

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

**DECLARATION OF STUART LIPOFF IN SUPPORT OF 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.	Qualifications.....	2
II.	Understanding of Relevant Legal Principles	12
III.	'681 Patent	14
	A. Technical Background.....	15
	1. Language and Dialog	17
	2. Interpreting Meaning in Dialog	20
	3. Interactive/Spoken Dialog Systems	24
	B. Overview of the '681 Patent.....	46
	C. Level of Ordinary Skill in the Art	48
	D. Prosecution History	49
	E. IPR2020-01367	50
	F. Claim Construction.....	53
	1. Speech Recognition Engine	54
	2. Amazon Litigation Claim Construction.....	54
IV.	GROUND 1: Combination of SmartKom and Kobsa Renders Obvious Claims 1-9, 11-21, 23-33, And 35-36.....	57
	A. Overview of the Combination	57
	1. SmartKom	57
	2. Kobsa.....	60
	3. Motivation to Combine	61
	B. Independent Claims 1, 13, 25.....	64
	1. Preambles [1P]/[13P]/[25P]	67
	2. Receiving an Utterance [1A]/[13A]/[25A]	72
	3. Conversational Speech Engine	80
	4. Short-Term and Long-Term Knowledge	84
	5. Determining an Intended Meaning [1D].....	101

TABLE OF CONTENTS

(continued)

Page ii

6.	“Generating” Limitation [1E.1]-[1E.2]/[13G.1]- [13G.2]/[25G]	119
C.	Claims 2, 14, 26.....	124
D.	Claims 3, 15, 27.....	126
E.	Dependent Claims 4, 16, 28	129
F.	Claims 5, 17, 29.....	132
1.	Conversational Goal.....	132
2.	Conversation Type	137
G.	Claims 6-7, 18-19, 30-31.....	140
H.	Claims 8-9, 20-21, 32-33.....	146
I.	Dependent Claims 11-12, 23-24, 35-36	149
V.	GROUND 2: The Combination Of Barbara And Ross Renders Obvious Claims 1, 5-8, 11-13, 17-20, 23-25, 29-32, And 35-36	154
A.	Overview of the Combination	154
1.	Barbara	154
2.	Ross	161
3.	Motivation To Combine.....	169
B.	Independent Claims 1, 13, and 25	171
1.	Preambles [1P]/[13P]/[25P]	174
2.	Receiving An Utterance [1A]/[13A]/[25A]	179
3.	Conversational Speech Engine	185
4.	Short-Term Shared Knowledge [1B.1]-[1B.2]/[13B.1]- [13B.2]/[25C.1]-[25C.2]	189
5.	Long-Term Shared Knowledge [1C]/[13C]/[25D]	191
6.	Determining An Intended Meaning [1D].....	193
7.	Identifying A Context [1E/13D/25E]	194
8.	Establishing The Intended Meaning [1D.2]-[1D.3]/[13E]- [13F]/[25F].....	197

TABLE OF CONTENTS

(continued)

Page iii

9.	Generating A Response [1E.1]-[1E.2]/[13G.1]- [13G.2]/[25G]	200
C.	Claims 5, 17, and 29.....	205
1.	Conversational Goal [5A/17A/29A]	205
2.	Conversation Type [5B/17B/29B]	209
D.	Claims 6, 18, and 30.....	212
E.	Claims 7, 19, and 31	214
F.	Claims 8, 20, and 32	217
G.	Claims 11-12, 23-24, and 35-36.....	220
VI.	GROUND 3: The Combination Of Barbara, Ross, And O'Neill Renders Obvious Claims 2-4, 9, 14-16, 21, 26-28, And 33	223
A.	Overview of the Combination	223
1.	O'Neill.....	223
2.	Motivation To Combine.....	225
B.	Claims 2, 14, and 26.....	227
C.	Claims 3, 15, and 27	228
D.	Claims 4, 16, and 28.....	230
E.	Claims 9, 21, and 33.....	232
VII.	GROUND 4: The Combination Of Barbara, Ross, and Franco Renders Obvious Claims 10, 22, And 34	234
A.	Overview of the Combination	234
1.	Franco.....	234
2.	Motivation To Combine.....	237
B.	Claims 10, 22, 34.....	239
1.	Multiple Preliminary Interpretations [10A/22A/34A].....	239
2.	Update Short-Term Shared Knowledge.....	245
3.	Determine Intended Meaning	248
VIII.	Conclusion	249

I, Stuart Lipoff, declare as follows:

1. I have been retained by Goodwin Procter LLP on behalf of Petitioners Samsung Electronics Co. Ltd. and Samsung Electronics America, Inc. (“Petitioners”) to provide this Declaration concerning technical subject matter relevant to the petition for *Inter Partes* Review (“Petition”) of U.S. Patent 8,073,681 (“the ’681 patent”).

2. I am over 18 years of age. I have personal knowledge of the facts stated in this Declaration and could testify competently to them if asked to do so.

3. I have been asked to provide my technical opinions regarding how a person of ordinary skill in the art would have understood the claims of the ’681 patent at the time of the alleged invention, which I have been asked to assume is the 2006 timeframe. For purposes of whether the teachings of the prior art render the claims of the ’681 patent obvious, I have been asked to assume the date of October 16, 2006. I have also been asked to provide my technical opinions on how concepts in the ’681 patent specification relate to claim limitations of the ’681 patent.

4. In reaching the opinions provided herein, I have considered the ’681 patent, its prosecution history, and the references cited in my Declaration and the Appendix. I have also drawn on my own education, training, research, knowledge, and personal and professional experience.

5. In general, I have been asked to cite to the specification of a patent or patent publication using the following formats: (Patent, Col:Line Number(s)) or (Patent, Paragraph Number(s)). For example, the citation ('681 patent, 1:1-10) points to the '681 patent specification at column 1, lines 1-10. Also, for convenience, I have been asked to use italics to denote limitations from the challenged claims. Unless otherwise noted, all emphasis is added.

6. I have personal knowledge of the facts and opinions set forth in this declaration and believe them to be true. If called upon to do so, I would testify competently thereto. I have been warned that willful false statements and the like are punishable by fine or imprisonment, or both.

7. I am being compensated for my time at my standard consulting rate. I am also being reimbursed for expenses that I incur during the course of this work. My compensation is not contingent upon the results of my study and analysis, the substance of my opinions, or the outcome of any proceeding involving the '681 patent. I have no financial interest in the outcome of this matter or in any litigation involving the '681 patent.

I. Qualifications

8. I believe I am well qualified to render useful opinions on this matter. I will briefly summarize my knowledge, training, and experience here. A more

detailed summary of my background, education, experience, and publications is set forth in my curriculum vitae (CV), which is submitted as EX-1004.

9. I earned a Bachelor of Science degree in Electrical Engineering in 1968 and a second Bachelor of Science degree in Engineering Physics in 1969, both from Lehigh University. I earned a Master of Science degree in Electrical Engineering from Northeastern University in 1974, and then a Master of Business Administration degree from Suffolk University in 1983.

10. I hold a Federal Communications Commission (“FCC”) General Radiotelephone License. I also hold a Certificate in Data Processing from the Institute for the Certification of Computing Professionals (“ICCP”), which is supported by the Association for Computing Machinery (“ACM”).

11. I am also a registered professional engineer (PE) in the Commonwealth of Massachusetts and in the State of Nevada.

12. I am a fellow of the Institute of Electrical and Electronics Engineers (“IEEE”) Consumer Electronics, Communications, Computer, Circuits, and Vehicular Technology Groups. I have been a member of the IEEE Consumer Electronics Society National Board of Governors (formerly known as the Administrative Committee) since 1981, and I was Boston Chapter Chairman of the IEEE Vehicular Technology Society from 1974 to 1976. I served as the 1996-1997 President of the IEEE Consumer Electronics Society, and from 1999 to 2018, I

served as Chairman of the Society's Technical Activities and Standards Committee and as Vice President of Publications for the Society, and since 2018, I have served as Vice President of Standards and Industry Activities for the Society. Currently, I am The Historian for the Society. I have also served as an Ibuka Award committee member for the IEEE's Award in the field of consumer electronics.

13. I have prepared and presented numerous papers at the IEEE and at other professional meetings. For example, in fall 2000, I served as general program chair for IEEE's Vehicular Technology Conference on advanced wireless communication technology. I have organized sessions at The International Conference on Consumer Electronics, and I was the 1984 program chairman. I conducted an eight-week IEEE-sponsored short course on Fiber Optics System Design. I received IEEE's Centennial Medal in 1984, and I received IEEE's Millennium Medal in 2000.

14. As Vice President and Standards Group Chairman for the Association of Computer Users ("ACU") from 1980 to 1983, I served as the ACU representative to the ANSI X3 Standards Group. From 1976 to 1978, I served as Chairman of the task group on user rule compliance for the FCC's Citizens Advisory Committee on Citizen's Band Radio.

15. Over the last twenty-five years, I have been a member of the Society of Cable Television Engineers, the Association for Computing Machinery, and The

Society of Motion Picture and Television Engineers. From 2001 to 2004, I served as a member of the USA advisory board to the National Science Museum of Israel.

16. In 1998, I presented a short course on international product development strategies as a faculty member for Technion Institute of Management in Israel. From 2001 to 2003, I served as a member of the board of directors of The Massachusetts Future Problem Solving Program.

17. I am a named inventor on seven United States patents and have several publications on data communications in publications, including Electronics Design, Microwaves, EDN, the Proceedings of the Frequency Control Symposium, Optical Spectra, and IEEE publications.

18. During my professional career, dating from 1969 to the present, I have been heavily engaged in the study, analysis, evaluation, design, and implementation of products and technology associated with consumer electronics and electronic appliances. A particular focus of my professional activities has been improving the man-machine interface including voice, speech, and speaker recognition for man-machine interactions. I also have extensive experience in studying foundation technologies and the applications supporting home automation, home appliance control, residential energy management, and home security and monitoring.

19. For approximately three years, from 1969 to 1972, I served as Project Engineer for Motorola's Communications Division, where I had project design

responsibilities for paging and wireless communication products. Projects I worked on while employed at Motorola included work on paging systems that included digital voice storage, voice compression, and voice synthesis. I also worked on projects that interfaced wireless data communications terminals to public safety computer systems for mobile data retrieval and data entry.

20. For approximately four years, from 1972 to 1976, I served as Section Manager for Bell & Howell Communications Company, where I also had project design responsibilities for paging and wireless communication products. The projects I supported included covert audio intelligence systems that recognized speech and activated digital voice compression recording systems. I also led projects for voice-based radio paging systems that recorded speech input, processed the speech to remove silence, processed the speech to digitally compress the speech, and store and forward the speech upon demand from DTMF or computer keyboard retrieval from the servers.

21. For twenty-five years, from 1976 to 2001, I worked for Arthur D. Little, Inc. (ADL), where I became the Vice President and Director of Communications, Information Technology, and Electronics (CIE) and served in that role for 10 years, from 1991 to 2001. At ADL, I was responsible for the firm's global CIE practice in laboratory-based contract engineering, product development, and technology-based consulting. I was also involved in multiple pioneering efforts to identify and explore

customer-to-business and business-to-business electronic commerce and transactions information processing opportunities (e-commerce). These projects involved technology assessment and analysis, as well as developing architectures and systems to support multiple applications, and typically involved an information retrieval component.

22. While at ADL, I worked on several projects involving the combination of voice interfaces (including speech recognition and voice audio output) and information retrieval as well as working on projects for utilities, service providers, and consumer electronics OEMs for home automation and energy management. For example, over the course of three years in the early-1990s, I worked on a project for Bolt Beranek and Newman (BB&N), where I evaluated and benchmarked technology for a voice input/output application that allowed end users (e.g., travel agents) to use speech inputs to interact with airline reservation databases to retrieve information about travel reservation options, where the results were returned to the user in an audible message. This system included a natural language front-end speech-interface module with speech recognition that used pre-defined recognition grammars to convert the end user's speech into structured commands supported by an airline reservation system. As another example, over the course of three years in the mid-1990s, I worked on a project for Texas Instruments that applied a speech-recognition interface for a variety of applications that retrieved information from

database servers. My work for electric, gas, and water utilities included a focus on remote and automatic meter reading and energy management of appliances and residential HVAC systems.

23. Other projects that I worked on at various points in my 25 years at ADL and afterwards that involved speech recognition technologies included the following:

24. Over the course of three years, in the early 1990s, I worked on a voice-interface project developing spoken digit telephone number recognition and voiceprint matching for Sprint's long-distance alternative access telephone services.

25. Over the course of a year, in the late 1980s, I worked on a voice interface project evaluating the processing power needed to perform various voice recognition applications by Rockwell Semiconductor's signal processing technology.

26. Other projects that I worked on at various points in my twenty-five years at ADL that involved information-retrieval technologies included the following:

27. Over the course of fifteen years, starting in the early 1980s, I worked on a project for the United States Postal Service (USPS), where we developed a real-time automated postal teller system that served as an interface between end users

and the USPS's information systems. This system included voice prompts for the vision impaired.

28. Over the course of two years, in the early 1990s, I worked on a project for the grocery industry consortium of The Food Marketing Institute and The Grocery Manufacturers Association, where I developed standards used by the industry for direct exchange electronic data interchange (DEX/UCS EDI). This project involved developing a business model for vendors who make direct store delivery of merchandise to retail stores (e.g., fast-moving goods that do not come via a warehouse such as soda, meat, bread) so that legacy paper receipts and signature could be captured on hand-held portable computers and then uploaded to the vendors' billing computers at some later time to generate invoices.

29. Over the course of two years, in the early 1990s, I worked on a project for MasterCard and Visa, where I supported a project exploring the applications and security issues associated with the use of smart cards in eCommerce. This project explored both physical security properties of the card media as well as issues associated with the back-end information processing servers. For example, I explored electronic watermarks resident on a credit card where the watermark digital content was captured at point of sale and then uploaded to the back-end credit card processor so that the card media could be authenticated as genuine.

30. Over the course of two years, in the late 1970s, I worked on a project for a multi-client consortium of newspapers and information publishers, where I participated in a project to understand opportunities for electronic home information and transaction services using both dedicated videotext terminals as well as home computers. The project was focused on providing end consumers in ordinary households with the means to read newspapers, interact with classified advertising, send messages, access telephone directories, and search for information.

31. Over the course of two years, in the late 1980s, I worked on a project in support of a multi-client study of new opportunities for financial industry firms, where I studied the security and encryption requirements to support electronic banking. This work involved consideration of counterfeit projection for media, physical security of systems, and the development of security protocols for home banking videotex terminals. This project focused on providing ordinary end consumer households with the means to conduct home banking in a secure and simple interface via a key board and visual display.

32. I also have had extensive experience in public and private network wired and wireless voice telecommunications while employed by Motorola, Bell & Howell, and Arthur D. Little, and while self-employed. In the course of these telecommunications projects, ranging from 1969 to the present, I have encountered a number of applications where audio input and voice are used to activate devices,

for example, for the purpose of saving battery power by entering into low-power, so-called “sleep” modes. These projects have involved the design of cellular telecommunications systems that implement industry-standard means of entering lower-power modes in the absence of voice.

33. Examples of other projects that I worked on at various points in my 25 years at ADL and afterwards that involved home automation, control, and energy management include:

34. Support of a unique two-way power line carrier system from electric utility substations to served homes for automatic meter reading and load management to cycle air conditioners and heating systems. The system was deployed by New England Electric Systems and sold worldwide by Emerson Electric.

35. I have served for over thirty years as a member of the IEEE International Conference on Consumer Electronics (ICCE) annual technical conference, and as a member of its technical program committee track on home automation and control. In this capacity, I reviewed dozens of research paper contributions under consideration for presentation at the ICCE.

36. For Honeywell, I performed a project to study alternative technologies suitable for in-home network energy management and environmental control including wireless thermostats. The project involved consideration of power line carriers, wireless radio frequencies, and wired communications systems.

37. For Cambridge Silicon Radio (UK), I worked with the client to develop a prioritized list of applications for their Bluetooth component offerings. The project mapped applications into specific target customers and based upon an analysis that considered the market needs with CSR's capabilities, a prioritized roadmap of products was developed to steer the R&D portfolio. A primary focus of the project was to explore short-range wireless applications for home automation and control.

II. Understanding of Relevant Legal Principles

38. I am not a lawyer, and I will not provide any legal opinions. Although I am not a lawyer, I have been advised certain legal standards are to be applied by technical experts in forming opinions regarding the meaning and validity of patent claims.

39. I understand that a patent claim is invalid if it is anticipated or obvious in view of the prior art, and that a claim can be unpatentable even if all of the requirements of the claim cannot be found in a single prior-art reference. I further understand that invalidity of a claim requires that the claim be anticipated or obvious from the perspective of a person of ordinary skill in the art at the time the invention was made.

40. I have been informed that a patent claim is invalid if it would have been obvious to a person of ordinary skill in the art. In analyzing the obviousness of a claim, I understand the following factors may be taken into account: (1) the scope

and content of the prior art; (2) the differences between the prior art and the claims; (3) the level of ordinary skill in the art; and (4) any so called “secondary considerations” of non-obviousness, if they are present. I am not aware of any evidence of secondary considerations of non-obviousness relevant to the ’681 patent. I reserve the right to supplement this Declaration if Patent Owner (“PO”) introduces evidence of secondary considerations of non-obviousness.

41. I understand that to prove that prior art or a combination of prior art renders a patent obvious, it is necessary to:

- (1) identify the particular references that, singly or in combination, make the patent obvious;
- (2) specifically identify which elements of the patent claim appear in each of the asserted references; and
- (3) explain why a person of ordinary skill in the art would have combined the references, and how they would have done so, to create the inventions claimed in the patent. I further understand that exemplary rationales that may support a conclusion of obviousness include:

- combining prior art elements according to known methods to yield predictable results;
- simple substitution of one known element for another to obtain predictable results;
- use of known technique(s) to improve similar devices (methods or products) in the same way;

- applying a known technique to a known device (method or product) ready for improvement to yield predictable results;
- “obvious to try” – choosing from a finite number of identified, predictable solutions with a reasonable expectation of success; known work in one field of endeavor may prompt variations of the work for use in either the same field or a different field based on design incentives or other market forces if the variations are predictable to a person of ordinary skill in the art; and
- some teaching, suggestion, or motivation in the prior art that would have led a person of ordinary skill in the art to modify the prior art reference or to combine prior art reference teachings to arrive at the claimed invention.

42. I have been informed that, in considering obviousness, hindsight reasoning derived from the patent-at-issue may not be used.

III. ‘681 Patent

43. The ‘681 patent, filed on October 16, 2006, “relates to a cooperative conversational model for a human to machine voice user interface.” (EX-1001, 1:8-9.) By 2006, cooperative conversation systems, referred to as “spoken dialog systems” or “interactive dialog systems”, were well-known and in use, as I discuss further in §III.A below. Indeed, as I discuss below, the ‘681 patent and the challenged claims do nothing more than co-opt these well-known concepts.

A. Technical Background

44. Prior to the filing date of the '681 patent (October 16, 2006), conversational systems like the type described in the '681 patent had been known and in use for at least a decade. In fact, long before the filing date, an “increasing number of telephone services” were being “offered in a fully automatic way with the help of speech technology.” (EX-1021, ix.) The systems providing these services, “called spoken dialogue systems (SDSs), possess speech recognition, speech understanding, dialogue management, and speech generation capabilities, and enable a more-or-less natural spoken interaction with the human user.” (EX-1021, ix.) In such human-machine interactions¹, the users are able to “cope with the limitations” of the machine-side of the conversation and reach the intended goal, “provided that both interlocutors behave in a cooperative way.” (EX-1021, ix.)

45. A large number of books and papers, published prior to October 16, 2006, exist in the area of interactive/spoken dialog systems which describe various aspects of such systems. In this section, I cite primarily to the following:

¹ Human-to-human communication is commonly referred to as “human-to-human interaction” or “HHI,” and human-to-machine communication is commonly referred to as “human-to-machine interaction” or “HMI”. (Möller, 9.)

- *Quality of Telephone-Based Spoken Dialogue Systems* by Sebastian Möller, Springer Science + Business Media (2005)
- *Spoken Dialogue Technology: Toward the Conversational User Interface* by Michael F. McTear, Springer (2004)
- *Spoken Language Processing* by Xuedong Huang et al., Prentice Hall PTR (2001)

However, each of these books cites dozens of other books and papers providing additional detail on various topics.

46. In this technical background section I provide an introduction to the state of speech recognition systems as of the filing date of the '681 patent (October 16, 2006). I begin with an introduction to concepts of language and dialog that span both human-to-human interactions and human-to-machine interactions to provide an understanding of the concept of terms like “utterance,” “turn,” “dialog,” “lexicon,” and “grammar.” I then discuss the importance of interpreting the meaning of words, which, as I explain further below, often requires considering more than just the words spoken, but shared background knowledge and context. It is here that I introduce Paul Grice and the advancement he brought to the field through his lectures and books directed to our understanding of conversational speech. I then explain how, as of 2006, the idea of using concepts from the study of human conversation,

such as Grice’s work, to aid in designing dialog systems that used speech recognition was well known in the NLP and voice interface community.

47. I conclude the technical background section with an introduction to the state of the art of conversational systems as of 2006, with a technical description of the systems and their components.

1. Language and Dialog

48. Spoken language “is very different to written language” because “people typically do not follow rigid syntactic and morphological constraints in their utterances.” (EX-1016, 49.) The lack of written language formality “in spontaneous spoken language makes linguistic analysis by machine both more difficult than, and different to, analysis of written language.” (EX-1016, 49.)

49. As Huang explains “[i]t is natural to assume that the *sentence* is a clear and simple chunking unit for dialog, by analogy with written communication.” (EX-1018, 860 (emphasis in original).) But, because “sentences are artificially delimited in written text, researches in dialog communication usually speak of the *utterance* as the basic unit.” (EX-1018, 860 (emphasis in original).)

50. In the context of conversational speech, one or more utterances can comprise a “turn.” A turn in conversation refers to the time one speaker holds the floor, often including multiple utterances—the smaller units of speech. Turns are shaped by interactional dynamics and may bundle several utterances together (EX-

1016, 49.) Thus, turns structure the flow of dialogue, beyond just the speech segments themselves. Turns are viewed as “building blocks for constructing a common task-oriented understanding among participants.” (EX-1018, 861.) Turns are normally “easily recognized by the machine.” (EX-1016, 49.)

51. Dialog is best understood not as a series of isolated turns but as structured discourse in which utterances are interdependent and contextually grounded. A dialog “consists of at least two turns, one by each speaker”, and a coherent dialog “will exhibit discourse phenomena” whose interpretation “depends on the dialogue context.” (EX-1020, 46.) This requires systems to track references across turns, including maintaining a record of entities introduced into the dialog and interpreting pronouns accordingly. In some cases, this involves explicit dialogue history; in others, it draws on broader background or world knowledge. (EX-1020, 46.)

52. Two examples of how dialog presents these issues are anaphora and ellipsis. Anaphora is a linguistic phenomenon where a word or phrase refers back to another word or phrase mentioned earlier in the discourse. In dialog systems, anaphora resolution is essential for understanding references like pronouns. Ellipsis occurs when part of a sentence is omitted because it is understood from context. In dialogue, users often leave out information that has already been established, and the system must infer the missing parts. (EX-1026, 62, 64.)

53. A key structuring device in dialogue is the adjacency pair—predictable turn sequences like question–answer or greeting–greeting—that capture mutual expectations and conversational obligations (EX-1020, 65.) These pairs are not merely formalities but reflect deeper discourse coherence, and they often aggregate into larger goal-driven patterns known as dialog games. (EX-1020, 65-68.) These games model the intentional structure of conversation and accommodate embedded exchanges, such as clarifications or corrections, within a larger purpose. (EX-1020, 65-68.)

54. This discourse-based view supports the design of dialog systems that are sensitive to context, capable of tracking referents, and able to respond appropriately to sub-dialogues and conversational shifts. As McTear notes, modeling these structures is essential for preserving coherence, managing initiative, and aligning system behavior with user expectations in naturalistic interaction. (See EX-1020, 48-49.)

55. Common terminology exists in the field of interactive/spoken dialog systems. For example, a “*lexicon* is a list of words, a *vocabulary*, annotated with syntactic (including morphological) and semantic features.” (EX-1016, 49 (emphasis in original).) “*Grammars* describe how words may be combined into phrases and sentences.” (EX-1016, 50 (emphasis in original).) “*Semantics* are

abstract representations of the meanings of words, phrases and sentences.” (EX-1016, 50 (emphasis in original).)

2. Interpreting Meaning in Dialog

a. Grice’s Cooperative Principle

56. Understanding meaning in speech recognition systems involves far more than transcribing sounds into text. It includes interpreting a speaker’s **intent**, grounding expressions in context, and producing coherent responses. Central to this process is the application of pragmatic principles—most notably Paul Grice’s Cooperative Principle (CP) and its accompanying maxims, which offer a framework for making conversation intelligible and purposeful. While Grice originally articulated these ideas in the context of human-to-human dialogue, these principles have long been understood to be critical to the design of human-machine interaction, particularly in systems aiming for natural, task-oriented spoken interaction, such as spoken dialog systems.

57. The Cooperative Principle reflects the foundational idea that people engaged in a conversation will typically work together—that is, cooperate—to communicate effectively and understand each other. As Grice put it in a 1967 lecture delivered at Harvard University and later published as part of the book “Syntax and Semantics”:

Our talk exchanges do not normally consist of a succession of disconnected remarks, and would not be rational if they did. They are characteristically, to some degree at least, cooperative efforts; and each participant recognizes in them, to some extent, a common purpose or set of purposes, or at least a mutually accepted direction.

(EX-1017, 45; *see also* EX-1021, 47 (According to Grice, “communication is cooperative action which requires that both parties have a minimal common purpose, or at least a mutually accepted direction.”).) At bottom, Grice’s idea was the fact that we generally cooperate when speaking with one another can be leveraged to better understand the meaning of what has been said.

58. Grice identified four “maxims” that he considered to define cooperative speech: Quantity, Quality, Relation, and Manner. (EX-1017, 45.) Bernsen emphasizes the relevance of Grice’s framework to speech interface design, noting that the maxims help designers understand and predict user behavior in spoken systems. (*See generally* EX-1016.) For example, users often expect system outputs to be relevant (Relation), truthful (Quality), clear (Manner), and sufficiently informative (Quantity). (EX-1016, 91.)

59. As Grice pointed out, the cooperative nature of conversational speech not only affects what speakers say, but correspondingly requires hearers to consider both what is literally said **and** what a speaker meant by uttering those words. This requires, among other things, an evaluation of the context for the utterance and shared assumptions that the speaker considered in making an utterance. Grice used

the term **implicature** to refer to the meaning of an utterance that is implied but not explicitly stated.

60. To use an example offered by Grice, if person A says “I am out of gas,” and person B replies “There’s a station around the corner,” B’s statement is only relevant (and thus cooperative) if the station referred to is a place that sells gas, and is actually open—thus B’s reply carries the implicature “there is a gas station around the corner, and it is open.” (EX-1017, 51.) Of course, B did not explicitly *say* the station is open, but A reasonably infers that meaning by assuming B’s adherence to the principle of cooperation. Grice’s insight was that much of human communication relies on such shared understandings—speakers imply additional meaning and listeners infer it, guided by the cooperative principle.

61. Before Grice articulated these ideas it was commonplace in linguistics to think of utterances as standalone structures with sentence-level meaning only. The focus of the analysis on a utterance-by-utterance level tended to be highly formal, logical, and literal. Grice’s advancement was to clearly articulate a theory of linguistics that focused on the concept of meaning as opposed to grammar or logical truth.

b. Grice’s Principle Applied to Human-Machine Interactions

62. Although Grice developed his ideas to describe human-to-human interactions (HHI), “they have fruitfully been used for addressing the problem of

cooperativity” in human-machine-interactions (HMI) as well. (EX-1021, 48.) By October 2006, decades of prior work in speech recognition envisioned systems that behave as cooperative conversational partners. To accomplish this, researchers in human-computer interaction extended Grice’s principles explicitly into design guidelines for dialog systems that used speech recognition.

63. For example, as early as 1988, Professor Susan Brennan, who was then a graduate student at Stanford University and working on a “cooperative and conversational” natural language interface at Hewlett Packard Labs, observed how it was “useful to evaluate human-computer communication in light of Grice’s cooperative principle and maxims.” (EX-1022, 1.) She was not the first to have this insight, indeed she noted that in the field of human-computer interaction systems, at that time, “the most promising strategies ... involve applying insights gained from psycholinguistics research in order to create better conversational human/computer interfaces.” (EX-1022, 1.)

64. A recently published peer-reviewed study investigated the extent to which Grice’s ideas have permeated the field of natural language processing. (EX-1019.) In *The Gricean Maxims in NLP-A Survey*, the authors noted how it had been known since at least 1987 that “[i]n order to build NLP systems that are able to use language beyond just its literal content, they need to incorporate pragmatic

capabilities.” (EX-1019, 470.) Most notably, the authors remarked, this meant incorporating Grice’s cooperative principle.

65. In sum, by the filing date of the ’681 patent the NLP and voice interface community widely understood that a cooperative conversational approach, grounded in shared knowledge and contextual inference, was crucial for any natural language system. This forms an important backdrop when examining the ’681 patent: the concepts it embodies were already part of the prior art landscape through both academic literature and earlier patent applications. Indeed, I understand that the Mr. Freeman, one of the individual listed as an inventor on the face of the ’681 patent acknowledged during a prior litigation that the claimed concepts of cooperative conversation came from Grice’s paper. (EX-1027, 288:8-292:4.)

3. Interactive/Spoken Dialog Systems

66. At a high level, the task of an interactive dialog system “is to enable and support the spoken interaction between the human user and the service offered by the application system.” (EX-1021, 19.) This high-level tasks leads to a number of sub-tasks performed by the system, including:

- coherence of the user input has to be verified, taking into account linguistic and task- or domain-related knowledge;
- communicative and task goals have to be negotiated with the user, and problems occurring during the interaction have to be resolved;

- references like anaphora or ellipses in the user's utterances have to be resolved;
- inferences which are reasonable in the communicative and task context have to be drawn, and the most probable user reaction has to be predicted.

(EX-1021, 19.)

67. Möller generally presents four types of systems that can be differentiated “[d]epending on the complexity of the dialogic interaction”:

- **Command systems:** “They are characterized by a direct and deterministic interaction. To each stimulus from one agent corresponds a unique response from the other. The response is independent of the state or context of each agent. This type of interaction is normally not considered as a dialogue, and is called a ‘tool metaphor’. Example: Pressing a key on the keyboard results in a character appearing on the screen.” (EX-1021, 19.)
- **Menu dialog systems:** “To this class belong simple question-answer user interfaces, where dialogue and task models are merged. The interaction is mainly system-directed, permitting only very little user initiative (e.g. barge-in). In contrast to command systems, several exchanges may be necessary in order to provoke one action of the application system. On the other hand, one user input can provoke different responses, depending on the internal state of the system, e.g. the current level in the menu structure. Example: So-called ‘Interactive Voice Response’ (IVR) systems which enable an interaction via Dual Tone Multiple Frequency (DTMF) or keyword recognition.” (EX-1021, 19.)
- **Spoken Dialog System (SDS):** “This narrow class of systems disposes of distinct and independent models for task, user, system, and dialogue. Context

information is taken into account using a particular knowledge base or dialogue history. Multiple types of references can be processed. An SDS may be capable of reasoning, of error or incoherence detection, internal correction, anticipation, and prediction.” (EX-1021, 19-20.)

- **Multimodal dialogue systems including speech:** “Systems of this class show the same characteristics as SDSs do, but in addition they are able to process and synchronize different modality information. Appropriate modalities for the individual interactions have to be selected, both at the system input and output side.” (EX-1021, 20.)

68. An interactive/spoken dialog system has a “modality” which “is simply a form (or mode) of representing information as output from, or input to, a computer system (Hovy and Arens, 1990).” (EX-1016, 77.) Some systems only support a single modality (e.g., speech) whereas others support multiple and are referred to as multimodal systems.

69. Spoken dialog systems, including multimodal dialog systems, are “an interface between a human user and an application system which uses speech as the interaction modality (Fraser, 1997).” (EX-1021, 18.) As such, the interface must be able to process information “coming from and going to the user through the speech-technology-based interface (voice user interface, VUI)” and information “coming from and going to the application system through a specialized (e.g., SQL-based) interface.” (EX-1021, 18.) The system “must achieve a number of actions in order

to be able to give a response, and the response will depend on the internal state of the system, or on the context of the interaction.” (EX-1021, 18.)

70. “What is meant by meaning or understanding?” (EX-1018, 853 (emphasis in original).) “Meaning is often a constellation that emerges from a conversational environment.” (EX-1018, 853.) Huang identifies four main interacting areas in spoken language understanding systems from which meaning arises:

- **Intent:** goals of listener and speaker in the interaction
- **Context:** the pressures, opportunities, interruptions, etc. of the interaction scene and communication media
- **Content:** the propositional or literal content of each utterance and the discourse as a whole
- **Assumptions:** what each participant can assume about other participants’ mental state, abilities, limitations, etc.

(EX-1018, 854.)

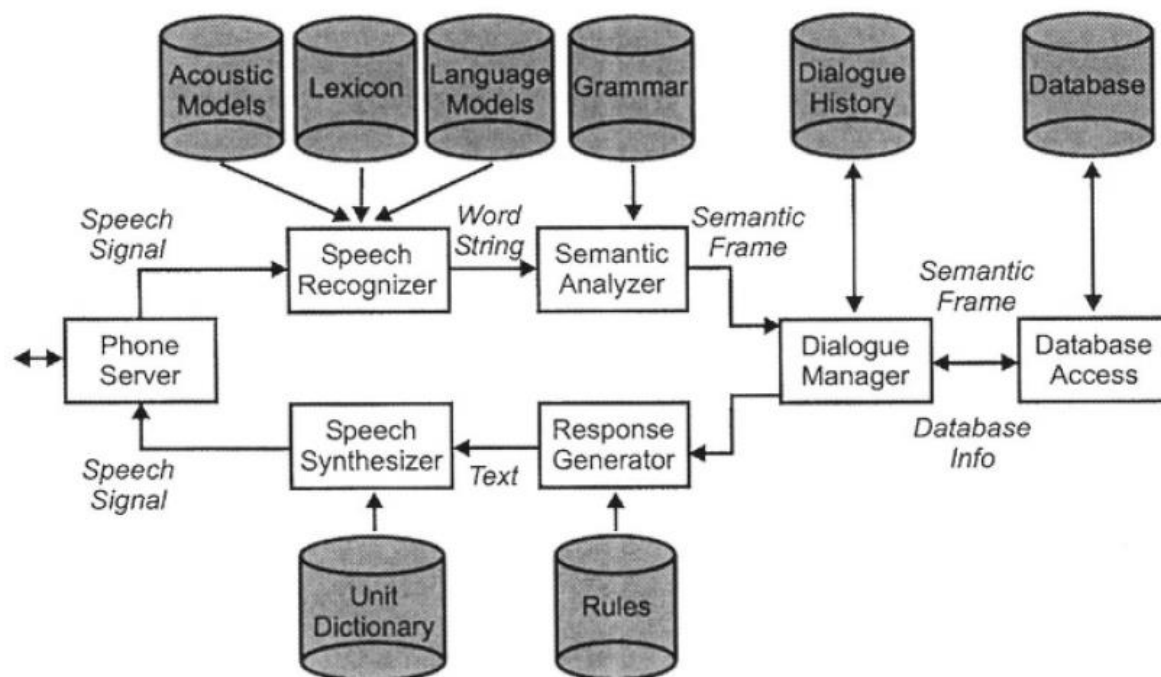
71. Most early spoken dialogue systems were designed for and operated over the telephone network, largely due to the widespread availability and standardized nature of telephony infrastructure. These systems typically employed a phone server interface to manage user access and interactions via voice calls, making them suitable for tasks like automated customer service and information retrieval. (EX-1021, 20.)

72. A spoken dialog system “consists of six major components which are accessed by the user” via an access device. (See, *e.g.*, EX-1021, 20.) These are:

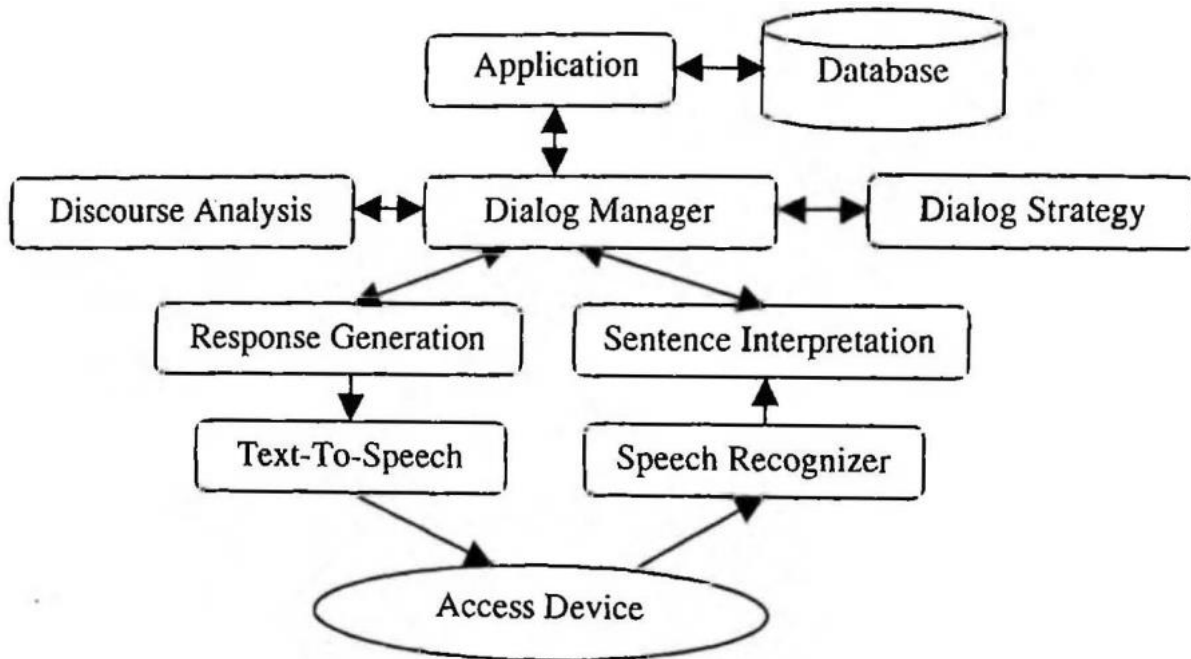
- **Phone Server:** Provides access to the user over the telephone network.
- **Speech Recognizer:** Converts the spoken input into a word string or word hypothesis graph.
- **Semantic Analyzer:** Interprets the recognized word string semantically, producing a semantic frame.
- **Dialogue Manager:** Interprets the semantic frame in context, manages dialogue history, and controls dialogue flow.
- **Response Generator:** creates natural language replies based on the dialogue manager’s output.
- **Speech Synthesizer:** Converts the response into speech, and transmits it back to the user.

