

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

SAMSUNG ELECTRONICS CO. LTD. and SAMSUNG ELECTRONICS
AMERICA, INC.,
Petitioners,

v.

VB ASSETS, LLC,
Patent Owner

IPR2025-00866

U.S. Patent No. 8,073,681

PETITION FOR *INTER PARTES* REVIEW OF U.S. PATENT NO. 8,073,681

Mail Stop PATENT BOARD
Patent Trial and Appeal Board
U.S. Patent & Trademark Office
P.O. Box 1450
Alexandria, VA 22313-1450

TABLE OF CONTENTS

I.	Grounds for Standing.....	1
II.	Identification of Challenge	1
	A. Prior Art.....	1
	B. Grounds For Challenge	2
III.	'681 Patent	2
	A. Overview	2
	B. Level of Ordinary Skill	3
	C. Prosecution History	4
	D. IPR2020-01367	4
	E. Claim Construction	5
IV.	GROUND 1: SmartKom-Kobsa Combination Renders Claims 1-9, 11-21, 23-33, 35-36 Obvious.....	5
	A. Combination Overview	5
	1. SmartKom	5
	2. Kobsa.....	7
	3. Motivation to Combine	8
	B. Independent Claims 1/13/25.....	10
	1. Preambles [1P]/[13P]/[25P]	10
	2. Limitations [1A]/[13A]/[25A]	12
	3. Limitations [1D.1]/[13D]/[25B]	15
	4. Short-Term and Long-Term Shared Knowledge	17
	a) Limitations [1B.1]/[13B.1]/[25C.1]	20

b)	Limitations [1C]/[13C]/[25D]	22
5.	Limitation [1D]	23
a)	Limitations [1D.1]/[13D]/[25E]	27
b)	Establishing The Intended Meaning [1D.2]- [1D.3]/[13E]-[13F]/[25F]	36
6.	Limitations [1E.1]-[1E.2]/[13G.1]-[13G.2]/[25G]	38
C.	Dependent Claims	43
1.	Claims 2/14/26	43
2.	Claims 3/15/27	44
3.	Claims 4/16/28	46
4.	Claims 5/17/29	48
a)	Limitations [5A]/[17A]/[29A]	48
b)	Limitations [5B]/[17B]/[29B]	50
5.	Claims 6-7/18-19/30-31	52
6.	Claims 8-9/20-21/32-33	56
7.	Claims 11-12/23-24/35-36	58
V.	GROUND 2: Barbara-Ross Combination Renders Obvious Claims 1, 5-8, 10-13, 17-20, 22-25, 29-32, 34-36	61
A.	Combination Overview	61
1.	Barbara	61
2.	Ross	63
3.	Motivation to Combine	65
B.	Independent Claims	67
1.	Preambles [1P]/[13P]/[25P]	67

2.	Receiving An Utterance [1A]/[13A]/[25A]	68
3.	Conversational Speech Engine [1D.1]/[13D]/[25B]	70
4.	Short-Term Shared Knowledge [1B.1]-[1B.2]/[13B.1]- [13B.2]/[25C.1]-[25C.2]	72
5.	Long-Term Shared Knowledge [1C]/[13C]/[25D]	73
6.	Determining An Intended Meaning [1D]	74
a)	Identifying A Context [1D.1]/[13D]/[25E]	74
b)	Establishing The Intended Meaning [1D.2]- [1D.3]/[13E]-[13F]/[25F]	75
7.	Generating A Response [1E.1]-[1E.2]/[13G.1]- [13G.2]/[25G]	76
C.	Claims 5/17/29	78
1.	Conversational Goal [5A]/[17A]/[29A]	78
2.	Conversation Type [5B]/[17B]/[29B]	79
D.	Claims 6-7/18-19/30-31	80
E.	Claims 8/20/32	82
F.	Claims 11-12/23-24/35-36	83
VI.	GROUND 3: Barabara-Ross-O'Neill Combination Renders Claims 2- 4, 9, 14-16, 21, 26-28, and 33 obvious.	85
A.	Combination Overview	85
1.	O'Neill	85
2.	Motivation To Combine	86
B.	Claims 2/14/26	88
C.	Claims 3/15/27	89
D.	Claims 4/16/28	89

E.	Claims 9/21/33	91
VII.	GROUND 4: Barbara-Ross-Franco Combination Renders Obvious Claims 10, 22, 34	91
VIII.	Stipulation.....	95
IX.	Mandatory Notices.....	96
A.	Real Party In Interest.....	96
B.	Related Matters.....	96
C.	Notice of Counsel and Service Information.....	96
X.	Conclusion	97

EXHIBIT LIST

Exh.	Reference
1001	U.S. Patent 8,073,681 to Freeman et al. (“the ’681 patent”)
1002	Prosecution History for U.S. Patent 8,073,681 (“’681 PH”)
1003	Declaration of Stuart Lipoff in Support of Inter Partes Review of U.S. Patent 8,073,681 (“Lipoff Declaration”)
1004	Curriculum Vitae of Stuart Lipoff (“Lipoff CV”)
1005	<i>SmartKom: Foundations of Multimodal Dialogue Systems</i> , Springer Science + Business Media, Inc. (Wolfgang Wahlster ed., 2006) (“SmartKom”)
1006	<i>User Models in Dialog Systems</i> , Springer-Verlag (Kobsa & Wahlster eds., 1989) (“Kobsa”)
1007	U.S. Publication 2004/0101198 to Barbara (“Barbara”)
1008	U.S. Publication 2002/0173960 to Ross et al. (“Ross”)
1009	U.S. Publication 2005/0278180 to O’Neill & Hanna (“O’Neill”)
1010	U.S. Patent Application Publication 2010/0125458 to Franco et al. (“Franco”)
1011	Susann Luperfoy, <i>Discourse PEGS: A Computational Analysis of Context-Dependent Referring Expressions</i> (Dec. 1991) (Ph.D. Dissertation, University of Texas at Austin) (“Luperfoy”)
1012	John McCarthy, <i>Notes on Formalizing Context</i> , Proceedings of the 13th International Joint Conference on Artificial Intelligence, Vol. 1 (1993) (“McCarthy”)
1013	Ian Pitt & Alistair Edwards, <i>Design of Speech-Based Devices</i> , Springer-Verlag (2003) (“Pitt”)

Exh.	Reference
1014	Robert Porzel & Iryna Gurevych, <i>Contextual Coherence in Natural Language Processing</i> , 2680 Lecture Notes in Computer Science 272 (“Porzel”)
1015	Abraham Silberschatz & Peter B. Galvin, <i>Operating System Concepts</i> , Addison-Wesley (1994) (“Silberschatz”)
1016	Niels Ole Bernsen et al., <i>Designing Interactive Speech Systems: From First Ideas to User Testing</i> , Springer (1998) (“Bernsen”)
1017	H.P. Grice, <i>Logic and Conversation in 3</i> Syntax and Semantics 41, Academic Press (1975) (“Grice”)
1018	Xuedong Huang et al., <i>Spoken Language Processing</i> , Prentice Hall PTR (2001) (“Huang”)
1019	Lea Krause & Piek Vossen, <i>The Gricean Maxims in NLP – A Survey</i> , Proceedings of the 17th International Natural Language Generation Conference (2024) (“Krause”)
1020	Michael F. McTear, <i>Spoken Dialogue Technology: Toward the Conversational User Interface</i> , Springer (2002)
1021	Sebastian Möller, <i>Quality of Telephone-Based Spoken Dialogue Systems</i> , Springer Science + Business Media (2005) (“Möller”)
1022	Susan E. Brennan, <i>The Multimedia Articulation of Answers in a Natural Language Database Query System</i> , Second Conference on Applied Natural Language Processing, (1988) (“Brennan”)
1023	Reserved
1024	Memorandum Order Regarding Claim Construction from <i>VB Assets, LLC v. Amazon.com Servs. LLC</i> , No. 1:19-cv-01410-MN (D. Del., filed July 29, 2019)
1025	Second Amended Joint Claim Construction Chart from <i>VB Assets, LLC v. Amazon.com Servs. LLC</i> , No. 1:19-cv-01410-MN (D. Del., filed July 29, 2019)

Exh.	Reference
1026	Andrew Kehler, <i>The Effect of Establishing Coherence in Ellipsis and Anaphora Resolution</i> , 31st Annual Meeting of the Association for Computational Linguistics (1993) (“Kehler”)
1027	Trial Transcript, Volume 2 from <i>VB Assets, LLC v. Amazon.com Servs. LLC</i> , No. 1:19-cv-01410-MN (D. Del., filed July 29, 2019)
1028	<i>Amazon.com et al. v. VB Assets LLC</i> , IPR2020-01367, Paper 1 (PTAB July 28, 2020).
1029	<i>Amazon.com et al. v. VB Assets LLC</i> , IPR2020-01367, Paper 13 (PTAB May 21, 2021).
1030	<i>Amazon.com et al. v. VB Assets LLC</i> , IPR2020-01367, Paper 27 (PTAB March 7, 2022).
1031	Declaration of June Munford

Samsung Electronics Co. Ltd. and Samsung Electronics America, Inc.

(“Petitioners”) petition for *inter partes* review of claims 1-36 of U.S. Patent 8,073,681 (EX-1001).

I. Grounds for Standing

Petitioners certify the ’681 patent is available for IPR; Petitioners are not barred or estopped.

II. Identification of Challenge

A. Prior Art

The ’681 patent was filed on October 16, 2006. Each applied reference was filed or published before this date.

1. **“*SmartKom: Foundations of Multimodal Dialogue Systems*”**, (“SmartKom”; EX-1005) is prior art under 35 U.S.C. §102(a) because it was publicly available by August 4, 2006. (EX-1031, ¶¶6-9.)
2. **“*User Models in Dialog Systems*”, Kobsa & Wahlster eds.**, (“Kobsa”; EX-1006) is prior art under 35 U.S.C. §102(b) because it was publicly available by December 5, 1989. (EX-1031, ¶¶10-13.)
3. **U.S. Publication 2004/0101198 to Barbara** (“Barbara”; EX-1007), published May 27, 2004, is prior art under 35 U.S.C. §102(b).
4. **U.S. Publication 2002/0173960 to Ross** (“Ross”; EX-1008), published November 21, 2002, is prior art under 35 U.S.C. §102(b).

5. **U.S. Publication 2005/0278180 to O’Neill** (“O’Neill”; EX-1009), published December 15, 2005, is prior art under 35 U.S.C. §102(a).

B. Grounds For Challenge

Ground		Claims	Prior Art
1	103	1-9, 11-21, 23-33, 35-36	SmartKom+Kobsa
2	103	1, 5-8, 11-13, 17-20, 23-25, 29-32, 35-36	Barbara+Ross
3	103	2-4, 9, 14-16, 21, 26-28, 33	Barbara+Ross+O’Neill
4	103	10, 22, 34	Barbara+Ross+Franco

III. ’681 Patent

A. Overview

The ’681 patent discloses “a cooperative conversational voice user interface” processes an utterance (input 105) at speech recognition engine 110 to generate preliminary interpretations which are provided to conversational speech engine 115 for further processing. (EX-1001, 7:7-9, 7:24-28, 7:37-40, Figure 1 (below).) Conversational speech engine 115 generates a response. (EX-1001, 7:43-46.)

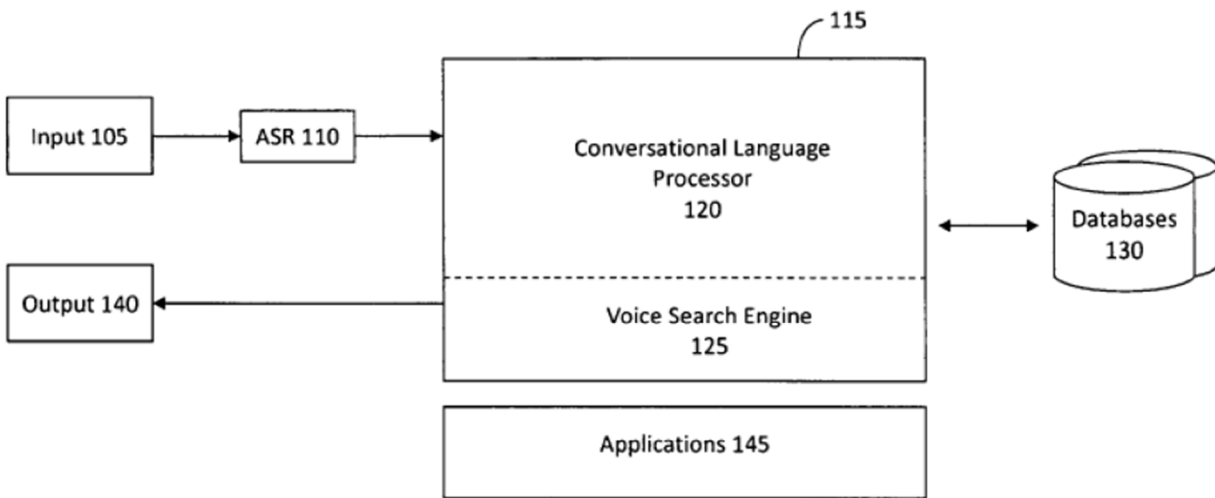


Figure 1

‘681 Patent, Figure 1

B. Level of Ordinary Skill

A person of ordinary skill in the art (“POSITA”) would have had a bachelor’s degree in computer science, computer engineering, electrical engineering, or related field in computing technology, and two years of experience with automatic speech recognition and natural language understanding, or equivalent education, research experience, or knowledge.¹ (EX-1003, ¶109.)

¹ The Board adopted this definition, which Patent Owner (“PO”) did not dispute, in IPR2020-01367, challenging the ’681 patent. (EX-1030, 8.)

C. Prosecution History

To overcome rejections, Applicants amended the claims and argued the art did not determine “an intended meaning for an utterance.” (EX-1002, 232.) Specifically, Applicants highlighted the claimed short-term shared knowledge and long-term shared knowledge used to create “profiles [that] may be particularly meaningful when attempting to disambiguate between contexts where a word has different meanings in different contexts.” (EX-1002, 233) (alterations in original).) The Examiner subsequently allowed the claims. (EX-1002, 923.)

D. IPR2020-01367

After institution of IPR2020-01367 challenging all '681 patent claims, PO disclaimed claims 37-42. The disclaimed claims have substantially overlapping limitations with independent claims 1/13/25 with the primary difference being the disclaimed claims were directed to determining an utterance contains insufficient information whereas claims 1/13/25 disambiguate an intent of an utterance having one or more words having different meanings in difference contexts. (EX-1003, ¶¶114-115.) In the final written decision, the Board determined the applied prior art failed to disclose the utterance includes “one or more words” that “have different meanings in different contexts.” (*See* EX-1030, 20-21.)

However, this limitation is directed to the well-known concept of “sense ambiguity.” (*See, e.g.*, EX-1003, ¶83.) Petitioners establish both SmartKom and Barbara explicitly disclose this limitation.

E. Claim Construction

In *VB Assets, LLC v. Amazon.com Servs. LLC*, No. 1:19-cv-01410-MN, the parties agreed the term “speech recognition engine” (claims 7/10/19/22/31/24) means “software or hardware that recognizes the words or phrases in the natural language utterance.” (EX-1024, 2.) Petitioners applied this construction.

Petitioner does not believe any other term requires construction to resolve patentability in this proceeding. (*See* EX-1003, ¶¶118-119.)

IV. GROUND 1: SmartKom-Kobsa Combination Renders Claims 1-9, 11-21, 23-33, 35-36² Obvious.

A. Combination Overview

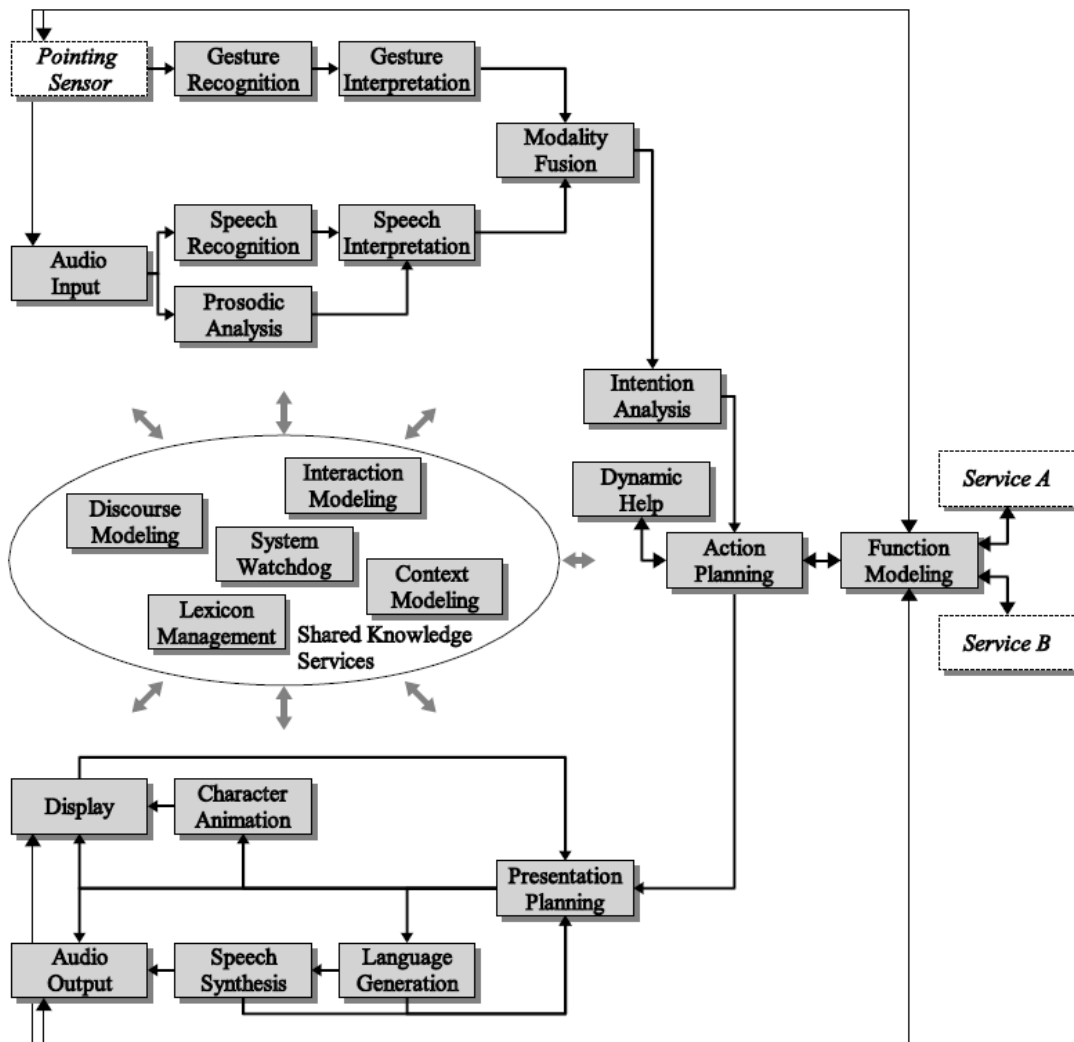
1. SmartKom

SmartKom “provides a comprehensive overview” of the SmartKom system and research results developed as part of the SmartKom initiative, begun in spring 1999. (EX-1005, VI, 31-32.) The SmartKom system “provides full symmetric multimodality in a mixed-initiative dialogue system with an embodied conversational agent.” (EX-1005, 3.) The generic software architecture provides a

² A listing of challenged claims is provided in the Appendix.

set of functional components, each represent a “processing unit[] of the executing system.” (EX-1005, 59, Figure 2 (60)³(below).) SmartKom describes three exemplary application scenarios SmartKom-Public (EX-1005, 63-64), SmartKom-Home (EX-1005, 65-66) and SmartKom-Mobile (EX-1005, 66-67)) derived from this generic architecture. The shaded components indicate components “reused in all application cases.” (EX-1005, 59.)

³ Because SmartKom is a series of papers, Figure numbers are reused. Petitioner therefore provides the page number in the Figure label.



SmartKom, Figure 2(60)

2. Kobsa

Kobsa “provide[s] a rather coherent survey of the field of user modeling.” (EX-1006, V.) Kobsa describes “user models” which are each a knowledge source “contain[ing] explicit assumptions of all aspects of the user that may be relevant to the dialog behavior of the system.” (EX-1006, 6.) Kobsa presents several exemplary user models storing user information (e.g., user domain expertise) and

describes user modeling components which “incrementally construct a user model,” “store, update and delete entries” and “supply other components of the system with assumptions about the user.” (EX-1006, 6.)

3. Motivation to Combine

A POSITA would have been motivated to combine Kobsa’s teachings regarding user models and user modeling with SmartKom’s dialogue system. (EX-1003, ¶¶126-131.) Kobsa is in the same field as the SmartKom and the ’681 patent—speech recognition systems. (EX-1001, 1:7-8; EX-1005, 3 (SmartKom “represents a new generation of multimodal dialogue systems”); EX-1006, 4 (describing user models “in natural-language dialog systems”); EX-1003, ¶126.)

