

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

MICROSOFT CORPORATION

IPR2025-00658

U.S. Patent No. 8,195,468

**DECLARATION OF DR. HENRY HOUH
UNDER 37 C.F.R. § 1.68 IN SUPPORT OF
PETITION FOR *INTER PARTES* REVIEW**

TABLE OF CONTENTS

	Page
i. Speech Recognition Engine	25
ii. Semantic Knowledge-Based Model	28
iii. Creating a Speech-Based Transcription	30
iv. Domain Agent Updating the Context Stack	71
v. Domain Agent Updating the semantic knowledge-based model	73
i. The Knowledge-enhanced Speech Recognition Engine	76
ii. Determining a most likely context for the follow-up multi-modal input	76
iii. Using the updated Context Stack	78
II. INTRODUCTION	12
III. QUALIFICATIONS AND PROFESSIONAL EXPERIENCE	14
IV. LEVEL OF ORDINARY SKILL IN THE ART	19
V. RELEVANT LEGAL STANDARDS	21
VI. OVERVIEW OF THE 468 PATENT	26
A. The Disclosures of the 468 Patent	26
VII. CLAIM CONSTRUCTION	28
VIII. IDENTIFICATION OF HOW THE CLAIMS ARE UNPATENTABLE	29
IX. THE PRINCIPAL PRIOR ART	29

A.	Maes	29
B.	The Combination of Maes’s Teachings with those of Ittycheriah and Coffman	32
X.	CLAIMS 19, 25-30, AND 32 ARE UNPATENTABLE	36
A.	Ground 1: Maes renders obvious claims 19, 25-30, and 32.....	36
1.	Independent Claim 19	36
a.	[19.0] A method for processing multi-modal natural language inputs, comprising:.....	36
b.	[19.1.0] receiving a multi-modal natural language input at a conversational voice user interface, the multi-modal input including a natural language utterance and a non-speech input provided by a user,....	42
c.	[19.1.1] wherein a transcription module coupled to the conversational voice user interface transcribes the non-speech input to create a non-speech based transcription;	52

d.	[19.2] identifying the user that provided the multi-modal input;	60
e.	[19.3.0] creating a speech-based transcription of the natural language utterance using a speech recognition engine and a semantic knowledge-based model, ...	66
i.	Speech Recognition Engine	68
ii.	Semantic Knowledge-Based Model	71
iii.	Creating a Speech-Based Transcription	76
f.	[19.3.1] wherein the semantic knowledge-based model includes a personalized cognitive model derived from one or more prior interactions between the identified user and the conversational voice user interface, ...	83
g.	[19.3.2] [wherein the semantic knowledge-based model includes ...] a general cognitive model derived from one or more prior interactions between a plurality of users and the conversational voice user interface, ...	90

- h. [19.3.3] [wherein the semantic knowledge-based model includes ...] an environmental model derived from an environment of the identified user and the conversational voice user interface; 94
- i. [19.4] merging the speech-based transcription and the non-speech-based transcription to create a merged transcription; 102
- j. [19.5] identifying one or more entries in a context stack matching information contained in the merged transcription; 112
- k. [19.6] determining a most likely context for the multi-modal input based on the identified entries; 121
- l. [19.7] identifying a domain agent associated with the most likely context for the multi-modal input; 124
- m. [19.8] communicating a request to the identified domain agent; and.....127
- n. [19.9] generating a response to the user from content provided by the identified domain agent as a result of processing the request. 128

2.	Dependent Claim 25	129
a.	[25.0] The method of claim 19, further comprising verifying an identity of the user based on voiceprint matching, password matching, or pass-phrase matching.	129
3.	Dependent Claim 26	130
a.	[26.0] The method of claim 19, further comprising receiving one or more follow-up multi-modal inputs at the conversational voice user interface, the follow-up multi-modal inputs including at least one of a follow-up natural language utterance or a follow-up non-speech input.	130
4.	Dependent Claim 27	134
a.	[27.0] The method of claim 26, wherein the identified domain agent updates the context stack and the semantic knowledge-based model in response to processing the request.	134
i.	Domain Agent Updating the Context Stack	134

ii.	Domain Agent Updating the semantic knowledge-based model.....	137
5.	Dependent Claim 28	144
a.	[28.0] The method of claim 27, wherein the knowledge-enhanced speech recognition engine determines a most likely context for the follow-up multi-modal input using the updated context stack.....	144
i.	The Knowledge-enhanced Speech Recognition Engine	144
ii.	Determining a most likely context for the follow-up multi-modal input	145
iii.	Using the updated Context Stack	148
6.	Dependent Claim 29	149
a.	[29.0] The method of claim 27, wherein the speech recognition engine creates a speech-based transcription of the follow-up natural language utterance using the updated semantic knowledge-based model.	149
7.	Dependent Claim 30	152
a.	[30.0] The method of claim 19, wherein the identified domain agent processes the request by	

	querying one or more local or network information sources.	152
8.	Dependent Claim 32	153
a.	[32.0] The method of claim 19, wherein the identified domain agent processes the request by directing a command to one or more local or remote devices.	153
B.	Ground 2: Maes, Coffman and Ittycheriah Render Obvious claims 19, 25-30, and 32	154
XI.	OBJECTIVE INDICIA OF NON-OBVIOUSNESS	154
XII.	CONCLUSION.....	155
	CLAIM LISTING	156

EXHIBIT LIST

No.	Exhibit Description
1001	U.S. Patent No. 8,195,468
1002	File History of U.S. Patent No. 8,195,468
1003	Declaration of Dr. Henry Houh
1004	CV of Dr. Henry Houh
1005	U.S. Patent No. 6,964,023 (“ <u>Maes</u> ”)
1006	U.S. Patent No. 5,937,383 (“ <u>Ittycheriah</u> ”)
1007	Second Amended Complaint
1008	U.S. Patent Application Publication No. 2002/0087315 (“ <u>Lee</u> ”)
1009	“A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Lawrence Rabiner, 1989 (“ <u>Rabiner</u> ”)
1010	Visser, I. (2011). “Seven things to remember about hidden Markov models: a tutorial on Markovian models for time series.” <i>J. Math. Psychol.</i> 55, 403–415. (“ <u>Visser</u> ”)
1011	W3C Working Draft 16 November 2001, “Semantic Interpretation for Speech Recognition”
1012	International Patent Application Publication No. WO 00/20962 (“ <u>Coffman</u> ”) filed as PCT/US99/22927
1013	U.S. Patent No. 7,137,126 (“ <u>Coffman126</u> ”)
1014	“Use and Acquisition of Semantic Language Model,” K. Wang et al., NAACL/HLT (Short Paper), Published by Association for Computational Linguistics, January 2004 (“ <u>Wang</u> ”)
1015	U.S. Patent Application Publication No. 20030055655 (“ <u>Suominen</u> ”)
1016	U.S. Patent Application Publication No. 20020198714 (“ <u>Zhou</u> ”)

No.	Exhibit Description
1017	Excerpts from Dafydd Gibbon et al., “Handbook of Multimodal and Spoken Dialogue Systems : Resources, Terminology, and Product Evaluation” (2000) (“ <u>Gibbon</u> ”)
1018	U.S. Patent No. 5,799,276 (“ <u>Komissarchik</u> ”)
1019	Excerpts from McGraw-Hill Dictionary of Scientific Terms (2003)
1020	RESERVED
1021	RESERVED
1022	RESERVED
1023	RESERVED
1024	RESERVED
1025	U.S. Patent No. 6,453,315 (“ <u>Weissman</u> ”)
1026	RESERVED
1027	RESERVED
1028	Shdaifat, I., Grigat, R.R. and Lütgert, S., 2001. Viseme recognition using multiple feature matching. In <i>INTERSPEECH</i> (pp. 2431-2434) (“ <u>Shdaifat</u> ”)
1029	EDTX Calendar, Judge Gilstrap
1030	U.S. Patent No. 6,510,415 (<u>Talmor</u>)
1031	PDF of webpage (https://www.speechtechmag.com/Articles/Editorial/Features/IBM-Announces-ViaVoice-98-29469.aspx)
1032	RESERVED

TABLE OF ABBREVIATIONS AND CONVENTIONS

Abbreviation	Meaning
468 Patent	Ex.1001: U.S. Patent No. 8,195,468
IPR	<i>inter partes</i> review
Petitioner	Bank of America, N.A. (“BofA”)
Patent Owner or PO	Dialect, LLC (“Dialect”)
Second Amended Complaint	Compl: Ex.1007
<i>xx:yy–zz</i>	column <i>xx</i> , lines <i>yy</i> to <i>zz</i>

1. I, Henry Houh, declare as follows:

II. INTRODUCTION

2. I am making this declaration in support of Petitioner’s request for an *Inter Partes* Review of U.S. Patent No. 8,195,468 (“the 468 Patent”) to Weider.

3. I am being compensated for my work in this matter at my standard hourly rate. I am also being reimbursed for reasonable and customary expenses associated with my work and testimony in this proceeding. My compensation is not contingent on the outcome of this matter or the specifics of my testimony.

4. I have been asked to provide my opinions regarding whether the subject matter of claims 19, 25-30, and 32 (“the Challenged Claims”) of the 468 Patent are anticipated or otherwise would have been obvious to a person having ordinary skill in the art (“POSITA”) at the time of the alleged invention, in light of the prior art. It is my opinion that the Challenged Claims would have been obvious to a POSITA.

5. In the preparation of this declaration, I have studied:

- the 468 Patent, Ex.1001
- the prosecution history of the 468 Patent (“468 File History”), Ex.1002
- U.S. Patent No. 6,964,023 (“Maes”), Ex.1005
- International Patent Application Publication No. WO 00/20962 (“Coffman”), Ex.1012

- U.S. Pat. No. 5,937,383 (“Ittycheriah”), Ex.1006
- Second Amended Complaint, Ex.1007
- the additional references I refer to below and/or identified in the exhibit list.

6. In forming the opinions expressed below, I have considered: the documents listed above; the relevant legal standards, including the standard for obviousness; and my own knowledge and experience based upon my work in the field of multimodal and conversational networking as described below.

7. Unless otherwise noted, all **emphasis** in any quoted material has been added. Claim terms are italicized.

8. Since I was retained, I have been in frequent contact with counsel while working on this matter to communicate my opinions. My opinions are presented both in this Declaration and in the Petition for Inter Partes Review, which it supports. In many, if not most, cases the language used to express my opinions in both documents is substantially the same. This is because I worked with counsel to capture my opinions, and generally speaking we captured them in the Petition itself as opposed to developing both the Petition and Declaration in tandem. My opinions, as reflected in the Petition we prepared, were then incorporated into my Declaration. I make this statement to explain the work process and underscore that the opinions in this Declaration are my own, were formed by me and were documented with the

assistance of counsel in the manner described. As detailed herein, it is my opinion that each of the challenged claims is rendered obvious by prior art references that predate the earliest priority date of the 468 Patent. If requested, I am prepared to testify about my opinions expressed herein.

III. QUALIFICATIONS AND PROFESSIONAL EXPERIENCE

9. I am over the age of 18 and am competent to write this declaration. I have personal knowledge, or have developed knowledge, of these technologies based upon education, training, or experience, of the matters set forth herein.

10. My complete qualifications and professional experience are described in my Curriculum Vitae, a copy of which can be found in Ex.1004. The following is a brief summary of my relevant qualifications and professional experience.