73. Below I provide illustrations of such a system from three textbooks: Möller, Huang, and McTear. I describe these in further detail in the sections that follow. In the system, a “speech signal from the user is first processed to the speech recognizer.” (EX-1021, 20.) During this process, the speech signal is “transformed into a word string or a word hypothesis graph which is then semantically analyzed” by the semantic analyzer. (EX-1021, 20.) Huang refers to the semantic analyzer as a sentence interpretation component and McTear refers to it as the language understanding component. The output from the semantic analyzer component “is a semantic frame representing what has been ‘understood’ from the user’s utterance.”

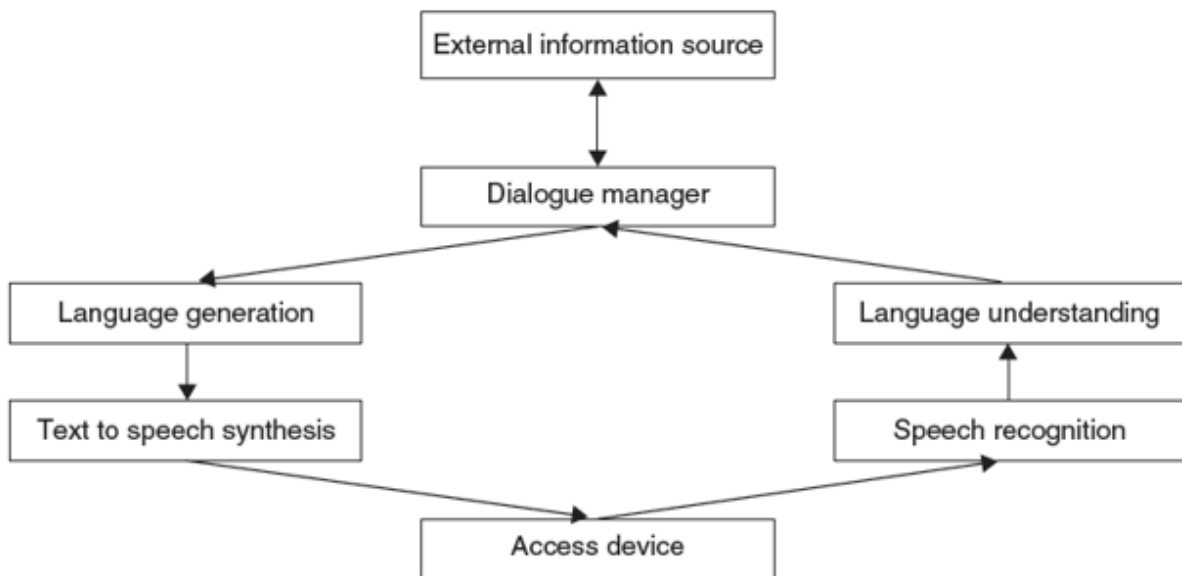
(EX-1021, 20.) The dialog manager “interpret[s] the semantic frame in the context of the dialogue and the task” and “keep[s] track of the dialogue history.” (EX-1021, 20.) When all relevant information has been collected, “a query to the underlying application (in this example a database) can be launched.” (EX-1021, 20.) The information from the application “ha[s] to be transformed into a response for the user” by the response generation module. (EX-1021, 20-21.) McTear refers to this component as the “language generation” component. This component “generates a response in text form, which is then transformed by the speech synthesizer into a speech signal which is transmitted to the human user.” (EX-1021, 21.)



Möller, Figure 2.4



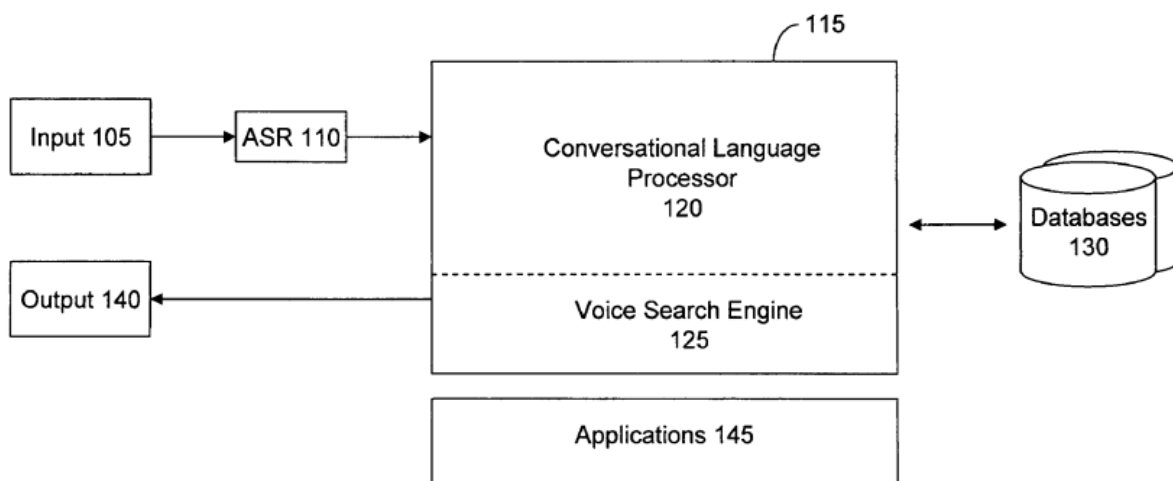
Huang, Figure 1.4



McTear, Figure 4.1

74. The six components described by Möller, McTear, and Huang “may be implemented in different ways.” (EX-1021, 21.) For example, the components can

be divided “into a speech platform (containing the phone server, the speech recognizer and the speech synthesizer), a voice application server (similar to a web application server, and containing the semantic analysis, the dialogue management and the response generation), and the application back-end (e.g., the database).” (EX-1021, 21.) I note this implementation is similar to the implementation described in the ’681 patent that I discuss in §III.B and is illustrated in Figure 1 of the ’681 patent (reproduced below).



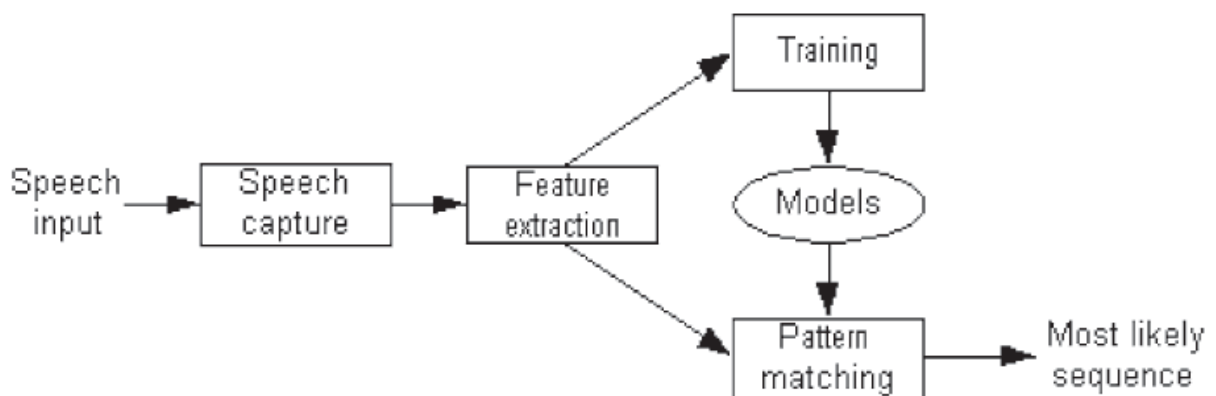
‘681 Patent, Figure 1

75. Further, in some implementations, some components may not be necessary, at least not as conceived by Möller, McTear, and Huang. For instance, while many systems at the relevant period were telephone based, that was not universally true and those that were not could use another voice-input device without a phone server.

76. In the next section I describe what are the key components of any speech dialogue system: the speech recognizer, the semantic analyzer, the dialogue manager, and the response generator and speech synthesizer.

a. Speech Recognition Component

77. The speech recognition component is commonly referred to as an automatic speech recognizer (ASR). The speech recognition process is illustrated by McTear's Figure 4.2 (below). Speech recognition begins when a user "speaks into a microphone or telephone handset." (EX-1020, 83.) The analog acoustic signal "is captured by the microphone and converted to an electric current, which is then passed to an analogue-to-digital converter within the sound card that translates the current into a stream of bits that represent the sound." (EX-1020, 83.) This part of the process, known as signal processing, "extract[s] the relevant set of features from the acoustic signal." (EX-1020, 83.)



McTear, Figure 4.2

78. More specifically, the acoustic signal is divided into frame, e.g., of about 10 ms, and “each frame is analysed for a number of features, such as the amount of energy at each of several frequency ranges, the overall energy in a frame, and differences from the previous frame.” (EX-1020, 83-84.)

79. The extracted features are then “classified as sequences of phonemes using an acoustic model, and these sequences of phonemes are then combined into words that represent the system’s estimate of what the user said.” (EX-1020, 83.) The process of phoneme and word identification involves an acoustic model “that shows how each word consists of a sequence of phonemes and how each phoneme relates to the values of the features extracted from the acoustic signal” and a “language model that specifies permissible sequences of words.” (EX-1020, 84.)

80. “The acoustic model captures variability in pronunciation using probabilities.” (EX-1020, 84.) Words may be pronounced in different ways by different individuals. (*See, e.g.*, EX-1020, 84.) For example, the word “tomato” may be pronounced “T AH M EY T OW” by American speakers or “T AH M AA T OW” by British speakers. (EX-1020, 84.) Hidden Markov Models (“HMMs”) are used in speech recognition systems “to represent [] variable patterns of speech.” (EX-1020, 85.)

81. The above acoustic model “is not sufficient on its own as a method for estimating a word or sequence of words given an acoustic signal.” (EX-1020, 86.)

Therefore, systems combine the acoustic model with “a language mode that contains knowledge about permissible sequences of words and which words are more likely in a given sequence.” (EX-1020, 86.) Two types of language models are commonly used—a “grammar (or finite state network) in which all the permissible word sequences in the application are specified” and an N-gram model “which provides statistical information on word sequences.” (EX-1020, 86.)

b. Language Understanding

82. The language understanding component “analyse[s] the output of the speech recognition component” and “assign[s] a meaning representation that can be used by the dialogue manager.” (EX-1020, 91.) Traditionally, language understanding involved “syntactic analysis, to determine the constituent structure of the recognised string (i.e., how the words group together), and semantic analysis, to determine the meanings of the constituents.” (EX-1020, 91.) Language understanding is problematic because of ambiguity in natural language and ill-formed input. (EX-1020, 91.)

83. McTear lists the following ways in which natural language can be ambiguous:

- **Lexical ambiguity:** “A word may belong to more than one part of speech, for example, ‘book’ can be a noun or a verb.” (EX-1020, 91.) This type of ambiguity “can usually be resolved within the context of

the other words in the sentence, as in the sentence ‘book a flight to London’.” (EX-1020, 91.)

- **Sense ambiguity:** “A word can have different meanings, for example, ‘bank’ can be a financial institution or the side of a river.” (EX-1020, 91.)
- **Structural ambiguity:** “The relationship between the phrases in the sentence is ambiguous, for example, ‘a flight to London arriving at 9.’ On a purely syntactic analysis, there are two possible readings, one in which the flight arrives at 9, and the other in which London arrives at 9.” (EX-1020, 91.)

84. In typical systems, the semantic analyzer uses predefined grammars, slot-filling methods, or statistical models to extract relevant information from user utterances. These might include details such as dates, locations, or actions, which are organized into semantic frames or attribute-value structures. For instance, in a travel booking system, an utterance like “I want to fly to Berlin next Monday” would be converted into a frame containing values for destination and travel date. (See EX-1021, 29-30.)

85. The language understanding component will often rely on a component known as a parser to assist in this process. The parser is a tool or algorithm that analyzes the grammatical structure of the recognized speech input. Its role is to break down the sentence into its syntactic components (like subjects, verbs, objects, etc.) and determine how those components relate to each other. This helps the system

understand what the user means, not just what words were spoken. For example, if a user says, “Book a flight to Paris next Friday,” the parser helps identify that:

“Book” is the action (verb),

“a flight” is the object,

“to Paris” indicates the destination,

“next Friday” refers to the date.

This structure can then be processed by the dialogue manager to determine the next system action. (*See* EX-1021, 26, 29-30.)

c. Dialog Management Component

In a spoken dialog system, the dialog manager is the “central component.” (EX-1020, 107.) The Dialog Manager “accepts spoken input from the user, produces messages to be communicated to the user, interacts with external knowledge sources, and generally controls the dialogue flow.” (EX-1020, 107.) Put differently, according to Möller, “It is the task of the dialogue manager to guarantee the smooth course of the dialogue, so that it is coherent with the task, the domain, the history of the interaction, with general knowledge of the ‘world’ and of conversational competence, and with the user.” (EX-1021, 27.) Core functions of the dialogue manager include:

- the collection of all information from the user which is needed for the task,
- the distribution of dialogue initiative,

- the provision of feedback and verification of information understood by the system,
- the provision of help to the user,
- the correction of errors and misunderstandings,
- the interpretation of complex discourse phenomena like ellipses and anaphoric references, and
- the organization of information output to the user.

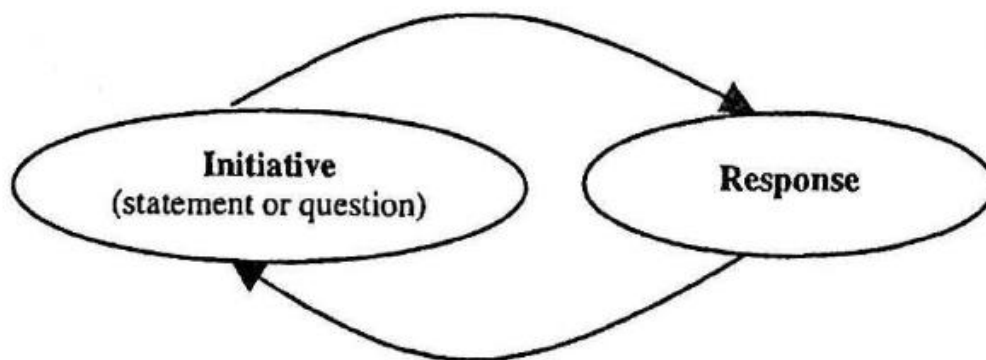
(EX-1021, 27.)

86. The Dialog Manager also serves as “a type of service controller which administers the follow of information between the different modules” in a spoken dialogue system, coordinating the flow of interaction by coordinating the flow of interaction by “interpret[ing] the semantic frame in the context of the dialogue history” and generating a system response. (EX-1021, 20-21.) Its primary role is to mediate between user input, system capabilities, and the application logic, enabling the system to conduct multi-turn conversations while managing ambiguity and error correction. As Huang put it, “the dialogue manager controls the flow of conversation” and is “the central component that communicates with applications and the spoken language understanding modules such as discourse analysis, sentence interpretation, and response generation.” (EX-1018, 7.)

87. While the dialog manager controls many parameters of the system, two in particular are worth emphasis: (1) control of initiative, and (2) control of flow.

(1) Control of Dialog Initiative/Participant Roles

88. In an interactive system, the “dialogue may be system-led, user-led or mixed initiative.” (EX-1020, 107; *see also* EX-1018, 860.) “In a system-led dialogue the system asks a sequence of questions to elicit the required parameters of the task from the user.” (EX-1020, 107.) “In a user-led dialogue the user controls the dialogue and asks the system questions in order to obtain information.” (EX-1020, 107.) “In a mixed initiative dialogue control is shared” such that the “user can ask questions at any time, but the system can also take control to elicit required information or to clarify unclear information.” (EX-1020, 107.) Regardless of the initiative type, “the fundamental structure of dialog consists of initiative-response pairs” shown in Huang’s Figure 17.2 (below). (EX-1018, 860.) Initiatives (I) “are often issued by users” while Responses (R) “are issued by the system.” (EX-1018, 860.)



Huang, Figure 17.2

89. “These distinctions are reflected in the methods that can be used to control the dialogue flow.” (EX-1020, 107.) For example, in system-led implementations, control involves determining what and when questions should be asked by the system. (EX-1020, 107.) In system-led implementations, the dialog flow “can be scripted as a sequence of choices.” (EX-1020, 107.) However, in the mixed-initiative, more open-ended implementations, “the choice of the next action is determined dynamically, based on the current state of the dialogue.” (EX-1020, 107-108.) In such systems, the dialog manager “acts as an intelligent agent that makes rational decisions about how to manage the dialog.” (EX-1020, 107-108.)

(2) Control of Dialog Flow

90. In spoken dialogue systems the control of dialogue flow is a core responsibility of the dialogue manager. Though they go by different names, there are largely three ways control of flow can be achieved. McTear refers to finite-state-based, frame-based, and agent-based dialogue controls. (EX-1020, 111.) Möller refers to dialogue grammar, plan-based approaches, and collaborative approaches. (EX-1021, 28.) Though there are differences between these two, I will focus on describing the approaches using Möller’s framework.

91. Starting with dialogue grammars, this model represents the most structured and rigid form of dialogue flow control, which implement dialogue flow

using structures like graphs or finite-state machines. As Möller explains, dialogue grammars implement a top-down model of dialogue management using structures such as graphs, finite-state machines, or declarative grammar rules. Each node in the graph corresponds to a specific system prompt—ranging from closed or open questions to explanations or choices delivered via “audible quoting”—and transitions are determined by the semantic interpretation of the user’s input, constrained by a context-free grammar. (EX-1021, 28.) This model supports simple, robust, and user-guided interactions, making it well-suited for well-structured tasks. However, Möller emphasizes its limitations: “a lack of flexibility, and a very close relation or mixture of task and dialogue models,” which renders it unsuitable for “ill-structured tasks” or “complex transactions.” (EX-1021, 28.) He notes that the rigid, system-driven structure of dialogue grammars can be mitigated through frame-based approaches, which offer greater adaptability by organizing application needs hierarchically. (EX-1021, 28.)

92. Next are plan-based approaches. These focus on modeling the communicative goals behind user interactions, including any relevant sub-goals. These systems use plan operators to analyze dialogue structure and infer the user’s underlying intentions, enabling them to interpret even indirect speech acts. While offering greater flexibility and depth of understanding than simpler models like dialogue grammars, plan-based systems are also significantly more complex. A key

requirement is alignment between the plans of the human and system agents—if these diverge, the dialogue may “head in the completely wrong direction.” (EX-1021, 28.)

93. Finally, collaborative approaches to dialogue management move beyond task structure to focus on the motivation and mechanics of dialogue itself. Unlike plan-based models that center on predefined goals, collaborative systems aim to model the shared beliefs and mutual understanding between participants. The dialogue manager tracks what both the system and the user believe to be true within the conversation, treating “accepted goals [as] shared beliefs.” (EX-1021, 28.) These approaches typically blend elements from agent theory, plan-based reasoning, and dialogue grammars to model generic, flexible properties of dialogue. However, this openness comes at a cost: because the dialogue is less constrained, users are more likely to speak in unanticipated ways, increasing the complexity of interpretation. As a result, collaborative systems require more advanced natural language understanding and reasoning capabilities. (EX-1021, 29.)

94. McTear provides additional detail on this “collaborative approach” system, which he refers to as “agent-based” control implementations. As McTear describes, these systems often “draw on techniques from Artificial Intelligence (AI) and focus on the modelling of dialogue as collaboration between intelligent agents to solve some problem or task.” (EX-1020, 116.) In these systems, communication

“is viewed as interaction between two agents, each of which is capable of reasoning about its own actions and beliefs, and sometimes also about the actions and beliefs of the other agent.” (EX-1020, 117.) “The dialogue model takes the preceding context into account with the result that the dialogue evolves dynamically as a sequence of related steps that build on each other.” (EX-1020, 117.) These systems may also “use expectations to predict and interpret the user’s next utterances.” (EX-1020, 117.)

95. Generally, agent-based, or collaborative, systems “require complex dialogue and user models as well as mechanisms for using these models as a basis for decisions on how to control the dialogue.” (EX-1020, 124.) “Information about the dialogue history and the user can be used 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. These sorts of modelling involve representations of discourse structure, of intentions, goals and beliefs, and of dialogue as a collaborative activity.” (EX-1020, 124.)

96. In the next section I discuss different sources of knowledge sources and user models that can be used to facilitate dialog management, particularly in collaborative / agent-based approaches to dialogue management.

(3) Knowledge Source for Dialog Management

97. The dialog manager may use a number of knowledge sources “which are sometimes referred to collectively as the dialogue model.” (EX-1020, 123.)

98. Collaborative or agent-based dialogue systems, “require complex dialogue and user models” in order to function effectively. (EX-1020, 124.) As McTear explains, these systems also need “mechanisms for using these models as a basis for decisions on how to control the dialogue.” (EX-1020, 124.) In such architectures, information from the dialogue history and user model is not merely recorded but actively used to shape system behavior. Specifically, this data is leveraged “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.)

99. These capabilities depend on sophisticated internal representations, including “representations of discourse structure, of intentions, goals and beliefs,” which help the system reason about the user’s communicative intent. (EX-1020, 124.) Crucially, dialogue is treated not as a sequence of exchanges, but “as a collaborative activity,” where meaning and mutual understanding are co-constructed by both participants. (EX-1020, 124.)

100. Examples of the type of knowledge relevant to dialogue management, as identified by McTear, include:

- **A dialogue history.** This includes “[a] record of the dialogue so far in terms of the propositions that have been discussed and the entities that have been mentioned. This representation provides a basis for conceptual coherence and for the resolution of anaphora and ellipsis.” (EX-1020, 123.)
- **A task record.** This model includes “a representation of the information to be gathered in the dialogue. This record, often referred to as a form, template or status graph, is used to determine what information has not yet been acquired. This record can also be used as a task memory (Aretoulaki and Ludwig, 1999) for cases where a user wishes to change the values of some parameters, such as an earlier departure time, but does not need to repeat the whole dialogue to provide the other values that remain unchanged.” (EX-1020, 123.)
- **A world knowledge model.** This model “contains general background information that supports any commonsense reasoning required by the system, for example, that Christmas day is December 25.” (EX-1020, 124.)
- **A domain model.** This model contains “specific information about the domain in question, for example, flight information.” (EX-1020, 124.)
- **A generic model of conversational competence.** This includes “knowledge of the principles of conversational turn-taking and discourse obligations, for example, that an appropriate response to a request for information is to supply the information or provide a reason for not supplying it.” (EX-1020, 124.)
- **A user model.** This model “may contain relatively stable information about the user that may be relevant to the dialogue – such as the user’s

age, gender and preferences – as well as information that changes over the course of the dialogue, such as the user’s goals, beliefs and intentions.” (EX-1020, 124.)

d. Response Generator and Speech Synthesizer

101. The response generator and speech synthesizer in a spoken dialogue system is responsible for converting internal representations—typically derived from the dialogue manager—into linguistically and pragmatically appropriate output. “They are described together, because the strict separation into a component which generates a textual version of the output for the user (response generation) and another one which generates an acoustic signal from the text (speech synthesizer) is not always appropriate.” (EX-1021 33.)

102. As Möller explains, “[r]esponse generation involves decisions about what information should be given to the user, how this information should be structured, and about the form of the message (words, syntax).” (EX-1021, 33.) The system must decide not just what to say, but how to say it—balancing informativeness, clarity, and conversational appropriateness. This process may be implemented via formal grammars or simpler “template” approaches, in which sentences are generated by filling predefined structures with contextually relevant content.

103. Möller further notes that at each dialogue act, “the response generator builds a template sentence ... filling gaps from the content of the current semantic

frame, the dialogue history, and the result of the database query” (EX-1021, 33.) Higher-level generation rules govern how much information can be included in a single utterance and how it should be structured, particularly when there is too much content to communicate all at once. Importantly, response generation should be informed by the user model, adapting language and content based on the user’s background knowledge and experience. As Möller puts it, this module must “respect the user model, e.g. with respect to his/her expected domain knowledge and experience.” Ultimately, the response generator ensures that the system communicates in a way that is coherent, concise, and aligned with user expectations.

B. Overview of the ’681 Patent

104. ‘681 patent is directed to “a cooperative conversational model for a human to machine voice user interface.” (EX-1001, 1:7–8.) The disclosed cooperative conversational voice user interface understands free-form human utterances, “freeing users from being restricted to a fixed set of commands and/or requests.” (EX-1001, 1:63–67.)

105. The ’681 patent discloses a system that “receive[s] an input, which may include a human utterance” (i.e., “words, syllables, phonemes, or any other audible sound made by a human being”), “where the utterance may include one or more requests (i.e., ”command, directive, other instruction for a device, computer or other machine, to retrieve information, perform a task, or take some other action“). (EX-

1001, 2:6–14.) ”The utterance component of the input ... [is] processed by a speech recognition engine ... to generate one or more preliminary interpretations of the utterance.“ (EX-1001, 2:16–20.) ”The one or more preliminary interpretations“ are then ”provided to a conversational speech engine for further processing, where the conversational speech engine ... communicate[s] with one or more databases to generate an adaptive conversational response, which may [then] be returned to the user as an output. (EX-1001, 2:20–25.)

106. Figure 1 of the ’681 patent, reproduced below, depicts an embodiment of a cooperative conversational voice user interface. (EX-1001, 6:64–65; 7:7–9.)

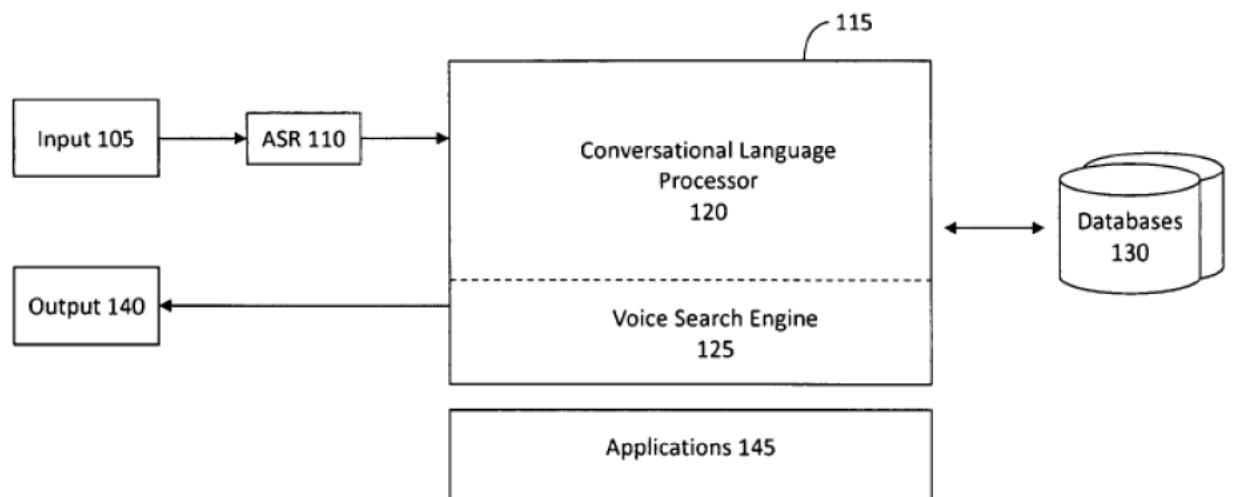


Figure 1

‘681 Patent, Figure 1

107. Figure 1 depicts a “system architecture for implementing a cooperative conversational voice user interface.” (EX-1001, 7:7–9.) The depicted system

“receive[s] an input 105 from a user”, where input 105 is “an utterance received by an input device (e.g., a microphone)”, and “where the utterance includes one or more requests.” (EX-1001, 7:10–13.) “The utterance component of input 105” is processed by “speech recognition engine 110” (i.e., “Automatic Speech Recognizer (“ASR”) 110”) to “generate one or more preliminary interpretations of the utterance.” (EX-1001, 7:24–28.) “The one or more preliminary interpretations generated by speech recognition engine 110” are then “provided to conversational speech engine 115 for further processing.” (EX-1001, 7:37–40.) “Conversational speech engine 115 ... communicate[s] with one or more databases 130 to generate an adaptive conversational response, which ... [is] returned to the user as output 140.” (EX-1001, 7:43–46.)

C. Level of Ordinary Skill in the Art

108. I understand that certain issues relating to the validity of the '681 patent must be judged from the perspective of a person of ordinary skill in the relevant art, as I discuss below. I have been asked to define the level of a “person of ordinary skill in the art” or “POSITA” at the time the alleged invention as claimed was made. In deciding the level of ordinary skill, I have considered the following factors, which I have been informed are relevant to the determination of ordinary skill in the art.

- levels of education and experience of persons working in the field;
- types of problems encountered in the art;
- prior art solutions to these problems;

- rapidity with which innovations are made; and
- sophistication of the technology.

109. In my opinion, a 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.

D. Prosecution History

110. During prosecution, the Examiner rejected the claims as obvious over U.S. 2005/0015256 ("Kargman") Kargman in view of U.S. Patent No. 5,748,841 ("Morin"), and Kargman in view of U.S. Patent No. 5,488,652 ("Bielby"). (EX-1002, 175-87.)

111. In response, Applicants amended the claims and argued that Kargman, alone or in combination, did not render the claims obvious because Kargman did not determine "an intended meaning for an utterance." (EX-1002, 232.) Applicants argued that Kargman only relates to "making statistical guesses or suggestions to avoid potential 'mismatches,'" but it does not "disclose, teach, or suggest disambiguating..." (EX-1002, 233.) For example, Applicants argued that "determining whether a particular spoken word was 'mushroom' or 'shrimp' relates, at best, to establishing what word was actually **said** rather than what a particular

word **means** in a particular context.” (EX-1002, 317 (emphasis in original).) Applicants further argued to distinguish over the applied art that the claims rely on short-term shared knowledge and long-term shared knowledge 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).)

112. The Examiner allowed the pending claims on December 6, 2011 without reasons for allowance. (EX-1002, 1012.)

E. IPR2020-01367

113. On July 28, 2020, Amazon.com, et al. filed a petition challenging claims 1-42 of the ’681 patent as obvious over U.S. Publication 2004/0193420 to Kennewick (“Kennewick”) alone or in combination with Huang, Seneff, et al., “Hypothesis Selection and Resolution in the Mercury Flight Reservation System” (“Seneff”), U.S. Patent 6,839,896 to Coffman et al. (“Coffman”), and U.S. Publication 2003/0182123 to Mitsuyoshi (“Mitsuyoshi”). (EX-1028, 15-64.)

114. The Board instituted trial. In its Patent Owner Response, Patent Owner disclaimed claims 37-42 with independent claims 37, 39, and 41 having substantially overlapping subject matter with claims 1, 13, and 25. (EX-1029, 1.) I illustrate the overlap between claim 1 and claim 37 below with differing language indicated with bold. Claims 13 and 39 and claims 25 and 41 have similar overlapping language. As

shown below, the preambles are identical as are limitations [1A.1] and [37A], [1B.1]-[1B.2] and [37B.1]-[37B.2], [1C.1]-[1C.2], [1D] and [37C.1]-[37D], [1D.1] and [37D.1], [1D.2] and [37D.2], and [1E.1]-[1E.2] and [37E].

Claim 1	Claim 37
[1P] A method for providing a cooperative conversational voice user interface, comprising	[37P] 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,	[37A] 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,	[37B.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;	[37B.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,	[37C.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;	[37C.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	[37D] determining an intended meaning for the utterance, wherein

Claim 1	Claim 37
determining the intended meaning for the utterance includes:	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	[37D.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
	[37D.2] inferring additional information about the utterance from the short-term shared knowledge and the long term shared knowledge in response to determining that the utterance contains insufficient information to complete a request in the identified context; and
[1D.2] establishing the intended meaning within the identified context,	[37D.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,	[37E] generating a response to the utterance based on the intended meaning established within the identified context.
[1E.2] wherein the conversational speech engine grammatically or syntactically adapts the response based on the intended meaning	

Claim 1	Claim 37
established within the identified context.	

115. The primary difference between the disclaimed claims and claims 1, 13 and 25 is that the disclaimed claim determine an utterance contains insufficient information to complete a request whereas claims 1, 13 and 25 disambiguate an intent from an utterance having one or more words that have different meanings in difference contexts.

116. In the Final Written Decision, the Board determined that Kennewick failed to disclose “one or more words ... have different meanings in different contexts.” (See EX-1030, 20-21.) This limitation is directed to the concept of sense ambiguity which was well-known in the art prior to October 2006. (See §III.A.3.d.) I establish in the grounds presented herein that this limitation was disclosed by both SmartKom and Barbara.

F. Claim Construction

117. I understand that “claim construction” is the process of determining a patent claim’s meaning. I understand that the Patent Office applies the same claim construction standard used in district courts which seeks to give claim terms their plain and ordinary meaning to a POSITA at the time of the claimed invention in view of the claims, specification, and prosecution history. I also understand that if the

patent applicant gave a term a special meaning in the specification or during prosecution, then the special meaning should be used.

1. Speech Recognition Engine

118. Claims 7, 10, 19, 22, 31, and 34 recite a “speech recognition engine.” I have been informed and understand that both parties in the Amazon litigation involving the ’681 patent proposed that the term “speech recognition engine” should be construed under a plain and ordinary meaning. (EX-1025, EX. A-4.) Further, I have been informed and understand that at the *Markman* hearing, PO and Amazon agreed that “speech recognition” means “software or hardware that recognizes the words or phrases in the natural language utterance.” (EX-1024, 2.) I have applied these constructions in this proceeding as well. I do not believe any other terms require construction in this proceeding.

2. Amazon Litigation Claim Construction

119. I have been informed and understand that other terms of the ’681 patent were at issue in the prior Amazon Litigation. I reproduce below the agreed-upon constructions, the disputed constructions, and the Court’s construction. As shown below, except for the “speech recognition engine” term, the Court found that no other term required construction. (See EX-1024, 2.)

a. Agreed-upon constructions

	Claim Term	Joint Proposed Construction	Court’s Construction
--	-------------------	------------------------------------	-----------------------------

1.	<p>“short-term shared knowledge”</p> <p>‘681 patent, claims 1-3, 10, 13-15, 22, 25-27, 34, 37, 39, 41</p>	<p>No construction necessary.</p> <p>The parties agree the claims of the ‘681 patents require “short-term shared knowledge” to be about the current conversation as opposed to just any conversation.</p>	<p>The Court agreed with parties that term does not require construction (EX-1024, 2.)</p>
----	---	---	--

b. Disputed Constructions

	Claim Term	VoiceBox’s Proposed Construction	Amazon’s Proposed Construction	Court’s Construction
1.	<p>“long-term shared knowledge” / “long-term knowledge”</p> <p>‘681 patent, claims 1, 4, 13, 16, 25, 28, 37, 39, 41</p> <p>‘049 patent, claims 4-6, 8, 14-16, 18</p>	<p>Plain and ordinary meaning: “shared knowledge built by identifying information with long-term significance that is user centric, rather than session-based”</p>	<p>No construction necessary; or, alternatively: “shared knowledge from one or more prior conversations”</p>	<p>Term does “not require construction beyond the clarification that both have the same meaning and are a type of shared knowledge” (EX-1024, 2.)</p>
2.	<p>“speech recognition engine” / “speech recognition”</p>	<p>Plain and ordinary meaning: “software or hardware that</p>	<p>Plain and ordinary meaning: “software or hardware that</p>	<p>Court’s order indicated that parties agreed to this construction at the <i>Markman</i></p>

	Claim Term	VoiceBox’s Proposed Construction	Amazon’s Proposed Construction	Court’s Construction
	‘681 patent, claims 7, 10, 19, 22, 31, 33, 34;	generates one or more preliminary interpretations of a given utterance” / “the process of generating one or more preliminary interpretations of a given utterance”	recognizes the words or phrases in the natural language utterance”	hearing: “speech recognition” means “software or hardware that recognizes the words or phrases in the natural language utterance” (EX-1024, 9.)
3.	“context” ‘681 patent, claims 1, 7, 13, 19, 22, 25, 31, 34, 37, 39, 41;	No construction necessary: “context”	“subject domain(s) that applies to a user utterance”	Term does not require construction (EX-1024, 2.)
4.	“identifying ... roles associated with the user and one or more participants” ‘681 patent, claims 5, 17, 29	No construction necessary: “identifying ... roles associated with the user or one or more other participants”	Indefinite	Term does not require construction (EX-1024, 2.)
5.	“classifying one or more of the utterance or the current conversation into a conversation type”	No construction necessary: “classifying one or more of the utterance or the current conversations into a	Indefinite	Term does not require construction (EX-1024, 2.)

