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

(19) World Intellectual Property Organization
International Bureau

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

PCT

(10) International Publication Number
WO 2008/118195 A2

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

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

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

(25) Filing Language: English

(26) Publication Language: English

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

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

(72) Inventors; and

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

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

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

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

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

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

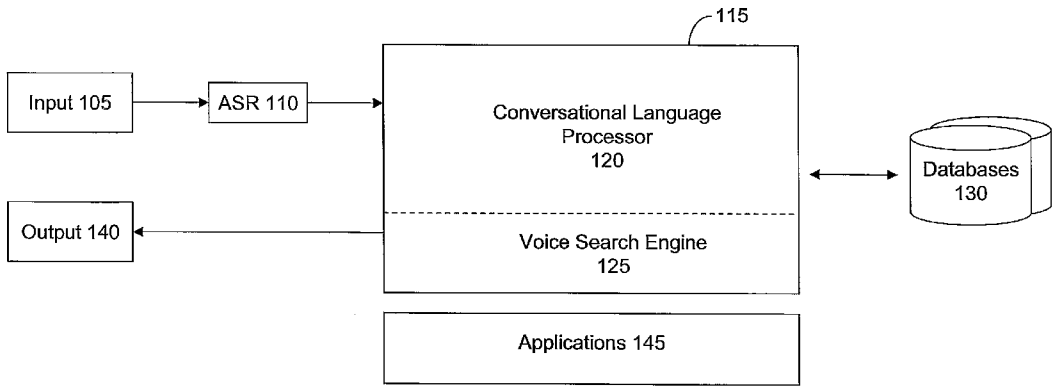

FIG. 1

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

WO 2008/118195 A2

SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

FIELD OF THE INVENTION

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

BACKGROUND OF THE INVENTION

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

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

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

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

SUMMARY OF THE INVENTION

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

convenient.

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

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

BRIEF DESCRIPTION OF THE DRAWINGS

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

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

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

DETAILED DESCRIPTION

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

those conditions.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

User	What's the weather like in Dallas?

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

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

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

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

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

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

CLAIMS

What is claimed is:

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

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

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

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

generating a response based on the determined intended meaning.

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

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

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

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

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

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

utterance.

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

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

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

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

12. The method of claim 1, further comprising:

learning that the determined intended meaning is incorrect;

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

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

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

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

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

instructions operable when executed to:

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

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

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

generate a response based on the determined intended meaning.

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

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

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

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

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

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

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

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

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

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

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

learn that the determined intended meaning is incorrect;

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

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

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

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

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

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

one or more processors collectively operable to:

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

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

generate a response based on the determined intended meaning.

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

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

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

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

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

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

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

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

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

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

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

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

learn that the determined intended meaning is incorrect;

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

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

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

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

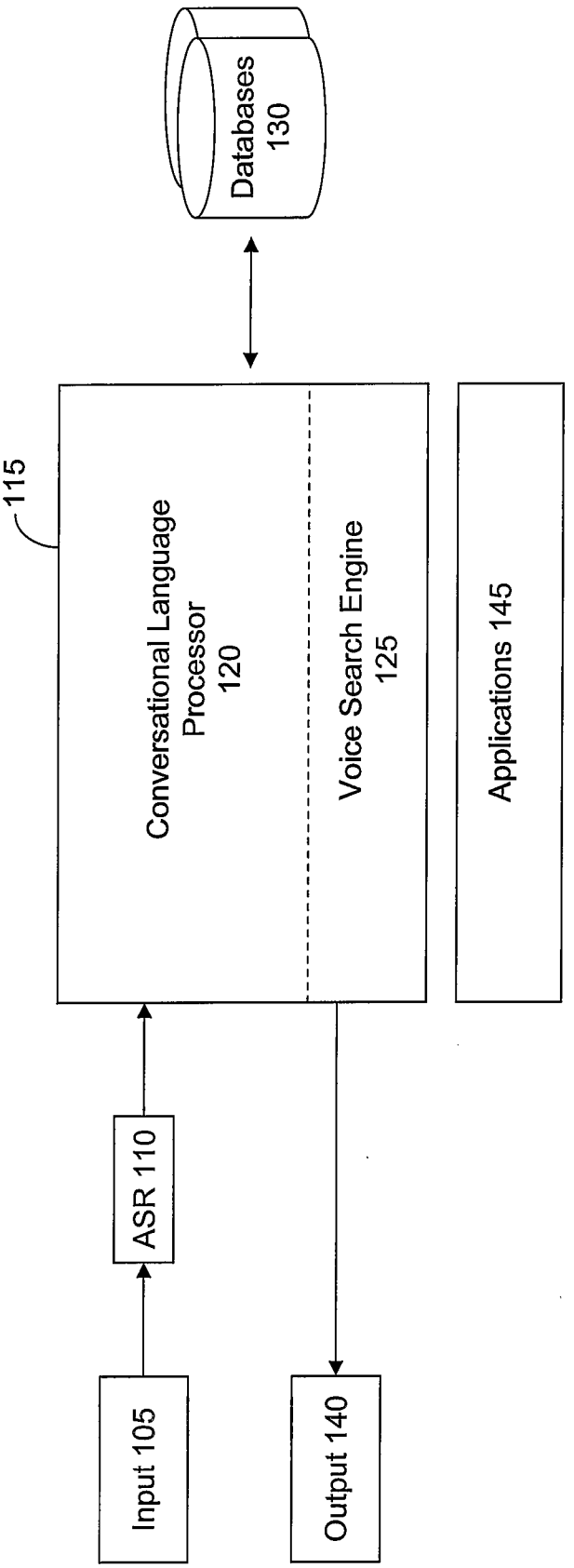
Databases
130
115
Conversational Language
Processor
120
Voice Search Engine
125
Applications 145
ASR 110
Input 105
Output 140

