US008073681B2

(12) **United States Patent**
Baldwin et al.

(10) **Patent No.: US 8,073,681 B2**
(45) **Date of Patent: Dec. 6, 2011**

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

(75) Inventors: **Larry Baldwin**, Maple Valley, WA (US); **Tom Freeman**, Mercer Island, WA (US); **Michael Tjalve**, Bellevue, WA (US); **Blane Ebersold**, Seattle, WA (US); **Chris Weider**, Everett, WA (US)

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

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 1217 days.

(21) Appl. No.: **11/580,926**

(22) Filed: **Oct. 16, 2006**

(65) **Prior Publication Data**

US 2008/0091406 A1 Apr. 17, 2008

(51) **Int. Cl.**
G06F 17/27 (2006.01)
G10L 21/00 (2006.01)
G10L 11/00 (2006.01)

(52) **U.S. Cl.** **704/9**; 704/270; 704/275

(58) **Field of Classification Search** 704/270–275, 704/9

See application file for complete search history.

(56) **References Cited**

U.S. PATENT DOCUMENTS

4,430,669 A		2/1984	Cheung	358/122
5,027,406 A		6/1991	Roberts et al.	381/43
5,155,743 A		10/1992	Jacobs	375/28
5,274,560 A		12/1993	LaRue	364/444
5,377,350 A		12/1994	Skinner	395/600
5,386,556 A		1/1995	Hedin et al.	395/600
5,424,947 A		6/1995	Nagao et al.	364/419.08
5,471,318 A		11/1995	Ahuja et al.	358/400
5,475,733 A		12/1995	Eisdorfer et al.	379/52
5,488,652 A	*	1/1996	Bielby et al.	379/88.03
5,499,289 A		3/1996	Bruno et al.	379/220
5,500,920 A		3/1996	Kupiec	395/2.79
5,517,560 A		5/1996	Greenspan	379/114
5,533,108 A		7/1996	Harris et al.	379/201
5,537,436 A		7/1996	Bottoms et al.	375/222

(Continued)

FOREIGN PATENT DOCUMENTS

EP 1 320 043 A2 6/2003

(Continued)

OTHER PUBLICATIONS

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*, Jun. 2002, pp. 37-46.

(Continued)

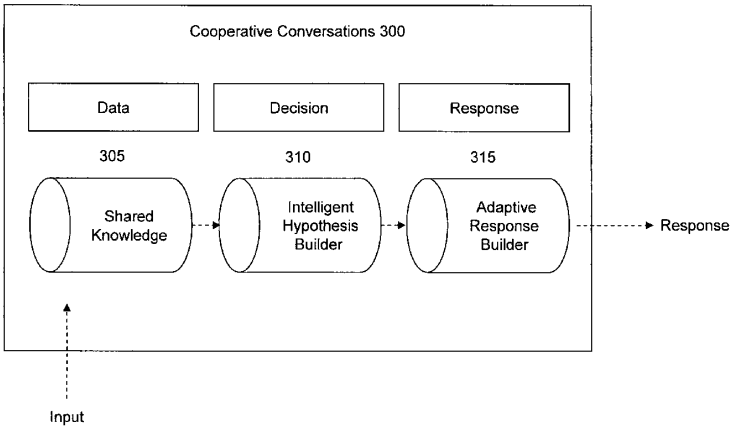
Primary Examiner — Eric Yen

(74) *Attorney, Agent, or Firm* — Pillsbury Winthrop Shaw Pittman LLP

(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.

