

**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

---

IPR2025-00870

U.S. Patent No. 10,755,699

---

**DECLARATION OF STUART LIPOFF IN SUPPORT OF PETITION FOR  
*INTER PARTES* REVIEW OF U.S. PATENT NO. 10,755,699**

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.....	3
II.	Understanding of Relevant Legal Principles.....	12
III.	Technical Background.....	15
	A. Language and Dialog .....	17
	B. Interpreting Meaning in Dialog.....	20
	1. Grice’s Cooperative Principle.....	20
	2. Grice’s Principle Applied to Human-Machine Interactions.....	23
	C. Interactive/Spoken Dialog Systems .....	24
	1. Speech Recognition Component.....	32
	2. Language Understanding .....	34
	3. Dialog Management Component .....	37
	4. Response Generator and Speech Synthesizer .....	45
IV.	GROUND 1: The Combination of SmartKom and Kobsa.....	47
	A. Overview of the Combination .....	47
	1. SmartKom .....	47
	2. Kobsa.....	59
	3. Motivation to Combine .....	61
	B. Independent Claims 1 and 12.....	64
	1. Preamble [1P] / Preamble [12P]/[12A].....	65
	2. Identifying a Context [1C]/[12D] .....	67
	3. Short-Term and Long-Term Knowledge Limitations.....	75
	4. Identifying a Manner [1G]/[12H] .....	93
V.	GROUND 2: The Combination Of Barbara, Ross And Kellner .....	100
	A. Overview of the Combination.....	100
	1. Barbara .....	100
	2. Ross .....	107

# TABLE OF CONTENTS

(continued)

Page ii

3.	Kellner.....	115
4.	Motivation To Combine Barbara With Ross .....	117
5.	Motivation To Combine Barbara And Ross With Kellner ....	119
B.	Independent Claims 1 and 12 .....	123
1.	Preamble [1P] / [12P] / [12A].....	124
2.	Identifying a Context [1C]/[12D] .....	126
3.	Accumulating Short-Term Knowledge [1E]/[12F] .....	129
4.	Accumulating Long-Term Knowledge [1F]/[12G] .....	132
5.	Identifying a Manner [1G]/[12H] .....	135
C.	Dependent Claims 4, 15 .....	137
1.	Contextual Signifiers And/Or Grammatical Rules [4A]/[15A] .....	138
2.	Response Based On Contextual Signifiers And/Or Grammatical Rules [4B]/[15B].....	139
D.	Dependent Claims 5, 16 .....	140
VI.	Conclusion .....	141

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 10,755,699 (“the ’699 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 ’699 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 ’699 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 ’699 patent specification relate to claim limitations of the ’699 patent.

4. In reaching the opinions provided herein, I have considered the ’699 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 ('699 patent, 1:1-10) points to the '699 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. In this declaration, I provide focused testimony to explain the background and terms of relevant art and provide helpful context regarding the applied prior art. I further provide focused testimony related to the limitation added by the Patent Owner to overcome the Examiner's rejection, "*identifying ... a manner in which the natural language utterance was spoken based on the short-term knowledge and the long-term knowledge about how a user utters a request*", and the "*short-term knowledge*" and "*long-term knowledge*" limitations added by Examiner's amendment to secure allowance of the claims.

7. 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.

8. 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 '699 patent. I have no financial interest in the outcome of this matter or in any litigation involving the '699 patent.

## **I. Qualifications**

9. 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.

10. 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.

11. 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").

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

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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:

25. 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.

26. 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.

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

28. 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.

29. 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.

30. 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.

31. 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 directors, and search for information.

32. 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.

33. 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.

34. 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:

35. 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.

36. 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.

37. 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.

38. 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**

39. 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.

40. 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.

41. 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 ’699 patent. I reserve the right to supplement this Declaration if Patent Owner (“PO”) introduces evidence of secondary considerations of non-obviousness.

42. 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.

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

### **III. Technical Background**

44. The '699 patent provides limited details regarding interactive dialogue systems and the concepts/terminology used in this field. I provide in this section a background of the field to provide context and also provide discussion of common terms used in the field. Prior to the claimed priority date of the '699 patent (October 16, 2006), conversational systems like the type described in the '699 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 interactions<sup>1</sup>, the users are able to “cope with the

---

<sup>1</sup> 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.” (EX-1021, 9.)

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:

- *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 '699 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.

#### **A. 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, researchers 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).)

## **B. Interpreting Meaning in Dialog**

### **1. 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.

## 2. 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 ’699 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 ’699 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 ’699 patent acknowledged during a prior litigation that the claimed concepts of cooperative conversation came from Grice’s paper. (EX-1027, 288:8-292:4.)

### **C. 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

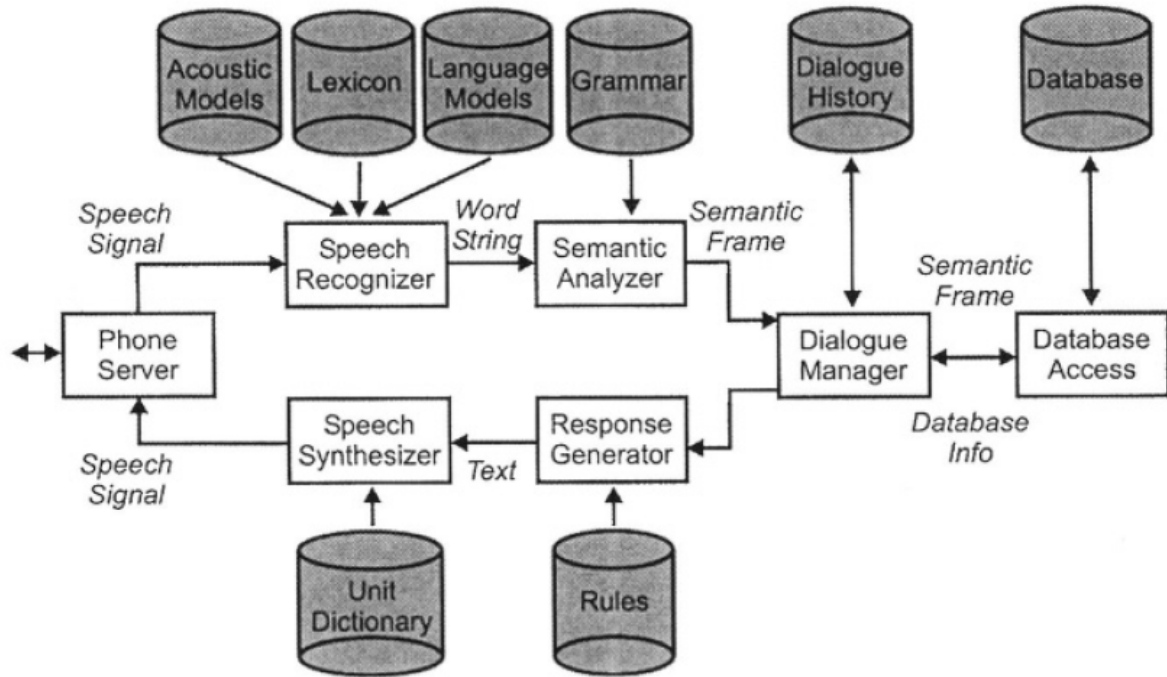
them suitable for tasks like automated customer service and information retrieval. (EX-1021, 20.)

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:

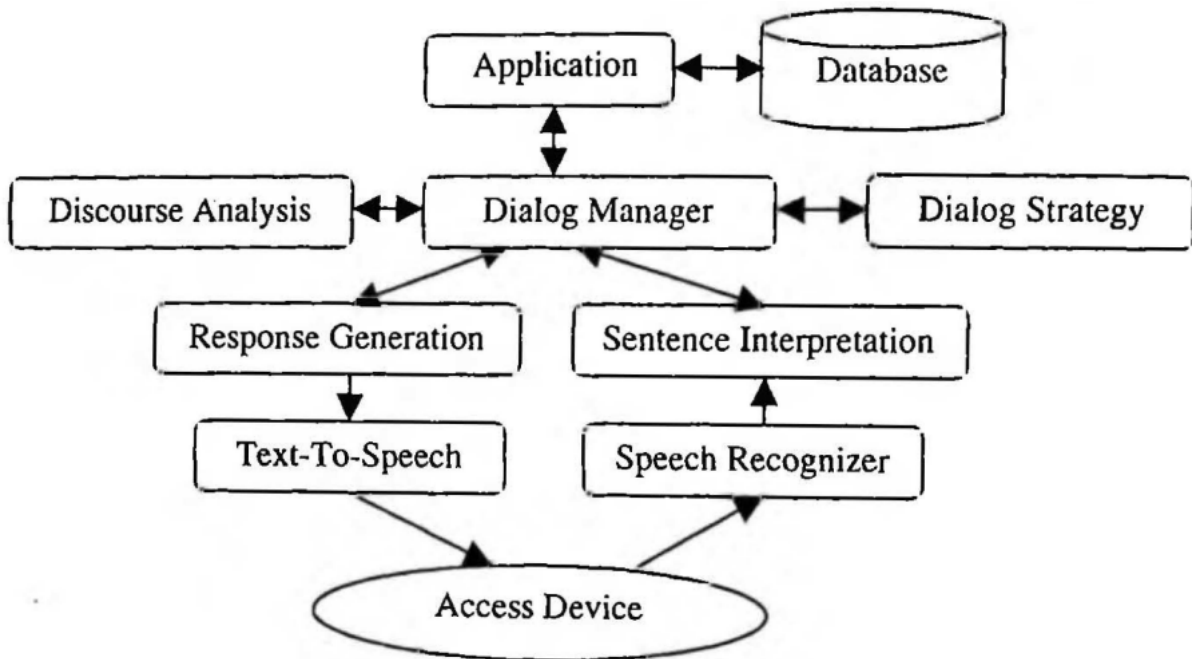
- **Phone Server:** Provides access to the user over the telephone network.
- **Speech Recognizer:** Converts the spoken input into a word string or word hypothesis graph.
- **Semantic Analyzer:** Interprets the recognized word string semantically, producing a semantic frame.
- **Dialogue Manager:** 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

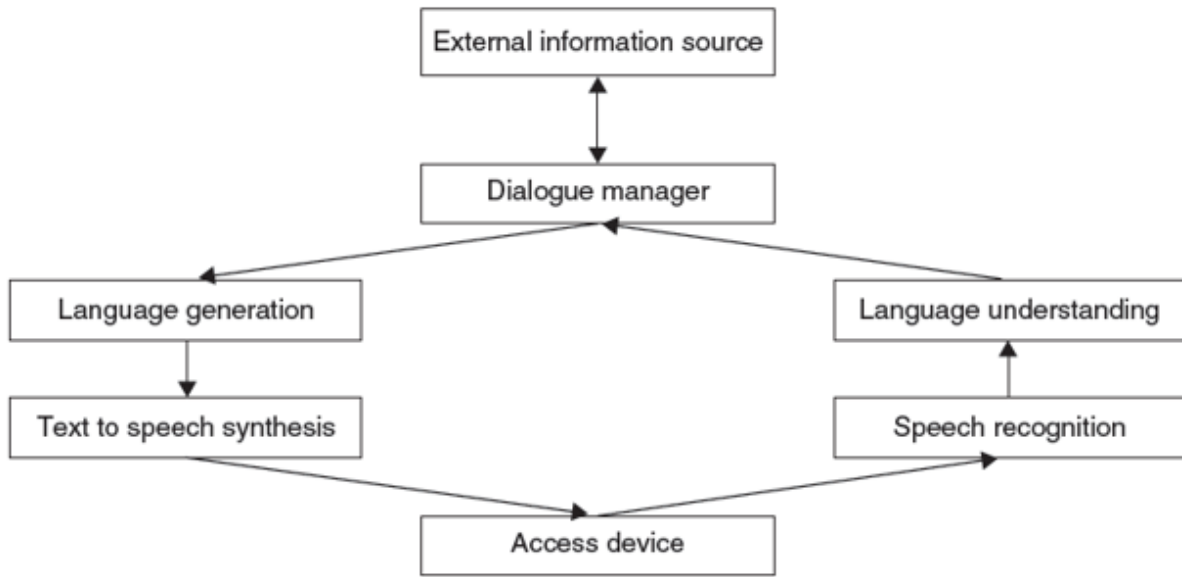
understanding component. The output from the semantic analyzer component “is a semantic frame representing what has been ‘understood’ from the user’s utterance.” (EX-1021, 20.) The dialog manager “interpret[s] the semantic frame in the context of the dialogue and the task” and “keep[s] track of the dialogue 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

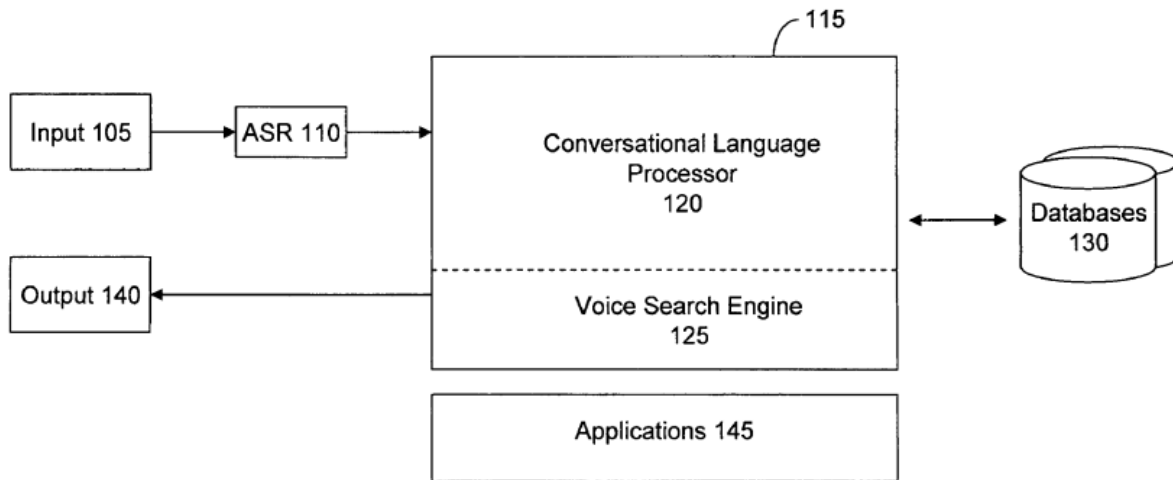


Huang, Figure 1.4



**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 ’699 patent is illustrated in Figure 1 of the ’699 patent (reproduced below).