11. I received a Ph.D. in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology (“MIT”) in 1998. Beforehand, I received a Master of Science degree in Electrical Engineering and Computer Science in 1991, a Bachelor of Science degree in Electrical Engineering and Computer Science in 1989, and a Bachelor of Science degree in Physics in 1990, all from MIT.

12. I am currently self-employed as an independent technical consultant. Until 2022, I was also president of a company that provides supplemental science, technology, engineering, and mathematics (“STEM”) education to children of all ages.

13. I first entered telecommunications in 1987 when I worked as a summer intern at AT&T Bell Laboratories as part of a five-year dual degree program at MIT. I continued to work at AT&T Bell Laboratories as part of this MIT program. While at MIT, I was a teaching assistant (“TA”) in the Electrical Engineering and Computer Science Department’s core Computer Architectures course. I first was a TA as a senior for a role typically reserved for graduate students. I later became head TA. The course covered various topics in computer architectures. As a TA, I helped write homework assignments, lab assignments, and exams. I also taught in the recitation sections.

14. Later, as part of my doctoral research at MIT from 1991-1998, I was a research assistant in the Telemedia Network Systems (“TNS”) group at the Laboratory for Computer Science. The TNS group built a high-speed gigabit network and created applications that ran over the network. Example applications included ones for remote video capture, processing, and display of video on computer terminals. In addition to working on the design of core network components, designing and building the high-speed links, and designing and writing the device drivers for the interface cards, I also set up the group’s web server.

15. I also helped to build the web pages that initiated the above- mentioned video sessions via a web interface. Vice President Al Gore visited our group in 1996 and received a demonstration of—and remotely drove—a radio- controlled toy car

with a wireless video camera mounted on it that was built by our group. This toy car device received commands transmitted over a network from a remote computer, and video data from the toy car was transmitted wirelessly then over a computer network back to the user controller. On occasion, we allowed users visiting our web site to drive the toy car from their remote computer while they watched the video on their computer. The video stream was encoded by TNS-designed hardware, streamed over the TNS-designed network, and displayed using TNS-designed software.

16. I defended and submitted my Ph.D. thesis, titled “Designing Networks for Tomorrow’s Traffic,” in January 1998. As part of my thesis research, I analyzed local area and wide area flows to show a more efficient method for routing packets in a network, based on traffic patterns at the time.

17. From 1997 to 1999, I was a Senior Scientist and Engineer at NBX Corporation, a start-up that made business telephone systems for streaming packetized audio over data networks instead of using traditional telephone lines. NBX was later acquired by 3Com Corporation and the phone system is still used today by numerous businesses.

18. As part of my work at NBX, I designed the core audio reconstruction algorithms for the telephones, as well as the packet transmission algorithms. I also designed and validated the core packet transport protocol used by the phone system.

The protocol was used for all signaling in the phone system, including for the setup of conference calls.

19. The NBX system also featured a computer interface for initiating phone calls, which could also initiate conference calls. The NBX system also supported the Telephony Application Programming Interface (“TAPI”) that allowed other computer programs to integrate with our system telephony features. We obtained U.S. Patent No. 6,967,963, entitled “Telecommunication method for ensuring on-time delivery of packets containing time-sensitive data,” as part of this work. During my time at NBX, I also evaluated and purchased VPN equipment in order to demonstrate that our system worked over long distances through VPN devices, so that our phones could be located remotely. We used the VPN equipment to connect from various trade shows, for example, in Las Vegas and Los Angeles, remotely into our home network so that phones in the remote network were able to connect to our home facility in Andover, MA.

20. From 1999-2004, I was employed by Empirix or its predecessor company, Teradyne. Empirix was a leader in test tools for telecommunications protocols and systems, providing functional testing tools as well as load testing tools. From 2000-2001, I conceived and built a test platform for testing Voice- over-IP (VoIP). The first application on this new test platform was a cloud emulator for simulating the effects of transmitting VoIP over a busy network.

21. In 2006, as part of my role at BBN Technologies, I helped found PodZinger Inc., now known as RAMP Inc. PodZinger utilized BBN’s speech recognition algorithms to search through the spoken words in audio and video segments. While I was Vice President of Operations and Technology, PodZinger followed its initial prototype with a full streaming audio and video search solution. I also created a social networking web site, which BBN sold to a venture-funded startup company. In the process of creating the web site, I designed and specified the authentication and authorization protocols.

22. I have been awarded several United States patents, and I have several patent applications pending including the following examples:

- U.S. Patent No. 7,975,296, “Automated security threat testing of web pages”;
- U.S. Patent No. 7,877,736, “Computer language interpretation and optimization for server testing”;
- U.S. Patent No. 7,801,910, “Method and apparatus for timed tagging of media content”;
- U.S. Patent 7,590,542, “Method of generating test scripts using a voice-capable markup language”;
- U.S. Patent No. 6,967,963, “Telecommunication method for ensuring on-time delivery of packets containing time-sensitive data”;

- U.S. Patent Application Publication No. 20070106685, “Method and apparatus for updating speech recognition databases and reindexing audio and video content using the same”;
- U.S. Patent Application Publication No. 20070106693, “Methods and apparatus for providing virtual media channels based on media search”;
- U.S. Patent Application Publication No. 20070106760, “Methods and apparatus for dynamic presentation of advertising, factual, and informational content using enhanced metadata in search-driven media applications”;
- U.S. Patent Application Publication No. 20070112837, “Method and apparatus for timed tagging of media content”;
- U.S. Patent Application Publication No. 20070118873, “Methods and apparatus for merging media content”; and
- U.S. Patent Application Publication No. 20090222442, “User-directed navigation of multimedia search results.”

IV. LEVEL OF ORDINARY SKILL IN THE ART

23. My opinions are provided based on what a person of ordinary skill in the art (“POSITA”) in the technical field of the invention would have understood at the time of the purported invention of the 468 Patent. I have been asked to assume the date of the purported invention for the 468 patent is August 29, 2005.

24. I understand there are multiple factors relevant to determining the level of ordinary skill in the pertinent art, including (1) the levels of education and experience of persons working in the field at the time of the invention; (2) the sophistication of the technology; (3) the types of problems encountered in the field; and (4) the prior art solutions to those problems.

25. Based on the materials and information I have reviewed, and on my experience in the technical areas relevant to the Asserted Patents, a POSITA with respect to the Asserted Patents as of the time of the invention, i.e., the earliest possible priority date of the 468 Patent, August 29, 2005, would have had a bachelor's degree in electrical engineering, computer science, computer engineering, or an equivalent, and two years of relevant experience involving computer science fundamentals, including natural language processing, speech recognition and transcription, non-speech recognition and transcription that is pertinent to the 468 Patent. Lack of professional experience can be remedied by additional education, and vice versa.

26. For purposes of this Declaration, in general, and unless otherwise noted, my statements and opinions, such as those regarding my own experience and what a POSITA would have understood or known generally (and specifically related to the references I consulted herein), reflect the knowledge that existed in the relevant field as of the priority date of the 468 Patent.

V. RELEVANT LEGAL STANDARDS

27. I am not an attorney. In preparing and expressing my opinions and considering the subject matter of the 468 Patent, I am relying on certain basic legal principles that counsel has explained to me.

28. I understand that prior art to the 468 Patent includes patents and printed publications in the relevant art that predate the priority date of the 468 Patent. For purposes of this Declaration, I am applying August 29, 2005, as the priority date of the 468 Patent, but I take no position on that issue.

29. I understand and have been informed that in order for an inventor to be entitled to a patent, the invention must be “new.” There are a number of ways that an invention can be “anticipated,” that is, not new.

30. A patent claim is anticipated if each limitation of the patent claim is found either expressly or inherently in a single item of prior art. While a prior art reference need not use the same words as a patent claim to anticipate the claim, the prior art reference must describe the requirements of the claim with sufficient clarity such that POSITA would have been able to make and use the claimed invention based on the reference and his or her knowledge in the applicable technical field.

31. The disclosure of a feature of a claimed invention can be either “express,” meaning that the text, figures, or other content of a reference actually teaches the aspects of the patent claim, or the disclosure can be “inherent.” An

express disclosure does not need to use the same words as the patent claim being considered, and does not need to be depicted in exactly the same way, but the disclosure must actually convey the claimed invention to POSITA using some combination of words, figures, data, or other portions of the patent. I have been informed and understand that in order to establish that an element of a claim is “inherent” in the disclosure in a prior art reference, it must be clear to a person skilled in the art that the missing element is an inevitable part of what is explicitly described in the reference, and that it would have been recognized as necessarily present by a person skilled in the art. Inherency cannot be established by probabilities or possibilities, and the mere fact that something may result from a given set of circumstances is not enough to establish inherency.

32. I further understand that to anticipate a claim, the elements of the claim must be found in the prior art reference “arranged as in the claim.” I have been informed that this means that the elements of the prior art reference may not be recombined in a way that is not taught in the prior art to arrive at the invention; instead, the prior art must teach the claimed subject matter in the arrangement required by the claims.

33. I understand that if a reference incorporates other documents by reference, the incorporating reference and the incorporated reference(s) should be treated as a single prior art reference for purposes of analyzing anticipation.

34. I understand that it is acceptable to consider evidence other than the information in a particular prior art document to determine if a feature is necessarily present in or inherently described by that reference.

35. I have been informed and understand that subject matter claimed in a patent is obvious under 35 U.S.C. § 103 if a person of ordinary skill in the art at the time the alleged invention was made would have had reason to combine or modify the disclosures of one or more prior art references to arrive at the claimed subject matter. A claim may be unpatentable even if each claim limitation is not present or disclosed in a single prior-art item.

36. I have been informed and understand that, under the doctrine of obviousness, a claim is unpatentable if the differences between the invention and the prior art are such that the subject matter as a whole would have been obvious at the time the invention was made to a person having ordinary skill in the art to which the subject matter pertains. A person of ordinary skill in the art is presumed to have knowledge of the relevant prior art at the time of the claimed invention.

37. I have been informed and understand that obviousness is based on the scope and content of the prior art, the differences between the prior art and the claim, the level of ordinary skill in the art and secondary indicia of obviousness and nonobviousness to the extent such indicia exist. I have been advised that the following evidence, when present, must be considered before forming a conclusion

that a claimed invention is obvious: (1) the invention’s commercial success, (2) long felt but unresolved needs, (3) the failure of others, (4) skepticism by experts, (5) praise by others, (6) teaching away by others, (7) recognition of a problem, and (8) copying of the invention by competitors. In this instance, however, I am not aware of any evidence of any of these factors. I will of course consider any evidence on these issues which is presented to me.

38. I have been informed and understand the determination of whether the asserted claims would have been obvious to a person of ordinary skill in the art and, therefore, invalid, is not governed by any rigid test or formula. A determination that a claim is obvious is, instead, based on a common-sense determination that the claimed invention is merely a combination of known limitations to achieve predictable results. Any of the following rationales are acceptable justifications to conclude that a claim would have been obvious: (1) the claimed invention is simply a combination of known prior art methods to yield predictable results; (2) the claimed invention is a simple substitution of one known element for another to obtain predictable results; (3) the claimed invention uses known techniques to improve similar devices (methods, or products) in the same way; (4) the claimed invention applies a known technique to a known device (method, or product) ready for improvement to yield predictable results; (5) the claimed invention was “obvious to try”—choosing from a finite number of identified, predictable solutions, with a

reasonable expectation of success; (6) there is known work in one field of endeavor that may prompt variations of it for use in either the same field or a different one based on design incentives or other market forces if the variations would have been predictable to one of ordinary skill in the art; or, (7) there is some teaching, suggestion, or motivation in the prior art that would have led one of ordinary skill in the art to modify the prior art reference to combine prior art teachings to arrive at the claimed inventions.