42 Claims, 3 Drawing Sheets



U.S. PATENT DOCUMENTS

5,539,744	A	7/1996	Chu et al.	370/60
5,557,667	A	9/1996	Bruno et al.	379/201
5,563,937	A	10/1996	Bruno et al.	379/201
5,590,039	A	12/1996	Ikeda et al.	395/759
5,617,407	A	4/1997	Bareis	369/275.3
5,633,922	A	5/1997	August et al.	379/220
5,675,629	A	10/1997	Raffel et al.	379/58
5,696,965	A	12/1997	Dedrick	395/610
5,708,422	A	1/1998	Blonder et al.	340/825.34
5,721,938	A	2/1998	Stuckey	395/754
5,722,084	A	2/1998	Chakrin et al.	455/551
5,742,763	A	4/1998	Jones	395/200.3
5,748,841	A *	5/1998	Morin et al.	704/257
5,752,052	A	5/1998	Richardson et al.	395/759
5,754,784	A	5/1998	Garland et al.	395/200.49
5,761,631	A	6/1998	Nasukawa	704/9
5,774,859	A	6/1998	Houser et al.	704/275
5,794,050	A	8/1998	Dahlgren et al.	395/708
5,794,196	A	8/1998	Yegnanarayanan et al.	704/255
5,797,112	A	8/1998	Komatsu et al.	701/201
5,799,276	A	8/1998	Komissarchik et al.	704/251
5,802,510	A	9/1998	Jones	707/2
5,832,221	A	11/1998	Jones	375/200.36
5,839,107	A	11/1998	Gupta et al.	704/270
5,867,817	A	2/1999	Catallo et al.	704/255
5,878,385	A	3/1999	Bralich et al.	704/9
5,878,386	A	3/1999	Coughlin	704/10
5,892,813	A	4/1999	Morin et al.	379/88.01
5,895,464	A	4/1999	Bhandari et al.	707/3
5,895,466	A	4/1999	Goldberg et al.	707/5
5,897,613	A	4/1999	Chan	704/210
5,902,347	A	5/1999	Backman et al.	701/200
5,911,120	A	6/1999	Jarett et al.	455/417
5,918,222	A	6/1999	Fukui et al.	707/1
5,926,784	A	7/1999	Richardson et al.	704/9
5,933,822	A	8/1999	Braden-Harder et al.	707/5
5,953,393	A	9/1999	Culbreth et al.	379/88.25
5,960,399	A	9/1999	Barclay et al.	704/270
5,960,447	A	9/1999	Holt et al.	707/500
5,963,894	A	10/1999	Richardson et al.	704/9
5,963,940	A	10/1999	Liddy et al.	707/5
5,987,404	A	11/1999	Della Pietra et al.	704/9
5,991,721	A	11/1999	Asano et al.	704/257
5,995,119	A	11/1999	Cosatto et al.	345/473
5,995,928	A	11/1999	Nguyen et al.	704/251
6,009,382	A	12/1999	Martino et al.	704/1
6,014,559	A	1/2000	Amin	455/413
6,018,708	A	1/2000	Dahan et al.	704/244
6,021,384	A	2/2000	Gorin et al.	704/1
6,044,347	A	3/2000	Abella et al.	704/272
6,049,602	A	4/2000	Foladare et al.	379/265
6,049,607	A	4/2000	Marash et al.	379/410
6,058,187	A	5/2000	Chen	380/21
6,078,886	A	6/2000	Dragosh et al.	704/270
6,081,774	A	6/2000	De Hita et al.	704/9
6,085,186	A	7/2000	Christianson et al.	707/3
6,101,241	A	8/2000	Boyce et al.	379/88.01
6,108,631	A	8/2000	Ruhl	704/270
6,119,087	A	9/2000	Kuhn et al.	704/270
6,134,235	A	10/2000	Goldman et al.	370/352
6,144,667	A	11/2000	Doshi et al.	370/401
6,144,938	A	11/2000	Surace et al.	704/257
6,154,526	A *	11/2000	Dahlke et al.	379/88.03
6,160,883	A	12/2000	Jackson et al.	379/230
6,167,377	A	12/2000	Gillick et al.	704/240
6,173,266	B1	1/2001	Marx et al.	704/270
6,173,279	B1	1/2001	Levin et al.	707/5
6,175,858	B1	1/2001	Bulfer et al.	709/206
6,185,535	B1	2/2001	Hedin et al.	704/270
6,192,110	B1	2/2001	Abella et al.	379/88.01
6,192,338	B1	2/2001	Haszto et al.	704/257
6,195,634	B1	2/2001	Dudemaine et al.	704/231
6,195,651	B1	2/2001	Handel et al.	707/2
6,208,964	B1	3/2001	Sabourin	704/244
6,208,972	B1	3/2001	Grant et al.	704/275
6,219,346	B1	4/2001	Maxemchuk	370/338
6,219,643	B1	4/2001	Cohen et al.	704/257
6,226,612	B1	5/2001	Srenger et al.	704/256
6,233,556	B1	5/2001	Teunen et al.	704/250
6,233,559	B1	5/2001	Balakrishnan	704/275
6,246,981	B1	6/2001	Papineni et al.	704/235
6,269,336	B1	7/2001	Ladd et al.	704/270
6,272,455	B1	8/2001	Hoshen et al.	704/1
6,278,968	B1 *	8/2001	Franz et al.	704/3
6,288,319	B1	9/2001	Catona	84/609
6,292,767	B1	9/2001	Jackson et al.	704/1
6,308,151	B1	10/2001	Smith	704/235
6,314,402	B1	11/2001	Monaco et al.	704/275
6,366,882	B1	4/2002	Bijl et al.	704/235
6,366,886	B1	4/2002	Dragosh et al.	704/270.1
6,374,214	B1 *	4/2002	Friedland et al.	704/235
6,377,913	B1	4/2002	Coffman et al.	704/8
6,381,535	B1	4/2002	Durocher et al.	701/202
6,385,596	B1	5/2002	Wiser et al.	705/51
6,385,646	B1	5/2002	Brown et al.	709/217
6,393,428	B1	5/2002	Miller et al.	707/102
6,397,181	B1	5/2002	Li et al.	704/256
6,404,878	B1	6/2002	Jackson et al.	379/221.01
6,405,170	B1	6/2002	Phillips et al.	704/270
6,408,272	B1	6/2002	White et al.	704/270.1
6,411,810	B1	6/2002	Maxemchuk	455/453
6,415,257	B1	7/2002	Junqua et al.	704/275
6,418,210	B1	7/2002	Sayko	379/142.15
6,420,975	B1	7/2002	DeLine et al.	340/815.4
6,429,813	B2	8/2002	Feigen	342/357.13
6,430,285	B1	8/2002	Bauer et al.	379/265.01
6,430,531	B1	8/2002	Polish	704/257
6,434,523	B1	8/2002	Monaco	704/257
6,434,524	B1	8/2002	Weber	704/257
6,442,522	B1	8/2002	Carberry et al.	704/257
6,446,114	B1	9/2002	Bulfer et al.	709/206
6,453,153	B1	9/2002	Bowker et al.	455/67.4
6,453,292	B2	9/2002	Ramaswamy et al.	704/235
6,456,711	B1	9/2002	Cheung et al.	379/265.09
6,466,654	B1	10/2002	Cooper et al.	379/88.01
6,466,899	B1	10/2002	Yano et al.	704/1
6,470,315	B1	10/2002	Netsch et al.	704/256
6,498,797	B1	12/2002	Anerousis et al.	370/522
6,499,013	B1	12/2002	Weber	704/257
6,501,833	B2	12/2002	Phillips et al.	379/88.07
6,501,834	B1	12/2002	Milewski et al.	379/93.24
6,513,006	B2	1/2003	Howard et al.	704/257
6,522,746	B1	2/2003	Marchok et al.	379/406.03
6,523,061	B1	2/2003	Halverson et al.	709/202
6,532,444	B1	3/2003	Weber	704/257
6,539,348	B1	3/2003	Bond et al.	704/9
6,549,629	B2	4/2003	Finn et al.	381/92
6,553,372	B1	4/2003	Brassell et al.	707/5
6,556,970	B1	4/2003	Sasaki et al.	704/257
6,556,973	B1	4/2003	Lewin	704/277
6,560,576	B1	5/2003	Cohen et al.	704/270
6,567,778	B1	5/2003	Chao Chang et al.	704/257
6,567,797	B1	5/2003	Schuetze et al.	707/2
6,570,555	B1	5/2003	Prevost et al.	345/156
6,570,964	B1	5/2003	Murveit et al.	379/67.1
6,574,597	B1	6/2003	Mohri et al.	704/251
6,574,624	B1	6/2003	Johnson et al.	707/5
6,581,103	B1	6/2003	Dengler	709/231
6,587,858	B1	7/2003	Strazza	707/102
6,594,257	B1	7/2003	Doshi et al.	370/352
6,598,018	B1	7/2003	Junqua	704/251
6,604,075	B1	8/2003	Brown et al.	704/270.1
6,604,077	B2	8/2003	Dragosh et al.	704/270.1
6,606,598	B1	8/2003	Holthouse et al.	704/275
6,611,692	B2	8/2003	Raffel et al.	455/552
6,614,773	B1	9/2003	Maxemchuk	370/337
6,615,172	B1	9/2003	Bennett et al.	704/257
6,622,119	B1	9/2003	Ramaswamy et al.	704/9
6,629,066	B1	9/2003	Jackson et al.	704/9
6,631,346	B1	10/2003	Karaorman et al.	704/9
6,643,620	B1	11/2003	Contolini et al.	704/270
6,650,747	B1	11/2003	Bala et al.	379/265.06
6,658,388	B1	12/2003	Kleindienst et al.	704/275
6,678,680	B1	1/2004	Woo	707/6
6,681,206	B1	1/2004	Gorin et al.	704/243
6,691,151	B1	2/2004	Cheyer et al.	709/202
6,701,294	B1	3/2004	Ball et al.	704/257