**‘699 Patent, Figure 1**

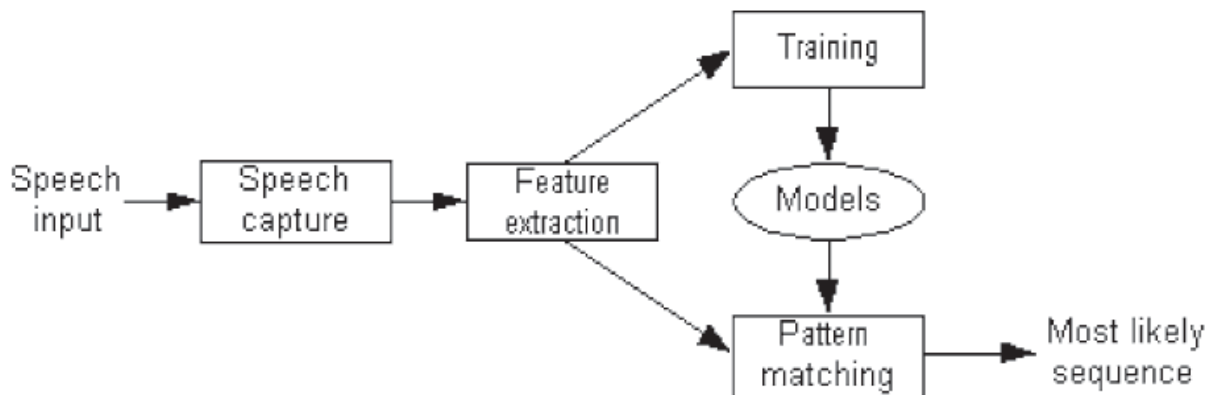
75. Further, in some implementations, some components may not be necessary, at least not as conceived by Möller, McTear, and Huang. For instance, while many systems at the relevant period were telephone based, that was not universally true and those that were not could use another voice-input device 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.

### **1. 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.)

## **2. 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.)

### **3. 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.)

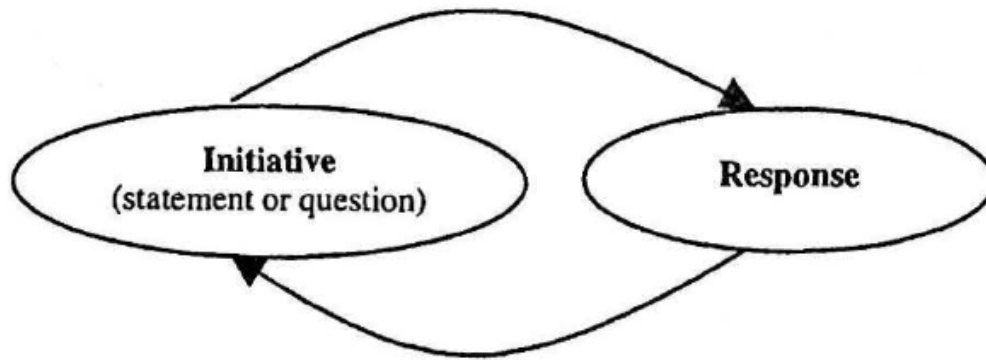
86. The Dialog Manager also serves as “a type of service controller which administers the flow of information between the different modules” in a spoken dialogue system, coordinating the flow of interaction by coordinating the flow of interaction by “interpret[ing] the semantic frame in the context of 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.

**a) 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.)

### **b) 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.

### **c) 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.)

#### **4. 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.

#### **IV. GROUND 1: The Combination of SmartKom and Kobsa**

##### **A. Overview of the Combination**

###### **1. SmartKom**

104. 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.) Because SmartKom is a collection of papers, I provide context in this section regarding the end-to-end system.

105. The SmartKom initiative, begun in spring 1999, over 7 years before the earliest possible priority date of the ’699 patent, 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 earliest possible priority date of the ’699 patent. (EX-1005, 35.)

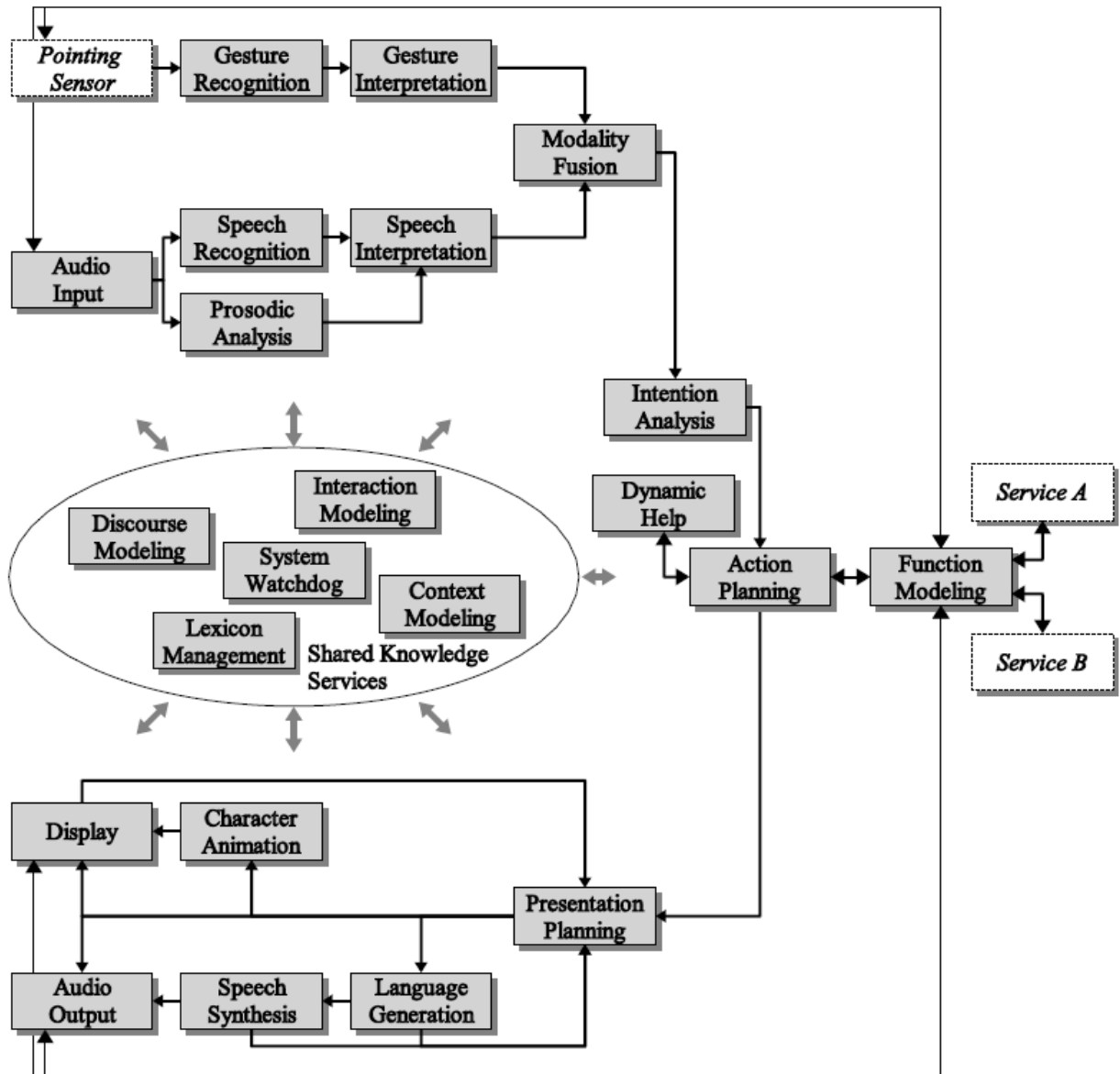
106. The SmartKom system “provides full symmetric multimodality in a mixed-initiative dialogue system with an embodied conversational agent.” (EX-

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.)

107. The generic software architecture of the SmartKom, illustrated in Figure 2(60)<sup>2</sup> 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 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.)

---

<sup>2</sup> 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.



SmartKom, Figure 2(60)

108. The basic SmartKom system “supports multimodal input using speech plus gestures and integrates some application-specific services.” (EX-1005, 60.) “Modality-specific input analysis can be modularized into separate components, and it has proven useful to differentiate further between modality-specific recognition,

i.e., processing of sensory data on the lexical and syntactical layer, and subsequent semantic interpretation of the derived symbolic representation. (EX-1005, 61.) For the speech modality, the input analysis components include the speech recognition and speech interpretation components.

109. In the SmartKom system, “[t]he input of a user is initially processed by recognizers like a speech recognition or a gesture recognition component.” (EX-1005, 286.) The recognizers “produce a semantic interpretation of a grammar applied to the speech recognition result, or the representation of an element presented on the screen which is likely to be addressed by the user.” (*Id.*) Further, SmartKom explains that “stochastic methods are applied in order to compute readings of the respective input.” (*Id.*) Therefore, “[i]nstead of computing one single ‘best’ result, a number of hypotheses with a rating of the quality of each hypothesis might be computed.” (*Id.*)

110. More specifically, SmartKom’s speech recognition engine “transforms the acoustic signal into a **sequence of hypothesized words** in orthographic representation.” (EX-1005, 86, Figure 2(60).) The speech recognition engine uses “**confidence measures**” to “estimate confidences for the correctness of the words hypothesized by the recognizer.” (EX-1005, 85, 96.) The natural language processing (speech interpretation) module “transform[s] the word lattice sent by the speech recognizer into a list of hypotheses representing ... possible user intentions.”

(EX-1005, 195.) The hypotheses, each with a corresponding score value (confidence in the speech recognition result), is provided to the modality fusion component. (EX-1005, 204, 224.)

111. SmartKom further includes a prosody<sup>3</sup> module that “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.)

112. 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 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.)

---

<sup>3</sup> Prosody “refers to the segments of speech larger than phonemes, e.g., syllables, words, phrases, and whole utterances.” (EX-1005, 139.)

Second, the probability of the whole utterance is processed “to decide one of the several classes.” (EX-1005, 150.)

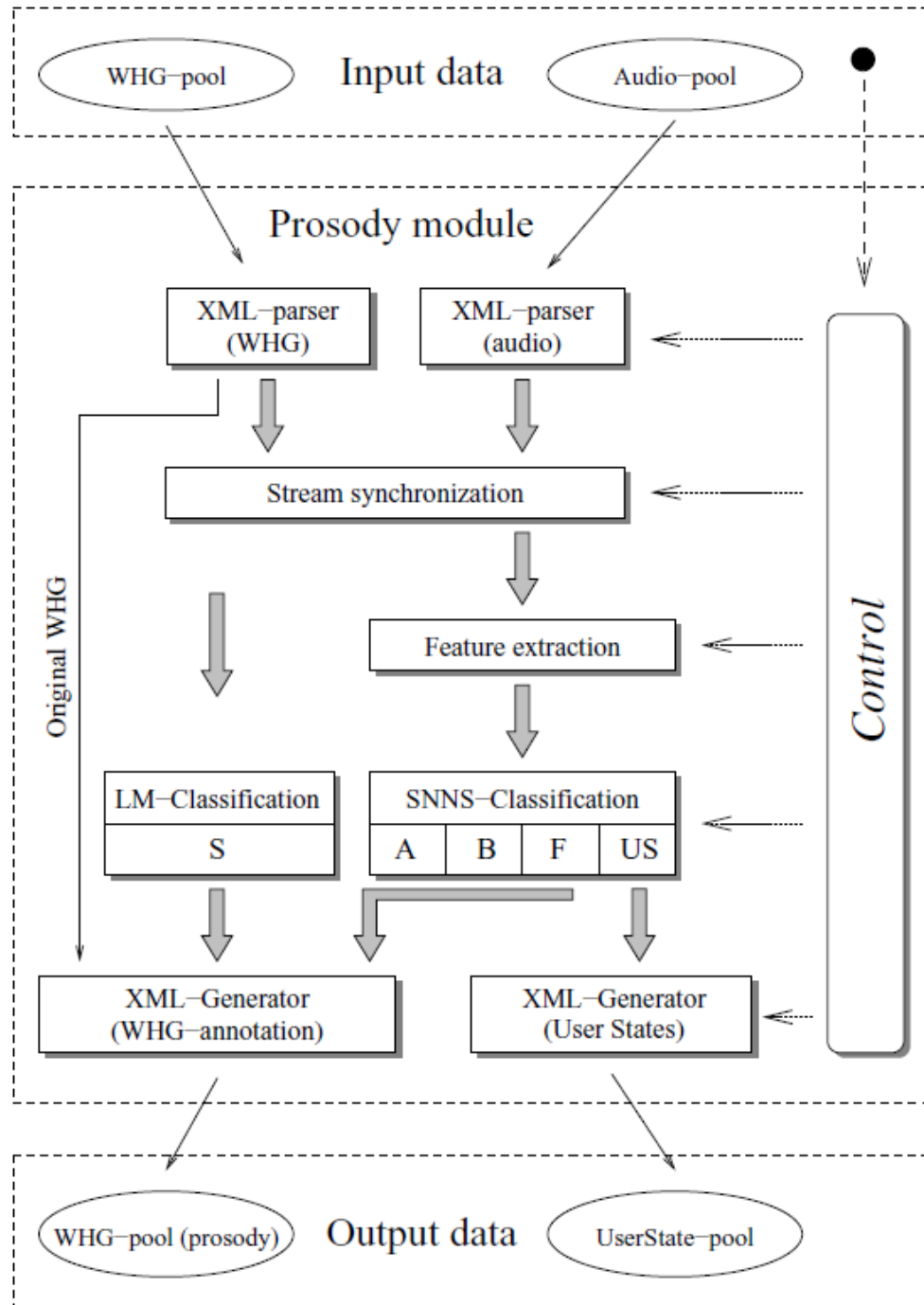


Fig. 1. Architecture of the prosody module

113. The 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.)

114. “Interaction management within SmartKom comprises several components.” (EX-1005, 61.) The intention analysis component “select[s] the most plausible input interpretation from the given set of hypotheses.” (*Id.*) 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.) “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. (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. (EX-1005, 285.) The intention recognizer then “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.)