FIG. 1

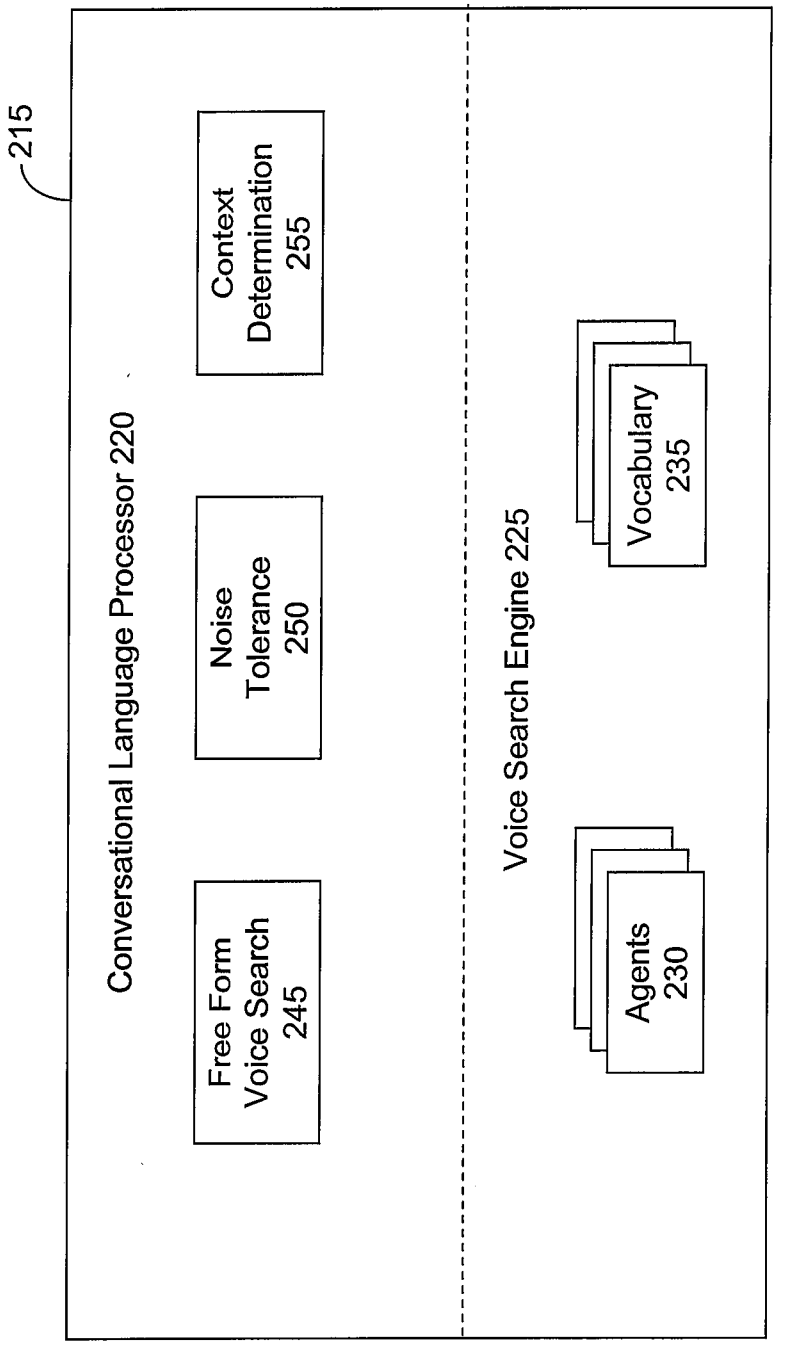
215
Conversational Language Processor 220
Free Form Voice Search 245
Noise Tolerance 250
Context Determination 255
Voice Search Engine 225
Agents 230
Vocabulary 235