SmartKom explicitly motivates the combination, mentioning use of a user model and/or user profiles/preferences in its context modeling and presentation planning. (*See, e.g.*, EX-1005, 16 (use of user model by the presentation planner), 274 (use of user model in context modeling), 407 (“Given an appropriate user model, personal preferences ... can be used.”); EX-1003, ¶127.) SmartKom also mentions storing preferences stated by a user during a conversation. (*See* EX-1005, 336.) However, SmartKom provides limited additional details regarding the content of the user models and techniques for user modeling. Accordingly, based on the suggestions in SmartKom, a POSITA would have been motivated to search for references that describe user models and would have been led to Kobsa which

provides a “coherent survey” of the topic. (EX-1003, ¶128; EX-1006, V.) Kobsa is co-edited by Wolfgang Wahlster who edited SmartKom and is Scientific Director of the SmartKom project, further leading a POSITA to Kobsa for further details of user models. (*Id.*)

Kobsa also motivates the combination describing benefits of user models for “interact[ing] with people in an intelligent and cooperative manner.” (EX-1006, 411, EX-1003, ¶129.) Kobsa stresses that user models allow the system to “tailor[] object descriptions to the user’s level of expertise” and “adapt[] an expert system’s response behavior to the background knowledge of its users.” (EX-1006, 196, 416.) A POSITA would have therefore been motivated to add Kobsa’s teachings into SmartKom to improve the user experience by providing user-specific and tailored responses. (EX-1003, ¶129.)

The combination is also nothing more than use of a known technique (Kobsa’s user models and user modeling) to improve similar devices (SmartKom’s dialogue system) in the same way (providing any additional knowledge source of interpretation and response generation). *KSR Int’l Co. v. Teleflex Inc.*, 550 U.S. 398, 417 (2007); (EX-1003, ¶130.).

A POSITA would have had a reasonable expectation of success and the results of the combination would have been predictable because SmartKom uses a standard software architecture and standard storage structures. (*See* EX-1005,

Figure 2(60); EX-1003, ¶131.) Kobsa’s user model and user modeling components are also merely storage and software constructs. A POSITA would have therefore been able to implement Kobsa’s user models/modeling based on the teachings of SmartKom with predictable results. (EX-1003, ¶131.)

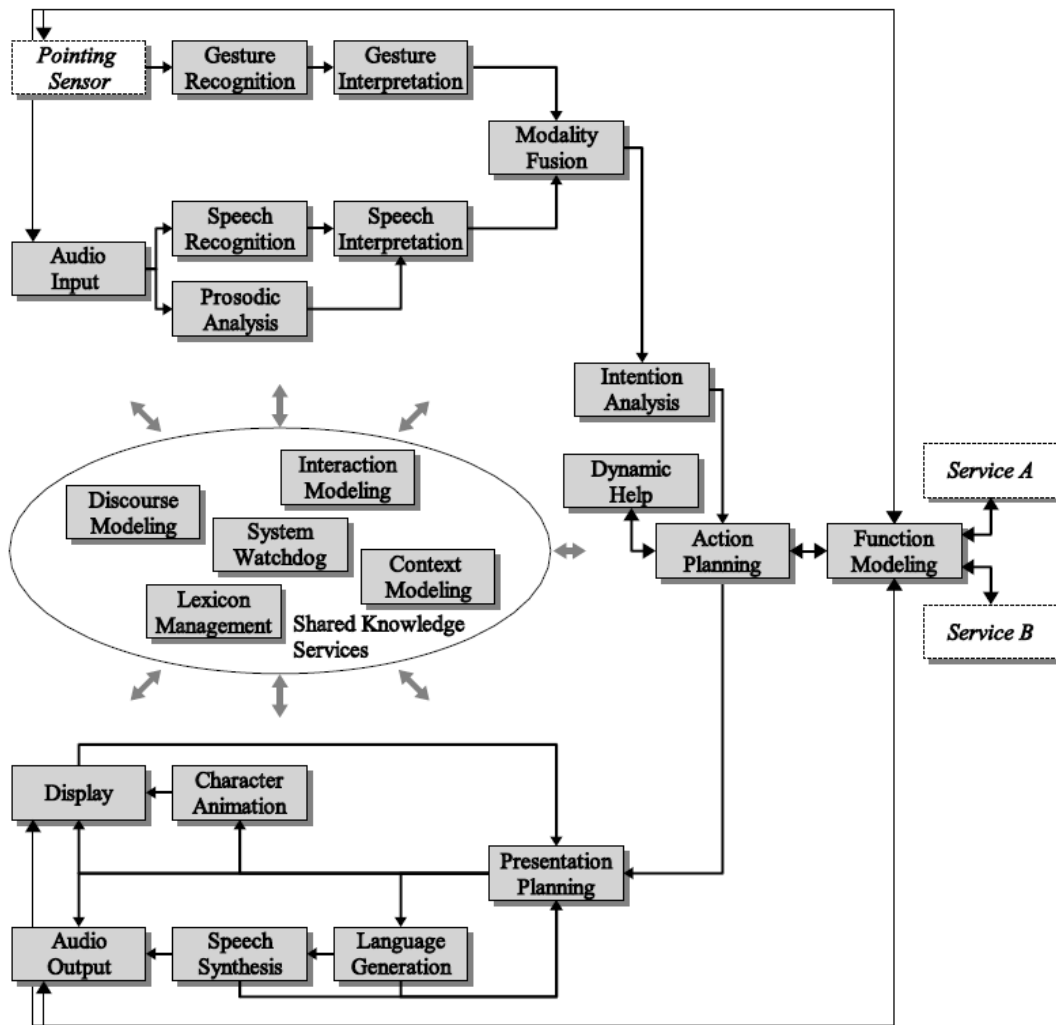
B. Independent Claims 1/13/25

1. Preambles [1P]/[13P]/[25P]

SmartKom discloses a “*system*”⁴ [25P], the SmartKom system which “represents a new generation of multimodal dialogue systems.” (EX-1005, 3; EX-1003, ¶133.)

The generic software architecture of SmartKom is illustrated in Figure 2(60) below. (EX-1005, 60.) This software provides a “*method*” [1P] performing the claim 1 functions. (EX-1003, ¶134.)

⁴ Claim language is indicated by italics throughout.



SmartKom, Figure 2(60)

SmartKom discloses implementations executing on a server having “main memory” and processor(s). (See EX-1005, 13, 440.) The main memory or a similar computer readable medium would necessarily store the SmartKom software providing SmartKom functionality. (EX-1003, ¶¶135-136.) Thus, although not explicit, a POSITA would have known, based on general knowledge in the field, that SmartKom’s software components include “computer-executable

instructions”⁵ stored in a “*non-transitory computer readable medium*” and executed by a processor. (EX-1003, ¶¶135-136, *citing* EX-1015, 29-30, 37.)

The SmartKom system provides a “*voice user interface*.” (EX-1003, ¶137.) Specifically, the system includes, e.g., a speech recognition component. (EX-1005, 167, *see also*, 85, Figure 2(60).)

The SmartKom system is “*cooperative*” and “*conversational*.” (EX-1003, ¶138.) The SmartKom system supports “**cooperative** dialogues with mixed initiative” using a “conversational agent.”⁶ (EX-1005, 8, 381 (referring to “cooperative communication”), 6.)

Accordingly, SmartKom discloses a “*method*” [1P], “*non-transitory computer readable medium*” [13P], and “*system*” [25P], each “*for providing a cooperative conversational voice user interface*.” (EX-1003, ¶¶133-139.)

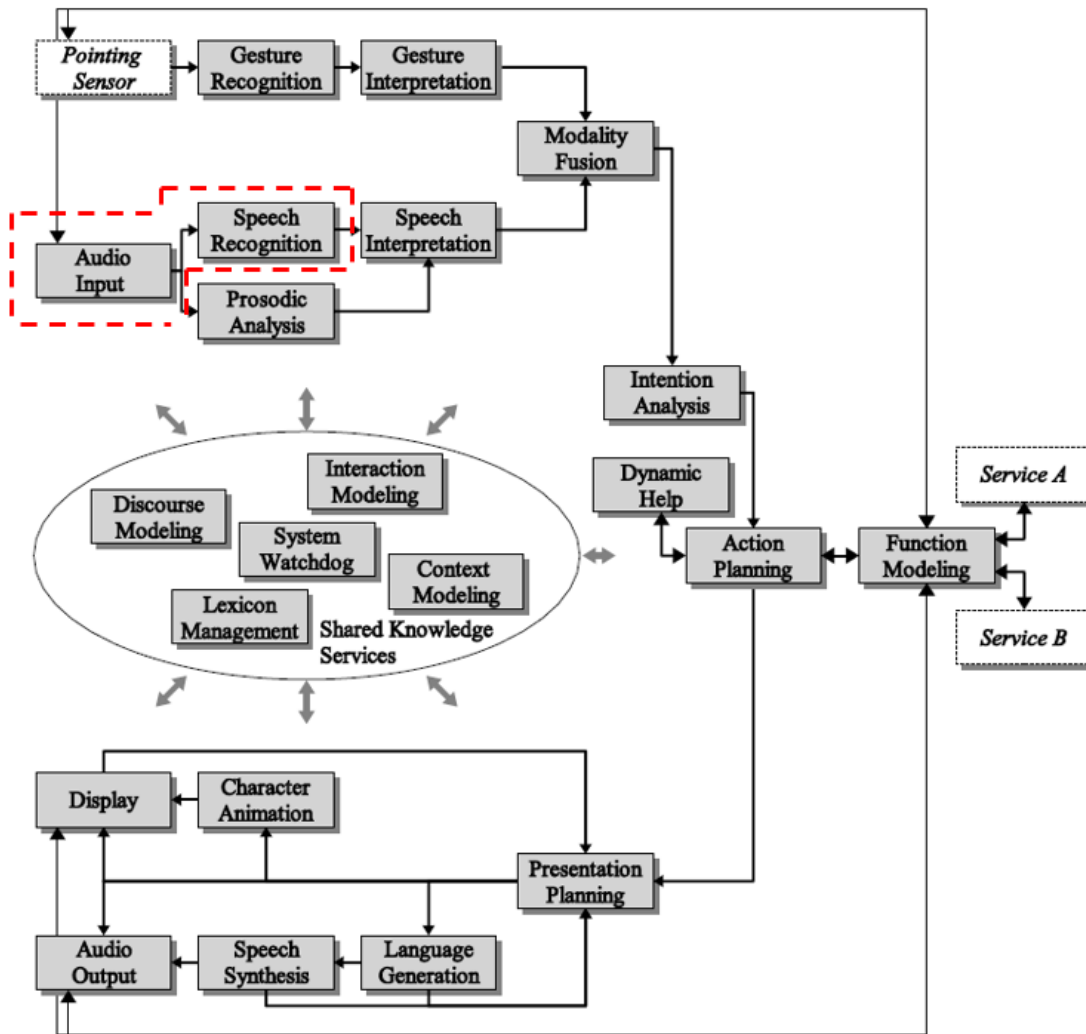
2. Limitations [1A]/[13A]/[25A]

SmartKom discloses “*a voice input device*” that “*receiv[es] an utterance*” “*during a current conversation with a user*” [1A.1]/[13A.1]/[25A.1]-[25A.2] (EX-1003, ¶142.)

⁵ The ’681 patent does not use the term “*instructions*.” At most, it mentions a “conversational language **processor**.” (See EX-1001, 7:40-43.)

⁶ Unless otherwise noted, all emphasis added.

The SmartKom system includes an audio input (e.g., a microphone) that receives voice (speech) input from a user and provides an acoustic input signal to a speech recognition component that “transforms the acoustic input signal into an orthographic transcription representing the utterance of the speaker.” (See EX-1005, 85, 440, 457, Figure 2(60).) The audio input component is “*a voice input device*” that “*receive[s] an utterance*” from a user “*during a current conversation*” between the system and the user. (EX-1003, ¶¶140-141.) Additionally or alternatively, because the speech recognizer receives the acoustic input voice signal, it is also a “*voice input device*.” (EX-1003, ¶142.)



SmartKom, Figure 2(60)

SmartKom's system accepts and handles user utterances that *"include[] one or more words that have different meanings in different contexts"* [1A.2]/[13A.2]/[25A.3]. (EX-1003, ¶¶143-151.) This limitation is directed to an ambiguity common in natural language processing referred to as "sense ambiguity." (EX-1003, ¶¶143-148; EX-1020, 91, EX-1013, 62, 161.) SmartKom's components, including its speech interpretation and intention analysis components,

handle a wide range of ambiguities and imprecisions common in spoken dialog through its contextual coherence processing, described further in §IV.B.5. One described example is the handling of the German verb “kommen” which can mean “to come” or “showing on” depending on the context. (EX-1005, 607; EX-1014, 281.) That is, “kommen” is a word hving different meanings in different contexts. Porzel, a companion article drafted by the same authors as the relevant SmartKom chapter and cited therein (EX-1005, 269, 283), provides two example input utterances using “kommen”:

U1: *was für Spielfilme kommen heute Abend*
what for films come today evening

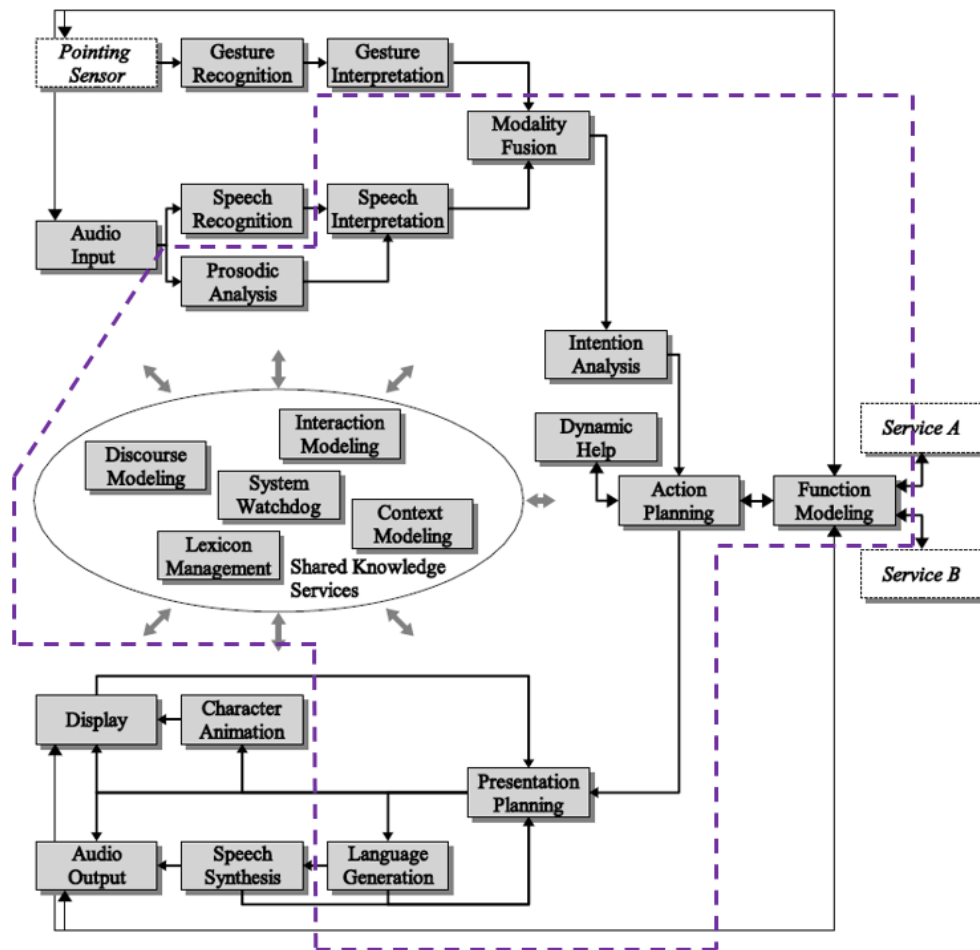
U2: *wie kann ich mir zur Schloss kommen*
how can I me to castle come

(EX-1014, 281.)

3. Limitations [1D.1]/[13D]/[25B]

SmartKom discloses “*a conversational speech engine*”, recited in limitations [25B]/[1D.1]/[13D]. (EX-1003, ¶¶152-156.) In the ’681 patent, the conversational speech processor includes a set of modules that process output from a speech recognition component and generate a response. (See EX-1001, 7:37-40, Figures 1-2.) The SmartKom system uses a similar modular architecture. (See EX-1005, Figure 2(60)(below).) Specifically, the SmartKom system includes components

that process the speech recognition engine output including the speech interpretation, modality fusion, and intention analysis components and components participating in response generation including the action planning, presentation planning, and language generation components. (EX-1003, ¶154.) These components leverage the shared knowledge service components. (*Id.*) Thus, SmartKom discloses components (e.g., within purple box below) that collectively form “a conversational speech engine” [1D.1]/[13D]/[25B].



SmartKom, Figure 2(60)

A POSITA would have generally understood software executes on a processor. (EX-1003, ¶155.) SmartKom discloses an implementation where its software components execute on a server “consist[ing] of 3 dual Xeon 2.8 GHz processor.” (EX-1005, 13, *see also*, Figure 1(440) (showing a server with an Intel Xeon dual-processor).) Thus, SmartKom discloses the “*conversational speech engine includes one or more processors*” [25B]. (EX-1003, ¶155.)

4. Short-Term and Long-Term Shared Knowledge

The SmartKom-Kobsa combination discloses use of “*short-term shared knowledge*” that “*includes knowledge about the utterance received [at the voice] during the current conversation*” [1B.2]/[13B.2]/[25C.2] and “*long-term shared knowledge*” that “*includes knowledge about one or more past conversations with the user*” [1C.2]/[13C.2]/[25D.2]. (EX-1003, ¶¶157-170.)

It was well-known, prior to the ’681 patent, that interactive dialogue systems used “information about the dialogue history and the user” “to constrain how the system interprets the user’s subsequent utterances and to determine what the system should say and how it should be said.” (EX-1020, 124.) These processes involve “representations of discourse structure, of intentions, goals and beliefs, and of dialogue as a collaborative activity.” (EX-1020, 124.) SmartKom-Kobsa similarly integrates “*short-term*” and “*long-term*” knowledge to enrich intention hypotheses associated with potential interpretations of an utterance with contextual

information and to provide a contextual coherence score for each. (*See, e.g.*, §IV.B.5.)

Specifically, SmartKom’s architecture provides “**shared** knowledge services” including modeling components that “enable[] the system to act analogously, i.e., to provide hitherto implicit knowledge concerning what is talked about.” (*See, e.g.*, EX-1005, 277-278, Figure 2(60).) The specific knowledge stores and their associated contexts are illustrated in SmartKom’s Table 1(274) (below), that provides a “broad categorization of the types of context relevant to spoken dialogue systems, their content and respective knowledge stores.” (EX-1005, 274; EX-1014, 273-274.)

Table 1. Contexts, content and knowledge sources

Types of context	Content	Knowledge store
Dialogical context	What has been said by whom	Dialogue model
Ontological Context	World/conceptual knowledge	Domain model
Situational context	Time, place, etc.	Situation model
Interlocutionary context	Properties of the interlocutors	User model

SmartKom, Table 1(274)

The dialogue model contains “what has been said by whom” and is associated with dialogical (discussion/conversation) context. (*See* EX-1005, Table 1(274).) That is, the dialogue model includes the dialogue history which is a type of “*short-term shared knowledge*” that “*includes knowledge about the utterance*

received” at the voice input “*during the current conversation*” consistent with the ’681 patent’s disclosure. (EX-1003, ¶162; EX-1001, 4:53-55 (short-term knowledge includes “input received during a single conversation”).) In SmartKom, the dialog history is referred to as the “discourse state” and is stored in discourse memory. (EX-1005, 239, 242, 246.)

The situation model, which records “time, place” of an utterance is also a type of “*short-term shared knowledge*” that “*includes knowledge about the utterance received*” (i.e., the time and location of the utterance), consistent with the ’681 patent’s disclosure. (EX-1003, ¶163, EX-1001, 4:55-60 (indicating short-term knowledge includes “location” data).)

The SmartKom system also stores the user state during a conversation which is “an extension of the well-known term of emotion with some internal states of a human like, e.g., ‘*hesitant*’, that are important in the context of human computer interaction (HCI).” (EX-1005, 139 (emphasis in original).) User state is a type of “*short-term shared knowledge*” that “*includes knowledge about the utterance received*.” (EX-1003, ¶164.)