115. After selection of an interpretation from the set of hypotheses, the intention recognizer provides the interpretation to the action planner, also referred to as the dialogue manager, which makes “decisions on the applications to contact and the next interaction with the user.” (EX-1005, 287.) “Action planning constitutes the heart of dialogue control and is backed by a supplementary help component that is activated whenever difficulties occur during the interaction or if additional help is needed.” (EX-1005, 61.) The main processing stages are supported by the following components that “actively maintain shared knowledge sources:

- The multimodal discourse model is utilized for the semantic and pragmatic interpretation during input and output processing. It is dynamically updated as system output progresses and performs contextual reasoning and scoring.
- Contextual information, as it is needed to handle references to situative parameters like current place and time, is provided by the context model.

- 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. The interaction model allows one to dynamically adapt the communicative behaviour of the system.
- The lexicon is a dynamic knowledge source, which is updated with additional lexical entries depending on dynamic application data as it is received from the external information services. Lexicon updates are propagated to all components that process natural language input and output.”

(EX-1005, 61.)

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

**Table 3.** Fully expanded plan for sending a fax message

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

**SmartKom, Table 3(313)**

117. 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 adapted based on the interpretation (“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). (EX-1005, 313 (Table 3).)

118. “The function modeling component realizes the application interface and in addition controls all input and output devices to coordinate access depending on an explicit state model.” (EX-1005, 61.) This is the component that, e.g., queries external resources for movie times, television listings, and/or directions.

119. The presentation planner, natural language generation [NLG] component, and speech synthesis cooperate to present the response multi-modally to the user. (See EX-1005, 396, 401.)

120. In SmartKom, “[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, 305 (emphasis in original); see also EX-1005, 287-288.) 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.)

121. 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,

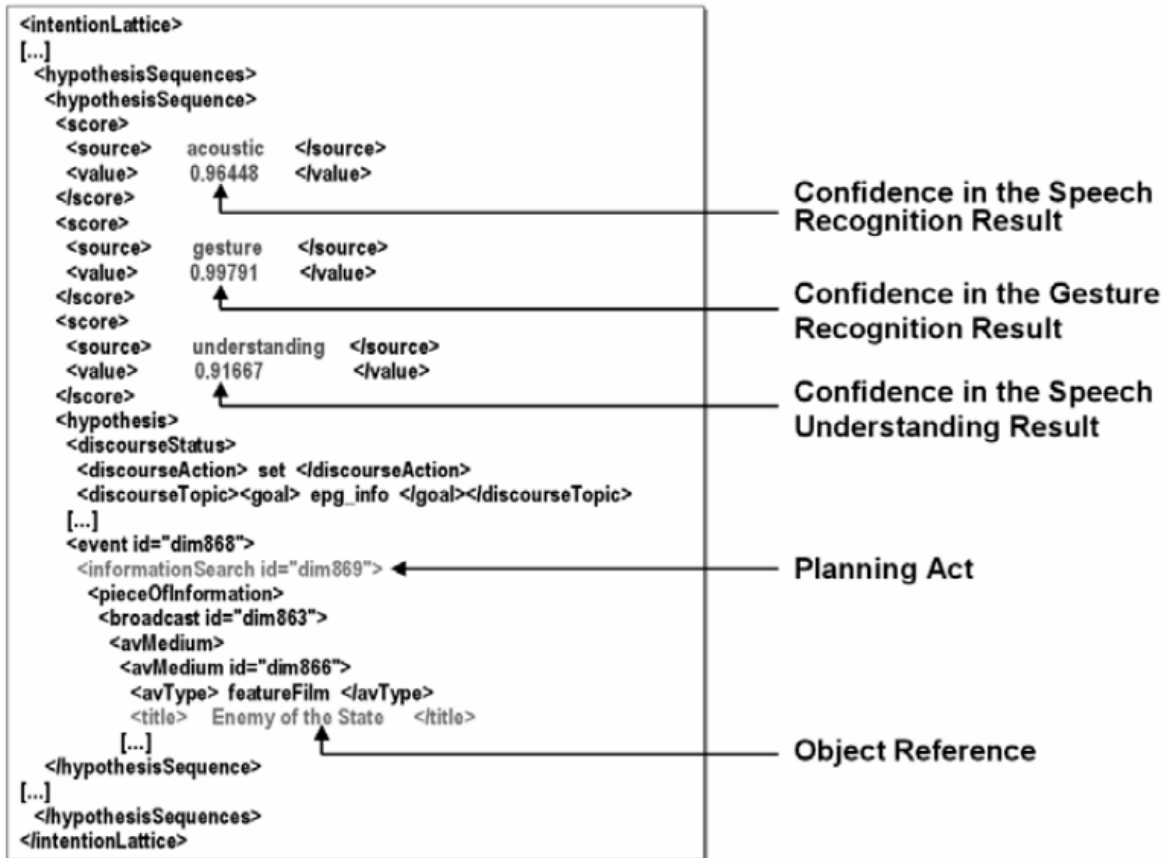
305.) “The constituents of each intention segment that are important for the dialogue manager are therefore” the following:

- 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.)

122. 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.) An exemplary intention lattice fragment is illustrated below in SmartKom’s Figure 8(15). 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 which works with other output components to present the response to the user. (EX-1005, 204.)



SmartKom, Figure 8(15)

## 2. Kobsa

123. “User Models in Dialog Systems” by Kobsa (“Kobsa”; EX-1006) published in 1989, 17 years prior to the earliest possible priority date of the ’699 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(16)

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 ’699 patent—speech

recognition systems. (EX-1001, 1:28-29; 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.)
- Identifying “UserKnowledge” including “user familiarity with task” and “user familiarity with system” as model used to adapt system responses. (EX-1005, 322-324.)
- “*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 and 12**

132. In this section, I focus my testimony on a subset of claim limitations including the limitation amended by the Patent Owner during prosecution to overcome a rejection (EX-1002, 419) and the limitations added by Examiner's

amendment to secure allowance (EX-1002, 444.) I also address aspects of the preamble and the “*identifying ... a context*” limitations to provide additional analysis and context. Despite this focused testimony, it is my opinion that the combination of SmartKom and Kobsa discloses all the limitations of the independent claims 1 and 12 and the dependent claims 2-11 and 13-22.

### **1. Preamble [1P.2] and Limitation [12A]**

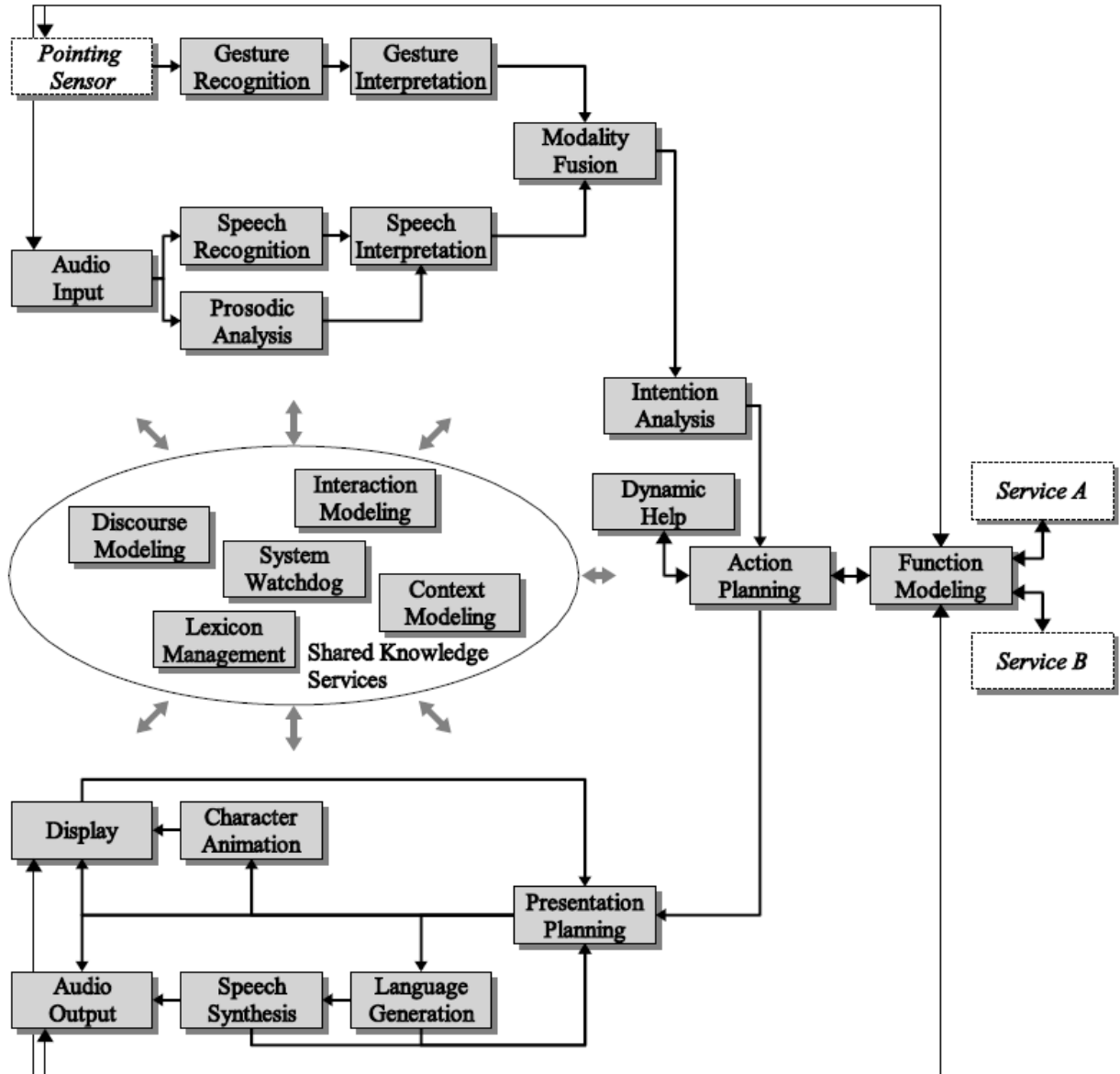
133. Claims 1 and 12 recite “*a computer system*” [1P.2] including “*one or more physical processors executing one or more computer program instructions<sup>4</sup> which, when executed*” [1P.2]/[12A] “*perform the method*” [1P.2] or “*configure the one or more physical processors.*” I note that, like the ’699 patent, SmartKom does not explicitly use the word “*instructions.*” However, this was a well-known term in the art.

134. SmartKom describes that the SmartKom software executing on a server “consist[ing] of 3 dual Xeon 2.8 GHz processor.” (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).) The generic software architecture of the SmartKom system, illustrated in Figure 2(60) below, includes a set of components with each

---

<sup>4</sup> The ’699 patent does not use the term “*instructions.*” At most, it mentions a “*conversational language processor.*” (*See* EX-1001, 8:23-28.)

“component represent[ing] one of the principal processing units of the executing system.” (EX-1005, 59.)



SmartKom, Figure 2(60)

135. A POSITA would have known, long before the earliest possible priority date of the '699 patent, that software is “computer programs; instructions that make

hardware work.” (EX-1013, 489.) Although SmartKom does not use the term “*instructions*,” a POSITA would understand that SmartKom’s software contains “*computer program instructions*.”

## 2. Identifying a Context [1C]/[12D]

[1C] identifying, by the computer system, a context for the natural language utterance based on the one or more words or phrases recognized from the natural language utterance;

[12D] identify a context for the natural language utterance based on the one or more words or phrases recognized from the natural language utterance;

136. The combination of SmartKom and Kobsa discloses a “*computer system*” that identifies “*a context for the natural language utterance based on the one or more words or phrases recognized from the natural language utterance*” [1C]/[12D]. SmartKom notes that “[s]peakers may not always be aware of the potential ambiguities inherent in their utterances” and “leave it to the context to disambiguate and specify the message.” (EX-1005, 275.) Instead, “[t]hey leave it to the context to disambiguate and specify the message.” (*Id.*) “In order to interpret the utterance correctly, the addressee must employ several context-dependent resources.” (*Id.*)

137. The intention recognizer, that I described in the SmartKom overview, 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.)

138. As part of this process, “the discourse modeler receives a set of hypotheses” and enriches each “with previous discourse information” (discourse state) and context-domain information from the domain model (“*long-term 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.)

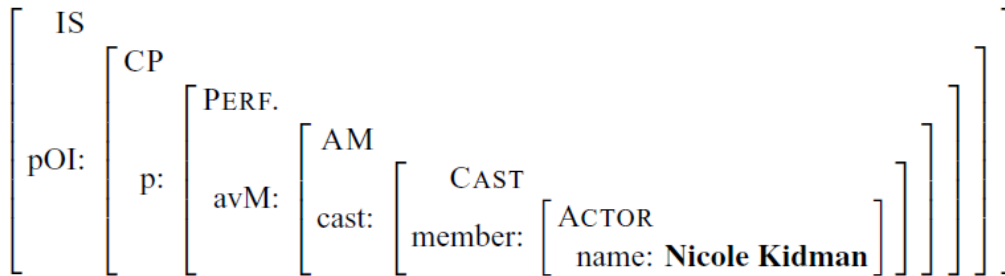
139. The following example illustrates this context identification. In this example, “the user has requested information about tonight’s television program,” in utterance-1. (EX-1005, 262-263.) The TFS representation of the utterance-1 (part of the discourse history in the discourse state) is shown in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

U-1: What’s on TV tonight?