6,704,708	B1	3/2004	Pickering	704/235
6,708,150	B1	3/2004	Hirayama et al.	704/243
6,721,001	B1	4/2004	Berstis	348/231.3
6,721,706	B1	4/2004	Strubbe et al.	704/275
6,735,592	B1	5/2004	Neumann et al.	707/101
6,741,931	B1	5/2004	Kohut et al.	701/209
6,742,021	B1	5/2004	Halverson et al.	709/218
6,751,591	B1 *	6/2004	Gorin et al.	704/257
6,751,612	B1	6/2004	Schuetze et al.	707/4
6,757,718	B1	6/2004	Halverson et al.	709/218
6,795,808	B1	9/2004	Strubbe et al.	704/275
6,801,604	B2	10/2004	Maes et al.	379/88.17
6,801,893	B1	10/2004	Backfried et al.	704/257
6,829,603	B1 *	12/2004	Wolf et al.	1/1
6,832,230	B1	12/2004	Zilliacus et al.	707/203
6,833,848	B1	12/2004	Wolff et al.	345/719
6,856,990	B2	2/2005	Barile et al.	707/10
6,865,481	B2	3/2005	Kawazoe et al.	701/211
6,877,134	B1	4/2005	Fuller et al.	715/500.1
6,901,366	B1	5/2005	Kuhn et al.	704/275
6,912,498	B2	6/2005	Stevens et al.	704/235
6,937,977	B2	8/2005	Gerson	704/201
6,944,594	B2	9/2005	Busayapongchai et al.	704/275
6,950,821	B2	9/2005	Faybishenko et al.	707/10
6,954,755	B2	10/2005	Reisman	707/10
6,973,387	B2	12/2005	Masclet et al.	701/211
6,975,993	B1	12/2005	Keiller	704/275
6,980,092	B2	12/2005	Turnbull et al.	340/425.5
6,990,513	B2	1/2006	Belfiore et al.	709/203
6,996,531	B2	2/2006	Korall et al.	704/270
7,003,463	B1	2/2006	Maes et al.	704/270.1
7,016,849	B2	3/2006	Arnold et al.	704/275
7,020,609	B2	3/2006	Thrift et al.	704/270.1
7,024,364	B2	4/2006	Guerra et al.	704/270
7,027,975	B1	4/2006	Pazandak et al.	704/9
7,058,890	B2	6/2006	George et al.	715/728
7,062,488	B1	6/2006	Reisman	707/8
7,082,469	B2	7/2006	Gold et al.	709/231
7,092,928	B1	8/2006	Elad et al.	706/60
7,127,400	B2	10/2006	Koch	704/270.1
7,136,875	B2	11/2006	Anderson et al.	707/104.1
7,137,126	B1	11/2006	Coffman et al.	719/328
7,146,319	B2	12/2006	Hunt	704/254
7,197,069	B2	3/2007	Agazzi et al.	375/233
7,203,644	B2	4/2007	Anderson et al.	704/246
7,228,276	B2	6/2007	Omote et al.	704/243
7,231,343	B1	6/2007	Treadgold et al.	704/9
7,236,923	B1	6/2007	Gupta	704/9
7,277,854	B2	10/2007	Bennett et al.	704/257
7,289,606	B2	10/2007	Sibal et al.	379/52
7,301,093	B2	11/2007	Sater et al.	84/615
7,340,040	B1	3/2008	Saylor et al.	379/67.1
7,366,669	B2	4/2008	Nishitani et al.	704/256
7,376,645	B2	5/2008	Bernard	707/3
7,386,443	B1	6/2008	Parthasarathy et al.	704/201
7,398,209	B2	7/2008	Kennewick et al.	704/255
7,415,414	B2	8/2008	Azara et al.	704/270
7,424,431	B2	9/2008	Greene et al.	704/270
7,461,059	B2	12/2008	Richardson et al.	707/5
7,472,020	B2	12/2008	Brulle-Drews	701/211
7,472,060	B1	12/2008	Gorin et al.	704/240
7,487,088	B1 *	2/2009	Gorin et al.	704/240
7,493,259	B2	2/2009	Jones et al.	704/257
7,493,559	B1	2/2009	Wolff et al.	715/727
7,502,738	B2	3/2009	Kennewick et al.	704/257
7,536,297	B2	5/2009	Byrd et al.	704/10
7,536,374	B2 *	5/2009	Au	706/55
7,558,730	B2	7/2009	Davis et al.	704/235
7,620,549	B2	11/2009	Di Cristo et al.	704/257
7,634,409	B2	12/2009	Kennewick et al.	704/257
7,640,160	B2	12/2009	Di Cristo et al.	704/257
7,676,365	B2	3/2010	Hwang et al.	704/240
7,693,720	B2	4/2010	Kennewick et al.	704/275
7,788,084	B2 *	8/2010	Brun et al.	704/7
7,809,570	B2	10/2010	Kennewick et al.	704/257
7,818,176	B2	10/2010	Freeman et al.	704/270
7,831,433	B1	11/2010	Belvin et al.	704/275
7,873,523	B2	1/2011	Potter et al.	704/275
7,917,367	B2	3/2011	Di Cristo et al.	704/270.1
7,949,529	B2	5/2011	Weider et al.	704/270
7,983,917	B2	7/2011	Kennewick et al.	704/257
8,015,006	B2	9/2011	Kennewick et al.	704/236
2001/0041980	A1	11/2001	Howard et al.	704/270
2001/0049601	A1 *	12/2001	Kroeker et al.	704/254
2002/0035501	A1	3/2002	Handel et al.	705/10
2002/0049805	A1	4/2002	Yamada et al.	709/202
2002/0065568	A1	5/2002	Silfvast et al.	700/94
2002/0069059	A1	6/2002	Smith	704/257
2002/0082911	A1	6/2002	Dunn et al.	705/14
2002/0087525	A1	7/2002	Abbott et al.	707/3
2002/0120609	A1	8/2002	Lang et al.	707/1
2002/0124050	A1	9/2002	Middeljans	709/203
2002/0138248	A1	9/2002	Corston-Oliver et al.	704/1
2002/0143535	A1	10/2002	Kist et al.	704/251
2002/0188602	A1	12/2002	Stubler et al.	707/3
2002/0198714	A1	12/2002	Zhou	704/252
2003/0014261	A1	1/2003	Kageyama	704/275
2003/0046346	A1	3/2003	Mumick et al.	709/205
2003/0064709	A1	4/2003	Gailey et al.	455/412
2003/0088421	A1	5/2003	Maes et al.	704/270.1
2003/0097249	A1 *	5/2003	Walker et al.	704/1
2003/0112267	A1	6/2003	Belrose	345/728
2003/0120493	A1	6/2003	Gupta	704/270.1
2003/0135488	A1	7/2003	Amir et al.	707/3
2003/0144846	A1	7/2003	Denenberg et al.	704/277
2003/0182132	A1	9/2003	Niemoeller	704/275
2003/0212550	A1	11/2003	Ubale	704/215
2003/0236664	A1	12/2003	Sharma	704/251
2004/0006475	A1	1/2004	Ehlen et al.	704/270.1
2004/0025115	A1	2/2004	Sienel et al.	715/513
2004/0044516	A1	3/2004	Kennewick et al.	704/5
2004/0166832	A1	8/2004	Portman et al.	455/412.1
2004/0167771	A1 *	8/2004	Duan et al.	704/10
2004/0193408	A1	9/2004	Hunt	704/209
2004/0193420	A1	9/2004	Kennewick et al.	704/257
2004/0199375	A1	10/2004	Ehsani et al.	704/4
2004/0243417	A9	12/2004	Pitts, III et al.	704/276
2005/0015256	A1 *	1/2005	Kargman	704/272
2005/0021334	A1	1/2005	Iwahashi	704/240
2005/0021826	A1	1/2005	Kumar	709/232
2005/0033574	A1	2/2005	Kim et al.	704/251
2005/0043940	A1	2/2005	Elder	704/9
2005/0114116	A1	5/2005	Fiedler	704/201
2005/0137850	A1	6/2005	Odell	704/4
2005/0246174	A1	11/2005	DeGolia	704/270
2006/0206310	A1	9/2006	Ravikumar et al.	704/9
2007/0033005	A1	2/2007	Cristo et al.	704/9
2007/0033020	A1	2/2007	(Kelleher) Francois et al.	704/226
2007/0038436	A1	2/2007	Cristo et al.	704/9
2007/0043574	A1	2/2007	Coffman et al.	704/275
2007/0050191	A1	3/2007	Weider et al.	704/275
2007/0055525	A1	3/2007	Kennewick et al.	704/257
2007/0118357	A1 *	5/2007	Kasravi et al.	704/10
2007/0179778	A1	8/2007	Gong et al.	704/9
2007/0186165	A1	8/2007	Maislos et al.	715/728
2007/0214182	A1	9/2007	Rosenberg	707/104.1
2007/0250901	A1	10/2007	McIntire et al.	725/146
2007/0265850	A1	11/2007	Kennewick et al.	704/257
2007/0299824	A1	12/2007	Pan et al.	707/3
2008/0091406	A1	4/2008	Baldwin et al.	704/4
2008/0103761	A1	5/2008	Printz et al.	704/9
2008/0115163	A1	5/2008	Gilboa et al.	725/34
2008/0133215	A1	6/2008	Sarukkai	704/2
2008/0140385	A1	6/2008	Mahajan et al.	704/9
2008/0177530	A1	7/2008	Cross et al.	704/4
2008/0189110	A1	8/2008	Freeman et al.	704/251
2008/0235023	A1	9/2008	Kennewick et al.	704/257
2008/0235027	A1	9/2008	Cross	704/270.1
2008/0319751	A1	12/2008	Kennewick et al.	704/257
2009/0117885	A1	5/2009	Roth	455/414.3
2009/0144271	A1	6/2009	Richardson et al.	707/5
2009/0150156	A1	6/2009	Kennewick et al.	704/257
2009/0171664	A1	7/2009	Kennewick et al.	704/257
2009/0216540	A1	8/2009	Tessel et al.	704/275
2009/0271194	A1	10/2009	Davis et al.	704/235
2009/0299745	A1	12/2009	Kennewick et al.	704/257
2010/0023320	A1	1/2010	Di Cristo et al.	704/9