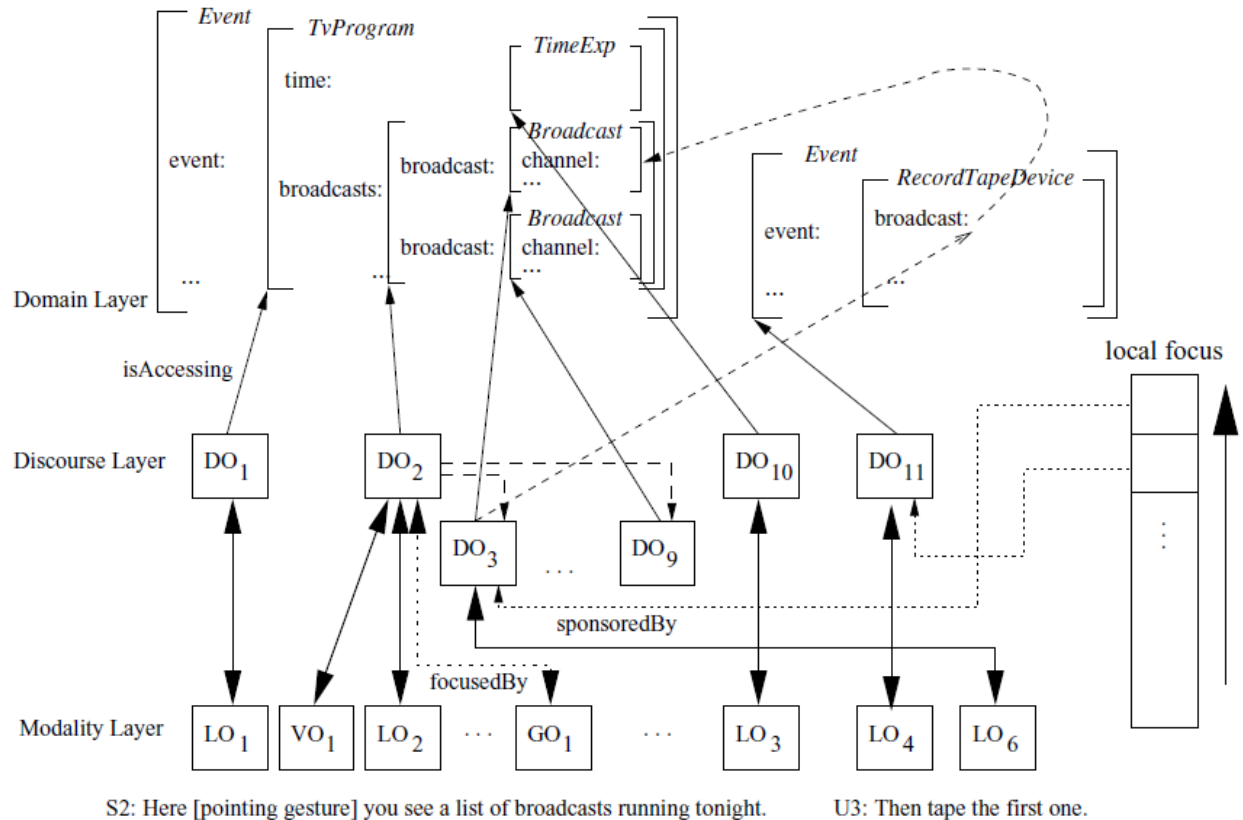
The domain model “hold[s] knowledge of the world that is talked about.” (EX-1005, 274.) The domain model is a type of “*long-term shared knowledge*.” (EX-1003, ¶165.)

SmartKom's Table 1(274) also includes the "user model" storing the "properties of the interlocutors" (i.e., user information) and providing "interlocutionary" (or "user") context. (EX-1005, 274.) As explained by Kobsa, "long-term" user models "describe relatively static characteristics of users" such as "his/her knowledge about a domain" and the system's "assumptions about the user's experience with respect to a domain of discourse." (EX-1006, 5, 17, 29, 52, 199; *see also*, 6, 200-205, 411-413.) Because the long-term user model includes knowledge derived from a user's prior conversations as discussed in §IV.A.2, the long-term user model includes "*knowledge about one or more past conversations with the user.*" (EX-1003, ¶¶166-170; EX-1001, 5:4-14 (long-term shared knowledge includes "explicit and/or implicit user preferences").)

a) Limitations [1B.1]/[13B.1]/[25C.1]

SmartKom discloses "*accumulat[ing]/[e] short-term shared knowledge about the current conversation*" [1B.1]/[13B.2]/[25C] in at least two ways. (EX-1003, ¶¶171-176.) First, SmartKom accumulates dialog history in the discourse state stored in discourse memory used to enrich and score intention hypotheses (representations of the possible interpretations of an utterance). (*See* EX-1005, 237.) In SmartKom, the "context representation" used "for representing the discourse state" consists of three levels: the discourse object layer, the modality layer and the domain object layer." (EX-1005, 239, 242, Figure 2(243)(below).)

The “discourse object layer is the central layer of the discourse representation” and “comprises the concepts introduced into the discourse.” (EX-1005, 242.) The modality layer includes modality objects (MOs), including linguistic objects (LOs) linked to discourse objects. (EX-1005, 244-245.) The local focus structure provides “access to all discourse objects” during interpretation of later utterances. (EX-1005, 245.)



SmartKom, Figure 2(243)

The discourse model “is dynamically updated as system output progresses.” (EX-1005, 61.) That is, the “*short-term shared knowledge*” contained in the

discourse state stored in discourse memory is continuously accumulated during a conversation. (EX-1003, ¶173.)

Second, SmartKom accumulates user state knowledge during a conversation. (EX-1003, ¶¶174-175.) Specifically, the prosody (speech segment) module takes speech signals and word lattices (word hypothesis graphs) from the speech recognizer and generates a “user state lattice.” (EX-1005, 141; EX-1003, ¶175.)

For these reasons and those discussed in §IV.B.4, the SmartKom-Kobsa combination discloses [1B.1]-[1B.2]/[13B.1]-[13B.2]/[25C.1]-[25C.2].

b) Limitations [1C]/[13C]/[25D]

The SmartKom-Kobsa combination discloses “*accumulat[e]/[ing] long-term knowledge about the user*” [1C.1]/[13C.1]/[25D.1]. (EX-1003, ¶¶177-182.)

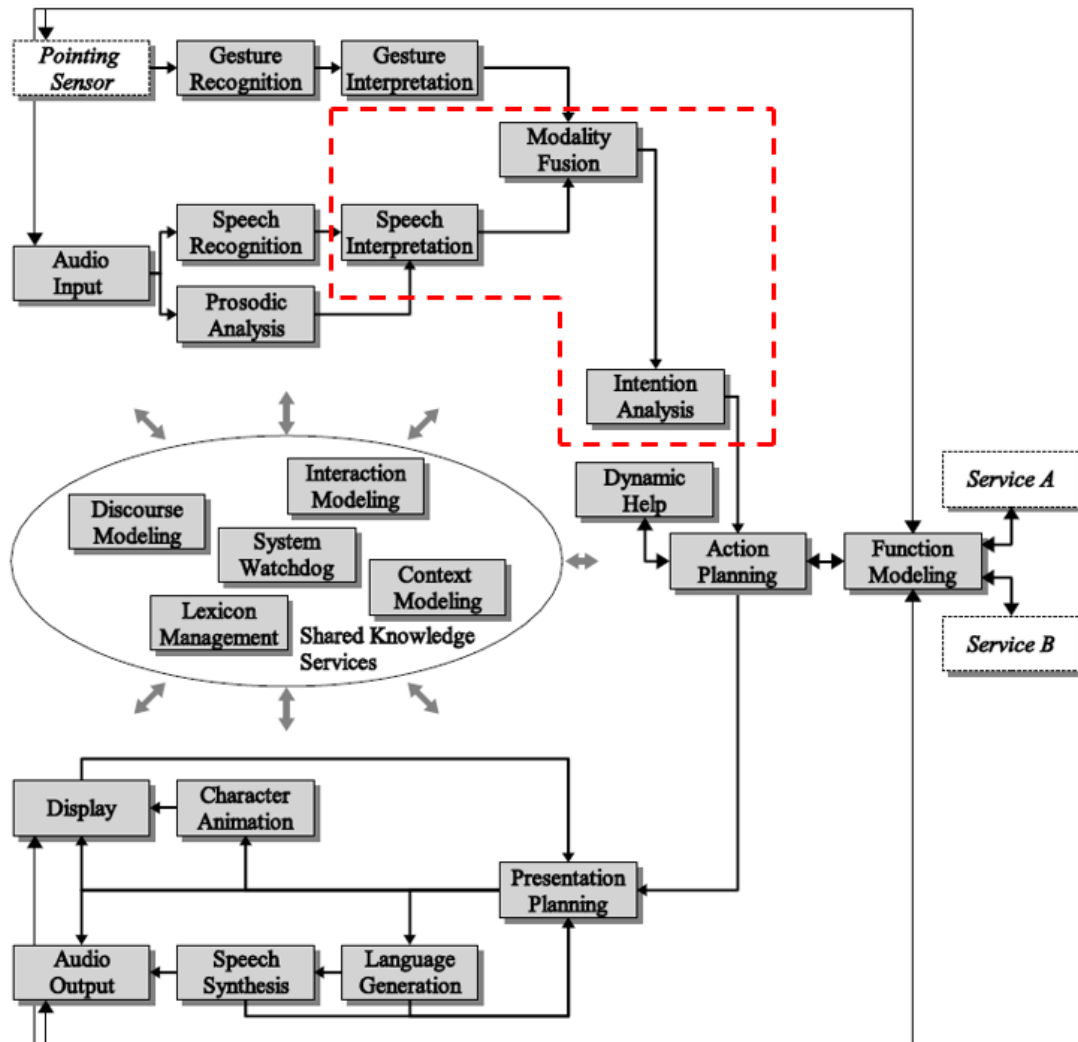
As I discussed in §IV.B.4, SmartKom’s user model provides “properties of the interlocutors” to support interlocutionary/user context. (EX-1005, 274.) The knowledge stored in the user model can be acquired (accumulated) “either *explicitly or implicitly*.” (EX-1006, 416 (emphasis in original).) For example, explicit knowledge can come from interviewing the user. (EX-1006, 416.) “Implicit acquisition involves ‘eavesdropping’ on the user-system interaction in order to observe the user’s behavior and from it to infer facts that go into the model.” (EX-1006, 416.) Accordingly, the stored user model “*includes knowledge about one or more past conversations with the user*.” (EX-1003, ¶178.)

Both SmartKom and Kobsa describe techniques for “*accumulating long-term shared knowledge*.” SmartKom describes an interaction model that “computes information on the user” from the system’s interpretation of an utterance. (EX-1005, 287.) Kobsa teaches the “*user modeling component* is the part of a dialog system whose function is to incrementally construct a user model; to store, update, and delete entries; to maintain the consistency of the model; and to supply other components of the system with assumptions about the user.” (EX-1006, 6 (emphasis in original).) As an example, Kobsa provides an overview of the General User Modeling System (“GUMS”) which “is designed for building long term models of individual users.” (EX-1006, 417.) Kobsa confirms the user model is persistent (long-term), teaching that, “at the end of a dialog session”, the system “records all information about the user inferred from his/her dialog behavior in a corresponding file.” (EX-1006, 10-11; EX-1003, ¶181.)

Thus, the SmartKom-Kobsa combination discloses [1C.1]-[1C.2]/[13C.1]-[13C.2]/[25D.1]-[25D.2].

5. Limitation [1D]

SmartKom discloses “*determining an intended meaning for the utterance*” [1D]. (EX-1003, ¶¶183-189.) SmartKom’s intention recognizer “has the task to finally rank the remaining interpretation hypotheses and **to select the most likely one**, which is then passed on to the action planner.” (EX-1005, 14-15.)

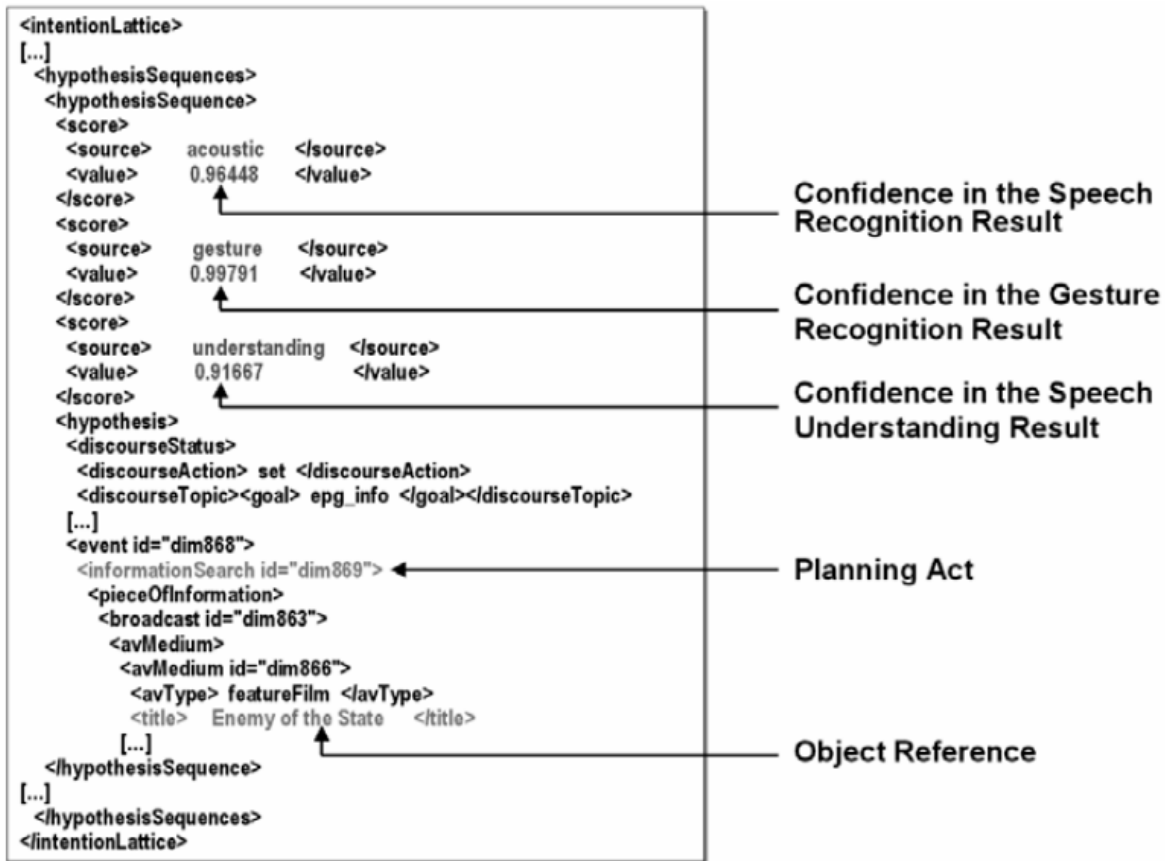


SmartKom, Figure 2(60)

The SmartKom modules transfer information through an “intention lattice,” containing a number of hypotheses that “stand for alternative readings of one and the same user input.” (EX-1005, 287-288.) In SmartKom, the “basic information flow from user input to system output continuously adds further processing results so that [this] representational structure [i.e., intention lattice] will be refined step-by-step.” (EX-1005, 62-63.) That is, the intention recognizer receives an intention

lattice from the modality fusion components including a set of hypotheses. (EX-1003, ¶184.)

The intention lattice shared between components contains one or more intention segments that each includes, e.g., a discourse object, goal manipulations (e.g., setting, retracting, or retaining a goal), slot manipulations (setting slots to establish new information, retracting to invalidate current information, or retaining slots to confirm information), annotation of the intention as positive or negative feedback (e.g., based on user state), and confidence scores from the analysis modules. (EX-1005, 306.) An exemplary intention lattice fragment is illustrated below. (EX-1005, Figure 8(15); EX-1003, ¶186.)



SmartKom, Figure 8(15)

In operation, the natural language processing (speech interpretation) module “transform[s] the word lattice sent by the speech recognizer into a list of hypotheses representing ... possible user intentions.” (EX-1005, 195.) The hypotheses, each with a corresponding score value (confidence in the speech recognition result), is provided to the modality fusion component which provides “integration and disambiguation of the different modalities.” (EX-1005, 204, 224.) Then, “a maximum of three hypotheses is sent to the intention analyzer”, each with a “fusion score.” (EX-1005, 233.) The intention recognizer engages a discourse

modeler that works in conjunction with a context modeler to enrich and validate/score the received hypotheses. (EX-1005, 286-287, Figure 1(286) (below).) A contextual coherence score for each hypothesis in the intention lattice is provided to the intention recognizer and used to select one hypothesis. (EX-1005, 285.) The selected hypothesis is considered the user's "*intended meaning for the utterance*" and therefore SmartKom discloses "*determining an intended meaning for the utterance*" [1D]. (EX-1003, ¶¶187-89.)

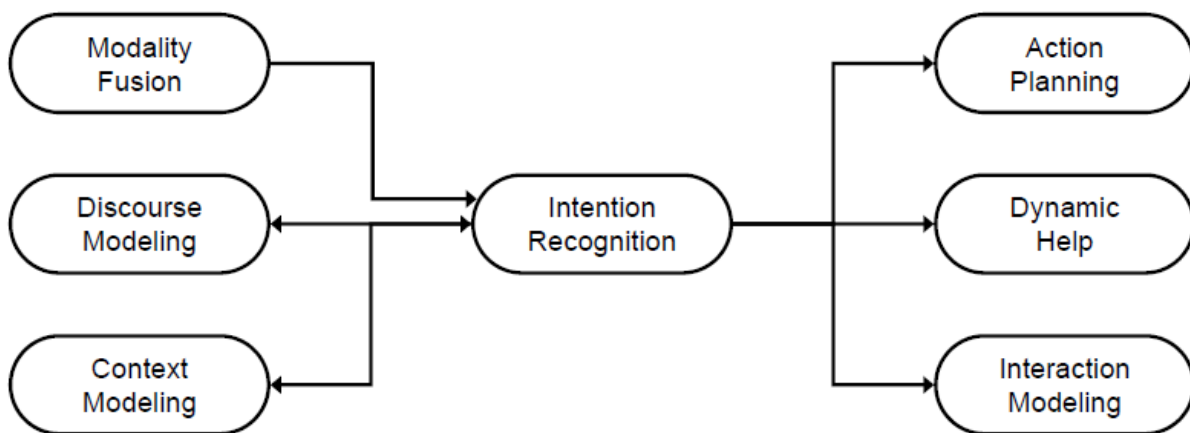


Fig. 1. Section of the multimodal interaction from the intention recognition point of view

SmartKom, Figure 1(286)

a) Limitations [1D.1]/[13D]/[25E]

The SmartKom-Kobsa combination discloses "*identify[ing]*" "*a context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge*" [1D.1]/[13D]/[25E]. (EX-1003, ¶¶190-203.) SmartKom notes that "[s]peakers may not always be aware of the potential ambiguities

inherent in their utterances” and “leave it to the context to disambiguate and specify the message.” (EX-1005, 275.) Specifically, “to interpret the utterance correctly, the addressee must employ several context-dependent resources.” (EX-1005, 275.)

In SmartKom, the intention recognizer employs “context-dependent resources” by engaging the discourse modeler and context modeler which work together “to enrich the information in the hypotheses with context knowledge.” (EX-1005, 287.) The “enrichments are evaluated by the discourse model and the context model to rate the quality of the augmentation of a hypothesis with knowledge from the discourse and the surrounding world.” (EX-1005, 287.)

As part of this process, “the discourse modeler receives a set of hypotheses” and enriches each “with previous discourse information” (discourse state; “*short-term shared knowledge*”) and context-domain information from the domain model (“*long-term shared knowledge*”). (EX-1005, 20, 238 (“each new user contribution has to be interpreted in the light of the previous discourse context”), 240 (discourse modeler “interprets [hypotheses] with respect to the current discourse context”).) Specifically, the discourse modeler associates domain-context information with the utterance through a typed feature structure (TFS). (EX-1005, 256.)

The following example illustrates this context identification. Here, “the user has requested information about tonight’s television program” (utterance-1). (EX-

1005, 262-263.) The TFS representation of utterance-1 (part of the discourse history in the discourse state) is shown in Figure 2(264).

U-1: *What's on TV tonight?*

SYS: (Displays the TV program listings) *Here you see a list of movies running tonight.*

U-2: *Is there a movie with Nicole Kidman?*

(EX-1005, 263.)

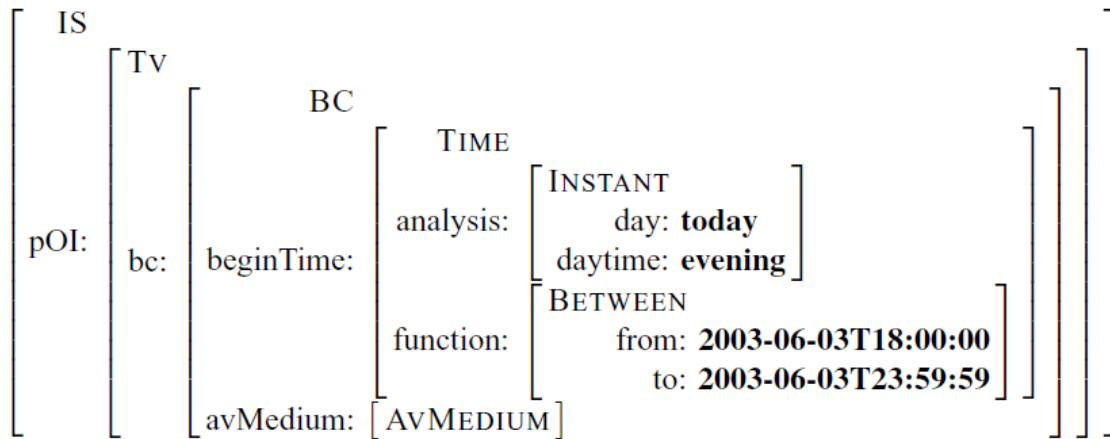


Fig. 2. TFS representation for the utterance *What's on TV tonight?* (*IS* = InformationSearch, *pOI* = pieceOfInformation, *Tv* = TvProgram, *bc* = broadcast)

SmartKom, Figure 2(264)

The user then “wants to reduce the number of displayed movies by uttering *Is there a movie with Nicole Kidman?*” (utterance-2). (EX-1005, 263.) “This user request, if interpreted separately, leads to at least two intention hypotheses— one where the user requests information about the current TV program” (U-2-hypoB)

and “one where the user requests information about the current cinema program” (U-2-hypoA). (EX-1005, 263.) The TFS for the first interpretation is shown in Figure 3(264). As shown, U-2-hypoB is associated with the context-domain of a cinema program. (EX-1005, 264.) The TFS for the second interpretation is shown in Figure 4(264). As shown, U-2-hypoA is associated with the context-domain of a broadcast TV performance. (EX-1005, 264.) That is, the discourse modeler enriches each with context-domain information. (EX-1003, ¶¶193-95.)