39. I have been informed and understand that an analysis of whether a claimed invention is obvious must not rely on a hindsight combination of prior art. The analysis must proceed in the context of the time of the invention or claimed priority date and consider whether the invention as a whole would have been obvious to a person of ordinary skill in the art, taking into consideration any interrelated teachings of the prior art, the effects of demands known to the design community or present in the marketplace, and the background knowledge possessed by a person having ordinary skill in the art, all in order to determine whether there was an apparent reason to combine any known elements in the fashion claimed by the patent at issue.

VI. OVERVIEW OF THE 468 PATENT

40. The 468 Patent was based on U.S. Patent Appl. No. 13/084,197 filed on April 11, 2011, and claims priority to U.S. Patent Appl. No. 11/212,693, filed on August 29, 2005 (now U.S. Patent No. 7,949,529). Ex.1001, Cover.

41. I have been asked to assume that it has an effective filing date of August 29, 2005.

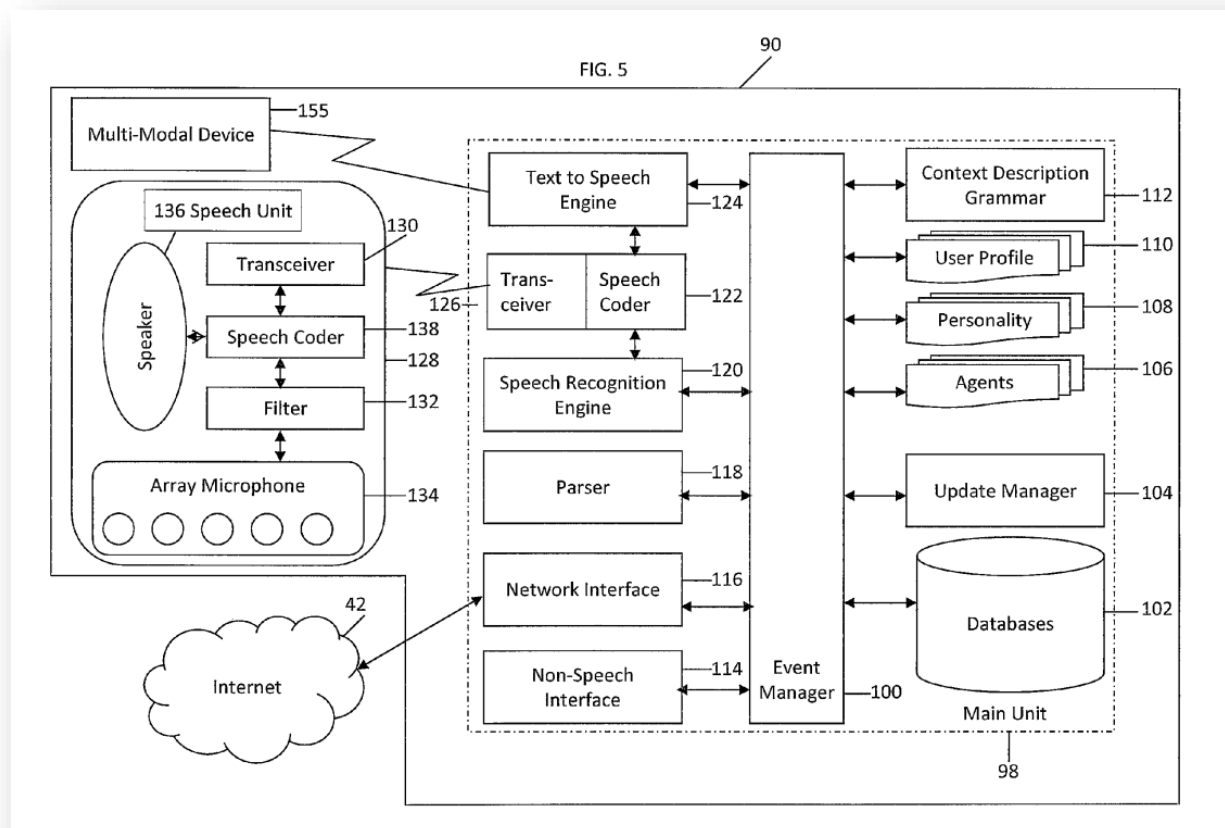
A. The Disclosures of the 468 Patent

42. The 468 Patent is directed to a system that “includes speech-based and non-speech-based interfaces for telematics applications,” and “identifies and uses context, prior information, domain knowledge, and user specific profile data to achieve a natural environment for users that submit requests and/or commands in multiple domains.” Ex.1001, Abstract. The 468 Patent explains that it “provides mobile devices having a speech interface and/or a combination of a speech interface and a non-speech interface to enable natural language human-machine interactions” and “enables mobile users to submit natural language speech and/or non-speech questions or commands in a wide range of domains.” Ex.1001, 1:17–22.

43. The 468 Patent notes that “[s]peech recognition has ... improved in accuracy and ... used in a wide range of applications,” and that “[n]atural language processing has been applied to the parsing of speech queries.” Ex.1001, 1:62–1:66. The 468 Patent also states that there apparently “remain a number of significant

barriers to creation of a complete natural language verbal and/or textual-based query and response environment.” Ex.1001, 2:3–5. The 468 Patent purportedly overcomes the “drawbacks” of prior art systems. Ex.1001, 2:54–55.

44. Figure 5 of the 468 Patent is characterized as “an overall diagrammatic view of the interactive natural language processing system according to an embodiment of the invention.” Ex.1001, 15:38–40.



Ex.1001, Figure 5. The system shown in Figure 5 comprises speech unit ..., speech recognition engine ..., context description grammar module ..., parser ..., agents

..., multi-modal device ... and non-speech interface Speech unit ... includes a microphone to receive a spoken utterance from a user. Ex.1001, 25:48–52.

45. After initial processing, the speech is passed to speech recognition engine for processing and subsequently, “[a]ny recognized information may be processed by the parser ..., which transforms information into complete algorithms and questions using data supplied by knowledge agents” which further process the commands or questions and return results to the user. *See generally* Ex.1001, 26:18–37. Multimodal support may be provided to maintain the context during both voice interaction and interaction through the non-speech interface. *See generally* Ex.1001, 27:33–41.

VII. CLAIM CONSTRUCTION

46. It is my understanding that in order to properly evaluate Challenged Claims, the terms of the claims must first be interpreted. It is my understanding that for the purposes of this inter partes review, the claims are to be construed under the so-called *Phillips* standard, under which claim terms are given their ordinary and customary meaning as would have been understood by a POSITA in light of the specification and prosecution history, unless the inventor has set forth a special meaning for a term. I have also been informed that claim terms only need to be construed to the extent necessary to resolve the unpatentability inquiry.

VIII. IDENTIFICATION OF HOW THE CLAIMS ARE UNPATENTABLE

47. The discussion in this Declaration provides a detailed analysis of how the asserted prior art teaches each limitation of the Challenged Claims.

48. As part of my analysis, I have considered, and discuss in detail, the scope and content of the prior art and any differences between the alleged invention and the prior art.

49. It is my opinion that the alleged invention recited in the Challenged Claims would have been obvious in view of the teachings of the asserted prior art and the knowledge of a POSITA, as I discuss in detail below.

IX. THE PRINCIPAL PRIOR ART

A. Maes

50. U.S. Patent No. 6,964,023 to Maes et al. (Ex.1005) was filed on February 5, 2001, and issued on November 8, 2005. I understand that it is prior art to the 468 Patent.

51. Maes explains that although “multi-modal systems would appear to have inherent advantages over systems that use only one data input mode, the existing multi-modal techniques fall significantly short of providing an effective conversational environment between the user and the computing system with which the user wishes to interact.” Ex.1005, 1:60–63. Maes thus describes “systems and methods are provided for performing focus detection, referential ambiguity resolution and mood classification in accordance with multi-modal input data, in

varying operating conditions, in order to provide an effective conversational computing environment for one or more users.” Ex.1005, Abstract.

52. In particular, Maes explains that “the multi-modal conversational computing system 10 of the present invention receives multi-modal input in the form of audio input data, video input data, as well as other types of input data (in accordance with the I/O subsystem 12), processes the multi-modal data (in accordance with the I/O manager 14), and performs various recognition tasks (e.g., speech recognition, speaker recognition, gesture recognition, lip reading, face recognition, etc., in accordance with the recognition engines 16), if necessary, using this processed data. The results of the recognition tasks and/or the processed data, itself, is then used to perform one or more conversational computing tasks, e.g., focus detection, referential ambiguity resolution, and mood classification (in accordance with the dialog manager 18, the context stack 20 and/or the classifier 22).” Ex.1005, 4:7–22.

53. Moreover, Maes expressly incorporates by reference Coffman and Ittycheriah in their entireties. In particular, Maes states that “[i]t is to be appreciated the conversational virtual machine disclosed in PCT patent application identified as PCT/US99/22927 filed on Oct. 1, 1999 and entitled “Conversational Computing Via Conversational Virtual Machine,” [*i.e.*, Coffman] the disclosure of which is incorporated by reference herein, may be employed to provide a framework for the

I/O manager, recognition engines, dialog manager and context stack of the invention.” Ex.1005, 9:30–37. Maes also states “another method to detect incorrectly recognized words is disclosed in U.S. Pat. No. 5,937,383 to Ittycheria entitled “Apparatus and Methods for Speech Recognition Including Individual or Speaker Class Dependent Decoding History Caches for Fast Word Acceptance or Rejection,” the disclosure of which is incorporated herein by reference.” Ex.1005, 19:43–49.

54. Coffman and Ittycheriah are, therefore, considered effectively part of Maes.

55. As the entire disclosure of Coffman is directed to “Conversational Computing Via Conversational Virtual Machine,” a POSITA would have understood Maes to be incorporating with detailed particularity the entire disclosure of Coffman. At a minimum, the disclosures related to the “conversational virtual machine” that “provide a framework for the I/O manager, recognition engines, dialog manager and context stack,” are expressly incorporated with particularity, which is the material relied upon below. Similarly, the entire disclosure of Ittycheriah is also expressly incorporated with particularity given its focus on “detect[ing] incorrectly recognized words”—the focus of the entire disclosure of Ittycheriah—the material of which is relied upon below.

B. The Combination of Maes’s Teachings with those of Ittycheriah and Coffman

56. To the extent one asserts that Coffman and Ittycheriah encompass different embodiments from that of Maes, a POSITA would have been motivated to include Ittycheriah’s teachings of maintaining a decoding history cache database storing keywords (vocabulary such as names and/or commands) frequently used by a user or a class of speakers, and score information generated from decoded keywords in accordance with Hidden Markov Models (HMMs) (Ex.1006, 1:55–58, 4:19–32, 4:54–63, and 5:5–19) in Maes’s semantic module for the reasons below. Further, as also explained below, a POSITA would have been additionally motivated to combine Coffman’s teachings related to searching a context stack to find entries that match transcribed user input and merging/combining speech and non-speech (e.g., keyboard input) transcriptions (Ex.1012, 41:2–5, 42:1–3, and 38:1–19) with Maes’s teachings of processing multi-modal inputs.

57. For example, as explained above, Maes expressly incorporates by reference Coffman and Ittycheriah in their entireties. In addition, all three references name Stephane Maes as a common inventor, and have IBM as the common assignee. Ex.1005, Abstract, 9:30–37 and 19:43–49; Ex.1012, Cover, Abstract; and Ex.1006, Cover, Abstract. I understand that for these reasons alone, a POSITA would have

been motivated to combine the teachings of Coffman and Ittycheriah with those of Maes.