	Claim Term	VoiceBox’s Proposed Construction	Amazon’s Proposed Construction	Court’s Construction
	‘681 patent, claims 5, 7-8, 17, 19-20, 29, 31-32	conversation type”		
6.	“identifying ... information allocation among the user and one or more other participants” ‘681 patent, claims 5, 17, 29	No construction necessary: “identifying . . . information allocation among the user and one or more other participants”	Indefinite	Term does not require construction (EX-1024, 2.)

IV. GROUND 1: Combination of SmartKom and Kobsa Renders Obvious Claims 1-9, 11-21, 23-33, And 35-36.

A. Overview of the Combination

1. SmartKom

120. I understand that “SmartKom: Foundations of Multimodal Dialogue Systems” by Wahlster (“SmartKom”; EX-1005) was published by August 2006. SmartKom “provides a comprehensive overview” of the SmartKom system and associated research results developed as part of the SmartKom research project. (See EX-1005, VI.) The SmartKom initiative, begun in spring 1999, included twelve research partners spanning research centers, industry (DaimlerChrysler, Philips, Siemens, Sony), and universities. (EX-1005, 31-32.) As of the publication date of

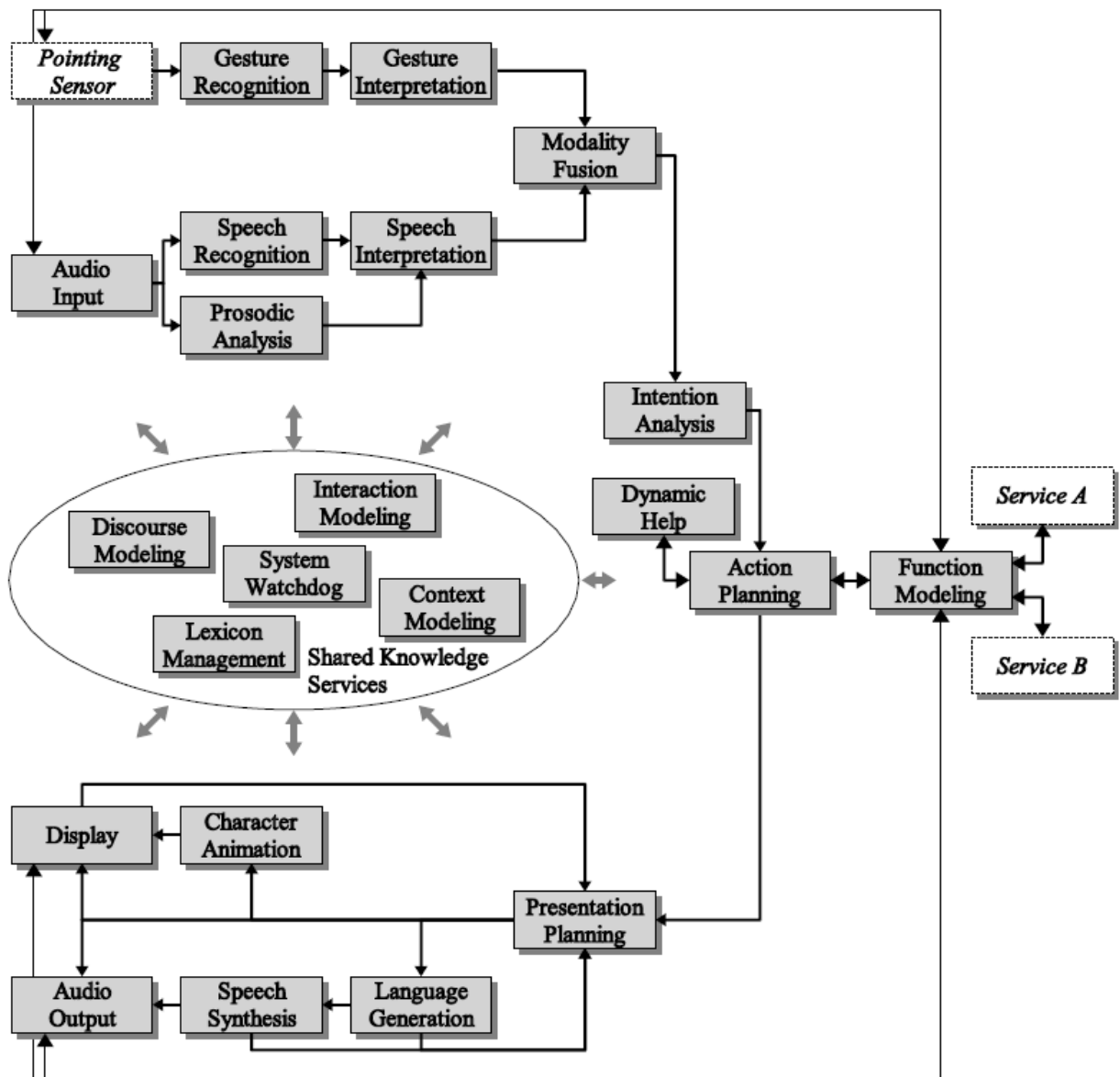
SmartKom, 52 patents concerning the SmartKom system were filed and 59 products and prototypes were released with 29 products released through the industry partners. (EX-1005, 24.) The final SmartKom demonstrator product was presented in June 2003, over three years before the '681 patent. (EX-1005, 35.)

121. The SmartKom system “provides full symmetric multimodality in a mixed-initiative dialogue system with an embodied conversational agent.” (EX-1005, 3.) SmartKom “represents a new generation of multimodal dialogue systems that deal not only with simple modality integration and synchronization but cover the full spectrum of dialogue phenomena that are associated with symmetric multimodality (including crossmodal references, one-anaphora, and backchannelling).” (EX-1005, 3.)

122. The generic software architecture of the SmartKom, illustrated in Figure 2(60)² below, includes a set of components with each “component represent[ing] one of the principal processing units of the executing system.” (EX-1005, 59.) This architectural framework “is designed to support a wide range of collaborative and multimodal dialogues that allow users to intuitively and efficiently access the functionalities needed for their task.” (EX-1005, 63.) SmartKom

² Because SmartKom is a collection of papers, it reuses the same Figure numbers in different chapters. For ease of discussion, I provide the page number in the Figure label.

describes three different application scenarios including public information and communication kiosk (SMARTKOM-Public), an infotainment assistant for the living room at home (SMARTKOM-Home), and a mobile travel companion (SMARTKOM-Mobile). (EX-1005, 63, 65-66.) The shaded components are “reused in all application cases.” (EX-1005, 59.)



SmartKom, Figure 2 (60)

2. Kobsa

123. “User Models in Dialog Systems” by Kobsa (“Kobsa”; EX-1006) published in 1989, 17 years prior to the filing date of the ’681 patent. Kobsa is a collection of conference papers from an international workshop on user modeling intended to provide a “coherent survey of the field of user modeling.” (EX-1006, V.) Kobsa notes that user models and user modeling was not known at the time of its 1989 publication date, stating that a “dozen major and several more minor user modeling systems have been designed and implemented in the last decade, mostly in the context of natural-language dialog systems.” (EX-1006, V.)

124. Kobsa generally describes a “user model” which “is a knowledge source in a natural-language dialog system which contains explicit assumptions on all aspects of the user that may be relevant to the dialog behavior of the system.” (EX-1006, 6.) Kobsa describes categories of knowledge for modeling including a user’s capabilities/experience and a user’s bias, preferences and attitudes. (See EX-1006, 415.) Kobsa describes several exemplary user models, including a model storing a user’s knowledge state, illustrated in Figure 3 below. (EX-1006, 16.)

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

125. Kobsa also describes a “user modeling component” that is “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 model.” (EX-1006, 6.) Kobsa further provides an overview of a general user modeling shell for building and maintaining long term models of individual users. (EX-1006, 411.)

3. Motivation to Combine

126. A POSITA would have been motivated to combine Kobsa’s teachings regarding user models and user modeling with SmartKom’s multimodal dialogue system. Kobsa is in the same field as the SmartKom and the ’681 patent—natural

language dialog systems/voice user interfaces. (EX-1001, 1:7-8; EX-1005, 3 (SmartKom “represents a new generation of multimodal dialogue systems”); EX-1006, V (Kobsa describes user models “in natural-language dialog systems”).)

127. SmartKom explicitly motivates the combination, mentioning use of a user model and/or user profiles/preferences in its context modeling and presentation planning:

- “The presentation planner recursively decomposes the presentation goal into primitive presentation tasks using 121 presentation strategies that vary with the discourse context, **the user model**, and ambient conditions.” (EX-1005, 16.)
- “Interaction modeling is concerned with different aspects like available modalities and user preferences for specific forms of communication as well as the affective state of the user.” (EX-1005, 61.)
- Identifying “User model” as the knowledge store holding “[p]roperties of the interlocutors” and associated with “[i]nterlocutionary context” (EX-1005, 274.)
- “*User preferences*: Given an appropriate **user model**, personal preferences, e.g., about the level of verbosity in speech output, can be used.” (EX-1005, 407 (emphasis in original).)
- “This way the system tries to react to user requests in an autonomous, independent, and context-aware way ... while taking into account the user preferences and certain conditions of a given situation or location.” (EX-1005, 509.)

128. SmartKom also mentions the concept of storing user preferences: “If the user verbally expresses a *like* or *dislike*, the semantic representation is passed to the nonstandard component” and the system “asks if it should store this positive or negative preference.” (EX-1005, 336 (emphasis in original).) However, SmartKom provides limited additional details regarding the content of the user models and 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. Kobsa is co-edited by Wolfgang Wahlster who is the editor of SmartKom and the Scientific Director of the SmartKom project, further leading a POSITA to Kobsa for further details of user models.

129. Kobsa also motivates the combination describing benefits of user models for “interact[ing] with people in an intelligent and cooperative manner.” (EX-1006, 411.) Kobsa explains user models support “(1) the task of recognizing and interpreting the information seeking behavior of a user, (2) providing the user with help and advice, (3) eliciting information from the user and (4) providing information to him.” (EX-1006, 416.) 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.) 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.

130. The combination is also nothing more than the 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). I provide the motivations above.

131. 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).) 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.

B. Independent Claims 1, 13, 25

132. The independent claims (claims 1, 13, and 25) of the '681 patent include substantially overlapping subject matter as highlighted by the chart below. Because this overlap, I address the independent claims together in this section.

Claim 1	Claim 13	Claim 25
[1P] A method for providing a cooperative	[13P] A non-transitory computer readable	[25P] A system for providing a cooperative

Claim 1	Claim 13	Claim 25
conversational voice user interface, comprising	medium containing computer-executable instructions for providing a cooperative conversational voice user interface, the computer-executable instructions operable when executed to:	conversational voice user interface, comprising:
		[25A.1] a voice input device configured to
[1A.1] receiving an utterance at a voice input device during a current conversation with a user,	[13A.1] receive an utterance at a voice input device, during a current conversation with a user,	[25A.2] receive an utterance during a current conversation with a user,
[1A.2]/[13A.2]/[25A.3] wherein the utterance includes one or more words that have different meanings in different contexts;		
		[25B] a conversational speech engine, wherein the conversational speech engine includes one or more processors configured to:
[1B.1] accumulating short-term shared knowledge about the current conversation,	[13B.1]/[25C.1] accumulate 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;	[13B.2] wherein the short-term shared knowledge includes knowledge about the utterance received at the	[25C.2] wherein the short-term shared knowledge includes knowledge about the utterance received during the current conversation;

Claim 1	Claim 13	Claim 25
	voice [sic] during the current conversation;	
[1C.1] accumulating long-term shared knowledge about the user,	[13C.1]/[25D.1] accumulate long-term shared knowledge about the user,	
[1C.2]/[13C.2]/[25D.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	[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;	[25E] identify a context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge;
[1D.2] establishing the intended meaning within the identified context,	[13E] establish an intended meaning for the utterance within the identified context,	[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
[1D.3]/[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		

Claim 1	Claim 13	Claim 25
[1E.1] generating a response to the utterance,	[13G.1] generate a response to the utterance,	[25G] generate a grammatically or syntactically adapted response to the utterance based on the intended meaning established within the identified context
[1E.2]/[13G.2]wherein the conversational speech engine grammatically or syntactically adapts the response based on the intended meaning established within the identified context.		

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

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

[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:

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

133. SmartKom discloses a “system” [25P]—the SmartKom multimodal dialogue system. The SmartKom system “provides full symmetric multimodality in a mixed-initiative dialogue system with an embodied conversational agent.” (EX-1005, 3.) “SMARTKOM represents a new generation of **multimodal dialogue systems** that deal not only with simple modality integration and synchronization but cover the full spectrum of dialogue phenomena that are associated with symmetric

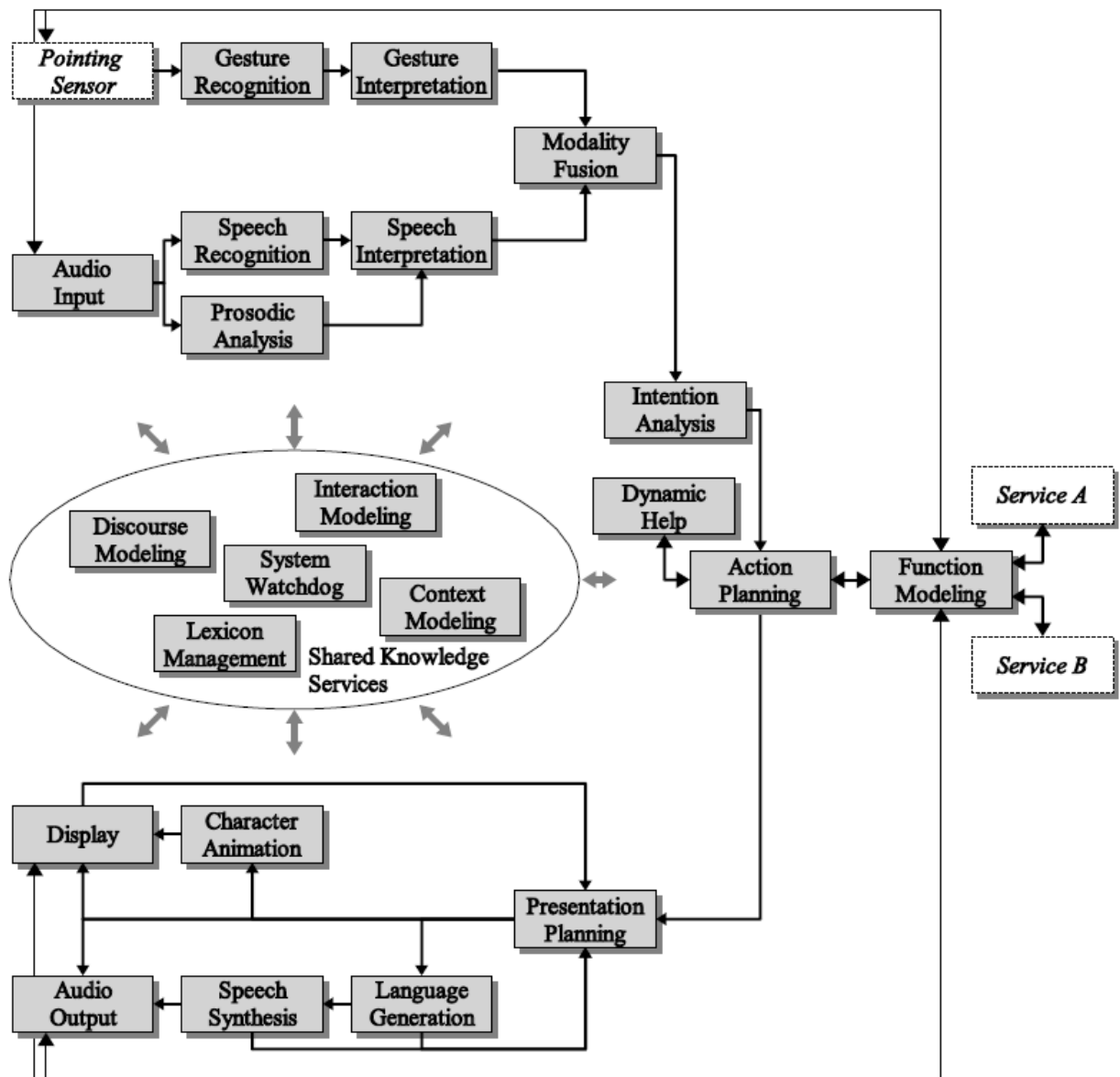
multimodality (including crossmodal references, one-anaphora, and backchannelling).” (EX-1005, 3.)

134. “One of the central ideas embedded within the” SmartKom framework “is to develop a kernel NLP [natural language processing] system that can be used in a variety of situations, i.e., scenarios, domains and modalities.” (EX-1005, 275.)

The generic software architecture of SmartKom is illustrated in Figure 2(60) below.

This software provides a “*method*” [1P] performing the functions specified in claim

1.



SmartKom, Figure 2(60)

135. SmartKom further discloses a “non-transitory computer readable medium containing computer-executable instructions³ ... operable when executed

³ The '681 patent does not use the term “instructions.” At most, it mentions a “conversational language processor” included in the conversational speech engine. (See EX-1001; 7:40-43.)

to” perform the steps specified in claim 13. SmartKom’s “**software** architecture” is illustrated in Figure 2(60) (above). (EX-1005, 59, 439 (describing that SmartKom-Mobile “incorporates more than 40 software components”).) Although SmartKom does not use the term “*instructions*”, a POSITA would understand that SmartKom’s software contains “*computer-executable instructions*” that are stored on a “*non-transitory computer readable medium*” such as a computer memory and are executed by a processor.

136. SmartKom describes implementations that execute on a server having a “main memory” and at least one processor. (See EX-1005, 13, 440.) The main memory or a similar computer readable medium would necessarily store the SmartKom software processing units executed to provide the SmartKom functionality. The 1994 textbook on Operating System Concepts by Silberschatz (“Silberschatz”; EX-1015) confirms this understanding. Silberschatz explains that a “computer system consists of a CPU and a number of device controllers that are connected through a common bus that provides access to shared memory.” (EX-1015, 29-30.) Silberschatz stresses that “[f]or a computer to do its job of executing program,” the programs must be stored in a memory. (EX-1015, 37.) Therefore, a POSITA would understand that the system includes a “*non-transitory computer readable medium*” that “*contain[s] computer-executable instructions*” which when executed perform the recited steps for providing a voice user interface.

137. The SmartKom system provides a “*voice user interface*.” Specifically, the system includes, e.g., speech and gesture recognitions components that “analyze[] facial expression, **speech** and gesture simultaneously.” (EX-1005, 167, *see also* EX-1005 493 (describing “intelligent multimodal interface” for SmartKom-Home).) In Smartkom, “[s]peech recognition is an integral part of the multimodal dialogue system.” (EX-1005, 85.) As shown in Figure 2(60), the audio input receives voice input from a user and provides the resulting acoustic input signal to the speech recognition component that “transforms the acoustic input signal into an orthographic transcription representing the utterance of the speaker.” (EX-1005, 85.)

138. The SmartKom system is “*cooperative*” and “*conversational*.” The SmartKom system “provides full symmetric multimodality in a mixed-initiative dialogue system with an embodied **conversational** agent.” (EX-1005, 3.) For example, SmartKom uses “an anthropomorphic and affective **user interface** through an embodiment **conversational agent** called Smartakus.” (EX-1005, 6.) The SmartKom architecture, shown above, supports “**cooperative** dialogues with mixed initiative and a synergistic use of multiple modalities.” (EX-1005, 8, 368 (“The implemented architecture of such a system consists of a set of software agents to allow cooperative, adaptive, and proactive behavior of the overall system”), 381 (referring to “cooperative communication”).)

139. Thus, SmartKom discloses a “*method for providing a cooperative conversational voice user interface*” [1P], 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*” perform the recited steps [13P], and a “*system for providing a cooperative conversational voice user interface*” [25P].

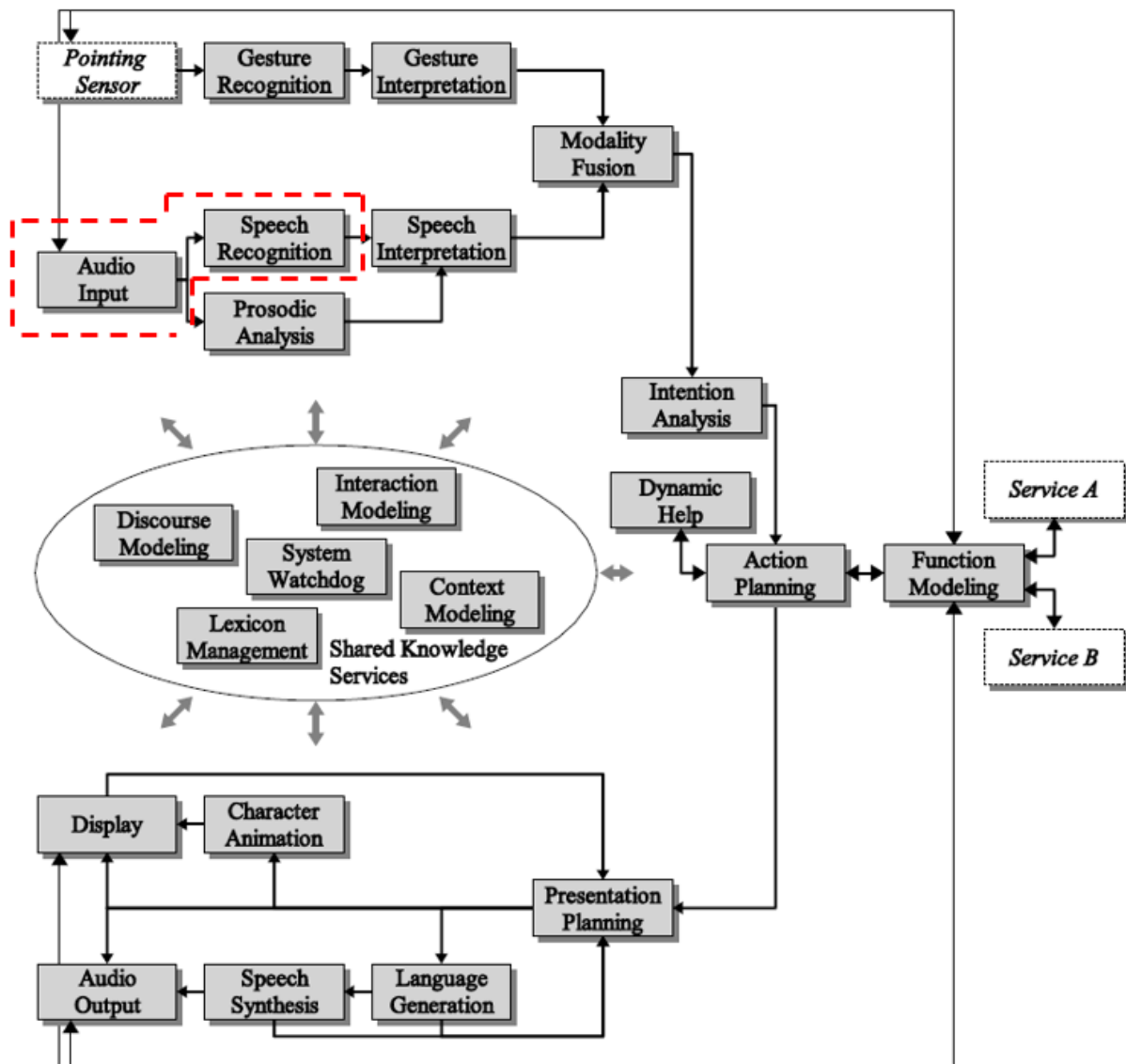
2. Receiving an Utterance [1A]/[13A]/[25A]

<p>[1A.1] receiving an utterance at a voice input device during a current conversation with a user,</p> <p>[13A.1] receive an utterance at a voice input device, during a current conversation with a user,</p> <p>[25A.1] a voice input device configured to [25A.2] receive an utterance during a current conversation with a user,</p> <p>[1A.2]/[13A.2]/[25A.3] wherein the utterance includes one or more words that have different meanings in different contexts; and</p>
--

140. The SmartKom software architecture, illustrated in Figure 2(60), includes an audio input, e.g., “a microphone array” and a speech recognition engine, also referred to as a “speech recognizer.” (See EX-1005, 85-108 (describing “Speech Recognition”), 212.) For example, in the SmartKom-Mobile Car application scenario, the system includes an “array of four microphones” for speech input. (EX-1005, 530.) SmartKom-Public and SmartKom-Mobile also include microphone inputs. (See EX-1005, 440 (showing a microphone used with SmartKom-Public),

457 (mentioning microphones used in SmartKom-Mobile.) As shown in Figure 2(60), the audio input receives voice (speech) input from a user and provides an acoustic input signal to the speech recognition component. (EX-1005, 85.)

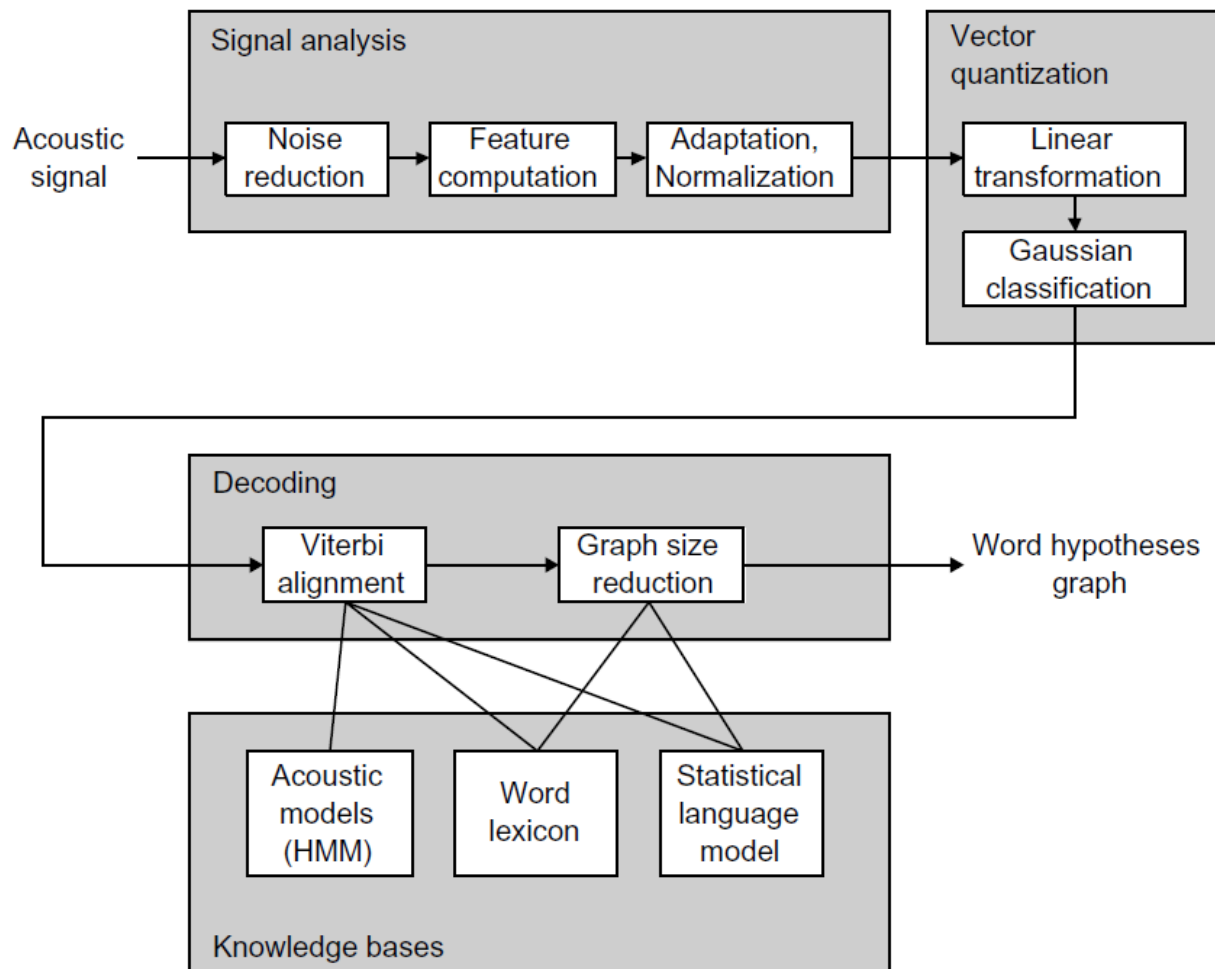
141. The audio input component is “*a voice input device*” that “*receive[s] an utterance*” from a user “*during a current conversation*” between the system and the user. Additionally or alternatively, because the speech recognizer receives the acoustic input voice signal, it is also “*a voice input device.*”



SmartKom, Figure 2(60)

142. The speech recognition engine “transforms the acoustic input signal into an orthographic transcription representing the utterance of the speaker.” (EX-1005, 85.) To perform this function, the speech recognizer uses “three main processing stages: signal analysis, vector quantization, and decoding” illustrated in Figure 1(86) (below). (EX-1005, 86.) Thus, SmartKom discloses “a voice input

device [audio input and/or speech recognition engine] configured to receive an utterance during a current conversation with a user” [25A.1]-[25A.2], “receiving an utterance at a voice input device [audio input and/or speech recognition engine] during a current conversation with a user” [1A.1], and “receive an utterance at a voice input device, during a current conversation with a user” [13A.1].



SmartKom, Figure 1(86)

143. The claim limitation further recites that the “*received utterance*” includes “*one or more words that have different meanings in different contexts*”

[1A.2]/[13A.2]/[25A.3]. This limitation recites a type of ambiguity common in natural language processing, referred to as “sense ambiguity”, where a “word can have different meanings for example, ‘bank’ can be a financial institution or the side of a river.” (EX-1020, 91.)

144. This type of ambiguity was well-known and well-understood prior to the '681 patent. For example, the book, “Design of Speech-based Devices” to Pitt, et al. (“Pitt”; EX-1013), describes that this ambiguity relates to words that have “more than one meaning” but “no difference in pronunciation to indicate the change of meaning.” (EX-1013, 62.) Pitt separates this ambiguity into two categories. In the first, “each different meaning is also a different part of language” which will therefore “be distinguished by syntactic as well as semantic changes in context.” (EX-1013, 62.) In the second, the “word can take more than one meaning while functioning as the same part of language (for example jet which, when used as a noun, can mean either a stream of liquid or an aircraft).” (EX-1013, 161.) Pitt provides an appendix of words falling into this second category including, e.g., the word “bill” which may either be a “demand for money, act of parliament, [or] a beak of web-footed bird.” (EX-1013, 161.)

145. The '681 patent provides two examples directed only to single word sense ambiguity. In a first example, the '681 patent presents the “user utterance of ‘What about traffic?’” and posits the word traffic has “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), [or] 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.) I note that in this example, the meaning of the single word appears to be resolved by the system of the '681 patent based on the current dialog history.

146. In a second example in the '681 specification, "a user may utter 'What's the weather in Seattle?' and context determination process 255 may establish Weather as a context, as well as establishing Seattle as an environmental context." (EX-1001, 11:64-67.) "The user may then utter 'and Portland?' and context determination process 255 may return a weather report for Portland, Oreg. based on the Weather and an environmental proximity between Portland, Oreg. and Seattle, Wash." (EX-1001, 11:67-12:4.) That is, although the term "Portland" generally refers to a city, in this example, it specifically means a city in Oregon in the first environmental context and a city in Maine in a different environmental context (e.g., when user utters "Portsmouth N.H.") in the first question. (See EX-1001, 12:8-11.)

147. SmartKom cites to a 1986 Paper by John McCarthy, "Notes on Formalizing Context" ("McCarthy"; EX-1012) which presents a similar example as

the “Portland” example of the ’681 patent. (EX-1005, 269.) McCarthy discusses “context” generally with a particular goal of “treat[ing] the context associated with a particular circumstance, e.g., the context of a conversation in which terms have particular meanings that they wouldn’t have in the language in general.” (EX-1012, 555.) An example that McCarthy provides is the word “Holmes” which although generally refers to an individual, in the context domain of “Sherlock Holmes stories” means an individual who is a detective whereas in the context domain of “U.S. legal history” means the individual who was a Supreme Court Justice. (EX-1012, 555-556.)

148. This type of “sense ambiguity” also occurs in phrases or entire utterances. The paper, “Contextual Coherence in Natural Language Processing” by Porzel and Gurevych (“Porzel”; EX-1014), which is also cited in SmartKom (EX-1005, 283), notes that “[w]ithout contextual knowledge[,] utterances, or fragments thereof, become susceptible of interpretation in more than one way.” (EX-1014, 275.) I note that Porzel and Gurevych are also the authors of the SmartKom Chapter “In Context: Integrating Domain- and Situation-Specific Knowledge.” (See EX-1005 269.) Porzel notes that the “more conversational a dialogue system becomes, the more difficult and unreliable become recognition and processing.” (EX-1014, Abstract.) Specifically, Porzel “introduce[s] a set of contextual coherence measurements that can improve the reliability of spoken dialogue systems, by

including contextual knowledge at various stages in the natural language pipeline.” (EX-1014, Abstract.) These are used in the SmartKom system as I discuss in more detail in §IV.B.5.

149. SmartKom provides a sophisticated processing system which handles a wide range of ambiguities and other issues common in natural language processing. SmartKom addresses the claimed sense ambiguity reference in the claim in part through the contextual coherence processing which uses its discourse and context models to provide a score for each interpretation of an ambiguous utterance. (EX-1005, 271.) Specifically, in cases “of pragmatic ambiguity, the model resulting from the analyses of the corresponding data ... comput[es] a contextual coherence score for the competing interpretations.” (EX-1005, 273.)

150. SmartKom provides an example of utterances using the German verb “kommen” which can mean “to come” or “showing on.” (EX-1005, 607; EX-1014, 281.) Porzel, which I mentioned above, provides two specific example utterances using “kommen”:

(6) *was für Spielfilme kommen heute Abend*
what for films come today evening

(7) *wie kann ich mir zur Schloss kommen*
how can I me to castle come

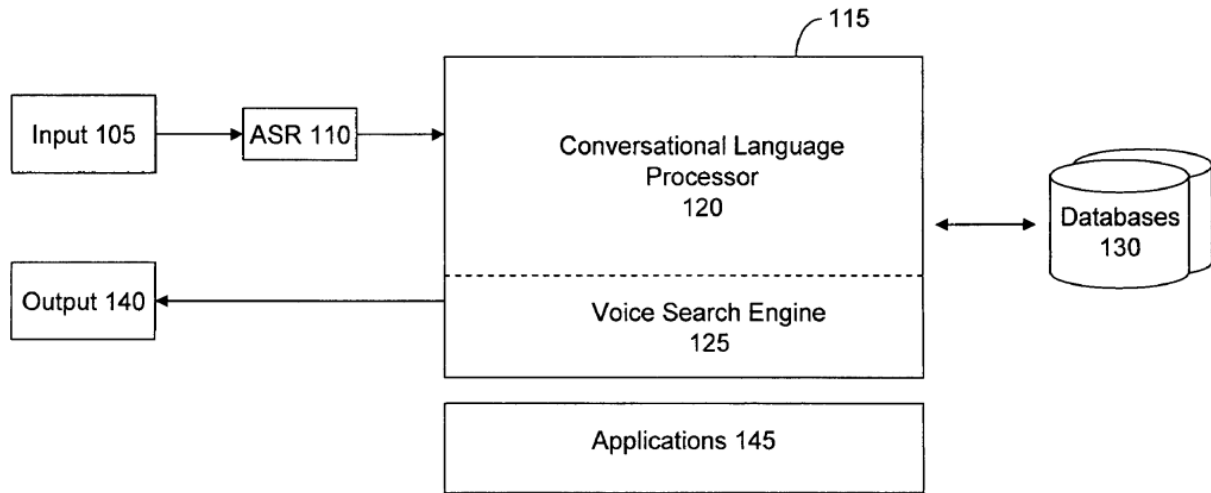
(EX-1014, 281.) SmartKom describes that “[d]ue to lexical ambiguity of individual words” such as *kommen*, “each SRH [speech recognition hypothesis] can be mapped to a set of possible interpretations called *conceptual representations* (CR).” (EX-1005, 607 (emphasis in original).) In each of the above examples, the “German word *kommen* can be mapped either to the ontological concept *MotionDirectedProcess*” (as in utterance 6) “or the concept *WatchPerceptualProcess*” (as in utterance 7). (EX-1005, 607; EX-1014, 281-282.)

151. Thus, SmartKom’s system accepts and handles user utterances that “include[] one or more words that have different meanings in different contexts.” (See also §IV.B.5.)