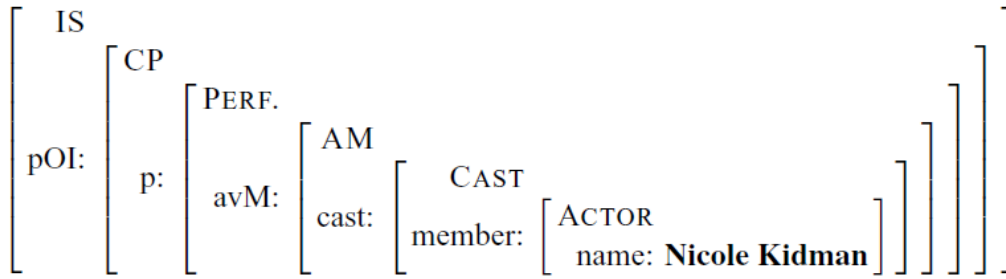


Fig. 3. One possible interpretation of the utterance *Is there a movie with Nicole Kidman?* (*IS* = InformationSearch, *pOI* = pieceOfInformation, *CP* = CinemaProgram, *p* = performance, *P* = Performance, *bc* = Broadcast, *Bc* = Broadcast)

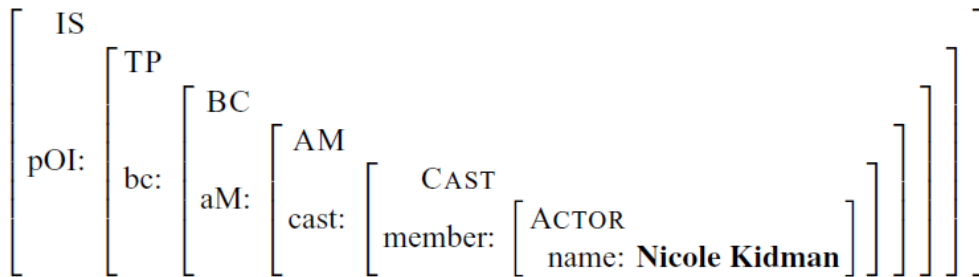


Fig. 4. A second possible interpretation of the utterance *Is there a movie with Nicole Kidman?* (*IS* = InformationSearch, *pOI* = pieceOfInformation, *CP* = CinemaProgram, *p* = performance, *P* = Performance, *aM* = avMedium, *AM* = AvMedium)

SmartKom, Figures 3-4(264)

Each of these hypotheses, enriched with context-domain information, is then overlay[ed] with the prior utterance (U-1) and scored. (EX-1005, 266.) Because U-1 is associated with TV context domain, the TV intention hypothesis U-2-hypoB is more likely and receives a higher score. (EX-1005, 266.) The interpretation/enrichment of U-2 with respect to the previous discourse context is shown in Figure 5(265). As shown, the hypothesis has been enriched with the

identified context based on the discourse history and context-domain (broadcast-TV). (EX-1003, ¶195.)

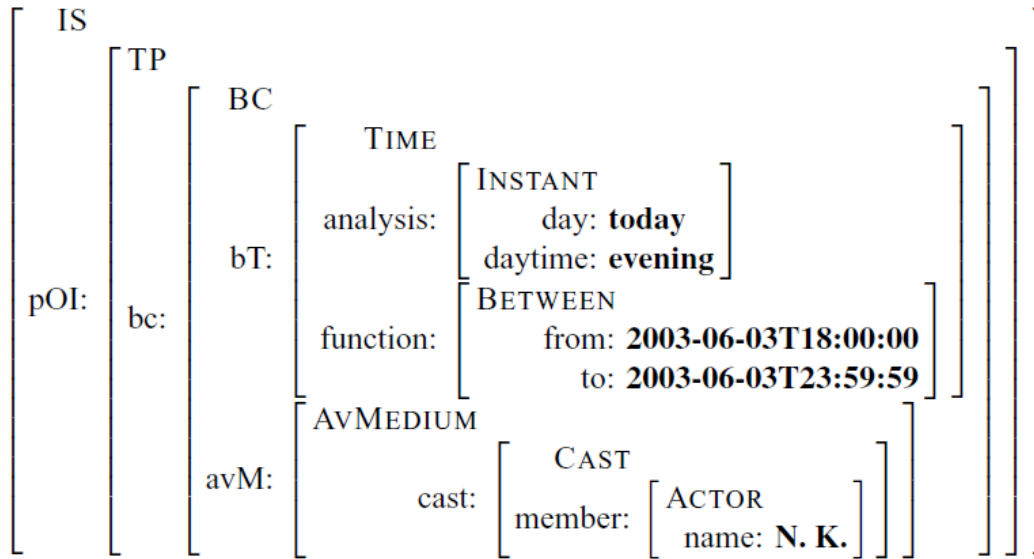


Fig. 5. Interpretation of the utterance *Is there a movie with Nicole Kidman?* with respect to the previous discourse context

SmartKom, Figure 5(265)

In addition to the above enrichment through discourse state and context-domain, the hypotheses are enriched with context from the context model. (EX-1005, 285, 287.) The SmartKom-Kobsa system uses this contextual knowledge for “**lexical and pragmatic disambiguation**, decontextualization [sic] of domain and common-sense knowledge that was left implicit by the user” and “for estimation of] an overall coherence score that is used in intention recognition.” (EX-1005,

269.) To perform these actions, SmartKom “link[s] the **context model to interfaces providing contextual information.**” (EX-1005, 277-278.)

Specifically, the context model uses a set of contexts, illustrated in Table 1(275), associated with a knowledge store (referred to as a “model”). (EX-1005, 274-275.) The “*context[s]*” provided by the model are used to enrich the “intention lattice” hypotheses through the discourse/context modelers, as discussed below. (EX-1005, 305-306.)

Table 1. Contexts, content and knowledge sources

Types of context	Content	Knowledge store
Dialogical context	What has been said by whom	Dialogue model
Ontological Context	World/conceptual knowledge	Domain model
Situational context	Time, place, etc.	Situation model
Interlocutionary context	Properties of the interlocutors	User model

SmartKom, Table 1(274)

The context model provides “context-specific insertions” to enrich each hypothesis (e.g., “provide hitherto implicit knowledge concerning what is talked about”). (EX-1005, 277.) For example, the “emotional state of the user” (i.e., the user state) which is part of the “*short-term shared knowledge*” is added. (EX-1003, ¶200.) Situational context, including location and time of the utterance, is also added, as shown in exemplary Table 2(277) below. In the combination of SmartKom-Kobsa, a hypothesis is enriched with “*long-term shared knowledge*”

from the user model such as the user’s preferences or expertise in the identified context-domain. (See §IV.B.4, EX-1005, 274.) These enrichments assist with interpretation of user intent as well as shaping response generation. (§IV.B.6.)

Table 2. Context-specific insertions into a sample intention hypothesis resulting from the interpretation of a speech recognition hypothesis

<pre> <informationSearchProcess> <entertainment> <performance> <cinema> <contact> <address> <town> here </town> </address> </contact> </cinema> <time> <beginTime> <at> now </at> </beginTime> </time> </performance> </entertainment> </informationSearchProcess> </pre>	<pre> <contact> <x> 70.345 </x> <y> 49.822 </y> <town> Heidelberg </town> </contact> <time> <at> 19:00:00T26:08:03 </at> </time> </pre>
	<pre> <scores> <contextualCoherence> 0.46 </contextualCoherence> </scores> </pre>

SmartKom, Table 2(277)

Thus, the context modeler of the SmartKom-Kobsa combination identifies a context based on dynamic situational knowledge (“*short-term shared knowledge*”), dynamic instances (e.g., position of user when providing current utterance; “*short-term shared knowledge*”), and common ground knowledge (“*long-term shared-*

knowledge”), and user-specific profile information (e.g., experience, preferences, etc.; “*long-term shared-knowledge*”) and incorporates this context with the discourse and context-domain context (e.g., television, cinema program, etc.) in each hypothesis. (EX-1003, ¶201; *see, e.g.*, EX-1005, 364 (“SmartKom domains such as electronic program guides (EPG) for TV’s, cinema programs, and movie information”), 21.)

The context modeler also “compute[s] [a] contextual coherence score[.]” (EX-1005, 276, 271.) The scoring task involves using the “ontological domain context to measure the semantic coherence of the individual interpretation” and “using dynamic situational and discourse information, e.g., previous ontological contexts of prior turns.” (EX-1005, 271.) For example, a companion paper by Porzel, explains that “a panel of experts” including a “*situation expert* observing, e.g., time, data, proximity, accessibility”, “*user expert* observing, e.g., interests, transportation, thrift”, “*discourse expert* observing, e.g., referential status, discourse accessibility,” and an “*ontological expert* observing, e.g., object types and object roles” vote and these votes are combined to obtain the contextual coherence score. (EX-1014, 283.) Table 2(277) above includes an example of the contextual coherence score added into hypothesis of the intention lattice.

SmartKom-Kobsa combination therefore discloses “*identify[ing]*” “*a context associated with the utterance* [context-domain, situational context, short-term user

state, long-term user context] *from the short-term shared knowledge* [user state, short-term user model, situational model] *and the long-term shared knowledge* [long-term user model, context-domain model]” [1D.1]/[13D]/[25E].

b) Establishing The Intended Meaning [1D.2]-[1D.3]/[13E]-[13F]/[25F]

In the SmartKom architecture, the intention recognizer establishes the intended meaning of an utterance within the identified context—i.e., disambiguates an intent that the user had in speaking the one or more words that have different meanings in different contexts. (EX-1003, ¶¶204-10.) As discussed in §IV.B.5.a, the intention recognizer engages the discourse and context modelers to enrich received intention hypotheses and provide associated contextual coherence scores. For example, the contextual domain (e.g., from the domain model), situational knowledge, and user profile/preference knowledge are used “for augmenting” the “intention hypotheses with implicit information” in the SmartKom-Kobsa system “to spell out their underlying intentions, and finally, to define a common background representation for the processed content, i.e., intention lattices.” (EX-1005, 271, 274.) Using the scores and enriched hypotheses, the intention recognizer “select[s] the interpretation of the user input which is considered to be the best matching one.” (EX-1005, 287.) Specifically, the intention recognizer includes a “probabilistic model” that “combines various scores, based on features

in the representation and computed by the SmartKom modules” to support its selection of the intention hypothesis (“*intended meaning*”). (EX-1005, 285.) Thus, SmartKom discloses “*establish[ing] an intended meaning for the utterance within the identified context.*” (EX-1003, ¶204.)

For utterances containing a sense ambiguity, SmartKom discloses the system “*disambiguate[s] an intent that the user had in speaking the one or more words that have the different meanings in the different contexts*” using the enrichment and scoring process discussed above. (EX-1003, ¶¶205-10.) Specifically, in cases “of pragmatic ambiguity” such as sense ambiguity, “the model resulting from the analyses of the corresponding data has to embed the utterance at hand into a greater situational context, e.g., by computing a contextual coherence score for the competing interpretations.” (EX-1005, 273.)

Thus, the disambiguation of words that have the different meanings in the different contexts occurs through use of the identified context, integrated into each intention hypothesis. An example of such disambiguation occurs in an utterance containing the German word “*kommen*.” SmartKom discloses an “evaluation of the contextual coherence scoring in terms of its disambiguation performance meant calculating how often contextual coherence picks the appropriate reading, given an ambiguous lexicon entry such as *kommen* associated in the lexicon with both

WatchPerceptualProcess and MotionDirectedTransliterated.” (EX-1005, 280, *see also* EX-1005, 607; EX-1014, 281-282.)

SmartKom’s system thus considers more contexts when disambiguating sense ambiguities than the ’681 patent which only describes using recent dialog history:

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).

(EX-1001, 11:10-20.)

6. Limitations [1E.1]-[1E.2]/[13G.1]-[13G.2]/[25G]

SmartKom discloses “*generat[ing]/[e] 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*” [1E.1]/[13G.1] and “*generat[ing] a grammatically or syntactically adapted*

response to the utterance based on the intended meaning established within the identified context” [25G]. (EX-1003, ¶¶211-17.)

After selection of an interpretation from the set of hypotheses, the intention recognizer provides the interpretation to the action planner which makes “decisions on the applications to contact and the next interaction with the user.” (EX-1005, 287.) The action planner, also referred to as the dialogue manager, “*generat[es the] response to the utterance.*” (See EX-1005, 302, Figure 1(303)(below); EX-1003, ¶212.) The presentation planner, natural language generation [NLG] component, and speech synthesis cooperate to present the response multi-modally to the user. (See EX-1003, ¶212; EX-1005, 396, 401.)

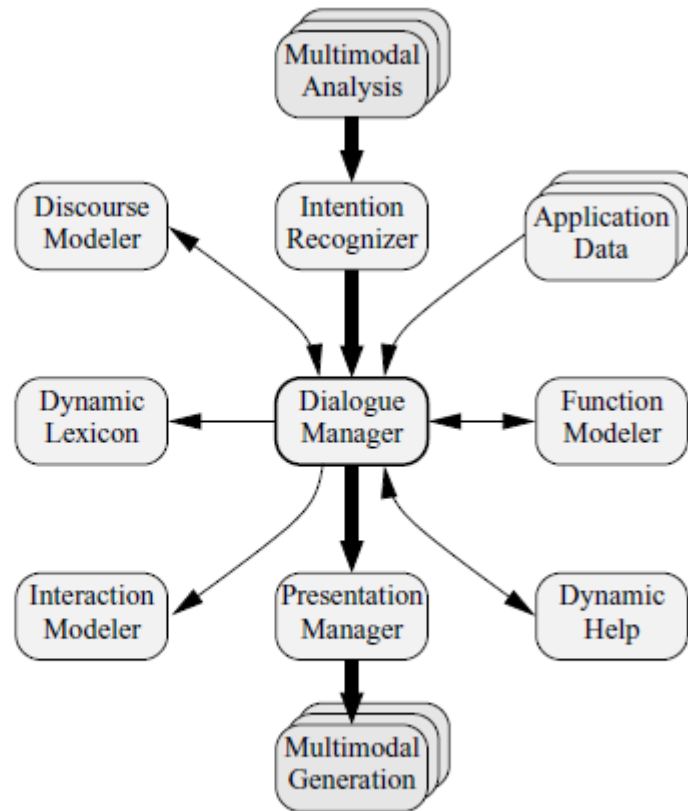


Fig. 1. Modules in direct communication with the dialogue manager

The action planner “*generates a response to the utterance*” through development of an action plan “*specify[ing] a possible course of subsequent communicative acts to reach*” a goal articulated by the user in the utterance. (EX-1005, 311; EX-1003, ¶213.) An action plan, illustrated below for responding the utterance “I want to send a document”, includes a set of steps (also referred to as “games”/“moves”). Each step includes the utterance/conversation type (e.g., instruct, inform, request-response, graphical action) and the channel which specifies the roles of the parties for the step. (See EX-1005, 305-313, Table 3(313).)

Table 3. Fully expanded plan for sending a fax message

Step	Task	Channel	Application
1	Present clear screen	→ user (g)	fax
2-1	Present scanning area	→ user (g,s)	realDocument
2-2	Instruct to place the document	→ user (s)	
2-3	Initialize document scanner	→ realDocument	
2-4	Response initialization complete	← realDocument	
2-5	Request start scanning	→ realDocument	
2-6	Response scanning complete	← realDocument	
2-7	Request to remove document	→ user (s)	
2-8	Response document removed	← realDocument	
2-9	Request scanned image	→ realDocument	
2-10	Response scanned image	← realDocument	
2-11	Present scanned image	→ user (g)	
3-1	Present keypad and request number	→ user(g,s)	phone
3-2	Collect number	← user (reactive)	
3-3	Response number	← internal	
4	Inform “fax being sent”	→ user (s)	fax
5-1	Request transcribe picture	→ realDocument	realDocument
5-2	Response transcribed image	← realDocument	
6	Request sending of fax	→ telephony	fax
7	Response sending complete	← telephony	
8	Inform about completion of task	→ user (g,s)	

In the above plan, the “*intended meaning* [of the utterance] *established within the identified context*” is “I want to send a fax” which sets the goal of sending a fax. (See EX-1005, 312-313 (Table 3).) The responses to this utterance are “*grammatically or syntactically adapt[ed]*” based on the meaning (“I want to send a fax”). (EX-1003, ¶214.) Specifically, the system adapts a response, e.g., to “[i]nstruct” the user “to place the document” (step 2-2) in the scanning area and

later to request the user input a telephone number (step 3-1). (EX-1005, 313 (Table 3).)

Considering the phrase “was für Spielfilme kommen heute Abend” which includes the word “kommen” meaning either “showing on” or “to come” (§IV.B.2)—when the system establishes the “*intended meaning ... within the identified context*” for utterance to be “what films are showing this evening,” the action planner prepares a responsive plan to achieve the goal (providing a film listing). (EX-1003, ¶215.)

SmartKom provides a similar action plan for an utterance to provide TV listings to a user. In this example, the utterance (and its goal) was interpreted as “show me the TV program.” (EX-1005, 308, Figure 4(309).) The response has two alternatives—if the TV database finds no results, the response will “be an intention of type ‘noDataAvailable,’ generated as ‘I have not found any data for your request’” OR if broadcasts are found, the “results [provided] in an intention containing the results of the database query” which “is converted by the output channel to a TV program presentation, which is then passed on to the presentation manager.” (EX-1005, 308-309.)

Associating this example with the response to the “kommen” utterance, if the movie database finds no results for films showing tonight, the system response is “I have not found any data for your request” and if the system finds movies, the

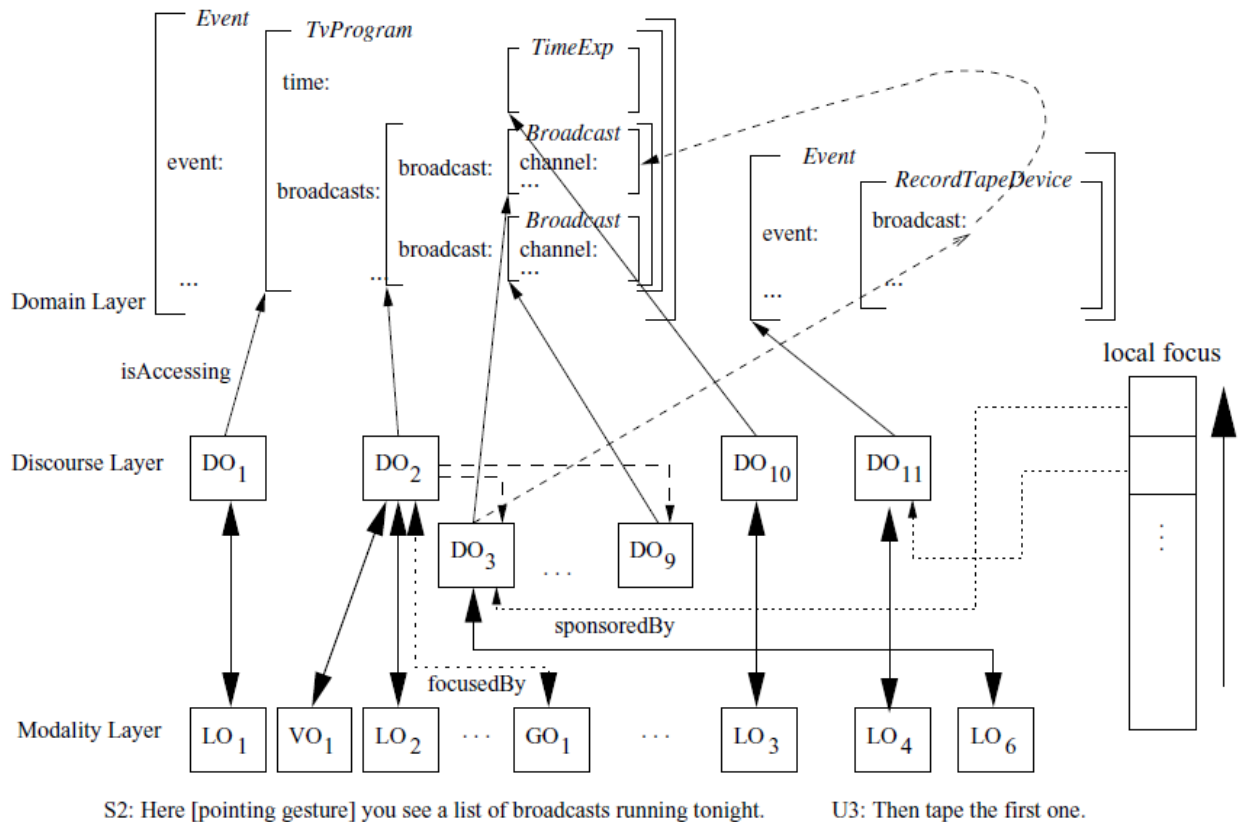
results are provided. (See EX-1005, 308-309; EX-1003, ¶217.) That is, “*the response*” is “*grammatically or syntactically*” adapted “*based on the intended meaning established within the identified context.*” (EX-1003, ¶217.)