58. Additionally, a POSITA would have found Ittycheriah provides a motivation to combine its teachings with Maes. For example, a POSITA would have found Ittycheriah's teachings of cache database (“*cognitive model[s]*”) storing words/keywords (vocabulary such as names and/or commands) frequently used by a user or a class of users (Ex.1006, 1:55–58, 4:54–63, and 5:5–19) would reduce the amount of hypotheses (e.g., limiting the set of possible hypotheses) to consider in Maes's decoding process, thereby enabling faster matching/decoding of keywords, and improving the responsiveness of Maes's system, which is consistent with Maes's goal of faster executions. Ex.1005, 11:64–12:30, 5:29–34; Ex.1006, 2:13–19, 2:58–3:21, and 1:37–43.

59. A POSITA would recognize that using HMMs for decoding words, as Ittycheriah teaches (*see* Ex.1006, 4:19–32), were well-known and utilized in language understanding and determined pertinent for analyzing cognitive processes such as in speech recognition models at the time of the 468 Patent. *See*, Ex.1008, [0013] (“**The multiple language models ... may be hidden Markov language models that are domain specific and at different levels of specificity.** [HMMs] are described generally in ... “Robustness In Automatic Speech Recognition”, Jean Claude Junqua et al., Kluwer Academic Publishers, Norwell, Mass., 1996, pages 90-

102.”); *see also* Ex.1009, 258 (“Neither **the theory of [HMMs] nor its applications to speech recognition is new**. The basic theory was published in a series of classic papers by Baum and his colleagues [] in the late 1960s and early 1970s and was implemented for speech processing applications by Baker ... at CMU, and by Jelinek and his colleagues at IBM [] in the 1970s.”)

60. Furthermore, Visser identifies HMMs as “especially suitable for studying psychological or cognitive processes in which qualitatively different cognitive states unfold over time” and therefore, considers HMMs as “the model of choice for analyzing cognitive processes based on time serial data.” Ex.1010, 403 and 413. Thus, these references reflect the POSITA’s knowledge as it relates to popularity of HMMs and their benefits to speech recognition techniques at the time of the claimed invention in support of Ittycheriah’s teachings.

61. Furthermore, Ittycheriah discloses an example of a voice-controlled navigation application and Maes discloses an example implementation of an in-vehicle system. Ex.1006, 2:49–53 and 4:54–5:4; Ex.1005, 6:15–28, 6:43–50, 7:60–8:4 and 8:37–42 . A POSITA would be further motivated to combine the teachings of Maes and Ittycheriah because Ittycheriah’s voice-controlled navigation would significantly benefit Maes’s in-vehicle system by providing voice-based navigation assistance to drivers. *Id.*

62. In addition, a POSITA would have been further motivated to combine Coffman's teachings of organizing its context stack and updating it such that the most-recent context is at the top of the context with the teachings of Maes and Ittycheriah because to do so would benefit Maes's system by providing efficient identification and retrieval of the most-recent context, allowing for faster executions and improving the responsiveness of Maes's system, and would be inline with Ittycheriah's goals of quickly accepting or rejecting keywords. Ex.1012, 21, 23–28, 41:1–5, 40:15–19, and 42:7–9; Ex.1005, 7:60–8:4, 8:37–42, 37:51–62, and 5:29–34; Ex.1006, 2:58–60.

63. Coffman, Maes and Ittycheriah are analogous art as the 468 Patent because each are directed to functionality of natural language processing systems and are reasonably pertinent to the problem addressed in the 468 Patent. Ex.1001, 1:16–24; Ex.1005, 6:15–28 and 6:43–50; Ex.1006, 1:55–58 and 2:13–19; Ex.1012, 1–4 and 29:1–4. Furthermore, each of Maes and Coffman expressly disclose the use of a context stack and notably, Maes's context stack disclosures are almost verbatim as Coffman's. Ex.1012, 21, 23–28 and Ex.1005, 37:51–62. Furthermore, each of Maes and Coffman teach merging/combining speech and non-speech-based transcriptions or otherwise synchronize information streams. Ex.1005, 20:5–7, 11:55–63, 21:51–54, 29:30–36; Ex.1012, 38:2–19.

64. A POSITA would have had a reasonable expectation of success in combining the teachings of these references because to do so would be a simple combination of analogous teachings that would be well within the skill of a POSITA. For example, modifying Maes's semantic module to include Ittycheriah's teachings of a cache database storing scores and words decoded in accordance with HMMs from previous interactions with a user or a class of users and to further modify Maes's system to include Coffman's teachings related to searching a context stack to find entries that match transcribed user input and merging/combining speech and non-speech-based transcriptions would add to the functionality of Maes's system and would have been a simple implementation and predictably resulted without undue experimentation a recognition system with improved accuracy, as Maes seeks to implement (Ex.1005, 11:62–63).

X. CLAIMS 19, 25-30, AND 32 ARE UNPATENTABLE

A. Ground 1: Maes renders obvious claims 19, 25-30, and 32

65. Claims 19, 25-30, and 32 of the 468 Patent are rendered obvious over Maes.

1. Independent Claim 19

a. [19.0] A method for processing multi-modal natural language inputs, comprising:

66. Maes teaches or suggests this claim element.

67. Maes describes “[s]ystems and methods are provided for performing focus detection, referential ambiguity resolution and mood classification **in accordance with multi-modal input data.**” Ex.1005, Abstract (emphasis added⁶). Maes continues: “The present invention relates to **multi-modal data processing techniques** and, more particularly, to systems and methods for performing focus detection, referential ambiguity resolution and mood classification in accordance **with multi-modal input data.** ... Systems that employ such ‘multi-modal’ input techniques have inherent advantages over systems that use only one data input mode.” Ex.1005, 1:9–23.

68. Further, Maes’s systems and methods are directed at “*processing multi-modal natural language inputs.*” For example, with reference to Figure 1 (reproduced below), Maes discloses:

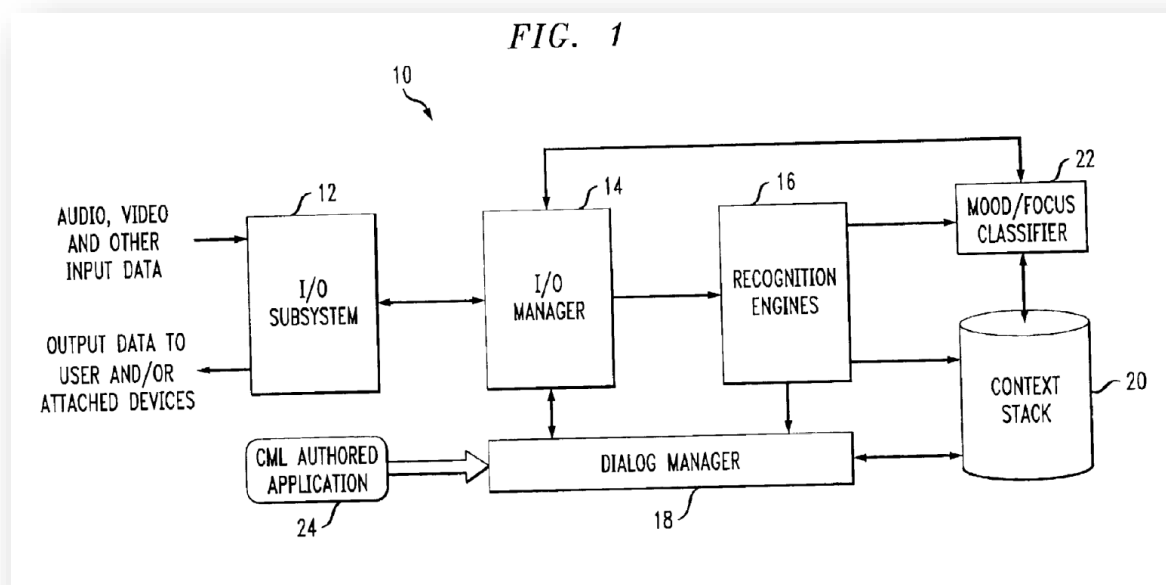
[A] multi-modal conversational computing system ...
compris[ing] an input/output (I/O) subsystem 12, an I/O manager module 14, one or more recognition engines 16, a dialog manager module 18, a context stack 20 and a mood/focus classifier 22.

Generally, the multi-modal conversational computing system 10 of the present invention receives multi-modal input in the form of audio input data, video

⁶ Emphasis added throughout unless otherwise specified.

input data, as well as other types of input data (in accordance with the I/O subsystem 12), processes the multi-modal data (in accordance with the I/O manager 14), and performs various recognition tasks (e.g., speech recognition, speaker recognition, gesture recognition, lip reading, face recognition, etc., in accordance with the recognition engines 16), if necessary, using this processed data. The results of the recognition tasks and/or the processed data, itself, is then used to perform one or more conversational computing tasks, e.g., focus detection, referential ambiguity resolution, and mood classification (in accordance with the dialog manager 18, the context stack 20 and/or the classifier 22).

Ex.1005, 3:66-4:21.

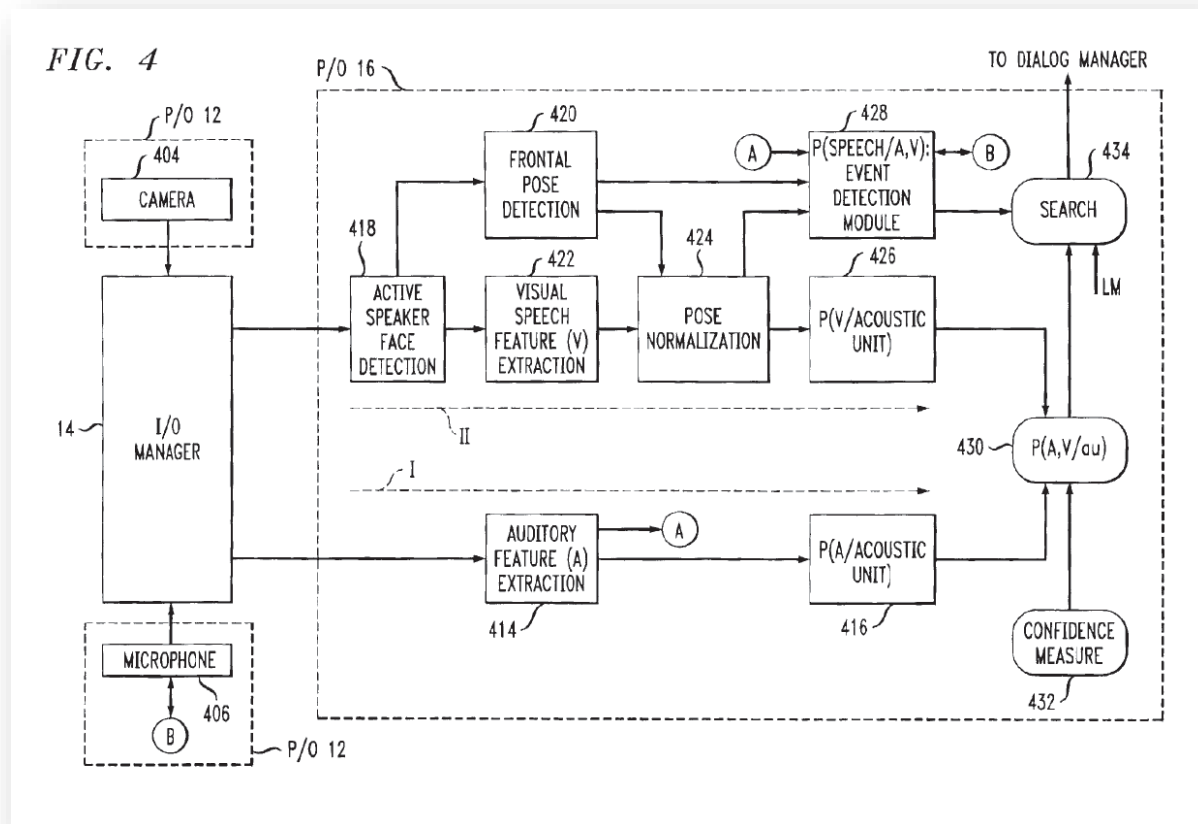


Ex.1005, Figure 1.

