	11/06/2023
Name	Jeff Ronaldi	Telephone	(571) 334-1112
Title and Company	Manager, Dialect, LLC		

NOTE: Signatures of all the applicants or patent owners of the entire interest or their representative(s) are required. If more than one signature is required, submit multiple forms, check the box below, and identify the total number of forms submitted in the blank below.

☐ A total of _____ forms are submitted.

This collection of information is required by 37 CFR 1.31, 1.32, and 1.33. The information is required to obtain or retain a benefit by the public, which is to update (and by the USPTO to process) the file of a patent or reexamination proceeding. Confidentiality is governed by 35 U.S.C. 122 and 37 CFR 1.14. This collection is estimated to take 3 minutes to complete, including gathering, preparing, and submitting the completed application form to the USPTO. Time will vary depending upon the individual case. Any comments on the amount of time you require to complete this form and/or suggestions for reducing this burden, should be sent to the Chief Information Officer, U.S. Patent and Trademark Office, U.S. Department of Commerce, P.O. Box 1450, Alexandria, VA 22313-1450. DO NOT SEND FEES OR COMPLETED FORMS TO THIS ADDRESS. **SEND TO: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.**

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

Privacy Act Statement

The **Privacy Act of 1974 (P.L. 93-579)** requires that you be given certain information in connection with your submission of the attached form related to a patent application or patent. Accordingly, pursuant to the requirements of the Act, please be advised that: (1) the general authority for the collection of this information is 35 U.S.C. 2(b)(2); (2) furnishing of the information solicited is voluntary; and (3) the principal purpose for which the information is used by the U.S. Patent and Trademark Office is to process and/or examine your submission related to a patent application or patent. If you do not furnish the requested information, the U.S. Patent and Trademark Office may not be able to process and/or examine your submission, which may result in termination of proceedings or abandonment of the application or expiration of the patent.

The information provided by you in this form will be subject to the following routine uses:

1. The information on this form will be treated confidentially to the extent allowed under the Freedom of Information Act (5 U.S.C. 552) and the Privacy Act (5 U.S.C 552a). Records from this system of records may be disclosed to the Department of Justice to determine whether disclosure of these records is required by the Freedom of Information Act.
2. A record from this system of records may be disclosed, as a routine use, in the course of presenting evidence to a court, magistrate, or administrative tribunal, including disclosures to opposing counsel in the course of settlement negotiations.
3. A record in this system of records may be disclosed, as a routine use, to a Member of Congress submitting a request involving an individual, to whom the record pertains, when the individual has requested assistance from the Member with respect to the subject matter of the record.
4. A record in this system of records may be disclosed, as a routine use, to a contractor of the Agency having need for the information in order to perform a contract. Recipients of information shall be required to comply with the requirements of the Privacy Act of 1974, as amended, pursuant to 5 U.S.C. 552a(m).
5. A record related to an International Application filed under the Patent Cooperation Treaty in this system of records may be disclosed, as a routine use, to the International Bureau of the World Intellectual Property Organization, pursuant to the Patent Cooperation Treaty.
6. A record in this system of records may be disclosed, as a routine use, to another federal agency for purposes of National Security review (35 U.S.C. 181) and for review pursuant to the Atomic Energy Act (42 U.S.C. 218(c)).
7. A record from this system of records may be disclosed, as a routine use, to the Administrator, General Services, or his/her designee, during an inspection of records conducted by GSA as part of that agency's responsibility to recommend improvements in records management practices and programs, under authority of 44 U.S.C. 2904 and 2906. Such disclosure shall be made in accordance with the GSA regulations governing inspection of records for this purpose, and any other relevant (*i.e.*, GSA or Commerce) directive. Such disclosure shall not be used to make determinations about individuals.
8. A record from this system of records may be disclosed, as a routine use, to the public after either publication of the application pursuant to 35 U.S.C. 122(b) or issuance of a patent pursuant to 35 U.S.C. 151. Further, a record may be disclosed, subject to the limitations of 37 CFR 1.14, as a routine use, to the public if the record was filed in an application which became abandoned or in which the proceedings were terminated and which application is referenced by either a published application, an application open to public inspection or an issued patent.
9. A record from this system of records may be disclosed, as a routine use, to a Federal, State, or local law enforcement agency, if the USPTO becomes aware of a violation or potential violation of law or regulation.