C. Dependent Claims

1. Claims 2/14/26

SmartKom discloses “*populat[ing] a short-term context stack with information about the utterance received during the current conversation.*” (EX-1003, ¶¶218-19.)

SmartKom’s discourse modeling module (DIM) represents the discourse state (“*short-term shared knowledge*”) using a “three-tiered” structure: “the discourse object layer, the modality layer, and the domain object layer.” (EX-1005, 242.) The discourse object layer “comprises the concepts introduced into the discourse.” (EX-1005, 242.) A focus space “covers all turns belonging to the same topic and enables access to a corresponding local focus structure.” (EX-1005, 245.) The local focus structure is a stack that “provides and restricts access to all discourse objects that are antecedent candidates for later reference.” (EX-1005, 245, 248 (referencing the “local focus stack”).) Thus, SmartKom populates the local focus stack (“*short-term context stack*”) “*with information about the utterance received during the current conversation.*” (EX-1003, ¶218.)



SmartKom, Figure 2(244)

2. Claims 3/15/27

SmartKom discloses or suggests “*expir[ing]/[e] the information about the utterance from the short-term context stack after a psychologically appropriate amount of time*” [3]/[15]/[27]. (EX-1003, ¶¶220-23.) The ’681 patent lacks details regarding the extent or metrics associated with “*a psychologically appropriate amount of time.*” (See EX-1001, 4:65-5:2.) Regardless, SmartKom discloses dialog history information expires during or after a session.

SmartKom’s discourse state is “based on the three-tiered context representation presented in Luperfoy (1992).” (EX-1005, 239.) Luperfoy notes an

“agent participating in a dialogue experiences information decay[ing] over the course of the conversation” with “information from the linguistic, discourse, and belief system tiers decays at different rates and in response to different cognitive forces/limitations.” (EX-1011, 24-25.) Specifically, LOs “become old and vanish at an approximately linear rate as a function of time counted from the point of their introduction into the discourse history, i.e., as LOs get older, they fade from the discourse and can no longer serve as linguistic sponsors for anaphors.” (EX-1011, 25.) Discourse objects (referred to by Luperfoy as “pegs”) “decay as a function of attentional focus, so that as long as an individual or concept is being attended to in the dialogue, the discourse peg will remain near the top of the focus stack and available as a potential discourse sponsor for upcoming dependent referring expressions.” (EX-1011, 25.)

SmartKom stores its discourse state in “discourse memory.” (EX-1005, 242.) SmartKom acknowledges this decay of information, explaining that “[f]or longer dialogues (more than half an hour of discourse), the discourse memory runs out of memory” and in such cases, “even for most humans,” it is “necessary to forget information.” (EX-1005, 252.) Thus, SmartKom describes or suggests a “*psychologically appropriate amount of time*” before expiring information from the discourse model. (EX-1003, ¶222.)

3. Claims 4/16/28

The SmartKom-Kobsa combination discloses “*update[ing]/[e] 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*” [4]/[16]/[28]. (EX-1003, ¶¶224-27.)

Kobsa teaches accumulating knowledge stored in the user model (“*long-term profile associated with the user*”) either explicitly from asking the user questions or implicitly by eavesdropping on the conversation “to observe the user’s behavior and from it to infer facts.” (EX-1006, 416.) This information is “*about the utterance received during the current conversation.*” (EX-1003, ¶225.) Kobsa describes exemplary user experience/knowledge data (Figure 3) accumulated through interpreting utterances in prior conversations reflecting “whether the user understands” basic concepts in the discourse domain. (EX-1006, 197, 17.) The concepts and their state (known/not-know/no-information) reflect “*information about the utterance received during the current conversation.*” (EX-1003, ¶225.) The state and certainty are also “*relevant data.*”

Concept hierarchy of the system	User model	
	user's knowledge state	certainty of assumption
INFECTIOUS-PROCESS	KNOWN	100
HEAD-INFECTION	KNOWN	100
SUBDURAL-INFECTION	NOT-KNOWN	100
OTITIS-MEDIA	NO-INFORMATION	100
SINUSITIS	NO-INFORMATION	100
MENINGITIS	KNOWN	100
BACTERIAL-MENINGITIS	KNOWN	90
MYCOBACTERIUM-TB	NOT-KNOWN	70
VIRUS	NOT-KNOWN	90
FUNGAL-MENINGITIS	NOT-KNOWN	100
MYCOTIC-INFECTION	NOT-KNOWN	100
ENCEPHALITIS	NOT-KNOWN	90
SKIN-INFECTION	KNOWN	100

Figure 3. An example of an overlay model

Kobsa, Figure 3

Kobsa further teaches “at the end of a dialog session,” the system “records all information about the user inferred from his/her dialog behavior in a corresponding file.” (EX-1006, 10-11.) That is, the system updates the long-term user model with information/relevant data before the session ends and short-term session knowledge is deleted. (EX-1003, ¶226.) Because the short-term dialog history expires (§IV.C.2), information in that dialog history and its “*relevant data*” are “*associated with the information expired from the short term context stack.*” (EX-1003, ¶226.)

4. Claims 5/17/29

a) Limitations [5A]/[17A]/[29A]

SmartKom discloses “*identify[ing] 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*” [5A]/[17A]/[29A]. (EX-1003, ¶¶228-33.)

The user’s goal “is a state of affairs he wishes to achieve.” (EX-1006, 414.) SmartKom identifies a “*conversational goal associated with the utterance,*” consistent with the ’681 patent’s disclosure. (EX-1003, ¶230; EX-1001, 14:39-47 (goal includes “getting a discrete piece of information or performing a discrete task”).) SmartKom’s intention lattice is associated with a hypothesis for an utterance “*specif[ying] a set of dialogue acts, which in turn can be either goal or slot manipulations.*” (EX-1005, 305; *see also*, 306 (specifying goal manipulations as “setting”, “retracting” or “retaining” a goal).) The system identifies the “*conversational goal*” in the action plan. (*See* EX-1005, 308 (showing example of “goal tag” in plan as signifying “**a user-specifiable goal** (i.e., “show me the TV program” or equivalent)”).)

SmartKom also identifies “*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” when creating its plan for meeting a user specified goal. (EX-1003, ¶231.) For example, the action planner specifies a channel and task associated with each “move” in the plan for addressing the “*conversational goal*.” (See, e.g., EX-1005, 311-312.) As shown in the example below, a left arrow next to the term “user” indicates a user output channel (e.g., response provided via an output modality such as speech). (EX-1003, ¶231.) This channel is associated with a task performed by the system such that the system role is leader and the user is the listener/consumer. (*Id.*) For example, in step 4, which is a speech-based (s) “inform” conversation type, the system is the leader and the user is the information consumer. The information in this step is allocated (provided) to the user (e.g., information that “fax being sent”). (*Id.*)

Table 3. Fully expanded plan for sending a fax message

Step	Task	Channel	Application
1	Present clear screen	→ user (g)	fax
2-1	Present scanning area	→ user (g,s)	realDocument
2-2	Instruct to place the document	→ user (s)	
2-3	Initialize document scanner	→ realDocument	
2-4	Response initialization complete	← realDocument	
2-5	Request start scanning	→ realDocument	
2-6	Response scanning complete	← realDocument	
2-7	Request to remove document	→ user (s)	
2-8	Response document removed	← realDocument	
2-9	Request scanned image	→ realDocument	
2-10	Response scanned image	← realDocument	
2-11	Present scanned image	→ user (g)	
3-1	Present keypad and request number	→ user(g,s)	phone
3-2	Collect number	← user (reactive)	
3-3	Response number	← internal	
4	Inform “fax being sent”	→ user (s)	fax
5-1	Request transcribe picture	→ realDocument	realDocument
5-2	Response transcribed image	← realDocument	
6	Request sending of fax	→ telephony	fax
7	Response sending complete	← telephony	
8	Inform about completion of task	→ user (g,s)	

b) Limitations [5B]/[17B]/[29B]

SmartKom discloses “*classify[ing] 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 ...*” [5B]/[17B]/[29B]. (EX-1003, ¶¶234-38.)

SmartKom’s action planner classifies its utterances into communicative games containing actions based on the conversational goal articulated and understood by the system. (EX-1003, ¶236.) For example, SmartKom labels

“communicative games” as, e.g., “*inform, instruct, request-response*, and the special type *graphicalAction*.” (EX-1005, 312, Table 3(below).) A POSITA would recognize these to also be conversation/utterance types—query (request-response), instruction, and informative. For example, as shown below, the system selects the “games” (i.e., conversation types) needed to achieve the user’s articulated goal—i.e., the utterances in the conversation are assigned a conversation type based on the “*identified conversational goal*.” (EX-1003, ¶236.)

Table 3. Fully expanded plan for sending a fax message

Step	Task	Channel	Application
1	Present clear screen	→ user (g)	fax
2-1	Present scanning area	→ user (g,s)	realDocument
2-2	Instruct to place the document	→ user (s)	
2-3	Initialize document scanner	→ realDocument	
2-4	Response initialization complete	← realDocument	
2-5	Request start scanning	→ realDocument	
2-6	Response scanning complete	← realDocument	
2-7	Request to remove document	→ user (s)	
2-8	Response document removed	← realDocument	
2-9	Request scanned image	→ realDocument	
2-10	Response scanned image	← realDocument	
2-11	Present scanned image	→ user (g)	
3-1	Present keypad and request number	→ user(g,s)	phone
3-2	Collect number	← user (reactive)	
3-3	Response number	← internal	
4	Inform “fax being sent”	→ user (s)	fax
5-1	Request transcribe picture	→ realDocument	realDocument
5-2	Response transcribed image	← realDocument	
6	Request sending of fax	→ telephony	fax
7	Response sending complete	← telephony	
8	Inform about completion of task	→ user (g,s)	

SmartKom, Table 3(313)

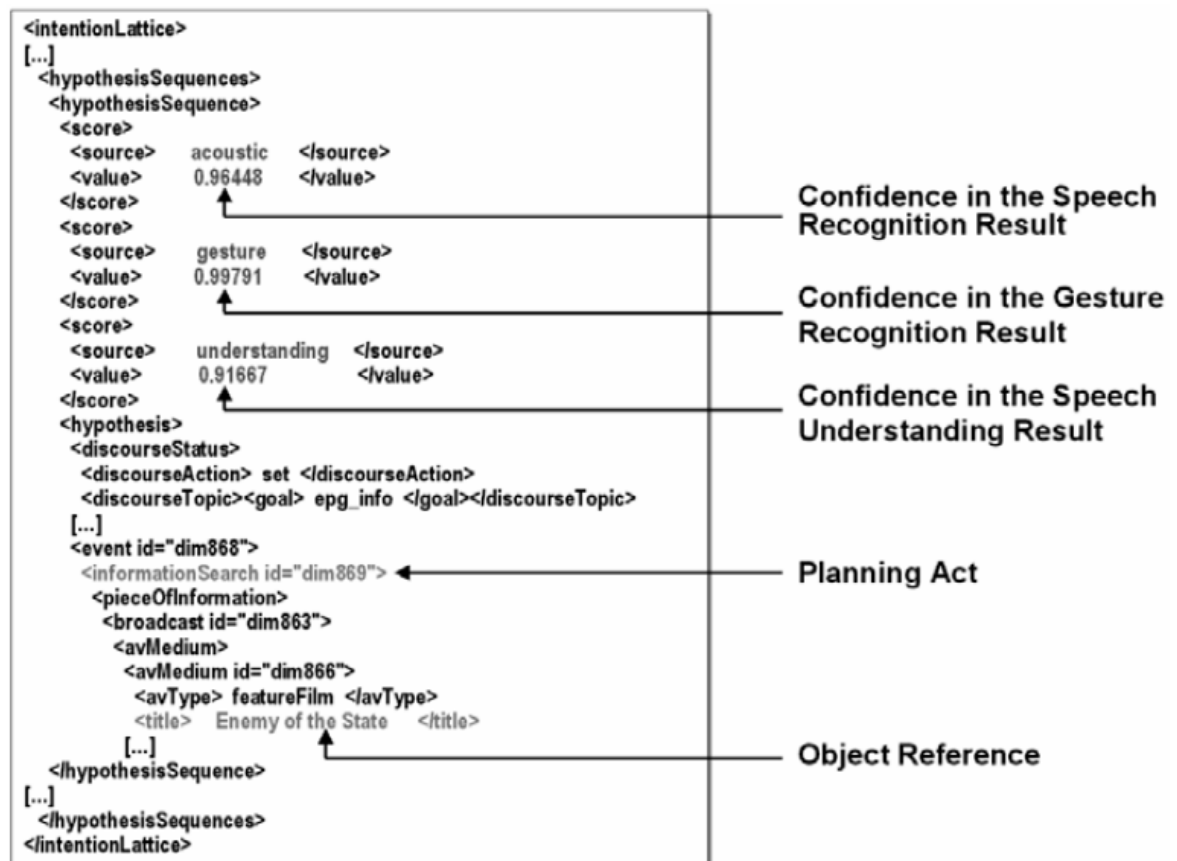
SmartKom’s “*conversational speech engine*” “*establishes the intended meaning based on the conversation type.*” (EX-1003, ¶237.) Because the action plan includes the “moves” needed to achieve the user specified goal, the plan predicts “future moves of the dialogue” (e.g., need to request information from a user, need for the user to supply information). (See EX-1005, 308-312.) These expected moves are used by analysis modules “to resolve ambiguities or to prime the lexicon for words that may occur in the input. (EX-1005, 312.) Accordingly, “[a]fter each dialogue turn, the dialogue manager publishes an updated data structure, called an *expectation*, to inform other modules about likely subsequent utterances.” (EX-1005, 312 (emphasis in original).) Therefore, when interpreting an utterance in the conversation, the system has the expectations associated with the received utterance including the “*conversation type.*” (EX-1003, ¶237.)

5. Claims 6-7/18-19/30-31

Claims 7/19/31 depend from claims 6/18/30. Based on their interdependence Petitioner addresses [7A]/[19A]/[31A] first, followed by claims 6/18/30 and [7B]/[19B]/[31B]. SmartKom discloses “*generating a preliminary interpretation of the utterance at a speech recognition engine coupled to the voice input device and the conversational speech engine*” [7A]/[19A]/[31A]. (EX-1003, ¶¶239-40.)

SmartKom’s system includes a “**speech recognition engine**”, coupled to the audio input and “*conversational speech engine*” (§IV.B.3), that “transforms the

acoustic signal into a **sequence of hypothesized words** in orthographic representation.” (EX-1005, 86, Figure 2(60).) The speech recognition engine uses “**confidence measures**” to “estimate confidences for the correctness of the words hypothesized by the recognizer.” (EX-1005, 85, 96.) The speech recognition engine adds this “confidence” score to each hypothesis in the intention lattice, as shown in the exemplary fragment below. As shown, the speech interpretation (natural language understanding) component also add a “confidence” score for its result. (See EX-1005, 204.)



SmartKom, Figure 8(15)

As discussed in §IV.B.5, when establishing the “*intended meaning*” of an utterance, the context module “compute[s] [a] contextual coherence score[.]” (EX-1005, 276, 271, Table 2(277).) The contextual coherence score represents “*a degree of certainty about the intent*” (EX-1003, ¶242.) As shown in the Table below, this score is included with the hypothesis in the intention lattice. (EX-1005, Table 2(277).) Thus, SmartKom discloses “*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*” [6]/[18]/[30]. (EX-1003, ¶¶241-42.)

Table 2. Context-specific insertions into a sample intention hypothesis resulting from the interpretation of a speech recognition hypothesis

<pre> <informationSearchProcess> <entertainment> <performance> <cinema> <contact> <address> <town> here </town> </address> </contact> </cinema> <time> <beginTime> <at> now </at> </beginTime> </time> </performance> </entertainment> </informationSearchProcess> </pre>	<pre> <contact> <x> 70.345 </x> <y> 49.822 </y> <town> Heidelberg </town> </contact> <time> <at> 19:00:00T26:08:03 </at> </time> </pre>
	<pre> <scores> <contextualCoherence> 0.46 </contextualCoherence> </scores> </pre>

The scoring task involves using the “ontological domain context to measure the semantic coherence of the individual interpretation” and “using dynamic situational and discourse information, e.g., previous ontological contexts of prior turns.” (EX-1005, 271.) This “dynamic situational and discourse information” includes “*information associated with the identified context*” (i.e., dialog history, situational/user states) and the “confidence” values which are a type of information about the discourse. As noted in §IV.B.3, software executing on the processors of the “*conversational speech engine*” perform these process steps. Thus, “*the degree*

of certainty to the hypothesis” is assigned by the “conversational speech engine” and its associated processors, “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” [7B]/[19B]/[31B]. (EX-1003, ¶¶243-44.)

6. Claims 8-9/20-21/32-33

SmartKom discloses *“the conversational speech engine further grammatically or syntactically adapts the response based on conversation type” [8]/[20]/[32]. (EX-1003, ¶¶245-46.)*

As discussed in §IV.B.6, the system adapts its response based on the conversation type established in the action plan. For example, in acting on the user goal of “I want to send a fax”, the system generates a series of responses to the user including an instruction: place the document on the scanning area and a request “provide a number.” (EX-1005, 312, Table 3(313).) These responses are therefore *“grammatically or syntactically adapt[ed] based on the conversation type”* (instruction/query-response) identified in the plan.

SmartKom further discloses *“the response” is “grammatically or syntactically adapted”* by the *“conversational speech engine”* and its processors *“to influence a subsequent reply utterance that the conversational speech engine*

expects from the user during the current conversation” [9]/[21]/[33]. (EX-1003, ¶¶247-50.)

The action plan designs its responses to influence the user to provide information or perform tasks necessary to achieve user’s articulated goal. (*See, e.g.,* EX-1005, 307-313.) As shown in the example below, the system recognizes that to achieve the goal of sending a fax, it expects the user to provide a telephone number and therefore adapts its response to the user to request the number. (*See* EX-1005, Table 3(313).) That is, SmartKom’s “*conversational speech processor*” adapts its response to influence’s the user’s reply to gather the information needed to achieve the goal being processed. (EX-1003, ¶249.)

Table 3. Fully expanded plan for sending a fax message

Step	Task	Channel	Application
1	Present clear screen	→ user (g)	fax
2-1	Present scanning area	→ user (g,s)	realDocument
2-2	Instruct to place the document	→ user (s)	
2-3	Initialize document scanner	→ realDocument	
2-4	Response initialization complete	← realDocument	
2-5	Request start scanning	→ realDocument	
2-6	Response scanning complete	← realDocument	
2-7	Request to remove document	→ user (s)	
2-8	Response document removed	← realDocument	
2-9	Request scanned image	→ realDocument	
2-10	Response scanned image	← realDocument	
2-11	Present scanned image	→ user (g)	
3-1	Present keypad and request number	→ user(g,s)	phone
3-2	Collect number	← user (reactive)	
3-3	Response number	← internal	
4	Inform “fax being sent”	→ user (s)	fax
5-1	Request transcribe picture	→ realDocument	realDocument
5-2	Response transcribed image	← realDocument	
6	Request sending of fax	→ telephony	fax
7	Response sending complete	← telephony	
8	Inform about completion of task	→ user (g,s)	

7. Claims 11-12/23-24/35-36

Barbara discloses or suggests the “*multi-modal input*” claims [11]/[23]/[35] and “*multi-modal output*” claims [12]/[24]/[36]. (EX-1003, ¶¶393-98.)

Barbara describes a “first system, in which documents are electronically scanned” (i.e., an OCR input; “*non-voice input*”). (EX-1007, ¶26, Figure 1.)

Barbara’s system further includes a display screen at the user terminal (i.e., a visual output; “*non-voice output*”). (See, e.g., EX-1007, Figures 1, 6.) Barbara teaches this system is “additionally or alternatively” used “to create a voice

interface where the end user system responds to spoken natural language commands” (i.e., a speech (“voice”) input/output). (EX-1007, ¶82, *see also*, ¶166.) That is, Barbara teaches a system with multiple modalities for input and output—a “*multi-modal*” system.

Although Barbara does not provide an explicit example use of a “*multi-modal input*” that “*includes one or more voice inputs relating to the utterance*”, Barbara’s teachings are suggestive of such a use. (EX-1003, ¶386.) For example, a POSITA would understand, based on Barbara’s teaching, the “user system responds to spoken natural language commands” to indicate the user could speak the command “scan” to cause the OCR input (“*non-voice input*”) to the system. (EX-1003, ¶386.) That is, the spoken “utterance” relates to the “*non-voice input*” (OCR-scan). (EX-1003, ¶386.) Moreover, the use of multi-modal inputs as disclosed in Barbara were well known to those skilled in the art by the filing date of the ’681 patent. (EX-1020, 381 (stating “[m]ultimodal systems provide a more flexible interface for users who can choose the most appropriate interface to meet their requirements and preferences” and “[m]ultimodal systems that permit more than one input mode can also provide for better handling of errors, particularly errors produced by the speech recognition component”).)

Barbara teaches or suggests claims [11]/[23]/[35]. (EX-1003, ¶387.)

Barbara describes the interplay between these modalities in exemplary systems for providing user responses. For example, in response to an utterance, the user terminal “may either speak back or perform some action ... or both.” (EX-1007, ¶107.) That is, the multi-modal output provides “one or more tasks executed to process a request identified from the intended meaning.” (EX-1003, ¶388.)