69. With reference to Figures 1 and 4, Maes describes Figure 4 (reproduced below) as “an audio-visual speech recognition Maes describes Figure 4 (reproduced below) as “an audio-visual speech recognition module⁷ that may be employed as one of the recognition modules⁸ of FIG. 1 **to perform speech recognition using multi-modal input data received** in accordance with the invention.” Ex.1005, 10:50–52. Specifically, Figure 4 shows that the audio-visual speech recognition module receives multi-modal input data including an audio signal (e.g., speech provided by the speaker and/or background noise) (“*natural language inputs*”) and a video signal (e.g., the speaker’s face including lip movement and/or background objects in the environment) and processes the multi-modal input data (e.g., comprising “*multi-modal natural language inputs*”) by passing through several processing modules (e.g., the video signal is processed by blocks/modules 418, 422, 424, 426 and the audio signal is processed by blocks/modules 414, 416). Therefore, Maes discloses “[a] method for processing multi-modal natural language inputs.”

⁷ As used herein, the terms “Maes’s system” and “audio-visual speech recognition module” are used interchangeably.

⁸ Maes uses the term “block” and “module” interchangeably. For example, see 12:43-38 and 17:46-50 referring to “414” as “block[] 414” and “module 414.”



Ex.1005, Figure 4.

70. Maes further explains that “multi-modality” “may comprise a combination of other modalities other than voice and video. For example, multi-modality may include keyboard/pointer/mouse (or telephone keypad) and other sensors, etc. Thus, a general principle of the present invention of the combination of modality through at least two different sensors (and actuators for outputs) to disambiguate the input, and guess the mood or focus, can be generalized to any such combination. Engines or classifiers for determining the mood or focus will then be

specific to the sensors but the **methodology** of using them is the same as disclosed herein.” Ex.1005, 3:4–11.

71. Accordingly, in Maes, “**computer software including instructions or code for performing the methodologies of the invention**, as described herein, may be stored in one or more of the associated memory devices (e.g., ROM, fixed or removable memory) and, when ready to be utilized, loaded in part or in whole (e.g., into RAM) and executed by a CPU. ... [I]t should be understood that the elements illustrated in FIGS. 1 through 9C may be implemented in various forms of hardware, software, or combinations thereof, e.g., one or more digital signal processors with associated memory, application specific integrated circuit(s), functional circuitry, one or more appropriately programmed general purpose digital computers with associated memory, etc. Given the teachings of the invention provided herein, one of ordinary skill in the related art will be able to contemplate other implementations of the elements of the invention.” Ex.1005, 45:45–60.

72. Maes describes examples in which its methods are applicable to a broadcast news system which uses spoken utterance⁹ inputs which further confirms that its method applies to “*natural language inputs*.” For example, Maes states that

-
- ⁹ As used herein, it will be understood that a “spoken utterance” received using Maes’s system, would encompass both an audio portion and a video (*i.e.*, non-speech) portion (e.g., the accompanying lip movement) of the utterance. Ex.1005, 6:43–47.

“the audio-visual speech recognition module ... process[es] arbitrary content video [such as] in the context of broadcast news ... [which] contain a newsperson speaking at a location where there is arbitrary activity and noise in the background.” Ex.1005, 10:63–11:15. *See* Gibbon defining “*natural language*” as “a language used by humans for communication.” Ex.1017, 444.

73. As another example, Maes states that “the multi-modal conversational computing system 10 may be employed within a vehicle” in which a user may say the “spoken utterance ‘turn it on’” and Maes’s system processes inputs relating to an I/O event representative of the user’s spoken utterance (audio portion) and the accompanying lip movement of the spoken utterance (video portion). Ex.1005, 4:30–32, 6:43–50, and 7:63–8:29.

74. Therefore, Maes discloses “[a] *method for processing multi-modal natural language inputs, comprising:*”

- b. **[19.1.0] receiving a multi-modal natural language input at a conversational voice user interface, the multi-modal input including a natural language utterance and a non-speech input provided by a user,**

75. Maes discloses this claim element.

76. For example, Maes describes Figure 4 (reproduced below) as “an audio-visual speech recognition module that may be employed as one of the recognition modules of FIG. 1 **to perform speech recognition using multi-modal input data received** in accordance with the invention.” Ex.1005, 10:50–52. Specifically,

Maes's Figure 4 discloses “an audio-visual speech recognition module” that “receive[s]” (“*receiving*”) multi-modal input data—such as an audio signal and a video signal—from a speaker (“*a multi-modal natural language input ... provided by a user*”) “at” an “input/output (I/O) subsystem ... compris[ing] one or more microphones for capturing audio input data” (e.g., “*a conversational voice user interface*”), which “one of ordinary skill in the art will realize that other user interfaces and devices ... may be included for capturing user activity.” Ex.1005, 6:21–29, 6:5–8, and 10:47–11:23.

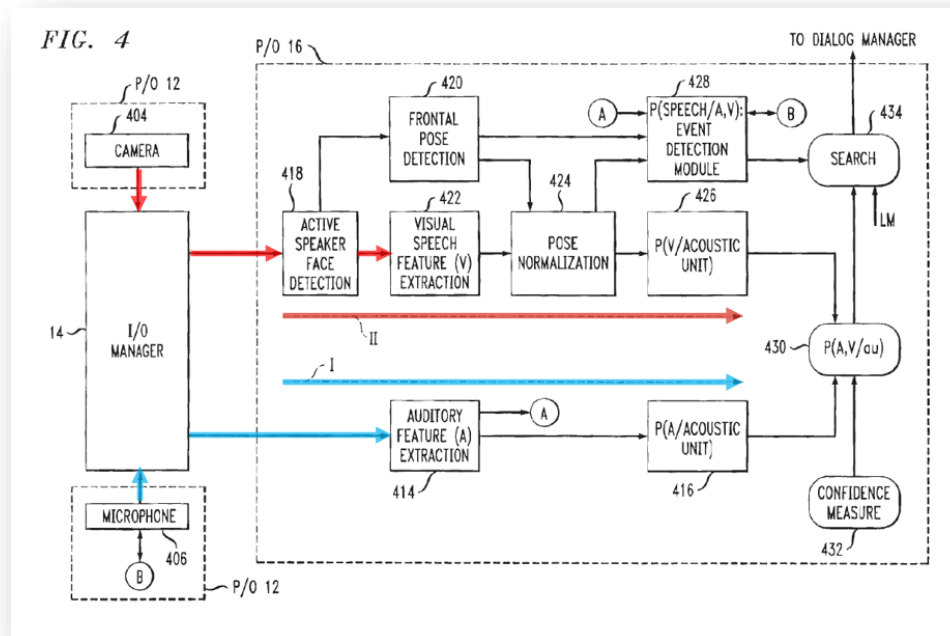
77. Maes also discloses that the “multi-modal input data” includes “an audio signal” (e.g., speech provided by the speaker and/or background noise) (“*a natural language utterance*”) “and” “a video signal” (“*a non-speech input*”), e.g., including the speaker’s face including lip movement and/or background objects in the environment. Ex.1005, 11:8–63.

78. With reference to Figure 4¹⁰, Maes depicts “[a] phantom line denoted by Roman numeral I represents the processing path the audio information signal takes within the module, while a phantom line denoted by Roman numeral II

¹⁰ There appears to be a typographical error in Figure 4—the labels “P/O” are incorrect and should have been denoted “I/O,” e.g., “I/O subsystem 12.” Ex.1005, 4:8–16 and Figure 1 confirm this.

represents the processing path the video information signal takes within the module.”

Ex.1005, 11:55–63. Paths I and II are annotated in blue and red in Figure 4 (below).



Ex.1005, Figure 4 (annotated).

79. Maes explains “receiving real-time arbitrary content from a video camera 404 and microphone 406 via the I/O manager 14 ... While the video signals received from the camera 404 and the audio signals received from the microphone 406 are shown in FIG. 4 as not being compressed, they may be compressed and therefore need to be decompressed in accordance with the applied compression scheme.” Ex.1005, 11:23–31. Specifically, the feature extractor 414 of the audio-visual speech recognition module “receives an audio or speech signal” and the active speaker face detection module 418 “receives video input [from] camera 404” (e.g.,

“the visual portion (e.g., lip movement) of the utterance,” (Ex.1005, 6:46-47)).
Ex.1005, 11:64–12:2, 12:49–61.

80. Maes explains in detail what constitutes the “audio signal” (“*natural language utterance*”) and the “video signal” (“*non-speech input*”). For example, the “audio signal” (“*natural language utterance*”) can include “silence,” “speech,” and “noise.” Ex.1005, 13:3–15. And, furthermore, with regards to the “video signal” (“*non-speech input*”) comprising “the visual portion (e.g., lip movement) of the utterance,” (Ex.1005, 6:46–47) Maes describes that:

the video signal captured by the camera 404 can be of any particular type. As mentioned, the face and pose detection techniques may process images of any wavelength such as, e.g., visible and/or non- visible electromagnetic spectrum images. By way of example only, this may include infrared (IR) images (e.g., near, mid and far field IR video) and radio frequency (RF) images. Accordingly, the module may perform audio-visual speech detection and recognition techniques in poor lighting conditions, changing lighting conditions, or in environments without light. For example, the system may be installed in an automobile or some other form of vehicle and capable of capturing IR images so that improved speech recognition may be performed. Because video information (i.e., including visible and/or non-visible electromagnetic spectrum images) is used in the speech recognition process, the system is less susceptible to recognition errors due to noisy conditions, which significantly hamper conventional recognition systems that use only audio information. In addition, due to the methodologies for processing the visual information described herein, the module provides the capability to perform accurate LVCSR (large vocabulary continuous speech recognition).

Ex.1005, 11:32–54.

81. Further, Maes explains that, in accordance with its invention, “**multi-modality ... may comprise a combination of other modalities other than voice and video**. For example, multi-modality may include keyboard/pointer/mouse (or telephone keypad) and other sensors, etc. Thus, a general principle of the present invention of the combination of modality through at least two different sensors (and actuators for outputs) to disambiguate the input, and guess the mood or focus, can be generalized to any such combination. Engines or classifiers for determining the mood or focus will then be specific to the sensors but the methodology of using them is the same as disclosed herein.” Ex.1005, 3:1–12. Maes’s system therefore complements conventional interfaces and user input/output rather than replacing them and as Maes describes “[t]his is the notion of “multi-modality” whereby speech, and video ... may be used in parallel with a mouse, keyboard, and other input devices such as a pen.” Ex.1005, 31:61–66. The above-mentioned disclosures of Maes relating to inputs to its system received via keyboard/pointer/mouse (or telephone keypad) are the same non-speech inputs described as the 468 Patent’s description of its system coupled to “keypads 14 for receiving textual data input” and “users may interact with the mobile device 36 through the speech unit 128, the keypad 74 or keyboard, a display 72 that displays text, graphics, video, or other peripheral” such as “a stylus and digitizer pad for capturing electronic handwriting.”