3. Conversational Speech Engine

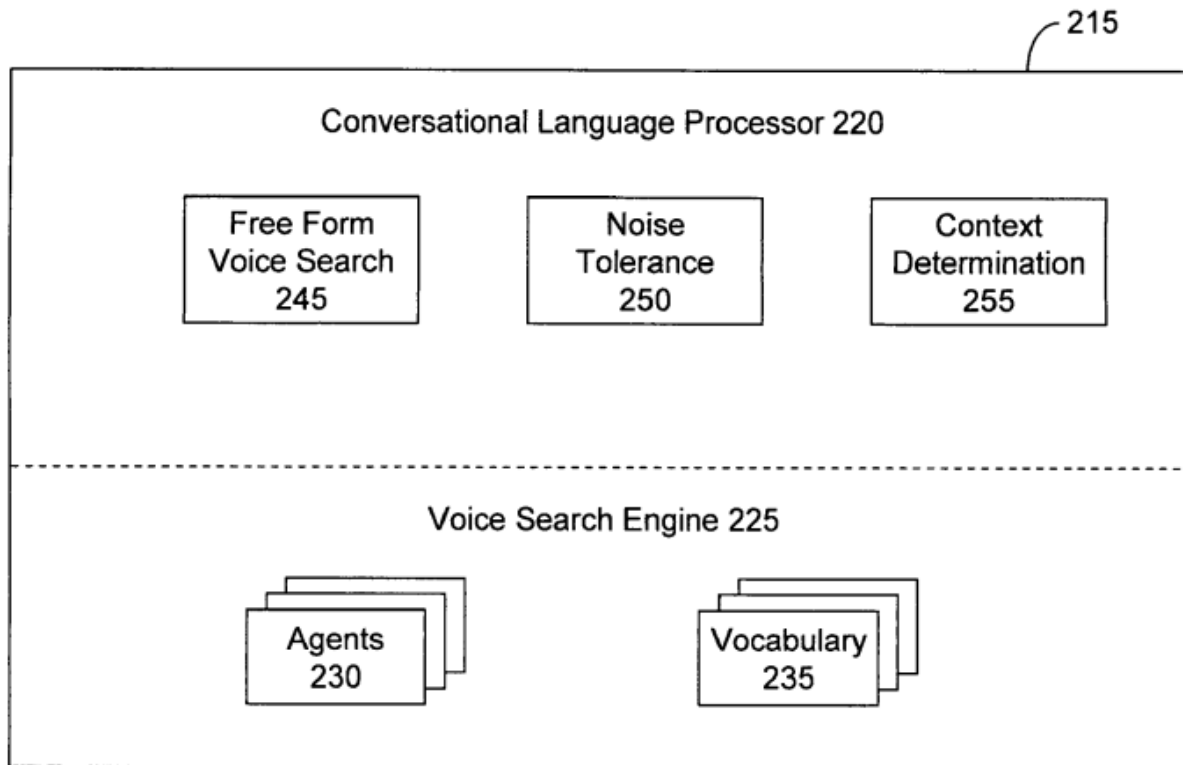
<p>[1D.1] identifying, at a conversational speech engine, ...</p> <p>[13D] identify a context associated with the utterance, wherein a conversational speech engine ...</p> <p>[25B] a conversational speech engine, wherein the conversational speech engine includes one or more processors configured to:</p>

152. Each of the independent claims recites “*a conversational speech engine*.” In the ’681 patent, the conversational speech engine receives the output of the speech recognition engine and provides “further processing” including language processing and response generation. (EX-1001, 7:37-40, Figure 1.)



‘681 Patent, Figure 1

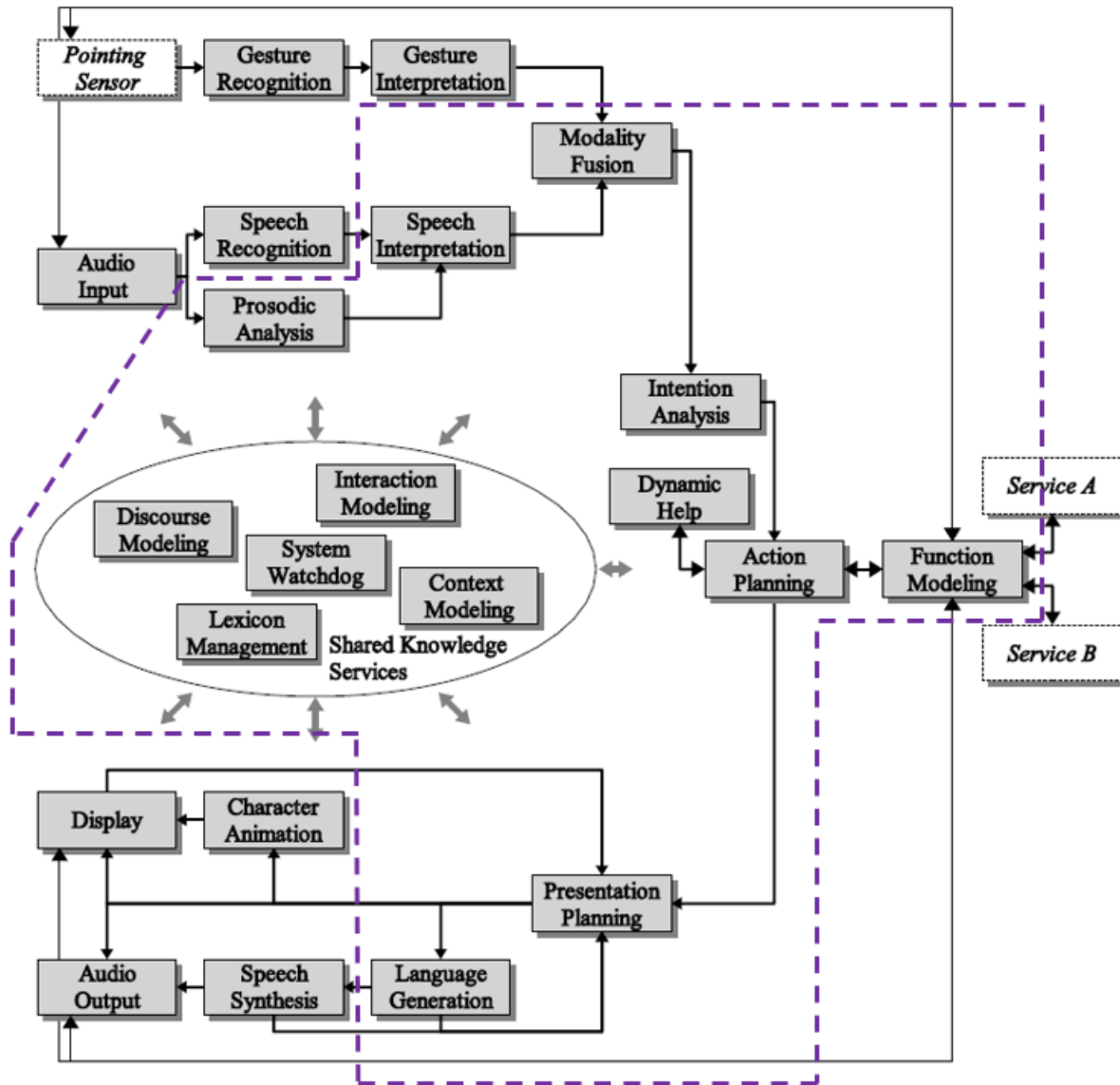
153. The conversational speech engine, illustrated in Figure 2 of the '681 patent below, includes free form voice search module 245, noise tolerance module 250, and context determination process module 255. (EX-1001, Figure 2.) These modules 245-255 communicate with voice search engine 255, also within the conversational speech engine, “that includes one or more context domain agents 230 and/or one or more vocabularies 235.” (EX-1001, 8:1-5.)



‘681 Patent, Figure 2

154. The SmartKom system has a similar architecture as the ‘681 patent’s system. The SmartKom system, illustrated in Figure 2(60) below, has a set of components that receive the output of the speech recognition component and provide further processing including the speech interpretation (natural language processing), modality fusion, and intention analysis components and components that participate in generating a response including the action planning, presentation planning, and language generation components. These components leverage functionality provided by the shared knowledge service components. (EX-1005, Figure 2(60).) Thus, SmartKom discloses components (e.g., those within purple box below) which

155. As discussed in §IV.A.1, each component in the software architecture “represents one of the principal processing units of the executing system.” (EX-1005, 59.) A POSITA would have understood that software executes on a processor.



(EX-1015, 29-30 (Silberschatz explains that a “computer system consists of a CPU and a number of device controllers that are connected through a common bus that provides access to shared memory”).) SmartKom describes a implementation where the software components execute on a server “consist[ing] of 3 dual Xeon 2.8 GHz processors.” (EX-1005, 13, *see also* EX-1005, Figure 1(440), showing a server box with “3 high-end PC systems” with an Intel Xeon dual-processor.) Thus, SmartKom discloses the “*conversational speech engine includes one or more processors*” [25B].

156. I discuss the process steps performed by the “*conversational speech engine*” in §§IV.B.5-6.

4. Short-Term and Long-Term Knowledge

157. The independent claims recite “*short-term shared knowledge*” and “*long-term shared knowledge.*” According to the ’681 patent “[s]hort-term knowledge may accumulate during a single conversation, where input received during a single conversation may be retained” and “may include cross-modality awareness, where in addition to accumulating input relating to user utterances, requests, locations, etc., the shared knowledge may accumulate a current user interface state relating to other modal inputs to further build shared knowledge models.” (EX-1001, 4:53-60.) In contrast, “[l]ong-term shared knowledge may generally be user-centric, rather than session-based, where inputs may be

accumulated over time to build user, environmental, cognitive, historical, or other long-term knowledge models” and “may include explicit and/or implicit user preferences, a history of recent contexts, requests, tasks, etc., user-specific jargon related to vocabularies and/or capabilities of a context, most often used word choices, or other information.” (EX-1001, 5:4-14.)

158. It was well-known, prior to the ’681 patent, that in a interactive dialogue system such as the SmartKom-Kobsa system, “information about the dialogue history and the user” is “used 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 sorts of modelling involve representations of discourse structure, of intentions, goals and beliefs, and of dialogue as a collaborative activity.” (EX-1020, 124.) SmartKom-Kobsa integrates “*short-term*” and “*long-term*” knowledge into the intention modeling components engaged by its intention recognizer to enrich the hypotheses associated with an utterance and to provide a contextual coherence score for potential interpretations of an utterance. (See, e.g., EX-1005, 271-75.)

159. SmartKom uses both “*short-term shared knowledge*” and “*long-term shared knowledge*.” SmartKom’s architecture contains “**shared** knowledge services” including discourse and context modeling components which are collectively part of SmartKom’s “context model” which operate in concert to enrich

and score hypotheses. (EX-1005, Figure 2(60).) These components therefore “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.)

160. SmartKom explains that “[u]tterances in dialogues, whether in human-human interaction or human-computer interaction, occur in a specific situation that is composed of different types of contexts.” (EX-1005, 274.) The specific knowledge stores and their associated context are illustrated in SmartKom’s Table 1 (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; *see also* EX-1014, 273-74.)

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)

161. For example, SmartKom describes a “domain and situational model” module containing domain-specific knowledge being implemented as `modeler.knowledge` and a “function model” module containing application-specific knowledge implemented in module called `modeler.function`. (EX-1005, 276.)

162. The dialogue model contains “[w]hat has been said by whom” and is associated with the dialogical context (discussion/conversation). (See EX-1005, Table 1(274).) The dialogue model in SmartKom includes the dialog history, which represents the state of the dialogue—what has been talked about and what is being talked about at the moment. (EX-1005, Table 1(274).) The dialogue history is used by the SmartKom system for disambiguation of context dependent utterances, and context sensitive interpretation, e.g., reference resolution and handling of ellipses. (EX-1005, 274.) The dialogue history is therefore 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. (See, e.g., EX-1001, 4:53-55 (short-term knowledge includes “input received during a single conversation”).) In SmartKom, the dialogue history is referred to as the “discourse state” and is stored in discourse memory. (EX-1005, 239, 242, 246.)

163. The situation model “monitor[s] ... corresponding situational factors relevant to resolving [] pragmatic ambiguities.” (EX-1005, 273.) Situational context 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. (See EX-1001, 4:55-60 (indicating short-term knowledge includes “location” data).)

164. Further, SmartKom explains that the system stores “[u]ser states,” which are “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.) “This extension of emotion refers to the interaction of users with the system, for instance, if the user shows hesitance or uncertainty because he does not know how the machine can help him.” (EX-1005, 139.) The user state is associated with emotion or state of the user when making the utterance and there is a type of “*short-term shared knowledge*” that “*includes knowledge about the utterance received.*”

165. The domain model “hold[s] knowledge of the world that is talked about.” (EX-1005, 274.) The SmartKom system uses information from the domain model “to guide the semantic interpretation of user’s utterances; to find the relevant items and relations that are discussed, to supply default values, etc.” (EX-1005, 274.) The domain knowledge is also “used to map information in the user’s utterance to concepts suitable for database search.” (EX-1005, 274.) In SmartKom, the domain model is “modeled as an ontology proper.” (EX-1005, 274.) The domain model is therefore a type of “*long-term shared knowledge.*”

166. SmartKom also discloses a user model. Table 1 shows that the user model stores “properties of the interlocutory” (i.e., information about the system users) and provides “interlocutionary context” associated with the content stored in

the “User model”. (EX-1005, 274.) SmartKom explains the “**role of the interlocutionary context**” is “of importance” in handling “certain ambiguities.” (EX-1005, 275.) SmartKom discloses, at a minimum, the “interaction preferences of the users are monitored actively by the system.” (EX-1005, 276.) While SmartKom discusses “general user model information⁴ is supplied via external sources, e.g., via a user’s SmartCard” (EX-1005, 276), Kobsa provides details of user models and of accumulating additional information regarding a user’s preferences, experience, and environment to enhance the user model. As discussed in Kobsa and in §IV.B.4.b, it was well-known to derive user profile information from user conversations. One example is provided in Barbara where the system derives that a user has a dog named “Toto” based on frequent references to “Toto” in the utterances related to the pet domain-context. (See EX-1007, ¶¶90-98.) Because the long-term user model includes knowledge derived from a user’s prior conversation the long-term user model includes “*knowledge about one or more past conversations with the user.*”

167. Kobsa explains that “[a] cooperative system must certainly take into account the user’s goals and plans, [and] **his/her prior knowledge about a**

⁴ This “general user model information” is understood to include, e.g., foundational profile information such as the user’s name, address, occupation, etc. For mobile applications, the information would also include the user’s telephone number.

domain.” (EX-1006, 5.) Further, “it is particularly in the above conversational settings that the construction and use of an *explicit model of the user’s beliefs, goals, and plans* becomes a central problem.” (EX-1006, 5 (emphasis in original).) Kobsa therefore describes “the general architecture of a domain independent system for building and maintaining long term models of individual users.” (EX-1006, 411.)

168. As explained by Kobsa, a “*user model* is a knowledge source in a natural-language dialog system which contains explicit assumptions on all aspects of the user that may be relevant to the dialog behavior of the system.” (EX-1006, 6 (emphasis in original).) And “[a] *user modeling component* is that 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).) Kobsa’s user models “record[] all information about the user inferred from his/her dialog behavior in a corresponding file.” (EX-1006, 10-11.) Kobsa explains that “[a]s the application system interacts with a user it can acquire knowledge about him and pass that knowledge on to the user model maintenance system for incorporation.” (EX-1006, 412.) This forms a “long-term knowledge base about the user.” (EX-1006, 413.) Thus, Kobsa, discloses a user modeling component that constructs a user model based on previous dialogues and retains that knowledge after the conversation ends. (EX-1006, 6-10, 411-413.)

169. Kobsa explains that a user model can be short-term or long-term. (EX-1006, 39, 41.) “Long-term models describe relatively static characteristics of users, while short-term models describe specific topics and goals in the current discourse.” (EX-1006, 39.) In this way, “a short-term model probably overlaps with what is usually meant by the term discourse model.” (EX-1006, 39.) A POSITA would therefore understand Kobsa’s short-term user model corresponds to SmartKom’s dialogue history. Because the long-term user model includes knowledge derived from a user’s prior conversations as I discuss in further detail in §IV.B.4.b, the long-term user model includes “*knowledge about one or more past conversations with the user.*” (EX-1001, 5:10-14 (long-term shared knowledge includes “explicit and/or implicit user preferences”).)

170. Kobsa provides an example of a user model in the TAILOR dialog system which “contains individual assumptions about the user’s experience with respect to a domain of discourse.” (EX-1006, 197.) The assumptions “are represented by a list of those items in the system’s knowledge base which are known to the user, and by information about whether the user understands the basic concepts underlying the domain of discourse.” (EX-1006, 197.) To represent concepts a user is “probably familiar with”, an overlay technique is employed, such as shown in Figure 3 below. (EX-1006, 15.) This information about the user is retained and used in future conversations with the user.

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

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

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

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

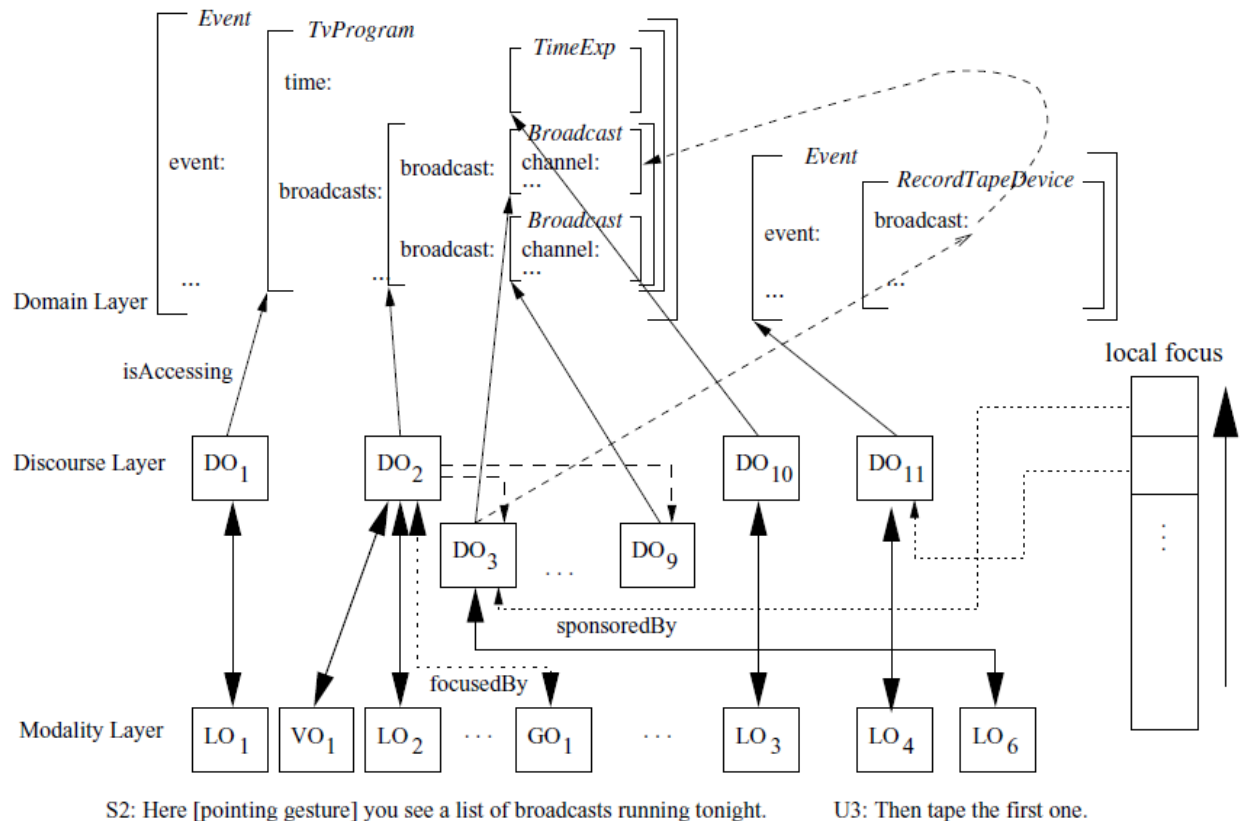
[25C] accumulate short-term shared knowledge about the current conversation,

[1B.2]/[25C.2] wherein the short-term shared knowledge includes knowledge about the utterance received during the current conversation

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

171. SmartKom discloses “*accumulat[ing]/[e] short-term shared knowledge about the current conversation*” [1B.1]/[13B.2]/[25C] in at least two ways. First, during a conversation, SmartKom accumulates dialog history in the discourse state stored in the discourse memory that is used by the discourse modeling module (DiM) to enrich and score intention hypotheses. (EX-1005, 237.) Generally, the discourse model is a knowledge source that contains the system’s description of the syntax, semantics and pragmatics of a dialog as it proceeds. (EX-1005, 237-38.) As explained by SmartKom, “[d]uring the course of a dialogue, each new user contribution has to be interpreted in the light of the previous discourse context.” (EX-1005, 238.) SmartKom uses “context representation” “for representing the discourse state.” (EX-1005, 239.) Context representation “consists of three levels: the discourse object layer, the modality layer and the domain object layer.” (EX-1005, 239, 242, Figure 2(243).) 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), e.g., linguistic objects, visual objects, and gesture objects, “linked to [their] corresponding DO[s].” (EX-1005, 244-45.) “The domain object layer encapsulates the instances of the domain model and provides access to the semantic information of objects, processes and actions.” (EX-1005, 245.) The local focus structure “provides and restricts access to all discourse object that are antecedent candidates

for later reference.” (EX-1005, 245.) On this level, “the content of the structure, i.e., discourse objects, are ordered by salience.” (EX-1005, 245.) That is, the local focus provides access to Dos during the interpretation of later utterances.



SmartKom, Figure 2(243)

172. As represented in the discourse state, the modality type and discourse object as well as their relationship to the domain layer entries included in the discourse model are the dialogue history and are each “*knowledge about the utterance received during the current conversation.*”

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

in the discourse memory is continuously accumulated over the course of a conversation. For example, when the intention recognition component selects an “intention hypothesis” for the current utterance (i.e., the meaning of the utterance), this “intention hypothesis is incorporated into the discourse history representing a user turn.” (EX-1005, 240-41.)

174. Second, SmartKom accumulates knowledge of a user state during an ongoing conversation. SmartKom explains that “[u]ser states are 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).) To gather user state information, SmartKom’s prosody⁵ module “analyze[s] the speech as a modality of the user input in order to detect the prosodic events as well as the most likely emotional state of the user.” (EX-1005, 141.) SmartKom “use[s] the features describing such prosodic cues as energy, pitch, and duration” and “some linguistic information” “to recognize the prosodic characteristics of a speech signal.” (EX-1005, 145.)

175. The prosody module, illustrated in Figure 1(142) below, “has two main inputs: the speech signals from the *audio module* and the word lattices (*word hypothesis graphs*, WHGs) from the *speech recognizer*.” (EX-1005, 141 (emphasis

⁵ Prosody “refers to the segments of speech larger than phonemes, e.g., syllables, words, phrases, and whole utterances.” (SmartKom, 139.)

in original).) As shown, “[a]fter running through feature extraction and classification steps, the detected prosodic events are added to the original input WHG and the **user state lattice** is generated.” (EX-1005, 141.) The “[u]ser state classification is done in two steps.” (EX-1005, 150.) First, word-based classification is used “to compute a probability to assign one of several user states to each word.” (EX-1005, 150.) Second, the probability of the whole utterance is processed “to decide one of the several classes.” (EX-1005, 150.) The user state is derived by the prosody module from the utterance and therefore is “*knowledge about the utterance received during the current conversation.*” (§IV.B.4.)

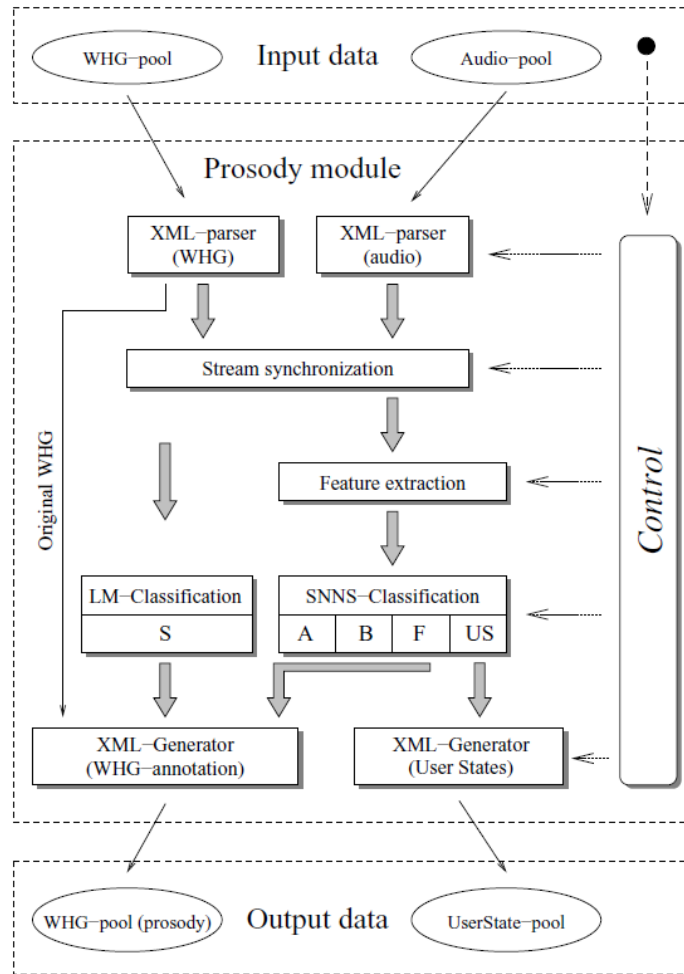


Fig. 1. Architecture of the prosody module

SmartKom, Figure 1(142)

176. Thus, for the above reasons and the reasons discussed in §IV.B.4, the accumulated “*short-term shared knowledge includes knowledge about the utterance received [at the voice] during the current conversation*” [1B.2]/[13B.2]/[25C.2]. And, the SmartKom-Kobsa combination discloses “*accumulating short-term shared knowledge about the current conversation*” [1B.1] and “*accumulate short-term shared knowledge about the current conversation*” [13B.1]/[25C].

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

<p>[1C.1] accumulating long-term shared knowledge about the user, [13C.1]/[25D.1] accumulate long-term shared knowledge about the user, [1C.2]/[13C.2]/[25D.2] wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;</p>
--

177. The combination of SmartKom and Kobsa discloses “*accumulat[e]/[ing] long-term knowledge about the user*” [1C.1]/[13C.1]/[25D.1]. As I discussed in §IV.B.4, SmartKom’s user model provides “[p]roperties of the interlocutors” to support “[i]nterlocutionary context.” (EX-1005, 274.) The user model therefore is a knowledge source “contain[ing] explicit assumptions on all aspects of the user that may be relevant for the dialog behavior of the system.” (See EX-1006, 6.)

178. The knowledge stored in a user model can be acquired (accumulated) “either *explicitly* or *implicitly*.” (EX-1006, 416 (emphasis in original).) For example, explicit knowledge can come from the application interviewing the user for certain information. (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.) The user model acquires user information from conversations between the user and system and therefore the stored

user model “*includes knowledge about one or more past conversations with the user.*”

179. Both SmartKom and Kobsa describes techniques for “*accumulating long-term shared knowledge.*” SmartKom describes that the “interaction model” of the shared knowledge services “computes information on the user from the hypothesis reflecting the user input.” (EX-1005, 287.) Kobsa teaches that the “*user modeling component* is that 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).) Kobsa describes a “general architecture of a domain independent system for building and maintaining *long term models of individual users.*” (EX-1006, 411 (emphasis in original).)

180. Specifically, 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.) With GUMS, the application “select[s] the initial stereotypes for the user and add[s] new facts about the user as it learns them.” (EX-1006, 418.) Each application knowledge base in GUMS has two parts, illustrated in Figure 3 (below): “(1) a collection of *stereotypes* organized into a taxonomy, and (2) a collection of models for the individuals.” (EX-1006, 418.) In user modeling, “a stereotype is a collection of facts and rules that are applicable for any person who is

seen as belonging to that stereotype.” (EX-1006, 418.) Some systems also expect the user “to mention enough self-characteristics at the start of a session so that the system can build up the user model.” (EX-1006, 83.)

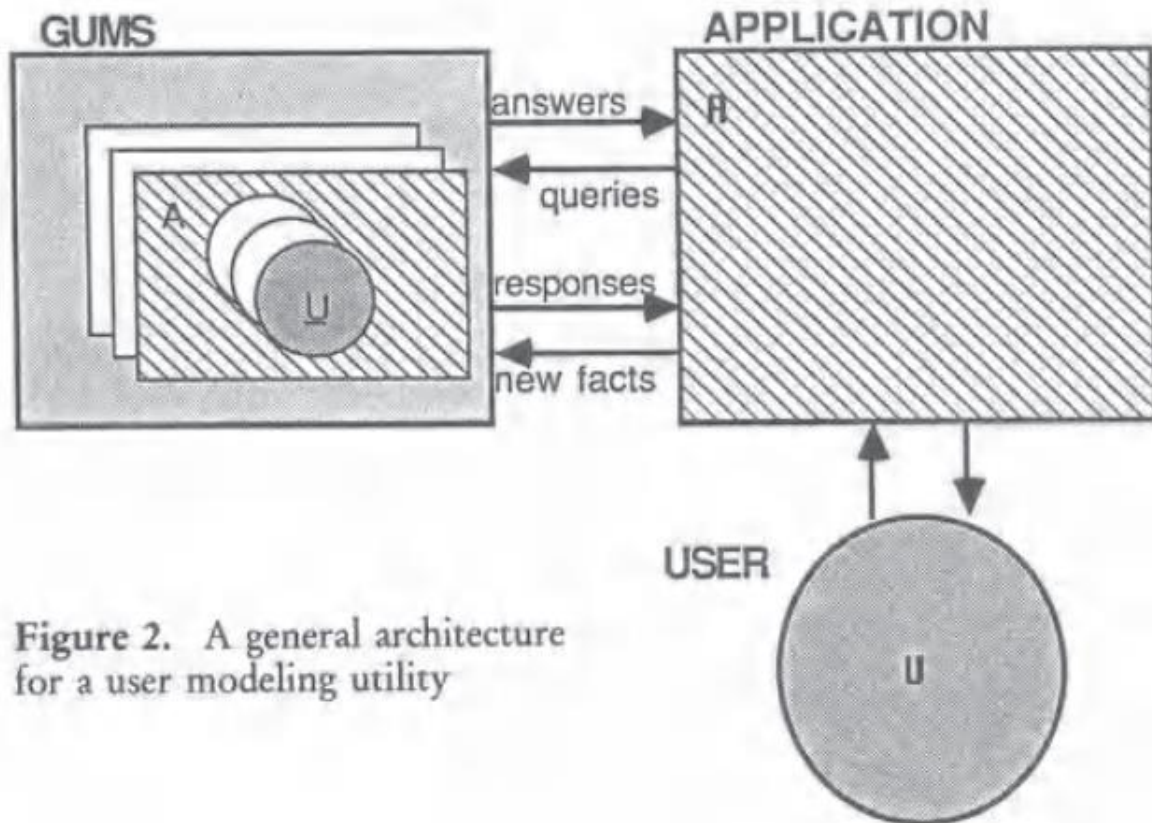


Figure 2. A general architecture for a user modeling utility

Kobsa, Figure 2

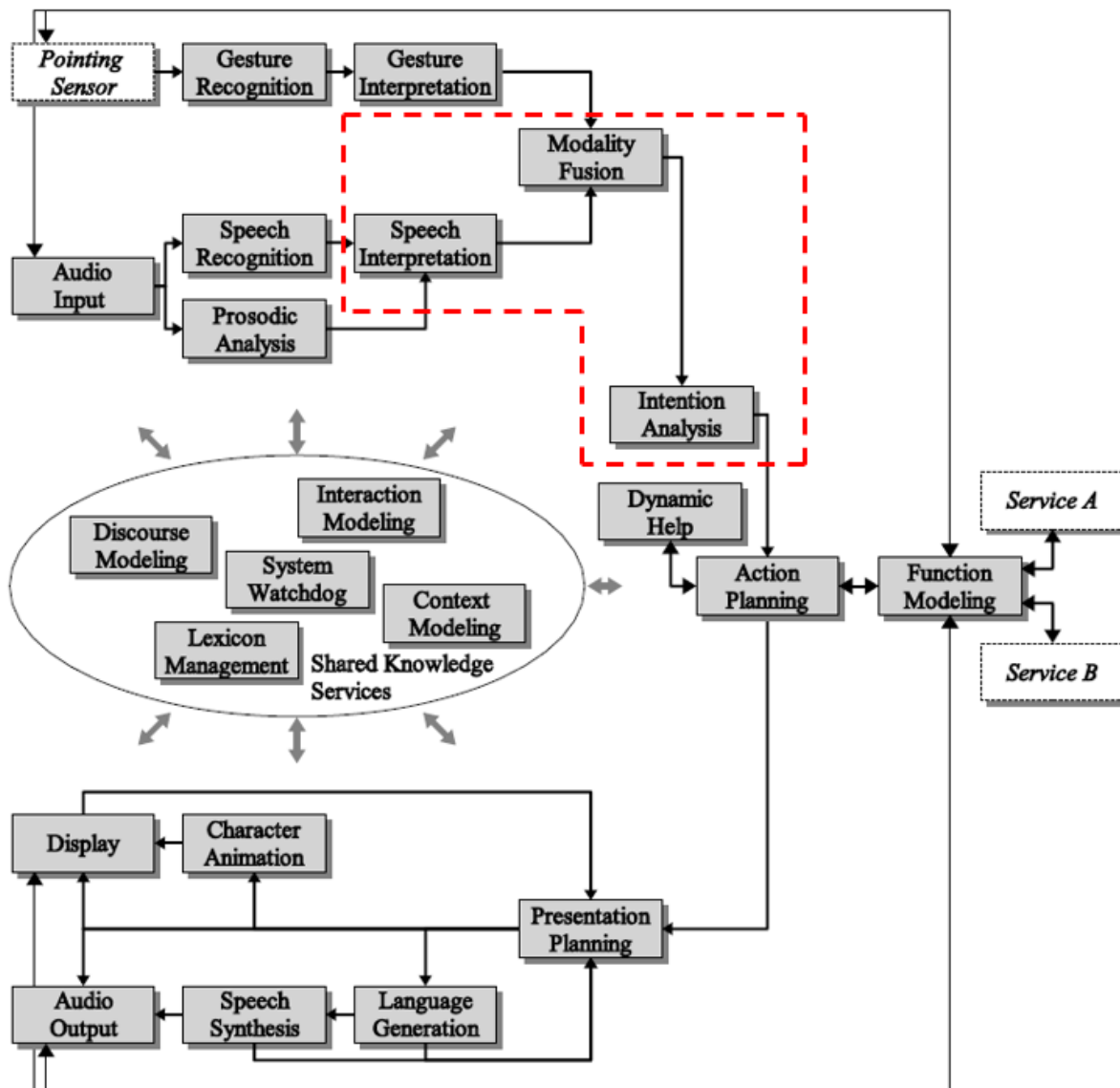
181. Kobsa confirms the user model is persistent (long-term), explaining that a “system can, at the beginning of a dialog, access an already existing model of a particular user which it had constructed during previous interactions” with the user. (EX-1006, 10.) For example, “at the end of a dialog session”, the system “records all information about the user inferred from his/her dialog behavior in a corresponding file.” (EX-1006, 10-11.)

182. Thus, for the above reasons and the reasons discussed in §IV.B.4, the accumulated “*long-term shared knowledge includes knowledge about one or more past conversations with the user*” [1C.2]/[13C.2]/[25D.2]. And the SmartKom-Kobsa combination disclosed “*accumulat[e][ing] long-term shared knowledge about the user*” [1C.1]/[13C.1]/[25D.1].

5. Determining an Intended Meaning [1D]

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

183. SmartKom discloses “*determining an intended meaning for the utterance*” [1D]. 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)

184. SmartKom explains that “[u]ser utterances are represented in a data structure called an *intention lattice*” which contains a number of hypothesis sequences which “stand for alternative readings of one and the same user input.” (EX-1005, 287-288, 305.) The above SmartKom modules transfer information to one another through this “intention lattice” structure. (EX-1005, 287-288.) In

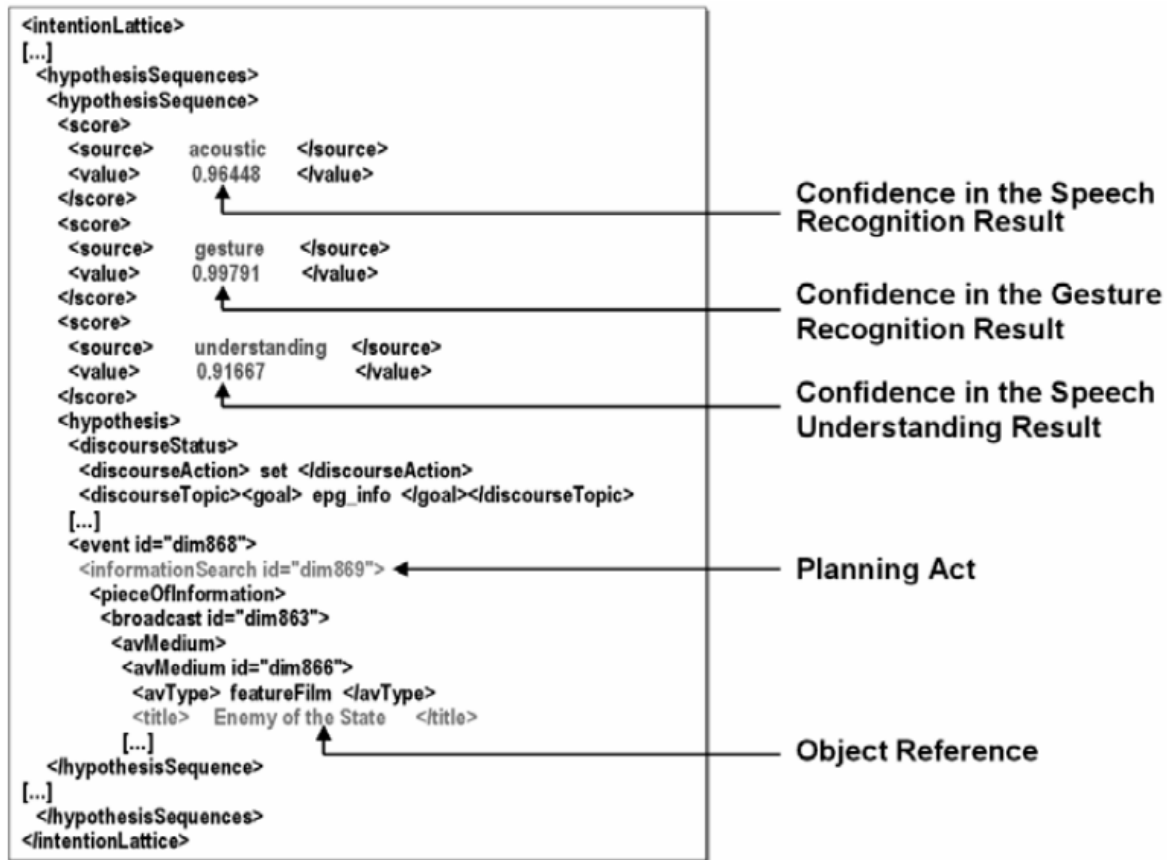
SmartKom, the “basic information flow from user input to system output continuously adds further processing results so that [this] representational structure [i.e., intention lattice] will be refined step-by-step.” (EX-1005, 62-63.) Therefore, the intention recognizer receives an intention lattice from the modality fusion components including a set of hypotheses.