Barbara also provides extensive discussion of an e-mail interface where the user speaks, e.g., “Show me incoming messages” and messages meeting the query are displayed at the user terminal (“*non-voice output*” that “*relates to the utterance*”). (See, e.g., EX-1007, ¶¶119-162, *see also*, ¶107 (“feedback is given by the application visually responding”).) As an example, Barbara explains “if a user asks their phone if they have any e-mail, the system may reply with a brief [spoken] summary of the new messages” whereas “[i]f the user asks their television/computer”, which has a display screen, “the user’s inbox may also be displayed.” (EX-1007, ¶107.) That is, the system provides a spoken summary and visual display (i.e., “*multi-modal*” output).

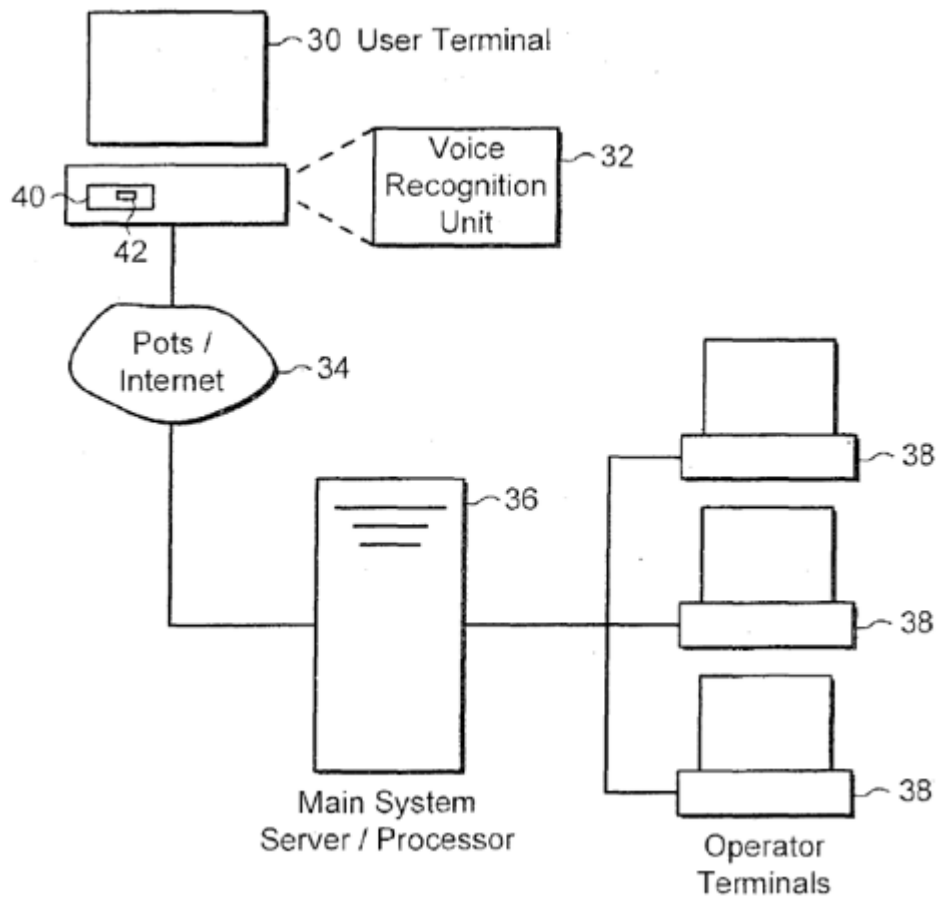
Barbara discloses claims [12]/[24]/[36] (EX-1003, ¶¶384-389.)

V. GROUND 2: Barbara-Ross Combination Renders Obvious Claims 1, 5-8, 10-13, 17-20, 22-25, 29-32, 34-36

A. Combination Overview

1. Barbara

Barbara's system (figure 6, below) interprets a user's spoken word (e.g., the utterance) "to determine what the user wanted, i.e. the intent of the spoken word." (EX-1007, ¶82.) The system includes user terminal 30 with voice interface equipment 32 "operable to receive the spoken word and translate it into an electronic format" that is passed to processor(s) "having software for receiving, interpreting and correcting the incoming information." (EX-1007, ¶¶ 83-85.) Contextual information is "captured as the user interacts with the system." (EX-1007, ¶98.)



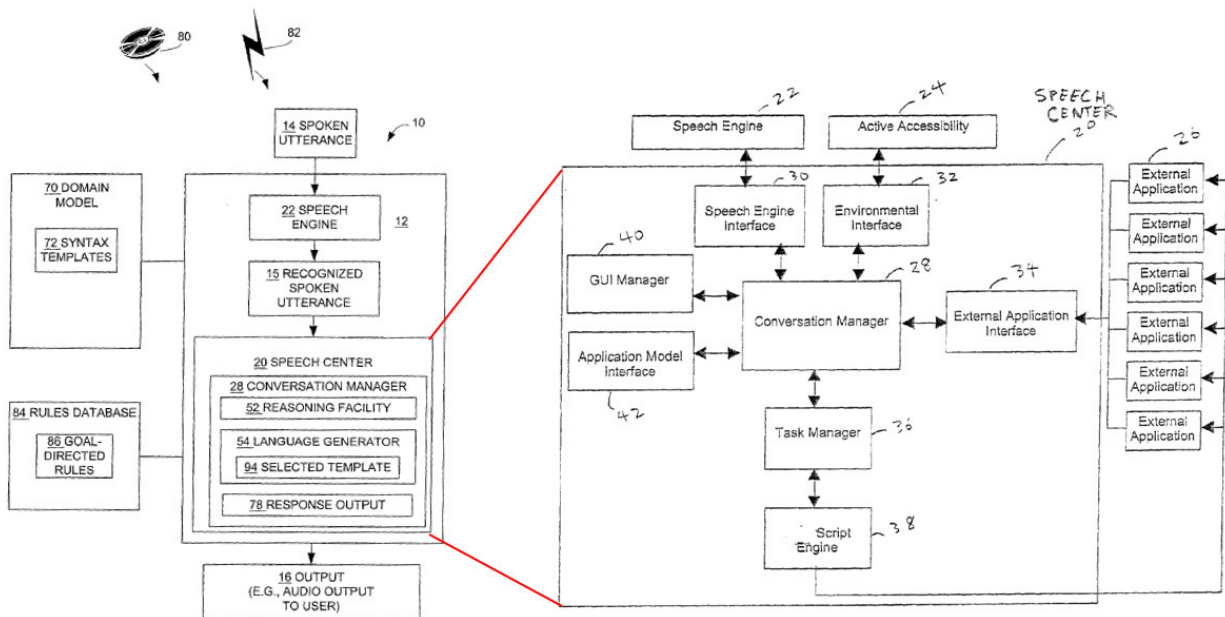
Barbara, Figure 6

When a user's "utterance has been detected," voice recognition engine 32 provides its "best guesses" "as to what was said," and what was meant. (EX-1007, ¶86.) After the correct intent is determined, the system provides a response. (EX-1007, ¶¶105-107, Fig. 7.)

Barbara also provides a manual correction technique when utterances "fail[] to meet the pre-determined criteria for automatic acceptance." (EX-1007, ¶¶100-104.) Notably, this operator intervention and correction "will be used heavily during" an "initial period." (EX-1007, ¶102.)

2. Ross

Ross discloses a “conversation manager [that] processes spoken utterances from a user of a computer, and develops responses to the spoken utterances.” (EX-1008, Abstract.) Ross’s system 10 comprises digital processor 12 including speech engine speech center 20 having conversational manager 28. (EX-1008, ¶¶22-24, Figs. 1-2 (below).)

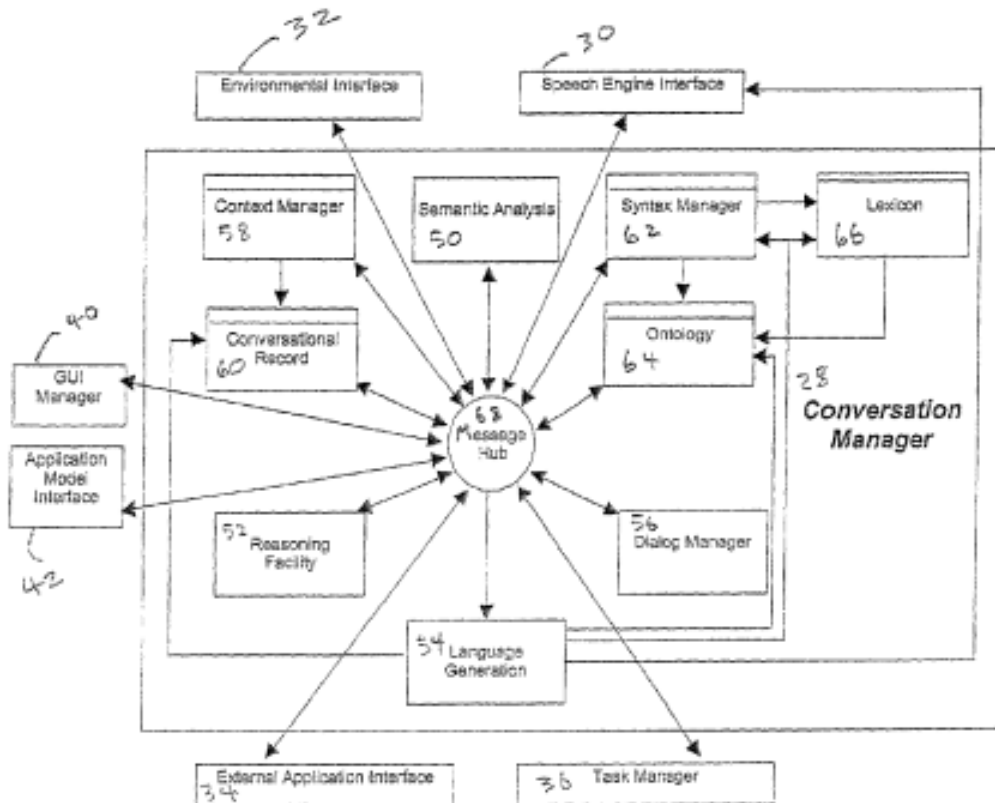


Ross, Figure 1

Ross, Figure 2

Conversational manager 28 includes reasoning facility 52 that “performs the reasoning process” for conversation manager 28 and is “primarily concerned with the determination of how to achieve the goals derived from the user’s questions and command.” (EX-1008, ¶56, Fig. 3 (below).) Conversation manager 28 also

stores conversational record 60 including each utterance in a conversation and domain model 70. (EX-1008, ¶57.)



Ross, Figure 3

Ross's systems uses the conversational records and domain model to process utterances. (EX-1008, ¶22 “a natural language response output 78 to the recognized spoken utterance 15 [is] based on” domain model 70, rules database 84, and syntax template 94).) Ross however notes that “[c]onversational speech is full of implicit and explicit references back to people and objects that were mentioned earlier.” (EX-1008, ¶57.) To understand such sentences, speech center 20 “looks at

the **conversational record 60**, and finds the missing information.” (EX-1008, ¶57.)

3. Motivation to Combine

While Barbara discloses a “server” to process utterances, it lacks details regarding the server’s architecture. A POSITA would have been motivated to combine Ross’s teachings regarding the architecture of a conversation manager with Barbara’s voice interface system. (EX-1003, ¶285.)

Barbara and Ross are analogous art to the ’681 Patent. Each pertains to the same field of endeavor, i.e., speech recognition systems. (*Compare* EX-1001, 7:24-33 *with* EX-1007, ¶7 *and* EX-1008, Abstract; EX-1003, ¶286.)

Barbara motivates the combination. For example, Barbara discloses the system “may either speak back or perform some action ... or both” in response to an utterance. (EX-1007, ¶107.) But, Barbara lacks details about how to generate such a response. A POSITA would have been motivated to search for a reference with such implementation details and would have been led to Ross which provides robust disclosure of its conversation manager including response generation. (EX-1003, ¶287.) That is, a POSITA would have been motivated to combine Barbara with Ross to provide a system with flexible response generation capabilities.

While Barbara discloses use of an “information database” and “knowledge base” to interpret utterances, Barbara provides limited detail regarding managing

the information in those data stores. Ross discloses creating conversational records for storing a dialog history between the user and system including each user utterance and the system's interpretation. (EX-1008, ¶57.) Ross further teaches that this conversational information "is eventually purged from the conversational record when it is no longer relevant to active goals." (*Id.*) Implementing Ross's teachings regarding purging information from the "conversational record" allows Barbara's dialog history to be expired and long-term data to be retained. (EX-1003, ¶288.) A POSITA would understand this data management approach improves storage efficiency and reduces need for extensive hardware storage. (EX-1003, ¶288)

Finally, the combination is nothing more than the application of a known technique (Ross's conversation manager and data storage) to a known device (Barbara's voice interface system) ready for improvement for the reasons discussed above. (EX-1003, ¶289.)

A POSITA would have had a reasonable expectation of success and the results of the combination would have been predictable. The Barbara-Ross combination would merely implement Barbara's knowledge base to use Ross's conversation records which expire, releasing memory space and Ross's conversation manager teachings including response generation components. (EX-1003, ¶290.) Integrating Ross's teachings into Barbara's system would have been

well within the skill set of a POSITA because the combination involves software and data storage, concepts well understood before the earliest priority date of the '681 patent. (EX-1003, ¶290.)

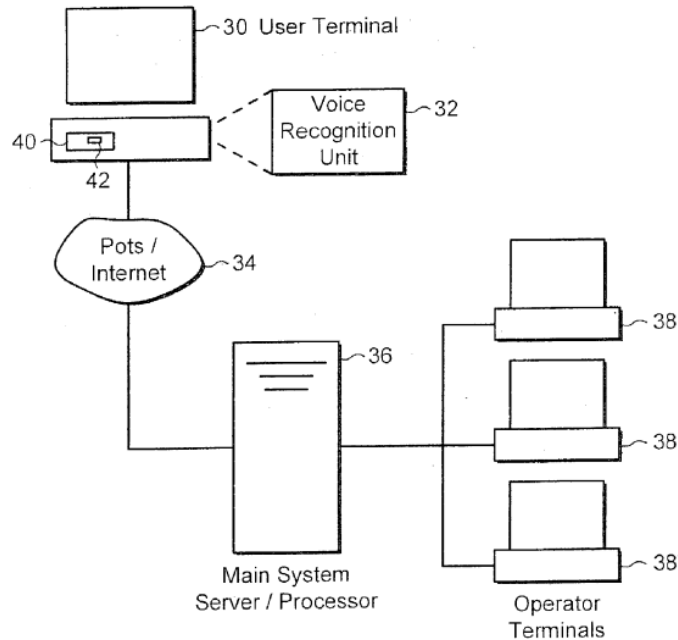
B. Independent Claims

1. Preambles [1P]/[13P]/[25P]

The Barbara-Ross combination discloses the preambles [1P]/[13P]/[25P]. (EX-1003, ¶¶292-297.) Barbara discloses a “*method*” [1P] and “*system*” [25P]: “**system and method** for improving accuracy of signal interpretation.” (EX-1007, ¶7, ¶86 (describing “**process steps** for using the” system), Figure 7.)

Barbara’s system has a main server with “one or more service processors” including “**software** for receiving, interpreting and correcting the incoming information.” (EX-1007, ¶85, Figure 6 (below).) Although Figure 6 shows voice interface equipment 32 at a user terminal, in an embodiment, end-user’s terminal 30 is “used as a telephone” with voice recognition software “provided at a remote location.” (EX-1007, ¶83.) Thus, the software associated with voice recognition is also included in the main server. (EX-1003, ¶¶295-296.) Barbara therefore discloses “**a computer program** [“*computer-executable instructions*”], **preferably on a data carrier or a computer readable medium** [“*non-transitory computer readable medium*”], for controlling [the] method of interpreting

electronic signals, the computer program comprising **instructions.**” (EX-1007, ¶24, claim 30; §IV.B.1.)



Barbara, Figure 6

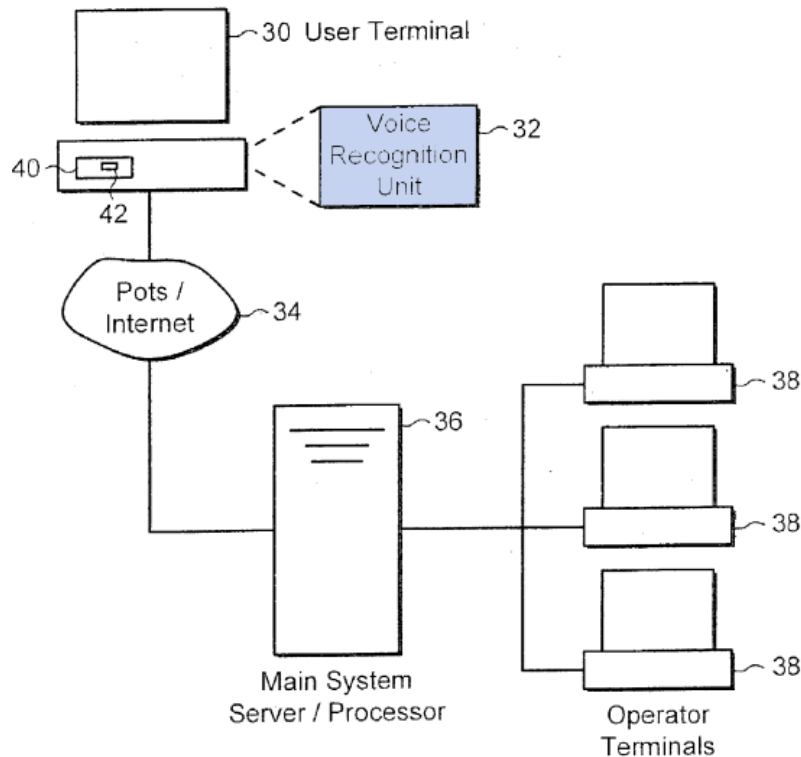
Barbara’s system is a “*voice user interface*” that is both “*cooperative*” and “*conversational.*” (EX-1003, ¶297.) Barbara teaches its system “create[s] a **voice interface** where the end system **responds to spoken natural language commands.**” (EX-1007, ¶82; *see also*, EX-1007, ¶86.)

2. Receiving An Utterance [1A]/[13A]/[25A]

Barbara discloses “*receiving an utterance*” [1A.1]/[13A.1]/[25A.1]/[25A.2]. (EX-1003, ¶¶298-307.)

Barbara’s system includes voice interface equipment 32 (shaded blue) “**operable to receive the spoken word** and translate it into an electronic format.”

(EX-1007, ¶83, Fig. 6 (below).) The voice recognition software, whether in the user device or in the server, alone or in combination with the component in the user device (e.g., a microphone) is a “voice input device” that “receive[s] an utterance” “during a current conversation with a user.” (EX-1003, ¶299.)



Barbara, Figure 6

Barbara discloses limitations [1A.2]/[13A.2]/[25A.3], Which are directed to the well-known concept of sense ambiguity. (EX-1003, ¶310; §IV.B.2.) Barbara provides an example of such sense ambiguity. Specifically, in Barbara, the word “Toto” means (1) name of a dog in the pet context or (2) name of a restaurant in the restaurant context. (EX-1007, ¶96 (“How’s Toto doing? (Toto could be the user’s dog)”), ¶148 (referring to Toto the restaurant).)

In Barbara's system "[w]hen each utterance is received at the server, it is evaluated into a set of possible requests that the user may have actually meant" (EX-1007, ¶90) and ambiguities are "resolved when the system evaluates the likelihood that the user really is asking" each hypothetical interpretation. (See EX-1007, ¶98.) Accordingly, the system handles "sense ambiguities" although Barbara provides only one example (Toto-dog versus Toto-restaurant). (EX-1003, ¶¶301-303.)

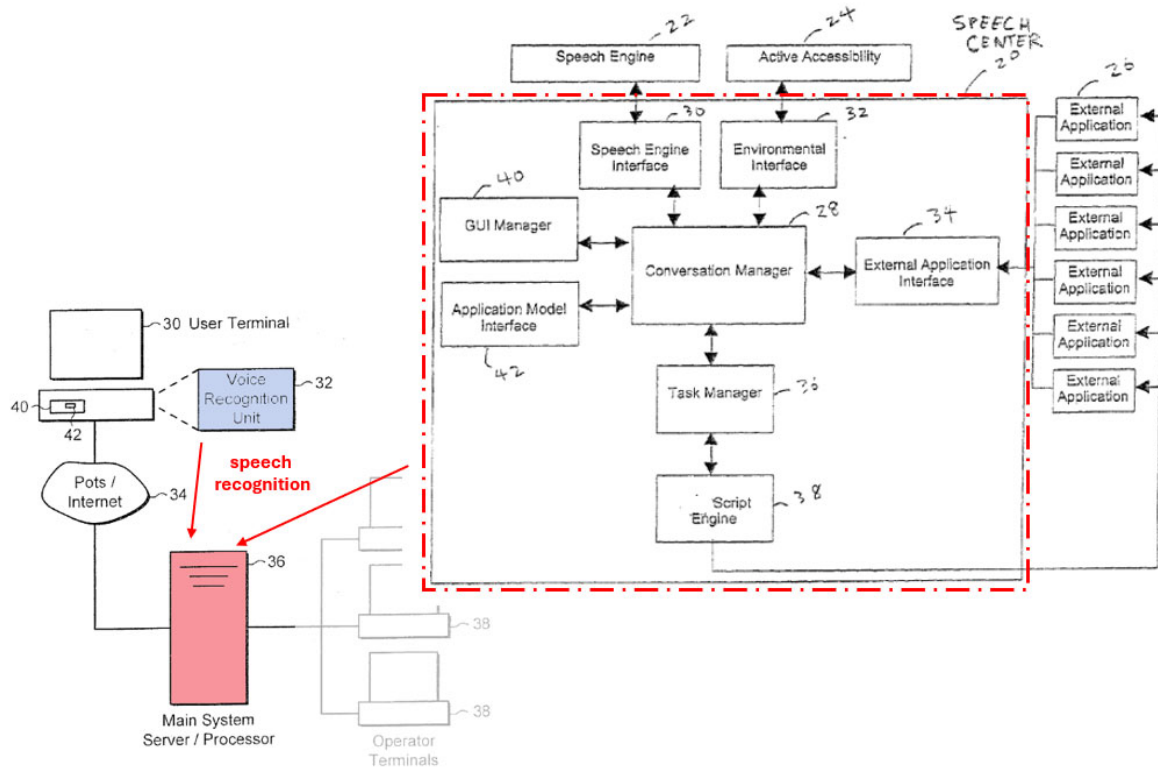
Thus, Barbara's system is capable of receiving an "*utterance*" from a user that "*includes one or more words that have different meanings in different contexts*" [1A.2]/[13A.2]/[25A.3].