2010/0049501	A1	2/2010	Kennewick et al.	704/9
2010/0049514	A1	2/2010	Kennewick et al.	704/233
2010/0057443	A1	3/2010	Di Cristo et al.	704/9
2010/0063880	A1	3/2010	Atsmon et al.	705/14.53
2010/0145700	A1	6/2010	Kennewick et al.	704/ 257
2010/0204986	A1	8/2010	Kennewick et al.	704/226
2010/0204994	A1	8/2010	Kennewick et al.	704/257
2010/0217604	A1	8/2010	Baldwin et al.	704/275
2010/0286985	A1	11/2010	Kennewick et al.	704/257
2010/0299142	A1	11/2010	Freeman et al.	704/9
2011/0112827	A1	5/2011	Kennewick et al.	704/9
2011/0112921	A1	5/2011	Kennewick et al.	705/26.1
2011/0131036	A1	6/2011	Di Cristo et al.	704/9
2011/0131045	A1	6/2011	Cristo et al.	704/249

FOREIGN PATENT DOCUMENTS

EP	1 646 037	4/2006
WO	WO 99/46763	9/1999
WO	WO 00/21232	4/2000
WO	WO 00/46792	8/2000
WO	WO 2004/072954	8/2001
WO	WO 01/78065	10/2001
WO	WO 2007/019318	2/2007
WO	WO 2007/021587	2/2007
WO	WO 2007/027546	3/2007
WO	WO 2007/027989	3/2007
WO	WO 2008/098039	8/2008
WO	WO 2008/118195	10/2008
WO	WO 2009/075912	6/2009
WO	WO 2009/145796	12/2009
WO	WO 2010/096752	8/2010

OTHER PUBLICATIONS

Cheyer et al., "Multimodal Maps: An Agent-Based Approach", *International Conference on Cooperative Multimodal Communication* (CMC/95), May 24-26, 1995, pp. 111-121.

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.

Turunen, "Adaptive Interaction Methods in Speech User Interfaces", Conference on Human Factors in Computing Systems, Seattle, Washington, 2001, pp. 91-92.

Mao, Mark Z., "Automatic Training Set Segmentation for Multi-Pass Speech Recognition", Department of Electrical Engineering, Stanford University, CA, copyright 2005, IEEE, pp. I-685 to I-688.

Vanhoucke, Vincent, "Confidence Scoring and Rejection Using Multi-Pass Speech Recognition", Nuance Communications, Menlo Park, CA, [no date], 4 pages.

Weng, Fuliang, et al., "Efficient Lattice Representation and Generation", Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, [no date], 4 pages.

El Meliani et al., "A Syllabic-Filler-Based Continuous Speech Recognizer for Unlimited Vocabulary", Canadian Conference on Electrical and Computer Engineering, vol. 2, Sep. 5-8, 1995, pp. 1007-1010.

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/>, Oct. 28, 2009, 4 pages.

Reuters, "IBM to Enable Honda Drivers to Talk to Cars", Charles Schwab & Co., Inc., Jul. 28, 2002, 1 page.

Lin, Bor-shen, et al., "A Distributed Architecture for Cooperative Spoken Dialogue Agents with Coherent Dialogue State and History", ASRU'99, 1999, 4 pages.

Kuhn, Thomas, et al., "Hybrid In-Car Speech Recognition for Mobile Multimedia Applications", Vehicular Technology Conference, IEEE, Jul. 1999, pp. 2009-2013.

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, pp. 1-5.

Lind, R., et al., "The Network Vehicle—A Glimpse into the Future of Mobile Multi-Media", *IEEE Aerosp. Electron. Systems Magazine*, vol. 14, No. 9, Sep. 1999, pp. 27-32.

Zhao, Yilin, "Telematics: Safe and Fun Driving", *IEEE Intelligent Systems*, vol. 17, Issue 1, 2002, pp. 10-14.

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*, vol. 3, Jun. 5-9, 2000, XP010507574, pp. 1257-1260.