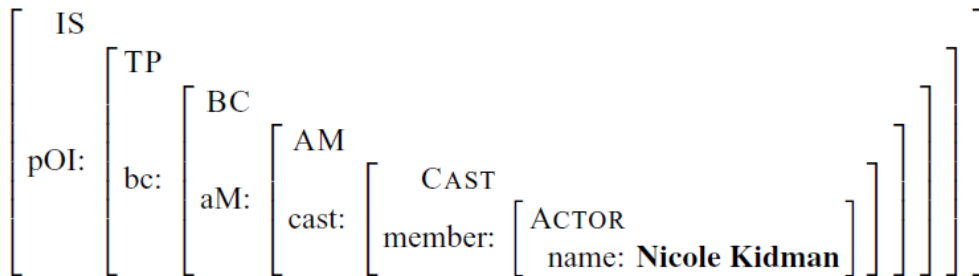
SYS: (Displays the TV program listings) Here you see a list of movies running tonight.



Figure 4(264) below. 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)



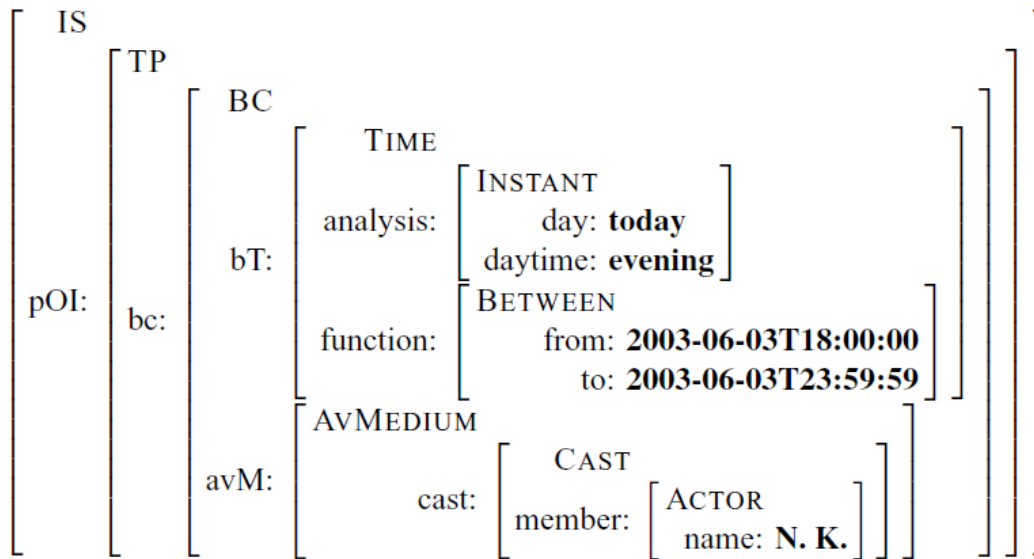
**Fig. 4.** A second possible interpretation of the utterance *Is there a movie with Nicole Kidman?* (*IS* = InformationSearch, *pOI* = pieceOfInformation, *CP* = CinemaProgram, *p* = performance, *P* = Performance, *aM* = avMedium, *AM* = AvMedium)

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

141. Each of these hypotheses, enriched with context-domain information, is then “overlay[ed]” with the prior utterance (U-1) and scored. (EX-1005, 263.)

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, 263.) The interpretation of U-2 with respect to the previous discourse context is shown in Figure 5(265) below. 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)

142. 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.)

143. 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.)

144. 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.)

145. 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)**

146. 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) is added. (See, e.g., EX-1005, 306, 317-319.) 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> &lt;informationSearchProcess&gt;   &lt;entertainment&gt;     &lt;performance&gt;       &lt;cinema&gt;         &lt;contact&gt;           &lt;address&gt;             &lt;town&gt;               here             &lt;/town&gt;           &lt;/address&gt;         &lt;/contact&gt;       &lt;/cinema&gt;     &lt;time&gt;       &lt;beginTime&gt;         &lt;at&gt;           now         &lt;/at&gt;       &lt;/beginTime&gt;     &lt;/time&gt;   &lt;/performance&gt; &lt;/entertainment&gt; &lt;/informationSearchProcess&gt; </pre>	<pre> &lt;contact&gt;   &lt;x&gt; 70.345 &lt;/x&gt;   &lt;y&gt; 49.822 &lt;/y&gt;   &lt;town&gt;     Heidelberg   &lt;/town&gt; &lt;/contact&gt; &lt;time&gt;   &lt;at&gt; 19:00:00T26:08:03 &lt;/at&gt; &lt;/time&gt; </pre>
	<pre> &lt;scores&gt;   &lt;contextualCoherence&gt;     0.46   &lt;/contextualCoherence&gt; &lt;/scores&gt; </pre>

**SmartKom, Table 2(277)**

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

### **3. Short-Term and Long-Term Knowledge Limitations**

148. The independent claims recite “*short-term knowledge*” and “*long-term knowledge*.” According to the ’699 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, 5:10-17.) 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, 5:29-39.)

149. It was well-known, prior to the ’699 patent, that in interactive dialogue systems 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.)

150. SmartKom uses both “*short-term knowledge*” and “*long-term 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.)

151. 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.) SmartKom’s architecture provides “shared knowledge services” including modeling components that “enable[] the system to act analogously, i.e., to provide hitherto implicit knowledge concerning what is talked about.” (See, e.g., EX-1005, 277-278, Figure

2(60).) The specific knowledge stores and their associated context are illustrated in SmartKom’s Table 1(274) (below), that provides a “broad categorization of the types of context relevant to spoken dialogue systems, their content and respective knowledge stores.” (EX-1005, 274; *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)**

**a) Accumulating Short-Term Knowledge [1E.1]-[1E.2]/[12F.1]-[12F.2]**

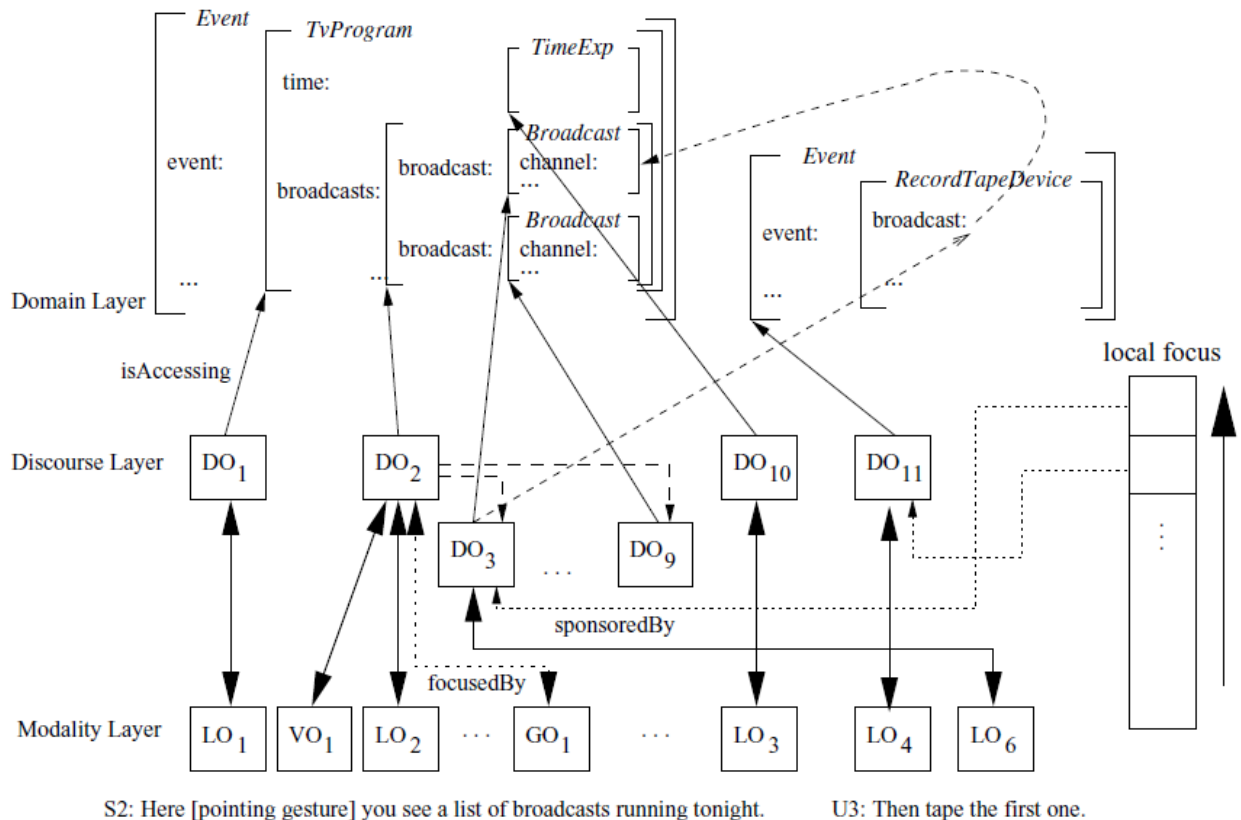
<p>[1E.1] accumulating, by the computer system, short-term knowledge based on one or more natural language utterances received during a predetermined time period,</p> <p>[12F.1] accumulate short-term knowledge based on one or more natural language utterances received during a predetermined time period,</p> <p>[1E.2] wherein the one or more natural language utterances received during the predetermined time period are related to a single conversation between a user and the computer system;</p> <p>[12F.2] wherein the one or more natural language utterances received during the predetermined time period are related to a single conversation between a user and the computer system;</p>
--

152. The SmartKom-Kobsa combination discloses the “*computer system*” “*accumulate[es]... short-term knowledge based on one or more natural language utterances ...*” [1E.1]/[12F.1] in two ways.

153. 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.)

154. 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 knowledge*” that is “*based on one or more [received] natural language utterances*” consistent with the ’699 patent’s disclosure. (See, e.g., EX-1001, 5:10-12 (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.)

155. 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)**

156. 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 “*based on one or more [received] natural language utterances.*”

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

158. Second, SmartKom accumulates situational and user state knowledge during an ongoing conversation. 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 knowledge*” that is “*based on one or more [received] natural language utterance*” (i.e., the time and location of the utterance), consistent with the ’699 patent’s disclosure. (See EX-1001, 5:12-17 (indicating short-term knowledge includes “location[.]” data).)

159. 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).) “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 knowledge*” that is “*based on one or more [received] natural language utterances*.”

160. To gather user state information, SmartKom’s prosody<sup>5</sup> 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.)

161. 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 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 “*based on one or more [received] natural language utterances*.”

---

<sup>5</sup> Prosody “refers to the segments of speech larger than phonemes, e.g., syllables, words, phrases, and whole utterances.” (EX-1005, 139.)

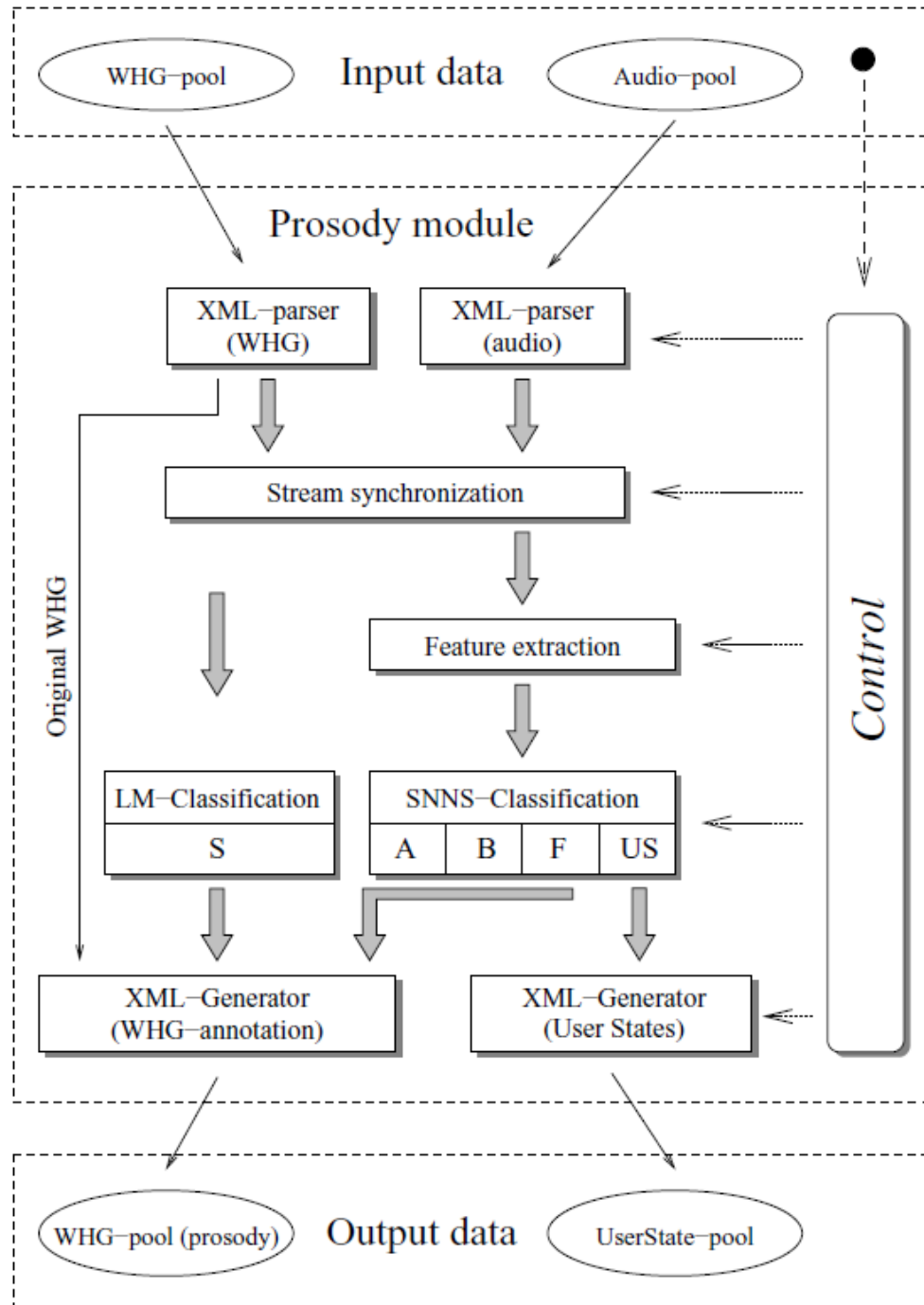


Fig. 1. Architecture of the prosody module

SmartKom, Figure 1(142)

162. The SmartKom-Kobsa combination discloses the “*short-term knowledge*” accumulates “*during a predetermined time period,*” consistent with the ’699 patent.