3. Conversational Speech Engine [1D.1]/[13D]/[25B]

The Barbara-Ross combination discloses "*a conversational speech engine*", consistent with the '681 patent. (EX-1003, ¶¶308-313; §IV.B.3; EX-1001, 7:37-40, Figures 1-2.)

The system of the Barbara-Ross combination has a similar structure where the "voice recognition engine" provides the "best guesses of the voice recognition engine as to what was said, with some indications of their likelihood" to the "software for receiving, interpreting and correcting the incoming information from the user terminal." (EX-1007, ¶¶85-87, 90.) These functions are provided by a

conversation manager, as taught by Ross. (EX-1008, ¶¶22, 32.) The conversation manager is “a conversational speech engine.” (EX-1003, ¶311.)



Barbara, Figure 1; Ross, Figure 2

As taught by Barbara and Ross, the “voice input unit” (e.g., microphone and/or speech recognition engine) is separate from the “conversational speech engine.” (EX-1003, ¶312.) Specifically, Ross teaches conversation manager 28 interfaces with speech engine 22, through speech engine interface 30. (See EX-1008, ¶24; EX-1003, ¶312.)

Ross discloses a digital processor that “hosts and executes a speech center system 20, conversation manager 28, and speech engine 22 in working memory.”

(EX-1008, ¶22; *see also* EX-1007, ¶85 (“[i]ncluded in the main server are one or more service processors having software for receiving, interpreting and correcting the incoming information from the user terminal”).) Thus, the Barbara-Ross combination discloses “*the conversational speech engine includes one or more processors*” [25B]. (EX-1003, ¶313.)

4. Short-Term Shared Knowledge [1B.1]-[1B.2]/[13B.1]-[13B.2]/[25C.1]-[25C.2]

In the Barbara-Ross combination, the “best guesses of the voice recognition engine as to what was said are evaluated” in the conversational speech processing component (“*conversational speech engine*”) “against an information database(s) and knowledge base.” (EX-1007, ¶88.) The information database stores “personal details or preferences from what the user has said” and includes non-personal information, such as “weather reports or stock prices.” (EX-1007, ¶89.) Barbara also discloses “[t]he correct utterance text and the correct intention are stored in the knowledge base.” (EX-1007, ¶106.) Both are included in the main server. (EX-1007, ¶89, Fig. 7.) Barbara’s system uses the information from both the information database and knowledge base to evaluate possible requests that the user may have meant. (EX-1007, ¶¶90-93.)

The Barbara-Ross combination uses Ross’s data organization, for example the knowledge base is short-term storage that uses Ross’s conversational records to

store (“*accumulat[es]*”) each utterance “along with the results of its semantic analysis.” (EX-1008, ¶57; EX-1007, ¶106.) This information “is eventually purged from the conversational record **when it is no longer relevant to active goals** and after some predefined period of time has elapsed.” (EX-1008, ¶57.) Thus, the Barbara-Ross combination discloses “*accumulating short-term shared knowledge ...*” including “*knowledge about the utterance received during the current conversation*” [1B.1]-[1B.2]/[13B.1]-[13B.2]/[25C.1]-[25C.2]. (EX-1003, ¶314-318.)

5. Long-Term Shared Knowledge [1C]/[13C]/[25D]

Barbara records (“*accumulates*”) “**personal details or preferences from what has been said**” in its “information database(s).” (EX-1007, ¶88; EX-1007, ¶98 (“context information ... is captured as the user interacts with the system”).) For example, Barbara’s system infers a speaker “who asks about their dog ten times a day” is a dog owner (personal detail) and therefore “might be considered very likely to be referring to it” during a future conversation. (*See, e.g.*, EX-1007, ¶98.) As a further example, based on stored user profile data, the system determines “[s]omeone who is not involved in financial markets is unlikely to be asking about the Nikeii”, implying the system infers preference/employment information about the user. (EX-1007, ¶98.) Thus, this information, derived “from what has been said”, is “*knowledge about one or more past conversations with the*

user.” (EX-1003, ¶330.) Because this information database content “is built up over time”, the accumulated information is “*long-term shared knowledge.*” (EX-1007, ¶88; EX-1003, ¶321; EX-1001, 5:10-13 (“Long-term shared knowledge may **include explicit and/or implicit user preferences**”).)

The Barbara-Ross combination therefore discloses [1C]/[13C]/[25D]. (EX-1003, ¶¶319-322.)

6. Determining An Intended Meaning [1D]

In the Barbara-Ross combination, a received utterance is “**interpreted to determine what the user wanted, i.e. the intent of the spoken word,**” by the conversational processing component (“*conversational speech engine*”). (EX-1007, ¶82.) Thus, Barbara discloses limitation [1D]. (EX-1003, ¶¶323-324.)

a) Identifying A Context [1D.1]/[13D]/[25E]

Barbara discloses “*identifying a context.*” Barbara explains “[c]ontextual information ... is captured as the user interacts with the system, in much the same way as interactions work in real life.” (EX-1007, ¶98.) For example,

if the user does not have a dog called Toto (or at least hasn’t told the system about it), they’re unlikely to be referring to it.

...

A user who asks about their dog ten times a day might be considered very likely to be referring to it.

(EX-1007, ¶98.)

As a further example, if “the user asks for the date of Toto’s last flea vaccination, it is much more likely that they are referring to Toto the dog, rather than Toto the restaurant...” (EX-1007, ¶148.) That is, Barbara teaches an utterance’s context (pets versus restaurant) is identified from the recent dialog history which includes a reference associated with the context-domain pets (e.g., “flea vaccination”)(“*short-term shared knowledge*”). Barbara also teaches an utterance’s context is identified from user profile information accumulated from prior conversations, the “*long-term shared knowledge*” (e.g., “Toto” determined to mean “dog” due to user profile indication user has a dog named Toto. (EX-1003, ¶¶331-332.)

Thus, the Barbara-Ross combination discloses [1D.1]/[13D]/[25E]. (EX-1003, ¶¶325-333.)

b) Establishing The Intended Meaning [1D.2]-[1D.3]/[13E]-[13F]/[25F]

The Barbara-Ross combination discloses limitations [1D.2]-[1D.3]/[13E]-[13F]/[25F]. (EX-1003, ¶¶343-46.)

Barbara’s system interprets an utterance “to determine what the user wanted, i.e. the intent of the spoken word.” (EX-1007, ¶82.) Specifically, the utterance is “evaluated into a set of possible requests that the user may have actually meant.” (EX-1007, ¶¶90-93) “Contextual information” is then used to “evaluate[] the likelihood that the user really is asking each of those things.” (EX-1007, ¶98.) That

is, Barbara selects one of multiple potential meanings based on the context and in doing so, “[a]mbiguities are resolved.” (EX-1007, ¶98.)

For example, “[i]f the user asks for the date of Toto’s last flea vaccination,” Barbara’s system uses the identified context domain (“pet”) to determine “it is much more likely that they are referring to Toto the dog, rather than Toto the restaurant.” (EX-1007, ¶148.) Accordingly, the user’s “*intended meaning*”, e.g., the word Toto is “*disambiguated*” using the identified context domain, e.g., “pet.” (EX-1003, ¶¶334-37.)

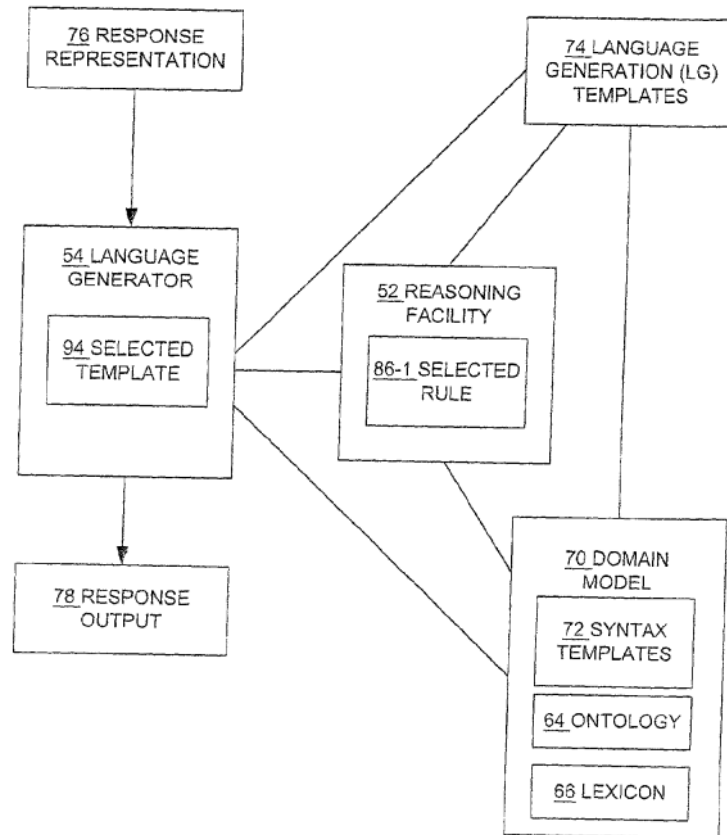
7. Generating A Response [1E.1]-[1E.2]/[13G.1]-[13G.2]/[25G]

The Barbara-Ross combination discloses limitations [1E.1]-[1E.2]/[13G.1]-[13G.2]/[25G]. (EX-1003, ¶¶338-348.)

In Barbara, after the intended meaning is established, the system “performs the correct intention,” e.g., by “*generating a response*” to the user. (EX-1007, ¶107.) For example, in Barbara, “feedback is given by the application visibly responding” or “the appliance speak[ing] back or perform[ing] some action or both.” (EX-1007, ¶107; *see also* EX-1007, ¶¶119-62.) As discussed in §V.A.3, Barbara lacks details regarding system component used to generate a response. Ross provides these details.

In Ross, the reasoning facility within the conversation manager generates a response representation 76 (e.g., perform action and speak) “in response to the

recognized spoken utterance 15.” (EX-1008, ¶¶61.) A natural language response 78 is then generated from the response representation. (EX-1008, ¶¶61, 63, Figure 4.)



Ross, Figure 4

When acting as a “medical expert system”, Barbara’s system “ask[s] the user questions” and after receiving and interpreting the user’s response “**continu[es] to ask relevant questions, narrowing down the diagnosis** until a conclusion of satisfactory confidence is reached.” (EX-1007, ¶168.) By asking **relevant questions** after interpreting the user’s response to narrow down the diagnosis, the Barbara-Ross system’s “*conversational speech engine*” “*grammatically adapt[s]*

the response based on the intended meaning established within the identified context.” [1E.2]/[13G.2]/[25G]. (EX-1007, ¶107; EX-1003, ¶¶346-348; *see also* EX-1001, 5:61-65.)

C. Claims 5/17/29

1. Conversational Goal [5A]/[17A]/[29A]

The Barbara-Ross combination discloses limitations [5A]/[17A]/[29A]. (EX-1003, ¶¶349-357.)

Barbara identifies “*a conversational goal*”, e.g., a discrete task to be performed (EX-1001, 14:39-47), in response to a user utterance. For example, in Barbara, when the “[u]ser says: ‘Show me incoming messages’” (EX-1007, ¶120), the system identifies that “[t]he phrase ‘Show me’ is taught as **a display query** for the sub-phrase which comprises the rest of the sentence (ie ”Show me x“->SELECT(x)).” (EX-1007, ¶122; EX-1003, ¶¶351-352.)

Barbara also identifies “*roles associated with the user and one or more other participants in the current conversation.*” (EX-1003, ¶353.) For example, Barbara discloses a user as a recipient, e.g., a consumer of information. (*See, e.g.*, EX-1007, ¶124 (“WHERE(Recipient=@@CURRENTUSER)”).) Barbara consequently identifies the voice interface system as a supporter that provides input as requested, e.g., the incoming messages, to the user, since the current conversation is a query

to display information to the user. (See, EX-1007, ¶122 (“The phrase ‘Show me’ is taught as a display query...”); EX-1003, ¶353.)

Barbara also identifies “*an information allocation among the user and the one or more other participants in the current conversation.*” (EX-1003, ¶354.)

“The phrase ‘Show me’ is taught as a **display query**...” for “objects with ObjectType Message.” (EX-1007, ¶¶122-23.) Accordingly, a POSITA would have understood that Barbara’s system holds the requested information because the user’s utterance was interpreted as a request (display query) for the objects with type “Message” Held by the system. (EX-1003, ¶354.)

As another example related to expert systems, Barbara discloses “*identifying a conversational goal*” e.g., obtaining a medical diagnosis, “*roles*,” e.g., system is information dispenser/user is consumer, and an “*information allocation*” e.g., the system holds the diagnosis information while the user holds his/her personal information. (EX-1007, ¶¶167-68; EX-1001, 15:21-26; EX-1003, ¶¶355-357.)

2. Conversation Type [5B]/[17B]/[29B]

The Barbara-Ross combination discloses limitations [5B]/[17B]/[29B]. (EX-1003, ¶¶358-362.)

A “query” is a “*conversation type*.” (EX-1001, 15:14-34, 14:61-63 (explaining “a query conversation may include a conversational goal of getting a discrete piece of information or performing a particular task...”)) Barbara discloses

“classifying an utterance/current conversation into a conversation type” (e.g., query) *“based on an identified conversational goal”* (e.g., request to perform a task). For example, “[t]he phrase ‘Show me’ is taught as a display **query** for the sub-phrase which comprises the rest of the sentence (ie ”Show me x“->SELECT(x)).” (EX-1007, ¶122; EX-1003, ¶360.)

Barbara also discloses *“establish[ing] the intended meaning based on the conversation type.”* (EX-1003, ¶¶361-362.) For example, when Barbara’s system identifies a phrase “as a display query,” the system understands the user intention is to display the items referenced in “the sub-phrase which comprises the rest of the sentence (ie ”Show me x“->SELECT(x)).” (EX-1007, ¶122.) That is, Barbara’s system establishes the user’s intention was to show items (“x”) based on the *“conversation type,”* e.g., query. (EX-1007, ¶123.)

D. Claims 6-7/18-19/30-31

The Barbara-Ross combination discloses limitations [7A]/[19A]/[31A]. (EX-1003, ¶¶363-377.)

In Barbara, the voice recognition unit (*“speech recognition engine”*) receives an utterance spoken into, e.g., a microphone (*“voice input device”*), and provides *“the best guesses ... as to what was said, with some indications of their likelihood...”* (EX-1007, ¶86.) As an example, if the user says “How’s Tokyo doing?” the voice recognition engine may “have outputted the following two

possible recognitions, with roughly equal probabilities” to the “*conversational speech engine*”: (1) “How’s Tokyo doing?” and (2) “How’s Toto doing?” (EX-1007, ¶¶90-96.) These interpretations “might be, hypothetically, roughly equally likely.” (EX-1007, ¶97.)

The Barbara-Ross system then “evaluate[s] [user utterances] into a set of possible requests that the user may have actually meant.” (EX-1007, ¶90.) For example, the possible recognition “How’s Tokyo doing” results in three potential intention hypotheses: “[1-a] a request for the status of the Tokyo production line, [1-b] the weather in Tokyo or [1-c] where the Nikei is trading.” (EX-1007, ¶97.) However, the second possible recognition “How’s Toto doing” results in only “[2-a] a request for the status of the operator’s pet dog Toto.” (EX-1007, ¶97.) Therefore, “[t]he single utterance was expanded to two possibilities for text, which were expanded to four possible intentions.” (EX-1007, ¶97.)

Barbara resolves ambiguities (such as the above) by evaluation the set of possible interpretations, scored by likelihood. (EX-1007, ¶¶98-99.) Barbara’s system calculates a “**combined likelihood** for each possible request” based on (1) “likelihood from the voice recognition engine that the text was actually what was said”, (2) “likelihood from the service’s knowledge base [*“short-term shared knowledge”*] that the request was what was meant given what was said”, and (3) “estimation from the user’s information base [*“long-term shared knowledge”*] of

which intents are likely.” (EX-1007, ¶¶90-93.) The combined likelihood is “*a degree of certainty about the intent*” and incorporates “*a confidence level associated with the preliminary interpretation*” [7B]/[19B]/[31B]. (EX-1003, ¶¶372-86.)

Therefore Barbara discloses claims [6]/[18]/[30] and limitations [7B]/[19B]/[31B]. (EX-1003, ¶¶363-377.)

E. Claims 8/20/32

The Barbara-Ross combination discloses “*grammatically or syntactically adapt[ing] the response based on the conversation type*” [8]/[20]/[32]. (EX-1003, ¶¶378-383.)

As noted above, a query is a “*conversation type*.” Barbara provides an example where “a user asks their phone if they have any e-mail.” (EX-1007, ¶110.) This utterance is classified as an “abstract intention of ‘*query* for message objects that have not yet been read’.” (EX-1007, ¶110.) The system “repl[ies] with a brief summary of the new messages.” (EX-1007, ¶110; EX-1003, ¶379.) By summarizing the messages, Barbara grammatically adapts its response based on the query. (EX-1003, ¶380.)

Further, asking only “relevant questions” when collecting information for medical diagnosis (EX-1007, ¶168) is another example where Barbara adapts a response based on conversation type. (EX-1003, ¶¶381-382.)

Additionally, Ross's "language generation (LG) module **uses syntax templates ... to generate questions** and responses to the user" and "uses rules to select [these] syntax templates to use for a given goal or propositions..." (EX-1008, ¶6.) Accordingly, the Barbara-Ross system adapts the response through use of rule-selected syntax templates, e.g., based on the "*conversation type*," e.g., query type. (EX-1008, ¶46; EX-1003, ¶383.)

F. Claims 11-12/23-24/35-36

Barbara discloses or at least suggests the "multi-modal input" claims [11]/[23]/[35] and "multi-modal output" claims [12]/[24]/[36]. (EX-1003, ¶¶384-389.)

Barbara describes a "first system, in which documents are electronically scanned" (i.e., an OCR input; "*non-voice input*"). (EX-1007, ¶26, Figure 1.) Barbara's system further includes a display screen at the user terminal (i.e., a visual output; "*non-voice output*"). (*See, e.g.*, EX-1007, Figures 1, 6.) Barbara teaches this system can be "additionally or alternatively" used "to create a voice interface where the end user system responds to spoken natural language commands" (i.e., a speech ("*voice*") input/output). (EX-1007, ¶82, *see also*, ¶166.) That is, Barbara teaches a system with multiple modalities for input and output—a "*multi-modal*" system.

Although Barbara does not provide an explicit example use of a “*multi-modal input*” that “includes one or more voice inputs relating to the utterance”, Barbara’s teachings are suggestive of such a use. (EX-1003, ¶386.) For example, a POSITA would understand based on Barbara’s teaching that the “user system responds to spoken natural language commands” to indicate the user could speak the command “scan” to cause the OCR input (“*non-voice input*”) to the system. (EX-1003, ¶386.) That is, the spoken “utterance” relates to the “*non-voice input*” (OCR-scan). (EX-1003, 386.) Moreover, the use of multi-modal inputs as disclosed in Barbara were well known to those skilled in the art by the filing date of the ’681 patent. (EX-1020, 381 (stating “[m]ultimodal systems provide a more flexible interface for users who can choose the most appropriate interface to meet their requirements and preferences” and “[m]ultimodal systems that permit more than one input mode can also provide for better handling of errors, particularly errors produced by the speech recognition component”).)

Barbara teaches or at least suggests claims [11]/[23]/[35]. (EX-1003, ¶387.)

Barbara describes the interplay between these modalities in exemplary systems for providing user responses. For example, in response to an utterance, the user terminal “may either speak back or perform some action ... or both.” (EX-1007, ¶107.) That is, the multi-modal output provides “*one or more tasks executed to process a request identified from the intended meaning.*” (EX-1003, ¶389.)

Barbara also provides extensive discussion of an e-mail interface where the user speaks, e.g., “Show me incoming messages” and messages meeting the query are displayed at the user terminal (a “non-voice output” that “relates to the utterance”). (See, e.g., EX-1007, ¶¶119-162, see also, ¶107 (“feedback is given by the application visually responding”).) As an example, Barbara explains “if a user asks their phone if they have any e-mail, the system may reply with a brief [spoken] summary of the new messages” whereas “[i]f the user asks their television/computer”, which has a display screen, “the user’s inbox may also be displayed.” (EX-1007, ¶107.) That is, the system both provides a spoken summary and a visual display (i.e., a “multi-modal” output).