185. SmartKom’s intention lattice “contain[s] one or more *intention* segments that partition the utterance sequentially (in case of, for example, multisentence utterances).” (EX-1005, 305.) Each user intention “can specify a set of dialogue acts, which in turn can be either goal or slot manipulations.” (EX-1005, 305.) “The constituents of each intention segment that are important for the dialogue manager are therefore” the following:

- A discourse object derived from the ontological framework, containing information about processes and roles the dialogue contribution is about.
- Goal manipulations:
 - setting a goal to adopt a new task, usually putting one or more processes on the agenda
 - retracting a goal to abort a task currently pursued
 - retaining a goal (this confirms that it is still actively being talked about, and an utterance occurs in the context of the goal)
- Slot manipulations:
 - setting slots to establish new information
 - retracting slots to take back / invalidate current information,
 - retaining slots to confirm informationSlot manipulations always occur in the context of a specific goal.
- Annotation of the intention as positive and negative feedback, which might either be an answer to a simple question, or an expression of the emotional state of the user.
- Confidence scores from the analysis modules.

(EX-1005, 306.) An exemplary intention lattice fragment is illustrated below in SmartKom's Figure 8(15).

186. In SmartKom, the “basic information flow from user input to system output continuously adds further processing results so that the representational structure will be refined step-by-step.” (EX-1005, 62-63.) Because this lattice includes a “planning act,” a POSITA would understand that it was generated by the action planning component. As shown, the other information was added as the lattice passed through the system. “The task of the natural language processing [speech interpretation] module is to transform the word lattice sent by the speech recognizer into a list of hypotheses representing in an abstract way possible user intentions.” (EX-1005, 195.) The list of hypotheses each with a corresponding score value (referred to as the confidence in the speech recognition result above) is provided to the modality fusion component. (EX-1005, 204.)



SmartKom, Figure 8(15)

187. This modality fusion component provides “integration and disambiguation of the different modalities.” (EX-1005, 224.) SmartKom’s modality fusion approach “generate[s] all meaningful combinations considering all hypotheses” from the speech recognition and gesture recognition components “and then select[s] the n best results, which are passed to the intention analyzer.” (EX-1005, 232.) In SmartKom, “a maximum of three hypotheses is sent to the intention analyzer”, each with “fusion score.” (EX-1005, 233.) If no gestures are available to

the modality fusion component, “the speech hypotheses are passed to the intention analyzer without any changes.” (EX-1005, 229.)

188. “Before the intention recognition takes a decision on the set of hypotheses for the user input,” the intention recognizer engages a discourse model that works in conjunction with a context modeler to enrich and validate/score the remaining hypotheses, illustrated in Figure 1 (286) below. (EX-1005, 286-287.) For example, the discourse model “contribute[s] to the selection of the most probable one [of the multiple hypotheses] on the basis of the previous disclosure context.” (EX-1005, 238.) The models consulted provide a contextual coherence score for each hypothesis in the intention lattice to the intention recognition component to select one hypothesis. (EX-1005, 285.) The selected hypothesis is considered the user’s “*intended meaning for the utterance.*”

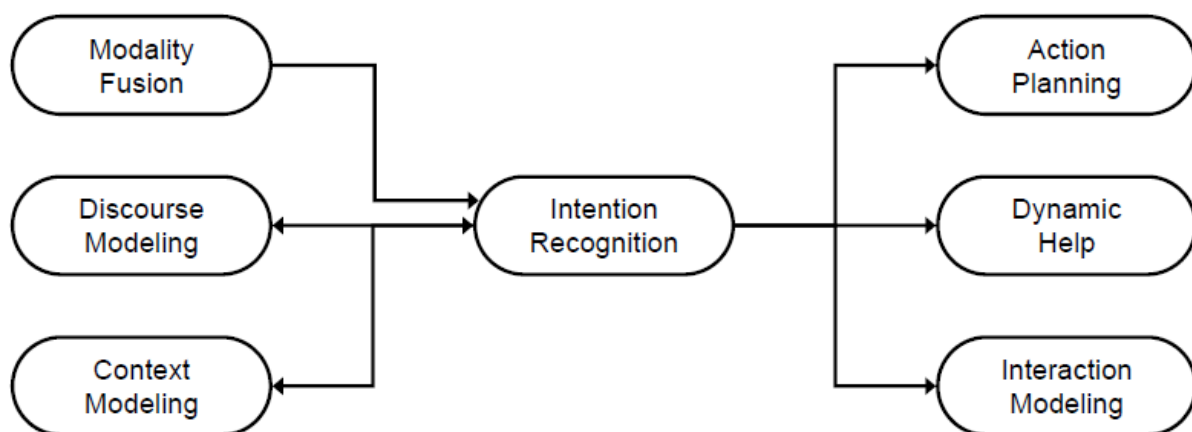


Fig. 1. Section of the multimodal interaction from the intention recognition point of view

SmartKom, Figure 1(286)

189. Thus, the combination of SmartKom and Kobsa discloses “*determining an intended meaning for the utterance*” [1D]. I discuss the “*identifying*” [1D.1] and “*establishing*” [1D.2] limitations of claim 1 below.

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

[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

[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;

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

190. The combination of SmartKom and Kobsa 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]. SmartKom notes that “[s]peakers may not always be aware of the potential ambiguities inherent in their utterances.” (EX-1005, 275.) Instead, “[t]hey leave it to the context to disambiguate and specify the message.” (EX-1005, 275.) “In order to interpret the utterance correctly, the addressee must employ several context-dependent resources.” (EX-1005, 275.)

191. 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.) “These

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, and to support the decisions to be taken by the intention recognition.” (EX-1005, 287.)

192. 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.)

193. The following example illustrates this context identification. In this example, “the user has requested information about tonight’s television program,” In utterance-1 (EX-1005, 262-263.) The TFS representation of the utterance-1 (part of the discourse history in the discourse state) is shown in Figure 2(264). (See, e.g., EX-1005, 335.)

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?

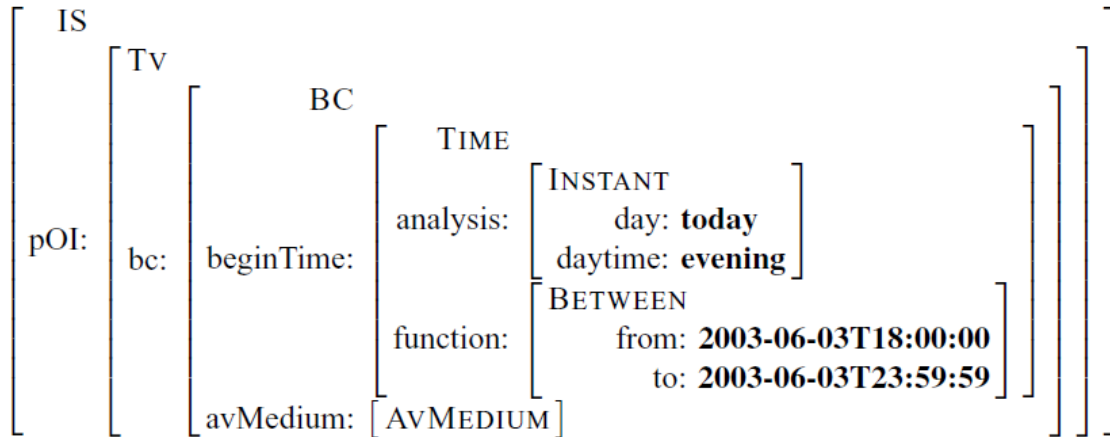


Fig. 2. TFS representation for the utterance *What's on TV tonight?* (*IS* = InformationSearch, *pOI* = pieceOfInformation, *Tv* = TvProgram, *bc* = broadcast)

SmartKom, Figure 2(264)

194. 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. 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. That is, the discourse modeler enriches each with context-domain information.

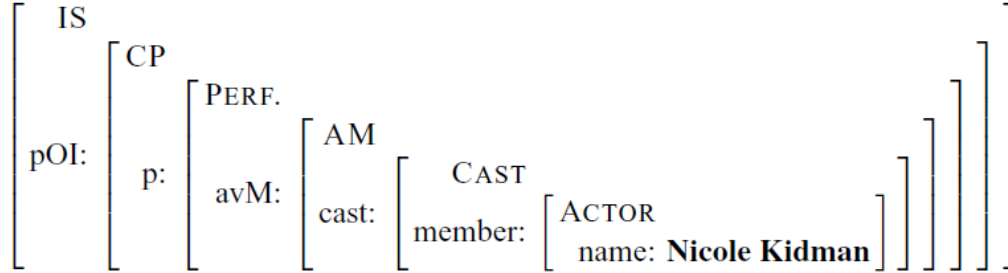


Fig. 3. One possible interpretation of the utterance *Is there a movie with Nicole Kidman?* (*IS* = InformationSearch, *pOI* = pieceOfInformation, *CP* = CinemaProgram, *p* = performance, *P* = Performance, *bc* = Broadcast, *Bc* = Broadcast)

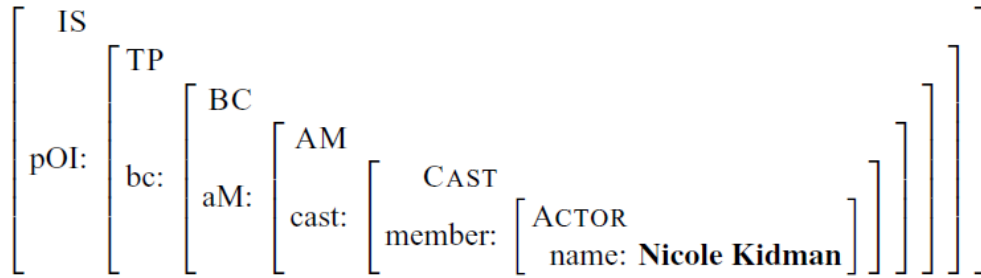


Fig. 4. A second possible interpretation of the utterance *Is there a movie with Nicole Kidman?* (*IS* = InformationSearch, *pOI* = pieceOfInformation, *CP* = CinemaProgram, *p* = performance, *P* = Performance, *aM* = avMedium, *AM* = AvMedium)

SmartKom, Figures 3(264), 4(264)

195. 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 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).

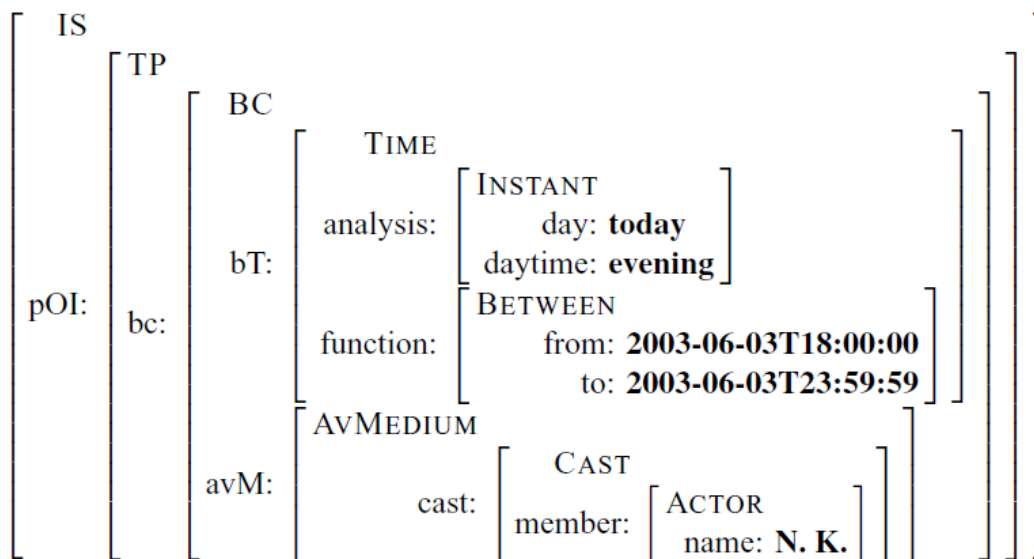


Fig. 5. Interpretation of the utterance *Is there a movie with Nicole Kidman?* with respect to the previous discourse context

SmartKom, Figure 5(265)

196. 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.)

197. SmartKom, as modified with disclosures of Kobsa, uses an “aggregate model of situative and domain knowledge [that] contains the SmartKom ontology (Gurevych et al., 2003b; Porzel et al., 2003b; Gurevych et al., 2006).” (EX-1005, 271.) The SmartKom-Kobsa system uses its contextual knowledge for “**lexical and pragmatic disambiguation**, decontextualization of domain and common-sense

knowledge that was left implicit by the user” and “for estimating an overall coherence score that is used in intention recognition.” (EX-1005, 269.)

198. The context model “assist[s] in evaluating the competing intention hypotheses against each other to find out what was said” and then “such contextual domain and situation knowledge can be used for augmenting such intention hypotheses with implicit information, to **spell out their underlying intentions** and, finally, to define a common background representation for the processed content, i.e., intention lattices in the case of the SmartKom system.” (EX-1005, 271.)

SmartKom summarizes that “a context model” is “employed in the following tasks:

- The explication of situationally implicit information. This task can be further differentiated into two subtasks:
 - provision of information that is indexical—such as time and place—based on common ground and common sense defaults and their dynamic instances, e.g., the current position of the user
 - provision of information that is pragmatic, such as speech acts and intentions and their dynamic instances, e.g., the actual open or closed state (accessibility) of the goal object
- The scoring of individual interpretations in terms of their contextual coherence. Again, this task can be further differentiated into two subtasks:
 - using the ontological domain context to measure the semantic coherence of the individual interpretations, e.g., the ranking *n*-best lists or semantic interpretations thereof

- using dynamic situational and discourse information, e.g., previous ontological contexts of prior turns”

(EX-1005, 271.)

199. Specifically, the context model uses a set of contexts, illustrated in Table 1, associated with a knowledge store (referred to as a “model”). (EX-1005, 274.) The identified “*context*” by the content model (and its associated modeling components) is used to enrich the “intention lattice.” (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)

200. 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. Situational context, including location and time of the utterance, is also added, as shown in exemplary Table 2(277) below.

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)

201. 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. (see, e.g., EX-1005, 364 (“SmartKom domains such as electronic program guides (EPG) for TV’s, cinema programs, and movie information”), 21.)

202. 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.

203. 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]

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

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

[1D.3]/[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

[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

204. The combination of SmartKom and Kobsa discloses “*establish[ing] the intended meaning [for the utterance] within the identified context.*” In the SmartKom architecture, the intention recognizer, engaging the context model, establishes the intended meaning of an utterance within the identified context to disambiguate an intent that the user had in speaking the one or more words that have different meanings in different contexts. As I discussed in §IV.B.5.a, the intention recognizer engages the discourse and context modelers and its associated modeling components to enrich received intention hypotheses and provide associated contextual coherence scores. From these, 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.)

205. As I further explained above in §IV.B.2, SmartKom contemplates sense ambiguity in user-supplied utterances. For utterances containing a sense ambiguity, SmartKom discloses the system “*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.*” 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.)

206. For each of these alternative intention hypotheses, the intention recognizer users the above process. The intention recognizer engages the discourse modeler and context modeler working in concert to “enrich the information in the hypotheses with context knowledge” and “[t]hese 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, and to support the decisions to be taken by the intention recognition.” (EX-1005, 287.) 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.)

207. Thus, the disambiguation of words that have the different meanings in the different contexts occurs through use of identified context, integrated into each intention hypothesis.

208. 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.)

209. 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.)

210. Thus, SmartKom discloses “*establishing the intended meaning within the identified context*” [1D.2] and “*establish an intended meaning for the utterance within the identified context*” [13E] “*wherein the conversational speech engine grammatically or syntactically adapts the response based on the intended meaning established within the identified context*” [1E.2]/[13G.2] and “*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*” [25F].

6. “Generating” Limitation [1E.1]-[1E.2]/[13G.1]-[13G.2]/[25G]

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

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

[1E.2]/[13G.2] wherein the conversational speech engine grammatically or syntactically adapts the response based on the intended meaning established within the identified context.

[25G] generate a grammatically or syntactically adapted response to the utterance based on the intended meaning established within the identified context.
--

211. 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].

212. 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.” (SmartKom, 287.) The action planner, also referred to as the dialogue manager, “*generat[es the] response to the utterance.*” (See SmartKom, 302, Figure 1(303)(below).) The presentation planner, natural language generation [NLG] component, and speech synthesis cooperate to present the response multi-modally to the user. (See SmartKom, 396, 401.)

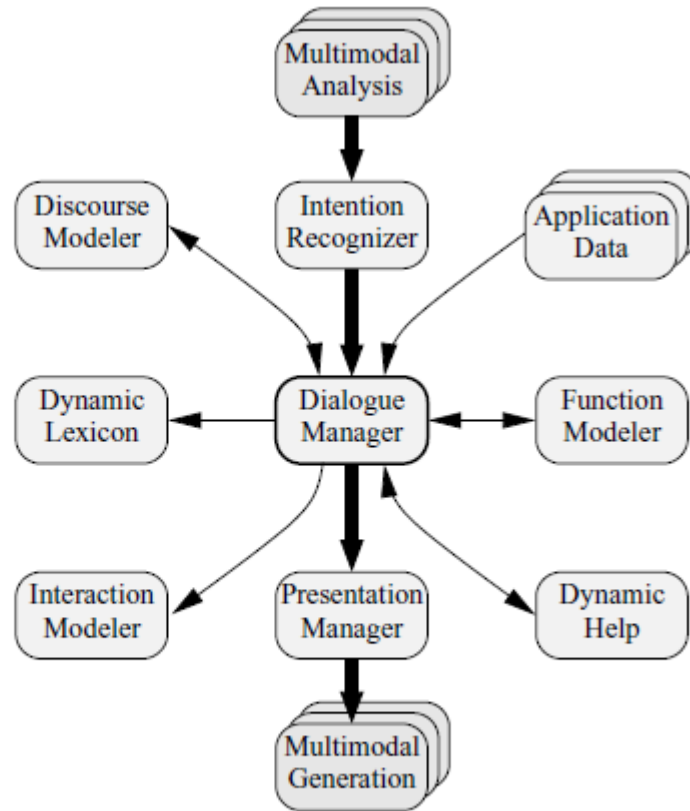


Fig. 1. Modules in direct communication with the dialogue manager

SmartKom, Figure 1(303)

213. 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. (SmartKom, 311.) 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 SmartKom, 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)	

SmartKom, Table 3(313)

214. 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 SmartKom, 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”). 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). (SmartKom, 313 (Table 3).)

215. 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).

216. 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.” (SmartKom, 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.” (SmartKom, 308-309.)

217. 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 SmartKom, 308-309.) That is, “*the response*” is

“grammatically or syntactically” adapted “based on the intended meaning established within the identified context.”

C. Claims 2, 14, 26

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.

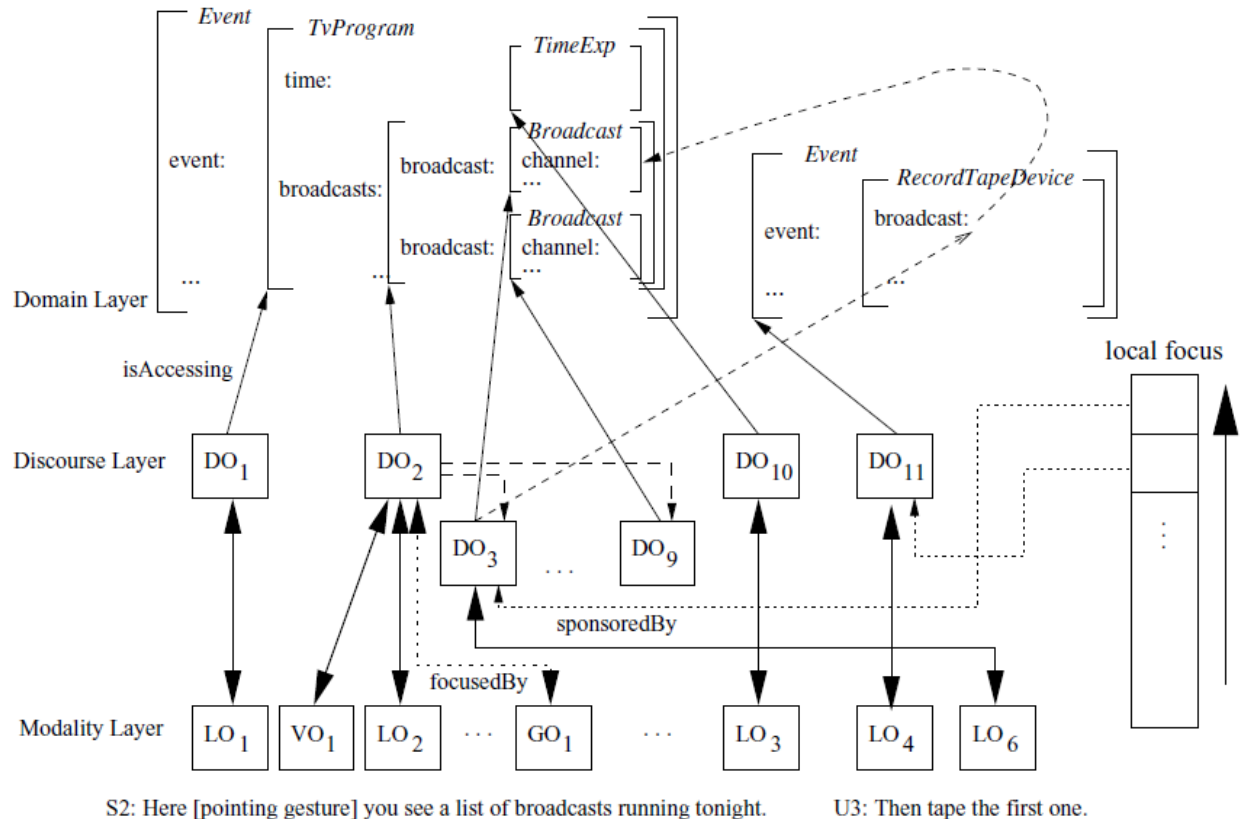
14. The non-transitory computer readable medium of claim 13, wherein to accumulate the short-term shared knowledge about the current conversation, the computer-executable instructions are further operable when executed to populate a short-term context stack with information about the utterance received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

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.

218. SmartKom discloses *“populat[ing]/[e] a short-term context stack with information about the utterance received during the current conversation.”*

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, illustrated below, in Figure 2(243) “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.) As shown below, 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”).) Therefore, SmartKom populates the local focus stack (“*short-term context stack*”) “with information about the utterance received during the current conversation.”



SmartKom, Figure 2(243)

219. Thus, SmartKom discloses “*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*” [2], “*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” [14], and “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” [26].

D. Claims 3, 15, 27

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.

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.

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.

220. 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]. The ’681 patent does not provide any insight into the extent of or the metrics associated with “*a psychologically appropriate amount of time.*” The ’681 Patent merely discloses:

Moreover, because cooperative conversations 300 model human conversations, **short-term session data may be expired after a psychologically appropriate amount of time**, thereby humanizing system behavior. For example, a human is unlikely to recall a context of a conversation from two years ago, but because the context would be identifiable by a machine, session context is expired after a predetermined amount of time **to reduce a likelihood of contextual confusion based on stale data.**

(EX-1001, 4:65-5:2.)

221. The SmartKom combination teaches that short-term information associated with a session decays/is forgotten during a lengthy conversation and is deleted at the end of any conversation. SmartKom explains that its discourse representation is “based on the three-tiered context representation presented in Luperfoy (1992).” (EX-1005, 239.) Luperfoy (EX-1011) describes the concept of “information decay” in a discourse model. (EX-1011, 24.) Luperfoy notes that an “agent participating in a dialogue experiences information decay over the course of the conversation.” (EX-1011, 24.) Luperfoy further explains that “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, linguistic objects (“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.)

222. SmartKom stores its discourse states in “discourse memory.” (EX-1005, 242.) SmartKom acknowledges the decay of information, described by Luperfoy, 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.) Moreover, because of the limited amount of discourse memory, a POSITA would have understood that short-term dialog history is deleted from this working memory at the end of a conversation. Thus, SmartKom describes or suggests the “*psychologically appropriate amount of time*” before expiring information from the discourse model.

223. Thus, SmartKom discloses “*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*” [3], “*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*” [15], and

“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” [27].

E. Dependent Claims 4, 16, 28

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.

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.

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.

224. The combination of SmartKom and Kobsa discloses *“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” [4]/[16]/[28].*

225. As I discussed in §IV.B.4.b, Kobsa teaches accumulating knowledge stored in the user model (“*long-term profile associated with the user*”) either explicitly from asking the user questions during a dialog or implicitly by eavesdropping on the conversation “to observe the user’s behavior and from it to infer facts.” (EX-1006, 416.) This information, acquired directly or implicitly from the spoken utterance, is “*about the utterance received during the current conversation.*” Kobsa describes exemplary user experience/knowledge data accumulated through interpreting utterances in prior conversations reflecting “whether the user understands” basic concepts in the discourse domain. (EX-1006, 179, 17.) This information is illustrated in Figure 3 below. The concepts and their state (known/not-know/no-information) reflect “*information about the utterance received during the current conversation*” because, e.g., the system assesses the state based on whether and how the user speaks about the concept in a dialog. The state and certainty of the state value 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

226. Kobsa further teaches 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.) Kobsa further notes that “at the beginning of a dialog,” the system can “access an already existing model of a particular user which it had constructed during previous interactions with him/her.” (EX-1006, 10.) That is, the system updates the long-term user model with information and any associated relevant data before the session ends and short-term session knowledge is deleted. 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.”

227. Thus, the SmartKom-Kobsa combination discloses “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” [4], “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” [16], and “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” [28].

F. Claims 5, 17, 29

1. Conversational Goal

[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

[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

[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.

228. 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.”*

229. The '681 Patent states “[c]onversational goals may broadly include: (1) getting a discrete piece of information or performing a discrete task” and “[p]articipant roles may broadly include: (1) a leader that controls a conversation, (2) a supporter that follows the leader and provides input as requested, and/or (3) a consumer that uses information.” (EX-1001, 14:39-47.)

230. SmartKom identifies a “*conversational goal*”, e.g., a discrete task to be performed, in response to a user utterance, consistent with the disclosure of the ‘681 patent. SmartKom explains that the intention lattice associated with a hypothesis for an utterance “specif[ies] a set of dialogue acts, which in turn can be either *goal* or *slot manipulations*.” (EX-1005, 305 (emphasis in original).) Accordingly, goal manipulations, such as “setting a goal to adopt a new task,” “retracting a goal to abort a task currently paused,” and “retaining a goal” are important for the dialogue manager when preparing an action plan. (EX-1005, 306.) The system for example identifies the goal when developing the action plan for responding to an utterance based on its established meaning (selected hypothesis). For example, as I explained above, an operator in the action plan includes a “**goal tag**” signifying “**a user-specifiable goal** (i.e., ‘show me the TV program’ or equivalent).” (EX-1005, 308.) Therefore, SmartKom identifies a “*conversational goal associated with the utterance*.”

231. 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. This is illustrated in the example plan (reproduced below) for meeting the goal of “sending a fax message.” (See EX-1005, 311-312.) The channel and task specify the roles and information

allocation in this portion of the conversation. For example, the left arrow next to the term “user” indicates a user output channel (e.g., response provided to a user via an output modality such as speech). (See EX-1005, Table 3(313).) This channel is associated with a task performed by the system such that the role of the system is the leader and the user is the listener/consumer. (See EX-1005, Table 3(313).) For example, in step 4, which is a speech-based (s) “inform” conversation type, the system is the leader and the user is the consumer of the information. (See EX-1005, Table 3(313).) The information in this step is allocated (provided) to the user (e.g., information that “fax being sent”). (See EX-1005, Table 3(313).) A right arrow next to the term “user” indicates a user input channel (e.g., user providing information or a request to the system). (See EX-1005, Table 3(313).) This channel is associated with a task performed by the system such that the role of the user is the leader (controlling the conversation) and the system is the listener. (See EX-1005, Table 3(313).) For example, in step 3-2, the user responds to the system request with information (a telephone number) and this information is therefore allocated to the system. (See EX-1005, 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)	

SmartKom, Table 3(313)

232. Further, In its SmartWeb application, SmartKom (SmartWeb) provides a context-aware user interface that “can support the user in different roles, e.g., as a car driver, a motor biker, a pedestrian or a sports spectator.” (EX-1005, 22.)

233. Thus, SmartKom discloses “*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*” [5A],

“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” [17A], and “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” [29A].

2. Conversation Type

[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.

[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.

[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.

234. 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, wherein the conversational speech engine further establishes the intended meaning based on the conversation type” [5B]/[17B]/[29B].

235. The '681 patent does not expressly define a “*conversation type*.” At most, the patent refers to concepts which may be considered as indicative of utterance types included in a longer conversation. For example, the '681 patent states that “a query conversation may include a conversational goal of getting a discrete piece of information or performing a particular task.” (EX-1001, 14:61-63.) The patent also references a “didactic conversation” in which “a leader of the conversation may control information desired by a supporter of the conversation” and an “exploratory conversation” in which “both participants share leader and supporter roles, and the conversation may have no specific goal, or the goal may be improvised as the conversation progresses.” (EX-1001, 14:67-15:8.) In a mixed-initiative, cooperative conversational dialogue system, a conversation between the user and system may include each of these types of conversation: queries (request-response) and information providing (didactic).

236. SmartKom’s action planner classifies its utterances based on the conversational goal articulated and understood by the system. As I discussed in §IV.B.6, the action plan includes “communicative games” labeled by SmartKom as “*inform, instruct, request-response, and the special type graphicalAction*.” (EX-

1005, 312 (emphasis in original).) A POSITA however would recognize these to also be conversation or utterance types—query (request-response), instruction, and informative. For example, as shown in the example plan (reproduced again below), the system selects the “games” (i.e., conversation types) needed to achieve the goal articulated by the user—i.e., the utterances in the conversation are assigned a conversation type based on the “*identified conversational goal*.” (EX-1005, 312-313.)

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)	

SmartKom, Table 3(313)

237. SmartKom also discloses that the “*conversational speech engine*” “*establishes the intended meaning based on the conversation type.*” The action plan includes the “moves” within the game(s) associated with achieving the user specified goal. (See EX-1005, 308-312.) The action plan thus predicts “future moves of the dialogue” (e.g., need to request information from a user, need for the user to supply information). (EX-1005, 312.) These expected moves “can be used to help the 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.*”

238. Thus, the conversation type associated with the utterance assists with “*establish[ing] the intended meaning*” of the utterance “*based on the conversation type*” and SmartKom discloses limitations [5B]/[17B]/[29B].

G. Claims 6-7, 18-19, 30-31

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.

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.

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,

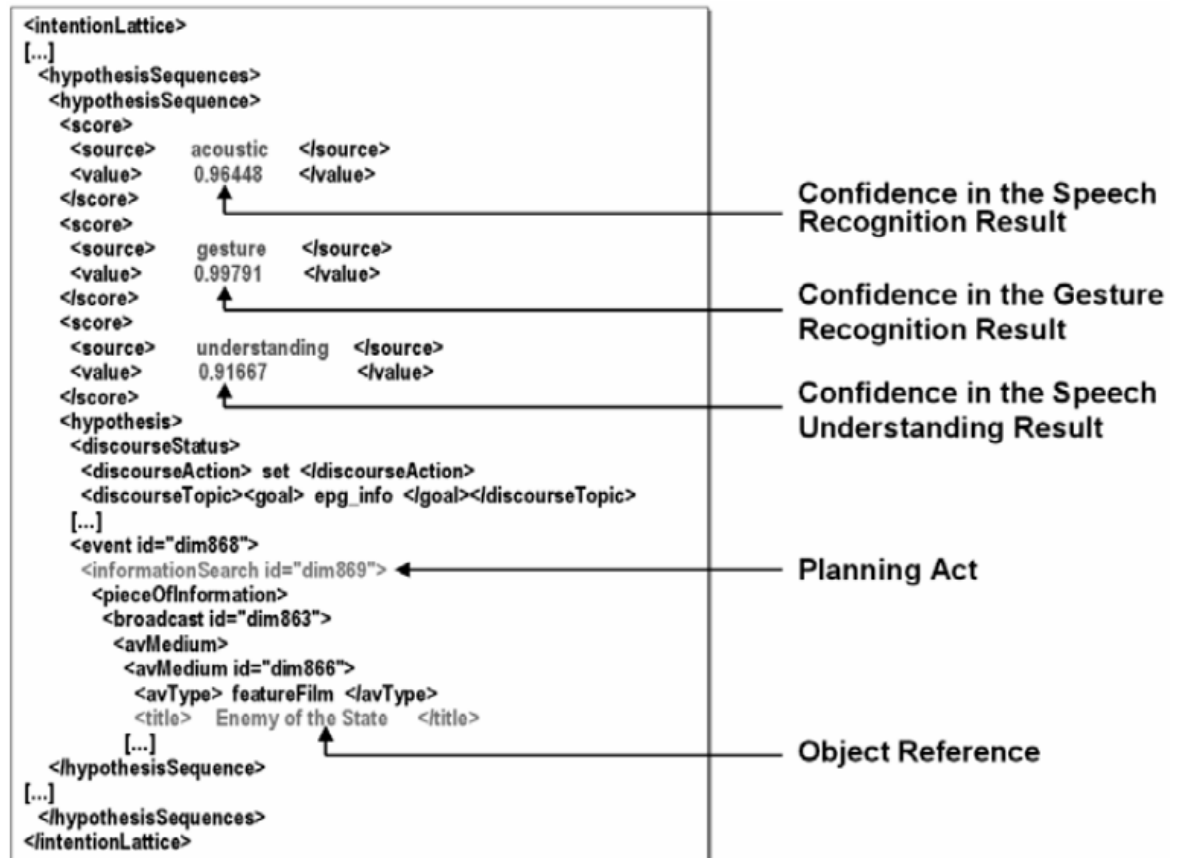
[31B] wherein the one or more processors are further configured to assign the degree of certainty to the hypothesis based on one or more of the conversation type, information associated with the identified context, a confidence level associated with the

preliminary interpretation generated at the speech recognition engine.
--

239. Claims 7/19/31 depend from claims 6.10.30 and based on their interdependence, I address [7A]/[19A]/[31A] first followed by claim [6]/[18]/[30] and then [7B]/[19B]/[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].

240. The SmartKom system includes a “**speech recognition engine**”, coupled to the audio input (“*voice input device*”) and the components that comprise the “*conversational speech engine*” (see § IV.B.5) that “transforms the acoustic signal” representing the user’s vocal utterance “into a **sequence of hypothesized words** in orthographic representation.” (EX-1005, 86, Figure 2(60).) SmartKom’s speech recognition system uses “**confidence measures**” to “estimate confidences for the correctness of the words hypothesized by the recognizer.” (EX-1005, 85; *see also* EX-1005, 96 (“confidence of each word is estimated as the posterior probability for the word being correct”).) SmartKom notes that such “confidence measures” were “common practice in state-of-the-art speech recognition systems.” (EX-1005, 96.) The speech recognition engine adds this “confidence” score for its results to each hypothesis in the intention lattice, as shown in the exemplary fragment below. (See EX-1005, Figure 8(15).) As shown in this example, the speech interpretation

(natural language understanding) component also add a “confidence” score for its result. (See, e.g., EX-1005, 204.)



SmartKom, Figure 8(15)

241. SmartKom therefore discloses that “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] and “the degree of certainty to the hypothesis” is assigned “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].

242. As I discussed in §IV.B.5, during the process of establishing the “*intended meaning*” of an utterance, the intention recognizer receives the intention lattice and engages the modeling components, including the context module, to “compute [a] contextual coherence score[]” of each hypothesis in the lattice. (EX-1005, 276, 271.) The contextual coherence score represents “*a degree of certainty about the intent ...*.” 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].

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)

243. 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, the software executing on the processors of the “*conversational speech engine*” perform these process steps.

244. Thus, SmartKom discloses “*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*” [7B]/[19B] and “*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*” [31B].

H. Claims 8-9, 20-21, 32-33

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.

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.

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.