O'Shaughnessy, Douglas, "Interacting with Computers by Voice: Automatic Speech Recognition and Synthesis", *Proceedings of the IEEE*, vol. 91, No. 9, Sep. 1. 2003. XP011100665, pp. 1272-1305.

Statement in Accordance with the Notice from the European Patent Office dated Oct. 1, 2007 Concerning Business Methods (OJ EPO Nov. 2007, 592-593), XP002456252.

* cited by examiner

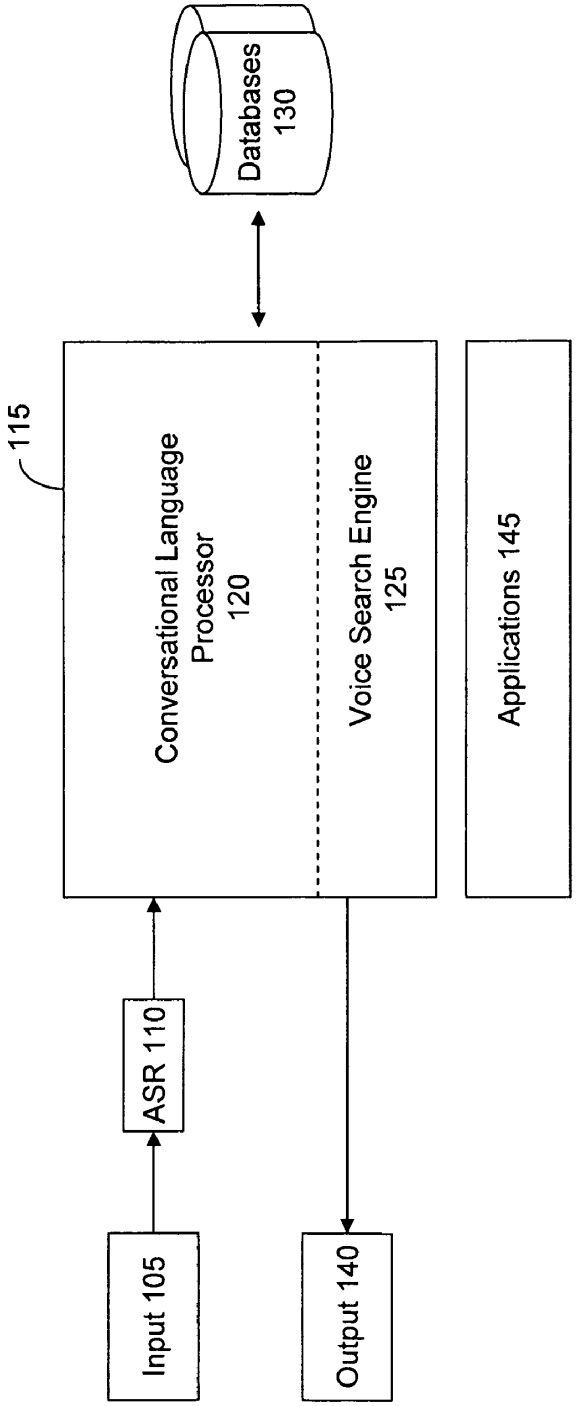

Input 105
ASR 110
115
Conversational Language Processor 120
Voice Search Engine 125
Output 140
Databases 130
Applications 145

FIG. 1

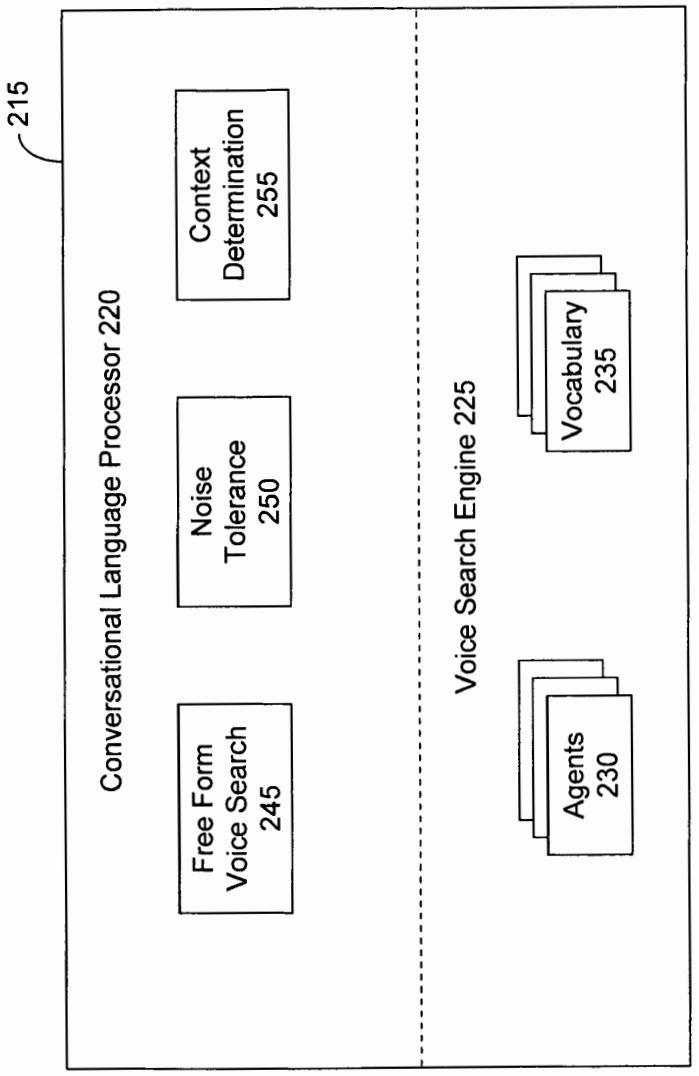

215
Conversational Language Processor 220
Free Form Voice Search 245
Noise Tolerance 250
Context Determination 255
Voice Search Engine 225
Agents 230
Vocabulary 235