Thus, Barbara discloses claims [12]/[24]/[36] (EX-1003, ¶¶384-89.)

VI. GROUND 3: Barabara-Ross-O’Neill Combination Renders Claims 2-4, 9, 14-16, 21, 26-28, and 33 obvious.

A. Combination Overview

1. O’Neill

O’Neill discloses a “system for conducting a dialogue with a user” using a DiscourseManager implementing “a strategy for confirming the semantic contents of a user’s utterances.” (EX-1009, Abstract, ¶¶105, 110.) For example, the DiscourseManager determines the response “to new, modified or negated information from the user” and “ensures that the information is at least implicitly confirmed.” (EX-1009, ¶111.)

O'Neill further discloses a structure for "knowledge-storing components." (EX-1009, ¶¶172-73.) Specifically, O'Neill discloses that the DiscourseManager "maintain[s] a record of the evolving dialogue in the form of a stack (i.e. the DiscourseStack 123) of dialogue frames." (EX-1009, ¶174.)

2. Motivation To Combine

A POSITA would have been motivated (i) to combine O'Neill's teachings to confirm what a user has spoken to verify a received utterance with the Barbara-Ross system and (ii) combine O'Neill's teaching regarding storing utterances of an on-going conversation in a stack with Ross's teaching of storing dialog history in a "conversational record." (EX-1003, ¶396.)

A POSITA would have been motivated to make these combinations for numerous reasons. First, the Barbara-Ross system invites the combination because it lacks the ability to confirm with the speaker a received utterance that the system is unable to interpret. (EX-1007, ¶110.) A POSITA would have understood such a capability is desirable to improve the user's experience when an utterance is incomplete, environmental noise impairs the system's ability to interpret the utterance, or the utterance is insolubly ambiguous. (EX-1003, ¶397.) McTear, for example, stresses it is "necessary in spoken dialogues with computers [to perform verification] given the possibility of recognition and understanding errors." (EX-1020, 119.) O'Neill further confirms the motivation, noting its system performs

verification of a user's intention in view of "technical difficulties with speech recognition." (EX-1009, ¶110; EX-1003, ¶397.)

Ross similarly lacks implementation details of its conversational record. A POSITA would have been motivated to implement the conversational record in the Barbara-Ross systems as a stack to improve processing efficiency in the system. O'Neill explains that the dialog history is maintained "in the form of a stack" such that "dialogue frames that can be added to or retrieved" from the stack "as required." (EX-1009, ¶174; EX-1003, ¶398.) Implementing Ross's conversational record as a stack would index utterances according to their recency, i.e., the last utterance would be at the top. (EX-1003, ¶398.) A POSITA would have been motivated to make the combination to improve processing efficiency because references (e.g., missing information/anaphora) are more likely to be found in more recent utterances than older ones, and therefore, the system need not parse many items in the conversational record. (EX-1003, ¶398.)

Both modifications are nothing more than use of known technique (well-known processes for implementing speech recognition) to improve a similar device (the Barbara-Ross speech recognition system) in the same way (implementing O'Neill's DiscourseManager to confirm the contents of a user's utterances and a stack to facilitate parsing). (EX-1003, ¶399.)

Barbara, Ross, and O’Neil are analogous art to the ’681 Patent. Each pertains to the same field of endeavor, i.e., speech recognition systems. (EX-1001, 1:8-9; EX-1007, ¶7; EX-1008, Abstract; EX-1009, ¶2; EX-1003, ¶400.)

A POSITA would have had a reasonable expectation of success in the combination and the results of the combination would have been predictable. Modifying the Barbara-Ross combination in the two ways discussed above are merely details on how to implement Barbara’s conversational system and would have been well within the skill set of a POSITA. (EX-1003, ¶401.)

B. Claims 2/14/26

The Barbara-Ross-O’Neil combination discloses “*populat[ing] a short-term context stack with information about the utterance received during the current conversation*” [2]/[14]/[26]. (EX-1003, ¶¶402-405.)

As explained in §V.A.3, the Barbara-Ross combination stores the current conversation utterances along with the results of their semantic analysis and the captured contextual information in a “conversational record.”

O’Neill “maintain[s] **a record of the evolving dialogue in the form of a stack (i.e. the DiscourseStack...) of dialogue frames** that can be added to or retrieved from the DiscourseStack 123 as required.” (EX-1009, ¶174.) In the Barbara-Ross-O’Neill combination the dialog history (utterances and associated information) are stored in a stack within Ross’s conversation record, such that each

utterance is added to or retrieved from the stack as required. That is, the first utterance is at the bottom and the most recent is at the top. (EX-1003, ¶404.) A POSITA would have also understood that the conversational record would be a “*short-term context stack*”, because the conversational record only holds information for a “predefined period of time.” (EX-1003, ¶404.)

C. Claims 3/15/27

The Barbara-Ross-O’Neill combination renders claims [3]/[15]/[27] obvious. (EX-1003, ¶¶406-409.)

Ross discloses “information is eventually **purged from the conversational record** when it is no longer relevant to active goals and after some predefined period of time has elapsed.” (EX-1008, ¶57; EX-1003, ¶406.) As noted above (§VI.A.2), the conversational record in the combination is implemented as a “*short term context stack*”, as taught by O’Neill. (EX-1003, ¶406.) A POSITA would have understood the time it takes for information to be “no longer relevant to active goals” is a “*psychologically appropriate amount of time*”, consistent with the disclosures of the ’681 patent. (EX-1001, 13:65-66; EX-1003, ¶408.)

D. Claims 4/16/28

Barbara-Ross-O’Neill combination renders claims obvious claims [4]/[16]/[28]. (EX-1003, ¶¶410-413.)

Barbara's **"information database(s) is built up over time**, by recording **personal details or preferences from what has been said**, or from information directly entered into the system by the user." (EX-1007, ¶88.) Accordingly, Barbara discloses *"updating a long-term profile associated with the user"* (e.g., the user's personal details/preferences) with *"information about the utterance received during the current conversation"* (e.g., from what has been said). In Barbara's system, the phrase "what has been said" would correspond to what has been said during the at-that-time current conversation, which would be recorded in the "conversational record." (EX-1003, ¶¶411-412.)

The "conversational record" includes "results of [an utterance's] semantic analysis," (i.e., what has been said) which "is eventually purged from the conversational record." (EX-1008, ¶57; EX-1003, ¶412.) Thus, the long-term shared knowledge derived from an utterance in a dialog that is stored in and expires from the stack. (EX-1003, ¶422; EX-1007, ¶411.)

Accordingly, Barbara-Ross-O'Neill combination discloses updating *a long-term profile associated with the user* (e.g., the user's personal details or preferences) with *relevant data associated with the information expired from the short, term context stack* (e.g., results of [an utterance's] semantic analysis). (EX-1003, ¶413.)

E. Claims 9/21/33

Barbara-Ross-O’Neil combination renders obvious claims [9]/[21]/[33].
(EX-1003, ¶¶414-419.)

O’Neill explains “any system conducting a complex transaction must have a strategy for confirming the semantic contents of a user’s utterances.” (EX-1009, ¶110.) O’Neill’s DiscourseManager implements such a strategy. (EX-1003, ¶¶416-418.) The DiscourseManager “ensures that the information is at least implicitly confirmed. ... (e.g. ‘So, you’re arriving at the Hilton in Belfast on Friday’)” and further “ensures that ... a confirmatory response ... is immediately followed by a question requesting further information (e.g. ‘what day are you departing?’).” (EX-1009, ¶¶111-12.) By immediately following the user’s response with a question requesting further information, O’Neill “*syntactically adapts the response to influence a subsequent reply utterance that the conversational speech engine expects from the user during the current conversation.*” (EX-1003, ¶¶414-419.)

VII. GROUND 4: Barbara-Ross-Franco Combination Renders Obvious Claims 10, 22, 34

Barbara-Ross do not explicitly disclose “*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*” [10B]/[22B]/[34B]. Franco provides this missing

teaching and the Barbara-Ross-Franco combination renders claims [10]/[22]/[34] obvious. (EX-1003, ¶422.)

For the reasons discussed in §V.D. for [7A]/[19A]/[31A], the Barbara-Ross combination also discloses “*generating multiple preliminary interpretations of the utterance at a speech recognition engine coupled to the voice input device and the conversational speech engine.*” [10A.1]/[22A.1]/[34A.1].

Barbara discloses that after each interpretation is scored for likelihood (“*confidence level*”), “a heuristic is adopted, which is able to differentiate between those interpretations that have a high probability of being correct and those that do not.” (EX-1007, ¶99; §V.D (discussing “*confidence level*” disclosure).) A POSITA would have understood based on Barbara’s teachings that the interpretation with the highest likelihood score is selected, among the possible interpretations, as an “*initial interpretation of the utterance.*” (EX-1003, ¶433.)

The Barbara-Ross system does not explicitly disclose permitting a user to correct the incorrect interpretation without manual intervention. Franco provides this missing teaching. In Franco, a first hypothesis is returned “to the speaker (e.g., via a textual display or a speech synthesis interface) ... for review/confirmation.” (EX-1010, ¶14.) Franco then “determines whether the speaker has indicated the need for a correction of the first hypothesis” where “the need for a correction is indicated by a subsequent speaker utterance” which “may repeat the entire first

utterance, or may repeat only one or more portions of the first utterance that were incorrectly recognized in the first hypothesis.” (EX-1010, ¶15.)

A POSITA would have been motivated to combine Franco’s teachings allowing a user to correct a “first hypothesis (i.e., a proposed recognition solution) based on the first utterance” (EX-1010, ¶14) with the Barbara-Ross system. (EX-1003, ¶242.) As noted above, the Barbara-Ross interface voice interface lacks a technique for correcting an improperly interpreted utterance without manual/operator intervention.

Both Barbara and Franco recognized the need to incorporate error correction in a voice interface system: (EX-1007, ¶3 (“the voice recognition software available at present is limited in its usefulness, primarily because it is not yet sophisticated or intelligent enough to be able to eliminate errors”); EX-1010, ¶5 (“need in the art for a method and apparatus for error correction in speech recognition applications.” (Franco, ¶5.) A POSITA would have been motivated to incorporate Franco’s automated solution to remove the need for human operators which reduces costs and improves the speed and useability of the Barbara-Ross combination. (EX-1003, ¶424.)

The results of the combination would have been predictable to a POSITA and the POSITA would have had a reasonable expectation of success because Franco discloses that its error recognition module can be “implemented as a

physical device or subsystem that is coupled to a processor through a communication channel.” (EX-1010, ¶27; EX-1003, ¶428.)

Barbara, Ross, and Franco are analogous art to the ’681 Patent. Each pertains to the same endeavor, i.e., speech recognition systems. (*Compare* the ’681 Patent, 7:24-33 *with* Barbara, ¶¶7, 12, 82 *and* Ross, ¶Abstract *and* Franco, ¶2.)

In the combination using Franco’s error correction technique, when the user recognizes an incorrect interpretation was made by the system, they could repeat either “How’s Tokyo” or “Tokyo.” A POSITA would have understood the “conversational record” in the Barbara-Ross-Franco system (“*short-term shared knowledge*”) retains the “*multiple preliminary interpretations of the utterance*” because Ross discloses “[e]ach utterance is indexed in the conversational record 60, **along with the results of its semantic analysis.**” (EX-1008, ¶57.) However, when the modified Barbara-Ross-Franco system determines “the need for a correction of the first hypothesis” (EX-1008, ¶15), the interpretation “How’s Toto doing” would be removed from the short-term memory, because Ross discloses “information is eventually purged from the conversational record **when it is no longer relevant to active goals**” (EX-1008, ¶57) and that initial interpretation is no longer relevant to the active goal of determining the interpretation of the utterance “How’s Tokyo doing.” (EX-1003, ¶444.)

Accordingly, Barbara-Ross-Franco discloses “*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*”

Because the initial interpretation “How’s Toto doing” is removed from short-term memory, it is no longer a possible interpretation of the received utterance. Franco discloses “if the phrase with the highest probability is identical to the first hypothesis, which the speaker has indicated as being incorrect, then the method ... selects the phrase with the second highest probability as an alternative combined hypothesis.” (EX-1010, ¶26.) Accordingly, the Barbara-Ross-Franco combination discloses “*the conversational speech engine determines the intended meaning based on one of the multiple preliminary interpretations having a next highest confidence level*” [10C]/[22C]/[34C].

VIII. Stipulation

To simplify the Fintiv analysis, and following the precedential decision in *Sotera Wireless, Inc. v. Masimo Corp.*, Petitioner stipulates that, if IPR is instituted, Petitioner will not pursue in the related district court proceeding any ground that Petitioner raised or reasonably could have raised against the challenged claims during the instituted IPR.

IX. Mandatory Notices

A. Real Party In Interest

The real parties-in-interest are Samsung Electronics Co. Ltd. and Samsung Electronics America, Inc.

B. Related Matters

To the best of Samsung's knowledge, the following is a list of judicial or administrative matter that would affect, or be affected by, a decision in the proceeding:

- *VB Assets, LLC v. Samsung Electronics Co. Ltd. et al.*, 2:24-cv-00828-JRG-RSP (E.D. Tex., filed Oct. 9, 2024)
- *VB Assets, LLC v. Amazon.com Servs. LLC*, No. 1:19-cv-01410-MN (D. Del., filed July 29, 2019)
- *VB Assets, LLC v. SoundHound AI, Inc.*, No. 1:24-cv-01279-MN (D. Del., filed Nov. 21, 2024)
- *VB Assets, LLC v. Apple, Inc.*, No. 1:24-cv-01368-MN (D. Del., filed Dec. 13, 2024)

C. Notice of Counsel and Service Information

Pursuant to 37 C.F.R. §§ 42.8(b)(3), 42.8, Petitioners designate the following lead and backup counsel:

Lead Counsel	Back Up Counsel
Lori Gordon (Reg. No. 50,633) GOODWIN PROCTER LLP 1900 N St. N.W. Washington, D.C. 20036 Phone: (202) 346-4000 Fax: (202) 346-4444 Gordon-ptab@goodwinlaw.com	Douglas Kline (Reg No. 35,574) GOODWIN PROCTER LLP 100 Northern Avenue Boston, MA 02210 Tel: (617) 570 1000 Fax: (617) 523-1231 dkline@goodwinlaw.com
	Srikanth K. Reddy (<i>pro hac vice</i> application to be filed) GOODWIN PROCTER LLP 100 Northern Avenue Boston, MA 02210 Tel: (617) 570 1000 Fax: (617) 523-1231 sreddy@goodwinlaw.com
	Brian T. Drummond (Reg. No. 68,414) GOODWIN PROCTER LLP 100 Northern Avenue Boston, MA 02210 Tel: (617) 570 1000 Fax: (617) 523-1231 bdrummond@goodwinlaw.com
	Theodoros Konstantakopoulos (Reg. No. 74,155) GOODWIN PROCTER LLP 620 Eighth Avenue New York, NY 10018 Phone: (212) 813-8800 Fax: (212) 355-3333 tkonstantakopoulos@goodwinlaw.com

Petitioners consent to electronic service at the email addresses listed in the table above.

X. Conclusion

IPR is respectfully requested.

Date: April 25, 2025

Respectfully submitted,

/Lori A. Gordon/

Lori A. Gordon (Reg. No. 50,633)
Goodwin Procter LLP
1900 N Street, N.W.
Washington, D.C. 20036
Phone: (202) 346 4435
Gordon-ptab@goodwinlaw.com

Lead Counsel for Petitioner

CERTIFICATE OF WORD COUNT UNDER 37 CFR §42.24(d)

Pursuant to 37 C.F.R. §42.24(a), Samsung hereby certifies that portions of the above-captioned Petition for *Inter Partes* Review of U.S. Patent No. 8,073,681 in accordance with and reliance on the word count provided by the word-processing system used to prepare this Petition, that the number of words in this paper is 13,994. Pursuant to 37 C.F.R. §42.24(a), this word count is in compliance and excludes the table of contents, table of authorities, mandatory notices under §42.8, certificate of service, certificate of word count, appendix of exhibits, and any claim listing. This word count was prepared using Microsoft Word.

Date: April 25, 2025

Respectfully submitted,

/Lori A. Gordon/

Lori A. Gordon (Reg. No. 50,633)
Goodwin Procter LLP
1900 N Street, N.W.
Washington, D.C. 20036
Phone: (202) 346 4435
Gordon-ptab@goodwinlaw.com

Lead Counsel for Petitioner

CERTIFICATE OF SERVICE

The undersigned certifies that a true copy of the Petition for *Inter Partes* Review of U.S. Patent No. 8,073,681 together with all exhibits identified in the above Table of Exhibits and Petitioners' Power of Attorney, have been served on the Patent Owner via FedEx Next Business Day Delivery on the below date, at the correspondence address of record as listed on the Patent Center:

Sheppard Mullin Richter & Hampton LLP
650 Town Center Drive
10th Floor
Costa Mesa, CA
Attorneys for VB Assets, LLC

Courtesy copies of the foregoing document were served by e-mail on the following counsel of record for Patent Owner in the concurrent litigation:

Theodore Stevenson, III (TX 19196650)
Brady Randall Cox (TX 24074084)
ALSTON & BIRD LLP
2200 Ross Avenue, Suite 2300
Dallas, TX 75201
Telephone: (214) 922-3400
Facsimile: (214) 922-3899
ted.stevenson@alston.com
brady.cox@alston.com

Erik John Carlson (CA 265167)
Caleb J. Bean (CA 299751)
Srishti Ghosh (CA 354393)
ALSTON & BIRD LLP
350 South Grand Avenue, 51st Floor

Los Angeles, CA 90071
Telephone: (213) 576-1000
Facsimile: (213) 576-1100
erik.carlson@alston.com
caleb.bean@alston.com
maddie.ghosh@alston.com

Natalie C. Clayton (NY 4409538)
ALSTON & BIRD LLP
90 Park Avenue, 15th Floor
New York, NY 10016
Telephone: (212) 210-9400
Facsimile: (212) 210-9444
natalie.clayton@alston.com

David Greenbaum (NY 2947455)
GREENBAUM LAW LLC
210 Allison Court
Englewood, NJ 07631
david@greenbaum.law

VBA-Samsung@alston.com

Attorneys for VB Assets LLC

Date: April 25, 2025

/Lori A. Gordon/

Lori A. Gordon

Lead Counsel for Petitioner

APPENDIX A – LIST OF CLAIMS

[1P] A method for providing a cooperative conversational voice user interface, comprising:

[1A.1] receiving an utterance at a voice input device during a current conversation with a user,

[1A.2] wherein the utterance includes one or more words that have different meanings in different contexts;

[1B.1] accumulating short-term shared knowledge about the current conversation,

[1B.2] wherein the short-term shared knowledge includes knowledge about the utterance received during the current conversation;

[1C.1] accumulating long-term shared knowledge about the user,

[1C.2] wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

[1D] determining an intended meaning for the utterance, wherein determining the intended meaning for the utterance includes:

[1D.1] 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

[1D.2] establishing the intended meaning within the identified context,

[1D.3] 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

[1E.1] generating a response to the utterance,

[1E.2] 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.

[5A] 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

[5B] 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

[7A] generating a preliminary interpretation of the utterance at a speech recognition engine coupled to the voice input device and the conversational speech engine,

[7B] 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.

[10A.1] 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,

[10A.2] wherein an initial interpretation of the utterance comprises one of the multiple preliminary interpretations having a highest confidence level; and

[10B] 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,

[10C] 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.

[13P] 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:

[13A.1] receive an utterance at a voice input device, during a current conversation with a user,

[13A.2] wherein the utterance includes one or more words that have different meanings in different contexts; and

[13B.1] accumulate short-term shared knowledge about the current conversation,

[13B.2] wherein the short-term shared knowledge includes knowledge about the utterance received at the voice during the current conversation

[13C.1] accumulate long-term shared knowledge about the user,

[13C.2] wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

[13D] 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;

[13E] establish an intended meaning for the utterance within the identified context,

[13F] 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

[13G.1] generate a response to the utterance,

[13G.2] 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.

[17A] 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

[17B] 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

[19A] generate a preliminary interpretation of the utterance at a speech recognition engine,

[19B] 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.

[22A.1] 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,

[22A.2] wherein an initial interpretation of the utterance comprises one of the multiple preliminary interpretations having a highest confidence level; and

[22B] 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,

[22C] 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.

[25P] A system for providing a cooperative conversational voice user interface, comprising:

[25A.1] a voice input device configured to

[25A.2] receive an utterance during a current conversation with a user,

[25A.3] wherein the utterance includes one or more words that have different meanings in different contexts; and

[25B] a conversational speech engine, wherein the conversational speech engine includes one or more processors configured to:

[25C.1] accumulate short-term shared knowledge about the current conversation,

[25C.2] wherein the short-term shared knowledge includes knowledge about the utterance received during the current conversation

[25D.1] accumulate long-term shared knowledge about the user,

[25D.2] wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

[25E] identify a context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge;

[25F] 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

[25G] 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.

[29A] 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

[29B] 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

[31A] generate a preliminary interpretation of the utterance,

[31B] 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.

[34A.1] The system of claim 25, further comprising a speech recognition engine configured to: generate multiple preliminary interpretations of the utterance,

[34A.2] wherein an initial interpretation of the utterance comprises one of the multiple preliminary interpretations having a highest confidence level; and

[34B] 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,

[34C] 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.