245. SmartKom discloses “*the conversational speech engine further grammatically or syntactically adapts the response based on conversation type*” [8]/[20]/[32].

246. As I discussed in §IV.B.6, the system adapts its response based on the conversation type established in the action plan. For example, if the conversation type is a request-response and the “move is on an output channel, a corresponding message is constructed [by the system] and sent.” (EX-1005, 308.) For example, in acting on the user goal of “I want to send a fax”, the system generates a series of responses 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.

247. 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].

248. The action plan designs its system responses to influence the user to provide information or perform tasks necessary to achieve goal articulated by user during the conversation. (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).)

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)	

249. SmartKom provides numerous additional examples of the system adapting the response to influence the user’s response, *e.g.*, to provide needed

information to achieve a goal which SmartKom refers to as “slot filling”. The following illustrates filling a slot:

(i) A simple case of filling a **previously unfilled slot**:

U: I want to see a movie.

S: **When do you want to see it?**

U: Tonight.

(EX-1005, 256.) In these slot-filling examples, SmartKom’s “*conversational speech processor*” adapting its response to influence’s the user’s reply to gather the information needed to achieve the goal being processed.

250. Thus, SmartKom discloses “*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*” [9], “*the conversational speech engine grammatically or syntactically adapts more [sic] the response to influence a subsequent reply utterance that the conversational speech engine expects from the user during the current conversation*” [21], and “*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*” [33].

I. Dependent Claims 11-12, 23-24, 35-36

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

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.

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.

251. SmartKom discloses “*the user speaks the utterance in a multi-mode input that further includes one or more non-voice inputs relating to the utterance*” [11]/[23]/[35].

252. SmartKom “introduce[s] the notion of symmetric multimodality for dialogue systems in which all input modes (e.g., speech, gesture, facial expression) are also available for output, and vice versa.” (EX-1005, 3.) SmartKom stresses that “the input channels are always open, i.e., the user can speak or perform gestures at any time.” (EX-1005, 229.)

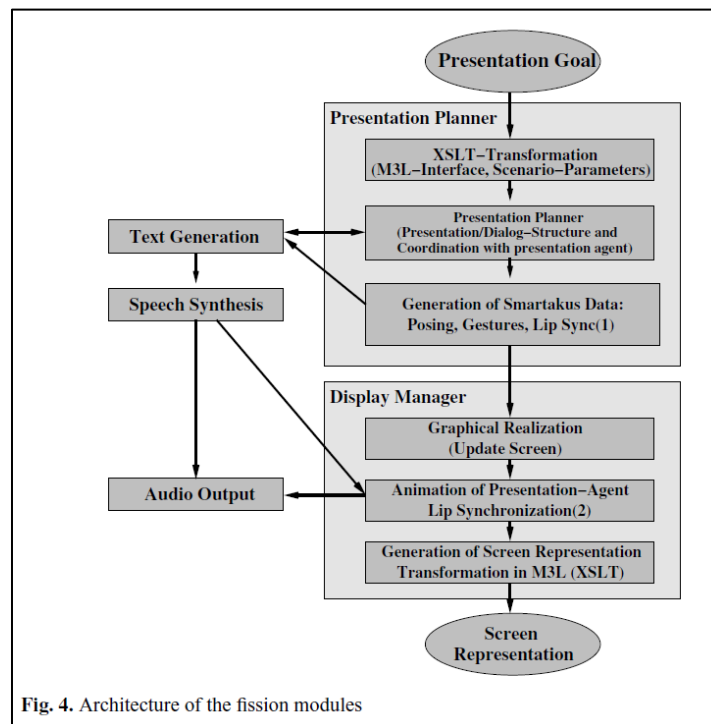
253. The SmartKom system receives a user's utterances, gestures, and facial expressions. SmartKom explains that "[t]he gesture recognition modules supply the SMARTKOM system with video-based interaction capabilities," where "[t]he users **pointing gestures are captured by a downward looking video camera.**" (EX-1005, 153.) SmartKom describes the importance of this multimodal input, stating that "**pointing gestures and speech disambiguate each other.**" (EX-1005, 153.) For instance, "[a] speech phrase 'I want this' and an appropriate pointing gesture provide a very efficient, multimodal approach to information transfer." (EX-1005, 153.)

254. SmartKom also discloses the generated "*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*" [12]/[24]/[36].

255. SmartKom recognizes that "[t]he best way to communicate with a human is to **use different information media simultaneously.**" (EX-1005, 385.) "For example, the use of graphics and speech in a presentation gives a better understanding level than a graphics-only presentation." (EX-1005, 385.) Accordingly, SmartKom's multi-modal output "consists of **synchronized presentations of appropriate graphics and speech output** accompanied and

moderated by corresponding Smartakus gestures and postures, including cross-modal references.” (EX-1005, 384.)

256. SmartKom further explains that “a crucial computational step” in response generation “is the distribution of the output elements over the available modalities (fission) that is based on a dynamic, context-dependent procedure taking care of several influencing factors, e.g., the dialogue context, the available media and modalities, or user preferences.” (EX-1005, 384, Figure 4 (below).) When responding to the user, “[b]esides gestures, mimics, and speech, SMARTKOM also uses graphics, which can be presented on different graphic devices.” (EX-1005, 386.)



SmartKom, Figure 4.

257. SmartKom provides an exemplary dialogue demonstrating multimodality on the input and output and highlighting the multi-modal output is related to the utterance:

(4) USR: Welche Filme mit Arnold Schwarzenegger laufen denn?
(*Are there any movies featuring Arnold Schwarzenegger?*)

...

The user query results in an information display that contains only movies with Arnold Schwarzenegger, as shown in Fig. 5 on the left side. The system utters:

(5) SYS: Hier sehen Sie die Filme die heute abend laufen.
(*These are the movies playing tonight.*)

Now the user query results in an information display that contains only movies with Arnold Schwarzenegger, as shown in Fig. 5 on the left side. The system utters:

(5) SYS: Hier sehen Sie die Filme die heute abend laufen.
(*These are the movies playing tonight.*)

Now the user employs multimodality and, while pointing to one of the movie title, says:

(6) USR: Gib mir Informationen zu diesem.
(*Give me information about this one.*)

(EX-1005, 46-48.)

258. Thus, SmartKom discloses “*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*” [12]/[24] and “*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” [36].

V. GROUND 2: The Combination Of Barbara And Ross Renders Obvious Claims 1, 5-8, 11-13, 17-20, 23-25, 29-32, And 35-36

A. Overview of the Combination

259. The combination of U.S. Patent Application Publication 2004/0101198 to Barbara (“Barbara”; EX-1007), which published May 27, 2004, and U.S. Patent Application Publication 2002/0173960 to Ross, et. al. (“Ross”; EX-1008), which published November 21, 2002, disclose or at least suggest every limitation of claims 1, 5-8, 11-13, 17-20, 23-24, 29-32, 35-36.

1. Barbara

260. Barbara discloses an “interpretation system for interpreting electronic signals.” (EX-1007, Abstract.) Barbara discloses two primary embodiments—(1) where the signal includes an image, e.g., derived from an electronic scanner (EX-1007, ¶¶11-12) and (2) where the signal “include[s] text derived from voice recognition software operable to convert spoken words into electronic text” (EX-1007, ¶23.) My analysis focuses on this second embodiment described in ¶¶82-179 and Figures 6-7.

261. Barbara describes that its system can be used “to create a voice interface where the end system responds to spoken natural language commands.” (EX-1007,

¶82.) When used as a voice interface, “the electronic version of the spoken word is checked for accuracy, corrected and interpreted to determine what the user wanted, i.e. the intent of the spoken word” and “[o]nce the intent is determined, a command signal can be sent to the user’s terminal to carry out that intent.” (EX-1007, ¶82.)

262. Barbara illustrates a exemplary voice interface system in Figure 6, which I reproduce below. The system includes “a user terminal 30 that has a voice interface equipment 32 that is operable to receive the spoken word and translate it into an electronic format.” (EX-1007, ¶83.) Barbara further teaches that the sound may also be recorded. (EX-1007, ¶83.) In an embodiment, “the end-user’s terminal 30 could merely be used as a telephone, indeed it may in fact be a telephone/mobile phone, or audio recorder and the voice recognition software could be provided at a remote location.” (EX-1007, ¶83.)

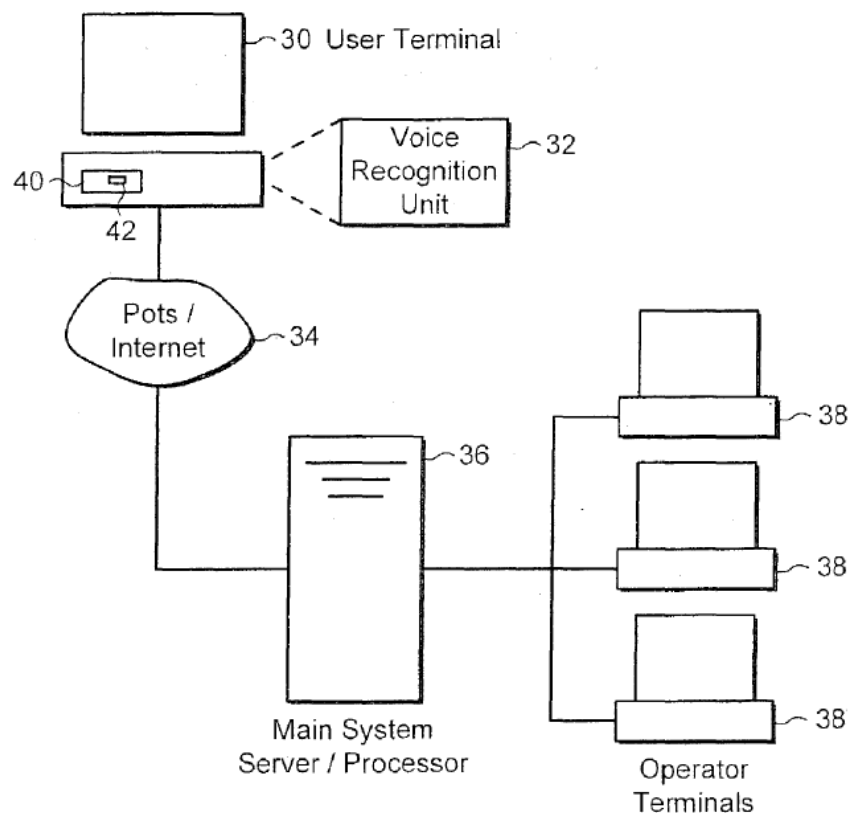


FIG. 6

Barbara, Figure 6

263. “[U]ser terminal 30 is connected via the internet or any other suitable communication system 34, for example a dedicated secure connection, to a main server(s) 36” (EX-1007, ¶84.) “Included in the main server are one or more service processors having software for receiving, interpreting and correcting the incoming information from the user terminal.” (EX-1007, ¶85.)

264. Contextual information “is captured as the user interacts with the system, in much the same way as interactions work in real life.” (EX-1007, ¶98.) The process when the system is used as a voice interface, illustrated in Figure 7

(below), “begins when the end-user says something at their machine” (e.g., user terminal 30). (EX-1007, ¶86.) When an “utterance has been detected,” the voice recognition engine provides a set of recognized text phrases. (EX-1007, ¶86.) For example, if “the user says ‘How’s Tokyo doing?’”, the “voice recognition engine might have detected this and have outputted the following two possible recognitions, with roughly equal probabilities: How’s Tokyo doing? [and] How’s Toto doing?” (EX-1007, ¶¶94-96.) “A request packet is then generated including the best guesses of the voice recognition engine as to what was said [e.g., How’s Toto doing and How’s Tokyo doing], with some indications of their likelihood and also an audio file of the utterance.” (EX-1007, ¶86.)

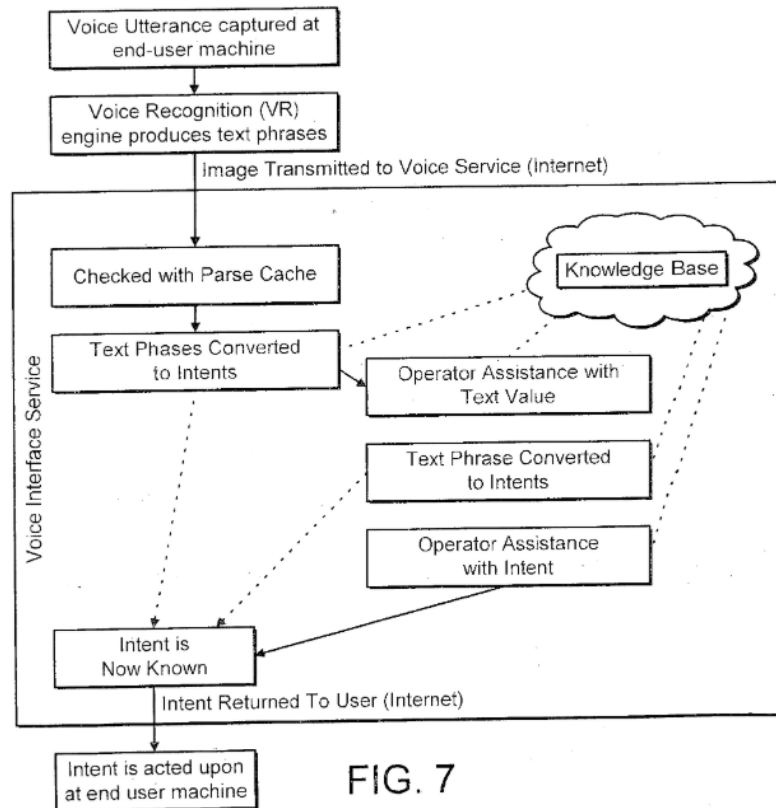


FIG. 7

Barbara, Figure 7

265. When the server receives an utterance, “it is evaluated into a set of possible requests that the user may have actually meant. A combined likelihood for each possible request is heuristically calculated.” (EX-1007, ¶90.) The likelihood is calculated based on a combination of the following:

the likelihood from the voice recognition engine that the text was actually what was said

the likelihood from the service’s knowledge base that the request was what was meant given what was said

an estimation from the user’s information base of which intents are likely.

(EX-1007, ¶¶91-93.)

266. Specifically, “[t]he best guesses of the voice recognition engine as to what was said are evaluated” by the main server “against an information database(s)” which is represented in Figure 7 as the Knowledge Base. (EX-1007, ¶88.) The information database(s)/Knowledge Base “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.)

267. Continuing with the above example, “[t]he knowledge base might interpret ‘How’s Tokyo doing’ as a request for the status of the Tokyo production line, the weather in Tokyo or where the Nikeii is trading” and both of these interpretations “might be, hypothetically, roughly equally likely.” (EX-1007, ¶97.) However, the system’s “interpretation of ‘How’s Toto doing’ might be strongly recognised as a request for the status of the operator’s pet dog Toto.” (EX-1007, ¶97.) Thus, in this example, “[t]he single utterance was expanded to two possibilities for text, which were expanded to four possible intentions.” (EX-1007, ¶97.)

268. During the language understanding processing performed by the main server, “[a]mbiguities are resolved when the system evaluates the likelihood that the user really is asking each of those things.” (EX-1007, ¶98.) As discussed above, contextual information “is captured as the user interacts with the system, in much the same way as interactions work in real life.” (EX-1007, ¶98.) Barbara does not

explain where the captured contextual information is stored and this information is stored in the knowledge base/information databases. In the dialog 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” in the captured utterance. (EX-1007, ¶98.) In contrast, a “user who asks about their dog ten times a day might be considered very likely to be referring to it” and “[s]omeone who is not involved in financial markets is unlikely to be asking about the Nikeii.” (EX-1007, ¶98.)

269. Thus, Barbara teaches that “[t]he history of what people have asked” which is stored in the Knowledge Base “is heuristically used to assist in scoring the possible intents.” (EX-1007, ¶98.) “In this way, a set of possible interpretations for the utterance are scored for likelihood” and “[o]nce this is done, 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.) In Barbara, an automatic acceptance heuristic “accepts interpretations where the interpretation’s score is better than a minimum threshold and the ratio of the best to the second-best score is greater than a set ratio threshold.” (EX-1007, ¶99.)

270. After the correct intent is determined, the system returns the correct intention to the client system which “performs the correct intention.” (EX-1007, ¶¶105-107, Figure 7 (above).) In Barbara, “[t]he exact way in which the system looks and responds depends on the nature of the system being operated.” (EX-1007, ¶107.)

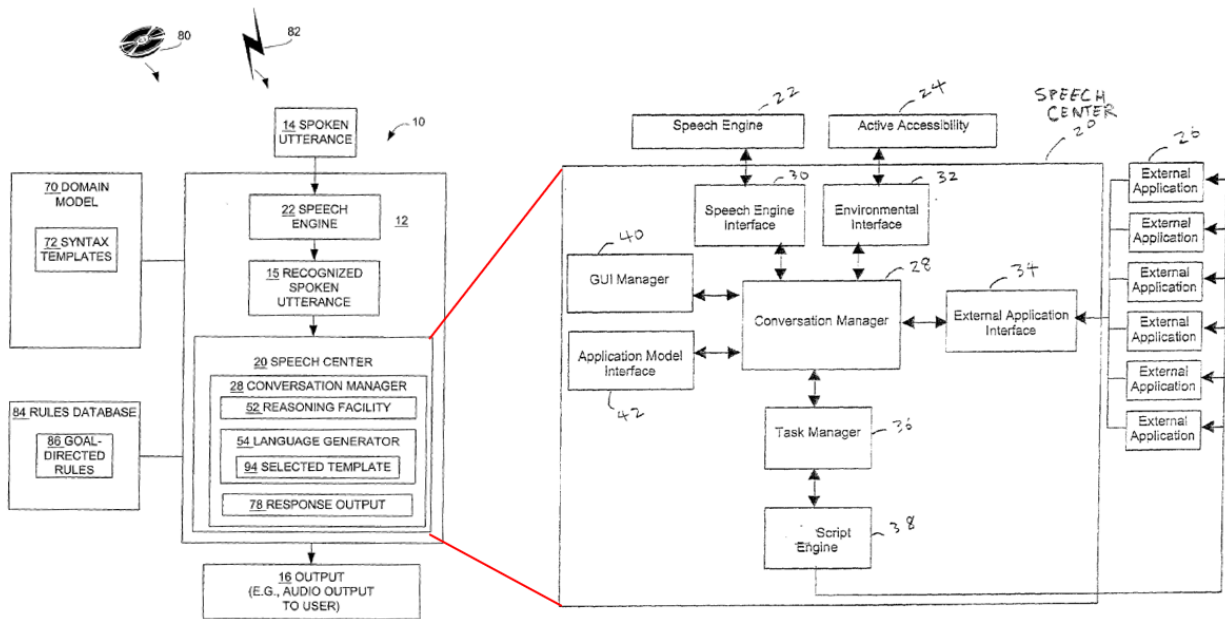
271. Barbara also provides a correction technique “[w]here the utterances have failed to meet the pre-determined criteria for automatic acceptance...” (EX-1007, ¶¶100-103.) When this occurs, the utterances “are sent to a human operator” who selects “the correct textual interpretation from the list produced by the voice recognition engine.” (EX-1007, ¶100.) Once the “text has been deduced, the interpretation process begins again,” but now because the system is “only considering one textual phrase with a much higher confidence, it may now be possible automatically to deduce the user’s intent with a higher degree of accuracy.” (EX-1007, ¶103.) Barbara’s system also involves a second human operator when “it is not possible to deduce what the user wanted to do despite knowing what they said.” (EX-1007, ¶104.) Barbara anticipates that operator intervention and correction “will be used heavily during” an “initial period, until the voice recognition engine becomes tuned to the user’s voice.” (EX-1007, ¶102.)

2. Ross

272. 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 “conversation manager includes a reasoning facility and a language generation module The reasoning facility selects a syntax template to use in generating a response output from the formal belief structure. The language generation module produces the response output

based on the formal structure, the selected syntax template, and the domain model.” (EX-1008, Abstract.) “The present invention provides for consistency between the input and output, without requiring the user to conform to a limited set of fixed phrases, as in conventional approaches.” (EX-1008, ¶7.)

273. Ross’s Figures 1-2 (below) “is an illustration of a preferred embodiment in a computer system 10 [that] includes a digital processor 12 which hosts and executes a **speech center system 20**, conversation manager 28, and speech engine 22 in working memory. The input spoken utterance 14 is a voice command or other audible speech input from a user of the computer system 10 (e.g., when the user speaks into a microphone connected to the computer system 10) based on common language words ... The speech center system 20 includes a conversation manager 28 which generates an output 16 based on the recognized spoken utterance 15.” (EX-1008, ¶22, Figs. 1-2 (below).) Ross’s Figure 2 (below) “shows the components of [the] **speech center system 20**, configured according to the present invention. The speech center 20 includes a **conversation manager 28**, speech engine interface 30, environmental interface 32, external application interface 34, task manager 36, script engine 38, GUI manager 40, and application module interface 42.” (EX-1008, ¶24.)



Ross, Figure 1

Ross, Figure 2

274. In Ross, “[t]he conversation manager 28 is the central component of the speech center 20 that integrates the information from all the other modules 30, 32, 34, 36, 38, 40, 42 When an utterance 15 is recognized, the conversation manager 28 combines an analysis of the utterance 15 with information on the state of the desktop and remembered context from previous recognitions to determine the intended target of the utterance 15.” (EX-1008, ¶32.)

275. Figure 3 of Ross “represents the structure of the conversation manager 28 in a preferred embodiment. Each of the functional modules, such as semantic analysis module 50, reasoning facility module 52, language generation module 54, and dialog manager 56, are indicated by plain boxes without a bar across the top. Data abstraction modules, such as the context manager 58, the **conversational**

record 60 [shaded green], the syntax manager 62, the ontology module 64, and the lexicon module 66 are indicated by boxes with a bar across the top.” (EX-1008, ¶133.)

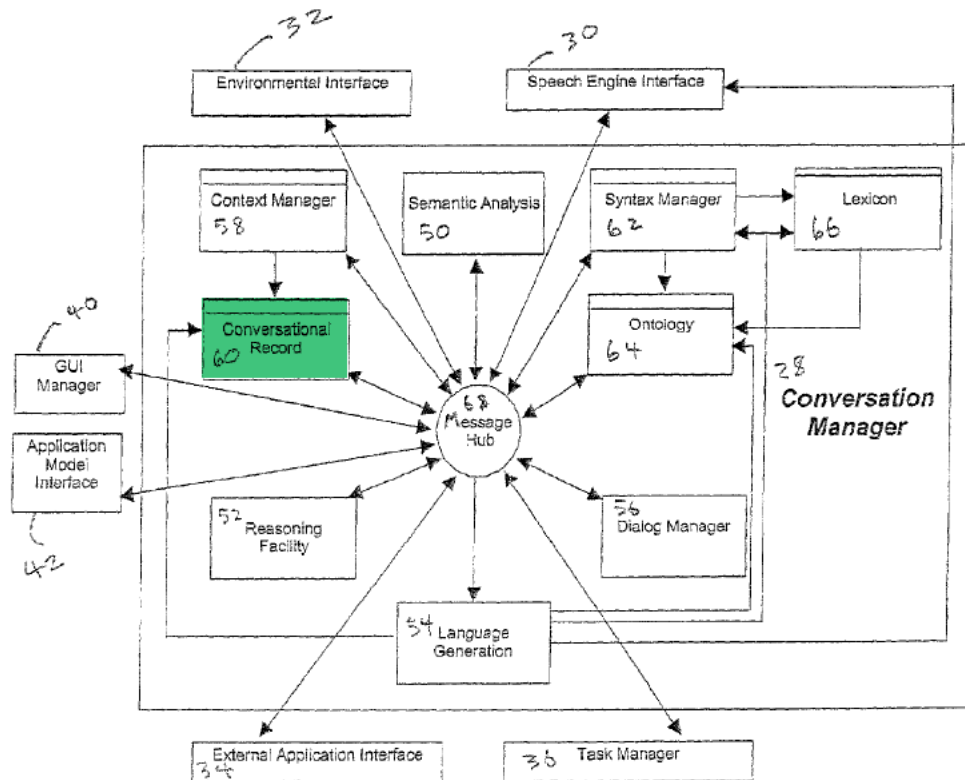


Fig. 3

Ross, Figure 3

276. Ross explains that the conversational manager 28 includes reasoning facility 52, which “performs the reasoning process for the conversation manager 28.” (EX-1008, ¶156.) Ross states that “[t]he reasoning facility 52 is primarily concerned with the determination of how to achieve the goals derived from the user’s questions and commands.” (EX-1008, ¶156.)

277. Ross explains that “dialog manager 56 serves as a traffic cop for information flowing back and forth” between reasoning facility 52 and the user including “deciding whether a speech center-generated response should be visible or audible” and “whether the response can be presented immediately.” (EX-1008, ¶¶59-60.) Conversation manager 28 also stores conversational record 60 including each utterance in a conversation and domain model 70. (EX-1008, ¶¶56-57.)

278. Ross discloses that its system uses conversational records and domain knowledge to process utterances, specifically that the reasoning facility 52 and language generation module 54 “generate[] a natural language response output 78 to the recognized spoken utterance 15 based on” domain model 70, rules database 84, and syntax template 94. (EX-1008, ¶22.) Ross notes that “[c]onversational speech is full of implicit and explicit references back to people and objects that were mentioned earlier.” (EX-1008, ¶57.) As Ross explains, “[t]o understand these sentences, the speech center system 20 looks at the **conversational record 60**, and finds the missing information. Each utterance is indexed in the conversational record 60, along with the results of its semantic analysis. The 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.)

279. Ross discloses an example where “after having said, ‘Create an appointment with Mark at 3 o’clock tomorrow’, a user might say ‘Change that to 4

o'clock.' The speech center system 20 establishes that a time attribute of something is changing, but needs to refer back to the conversational record 60 to find the appointment object whose time attribute is changing. Usually, the most recently mentioned object that fits the requirements will be chosen, but in some cases the selection of the proper referent is more complex, and involves the goal structure of the conversation.” (EX-1008, ¶58.)

280. Ross also “provides a language generation method that performs its work in the context of a domain model for a particular application. A domain model consists of several types of information. The most basic of these is the ontology, in which a developer specifies the entities, classes, and attributes that define the domain of discourse for a particular application. A lexicon provides information about the vocabulary used to talk about the domain.” (EX-1008, ¶5.) Ross also discloses that “[w]ith the addition of syntax templates expressed in terms of the ontology definitions, a grammar can be automatically generated for the domain, and output questions and responses in the domain can also be generated. Rules allow some simple automated reasoning within the domain, which provides an approach for the appropriate syntax template to be chosen for generating the output in response to the user.” (EX-1008, ¶5.)

281. Ross’s “language generation (LG) module uses syntax templates (in conjunction with information contained in the ontology and lexicon) to generate

questions and responses to the user. The language generation module uses rules to select which syntax templates to use for a given goal or propositions (goals and propositions are the formal belief structures manipulated by the reasoning component of the conversational system).” (EX-1008, ¶6.) In Ross’s system “[e]ither questions or answers can be generated. Questions are the natural output form for unrealized goals from the reasoning system; answers are the natural output form for propositions from the reasoning system.” (EX-1008, ¶6.)

282. Ross’s Figure 4 (below) “is a block diagram of the language generation module 54 (language generator) and associated components (reasoning facility 52, domain model 70, and language generation (LG) templates 74) according to the present invention.” (EX-1008, ¶61.)

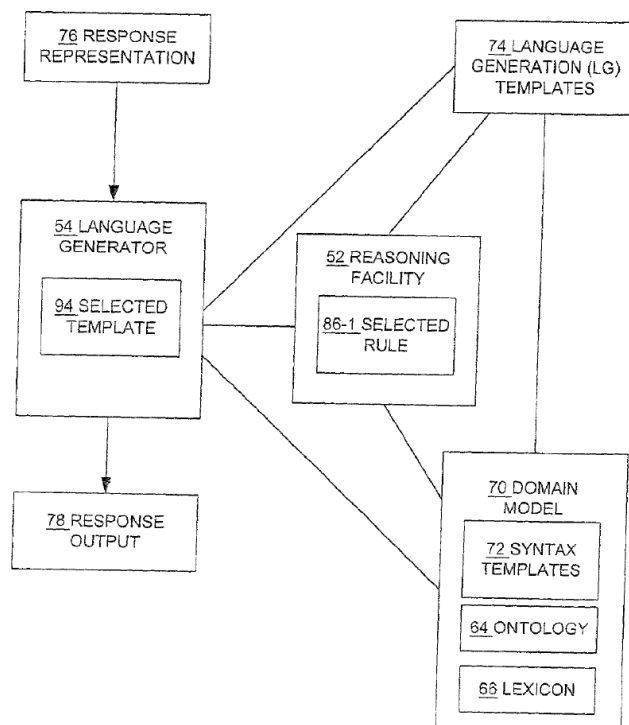


FIG. 4

Ross, Figure 4

283. In Ross, “[t]he response output 78 is a natural language response (e.g., text string), such as a statement or question, generated by the language generation module 54.” (EX-1008, ¶61.) Ross also discloses that “the language generation module 54 generates language from the same syntax templates 72 used for recognition, or from additional templates provided specifically for language generation. These additional templates are the language generation (LG) templates 74.” (EX-1008, ¶62.)

284. Ross explains that “the LG syntax templates 74 are defined at the top level for speech center-generated questions and assertions (these are distinguished with an ‘LGTemplate’ label from other syntax templates 72 in a syntax template file). These LG templates 74 can then reference new or existing (i.e. background or foreground) templates 72 in the domain model 70, where the majority of information about syntactic forms in the speech center 20 is represented. The special LG templates 74 are defined for the language generation module 54 for two reasons. One reason is to avoid having computer-generated questions and responses appear in the user input grammars. Another reason is to control the argument structure to pass arguments as needed.” (EX-1008, ¶76.)

3. Motivation To Combine

285. 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.

286. Barbara and Ross are analogous art to the ’681 Patent. All three pertain 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.).)

287. Barbara motivates the combination. For example, Barbara discloses the system “may either speak back or perform some action (e.g. start the spin cycle) 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. That is, a POSITA would have been motivated to combine Barbara with Ross to provide a system with flexible response generation capabilities.

288. While Barbara discloses use of an “information database” and “knowledge base” to evaluate meaning of 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. A POSITA would understand this data management approach improves storage efficiency and reduces need for extensive hardware storage.

289. 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.

290. 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. 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.

B. Independent Claims 1, 13, and 25

291. The independent claims (claims 1, 13, and 25) of the '681 patent includes substantially overlapping subject matter as highlighted by the chart below. Because this overlap, I address the independent claims together in this section.

Claim 1	Claim 13	Claim 25
[1P] A method for providing a cooperative conversational voice user interface, comprising	[13P] A non-transitory computer readable medium containing computer-executable instructions for providing a cooperative conversational voice user interface, the computer-	[25P] A system for providing a cooperative conversational voice user interface, comprising:

Claim 1	Claim 13	Claim 25
	executable instructions operable when executed to:	
		[25A.1] a voice input device configured to
[1A.1] receiving an utterance at a voice input device during a current conversation with a user,	[13A.1] receive an utterance at a voice input device, during a current conversation with a user,	[25A.2] receive an utterance during a current conversation with a user,
[1A.2]/[13A.2]/[25A.3] wherein the utterance includes one or more words that have different meanings in different contexts;		
		[25B] a conversational speech engine, wherein the conversational speech engine includes one or more processors configured to:
[1B.1] accumulating short-term shared knowledge about the current conversation,	[13B.1]/[25C.1] accumulate 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;	[13B.2] wherein the short-term shared knowledge includes knowledge about the utterance received at the voice [sic] during the current conversation;	[25C.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,	[13C.1]/[25D.1] accumulate long-term shared knowledge about the user,	

Claim 1	Claim 13	Claim 25
[1C.2]/[13C.2]/[25D.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	[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;	[25E] identify a context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge;
[1D.2] establishing the intended meaning within the identified context,	[13E] establish an intended meaning for the utterance within the identified context,	[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
[1D.3]/[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		
[1E.1] generating a response to the utterance,	[13G.1] generate a response to the utterance,	[25G] generate a grammatically or syntactically adapted response to the utterance based on the intended meaning established
[1E.2]/[13G.2] wherein the conversational speech engine grammatically or syntactically adapts the		

Claim 1	Claim 13	Claim 25
response based on the intended meaning established within the identified context.		within the identified context

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

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

[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:

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

292. The combination of Barbara and Ross discloses a “*method*” [1P], a “*non-transitory computer readable medium containing computer-executable*” [13P], and a “*system*” [25P] “*for providing a cooperative conversational voice user interface*” [25P].

293. Barbara discloses a “*method*” [1P] and “*system*” [25P]. Barbara “relates to a **system and method** for improving accuracy of signal interpretation.” (EX-1007, ¶7; *see also* EX-1007, Abstract (describing an “interpretation system”).) Barbara presents two primary embodiments—one directed to optical character recognition (*see* EX-1007, ¶¶33-81, Figs. 1-5) and one directed to voice recognition (*see* EX-1007, ¶¶82-179, Figs. 6-7.) My analysis focuses on the embodiment directed to voice recognition/spoken dialog. Regarding this embodiment, Barbara

describes use of its “system” “to create a voice interface where the end system responds to spoken natural language commands.” (EX-1007, ¶82.) Barbara further describes “the process” including “[t]he process steps for using the ”system“ which ”begins when the end-user says something at their machine.“ (EX-1007, ¶86.) Although Barbara does not use the term “*method*” in its discussion of the voice recognition system, Barbara repeatedly refers to its “present invention” as providing both a “system and method”:

An object of the present invention is to provide a system and method for providing corrected interpretations of electronic signals to users, which system and method are simple and cheap to implement, but highly effective.

(EX-1007, ¶5.)

The invention also relates to a system and method for interpreting natural-language text, either allowing language translation or alternatively allowing the intention in natural-language text to be deduced. The invention also relates to a system and method for construction and development of expert systems.

(EX-1007, ¶7.)

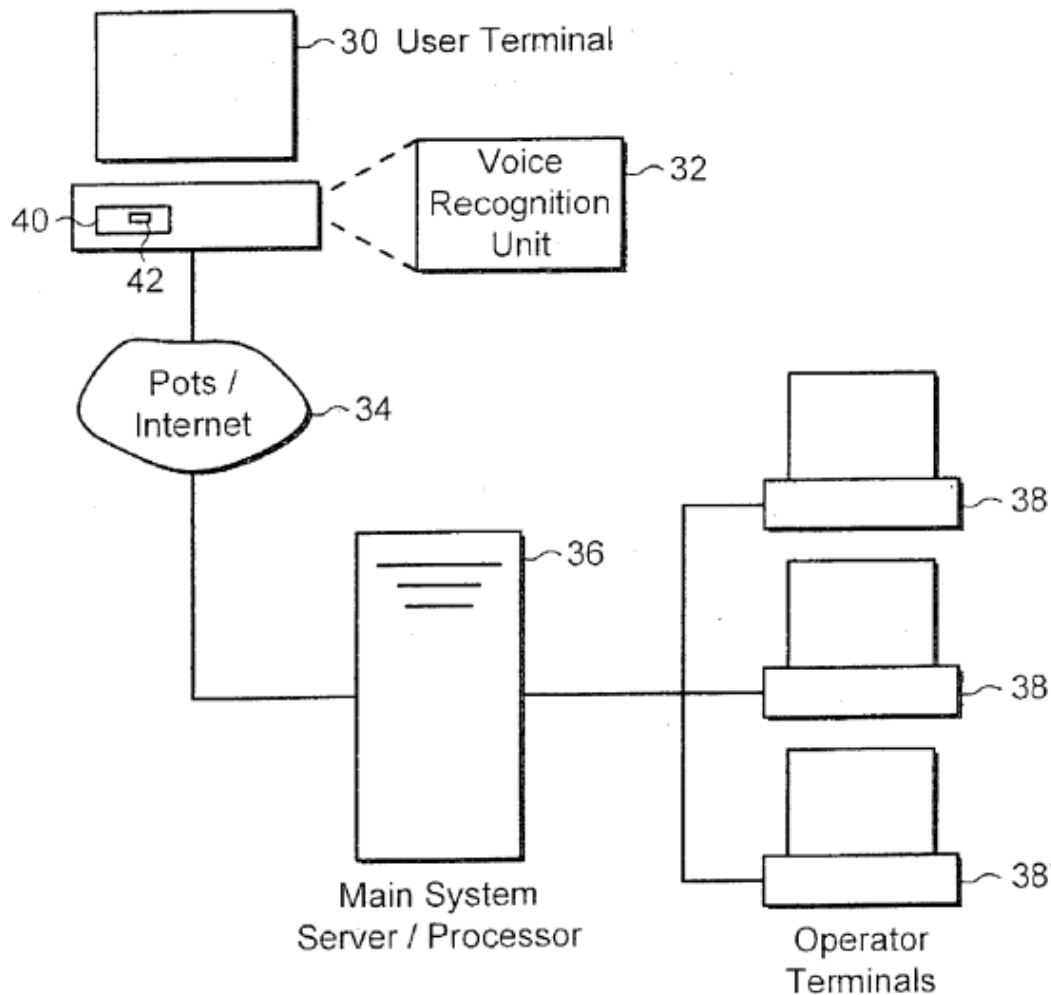
294. Moreover, Barbara recites in its claims a “method of interpreting electronic signals” including “receiving the electronic signal via a telecommunications system from a user terminal” and “processing the signal electronically to interpret it.” (EX-1007, claim 17.) Thus, a POSITA, and even a lay

person, would have understood a “process” referenced in the discussion of Figure 7 is a “method” [1P].