Ex.1001, 16:18–19, 6:22–26, and 19:1–4; *see also* Compl., ¶111 (describing “non-speech, keyboard-entered natural language input (“From my checking account”)” and “non-speech, icon-selection natural language input (by tapping one of the provided selection boxes containing the natural language input text)).

82. Additionally, with reference to the Figure 1¹¹, Maes discloses that “the multi-modal conversational computing system 10 of the present invention **receives multi-modal input in the form of audio input data, video input data, as well as other types of input data (in accordance with the I/O subsystem 12)**, processes the multi-modal data (in accordance with the I/O manager 14), and performs various recognition tasks (e.g., “speech recognition, speaker recognition, gesture recognition, lip reading, face recognition”). Ex.1005, 4:7–15. “The system 10 therefore performs referential ambiguity resolution with respect to multiple users **by taking both audio input data and image data input and processing it** to make a user resolution determination. This may include detecting speech activity and/or the identity of the user based on both audio and image cues.” Ex.1005, 4:67–5:5.

83. Maes explains how the multi-modal input comprising the audio signal (“*natural language utterance*”) and video signal (“*non-speech input*”) is received.

¹¹ Maes identifies Figure 4 as “a preferred embodiment of an audio-visual speech recognition module” that may be employed as one of the recognition modules of the Figure 1 embodiment. Ex.1005, 10:47–52.

As mentioned above, the data input portion of the subsystem may comprise one or more cameras or sensors for **capturing video input data representing the environment in which the system (or, at least, the I/O subsystem) is deployed**. The cameras/sensors may be capable of capturing not only visible image data (images in the visible electromagnetic spectrum), but also IR (near, mid and/or far field IR video) and/or RF image data. Of course, in systems with more than one camera, different mixes of cameras/sensors may be employed, e.g., system having one or more video cameras, one or more IR sensors and/or one or more RF sensors.

In addition to the one or more cameras, **the I/O subsystem 12 may comprise one or more microphones for capturing audio input data from the environment in which the system is deployed**. Further, the I/O subsystem may also include an analog-to-digital converter which converts the electrical signal generated by a microphone into a digital signal representative of speech uttered or other sounds that are captured. Further, the subsystem may sample the speech signal and partition the signal into overlapping frames so that each frame is discretely processed by the remainder of the system.

Ex.1005, 5:60–6:15.

84. Maes explains that “**the elements illustrated in FIGS. 1 through 9C may be implemented in various forms of hardware, software, or combinations thereof, e.g., one or more digital signal processors with associated memory, application specific integrated circuit(s), functional circuitry, one or more appropriately programmed general purpose digital computers with associated memory, etc.** Given the teachings of the invention provided herein, one of ordinary

skill in the related art will be able to contemplate other implementations of the elements of the invention.” Ex.1005, 45:50–60.

85. The 468 Patent describes an embodiment illustrated in Figure 1 in which the disclosed system (mobile structure 10) is “used in vehicles” and comprises “speech interface and non-speech interface.” Ex.1001, 15:58–64. Specifically, Figure 1 shows the “non-speech interface” of the inventive system including video camera 16 “for receiving multimedia communications or conferences” and “speech interface” / “speech-based interface” including “one or more speech units 128” that further “include one or more microphones ... to receive the utterances from the user.” Ex.1001, 16:19–20, 25:48–52 and Figure 1. This description of the 468 Patent is analogous to Maes’s disclosure of an in-vehicle system as explained below. *See generally* Ex.1005, 6:15–28 and 6:43–50.

86. Maes describes the “*receiving* ...” step by providing an example in which its multi-modal conversational computing system is employed within a vehicle. For instance, a user in the vehicle may say “turn it on,” while pointing at the vehicle radio. Maes’s system receives multi-modal inputs relating to an I/O event representative of the user’s spoken utterance and lip movement corresponding to the spoken utterance. *See* Ex.1005, 6:43–47. For example, “the microphone picks up the audible portion of the utterance and a camera picks up the visual portion (e.g., lip movement) of the utterance, [and] the event may be sent to an audio-visual speech

recognition engine to have the utterance recognized using both the audio input and the video input associated with the speech.” Ex.1005, 6:43-50. “The I/O manager [of the multi-modal conversational computing system] receives the raw multi-modal data and abstracts the data into a form that represents ... a spoken utterance ... [by] generalizing details associated with all or portions of the input data ... to yield a more generalized representation of the data for use in further operations.” Ex.1005, 6:35–38. “The dialog manager would therefore receive the results of the recognized events associated with the spoken utterance ‘turn it on’” including the accompanying lip movement. Ex.1005, 7:63–67.

87. Therefore, Maes teaches or suggests this claim element for at least this reason.

Additional Mapping

88. Maes additionally explains that “multi-modality ... may comprise a combination of other modalities other than voice and video [such as] ... **keyboard**/pointer/mouse (or telephone keypad) and other sensors” *See generally* Ex.1005, 3:1–12. Maes’s system “may also comprise other typical input devices for obtaining user input, e.g., **GUI-based devices** such as a keyboard, a mouse, etc., and/or other devices such as a stylus and digitizer pad for capturing electronic handwriting.” Ex.1005, 6:22–26 and 16:20–21. Maes’s system therefore complements conventional interfaces and user input/output rather than replacing

them, and as Maes describes “[t]his is the notion of ‘multi-modality’” whereby “speech, and video ... may be used in parallel with a mouse, **keyboard**, and other input devices such as a pen.” Ex.1005, 31:61–66. “The ‘multi-modality’ aspect of the [system] implies that various I/O resources such as voice, keyboard, pen, and pointing device (mouse), keypads, **touch screens**, ... and video ... can be used.” Ex.1005, 32:29–32. Thus, in Maes, keyboard-typed input and touch screen input comprised in the multi-modal input data constitute additional examples of “*a non-speech input*” included in the “*multi-modal natural language input*.”

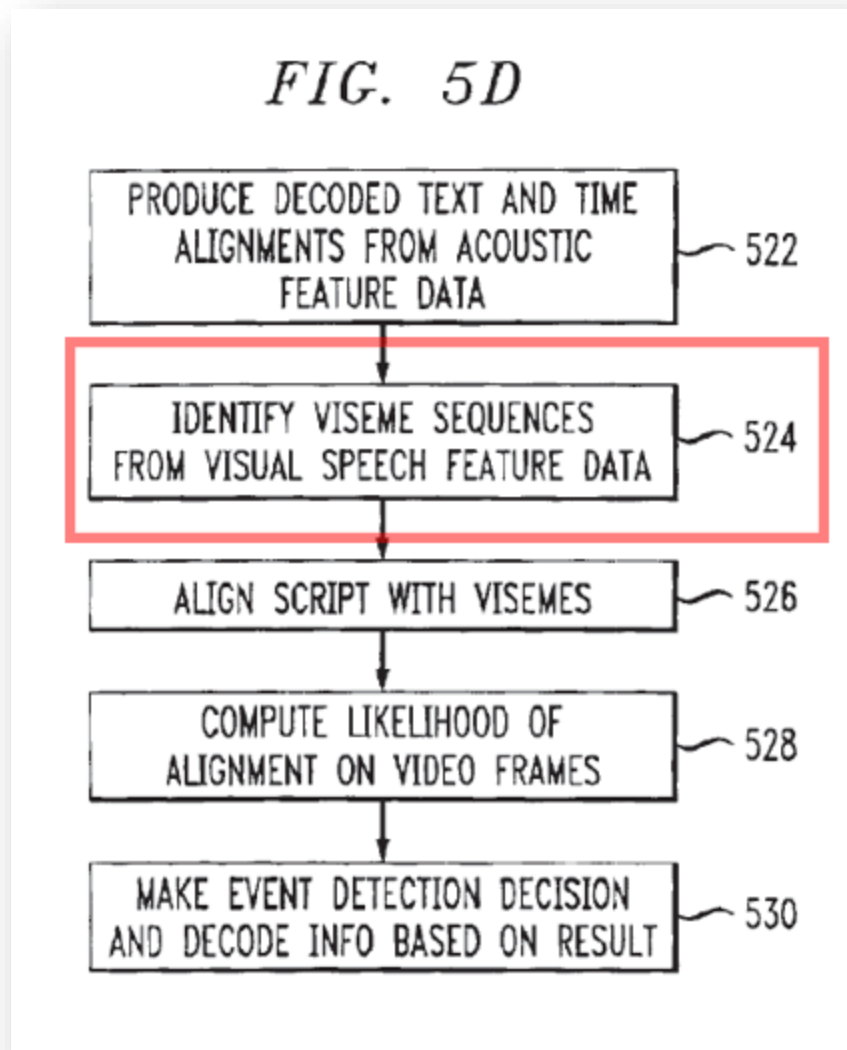
89. Maes’s above disclosures identifying system inputs received via keyboard/pointer/mouse/keypad/GUI-based device are the same non-speech inputs described in the 468 Patent: “user interface peripherals ... may include touch screen displays for displaying text, graphics and video; keypads 14 for receiving textual data input; ... a pointing device or stylus, or other user interface peripherals” and “users ... interact with the mobile device 36 through the speech unit ..., the keypad ... or keyboard, ... or other peripheral.” Ex.1001, 16:15–19 and 19:1–4; *see also* Dialect’s Complaint, Ex.1007, ¶111 (keyboard-entered natural language input (“From my checking account” and “non-speech, icon-selection natural language input (by tapping one of the provided selection boxes containing the natural language input text)”).

90. Therefore, Maes teaches or suggests this claim element for this additional reason.

- c. **[19.1.1] wherein a transcription module coupled to the conversational voice user interface transcribes the non-speech input to create a non-speech based transcription;**

91. Maes teaches or suggests this claim element.

92. For example, Maes depicts an audio-visual speech recognition module in Figure 4 comprising a visual speech feature extractor 422 that “extracts visual speech feature vectors (e.g., mouth or lip-related parameters)” (“*transcribes*”) to “produce” “a visual phonemes (visemes) sequence” (“*create a non-speech based transcription*”) “*from*” the video signal (“*the non-speech input*”), e.g. including the face detected in the video frame including the user’s lip movement. Ex.1005, 17:51–55, 21:47–50. Maes discloses the “*transcribes ...*” step as a processing step in block 524 in Figure 5D.



Ex.1005, Fig. 5D (annotated).

93. Maes further explains modules 422, 424 and 426 in Figure 4 extracting, normalizing, and labeling (e.g., “*transcrib[ing]*”) visemes, e.g., in a visemes sequence (“*a non-speech based transcription*”) using the extracted visual speech vectors.

The extracted visual speech feature vectors [produced by visual speech feature extractor 422] **are then normalized**

in block 424 with respect to the frontal pose estimates generated by the detection module 420. The normalized visual speech feature vectors are then provided to a probability module 426. Similar to the probability module 416 in the audio information path which labels the acoustic feature vectors with one or more phonemes, the probability module 426 labels the extracted visual speech vectors with one or more previously stored phonemes. ... Alternatively, the visual speech feature vectors may be labeled with visemes which, as previously mentioned, are visual phonemes or canonical mouth shapes that accompany speech utterances.

Ex.1005, 18:42–63.