UNITED STATES
PATENT AND TRADEMARK OFFICE

P.O. Box 1450
Alexandria, VA 22313 - 1450
www.uspto.gov

ELECTRONIC ACKNOWLEDGEMENT RECEIPT

APPLICATION #	RECEIPT DATE / TIME	ATTORNEY DOCKET #
12/617,506	**11/09/2023 12:30:14 PM Z ET**	**090611-0382843**

Title of Invention

SYSTEMS AND METHODS FOR RESPONDING TO NATURAL LANGUAGE SPEECH UTTERANCE

Application Information

APPLICATION TYPE	Utility - Nonprovisional Application under 35 USC 111(a)	PATENT #	7917367
CONFIRMATION #	4703	FILED BY	David Gerasimow
PATENT CENTER #	63206507	FILING DATE	11/12/2009
CUSTOMER #	909	FIRST NAMED INVENTOR	Philippe Di Cristo
CORRESPONDENCE ADDRESS	-	AUTHORIZED BY	-

Documents

TOTAL DOCUMENTS: 1

DOCUMENT	PAGES	DESCRIPTION	SIZE (KB)
7917367.pdf	2	Power of Attorney	207 KB

Digest

DOCUMENT	MESSAGE DIGEST(SHA-512)
7917367.pdf	FD790786427E8B5DF43AB18772967782A3F4200FBFD288E1F0 3CCC68D561558144221173B0DD204ED56D4F2C7FEE7E68A24 749E4A201B6E18E7BF377C37A1B79

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

described in MPEP 503.

New Applications Under 35 U.S.C. 111

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

National Stage of an International Application under 35 U.S.C. 371

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

New International Application Filed with the USPTO as a Receiving Office

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

UNITED STATES PATENT AND TRADEMARK OFFICE

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

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
12/617,506	11/12/2009	Philippe Di Cristo	9032-0009

CONFIRMATION NO. 4703

POA ACCEPTANCE LETTER

174899
The Law Offices of David A. Gerasimow, P.C.
211 W. Wacker Dr.
Ste. 1717
Chicago, IL 60606

OC000000064866832

Date Mailed: 11/22/2023

NOTICE OF ACCEPTANCE OF POWER OF ATTORNEY

This is in response to the Power of Attorney filed 11/09/2023.

The Power of Attorney in this application is accepted. Correspondence in this application will be mailed to the above address as provided by 37 CFR 1.33.

Questions about the contents of this notice and the requirements it sets forth should be directed to the Office of Data Management, Application Assistance Unit, at **(571) 272-4000** or **(571) 272-4200** or **1-888-786-0101**.

/lqchau/

UNITED STATES PATENT AND TRADEMARK OFFICE

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

APPLICATION NUMBER	FILING OR 371(C) DATE	FIRST NAMED APPLICANT	ATTY. DOCKET NO./TITLE
12/617,506	11/12/2009	Philippe Di Cristo	090611-0382843

CONFIRMATION NO. 4703

909
Pillsbury Winthrop Shaw Pittman, LLP
PO Box 10500
McLean, VA 22102

POWER OF ATTORNEY NOTICE

OC000000064866805

Date Mailed: 11/22/2023

NOTICE REGARDING CHANGE OF POWER OF ATTORNEY

This is in response to the Power of Attorney filed 11/09/2023.

- The Power of Attorney to you in this application has been revoked by the assignee who has intervened as provided by 37 CFR 3.71. Future correspondence will be mailed to the new address of record(37 CFR 1.33).

Questions about the contents of this notice and the requirements it sets forth should be directed to the Office of Data Management, Application Assistance Unit, at **(571) 272-4000** or **(571) 272-4200** or **1-888-786-0101**.

/lqchau/