163. Regarding “*predetermined time period,*” the claims recite that the “*short-term knowledge*” is accumulated “*during a predetermined time period*” and “*items of short-term knowledge*” are expired “*prior to the predetermined time period.*” The ’699 patent uses the term “*predetermined*” in the context of time only once and unrelated to the duration of the time period when the “*utterances*” are received:

For example, a human is unlikely to recall a context of a conversation from two years ago, but because the context would be identifiable by a machine, session context is **expired after a predetermined amount of time** to reduce a likelihood of contextual confusion based on stale data. However, relevant information from an expired session context may nonetheless be added to user, historical, environmental, cognitive, or other long-term knowledge models.

(EX-1001, 14:34-42.)

164. SmartKom-Kobsa discloses or suggests accumulating short term knowledge during a conversation between the user and system. SmartKom’s discourse state is “based on the three-tiered context representation presented in Luperfoy (1992).” (EX-1005, 239.) Luperfoy notes an “agent participating in a dialogue experiences information decay[ing] over the course of the conversation”

with “information from the linguistic, discourse, and belief system tiers decays at different rates and in response to different cognitive forces/limitations.” (EX-1012, 24-25.) Specifically, LOs “become old and vanish at an approximately linear rate as a function of time counted from the point of their introduction into the discourse history, i.e., as LOs get older, they fade from the discourse and can no longer serve as linguistic sponsors for anaphors.” (EX-1012, 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-1012, 25.)

165. SmartKom stores its discourse state in “discourse memory.” (EX-1005, 242.) SmartKom acknowledges this decay of information, explaining that “[f]or longer dialogues (more than half an hour of discourse), the discourse memory runs out of memory” and in such cases, “even for most humans,” it is “necessary to forget information.” (EX-1005, 252.) Kobsa also discloses “at the end of a dialog session” (e.g., the conversation), the system “records all information about the user inferred from his/her dialog behavior in a corresponding file.” (EX-1006, 10-11.)

166. Thus, SmartKom-Kobsa discloses that “*short-term knowledge*” is accumulated for a dialog between the user and system over the shorter of (1) the

duration of the conversation/dialog or (2) the size of short term storage (e.g., 30 minutes)—“*a predetermined time period.*”

167. The SmartKom-Kobsa combination therefore discloses “*accumulating, by the computer system, short-term knowledge based on one or more natural language utterances received during a predetermined time period*” [1E.1] and “*accumulate short-term knowledge based on one or more natural language utterances received during a predetermined time period*” [12F.1].

168. SmartKom discloses that “*the one or more natural language utterances received during the predetermined time period are related to a single conversation between the user and the computer system*” [1E.2]/[12F.2]. SmartKom provides numerous utterances occurring in a conversation between the user and the computer system. For example:

U1: *What’s on at the movies tonight?*

S1: (Displays a list of movies) *Here [↗]<sup>3</sup> you can see tonight’s cinema program.*

U2: *And what’s on TV?*

S2: (Displays a list of broadcasts) *Here [↗] are tonight’s broadcasts.*

U3: *Is there a movie with Arnold Schwarzenegger?*

(EX-1005, 238.) As discussed above, the “*predetermined time period*” is the shorter of (1) the duration of the conversation/dialog or (2) the size of short term storage (e.g., 30 minutes). As highlighted by the above example, the user utterances are

“related to a single conversation between the user and the computer system” and occur “during the predetermined time period.”

**b) Accumulating Long-Term Knowledge [1F]/[12G]**

[1F] accumulating, by the computer system, long-term knowledge, wherein the long-term knowledge is accumulated based on one or more natural language utterances received prior to the predetermined time period;

[12G] accumulate long-term knowledge, wherein the long-term knowledge is accumulated based on one or more natural language utterances received prior to the predetermined time period;

169. The SmartKom-Kobsa combination discloses “*long-term knowledge*.” SmartKom discloses a user model. Table 1(274) shows that the user model stores “properties of the interlocutory” (i.e., information about the system users) and provides “interlocutory context” associated with the content stored in the “User model”. (EX-1005, 274.) SmartKom explains the “**role of the interlocutory 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 information<sup>6</sup> is supplied via external sources, e.g., via a user’s

---

<sup>6</sup> 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.

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, 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 is “*based on one or more natural language utterances received.*”

**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)**

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

171. 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.)

172. 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, the long-term user model includes knowledge about prior conversations with the same user. (EX-1001, 5:35-39 (long-term knowledge includes “explicit and/or implicit user preferences”).)

173. 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, 199.) 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, 199.) To represent concepts a user is “probably familiar with,” an overlay technique is employed, such as shown in Figure 3 below. (EX-1006, 15.)

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(16)

174. The combination of SmartKom and Kobsa discloses “*accumulat[e]/[ing] ... long term knowledge.*” 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.)

175. Both SmartKom and Kobsa describes techniques for “*accumulating long-term knowledge.*” SmartKom describes that the “interaction model” of the Knowledge Base “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).)

176. 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(419) (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.)

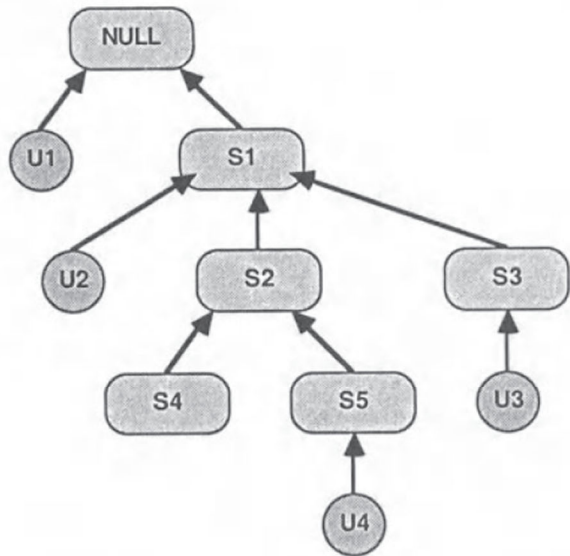


Figure 3. A hierarchy of stereotypes and individuals

Kobsa, Figure 3(419)

#### 4. Identifying a Manner [1G]/[12H]

[1G] identifying, by the computer system, a manner in which the natural language utterance was spoken based on the short-term knowledge and the long-term knowledge; and

[12H] identify a manner in which the natural language utterance was spoken based on the short-term knowledge and the long-term knowledge; and

177. The combination of SmartKom and Kobsa discloses a “computer system”/“physical processor[s]” that identifies “a manner in which the natural language utterance was spoken” [1G]/[12H]. SmartKom’s “interaction module” uses a three stage approach for “processing of emotions and user states”:

- “At the first stage the emotional state of the user is recognized from facial expression and prosody.”
- “At the second stage indications of problematic situations and **the emotional state of the user are collected from several sources and collectively evaluated.** The component also **analyzes the dialogue in respect to the style of interaction and the task and paradigm knowledge of the user** (Portele, 2003).”
- “The interpretation of emotions and user states, and the generation of reactions to these states build the third stage. It is realized by so-called dynamic help. This component is dedicated to manage subdialogues, and to provide presentation specification and intention analysis in problematic situations that are not handled by the standard dialogue component of SMARTKOM.”

(EX-1005, 320.)

178. SmartKom discloses certain “use cases” to demonstrate the “value of the user state classification” processing. (EX-1005, 320.) For example, SmartKom discloses that “emotion works as a semantic operation that turns a positive feedback into a negative one, which is considered as a form of sarcasm,” which the system “suspects that the emotional reaction may be caused by a like or a dislike concerning the properties of the presented objects.” (EX-1005, 320.) “If the system knows positive or negative preferences, it first presents objects that contain a preferred feature” and those “that show a disliked feature will be shown last.” (EX-1005, 320.)

179. SmartKom provides a number of examples of adapting the system response based on the manner of the user's speaking, including the following example:

**Second constellation: emotionally marked verbal communication**

**user:** *That's really a beautiful program!* (She produces this sentence with an angry prosody. The positive feedback is analyzed as being sarcastic.)

**system:** *You don't like science fiction? Shall I account for that in future presentations?*

**user:** *Yes./No.*

**system:** *OK. I'll take care of that!*

(EX-1005, 321 (emphasis in original).)

180. To identify the user's manner of speaking, the interaction module "collects and evaluates indications of emotions, problematic situations and other aspects of the interaction." (EX-1005, 321 (emphasis in original).) The interaction module is used "to estimate recognition results by taking other indicators into account that may support or devalue the result," and "introduces its own results concerning states of the user or characteristics [sic] of the interaction." (EX-1005, 321.)

181. The interaction module "operates by analyzing a set of possible *indicators*" (listed in Table 1(322) that "have values between 0 and 1." (EX-1005, 321 (emphasis in original).) For example, there are separate indicators associated with prosody module which convey the likelihood of "prosodically conveyed

anger”, “prosodically conveyed joy” and “prosodically conveyed dilatoriness.” There is also an indicator associated with speech recognition that conveys the likelihood of “linguistically conveyed anger.” As shown in the Table below, the indication are derived from data collected from the prosody recognizer, the user model, the discourse model, and the domain model. The “[i]ndicator values are mapped to models” which I discuss below. (EX-1005, 322.)

**Table 1.** List of indicators

Source	Description
Mimic recognizer	Mimically conveyed anger
Prosody recognizer	Prosodically conveyed anger
Mimic recognizer	Mimically conveyed joy
Prosody recognizer	Prosodically conveyed joy
Mimic recognizer	Mimically conveyed dilatoriness
Prosody recognizer	Prosodically conveyed dilatoriness
Speech recognition	Linguistically conveyed anger
Speech understanding	Ratio of unanalyzable words
Intention analysis	Overall score of the best hypothesis
Intention analysis	Difference in score between first and second best hypotheses
Intention analysis	Number of possible hypotheses (depth of lattice)
Speech recognition	Score of the speech recognizer
Gesture recognition	Score of the gesture analyzer
Speech understanding	Score of the language analyzer
Media integration	Score of multimodal integration
Discourse history	Score of the discourse module
Domain model	Score of the domain module
Intention analysis	Final score of the intention module
Intention analysis	Number of elements in the user input
Discourse history	Number of new (not previously mentioned) elements
Speech understanding	Number of elements addressed by speech
Gesture analysis	Number of elements addressed by gesture
Media integration	Number of elements addressed by speech and gesture
Intention analysis	Importance of speech recognition score for overall score
Intention analysis	Importance of gesture analysis score for overall score
Intention analysis	Importance of domain model score for overall score
Intention analysis	Importance of language understanding score for overall score
Intention analysis	Importance of discourse model score for overall score
Speech understanding	Relative number of sentence-like units in one turn
Speech understanding	Relative number of words in one turn
Speech understanding	Relative frequency of pronouns
Speech understanding	Relative frequency of verbs
Speech understanding	Relative frequency of adverbs
Speech understanding	Relative frequency of nouns
Speech understanding	Relative frequency of content words
Speech understanding, language generation	Relative frequency of content words appearing in the system output
Speech understanding, language generation	Relative frequency of content words not appearing in the system output

**SmartKom, Table 1(322)**

182. The indicator values are then “mapped to the models by means of a matrix multiplication.” (EX-1005, 322.) The matrix “design is motivated by the observation that most indicators can contribute to different models, and that the combination of simple indicators to complex models may be optimized by machine-learning algorithms.” (EX-1005, 322.) The interaction module “delivers four sets of models”: problem, user knowledge, modality, and linguistic (EX-1005, 322-23):

**Table 2.** List of models

Set	Description
Problem	Likelihood of a problem
Problem	Likelihood of an analysis problem
Problem	Discourse progress rate
Problem	Likelihood of the user being angry
Problem	Likelihood of the user being happy
UserKnowledge	Estimation of user familiarity with task
UserKnowledge	Estimation of user familiarity with system
Modality	Ratio of spoken input content
Modality	Ratio of gestural input content
Modality	Ratio of multimodal input content
ModalityContrastive	Ratio of contrastive usage of multimodal input
ModalityRedundant	Ratio of redundant usage of multimodal input
Linguistic	Adaptivity of user’s lexical choices to former system output
Linguistic	Likelihood of long turns
Linguistic	Likelihood of long sentences
Linguistic	Ratio of pronoun usage
Linguistic	Ratio of verb usage
Linguistic	Ratio of adverb usage
Linguistic	Ratio of noun and verb usage

**SmartKom, Table 2(323)**

183. The UserKnowledge set of model values “reflects the assumed task and paradigm knowledge of the user.” (EX-1005, 323.) “The task knowledge describes the user’s knowledge of the current task (e.g., programming a VCR), while the

paradigm knowledge indicates how well the user is accustomed to dealing with multimodal dialogue systems, and, especially, with SmartKom.” (EX-1005, 323.) A POSITA would understand that the UserKnowledge is obtained from the user model, which is “*long-term knowledge*.” As discussed above, in §IV.B.3, SmartKom’s user model provides “properties of the interlocutors” to support interlocutory/user context. (EX-1005, 274.) The knowledge stored in the user model can be acquired (accumulated) “either *explicitly* or *implicitly*.” (EX-1006, 416 (emphasis in original).) A user’s “familiarity with [a] task” and the user’s “familiarity with [the] system” are “properties of the interlocutor” as they are used to support user context, and therefore are stored in SmartKom’s user profile.