295. Barbara further discloses 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*” [13P] perform the steps specified in claim 13. I note that the specification of the ’681 patent does not use the language “*computer-executable instructions*” or discuss software, programs, or “*instructions.*” At most, the ’681 patent mentions inclusion of a “conversational language **processor**” included in the conversational speech engine. (See EX-1001, 7:40-43.)

296. As shown in Figure 6 below, Barbara’s system includes a main server having “one or more service processors having **software** for receiving, interpreting and correcting the incoming information from the user terminal.” (EX-1007, ¶85.) Although Figure 6 includes voice interface equipment 32 at a user terminal, Barbara teaches that “the end-user’s terminal 30 could merely be used as a telephone, indeed it may in fact be a telephone/mobile phone, or audio recorder and the voice recognition software could be provided at a remote location.” (EX-1007, ¶83.) Thus, the “software” associated with the “voice recognition” process is also included in the main server. Consistent with this disclosure Barbara discloses that “[a]ccording to a further aspect of the invention, there is provided ”**a computer program**

[“*computer-executable instructions*”], **preferably on** a data carrier or a **computer readable medium** [“*non-transitory computer readable medium*”], for controlling a method of interpreting electronic signals....“ (EX-1007, ¶24; *see also* EX-1007, claim 30 (“**computer program**, preferably on a data carrier or a computer readable medium, for controlling a method of interpreting electronic signals, the computer program comprising instructions for: receiving the electronic signal via a telecommunications system from a user terminal;“ and ”processing the signal electronically to interpret it;“); §IV.B.1.) As I discuss in the sections that follow, the server of the combination of Barbara and Ross performs the recited steps of claim 13 and therefore “*computer-executable instructions*” executed on cause the system to perform the recited steps.



Barbara, Figure 6

297. Barbara repeatedly refers to its system, illustrated in Figure 6 above, as providing a “*voice user interface*” that is both “*cooperative*” and “*conversational*.” As I mentioned above, Barbara teaches its system “create[s] **a voice interface** where the end system responds to spoken natural language commands.” (EX-1007, ¶82.) Specifically, Barbara describes “process steps for using the system” of Figure 6 “as a voice interface for carrying out a user intent.” (EX-1007, ¶86.) For example, in Barbara’s **system and method**, “the electronic version of the spoken word is

checked for accuracy, corrected and interpreted to determine what the user wanted, i.e. the intent of the spoken word” the “[o]nce the intent is determined, a command signal can be sent to the user’s terminal to carry out that intent.” (EX-1007, ¶82.)

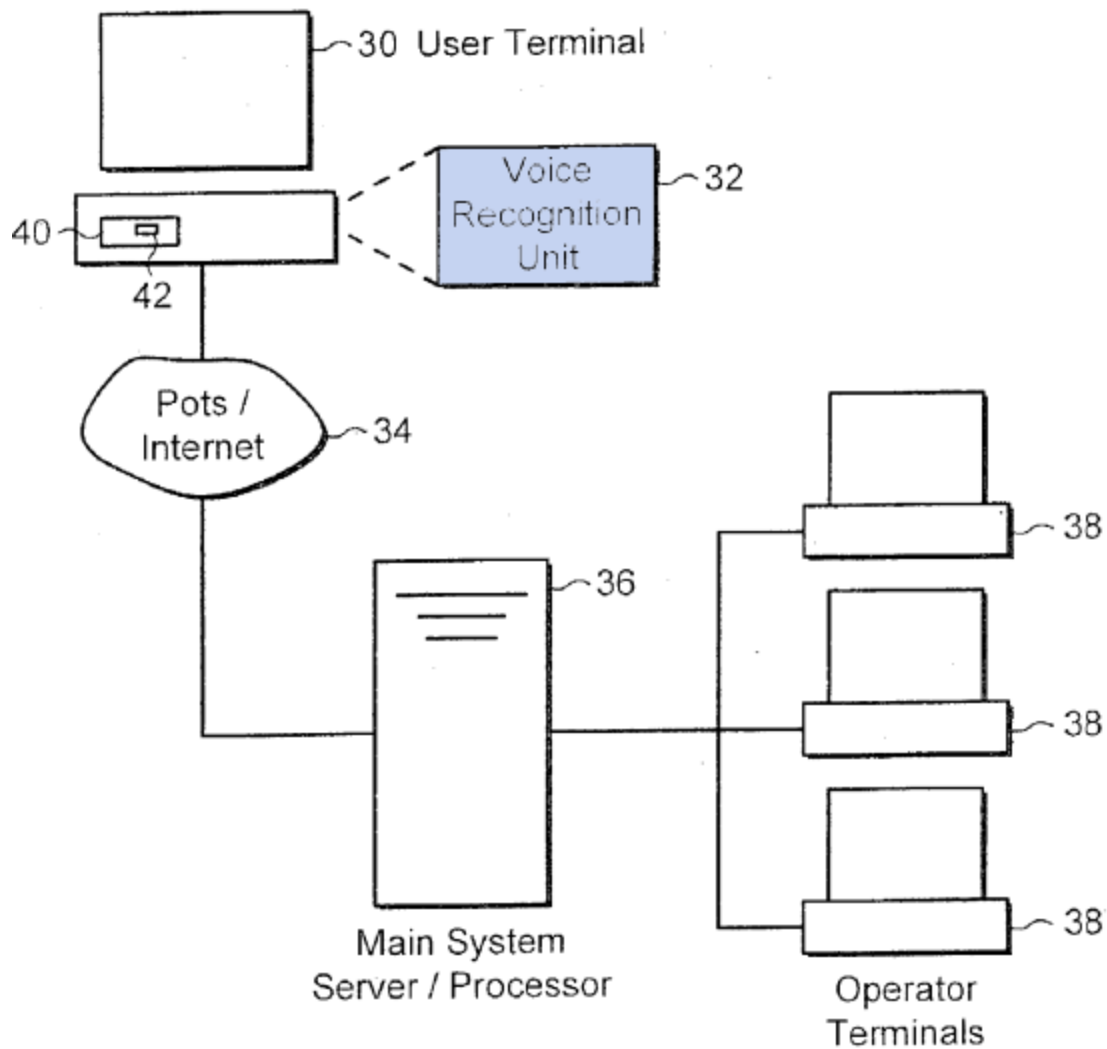
2. Receiving An Utterance [1A]/[13A]/[25A]

<p>[1A.1] receiving an utterance at a voice input device during a current conversation with a user,</p> <p>[13A.1] receive an utterance at a voice input device, during a current conversation with a user,</p> <p>[25A.1] a voice input device configured to</p> <p>[25A.2] receive an utterance during a current conversation with a user,</p> <p>[1A.2]/[13A.2]/[25A.3] wherein the utterance includes one or more words that have different meanings in different contexts; and</p>

298. Barbara discloses “*a voice input device*” that “*receive[s] an utterance*” “*during a current conversation with a user*” [1A.1]/[13A.1]/[25A.1]/[25A.2].

299. Barbara discloses “*a voice input device.*” In Barbara’s Figure 6, user terminal 30 “has a **voice interface equipment** 32 [shaded blue] that is **operable to receive the spoken word** and translate it into an electronic format.” (EX-1007, ¶83.) Barbara recognizes that “[t]he process of voice recognition” performed by the voice recognition unit is “standard.” (EX-1007, ¶86.) Barbara further teaches that user terminal 30 “could merely be used as a telephone, indeed it may in fact be a telephone/mobile phone, or audio recorder and the voice recognition software could

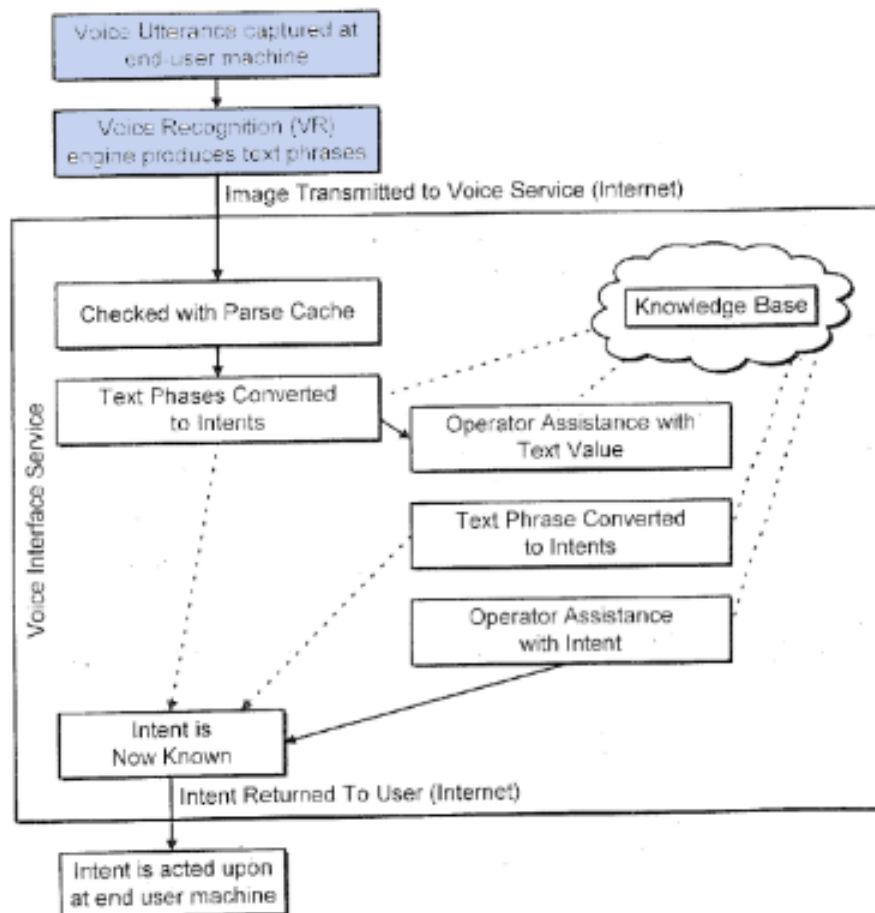
be provided at a remote location.” (EX-1007, ¶83.) That is, the voice recognition unit 32 is included in the server 36 (e.g., a remote location from the user terminal 30). The voice recognition unit 32, whether in the user device or in the centralized main system server, alone or in combination with the component in the user device, such as a telephone/mobile phone (e.g., a microphone), is the “*voice input device*.” Barbara explains that when the user speaks, the “voice recognition/interface service” “recognize[s] **that the user has said something** and determine[s] what has been said.” (EX-1007, ¶86.) Specifically, “software detects when the user is speaking” and “[w]hen the application is informed by the voice recognition package that **an utterance has been detected**, the software asks the voice recognition engine for its recognized text phrases.” (EX-1007, ¶86.) Therefore, Barbara’s “*voice input device*” “receive[s] *an utterance*” from a user “*during a current conversation*” between the system and the user.



Barbara, Figure 6

300. As I noted above, Barbara teaches that the voice recognition unit 32 is in the server 36. Accordingly, the steps highlighted in blue in Figure 7 are performed by the server. That is, the voice recognition engine at the server (“*voice input device*”) captures the voice utterance and produces the text phrases. Thus, the “*computer readable medium containing computer-executable instructions for*

providing a cooperative conversational voice user interface” includes “instructions” that perform the step of “receiv[ing] an utterance.”



Barbara, Figure 7

301. The combination of Barbara and Ross discloses “*the utterance includes one or more words that have different meanings in different contexts*” [1A.2]/[13A.2]/[25A.3]. As I discussed in §III.A.3.b, language understanding in a spoken dialog/natural language system is problematic because of ambiguity in the natural language and ill-formed input. (See, e.g., EX-1020, 91.) Natural language can be ambiguous due to lexical ambiguity where “[a] word may belong to more

than one part of speech, for example, ‘book’ can be a noun or a verb.” (EX-1020, 91.) Natural language can also be ambiguous due to sense ambiguity where “[a] word can have different meanings, for example, ‘bank’ can be a financial institution or the side of a river.” (EX-1020, 91.) Barbara teaches resolving both ill-formed input and these ambiguities where one or more words have different meanings in different contexts.

302. Barbara provides an example illustrating resolving ill-formed input and ambiguities based on words that sound similar and “*one or more words that have different meanings in different contexts.*” Specifically, Barbara presents the example of the user saying “How’s Tokyo doing?”. (EX-1007, ¶94.) The voice recognition engine “detect[s] this and [] output[s] the following two possible recognitions, with roughly equal probabilities: How’s Tokyo doing? [and] How’s Toto doing?” (EX-1007, ¶¶94-96.) That is, the user input is “ill-formed” in the sense that the voice recognition system is unable to distinguish clearly between “Tokyo” and “Toto.”

303. Both possible outputs (Tokyo versus Toto) are ambiguous because each is subject to different meanings depending on the context. For example, Barbara’s “knowledge base might interpret ‘How’s Tokyo doing’ as a request for the status of the Tokyo production line, the weather in Tokyo or where the Nikeii is trading.” (EX-1007, ¶97.) Each interpretation “might be, hypothetically, roughly equally likely.” (EX-1007, ¶97.) Similarly, the “interpretation of ‘How’s Toto doing’ might

be strongly recognised as a request for the status of the operator’s pet dog Toto.” (EX-1007, ¶97, *see also* EX-1007, ¶148.) However, the Barbara system recognizes this utterance may also relate to “Toto the restaurant.” (EX-1007, ¶148.)

304. 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.) A combined likelihood is calculated based on the following:

likelihood from the voice recognition engine that the text was actually what was said,

likelihood from the service’s knowledge base that the request was what was meant given what was said, and

estimation from the user’s information base of which intents are likely.

(EX-1007, ¶¶90-93.)

305. As an example, Barbara explains a “single utterance was expanded to two possibilities for text” by the voice recognition engine “which were expanded to four possible intentions.” (EX-1007, ¶97.)

306. In Barbara, “[a]mbiguities are resolved when the system evaluates the likelihood that the user really is asking each of those things.” (EX-1007, ¶98.) Accordingly, Barbara’s system is designed to handle “sense ambiguities” although Barbara provides only one example (Toto-dog versus Toto-restaurant). For example, for the output “How’s Toto doing”, “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.” (EX-

1007, ¶98.) Barbara’s system uses contextual information in resolving this type of ambiguity:

Contextual information such as this is captured as the user interacts with the system, in much the same way as interactions work in real life. If the user has never referred to their dog before, they are likely to introduce it in some way when first using in conversation. When they do not, and the system responds incorrectly or not at all, then the user may clarify their initial request and then the contextual information may be captured. A user who asks about their dog ten times a day might be considered very likely to be referring to it. Someone who is not involved in financial markets is unlikely to be asking about the Nikeii. The history of what people have asked is heuristically used to assist in scoring the possible intents.

(EX-1007, ¶98.)

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 or Tokyo the city, which the VR engine may have suggested as a possible hearing. The database has flea vaccination information for Toto but not for the others, so the score for the canine interpretation may be raised. In this way the system ‘knows’ about the real world and can resolve ambiguities in many difficult sentences.

(EX-1007, ¶148.)

307. That is, based on the appropriate context, e.g., “pets” or “dining,” the word “Toto” can either mean “dog” or “restaurant.”

3. Conversational Speech Engine

308. Independent claims 1, 13, and 25 each recite “*a conversational speech engine*,” consistent with the ’681 patent. (§IV.B.3; EX-1001, 7:37-40, Figure 2.)

System claim 25 recites the “*conversational speech engine*” as a standalone component “*configured to*” perform the actions of “*accumulate short-term knowledge*”, “*accumulate long-term knowledge*”, “*identify a context*” and “*establish an intended meaning*.” Method and computer program product claims recite the “*conversational speech engine*” as part of the “*identify a context*” and “*establish*” limitation. I recite each of these uses below.

[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;

[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

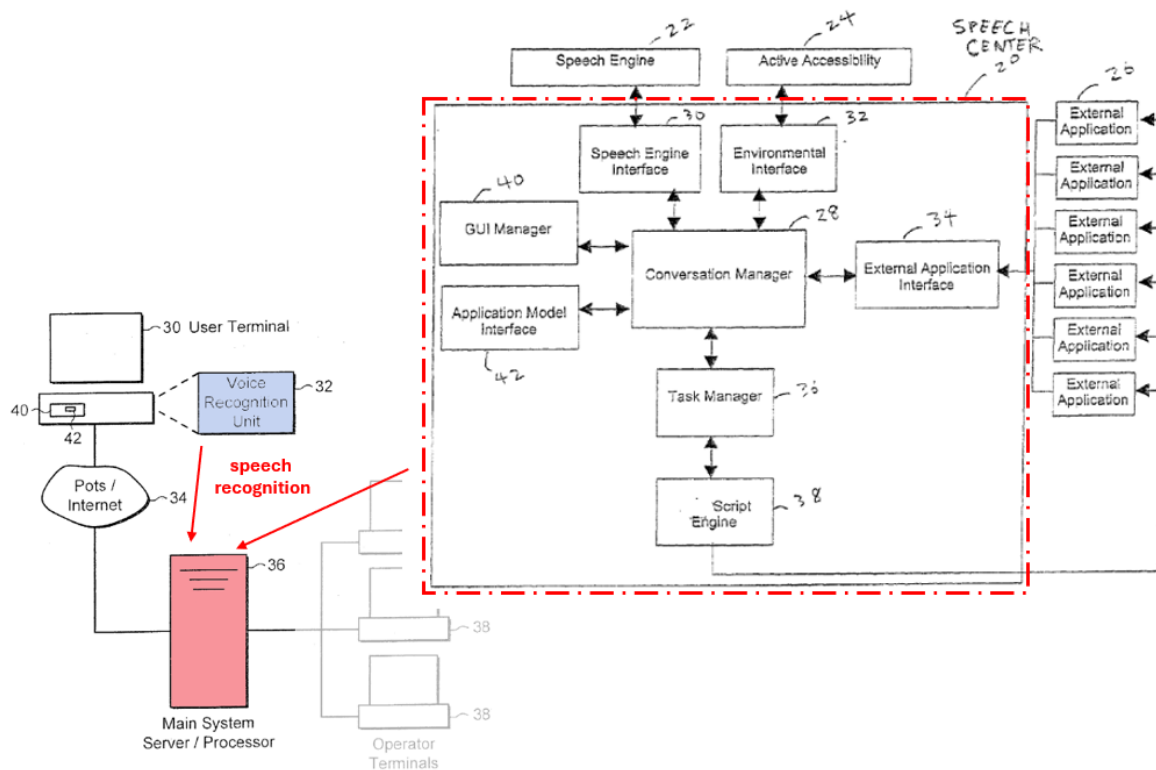
[25B] a **conversational speech engine**, wherein the conversational speech engine includes one or more processors configured to:

309. The combination of Barbara and Ross discloses “*a conversational speech engine*” recited in limitations [25B] and [1D.1]/[13D].

310. As I discussed in §IV.B.3, the ’681 patent discloses a conversational speech engine formed of a set of modules processing output from a speech recognition component and generating a response.

311. The system of Barbara-Ross has a similar structure. In the combined system, 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; *see also* EX-1007, ¶90.) These functions are provided by conversation manager 28, as taught by Ross. (EX-1008, ¶¶22, 32.) Conversation manager 28 shown in Ross’s Figure 2 (below) is “a *conversational speech engine*.”



Barbara, Figure 6; Ross, Figure 2 (annotated)

312. As taught by both Barbara and Ross, the “voice input unit” (e.g., microphone and/or speech recognition engine) is a separate component from the “conversational speech engine.” In Barbara, the “voice input unit” (e.g., voice recognition unit 32) is a separate component from the “conversational speech engine,” as shown in Barbara’s Figure 6. Similarly, Ross teaches conversation

manager 28 interfaces with speech engine 22, which is equivalent to Barbara's voice recognition unit, through speech engine interface 30, as shown in Ross's Figure 2. (See EX-1008, ¶24.)

313. Ross discloses that 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; 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, Barbara-Ross combination discloses "*the conversational speech engine includes one or more processors*" [25B].

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

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

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

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

[1B.2]/[25C.2] wherein the short-term shared knowledge includes knowledge about the utterance received during the current conversation

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

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

314. Barbara discloses “*accumulating short-term shared knowledge about the current conversation, wherein the short-term shared knowledge includes knowledge about the utterance received during the current conversation*” [1B.1]-[1B.2]/[13B.1]-[13B.2]/[25C.1]-[25C.2].

315. In Barbara, “[t]he best guesses of the voice recognition engine as to what was said are evaluated” in the conversational speech processing component of the server (“*conversational speech engine*” “against an information database(s)” and a “knowledge base.”) (EX-1007, ¶¶88, 92-93.) 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, ¶188.) Barbara also discloses that “[t]he correct utterance text and the correct intention are stored in the knowledge base.” (EX-1007, ¶106.) Barbara discloses an embodiment where both the information database(s) and the knowledge base are included in the main server. (EX-1007, ¶89, Figures 6-7.) 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.)

316. The combination of Barbara and Ross integrates Ross’s teaching regarding “conversational record” with Barbara’s “information database” teachings. Ross explains that “[c]onversational speech is full of implicit and explicit references back to people and objects that were mentioned earlier. To understand these sentences, the speech center system 20 looks at the conversational record 60, and finds the missing information.” (EX-1008, ¶57.) Ross achieves this benefit because “[e]ach utterance is indexed in the conversational record 60, along with the results of its semantic analysis.” (EX-1008, ¶57.)

317. 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.) Ross further discloses that “[t]he information [in the conversational record 60] 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.)

318. Accordingly, in the Barbara-Ross combination, the information database uses Ross’s conversational record that stores (“*accumulat[es]*”) each utterance “along with the results of its semantic analysis.” (EX-1008, ¶57.) 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].

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

[1C.1] accumulating long-term shared knowledge about the user,
[13C.1]/[25D.1] accumulate long-term shared knowledge about the user,
[1C.2]/[13C.2]/[25D.2] wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user;

319. Barbara discloses “*accumulating long-term shared knowledge about the user, wherein the long-term shared knowledge includes knowledge about one or more past conversations with the user*” [1C]/[13C]/[25D].

320. As I discussed in §V.B.4, in Barbara, the “best guesses of the voice recognition engine as to what was said are evaluated” in the conversational speech processing component of the server (“*conversational speech engine*” “against an **information database(s)**.”) (EX-1007, ¶88, ¶98 (“contextual information . . . is captured as the user interacts with the system”).) 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.) That is, Barbara records (“*accumulates*”) “**personal details or preferences from what has been said**” in the “information database(s).” (EX-1007, ¶88.)

321. For example, Barbara’s system infers that a speaker “who asks about their dog ten times a day” is a dog owner, e.g., a type of a 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 that “[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.*” Because this content of the information database “is built up over time”, the accumulated information is “*long-term shared*

knowledge.” (EX-1007, ¶88; EX-1001, 5:10-11 (“Long-term shared knowledge may **include explicit and/or implicit user preferences...**”).)

322. As another example, Barbara discloses that “[t]he database has flea vaccination information for Toto but not for the others, so the score for the canine interpretation may be raised. In this way the system ‘knows’ about the real world and can resolve ambiguities in many difficult sentences.” (EX-1007, ¶148.) This user information accumulated explicitly and/or implicitly corresponds to long-term knowledge. (See EX-1001, 5:10-14 (“Long-term shared knowledge may **include explicit and/or implicit user preferences**, a history of recent contexts, requests, tasks, etc., user-specific jargon related to vocabularies and/or capabilities of a context, most often used word choices, or other information.”).)

6. Determining An Intended Meaning [1D]

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

323. Barbara discloses “*determining an intended meaning for the utterance*” [1D].

324. Barbara determines an intended meaning for the utterance. For example, Barbara discloses that “[t]he system can however additionally or alternatively be used to allow text that is spoken into voice recognition equipment to be corrected or to create a voice interface where the end system responds to spoken natural language commands. In the case of voice recognition the electronic version

of the spoken word is checked for accuracy, corrected and then returned to the user. In the case of the voice interface, however, the electronic version of the spoken word is checked for accuracy, corrected and **interpreted to determine what the user wanted, i.e. the intent of the spoken word.**” (EX-1007, ¶82.) Barbara discloses that, “[o]nce the intent is determined, a command signal can be sent to the user’s terminal to carry out that intent.” (EX-1007, ¶82.) In addition, “[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, ¶82.) Thus, Barbara alone discloses “*determining an intended meaning for the utterance*” [1D].

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

[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

[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;

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

325. Barbara discloses “*identify[ing] . . . a context associated with the utterance*” during processing.

326. Barbara discloses that “[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.) Barbara further discloses that “[a]mbiguities are resolved when the system evaluates the likelihood that the user really is asking each of those things. For instance, 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. **Contextual information such as this is captured as the user interacts with the system, in much the same way as interactions work in real life.**” (EX-1007, ¶98.)

327. Using the same example discussed in §§V.B.2, V.B.5, above, “[i]f the user has never referred to their dog before, they are likely to introduce it in some way when first using in conversation. When they do not, and the system responds incorrectly or not at all, then the user may clarify their initial request and then the contextual information may be captured. A user who asks about their dog ten times a day might be considered very likely to be referring to it. Someone who is not involved in financial markets is unlikely to be asking about the Nikeii. The history of what people have asked is heuristically used to assist in scoring the possible intents.” (EX-1007, ¶98.)

328. In the particular example, when the user introduces their dog in the conversation, the “utterance is indexed in the conversational record 60, along with the results of its semantic analysis” (EX-1008, ¶57) (e.g., that Toto is the user’s dog) and then if within the bounds of the same conversation “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 or Tokyo the city, which the VR engine may have suggested as a possible hearing.” (EX-1007, ¶148.)

329. Accordingly, the Barbara-Ross system can identify a context associated with the utterance from the “*short-term shared knowledge*.”

330. Barbara also discloses identifying context based on person details of the user (“*long-term shared knowledge*”). Barbara explains that the information database (“*long-term shared knowledge*”) includes “**personal details**” about the user and “allow[s] users to **verbally enter personal details, such as the names of their children** or their favorite sport or their mother’s telephone number etc.” (EX-1007, ¶188.)

331. Further, Barbara explains that the information database includes personal information, such as a pets name, explaining that “[a] user who asks about their dog ten times a day might be considered very likely to be referring to it.” (EX-1007, ¶198) Therefore, the information database, e.g., the “*long-term knowledge*” in the Barbara-Ross system, would include the personal information that the user has a dog named “Toto” either by the frequent reference to it by the user or by allowing the user to enter their dog’s name. Therefore in connection with a user who interfaces with the system multiple times throughout the day about a dog named “Toto,” and inference can be drawn that the particular user has a dog named “Toto.” When the utterance includes a word is interpreted as “Toto,” the corresponding context may

be recognized as “pets.” Accordingly, the Barbara-Ross system can identify a context associated with the utterance from the “*long-term shared knowledge*.”

332. In summary, Barbara teaches context (pets versus restaurant) is identified from the recent dialog history, for example, dialog that includes a reference associated with the context-domain pets (e.g., “flea vaccination”) (“*short-term shared knowledge*”). Barbara also teaches context is identified from user profile information accumulated from prior conversations (e.g., Toto determined to mean dog because the user profile indicates that the particular user has a dog named Toto, information which could have been either derived from prior conversations or as an explicit input from the user).

333. Thus, the Barbara-Ross combination discloses “*identifying*” “*a context associated with the utterance from the short-term shared knowledge and the long-term shared knowledge.*”

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

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

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

[1D.3]/[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

[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

334. The combination of Barbara and Ross discloses “*establishing the intended meaning within the identified context, wherein the conversational speech engine establishes the intended meaning within the identified context to disambiguate an intent that the user had in speaking the one or more words that have the different meanings in different contexts*” [1D.2]-[1D.3]/[13E]-[13F]/[25F].

335. Barbara discloses that “the electronic version of the spoken word is checked for accuracy, corrected and interpreted to determine what the user wanted, i.e. the intent of the spoken word.” (EX-1007, ¶82.) Specifically, “[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. A combined likelihood for each possible request is heuristically calculated.” (EX-1007, ¶90.) A combined likelihood is based on the following:

the likelihood from the voice recognition engine that the text was actually what was said

the likelihood from the service’s knowledge base that the request was what was meant given what was said

an estimation from the user’s information base of which intents are likely.

(EX-1007, ¶¶90-93.)

336. “**Ambiguities are resolved** when the system evaluates the likelihood that the user really is asking each of those things. For instance, 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. **Contextual information** such as this is captured as the user interacts with the system, in much the same way as interactions work in real life. If the user has never referred to their dog before, they are likely to introduce it in some way when first using in conversation. When they do not, and the system responds incorrectly or not at all, then the user may clarify their initial request and then the contextual information may be captured. A user who asks about their dog ten times a day might be considered very likely to be referring to it. Someone who is not involved in financial markets is unlikely to be asking about the Nikeii. The history of what people have asked is heuristically used to assist in scoring the possible intents.” (EX-1007, ¶98.) That is, Barbara selects one of multiple potential meanings based on the context and in doing so, “[a]mbiguities are resolved” in Barbara’s system. (EX-1007, ¶98.)

337. Specifically, Barbara discloses that “[t]he information database may be used to help in determining which of the interpretations is correct. For instance, when asked to display something, the correct query is much more likely to have some values than to have none. If the user asks for the date of Toto’s last flea vaccination,” Barbara’s system may use the identified context domain of pet or dog to determine

that “it is much more likely that they are referring to Toto the dog, rather than Toto the restaurant or Tokyo the city, which the VR engine may have suggested as a possible hearing. The database has flea vaccination information for Toto but not for the others, so the score for the canine interpretation may be raised. In this way the system ‘knows’ about the real world and can resolve ambiguities in many difficult sentences.” (EX-1007, ¶148.) Accordingly, the user’s “*intended meaning*,” e.g., Toto the dog versus Toto the restaurant, is “*disambiguated*” within the identified context domain, e.g., “pet” versus “dining.”

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

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

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

[1E.2]/[13G.2] wherein the conversational speech engine grammatically or syntactically adapts the response based on the intended meaning established within the identified context.

[25G] generate a grammatically or syntactically adapted response to the utterance based on the intended meaning established within the identified context.

338. The combination of Barbara and Ross discloses “*generating a response to the utterance, wherein the conversational speech engine grammatically or syntactically adapts the response based on the intended meaning established within the identified context*” [1E.1]-[1E.2]/[13G.1]-[13G.2]/[25G].

339. Barbara discloses that after the intended meaning has been established, “[t]he client system now **performs the correct intention. The exact way in which**

the system looks and responds depends on the nature of the system being operated.” (EX-1007, ¶107.) Barbara discloses various ways to generate a response, including “feedback [] given by the application visibly responding” or “the appliance [] either speak[ing] back or perform[ing] some action (e.g. start the spin cycle) or both. Here the response could be a series of phonemes to be spoken and a command as to the state to enter (start rinse, cook toast, start heating).” (EX-1007, ¶107; *see also* EX-1007, ¶¶119-162 (describing examples related to an email application).)

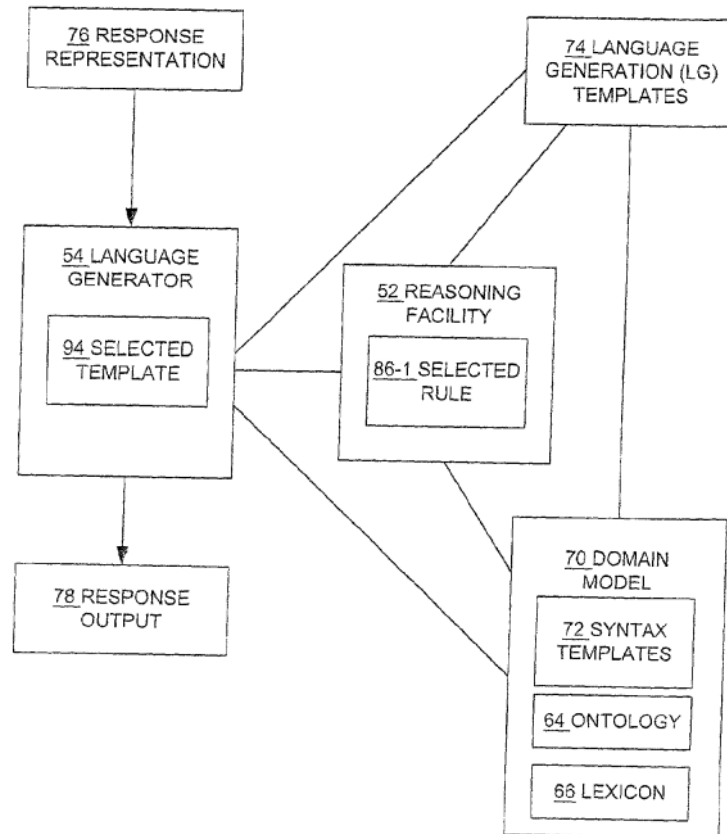
340. Ross also “provides a language generation method that performs its work in the context of a domain model for a particular application. A domain model consists of several types of information.” (EX-1008, ¶5.) Ross also discloses that “[w]ith the addition of syntax templates expressed in terms of the ontology definitions, a grammar can be automatically generated for the domain, and output questions and responses in the domain can also be generated.” (EX-1008, ¶5.)

341. Ross’s “language generation (LG) module uses **syntax templates ... to generate questions and responses to the user.** The language generation module **uses rules to select which syntax templates to use for a given goal or propositions** (goals and propositions are the formal belief structures manipulated by the reasoning component of the conversational system).” (EX-1008, ¶6.)

342. In Ross, “[t]he response output [] is a natural language response (e.g., text string), such as a statement or question, generated by the language

generation module []." (EX-1008, ¶61.) Ross also discloses that "the language generation module [] generates language from the same syntax templates [] used for recognition, or from additional templates provided specifically for language generation. These additional templates are the language generation (LG) templates []." (EX-1008, ¶62.)

343. Specifically, Ross discloses that "[b]ased on the goal provided by the language generation module 54, the reasoning facility 52 selects rule 86-1." (EX-1008, ¶69.) "When the above rule is invoked, the rule selects the response syntax template 94 (from the LG syntax templates 74)," (EX-1008, ¶70), as shown in Figure 4 (below).



Ross, Figure 4

344. Ross further explains that “the LG syntax templates [] are defined at the top level for speech center-generated questions and assertions ... These LG templates [] can then reference new or existing (i.e. background or foreground) templates [] in the domain model [], where the majority of information about syntactic forms in the speech center [] is represented. The special LG templates [] are defined for the language generation module [] for two reasons. One reason is to avoid having computer-generated questions and responses appear in the

user input grammars. Another reason is to control the argument structure to pass arguments as needed.” (EX-1008, ¶76.)

345. Accordingly, the Barbara-Ross combination “*generates a response to the utterance ... based on the intended meaning established within the identified context*” [1E.1/1E.2]/[13G.1/13G.2]/[25G]. Similarly, Barbara discloses that “the system could be adapted to interpret facts in what are called expert systems.” (EX-1007, ¶166.) Barbara notes that “[e]xpert systems attempt to distil the knowledge of an expert, such as a doctor or car mechanic, and produce a (large) set of rules so that the computer can do the job of the expert. For instance, an expert system might have rules like ‘if you have a cold and a fever you probably have flu’ in the medical domain. Or it may say things like ‘if your car doesn’t make any noise when you turn the key, and the radio also doesn’t work, then your battery is probably flat.’” (EX-1007, ¶167.)

346. As Barbara explains, when acting as a “medical expert system” the system works by asking the user questions, **and continuing to ask relevant questions, narrowing down the diagnosis** until a conclusion of satisfactory confidence is reached ... To do this, standard questions are entered into the system database, together with likely answers.” (EX-1007, ¶168.) By asking **relevant questions** to narrow down the diagnosis, the Barbara-Ross system’s “*conversational speech engine*” “*grammatically adapt the response based on the*

intended meaning established within the identified context” to perform[] the correct intention [1E.2]/[13G.2]/[25G]. (EX-1007, ¶167; *see also* EX-1001, 5:61-65.)

347. For example, following up in Barbara’s example, when the user wanted to engage a veterinarian expert system to ask questions about “Toto’s” medical condition, Barbara’s expert system would recognize that Toto refers to the user’s dog, and would engage in a conversation with the user formulating the questions to narrow the diagnosis.

348. Additionally, “[b]ased on the domain, the system can chose how strict to be in allowing differing responses to the questions. The more freedom that the user is allowed in responding to a question, the slower the system will learn. Against this, the field of medicine may require a lot more flexibility than computer technical support, given that the consequences of misdiagnosis are so much higher. An ideal application for this would be in computer technical support.” (EX-1007, ¶169.)

C. Claims 5, 17, and 29

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

[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

[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

[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.

349. The Barbara-Ross combination discloses “*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*” [5A]/[17A]/[29A].

350. The '681 Patent discloses that “[c]onversational goals may broadly include: (1) getting a discrete piece of information or performing a discrete task ... Participant roles may broadly include: (1) a leader that controls a conversation, (2) a supporter that follows the leader and provides input as requested, and/or (3) a consumer that uses information.” (EX-1001, 14:39-47.) The '681 Patent also discloses that “[i]nformation may be held by one or more of the participants at the outset of a conversation, where a participant may hold most (or all) of the information” (EX-1001, 14:47-51.)

351. Barbara identifies “*a conversational goal*,” e.g., a discrete task to be performed, 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.)

352. Additionally, “[t]he phrase ‘messages’ is taught as a restriction to those objects with ObjectType Message ... This may be represented as WHERE(ObjectType=‘Message’)” and “[t]he phrase ‘incoming’ is taught as a restriction to those messages sent to the current user. This may be represented as WHERE(Recipient=@@CURRENTUSER).” (EX-1007, ¶¶123-124.)