FIG. 2

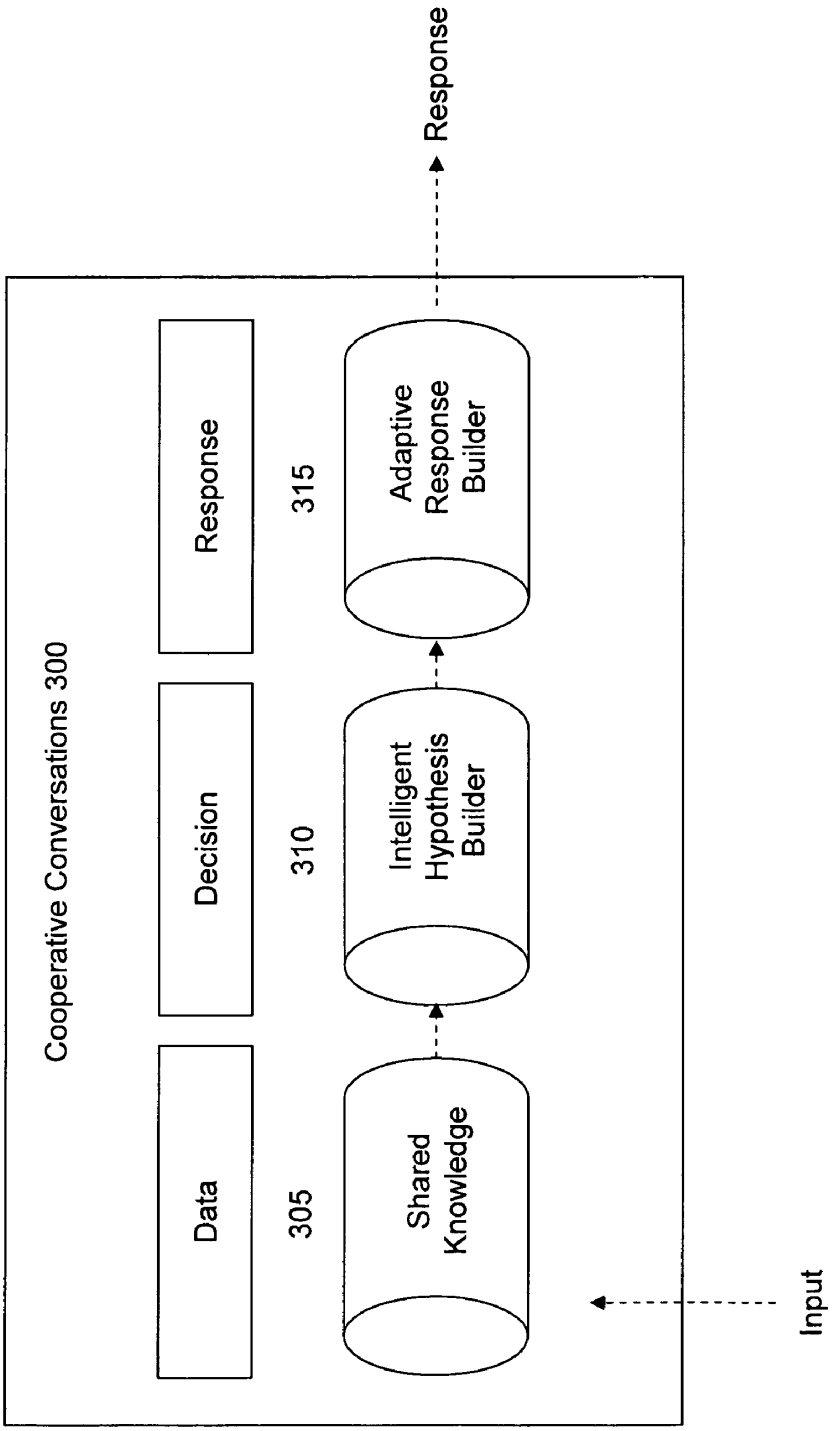

Cooperative Conversations 300
Data
Decision
Response
305
310
315
Shared Knowledge
Intelligent Hypothesis Builder
Adaptive Response Builder
Input
Response

FIG. 3

SYSTEM AND METHOD FOR A COOPERATIVE CONVERSATIONAL VOICE USER INTERFACE

FIELD OF THE INVENTION

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

BACKGROUND OF THE INVENTION

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.

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.

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

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.

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

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.

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.

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 grammars associated with the speech recognition engine and/or from possible interpretations determined by the conversational speech engine, thereby assuring that a mistake will not be repeated for an identical utterance.

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.

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).

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.

According to another aspect of the inventions 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).

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.

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).

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."

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.

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.

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.

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

FIG. **1** is an exemplary block diagram of a system architecture according to one aspect of the invention.

FIG. **2** is an exemplary block diagram of a conversational speech engine according to one aspect of the invention.

FIG. **3** is an exemplary block diagram of a cooperative conversational model according to one aspect of the invention.

DETAILED DESCRIPTION

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.

The utterance component of input **105** may be processed by a speech recognition engine **110** (which may alternatively be referred to herein as Automatic Speech Recognizer **110**, or as shown in FIG. **1**, ASR **110**) to generate one or more preliminary interpretations of the utterance. The speech recognition engine **110** may process the utterance using any suitable technique known in the art. For example, in one implementation, the speech recognition engine **110** may interpret the utterance using techniques of phonetic dictation to recognize a phoneme stream, as described in U.S. patent application Ser. No. 11/513,269, entitled "Dynamic Speech Sharpening," filed Aug. 31, 2006, which issued as U.S. Pat. No. 7,634,409 on Dec. 15, 2009, and which is hereby incorporated by reference in its entirety. The one or more preliminary interpretations generated by the speech recognition engine **110** may then be provided to a conversational speech engine **115** for further processing. Conversational speech engine **115** may include a conversational language processor **120** and/or a voice search engine **125**, described in greater detail in FIG. **2** below. Conversational speech engine **115** may communicate with one or more databases **130** to generate an adaptive conversational response, which may be returned to the user as an output **140**. In one implementation, output **140** may be a multi-modal output and/or an interaction with one or more applications **145** to complete the request. For example, output **140** may include a combination of an audible response and a display of a route on a navigation device. For example, the utterance may include a request to perform an action, and output **140** may include a conversational response reporting success or failure, as well as an execution of the action. In addition, in various implementations, the speech recognition engine **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).

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.

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.

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).

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.

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.

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.

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.

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.

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 U.S. patent application Ser. No. 11/197,504, entitled "Systems and Methods for Responding to Natural Language Speech Utterance," filed Aug. 5, 2005, which issued as U.S. Pat. No. 7,640,160 on Dec. 29, 2009 and U.S. patent application Ser. No. 11/212,

693, entitled "Mobile Systems and Methods of Supporting Natural Language Human-Machine Interactions," filed Aug. 29, 2005, both of which are hereby incorporated by reference in their entirety. Once a given context domain agent "wins" the competition, the winning agent may be responsible for establishing or inferring further contexts and updating short-term and long-term shared knowledge. If there is a deadlock between context domain agents, an adaptive conversational response may prompt the user to assist in disambiguating between the deadlocked agents. For example, a user utterance of "What about traffic?" may have a distinct meaning in various contexts. That is, "traffic" may have a first meaning when the user is querying a system's media player (i.e., "traffic" would be a Rock and Roll band led by singer/songwriter Steve Winwood), a second meaning when the user is querying a search interface regarding Michael Douglas films (i.e., "traffic" would be a film directed by Steven Soderbergh), a third meaning when the user is querying a navigation device for directions to an airport (i.e., "traffic" would be related to conditions on roads along a route to the airport).

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 grammars associated with the speech recognition engine and/or from possible subsequent interpretations determined by context determination process **255**, thereby assuring that a mistake will not be repeated for an identical utterance.

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.

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, Oreg. based on the Weather and an environmental proximity between Portland, Oreg. and Seattle, Wash. 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, N.H.," in the second utterance, context determination process **255** may instead retrieve a weather report for Portland, Me. 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, Me., or Portland, Oreg.?").