184. Another set of models (“Linguistic”) describes “the linguistic behavior of the user regarding the number of, e.g., referential expressions, usage of complete sentences and average length of input.” (EX-1005, 323.) These models “help to adapt the language generation in order to reflect the user’s style—based on the assumption that this is beneficial.” (EX-1005, 323.) “[T]he adaptivity of a user’s lexical choices to former system output is estimated, which can help adapting dynamic language models used in SMARTKOM and language generation in order to maximize this value as a measure of the common vocabulary.” (EX-1005, 323.)

185. Another set of models (“Modality”) “compares the use of different modalities by the user (for instance, a preference for gestures or spoken input).” (*Id.*)

Three models in the final set identify user state including “likelihood that the user is angry” or happy. (*Id.*) A POSITA would understand that the modality and linguistic sets and user state models include situational knowledge, discourse state, and user state information which are each “*short-term knowledge.*” As discussed above in §IV.B.3.a, SmartKom’s dialogue model contains “what has been said by whom” and is associated with dialogical (discussion/conversation) context, which forms a dialog history. (*See* EX-1005, Table 1(274).) SmartKom accumulates dialog history (discourse state) in discourse memory and uses this knowledge to enrich and score intention hypotheses. (*See* EX-1005, 237.) The Modality and Linguistic models rely on that dialog history to operate. For example, one Modality model describes the “[r]atio of spoken input content,” which would require knowledge of dialog history, i.e., “*short-term knowledge.*” (EX-1005, 323.) As another example, one Linguistic model describes the “ratio of noun and verb usage,” which also would require knowledge of the dialog history, i.e., “*short-term knowledge.*” (EX-1005, 323.)

186. The fourth set of models (“Problem”) addresses “[p]roblematic situations and user state information.” (EX-1005, 323.) One “problem” model “describes the likelihood that the user is angry by combining scores from mimic analysis, emotion extraction from prosody and use of certain words.” (EX-1005, 323.) A second model “indicate[s] problems in the analysis part of the system,” by “combin[ing] confidence values from recognizers and similar scores from speech

analysis, domain model, discourse history and intention recognition as well as differences in the distribution of these values among concurring hypotheses.” (EX-1005, 323.) And a third model “estimates the dialogue progress,” by using a number of indicators. (EX-1005, 323.) The indication of emotions/user states mapped into models by the SmartKom-Kobsa system is the “*manner in which the natural language utterance was spoken*” which is collected and evaluated using the prosody, user model, discourse mode, and domain model (“*identif[ie]d ... based on the short-term knowledge and the long-term knowledge*”).

## **V. GROUND 2: The Combination Of Barbara, Ross And Kellner**

### **A. Overview of the Combination**

187. The combination of U.S. Patent Application Publication 2004/0101198 to Barbara (“Barbara”; EX-1007), which published May 27, 2004, U.S. Patent Application Publication 2002/0173960 to Ross, et. al. (“Ross”; EX-1008), which published November 21, 2002, and U.S. Patent Application Publication 2002/0065651 to Kellner (“Kellner”; EX-1023), which published May 30, 2002, disclose or at least suggest every limitation of claims 1-22.

#### **1. Barbara**

188. Barbara discloses an “interpretation system for interpreting electronic signals”(EX-1007, Abstract) and, specifically, “a system and method for improving accuracy of signal interpretation.” (EX-1007, ¶7; *see also* EX-1007, ¶5.) Barbara discloses two primary embodiments—(1) where the signal includes an image, e.g.,

derived from an electronic scanner (EX-1007, ¶¶11-12, 38-81, Figures 1-5), 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.

189. 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.)

190. Barbara illustrates an 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.)

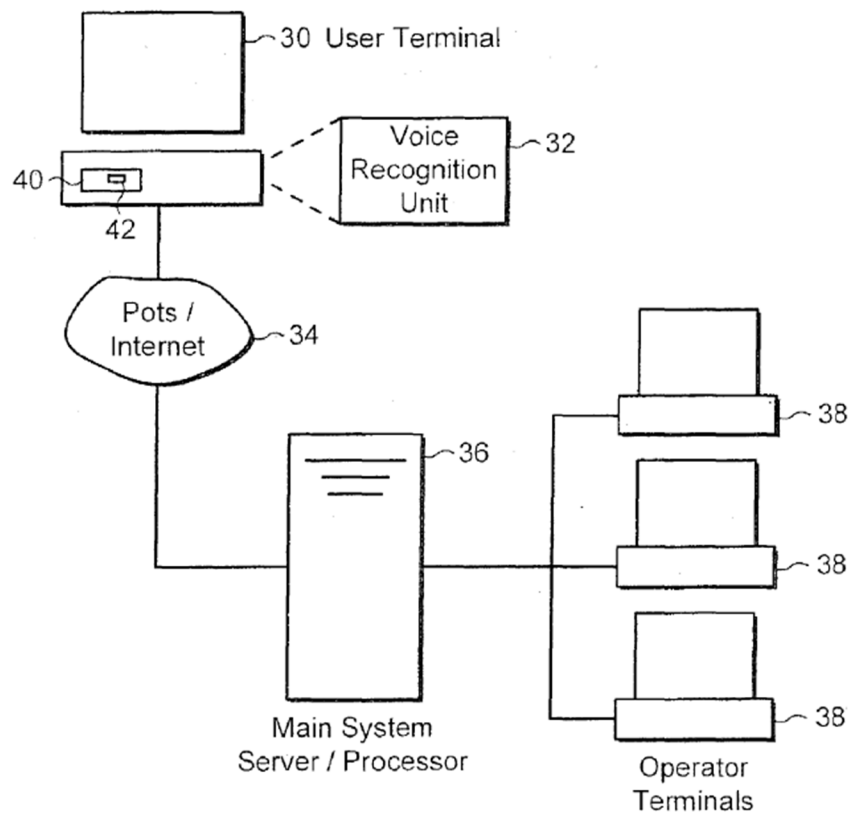


FIG. 6

**Barbara, Figure 6**

191. “[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.)

192. 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.)

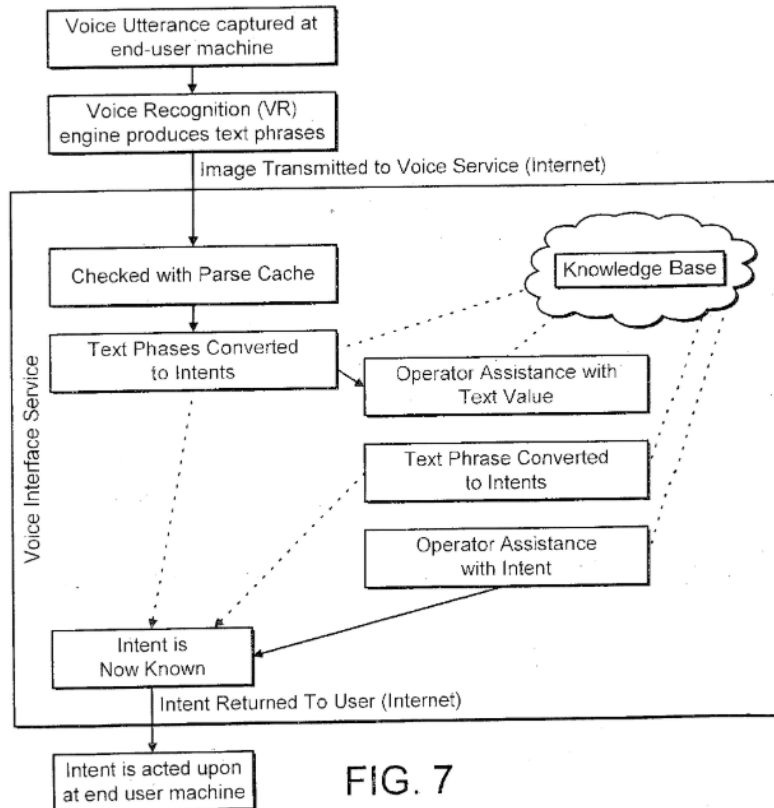


FIG. 7

**Barbara, Figure 7**

193. 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.)

194. 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)” and a “knowledge base.” (EX-1007, ¶¶88, 92-93.) 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.)

195. 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 Nikei 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.)

196. 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.)

197. 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.)

198. 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.) In Barbara, “feedback is given by the application visibly responding” or “the

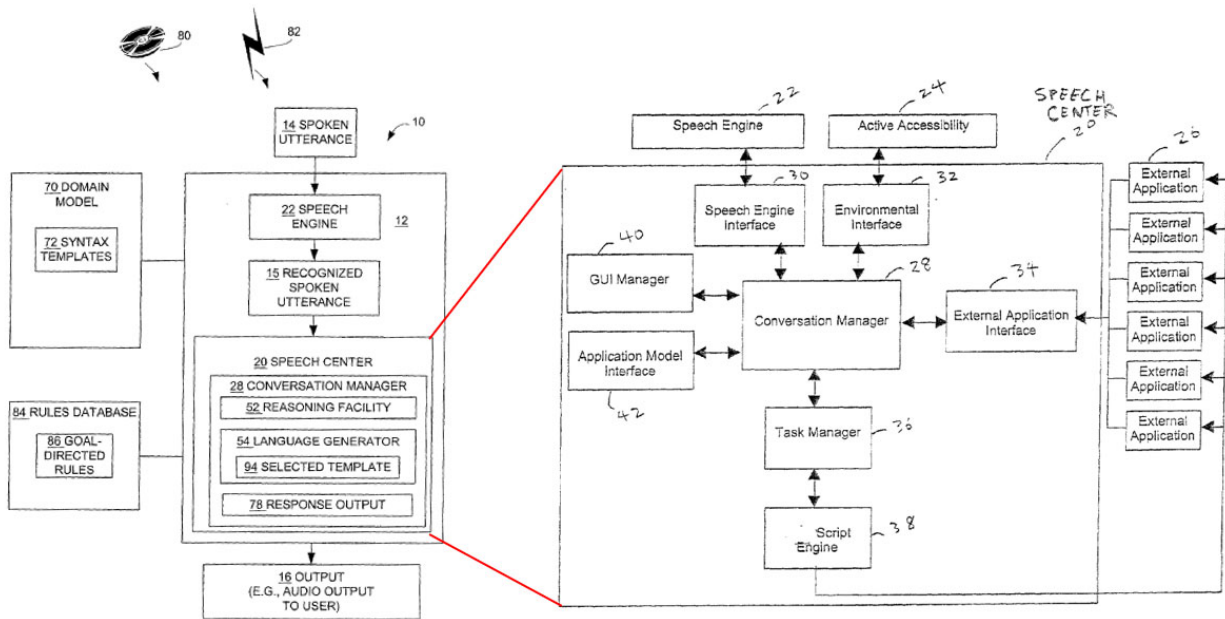
appliance speak[ing] back or perform[ing] some action ... or both.” (EX-1007, ¶107; *see also* EX-1007, ¶¶119-162.) Barbara also provides a correction technique “[w]here the utterances have failed to meet the pre-determined criteria for automatic acceptance...” (EX-1007, ¶¶100-103.) When this occurs, 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 automatically 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**

199. 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.)

200. 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, Figures 1-2 (below).) Ross’s Figure 2 “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

201. 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.)

202. 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, ¶33.)

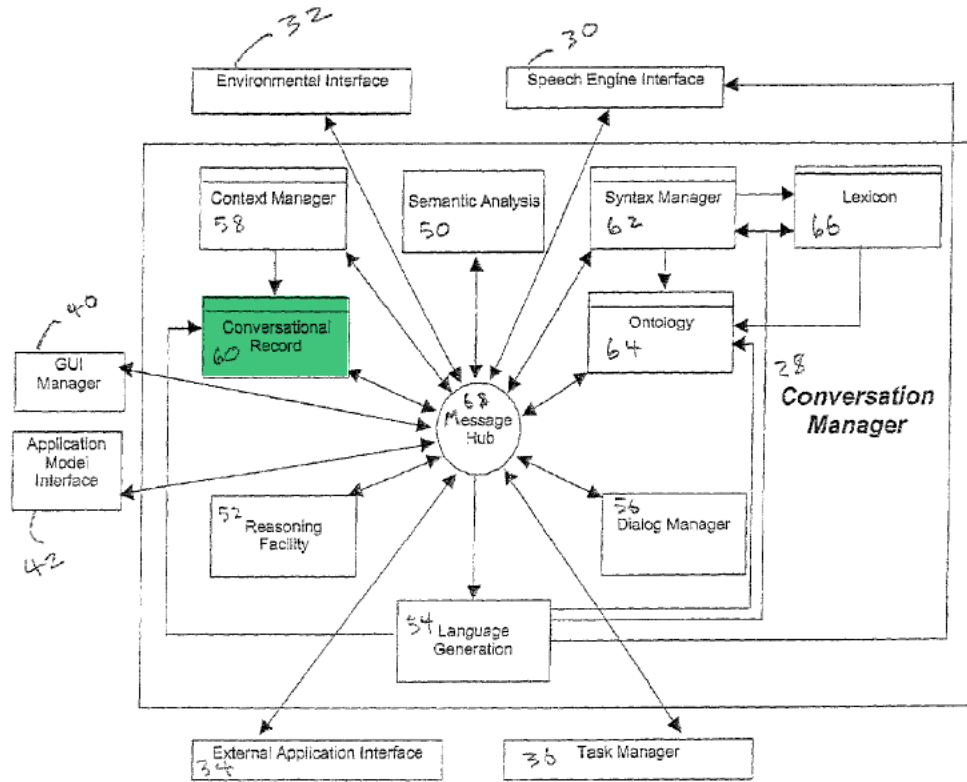


Fig. 3

Ross, Figure 3

203. Ross explains that the conversational manager 28 includes reasoning facility 52, which “performs the reasoning process for the conversation manager 28.” (EX-1008, ¶56.) 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, ¶56.)

204. 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.)

205. 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 a selected 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.)

206. 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.)

207. 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.)

208. 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.)

209. 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.)