353. Barbara also identifies “*roles associated with the user and one or more other participants in the current conversation*.” For example, Barbara discloses that the user is a recipient, e.g., a consumer of information. (See e.g., EX-1007, ¶124 (“WHERE(Recipient=@@CURRENTUSER)”).) A POSITA would have understood that Barbara identifies the voice interface system as a supporter that provides input as requested, e.g., the incoming messages, to the user, because Barbara’s system understood the current conversation as a query to display information to the user. (See EX-1007, ¶122 (“The phrase ‘Show me’ is taught as a **display query**...”).)

354. Barbara also identifies “*an information allocation among the user and the one or more other participants in the current conversation.*” As I discussed above, in Barbara’s system, “[t]he phrase ‘Show me’ is taught as a display query...” for “those objects with ObjectType Message.” (EX-1007, ¶¶122-23.) Accordingly, a POSITA would have understood that Barbara’s system would hold all of the requested information (ObjectType=‘Message’), because the user’s utterance was translated as a display query of the objects with type “Message.”

355. As another example, in connection with Barbara’s disclosure of expert systems, Barbara discloses “*identifying a conversational goal associated with the utterance*” e.g., get a medical diagnosis, “*roles associated with the user and one or more other participants in the current conversation,*” e.g., the system is a dispenser of information and the user is a consumer of information, “*and an information allocation among the user and the one or more other participants in the current conversation,*” e.g., the system has more information while the user has less information. (EX-1007, ¶¶167-68; EX-1001, 15:21-26.)

356. Ross similarly discloses identifying a conversational goal. For example, Ross discloses that “the speech center 20 can choose which functions to call in order to achieve its goals.” (EX-1008, ¶46.) Ross provides an example where “the user might ask the speech center system 20 to ‘Create an appointment with Mark tomorrow.’ Searching through its available rules the speech center 20 finds one that

states that it can create an appointment.” (EX-1008, ¶46.) Ross explains that “[e]xamining the rule description, the speech center 20 finds that it calls a function which has the following parameters: a person, date, time, and place. The speech center 20 then sets up goals to fill in these parameters, based on the information already available.” (EX-1008, ¶46.)

357. In the particular example, Ross discloses that “[t]he goal of finding the date will result in the location of another rule which invokes a function that can calculate a date based on the relative date ‘tomorrow’ information. The goal of finding a person results in the location of a rule that will invoke a function which will attempt to disambiguate a person’s full name from their first name. The goal of finding the time will not be satisfiable by any rules that the speech center 20 knows about, and so a question to the user will be generated to get the information needed.” And “[o]nce all the required information is assembled, the appointment creation function is called and the appointment scheduled.” (EX-1008, ¶46.)

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

[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.

[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.

[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.

358. The Barbara-Ross combination discloses “*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*” and “*establish[ing] the intended meaning based on the conversation type*” [5B]/[17B]/[29B].

359. The '681 Patent discloses that “a query conversation may include a conversational goal of getting a discrete piece of information or performing a particular task.” (EX-1001, 14:61-63.) Accordingly, one instance of a “*conversation type* is a “query.” (See also EX-1001, 15:14-34, disclosing three examples of “*conversation types*,” namely a “query,” “didactic,” and “exploratory,” “*conversation types*.”)

360. Barbara discloses that an utterance or the current conversation is classified into a “*conversation type*,” e.g., query, based on an “*identified conversational goal*,” e.g., the goal of performing a particular task. For example, Barbara discloses 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.)

361. Barbara also discloses that the "*conversational speech engine further establishes the intended meaning based on the conversation type.*" For example, when Barbara's system identifies that "[t]he phrase 'Show me' is taught as a display query," the system understands that the user intention was to display "the sub-phrase which comprises the rest of the sentence (ie "Show me x">SELECT(x))." (EX-1007, ¶122.) That is, when the system identifies that the user utterance is a display query, it understands that the user's intention was to display objects, for example, that are of type "Message." (EX-1007, ¶123.)

362. As another example, in connection with Barbara's disclosure of expert systems, Barbara discloses a conversational type that corresponds to a "didactic" conversational type. As the '681 Patent explains, "[i]n a didactic conversation, a leader of the conversation [system] may control information desired by a supporter of the conversation [user]. The supporter's role may be limited to regulating an overall progression of the conversation and interjecting queries for clarification." (EX-1001, 14:67-15:4.) Similarly, Barbara discloses that in a conversation with an expert system, e.g., for a medical diagnosis, a leader of the conversation [expert system] may control the information, e.g., mapping symptoms to potential diseases

or inflictions and making the diagnosis, desired by a supporter of the conversation [user].

D. Claims 6, 18, and 30

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.

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.

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.

363. Barbara discloses that “*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].

364. Barbara discloses that “[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. A **combined likelihood** for each possible request is heuristically calculated.” (EX-1007, ¶90.) A combined likelihood is based on the following:

the likelihood from the voice recognition engine that the text was actually what was said

the likelihood from the service’s knowledge base that the request was what was meant given what was said

an estimation from the user's information base of which intents are likely.

(EX-1007, ¶¶90-93.)

365. Barbara further discloses the following example:

[S]uppose the user says 'How's Tokyo doing?'. The voice recognition engine might have detected this and have outputted the following two possible recognitions, with roughly equal probabilities:

How's Tokyo doing?

How's Toto doing? (Toto could be the user's dog)

(EX-1007, ¶¶94-96.)

366. So in the particular example, Barbara's system hypothesizes that the user uttered either "How's Tokyo doing?" or "How's Toto doing?"

367. Barbara explains that "[t]he knowledge base might interpret 'How's Tokyo doing' as a request for the status of the Tokyo production line, the weather in Tokyo or where the Nikei is trading. These might be, hypothetically, roughly equally likely. The interpretation of 'How's Toto doing' might be strongly recognised as a request for the status of the operator's pet dog Toto. **The single utterance was expanded to two possibilities for text**, which were expanded to four possible intentions." (EX-1007, ¶97.) For example, the possible *intended meaning* "where the Nikei is trading" includes a hypothesis that the user intended to utter the phrase "How's Tokyo doing?"

368. Barbara further discloses that “a set of possible interpretations for the utterance are scored for likelihood. Once this is done, 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.)

369. Accordingly, Barbara discloses a hypothesis, e.g., what the voice recognition engine has detected as an utterance (“How’s Toto doing?/How’s Tokyo doing?”), and assigns a degree of certainty, e.g., combined likelihood score that is associated with a probability that the interpretation of the utterance is correct.

E. Claims 7, 19, and 31

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.

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.

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.

370. Barbara 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].

371. As I explained in §V.D, Barbara discloses the following example:

[T]he user says “How’s Tokyo doing?”. The voice recognition engine might have detected this and have outputted the following two possible recognitions, with roughly equal probabilities:

How’s Tokyo doing?

How’s Toto doing? (Toto could be the user’s dog)

(EX-1007, ¶¶94-96.)

372. Barbara explains that “[t]he knowledge base might interpret ‘How’s Tokyo doing’ as a request for the status of the Tokyo production line, the weather in Tokyo or where the Nikei is trading. These might be, hypothetically, roughly equally likely. The interpretation of ‘How’s Toto doing’ might be strongly recognised as a request for the status of the operator’s pet dog Toto. The single

utterance was expanded to two possibilities for text, **which were expanded to four possible intentions.**” (EX-1007, ¶97.)

373. These four “possible interpretations for the utterance are scored for likelihood.” (EX-1007, ¶99.) Barbara explains that “[a] **combined likelihood** for each possible request is heuristically calculated.” (EX-1007, ¶90.) A combined likelihood is based on the following:

the likelihood from the voice recognition engine that the text was actually what was said

the likelihood from the service’s knowledge base that the request was what was meant given what was said

an estimation from the user’s information base of which intents are likely.

(EX-1007, ¶¶90-93.)

374. Barbara explains that “a set of possible interpretations for the utterance are **scored for likelihood**. Once this is done, 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.) That is, the calculated “combined likelihood” corresponds to a “*degree of certainty*,” because it can provide a probability of whether an interpretation is correct.

375. Accordingly, Barbara discloses that the “*degree of certainty*,” e.g., combined likelihood, is based on “*a confidence level associated with the preliminary interpretation generated at the speech recognition engine*,” e.g., “the likelihood

from the voice recognition engine that the text was actually what was said.” (EX-1007, ¶91.)

376. Barbara also discloses that the *degree of certainty*, e.g., combined likelihood, is based on “*information associated with the identified context*.” For example, Barbara discloses that “[a]mbiguities are resolved **when the system evaluates the likelihood** that the user really is asking each of those things. For instance, 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.” (EX-1007, ¶98.)

377. Barbara further explains that “[c]ontextual information ... is captured as the user interacts with the system, in much the same way as interactions work in real life. If the user has never referred to their dog before, they are likely to introduce it in some way when first using in conversation. When they do not, and the system responds incorrectly or not at all, then the user may clarify their initial request and then the contextual information may be captured. A user who asks about their dog ten times a day might be considered very likely to be referring to it. Someone who is not involved in financial markets is unlikely to be asking about the Nikeii. The history of what people have asked is heuristically used to assist in scoring the possible intents.” (EX-1007, ¶98.)

F. Claims 8, 20, and 32

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

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.

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.

378. Barbara discloses “*grammatically or syntactically adapts the response based on the conversation type*” [8]/[20]/[32].

379. Barbara provides an example where “a user asks their phone if they have any e-mail.” (EX-1007, ¶110.) Barbara’s system identifies that the utterance can be classified as an “abstract intention of ‘**query** for message objects that have not yet been read’.” (EX-1007, ¶110.) Barbara further discloses that “the system may reply with a brief summary of the new messages” and that the “response[] would be generated by the abstract intention of ‘**query** for message objects that have not yet been read’.” (EX-1007, ¶110.)

380. By summarizing the messages, Barbara grammatically adapts the response based on the “*conversation type*,” e.g., “**query** for message objects that have not yet been read.”

381. Additionally, Barbara discloses that “the system could be adapted to interpret facts in what are called expert systems.” (EX-1007, ¶166.) Barbara notes

that “[e]xpert systems attempt to distil the knowledge of an expert, such as a doctor or car mechanic, and produce a (large) set of rules so that the computer can do the job of the expert. For instance, an expert system might have rules like ‘if you have a cold and a fever you probably have flu’ in the medical domain. Or it may say things like ‘if your car doesn’t make any noise when you turn the key, and the radio also doesn’t work, then your battery is probably flat’.” (EX-1007, ¶167.)

382. As Barbara explains, “[a] typical expert system works by asking the user questions, and **continuing to ask relevant questions, narrowing down the diagnosis** until a conclusion of satisfactory confidence is reached. This is very simple to map into the framework of the system described above. To do this, standard questions are entered into the system database, together with likely answers.” (EX-1007, ¶168.) A POSITA would have understood that by asking relevant questions to narrow down the diagnosis, Barbara’s system would grammatically adapt the response based on the *conversation type*, e.g., query type.

383. Additionally, Ross discloses that “language generation (LG) module **uses syntax templates ... to generate questions** and responses to the user. The language generation module uses rules to select which syntax templates to use for a given goal or propositions (goals and propositions are the formal belief structures manipulated by the reasoning component of the conversational system).” (EX-1008,

¶6.) Accordingly, the Barbara-Ross system would adapt the response to use syntax templates, e.g., based on the “*conversation type*,” e.g., query type.

G. Claims 11-12, 23-24, and 35-36

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.

24. 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.

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.

384. Barbara discloses “*the user speaks the utterance in a multi-modal input that further includes one or more non-voice inputs relating to the utterance*”

[11]/[24]/[35] and “*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*” [12]/[25]/[36].

385. 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..

386. 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. 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. That is, the spoken “utterance” relates to the “non-voice input” (OCR-scan). 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”).)

387. Therefore, Barbara teaches or at least suggests claims [11]/[23]/[35].

388. 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.”

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].

VI. GROUND 3: The Combination Of Barbara, Ross, And O’Neill Renders Obvious Claims 2-4, 9, 14-16, 21, 26-28, And 33

A. Overview of the Combination

390. U.S. Patent Application Publication 2005/0278180 to O’Neill et al. (“O’Neill”; EX-1009), which published December 15, 2005 discloses or at least suggests every limitation of claims 9, 21, and 33.

1. O’Neill

391. O’Neill discloses a “system for conducting a dialogue with a user.” (EX-1009, Abstract.) O’Neill’s system “receives a communication from a user, assigns the communication to a processing unit according to the communication’s semantic content ... the assigned processing unit processes the communication in accordance with the processing unit’s processing rules and provides a response to the user.” (EX-1009, Abstract.)

392. Specifically, O’Neill discloses a “DiscourseManager ... responsible for the improved dialogue manager’s 20 overall ‘conversational style’, (i.e. the generic discourse behavior of the improved dialogue manager 20).” (EX-1009, ¶110.) O’Neill explains that “[g]iven the current technical difficulties with speech recognition and the possibility that a user will misunderstand or change their mind

during a dialogue, any system conducting a complex transaction must have a strategy for confirming the semantic contents of a user's utterances.” (EX-1009, ¶110.)

393. O'Neill's DiscourseManager implements such strategy. For example, the DiscourseManager “determines the response of the improved dialogue manager 20 to new, modified or negated information from the user. In the case of new or modified information, the DiscourseManager 22 ensures that the information is at least implicitly confirmed. In an implicit confirmation strategy a recognized answer in a previous utterance is embedded in the system's next question (e.g. ‘So, you're arriving at the Hilton in Belfast on Friday’).” (EX-1009, ¶111.)

394. “The DiscourseManager 22 further ensures that the receipt of a confirmatory response by the user is immediately followed by a question requesting further information (e.g. ‘what day are you departing?’). However, it will of course be realized that towards the end of a dialogue session, the system may have all the information it requires to conclude the transaction. Consequently, in this case, the receipt of a confirmatory response is not followed by another question for further information.” (EX-1009, ¶112.)

395. O'Neill further discloses a structure for “knowledge-storing components.” (EX-1009, ¶¶172-73.) Specifically, O'Neill discloses that “DiscourseManager 20 (i.e. the generic functional element) uses a DiscourseHistory data class 122 **to maintain a record of the evolving dialogue in the form of a stack**

(i.e. the DiscourseStack 123) of dialogue frames that can be added to or retrieved from the DiscourseStack 123 as required.” (EX-1009, ¶174.)

2. Motivation To Combine

396. 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.”

397. 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. 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.)

398. 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.) 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. 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.

399. 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).

400. 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.)

401. 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.

B. Claims 2, 14, and 26

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.

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.

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.

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

403. As I explained in §V.B.4, the Barbara-Ross combination holds the current conversation utterances along with the results of their semantic analysis [i.e., “*information about the utterance*”] and the captured contextual information in a

“conversational record,” as disclosed by Ross. However, Ross does not disclose implementation details of the conversational record.

404. O’Neill discloses that “DiscourseManager ... uses a DiscourseHistory data class ... to maintain **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.) A POSITA would have been motivated to implement the conversational record—which holds “*information about the utterance received during the current conversation*”—as a stack, such that each utterance is added to or retrieved from the stack as required. 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.”

405. As I explained in §V.A.3, a POSITA would have been motivated to implement Ross’s “conversational record” in the form of a stack of utterances, because that would allow utterances to be added to or retrieved from the conversational record.

C. Claims 3, 15, and 27

<p>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.</p>
--

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.

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.

406. The Barbara-Ross system includes short-term shared knowledge with “information [that] 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.) As I noted above §VI.B, it would have been obvious to include the short term shared knowledge in Barbara-Ross in a “*short term context stack*,” as taught by O’Neill.

407. The ’681 patent contains no information about what constitutes a “*psychologically appropriate amount of time*.” Specifically, the ’681 Patent discloses:

Moreover, because cooperative conversations model human conversations, **short-term session data may be expired after a psychologically appropriate amount of time**, thereby humanizing system behavior. For example, a human is unlikely to recall a context of a conversation from two years ago, but because the context would be identifiable by a machine, session context is expired after a predetermined amount of time **to reduce a likelihood of contextual confusion based on stale data**

(EX-1001, 4:65-5:2.)

408. A POSITA would have understood that the time it takes for information to be “no longer relevant to active goals” to satisfy this limitation, because the ’681 Patent associates “*psychologically appropriate amount of time*” with the time data becomes stale. (EX-1001, 13:65-66.)

409. Accordingly, Barbara-Ross-O’Neill discloses “*expir[ing]/[e] the information about the utterance from the short-term context stack after a psychologically appropriate amount of time*” [3]/[15]/[27].

D. Claims 4, 16, and 28

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.

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.

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.

410. Barbara-Ross-O'Neill discloses *“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”* [4]/[16]/[28].

411. 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.”

412. The “conversational record” includes “results of [an utterance's] semantic analysis,” which “is eventually purged from the conversational record.” (EX-1008, ¶57.) A POSITA would have understood, however, that if the “results of [an utterance's] semantic analysis,” relate to a user's “personal details or preferences” (EX-1007, ¶88) this information would be included in Barbara's “information database” even after it has been purged (e.g., expired) from the “conversational database,” (e.g., short term context stack), because Ross teaches that

the particular information is purged “when it is no longer relevant **to active goals**” (EX-1008, ¶57.) A POSITA would have understood that—at least some—results of a semantic analysis would generally be relevant to the user, e.g., if it identifies person details—like a pet’s name, even after the information is irrelevant to the user’s active goals. (See, e.g., EX-1007, ¶148.)

413. Accordingly, Barbara-Ross-O’Neill 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).

E. Claims 9, 21, and 33

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.

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.

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.

414. The Barbara, Ross and O’Neill combination discloses “*grammatically or syntactically adapt[ing] the response to influence a subsequent reply utterance*

that the conversational speech engine expects from the user during the current conversation” [9]/[21]/[33].

415. O’Neill’s system “receives a communication from a user, assigns the communication to a processing unit according to the communication’s semantic content ... the assigned processing unit processes the communication in accordance with the processing unit’s processing rules and provides a response to the user.” (EX-1009, Abstract.)

416. Specifically, O’Neill discloses a “DiscourseManager ... responsible for the improved dialogue manager’s 20 overall ‘conversational style’, (i.e. the generic discourse behavior of the improved dialogue manager 20).” (EX-1009, ¶110.) O’Neill explains that “[g]iven the current technical difficulties with speech recognition and the possibility that a user will misunderstand or change their mind during a dialogue, any system conducting a complex transaction must have a strategy for confirming the semantic contents of a user’s utterances.” (EX-1009, ¶110.)

417. O’Neill’s DiscourseManager implements such strategy. For example, the DiscourseManager “determines the response of the improved dialogue manager 20 to new, modified or negated information from the user. In the case of new or modified information, the DiscourseManager 22 ensures that the information is at least implicitly confirmed. In an implicit confirmation strategy a recognized answer

in a previous utterance is embedded in the system’s next question (e.g. ‘So, you’re arriving at the Hilton in Belfast on Friday’).” (EX-1009, ¶111.)

418. “The DiscourseManager 22 further ensures that the receipt of a confirmatory response by the user is immediately followed by a question requesting further information (e.g. ‘what day are you departing?’). However, it will of course be realized that towards the end of a dialogue session, the system may have all the information it requires to conclude the transaction. Consequently, in this case, the receipt of a confirmatory response is not followed by another question for further information.” (EX-1009, ¶112.)

419. By immediately following the user’s response with a question requesting further information, O’Neill syntactically adapts the response (by requiring further information) to influence a subsequent reply utterance that the conversational speech engine expects from the user during the current conversation (e.g., the user’s reply including the requested information.)

VII. GROUND 4: The Combination Of Barbara, Ross, and Franco Renders Obvious Claims 10, 22, And 34

A. Overview of the Combination

420. I have described both Barbara and Ross above, and I discuss Franco below.

1. Franco

421. U.S. Patent Application Publication 2010/0125458 to Franco et al. (“Franco”; EX-1010) “relates generally to speech recognition applications and relates more particularly to correcting errors in speech recognition systems.” (EX-1010, ¶2.)

422. Franco includes multiple embodiments for determining that a word or phrase has been incorrectly interpreted. In one embodiment (shown in Figure 1 below), an error is determined based on feedback from a user. (EX-1010, ¶¶13-15.) The user provides a “‘correction’ utterance” and the system generates a “combined hypothesis” on what is being requested. (EX-1010, ¶17.)

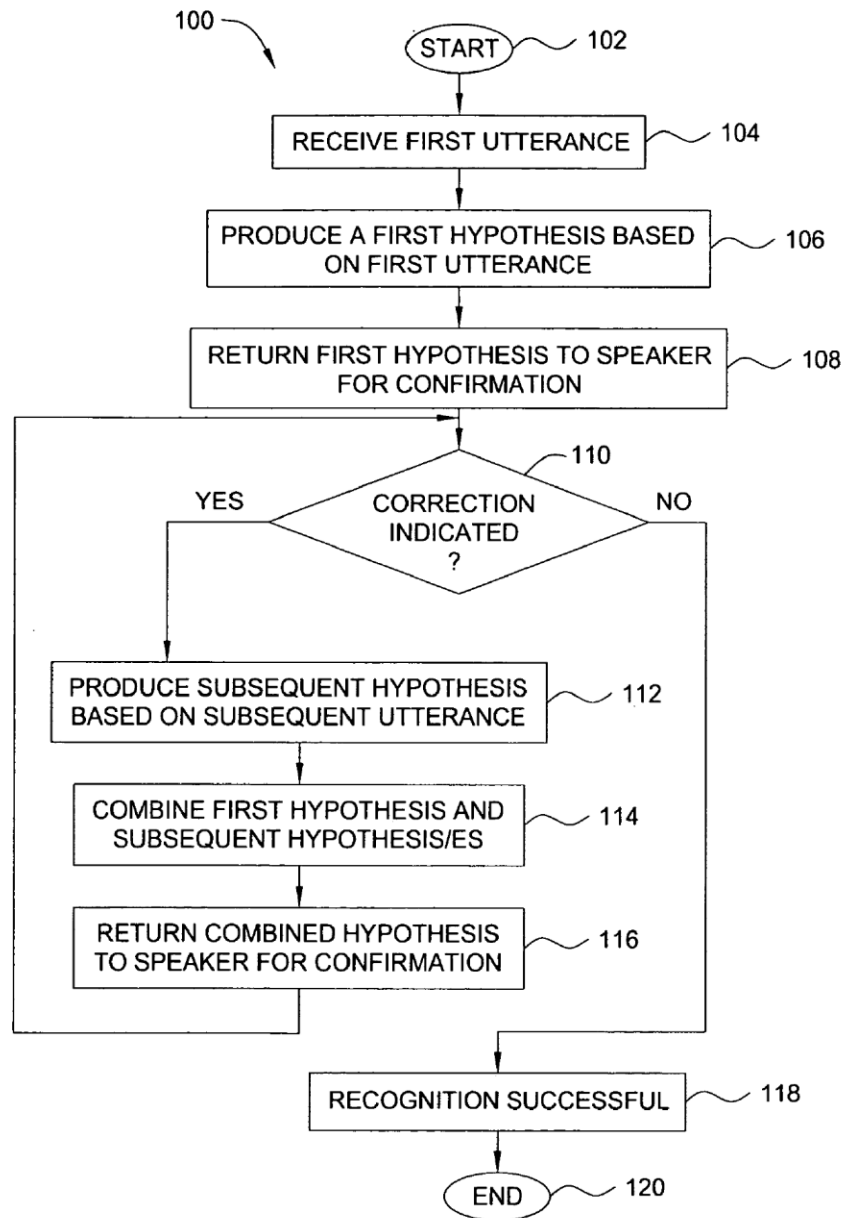


FIG. 1

Franco, Figure 1

423. Another embodiment relates to a scenario where a user offers an utterance, the utterance is incorrectly recognized, and the user repeats the same utterance without being prompted to do so by the system. (EX-1010, ¶18.) In such

scenario, the method combines hypothesized interpretations of the first and second (repeated) utterance to maximize the likelihood of correct interpretation. (EX-1010, ¶20 (“The method 100 thereby achieves improved error correction for speech recognition applications. By assuming that a subsequent speaker utterance will comprise all or part of a previously spoken (and incorrectly recognized) utterance, the acoustic evidence from the previous and subsequent utterances can be combined in an efficient manner to produce a more accurate corrected recognition.”).) For example, a user may state “[t]he number is 1234. I repeat, the number is 1234”. (EX-1010, ¶18.) “In this case, the method 100 would treat ‘The number is 1234’ as the first utterance and ‘I repeat, the number is 1234’ as the subsequent utterance, and combine respective first and subsequent hypotheses [] to produce the combined hypothesis” for recognizing the speech. (EX-1010, ¶18.)

2. Motivation To Combine

424. A POSITA would have been motivated to integrate Franco’s teachings of implementing error correction with voice interface in the Barbara-Ross combination (which explicitly discloses making corrections manually with the intervention of an operator) for numerous reasons.

425. First, Barbara recognizes the motivation to combine: “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-1007, ¶3.) A POSITA would understand Barbara expressing a motivation to combine the system disclosed in Barbara with additional systems, like Franco, that specialize in correcting errors in speech recognition systems.

426. Second, Franco also recognizes the motivation, stating that there is a “need in the art for a method and apparatus for error correction in speech recognition applications.” (EX-1010, ¶5.) A POSITA would understand that Franco is suggesting that the method of error correction disclosed in Franco was broadly applicable and could be implemented and combined with speech recognition applications, like Barbara’s.

427. Third, Barbara, Ross, and Franco are analogous art to the ’681 Patent. All three pertain to the same field of endeavor, i.e., speech recognition systems. (*Compare* EX-1001, 7:24-33 (“The utterance component of input 105 may be processed by a **speech recognition engine** 110 ... to generate one or more preliminary interpretations of the utterance. The **speech recognition engine** 110 may process the utterance using any suitable technique known in the art. For example, in one implementation, the speech recognition engine 110 may interpret the utterance using techniques of phonetic dictation to recognize a phoneme stream...”) *with* EX-1007, ¶7 (“The present invention relates to a system and method for improving accuracy of signal interpretation. In particular, the present invention relates to a system and method for improving text that is to be presented to a user,

said text collected ... using **voice recognition software**”), ¶12 (“[T]he signal may be derived from **voice recognition software** that is operable to record spoken words and convert those spoken words into electronic text”), ¶82 (“In the case of voice recognition the electronic version of the spoken word is checked for accuracy, corrected and then returned to the user.”) *and* EX-1008, Abstract (“[A] conversation manager [that] processes spoken utterances from a user of a computer, and develops responses to the spoken utterances.”) *and* EX-1010, ¶2 (“The present invention relates generally to speech recognition applications and relates more particularly to correcting errors in speech recognition systems.”).)

428. 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.) Barbara is implemented through a “user terminal 30 is connected via the internet or any other suitable communication system 34, for example a dedicated secure connection, to a main server(s) 36.” (EX-1007, ¶84.)

B. Claims 10, 22, 34

1. Multiple Preliminary Interpretations [10A/22A/34A]

[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,

[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,

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

[10A.2/22A.2/34A.2] wherein an initial interpretation of the utterance comprises one of the multiple preliminary interpretations having a highest confidence level; and

429. The Barbara-Ross-Franco combination discloses this limitation. Barbara discloses that “[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.) Barbara offers an example where “the user says ‘**How's Tokyo doing?**’”. The voice recognition engine might have detected this and have outputted the following two possible recognitions, with roughly equal probabilities:

How's Tokyo doing?

How's Toto doing? (Toto could be the user's dog)”

(EX-1007, ¶¶94-96.)

430. Barbara explains that “[t]he knowledge base might interpret ‘How's Tokyo doing’ as a request for the status of the Tokyo production line, the weather in Tokyo or where the Nikeii is trading. These might be, hypothetically, roughly equally likely. The interpretation of ‘How's Toto doing’ might be strongly

recognised as a request for the status of the operator's pet dog Toto. The single utterance was expanded to two possibilities for text, which were expanded to **four possible intentions.**” (EX-1007, ¶97.)

431. As I discussed in §V.B.2, Barbara teaches that voice recognition unit 32 is implemented within server 36, and is coupled to a “voice input device” (e.g., at user terminal 30) and the “*conversational speech engine*” which also is also implemented at the server.

432. Accordingly, Barbara discloses “*generating multiple preliminary interpretations (“four possible intentions”) of the utterance (“How's Tokyo doing?") at a speech recognition engine coupled to the voice input device and the conversational speech engine.*” [10A.1]/[22A.1]/[34A.1].

433. Barbara also discloses that “[a] **combined likelihood for each possible request** is heuristically calculated.” (EX-1007, ¶90; *see also* EX-1007, ¶99 (“[A] set of possible interpretations for the utterance are **scored for likelihood.**”)) That is, Barbara calculates for each of the “four possible intentions” a combined likelihood score. Barbara explains that once the interpretations are scored for likelihood, “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.) Therefore, a likelihood score correspond to a probability that the corresponding interpretation is correct, which corresponds to a “*confidence level*”

([10A.2/22A.2/34A.2]), in the sense that the higher the probability the interpretation is correct, the higher the “*confidence level*.” A POSITA would have understood that Barbara would select the interpretation with the highest likelihood score, among the four possible interpretations, as an “*initial interpretation of the utterance*,” because it would correspond to the highest probability it would be the user’s intention.

434. Accordingly, Barbara discloses “*an initial interpretation of the utterance comprises one of the multiple preliminary interpretations having a highest confidence level*.” [10A.2]/[22A.2]/[34A.2]

435. Franco likewise discloses this limitation. As noted above, Franco includes an embodiment directed to a scenario where the user offers an utterance, the utterance is incorrectly recognized, and the user repeats the same utterance without being prompted to do so by the system. (EX-1010, ¶18.) In another, the user makes the same utterance twice with a keyword (such as “repeat”) interspersed that triggers that a “need for correction” has occurred. (EX-1010, ¶15.) The system generates a “recognition hypothesis” based on “acoustic evidence” for each utterance. (EX-1010, ¶12.)

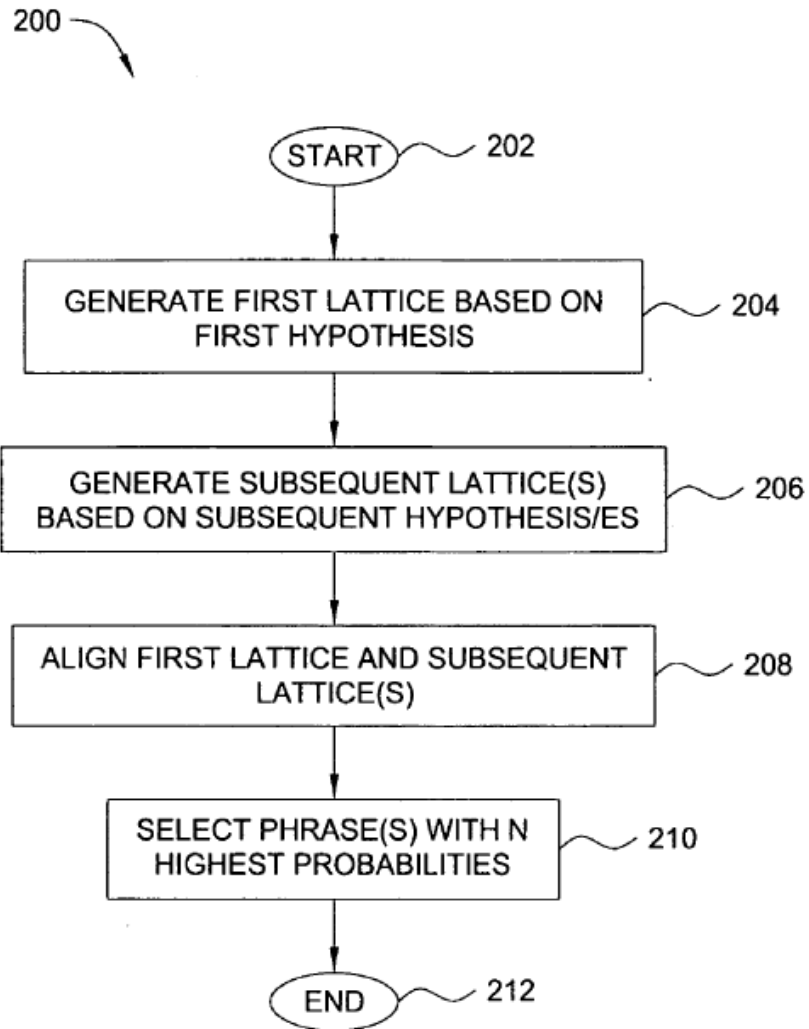
436. In Franco, the system:

generates a first word lattice based on the first hypothesis. The first word lattice is a network or diagram that graphically illustrates the first hypothesis. Specifically, the first lattice may be thought of as comprising a plurality of lattice positions or word “slots”, where each slot corresponds to a temporal interval in the first utterance. Each slot is associated with one or more words that may potentially have been uttered during the corresponding temporal interval. All of the slots in total comprise a phrase that may represent the first utterance (e.g., by selecting one word for each slot and then concatenating the selected words).

(EX-1010, ¶22.) The system also “generates a subsequent word lattice for the subsequent hypothesis.” (EX-1010, ¶23.)

437. Franco thus discloses “*generating multiple preliminary interpretations*” with each slot comprising a “*preliminary interpretation.*”

438. After creating the word lattices, the system next “aligns the first word lattice and the subsequent word lattice to form an aligned word lattice.” (EX-1010, ¶24.) “Each potential word is associated with a score that indicates the probability that the word was in fact, uttered ...” (EX-1010, ¶22.) “The probability of a phrase is the sum of the probabilities or scores of the individual words that comprise the phrase.” (EX-1010, ¶26.) The system “selects the phrase(s) with the N highest probabilities as the combined hypothesis/es to be presented to the speaker...” (EX-1010, ¶26.) This process is depicted in Figure 2 below.



Franco, Figure 2

439. Franco thus discloses an “*initial interpretation of the utterance comprises one of the multiple preliminary interpretations having a highest confidence level.*”

2. Update Short-Term Shared Knowledge

[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,

[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,

[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,

440. In connection to the example in the previous section, Barbara discloses that “[t]he interpretation of ‘How's Toto doing’ might be strongly recognised as a request for the status of the operator's pet dog Toto.” (EX-1007, ¶97.) Therefore, even though “the user sa[id] ‘How's Tokyo doing?’” (EX-1007, ¶94), Barbara’s system contemplates a possible interpretation about something unrelated, e.g., Toto, the user’s dog.

441. If Barbara’s “heuristic calculation” for “How’s Toto doing” resulted in a likelihood score that is higher than the likelihood scores of the other possible interpretations (because, for example, the user has had many prior interactions relating to their dog), then Barbara’s system would identify an incorrect interpretation of the user’s utterance, since the user said “How's Tokyo doing?”.

442. As I explained in §VII.A.2, a POSITA would have been motivated to combine the Barbara-Ross system with Franco to allow a user to correct a “first hypothesis (i.e., a proposed recognition solution) based on the first utterance.” (EX-1010, ¶14.)

443. Franco explains that in one embodiment, the method “returns the first hypothesis to the speaker (e.g., via a textual display or a speech synthesis interface) ... for review/confirmation.” (EX-1010, ¶14.) Then, Franco “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.” (EX-1010, ¶15.) Specifically, Franco explains that “[t]he subsequent utterance(s) 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.) Accordingly, when the user in Barbara’s example hears or sees the incorrect interpretation, they could repeat either “How’s Tokyo” or “Tokyo.”

444. A POSITA would have understood that the “conversational record” in the Barbara-Ross-Franco system (“*short-term shared knowledge*”) would hold the “*multiple preliminary interpretations of the utterance*” because the preliminary interpretations are results of semantic analysis of the utterance and Ross discloses that “[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-1010, ¶17), 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.”

445. 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*”

3. Determine Intended Meaning

[10C] wherein the conversational speech engine determines the intended meaning based on one of the multiple preliminary interpretations having a next highest confidence level.

[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.

[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.

446. As I discussed in §VII.B.2, the initial interpretation “How’s Toto doing” is removed from the short-term memory. Accordingly, it will no longer be a possible interpretation of the received utterance. Therefore, Barbara will use the likelihood scores (“*confidence levels*”) of the remaining possible interpretation to “differentiate between those interpretations that have a high probability of being correct and those that do not.” (EX-1007, ¶99.) Because a higher likelihood score would indicate a higher probability that the interpretation is correct, a POSITA would have understood that Barbara’s system would identify the possible interpretation having the likelihood score as the intended meaning.

447. Moreover, Franco discloses that “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.)

448. 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. Conclusion

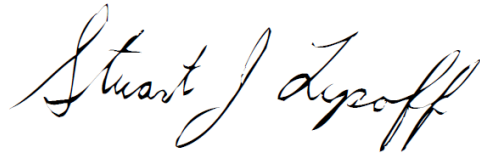
449. In signing this Declaration, I recognize that the Declaration will be filed as evidence in a contested case before the Patent Trial and Appeal Board of the United States Patent and Trademark Office. I also recognize that I may be subject to cross-examination in this proceeding. If required, I will appear for cross-examination at the appropriate time. I reserve the right to offer opinions relevant to the invalidity of the challenged claims at issue and/or offer testimony in support of this Declaration.

450. I hereby declare that all statements made herein of my own knowledge are true and that all statements are made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful

false statements and the like so made are punishable by fine or imprisonment, or both, under 18 U.S.C. § 1001.

Dated: April 25, 2025

Respectfully submitted,

A handwritten signature in black ink, reading "Stuart J. Lipoff". The signature is written in a cursive style with a large, stylized "S" and "L".

Stuart Lipoff

APPENDIX

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”)
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”)
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”)

Exh.	Reference
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 Syntax and Semantics</i> 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)
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)

Exh.	Reference
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