94. “As is known, visemes, or visual phonemes, are generally canonical mouth shapes that accompany speech utterances which are categorized and pre-stored similar to acoustic phonemes.” Ex.1005, 29:27–30. A POSITA would have understood that visemes sequence (“*a non-speech-based transcription*”) generated in Maes is essentially a string of visemes that is the tangible result of a visual recognition system’s interpretation of a user’s lip movement accompanying a spoken utterance and provides a visual representation of the non-speech content. For example, Shdaifat describes visemes as composed of features. Ex.1028, 1 (“The different features of the visemes are stored in a template”).



Figure 5: Features of the one viseme.

Ex.1028, Figure 5. Shdaifat further describes that visemes can be represented by images:

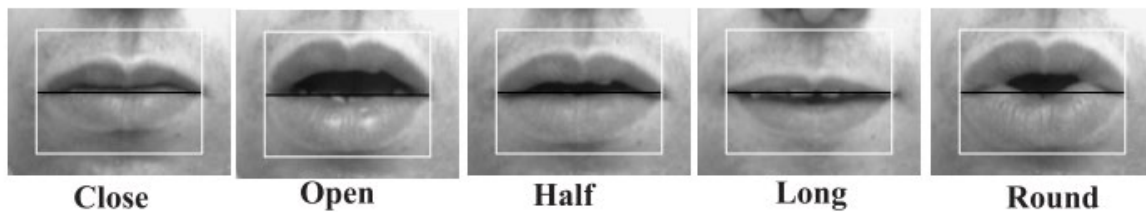


Figure 1: An example of images of five viseme classes.

Ex.1028, Figure 1. Referencing Figure 1, Shdaifat explains that “[f]or each viseme class the features of the lips are extracted and stored in templates. Each template contains the two images of the left and right corners of the mouth, the edge image of the viseme image, line segments of 10 pixel length fitted to the edges.”

Ex.1028, 2.

Ex.1005, Figure 4 (showing video information processing path (Path II) annotated in red). For example, the blocks annotated in red comprise functionality of “*a transcription module*”), which as Maes shows, is “*coupled to*” “the (I/O) subsystem 12 ... compris[ing] one or more microphones for capturing audio input data” (“*the conversational voice user interface*”). Ex.1005, 6:5–8.

96. Maes provides several examples of visual speech features. “Examples of visual speech features that may be extracted are grey scale parameters of the mouth region; geometric/model based parameters such as area, height, width of mouth region; lip contours arrived at by curve fitting, spline parameters of inner/outer contour; and motion parameters obtained by three dimensional tracking. Still another feature set that may be extracted via module 422 takes into account the above factors.” Ex.1005, 17:55–63.

97. In addition, Maes describes multiple ways of extracting visual speech feature vectors.

[W]hile the visual speech feature extractor 422 may implement one or more known visual feature extraction techniques, in one embodiment, the extractor extracts grey scale parameters associated with the mouth region of the image. Given the location of the lip comers, after normalization of scale and rotation, a rectangular region containing the lip region at the center of the rectangle is extracted from the original decompressed video frame. Principal Component Analysis (PCA), as is known, may be used to extract a vector of smaller dimension from this vector of grey-scale values.”

Ex.1005, 18:1–11.

Another method of extracting visual feature vectors that may be implemented in module 422 may include extracting geometric features. This entails extracting the phonetic/visemic information from the geometry of the lip contour and its time dynamics.

Ex.1005, 18:12–20. And as explained earlier, Maes describes that the visual speech feature vectors from the visual feature extractor 422 along with other modules depicted in Figure 4 (“*a transcription module*”) “*transcribes*” the video signal (“*the non-speech input*”), e.g. including the face detected in the video frame including the user’s lip movement to “*produce a visual phonemes (visemes) sequence*” (“*create a non-speech based transcription*”). Ex.1005, 21:43–50. Therefore, Maes discloses “*wherein a transcription module coupled to the conversational voice user interface transcribes the non-speech input to create a non-speech based transcription.*”

98. Referring to the example in which Maes’s multi-modal conversational computing system is employed within a vehicle, Maes explains the “*transcrib[ing]* ...” step in that example as follows. The raw multi-modal input data relating to an I/O event representative of the user’s spoken utterance and lip movement accompanying the spoken utterance received by the multi-modal conversational computing system is abstracted into a recognizable form. For example, “[t]he I/O manager [of the multi-modal conversational computing system] receives the raw multi-modal data and **abstracts the data into a form that represents ... a spoken utterance.** As is known, **a data abstraction operation may involve generalizing details associated with all or portions of the input data so as to yield a more generalized representation of the data for use in further operations.**” Ex.1005, 6:21–38. Thus, as a result of processing by several modules in Maes’s system, the

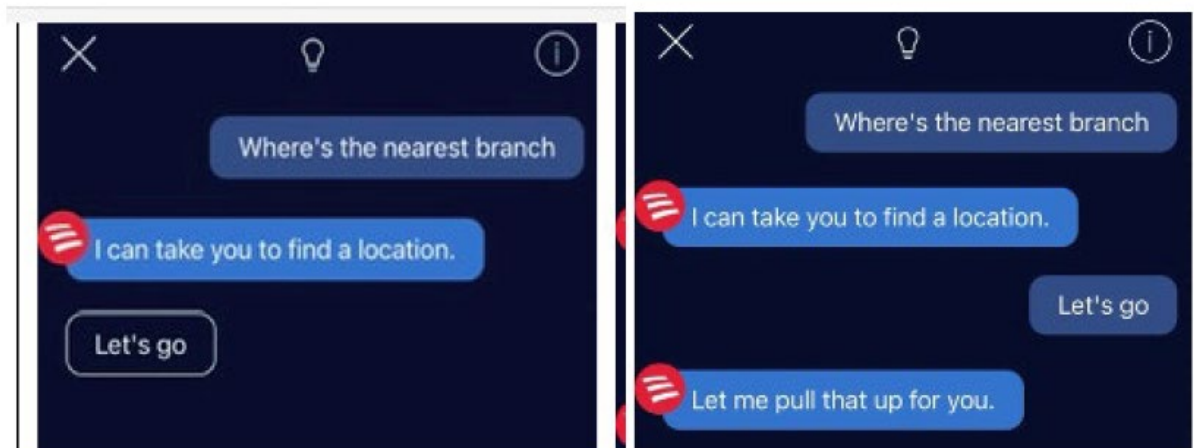
system produces a visemes sequence corresponding to the lip movement accompanying the spoken utterance of the user. *See also* Ex.1005, 6:43–50.

99. Accordingly, Maes teaches or suggests this element.

Additional Mapping

100. As explained in [19.1.0], in Maes, keyboard-typed inputs and touch screen input constitute additional examples of “*a non-speech input*” and further Maes teaches for “**typical GUI-based input** (e.g., **keyboard**, mouse, etc.), **the data may not necessarily need to be recognized since the data is already identifiable without recognition operations.**” Ex.1005, 6:62–65. Because the inputs are converted (transcribed) to their corresponding internal representations of such inputs, a POSITA, therefore, would have understood that each of the keyboard-typed input and touch screen input inherently serve as “*non-speech-based transcription.*” Ex.1005, 6:62–65.

101. Dialect’s Complaint, consistent with Maes’s teachings above, identifies: “Erica transcribed a non-speech input (tapping on box provided by Erica with the words “Let’s go”) into a non-speech-based transcription (displayed/represented to the user in the form of “Let’s go” in the user’s dialogue stream).” Ex.1007, ¶114.



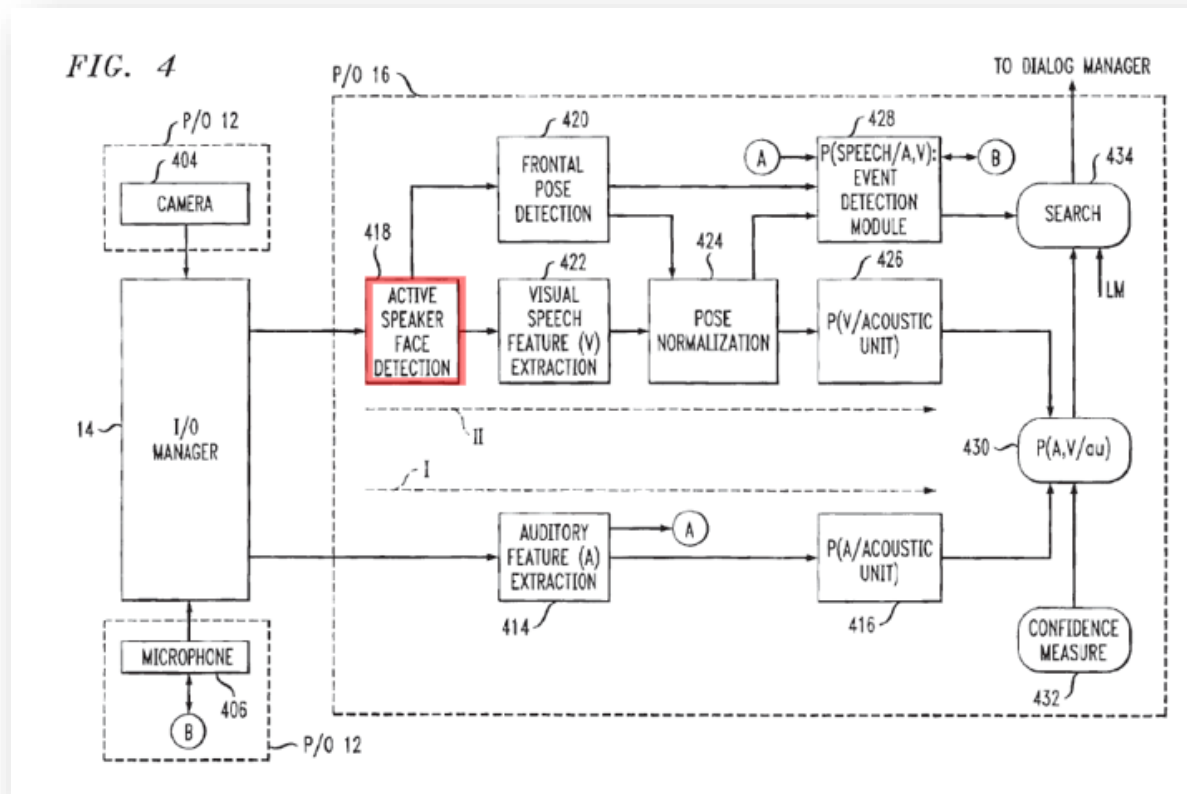
Id.

102. Accordingly, Maes teaches or suggests this element for this additional reason.

d. [19.2] identifying the user that provided the multi-modal input;

103. Maes teaches or suggests “*identifying the user that provided the multi-modal input.*”

104. For example, Maes discloses that the audio-visual speech recognition module shown in Figure 4 (reproduced below) includes an active speaker face detection module 418 for “detecting” (“*identifying*”) a face of the “speaker” (“*the user*”) “*who provided*” “the multi-modal input data” (“*the multi-modal natural language input*”). Ex.1005, 10:47–63, 12:49–62.



Ex.1005, Figure 4 (annotated).

105. Maes explains that “[t]he active speaker face detection module 418 receives video [from] input camera 404” and further that “speaker face detection can also be performed directly in the compressed data domain and/or from audio and video information rather than just from video information. In any case, module 418 generally locates and tracks the speaker’s face and facial features within the arbitrary video background.” Ex.1005, 12:49–62. Therefore, by detecting a speaker’s face in the video signal comprising the speaker’s face including lip movement and/or background objects in the environment, Maes’s audio-visual speech recognition

module “*identif[ies] the user that provided the multi-modal input.*” See Ex.1005, 6:50–53 (“[I]n addition, the event may be sent to an audio-visual speaker recognition engine to have the speaker of the utterance identified...”).