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.

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.

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**.

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.

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.

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.

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.

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).

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 conver-

sation, 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:

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

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.

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.

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.

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.

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 a confidence level associated with a preliminary interpretation generated at the speech recognition engine exceeds a predetermined threshold; (2) "pretty sure," where contextual signifiers and qualifiers relate to more than one task (select top-ranked task) and criteria relates to one request, and/or the confidence level associated with the preliminary interpretation generated at the speech recognition engine is below the predetermined threshold; (3) "not sure," where additional contextual signifiers or qualifiers are needed to indicate or rank a task; and (4) "no hypothesis," where little or no information can be deciphered. Each degree of certainty may further be classified as explicit or implicit, which may be used to adjust a response. The input received by Adaptive Response Builder **310** may also include a context, user syntax and/or grammar, context domain agent specific information and/or preferences (e.g., a travel context domain agent may know a user frequently requests information about France, which may be shared with a movie context domain agent so that responses may occasionally include French movies).

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.

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.

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?

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 d 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?

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.

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.

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 cer-

tainty, 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.

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.

What is claimed is:

1. A method for providing a cooperative conversational voice user interface, comprising:

receiving an utterance at a voice input device during a current conversation with a user, wherein the utterance includes one or more words that have different meanings in different contexts;

accumulating short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received during the current conversation;

accumulating long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

determining an intended meaning for the utterance, wherein determining the intended meaning for the utterance includes:

identifying, at a conversational speech engine, a context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge; and

establishing the intended meaning within the identified context, wherein the conversational speech engine establishes the intended meaning within the identified context to disambiguate an intent that the user had in speaking the one or more words that have the different meanings in the different contexts; and

generating a response to the utterance, wherein the conversational speech engine grammatically or syntactically adapts the response based on the intended meaning established within the identified context.

2. The method of claim **1**, wherein accumulating the short-term shared knowledge about the current conversation includes populating a short-term context stack with information about the utterance received during the current conversation.

3. The method of claim **2**, wherein accumulating the short-term shared knowledge about the current conversation further includes expiring the information about the utterance from the short-term context stack after a psychologically appropriate amount of time.

4. The method of claim **3** wherein accumulating the long-term shared knowledge about the user includes updating one or more long-term profiles associated with the user to include information about the utterance received during the current conversation and relevant data associated with the information expired from the short, term context stack.

5. The method of claim **1**, wherein determining the intended meaning for the utterance further includes:

identifying a conversational goal associated with the utterance, roles associated with the user and one or more other participants in the current conversation, and an information allocation among the user and the one or more other participants in the current conversation; and

classifying one or more of the utterance or the current conversation into a conversation type based on one or more of the identified conversational goal, the identified roles, or the identified information allocation, wherein the conversational speech engine further establishes the intended meaning based on the conversation type.

6. The method of claim **5** wherein the established intended meaning comprises a hypothesis having a degree of certainty about the intent that the user had in speaking the one or more words in the utterance.

7. The method of claim **6**, further comprising generating a preliminary interpretation of the utterance at a speech recognition engine coupled to the voice input device and the conversational speech engine, wherein the conversational speech engine assigns the degree of certainty to the hypothesis based on one or more of the conversation type, information associated with the identified context, or a confidence level associated with the preliminary interpretation generated at the speech recognition engine.

8. The method of claim **5**, wherein the conversational speech engine further grammatically or syntactically adapts the response based on the conversation type.

9. The method of claim **1**, wherein the conversational speech engine grammatically or syntactically adapts the response to influence a subsequent reply utterance that the conversational speech engine expects from the user during the current conversation.

10. The method of claim **1**, further comprising:

generating multiple preliminary interpretations of the utterance at a speech recognition engine coupled to the voice input device and the conversational speech engine, wherein an initial interpretation of the utterance comprises one of the multiple preliminary interpretations having a highest confidence level; and

updating the short-term shared knowledge about the current conversation to remove the initial interpretation from the multiple preliminary interpretations in response to determining that the initial interpretation was incorrect, wherein the conversational speech engine determines the intended meaning based on one of the multiple preliminary interpretations having a next highest confidence level.

11. The method of claim **1**, wherein the user speaks the utterance in a multi-modal input that further includes one or more non-voice inputs relating to the utterance.

12. The method of claim 1, wherein the conversational speech engine generates the response in a multi-modal output that includes one or more non-voice outputs that relate to the utterance or one or more tasks executed to process a request identified from the intended meaning.

13. A non-transitory computer readable medium containing computer-executable instructions for providing a cooperative conversational voice user interface, the computer-executable instructions operable when executed to:

receive an utterance at a voice input device, during a current conversation with a user, wherein the utterance includes one or more words that have different meanings in different contexts;

accumulate short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received at the voice during the current conversation;

accumulate long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

identify a context associated with the utterance, wherein a conversational speech engine identifies the context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge;

establish an intended meaning for the utterance within the identified context, wherein the conversational speech engine establishes the intended meaning within the identified context to disambiguate an intent that the user had in speaking the one or more words that have the different meanings in the different contexts; and

generate a response to the utterance, wherein the conversational speech engine grammatically or syntactically adapts the response based on the intended meaning established within the identified context.

14. The non-transitory computer readable medium of claim 13, wherein to accumulate the short-term shared knowledge about the current conversation, the computer-executable instructions are further operable when executed to populate a short-term context stack with information about the utterance received during the current conversation.

15. The non-transitory computer readable medium of claim 14, wherein to accumulate the short-term shared knowledge about the current conversation, the computer-executable instructions are further operable when executed to expire the information about the utterance from the short-term context stack after a psychologically appropriate amount of time.

16. The non-transitory computer readable medium of claim 15, wherein to accumulate the long-term shared knowledge about the user, the computer-executable instructions are further operable when executed to update one or more long-term profiles associated with the user to include information about the utterance received during the current conversation and relevant data associated with the information expired from the short term context stack.

17. The non-transitory computer readable medium of claim 13, wherein the computer-executable instructions are further operable when executed to:

identify a conversational goal associated with the utterance, roles associated with the user and one or more other participants in the current conversation, and an information allocation among the user and the one or more other participants in the current conversation; and

classify one or more of the utterance or the current conversation into a conversation type based on one or more of the identified conversational goal, the identified roles, or the identified information allocation, wherein the conversational speech engine further establishes the intended meaning based on the conversation type.

18. The non-transitory computer readable medium of claim 17, wherein the established intended meaning comprises a hypothesis having a degree of certainty about the intent that the user had in speaking the one or more words in the utterance.

19. The computer-readable medium of claim 18, wherein the computer-executable instructions are further operable when executed to generate a preliminary interpretation of the utterance at a speech recognition engine, wherein the conversational speech engine assigns the degree of certainty to the hypothesis based on one or more of the conversation type, information associated with the identified context, or a confidence level associated with the preliminary interpretation generated at the speech recognition engine.