FIG. 2

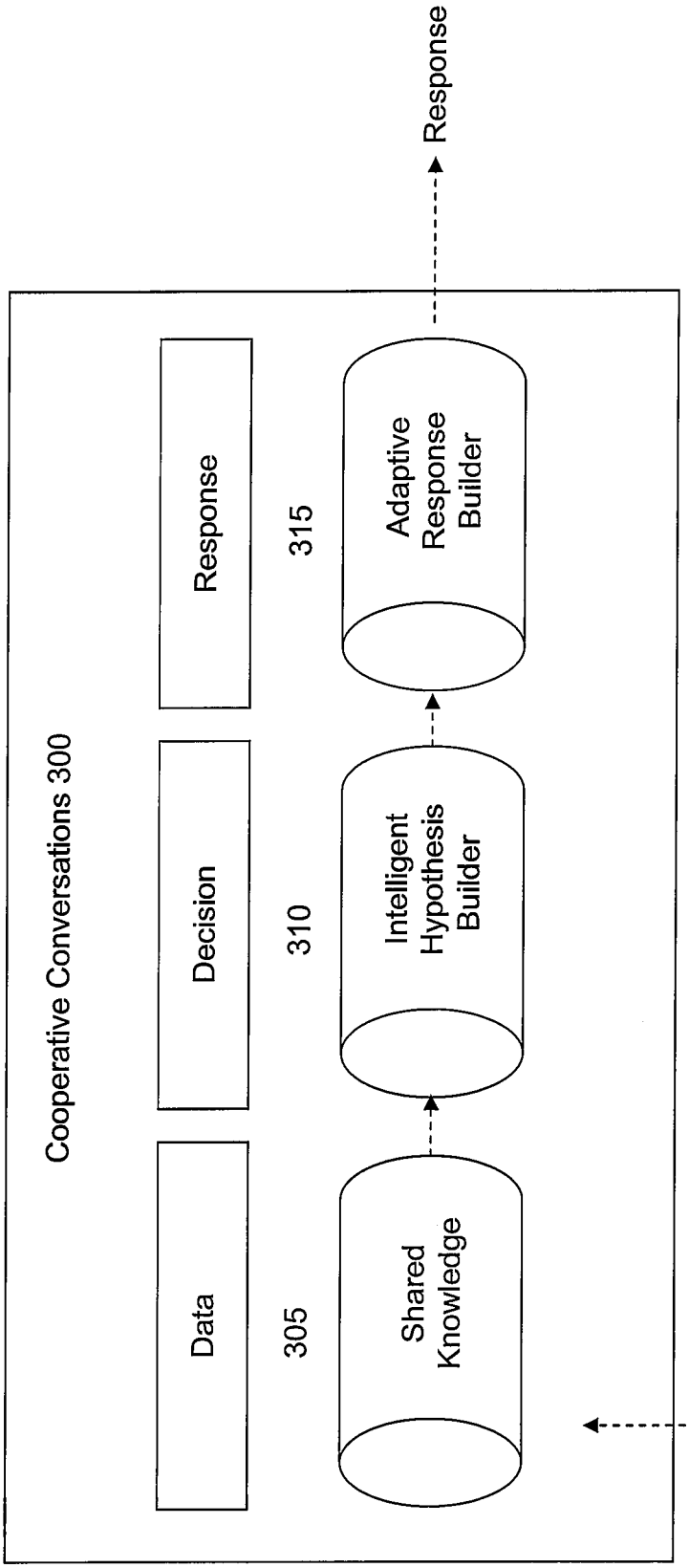
Response
Cooperative Conversations 300
Data
Decision
Response
305
310
315
Shared Knowledge
Intelligent Hypothesis Builder
Adaptive Response Builder
Input

FIG. 3