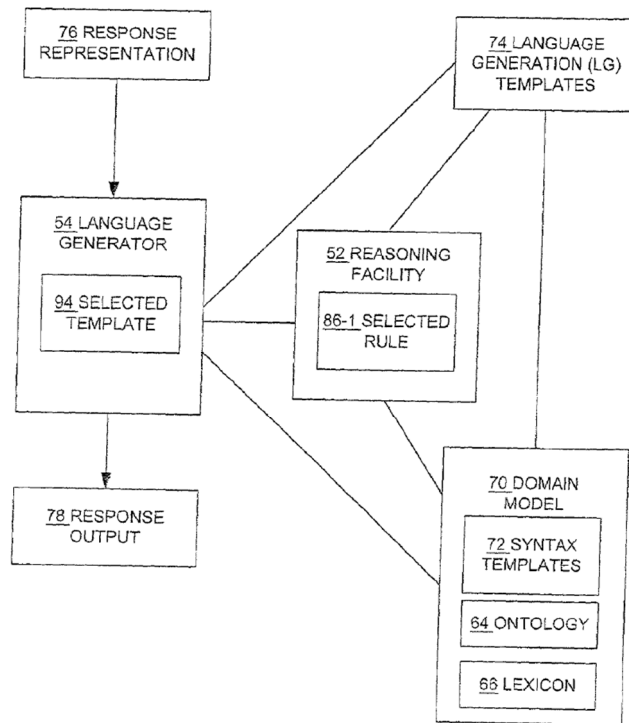


FIG. 4

**Ross, Figure 4**

210. 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.)

211. 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. Kellner**

212. Kellner discloses “a dialog system (1) which has a most comfortable and effective dialog structure for a user, comprising processing units for automatic speech recognition (3), natural language understanding (4), defining system outputs in dependence on information (7) derived from user inputs, generating acoustic and/or visual system outputs (9, 10, 11, 12), deriving user models, while the user models contain details about the style of speech of user inputs ....” (EX-1023, Abstract.)

213. In Kellner’s dialog system, “the contents and form of the system outputs **are adapted to the style of speech of user inputs and/or to the behavior**

of a user during a dialog with the dialog system.” (EX-1023, ¶5.) “The details about the style of speech and dialog interactions of a user are contained in an **associated user model** which is evaluated by the dialog system components.” (EX-1023, ¶5.) Specifically, “[w]hen user models are generated, more particularly the style of speech and interactions occurring between user and dialog system are taken into account.” (EX-1023, ¶14.)

214. “When the style of speech of a user input is determined, for example evaluations with respect to the following characterizing features are made:

- number of polite phrases used,
- address used (you),
- speech level (colloquial language, standard language, dialect)
- information density (number of words recognized as significant of a speech input in relation to the total number of words used),
- vocabulary and use of foreign words,
- number of different words in user inputs,
- classification of words of speech inputs with respect to rare occurrence.”

(EX-1023, ¶¶15-22.)

215. “Also a user's interaction behavior is incorporated in the associated user model so that, more particularly, the output modalities used by the dialog system during a dialog are used as well as possible in dependence on the use of various available input modalities ....” (EX-1023, ¶5.) Specifically, Kellner discloses that

“[i]n dependence on the determined style data, the style of speech outputs is adapted, i.e. respective polite phrases and the same address are used, the speech level is adapted to the detected speech level; in dependence on the information density in a speech input more or fewer extensive system outputs are generated and the vocabulary used for the speech outputs is selected accordingly.” (EX-1023, ¶23.)

216. As a result, “[t]he dialog system is thus in a position to generate system outputs adapted to a user's style of speech with a style of speech considered pleasant by the respective user. As a result, the inhibition threshold of the use of the dialog system can be lowered.” (EX-1023, ¶5.)

217. Kellner also discloses that “[b]oth the contents of the sentences and also the syntax of system outputs are adapted by the dialog system according to the invention, so that a dialog structure is made possible in which a dialog to be held with a user is experienced by him as pleasant and, furthermore, the dialog with the user is held highly effectively.” (EX-1023, ¶5.)

#### **4. Motivation To Combine Barbara With Ross**

218. 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.

219. Barbara and Ross are analogous art to the '699 patent. All three pertain to the same field of endeavor, i.e., speech recognition systems. (*Compare* EX-1001, 1:28-29 *with* EX-1007, ¶7 *and* EX-1008, Abstract.)

220. 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.

221. 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.” (EX-1008, ¶57.) 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.

222. 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.

223. 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 claimed priority date of the '699 patent.

## **5. Motivation To Combine Barbara And Ross With Kellner**

224. A POSITA would have also been motivated to combine Kellner's teachings of "adapt[ing] to the style of speech of user inputs and/or to the behavior of a user during a dialog with the dialog system" (EX-1023, ¶5) with Barbara's "voice interface where the end system responds to spoken natural language commands" (EX-1007, ¶82), as modified by Ross. For example, a POSITA would

have been motivated to include in Barbara's "information database" "[t]he details about the style of speech [of a user]," such as Kellner's "user model" to enable Barbara's system to be "in a position to generate system outputs adapted to a user's style of speech with a style of speech considered pleasant by the respective user." (EX-1023, ¶5.) A POSITA would have been motivated, because "[a]s a result, the inhibition threshold of the use of the dialog system can be lowered." (EX-1023, ¶5.)

225. A POSITA would have been motivated to make this combination for numerous reasons. *First*, Kellner explicitly motivates the combination teaching that adapting the style of speech output to a particular user's style of speech, creates a style of speech that is pleasant to the user. Specifically, Kellner discloses that "[i]n dependence on the determined style data [of the user], **the style of speech outputs is adapted....**" (EX-1023, ¶23.) As a result, "[t]he dialog system is thus in a position to **generate system outputs adapted to a user's style of speech with a style of speech considered pleasant by the respective user.**" (EX-1023, ¶5.) A POSITA would, therefore, have been motivated to implement Kellner's teachings of adapting the speech output to a style of a particular user in a voice interface system like Barbara's, because the modification would enhance the user experience with the system since the user will be interacting with a system that they would consider pleasant.

226. *Second*, a POSITA would have been motivated to adapt the style of speech output to a particular user's style of speech as taught by Kellner in Barbara's voice interface, because that would create a pleasant experience to a user by providing more tailored results to the user. Kellner discloses that "[t]he dialog system is thus in a position to generate system outputs adapted to a user's style of speech with a style of speech considered pleasant by the respective user. **As a result, the inhibition threshold of the use of the dialog system can be lowered.**" (EX-1023, ¶5.) A POSITA would have therefore understood Kellner to teach that any threshold that would inhibit, e.g., prevent, the use of the system, would be lowered by generating system outputs adapted to a user's style of speech. Accordingly, a POSITA would have understood that users would use the system more often and for different occasions. This is an additional motivation for the modification because the more a user interacts with the system, the system can capture more information about the user, and therefore, provide more tailored results to the user, e.g., tailored to the user's preferences and needs, which in turn would also enhance the user experience with the system.

227. *Third*, the combination is nothing more than use of a known technique (adapting an output of a dialog system to the style of speech of user inputs and/or to the behavior of a user during a dialog with the dialog system) to improve a similar

device (Barbara's voice interface system) in the same way (by adapting the output of Barbara's voice interface system to the user's style of speech and/or behavior).

228. *Fourth*, Barbara, Ross, and Kellner are analogous art to the '699 patent. All three pertain to the same field of endeavor, i.e., speech recognition systems. *Compare* EX-1001, 1:28-29 ("The invention relates to a cooperative conversational model for a human to machine voice user interface.") *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-1023, ¶1 ("The invention relates to a dialog system comprising processing units for **automatic speech recognition**, for natural language understanding, for defining system outputs in dependence on information derived from user inputs and for generating acoustic and/or visual system outputs.").

229. A POSITA would have had a reasonable expectation of success in the combination and the results of the combination would have been predictable. Combining Barbara and Ross with Kellner would merely implement Kellner's "user model" into Barbara's information database to capture the user's style of speech, such that the output of Barbara's system can be adapted according to the user's style of speech. (EX-1023, ¶¶5, 15-22). Such modification would merely consist an implementation detail on how to implement Barbara's response system and information database(s) and would have been well within the skill set of a POSITA before the earliest priority date of the '699 patent.

**B. Independent Claims 1 and 12**

230. In this section, I focus my testimony on a subset of claim limitations including the limitation amended by the Patent Owner during prosecution to overcome a rejection (EX-1002, 419) and the limitations added by Examiner's amendment to secure allowance (EX-1002, 444.) I also address aspects of the preamble and the "*identifying ... a context*" limitations and dependent claims 4, 5, 15, and 16, to provide additional analysis and context. Despite this focused testimony, it is my opinion that the combination of Barbara, Ross, and Kellner discloses all the limitations of the independent claims 1 and 12 and the dependent claims 2-11 and 13-22.

## 1. Preamble [1P] / [12P] And Limitation [12A]

[1P.1] A computer-implemented method of generating natural language system responses adapted based on a user's manner of speaking,

[1P.2] the method being implemented by a computer system that includes one or more physical processors executing one or more computer program instructions which, when executed, perform the method, the method comprising:

[12P] A system for generating natural language system responses adapted based on a user's manner of speaking, the system comprising:

[12A] one or more physical processors programmed with one or more computer program instructions which, when executed, configure the one or more physical processors to:

231. Barbara discloses a “*method*” [1P.1] and “*system*” [12P] including “*one or more physical processors [executing] [1P.2]/[programmed with] [12A] one or more computer program instructions which, when executed, configure the one or more physical processors*” recited in the claim. 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, Figures 1-5) and one directed to voice recognition (*see* EX-1007, ¶¶82-179, Figures 6-7). ) The system of Barbara’s Figure 6 (below) includes a main server (shaded red) with “**one or more service processors** having software for receiving,

interpreting and correcting the incoming information from the user terminal” [1P.2]/[12A]. (EX-1007, ¶85.)

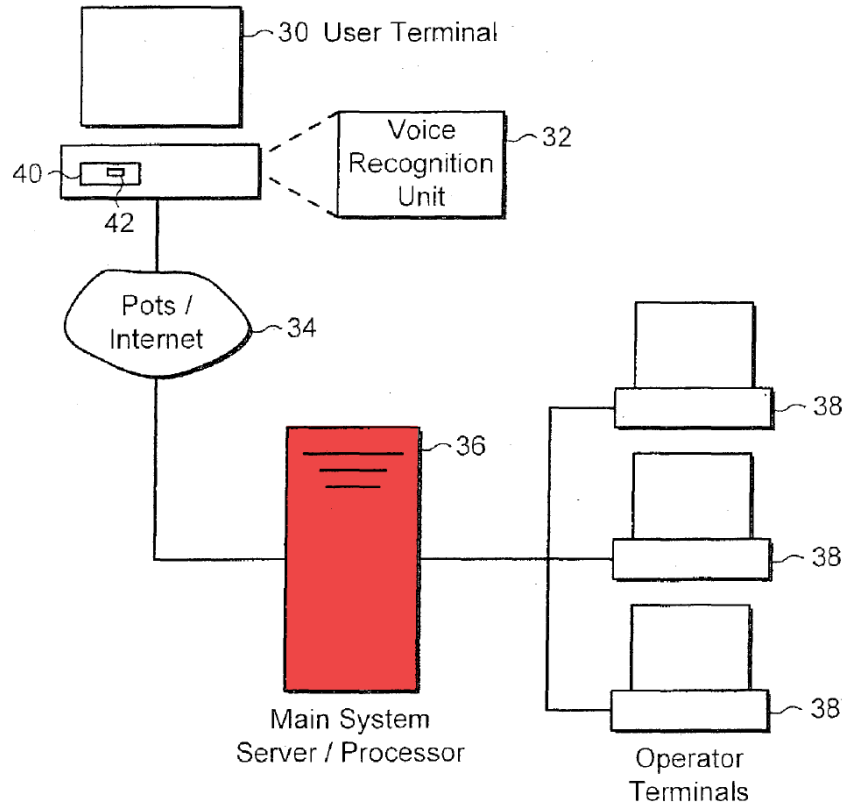


FIG. 6

**Barbara, Figure 6**

232. Barbara’s method and system “*generat[es] natural language system responses*” [1P.1]/[12P]. For example, Barbara discloses “[t]he system can ... 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.**” (EX-1007, ¶82.)

233. Kellner discloses “*generating natural language responses adapted based on a user's manner of speaking*” [1P.1]/[12P]: that “the contents and form of the system outputs **are adapted to the style of speech of user inputs** and/or to the behavior of a user during a dialog with the dialog system.” (EX-1023, ¶5.) Specifically, Kellner discloses that “[i]n dependence on the determined style data, **the style of speech outputs is adapted**, i.e. respective polite phrases and the same address are used, the speech level is adapted to the detected speech level; in dependence on the information density in a speech input more or fewer extensive system outputs are generated and the vocabulary used for the speech outputs is selected accordingly.” (EX-1023, ¶23.) [1P.1]/[12P].

234. Thus, the Barbara-Ross-Kellner combination discloses a software program executed on a processor that “*generat[es] natural language system responses adapted based on a user's manner of speaking*,” and therefore discloses the “*computer-implemented method*” and “*system*” of [12P] and [1P.1], with “*one or more physical processors programmed with one or more computer program instructions*” that—when executed—“*perform the method*” [1P.1] and “*configure*” the system [12A].

## 2. Identifying a Context [1C]/[12D]

[1C] identifying, by the computer system, a context for the natural language utterance based on the one or more words or phrases recognized from the natural language utterance;
--

[12D] identify a context for the natural language utterance based on the one or more words or phrases recognized from the natural language utterance;

235. Barbara discloses “*identifying, by the computer system, a context for the natural language utterance based on the one or more words or phrases recognized from the natural language utterance*” [1C]/[12D].

236. 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.** 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.)

237. 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, ¶0148.)