20. The non-transitory computer readable medium of claim 17, wherein the conversational speech engine further grammatically or syntactically adapts the response based on conversation type.

21. The non-transitory computer readable medium of claim 13, wherein the conversational speech engine grammatically or syntactically adapts more the response to influence a subsequent reply utterance that the conversational speech engine expects from the user during the current conversation.

22. The non-transitory computer readable medium of claim 13, wherein the computer-executable instructions are further operable when executed to:

generate multiple preliminary interpretations of the utterance at a speech recognition engine, wherein an initial interpretation of the utterance comprises one of the multiple preliminary interpretations having a highest confidence level; and

update the short-term shared knowledge about the current conversation to remove the initial interpretation from the multiple preliminary interpretations in response to determining that the initial interpretation was incorrect, wherein the conversational speech engine identifies the context associated with the utterance and establishes the intended meaning for the utterance based on one of the multiple preliminary interpretations having a next highest confidence level.

23. The non-transitory computer readable medium of claim 13, wherein the user speaks the utterance in a multi-modal input that further includes one or more non-voice inputs relating to the utterance.

24. The non-transitory computer readable medium of claim 13, wherein the conversational speech engine generates the response in a multi-modal output that includes one or more non-voice outputs that relate to the utterance or one or more tasks executed to process a request identified from the intended meaning.

25. A system for providing a cooperative conversational voice user interface, comprising:

a voice input device configured to receive an utterance during a current conversation with a user, wherein the utterance includes one or more words that have different meanings in different contexts; and

a conversational speech engine, wherein the conversational speech engine includes one or more processors configured to:

accumulate short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received during the current conversation;

accumulate long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

identify a context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge;

establish an intended meaning for the utterance within the identified context to disambiguate an intent that the user had in speaking the one or more words that have the different meanings in the different contexts; and

generate a grammatically or syntactically adapted response to the utterance based on the intended meaning established within the identified context.

26. The system of claim **25**, wherein to accumulate the short-term shared knowledge about the current conversation, the one or more processors are further configured to populate a short-term context stack with information about the utterance received during the current conversation.

27. The system of claim **26**, wherein to accumulate the short-term shared knowledge about the current conversation, the one or more processors are further configured to expire the information about the utterance from the short-term context stack after a psychologically appropriate amount of time.

28. The system of claim **27**, wherein to accumulate the long-term shared knowledge about the user, the one or more processors are further configured to update one or more long-term profiles associated with the user to include information about the utterance received during the current conversation and relevant data associated with the information expired from the short term context stack.

29. The system of claim **25**, wherein the one or more processors are further configured to:

identify a conversational goal associated with the utterance, roles associated with the user and one or more other participants in the current conversation, and an information allocation among the user and the one or more other participants in the current conversation; and

classify one or more of the utterance or the current conversation into a conversation type based on one or more of the identified conversational goal, the identified roles, or the identified information allocation, wherein the one or more processors are further configured to establish the intended meaning based on the conversation type.

30. The system of claim **29** wherein the established intended meaning comprises a hypothesis having a degree of certainty about the intent that the user had in speaking the one or more words in the utterance.

31. The system of claim **30**, further comprising a speech recognition engine configured to generate a preliminary interpretation of the utterance, wherein the one or more processors are further configured to assign the degree of certainty to the hypothesis based on one or more of the conversation type, information associated with the identified context, a confidence level associated with the preliminary interpretation generated at the speech recognition engine.

32. The system of claim **29** wherein the one or more processors are further configured to generate the grammatically syntactically adapted response based on the conversation type.

33. The system of claim **25**, wherein the one or more processors are further configured to generate the grammatically or syntactically adapted response to influence a subsequent reply utterance expected from the user during the current conversation.

34. The system of claim **25**, further comprising a speech recognition engine configured to:

generate multiple preliminary interpretations of the utterance, wherein an initial interpretation of the utterance comprises one of the multiple preliminary interpretations having a highest confidence level; and

update the short-term shared knowledge about the current conversation to remove the initial interpretation from the multiple preliminary interpretations in response to determining that the initial interpretation was incorrect, wherein the one or more processors are configured to identify the context associated with the utterance and establish the intended meaning for the utterance based on one of the multiple preliminary interpretations having a next highest confidence level.

35. The system of claim **25**, wherein the user speaks the utterance in a multi-modal input that further includes one or more non-voice inputs relating to the utterance.

36. The system of claim **25**, wherein the grammatically or syntactically adapted response comprises a multi-modal output that includes one or more non-voice outputs that relate to the utterance or one or more tasks executed to process a request identified from the intended meaning.

37. A method for providing a cooperative conversational voice user interface, comprising:

receiving an utterance at a voice input device during a current conversation with a user;

accumulating short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received during the current conversation;

accumulating long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

determining an intended meaning for the utterance, wherein determining the intended meaning for the utterance includes:

identifying, at a conversational speech engine, a context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge;

inferring additional information about the utterance from the short-term shared knowledge and the long-term shared knowledge in response to determining that the utterance contains insufficient information to complete a request in the identified context; and

establishing the intended meaning within the identified context based on the additional information inferred about the utterance; and

generating a response to the utterance based on the intended meaning established within the identified context.

38. The method of claim **37**, wherein the established intended meaning comprises an implicit hypothesis having a corresponding degree of certainty about an intent that the user had in speaking the utterance.

39. A non-transitory computer readable medium containing computer-executable instructions for providing a cooperative conversational voice user interface, the computer-executable instructions operable when executed to:

receive an utterance at a voice input device during a current conversation with a user;

accumulate short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received during the current conversation;

accumulate long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

identify a context associated with the utterance, wherein a conversational speech engine identifies the context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge;

infer additional information about the utterance from the short-term shared knowledge and the long-term shared knowledge in response to determining that the utterance contains insufficient information to complete a request in the identified context;

establish an intended meaning for the utterance within the identified context based on the additional information inferred about the utterance; and

generate a response to the utterance based on the intended meaning established within the identified context.

40. The non-transitory computer readable medium of claim **39**, wherein the established intended meaning comprises an implicit hypothesis having a corresponding degree of certainty about an intent that the user had in speaking the utterance.

41. A system for providing a cooperative conversational voice user interface, comprising:

a voice input device configured to receive an utterance during a current conversation with a user; and

a conversational speech engine, wherein the conversational speech engine includes one or more processors configured to:

accumulate short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received during the current conversation;

accumulate long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

identify a context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge;

infer additional information about the utterance from the short-term shared knowledge and the long-term shared knowledge in response to determining that the utterance contains insufficient information to complete a request in the identified context;

establish an intended meaning for the utterance within the identified identify a context based on the additional information inferred about the utterance; and

generate a response to the utterance based on the intended meaning established within the identified context.

42. The system of claim **41**, wherein the established intended meaning comprises an implicit hypothesis having a corresponding degree of certainty about an intent that the user had in speaking the utterance.

* * * * *