238. Moreover, for “[a] user who asks about their dog ten times a day might be considered very likely to be referring to it.” (EX-1007, ¶98.) The information database 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, or by providing flea vaccination information. Therefore in connection with a user who interfaces with the system multiple times throughout the day about a dog named “Toto,” Barbara’s system can draw an inference that the particular user has a dog named “Toto.” So when the utterance includes a word that is interpreted as “Toto,” the corresponding context may be recognized as “pets” or “canine.” (EX-1007, ¶98.) Accordingly, Barbara discloses “*identifying, by the computer system, a context for the natural language utterance based on the one or more words or phrases recognized from the natural language utterance.*”

### 3. Accumulating Short-Term Knowledge [1E]/[12F]

[1E.1] accumulating, by the computer system, short-term knowledge based on one or more natural language utterances received during a predetermined time period,

[1E.2] wherein the one or more natural language utterances received during the predetermined time period are related to a single conversation between a user and the computer system;

[12F.1] accumulate short-term knowledge based on one or more natural language utterances received during a predetermined time period,

[12F.2] wherein the one or more natural language utterances received during the predetermined time period are related to a single conversation between a user and the computer system;

239. In Barbara’s computer system, the “best guesses of the voice recognition engine as to what was said are evaluated” in the speech processing component of the server “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, ¶88.) 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.)

240. 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.)

241. 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.) That is, the “*short-term knowledge*” is accumulated in the conversational record during the duration of the conversation, e.g., when relevant to active goals, or for a predefined period of time—“*a predetermined time period.*” [1E.1]/[12F.1].

242. The Barbara-Ross-Kellner combination also discloses “*the one or more natural language utterances received during the predetermined time period are*

*related to a single conversation between a user and the computer system”*  
[1E.1]/[12F.1].

243. 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.” (EX-1007, ¶167.) 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.” (EX-1007, ¶168.) A POSITA would have understood that during the course of the diagnosis the one or more natural language utterances would be related to a “*single conversation between a user and the computer system,*” because the user’s interaction with the computer system had one goal, e.g., to narrow down the diagnosis. Accordingly, Barbara alone discloses “*the one or more natural language utterances are related to a single conversation between a user and the computer system*” [1E.1]/[12F.1].

244. Additionally, in the Barbara-Ross-Kellner system, during a predetermined time, “each utterance is indexed in the conversational record 60, along with the results of its semantic analysis” until “it is no longer relevant to active goals.” (EX-1008, ¶57.) Accordingly, a POSITA would have understood that the

received utterances that are indexed within the conversational record, would be related to a single conversation, e.g., a conversation associated with the user's current "active goals."

245. Accordingly, the Barbara-Ross-Kellner system discloses "*the one or more natural language utterances received during the predetermined time period are related to a single conversation between a user and the computer system*" [1E.2]/[12F.2].

#### 4. Accumulating Long-Term Knowledge [1F]/[12G]

[1F] accumulating, by the computer system, long-term knowledge, wherein the long-term knowledge is accumulated based on one or more natural language utterances received prior to the predetermined time period;

[12G] accumulate long-term knowledge, wherein the long-term knowledge is accumulated based on one or more natural language utterances received prior to the predetermined time period;

246. As I discussed in §V.B.3, in Barbara, the "best guesses of the voice recognition engine as to what was said are evaluated against an **information database(s)**." (EX-1007, ¶88.) Barbara's "**information database(s)** is built up over time, by recording [*"accumulating"*] **personal details or preferences from what has been said**, or from information directly entered into the system by the user." (EX-1007, ¶88.) Barbara also discloses that "[a] user who asks about their dog ten times a day might be considered very likely to be referring to it." (EX-1007, ¶98.)

As a further example, based on stored user profile data, the system determines “[s]omeone who is not involved in financial markets is unlikely to be asking about the Nikeii. The history of what people have asked is heuristically used to assist in scoring the possible intents.” (EX-1007, ¶98.) That is, Barbara implies that the system infers preference/employment information about the user. Barbara also 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-107, ¶148.) Accordingly, Barbara’s system knows when a user has asked a question multiple times, e.g., ten times in this example, and can accumulate knowledge about past conversations or utterances.

247. Thus, because this information is “built up over time,” Barbara alone discloses use of “*knowledge based on one or more natural language utterances*” and also discloses “*accumulat[ing]*” the “*knowledge based on one or more natural language utterances.*”

248. Barbara’s “information database(s)” would also hold “*long-term knowledge*” because it would include a user’s “personal details or preferences **from what has been said, or from information directly entered into the system by the user.**” (EX-1007, ¶88.) A POSITA would have understood that this user information accumulated explicitly and/or implicitly corresponds to long-term knowledge. (*See,*

EX-1001, 5:35-39 (“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.”).)

249. A POSITA would have further understood that the “*long-term knowledge based on one or more natural language utterances*” would relate to “*utterances*” that would have been “*received prior to the predetermined time period.*” As I explained above in §V.B.3, Ross’s conversational record holds “[e]ach utterance ..., along with the results of its semantic analysis” that is relevant to the user’s current conversation, e.g., “relevant to active goals” (EX-1008, ¶57.) For example, when a user engages in a conversation that introduces their dog “Toto,” the “conversational record” would hold the utterances of the current conversation and the results of their semantic analysis, for example, that the user is a dog owner or that they prefer dogs over cats. After, however, a predetermined time when the information is purged from the “conversational record,” a POSITA would have understood that, at least part of it, e.g., “personal details or preferences from what has been said” (EX-1007, ¶88) would be moved into Barbara’s informational database, because it constitutes “[t]he history of what people have asked [that is] used to assist in scoring the possible intents” ([1F]/[12G]). (EX-1007, ¶98.)

## 5. Identifying a Manner [1G]/[12H]

[1G] identifying, by the computer system, a manner in which the natural language utterance was spoken based on the short-term knowledge and the long-term knowledge; and

[12H] identify a manner in which the natural language utterance was spoken based on the short-term knowledge and the long-term knowledge; and

250. The Barbara, Ross, and Kellner combination discloses this limitation.

For example, Kellner discloses that “[w]hen the style of speech of a user input is **determined**, for example evaluations with respect to the following characterizing features are made:

number of polite phrases used,  
address used (you),  
speech level (colloquial language, standard language, dialect)  
information density (number of words recognized as significant of a speech input in relation to the total number of words used),  
vocabulary and use of foreign words,  
number of different words in user inputs,  
classification of words of speech inputs with respect to rare occurrence.”

(EX-1023, ¶¶15-22.)

251. In the Barbara, Ross, and Kellner combination, the “style of speech of a user input” (“*a manner in which the natural language utterance was spoken*”), is therefore evaluated based on different characterizing features, which include, among others, the “speech level” of the user input, e.g., one or more user’s utterances.

252. Kellner also discloses that “[t]he details about the style of speech and dialog interactions of a user are contained in an **associated user model** which is evaluated by the dialog system components.” (EX-1023, ¶5.) Specifically, “[w]hen user models are generated, more particularly the style of speech and interactions occurring between user and dialog system are taken into account.” (EX-1023, ¶14.) That is, Kellner discloses storing details about the style of speech and dialog interaction of a user in an **associated user model** akin to a user profile. As discussed in §V.A.4, Ross discloses creating a conversational record for storing utterances of a current conversation and the system’s interpretation. In the Barbara-Ross-Kellner, a POSITA would have found obvious to store “[t]he details about the style of speech and dialog interactions” (EX-1023, ¶5) of the current conversation in Ross’s conversational record—and then expire the information into the associated user model—because as explained in §V.A.4, this data management approach improves storage efficiency and reduces need for extensive hardware storage.

253. A POSITA would have understood that “*identifying [by the computer system] a manner in which the natural language utterance was spoken*” (determining the style of speech) is “*based on the short-term knowledge*” (characterizing features of the user input) and “*the long-term knowledge*” (user model) because, for example, in the case of the identifying a user “speech level,” Barbara-Ross-Kellner’s system would compare the speech level of the user input

(utterance in the current conversation) with the speech level stored within the associated user mode, so that the system can compare, for example, if the user is using colloquial language versus standard language. That is, the user model would include information about the speech levels (e.g., colloquial language, standard language, dialect) of its associated user, so that the system can compare whether the speech level of a current utterance maps closer to any of the colloquial, standard, or dialect speech levels.

**C. Dependent Claims 4, 15**

[4A] The method of claim 1, the method further comprising: obtaining, by the computer system, contextual signifiers and/or grammatical rules,

[4B] wherein generating the response based on the identified manner in which the natural language utterance was spoken comprises using the obtained contextual signifiers and/or grammatical rules to generate sentences for use as response sets to cooperate with the user.

[15A] The system of claim 12, wherein the one or more physical processors are further configured to: obtain contextual signifiers and/or grammatical rules,

[15B] wherein to generate the response based on the identified manner in which the natural language utterance was spoken, the one or more physical processors are further configured to use the obtained contextual signifiers and/or grammatical rules to generate sentences for use as response sets to cooperate with the user.

## 1. Contextual Signifiers And/Or Grammatical Rules [4A]/[15A]

254. The Barbara, Ross, Kellner combination discloses “*obtaining, by the computer system, contextual signifiers and/or grammatical rules,*” [4A]/[15A] in two ways.

255. First, Ross discloses “[a]n example of the generation of a response 8 ... shows the rule 86-1 for choosing the LG syntax template 74.” (EX-1008, ¶¶63, 77, Figure 4.)

256. Second, Kellner discloses that “[w]hen the style of speech of a user input is determined, for example evaluations with respect to the following characterizing features are made:

number of polite phrases used,  
address used (you),  
speech level (colloquial language, standard language, dialect)  
information density (number of words recognized as significant of a speech input in relation to the total number of words used),  
vocabulary and use of foreign words,  
number of different words in user inputs,  
classification of words of speech inputs with respect to rare occurrence.”

(EX-1023, ¶¶15-22.)

257. A POSITA would have understood that characterizing features, e.g., number of polite phrases used, way of addressing, speech level, which includes

dialect, are “*grammatical rules*,” because they are used to form sentences. Accordingly, Kellner discloses “*obtaining, by the computer system, contextual signifiers and/or grammatical rules*” [4A]/[15A].

## **2. Response Based On Contextual Signifiers And/Or Grammatical Rules [4B]/[15B]**

258. The Barbara, Ross, and Kellner combination also discloses “*using the obtained contextual signifiers and/or grammatical rules to generate sentences for use as response sets to cooperate with the user*,” [4B]/[15B] in two ways.

259. First, Ross discloses “[a]n example of the generation of a response 78 ... shows the rule 86-1 for choosing the LG syntax template 74.” (EX-1008, ¶¶63, 77, Figure 4.)

260. Second, Kellner discloses “[i]n dependence on the determined style data, **the style of speech outputs is adapted, i.e. respective polite phrases and the same address are used, the speech level is adapted to the detected speech level;** in dependence on the information density in a speech input more or fewer extensive system outputs are generated and the vocabulary used for the speech outputs is selected accordingly.” (EX-1023, ¶23.) That is Kellner discloses adapting the response using the “*grammatical rules*,” e.g., respective polite phrases and the same address are used, the speech level.

#### D. Dependent Claims 5, 16

5. The method of claim 1, wherein the long-term knowledge is associated with a first user, the method further comprising: generating, by the computer system, a profile associated with the first user based on the long-term knowledge, wherein the context for the natural language utterance is determined based further on the profile associated with the first user.

16. The system of claim 12, wherein the long-term knowledge is associated with a first user, and wherein the one or more physical processors are further configured to: generate a profile associated with the first user based on the long-term knowledge, wherein the context for the natural language utterance is determined based further on the profile associated with the first user.

261. Barbara renders obvious “*the long-term knowledge is associated with a first user*” and “*generating a profile associated with the first user based on the long-term knowledge.*” For example, Barbara discloses “[t]he 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.** For example, software may be provided **to allow users to verbally enter personal details, such as the names of their children or their favorite sport or their mother's telephone number etc.** Non-personal information (such as weather reports or stock prices) may be directly fed into the information database by system operators.” (EX-1007, ¶88.) Accordingly, Barbara discloses recording personal information, e.g., details and preferences, for a user. A POSITA would have found obvious to store the personal recorded information in a profile associated with each

user, so that the system can use the personal information for each user to disambiguate interpretations.

262. Barbara also renders obvious that “*the first context for the first natural language utterance is further determined based on the profile associated with the first user.*” For example, Barbara discloses an example where “the user asks for the date of Toto's last flea vaccination.” (EX-1007, ¶148.) Barbara explains that “[t]he database has flea vaccination information for Toto but not for the others,” (EX-1007, ¶148) and therefore “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.) A POSITA would have found obvious to hold the flea vaccination information for Toto in the user profile, because Toto is the user’s dog. Accordingly, Barbara discloses that the “*first context for the first natural language utterance,*” e.g., that the utterance refers to pets or canine (the score for the canine interpretation may be raised) (EX-1007, ¶148), is “*determined based on the profile associated with the first user,*” e.g., the flea vaccination information in the user’s profile.

## **VI. Conclusion**

263. 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.

264. 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: May 2, 2025

Respectfully submitted,

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

Stuart Lipoff

## APPENDIX

Exh.	Reference
1001	U.S. Patent 10,755,699 to Baldwin et al. (“the ’699 patent”)
1002	Prosecution History for U.S. Patent 10,755,699
1004	Curriculum Vitae of Stuart Lipoff
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”)
1012	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”)
1013	<i>Microsoft Computer Dictionary</i> , Microsoft Press (2002)
1014	Robert Porzel & Iryna Gurevych, <i>Contextual Coherence in Natural Language Processing</i> , 2680 Lecture Notes in Computer Science 272 (“Porzel”)
1016	Niels Ole Bernsen et al., <i>Designing Interactive Speech Systems: From First Ideas to User Testing</i> , Springer (1998) (“Bernsen”)
1017	H.P. Grice, <i>Logic and Conversation in 3</i> Syntax and Semantics 41, Academic Press (1975) (“Grice”)
1018	Xuedong Huang et al., <i>Spoken Language Processing</i> , Prentice Hall PTR (2001) (“Huang”)

Exh.	Reference
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	U.S. Publication 2002/0065651 to Kellner (“Kellner”)