106. In addition, Maes discloses use of an “audio-visual speaker recognition engine to have the speaker of the utterance identified, verified and/or authenticated.” Ex.1005, 6:50–55. Specifically, with reference to Figure 6¹², Maes elaborates on “a preferred embodiment of an audio-visual speaker recognition module that may be employed as one of the recognition modules of FIG. 1 to perform speaker recognition using multi-modal input data received in accordance with the invention.” Ex.1005, 22:41–46. “Like the audio-visual speech recognition module of FIG. 4, the speaker recognition module of FIG. 6 may receive the same types of arbitrary content video from the camera 604 and audio from the microphone 606 via the I/O manager 14.” Ex.1005, 22:60–64. For example, **“the [audio speaker recognition] module 616 may perform speaker identification and/or speaker verification using the extracted acoustic feature vectors. The processes of speaker identification and verification may be accomplished via any conventional acoustic information speaker recognition system ... [e.g.,] described in the U.S.**

¹² There appears to be a typographical error in Figure 6—the labels “P/O” are incorrect and should have been denoted as “I/O,” e.g., “I/O subsystem 12.” Ex.1005, 4:8–16 and Figure 1 confirm this.

patent application identified by Ser. No. 08/788,471, filed on Jan. 28, 1997. ... The illustrative speaker identification system may use two techniques: a model-based approach and a frame-based approach. In the experiments described herein, we use the frame-based approach for speaker identification based on audio.” Ex.1005, 23:44–24:3 .

107. Maes provides a detailed description of the frame-based approach as follows:

Let M_i be the model corresponding to the i th enrolled speaker. M_i is represented by a mixture Gaussian model defined by the parameter set $\{\mu_{i,j}, \Sigma_{i,j}, p_{i,j}\}_{j=1, \dots, n_i}$, consisting of the mean vector, covariance matrix and mixture weights for each of the n_i components of speaker i 's model. These models are created using training data consisting of a sequence of K frames of speech with d -dimensional cepstral feature vectors, $\{f_m\}_{m=1, \dots, K}$. The goal of speaker identification is to find the model, M_i , that best explains the test data represented by a sequence of N frames, $\{f_n\}_{n=1, \dots, N}$. We use the following frame-based weighted likelihood distance measure, $d_{i,n}$, in making the decision:

$$d_{i,n} = -\log\left[\sum_{j=1}^{n_i} p_{i,j} p(f_n | \mu_{i,j}, \Sigma_{i,j})\right]$$

The total distance D_i of model M_i from the test data is then taken to be the sum of the distances over all the test frames:

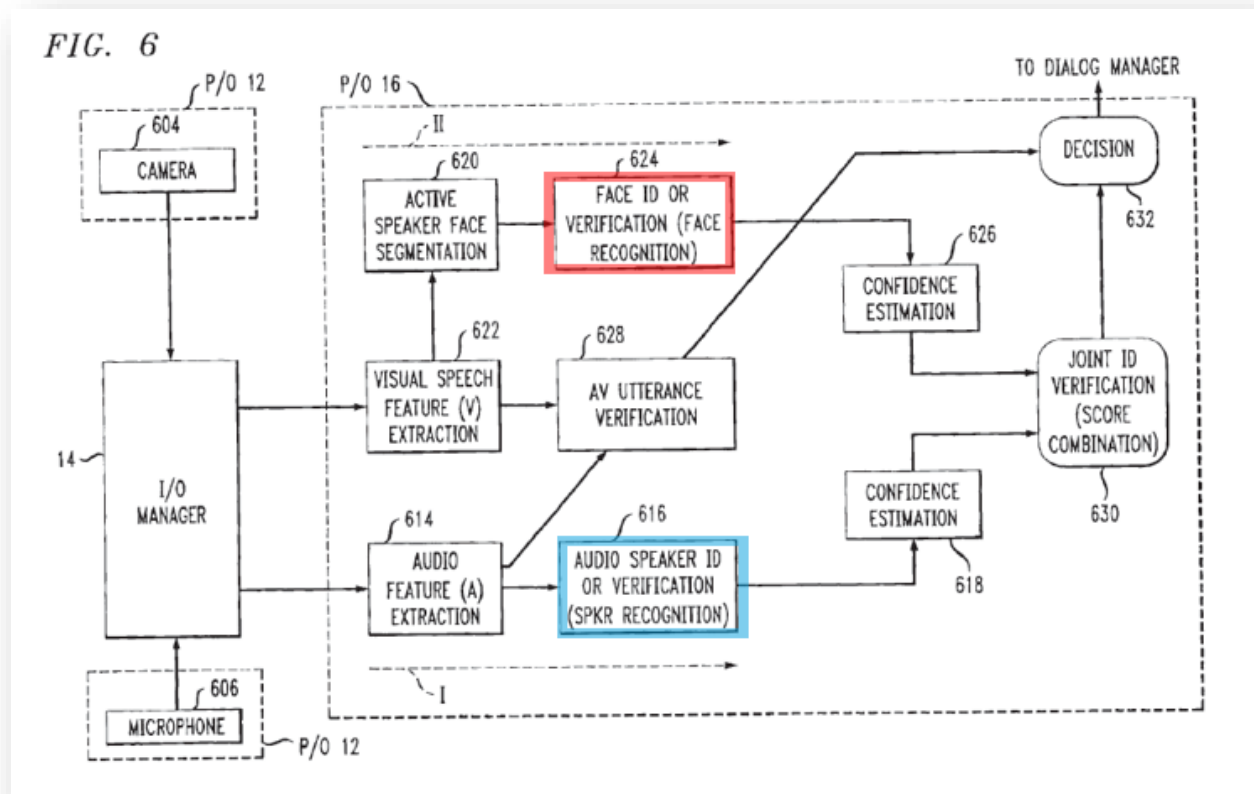
$$D_i = \sum_{n=1}^N d_{i,n}$$

Thus, the above approach finds the closest matching model and the person whose model that represents is

determined to be the person whose utterance is being processed.

Ex.1005, 24:2–30.

108. In addition to performing speaker identification using audio signals by the audio speaker recognition module 616, Maes's audio-visual speaker recognition module in Figure 6 (reproduced below) also includes a face recognition module 624 that uses video signals to “*identify[] the user that provided the multi-modal input.*” Specifically, the face recognition module 624 “*identif[ies] and/or verif[ies] the face of the person assumed to be the speaker in the video. Verification can also be performed by adding score thresholding or competing models. Thus, the visual mode of speaker identification is implemented as a face recognition system where faces are found and tracked in the video sequences, and recognized by comparison with a database of candidate face templates.*” Ex.1005, 24:44–25:3.



Ex.1005, Figure 6 (annotated). “**Speaker verification may be performed in a similar manner**, ... the input acoustic data is compared to determine if the data matches closely enough with stored models. If the comparison yields a close enough match, the person uttering the speech is verified. The match is accepted or rejected by comparing the match with competing models.” Ex.1005, 24:31–36. Therefore, in view of the above disclosures, Maes discloses “*identifying the user that provided the multi-modal input.*”

109. Referring to the example in which Maes’s multi-modal conversational computing system is employed within a vehicle, the “*identifying ...*” step would

result in the system recognizing/identifying the specific user who provided the multi-modal natural language inputs comprising an audible portion (e.g., picked up by a microphone) of the spoken utterance of “turn it on” and a visual portion (e.g., picked up by a camera) of lip movement. Ex.1005, 6:43–47, 7:64.

110. Accordingly, Maes discloses “*identifying the user that provided the multi-modal input.*”

- e. **[19.3.0] creating a speech-based transcription of the natural language utterance using a speech recognition engine and a semantic knowledge-based model,**

111. Maes discloses this claim element.

112. In the context of speech recognition systems, a “*semantic knowledge-based model*” and “*a speech recognition engine*” were well-known at the time of the 468 Patent. As early as 1998, a patent issued to Komissarchik (entitled “**Knowledge-based speech recognition system and methods** having frame length computed based upon estimated pitch period of vocalic intervals) discloses “**knowledge-based speech recognition system** and methods ... includ[ing] a series of ... steps which provide real-time analysis and transcription of spoken utterances.” Ex.1018, 8:4–7. Komissarchik describes that its invention includes “**semantic control of the recognition results.**” Ex.1018, 6:45–47; *see also* Ex.1018, 2:60–3:62 (describing several prior art references implementing knowledge-based systems).

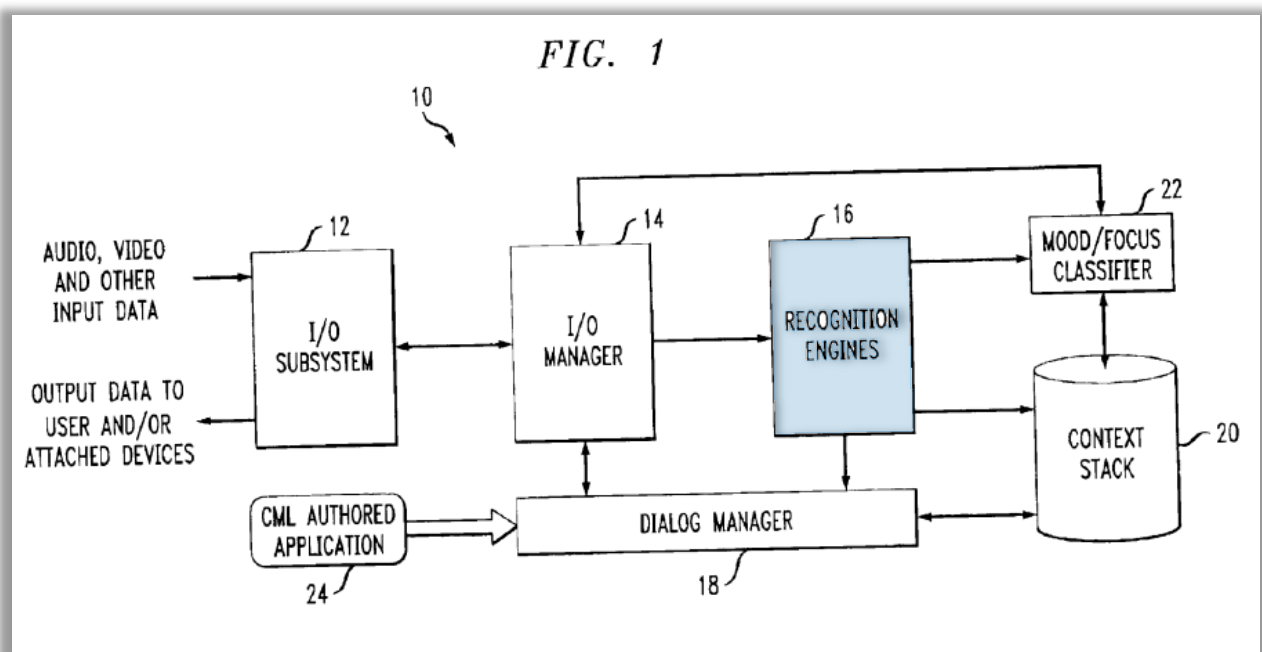
113. As another example, the W3C—the World Wide Web Consortium—described as “an international public-interest non-profit organization where Member organizations, a full-time staff, and the public work together to develop Web standards” and founded by “Web inventor Tim Berners-Lee—worked on a standard called “Semantic Interpretation for Speech Recognition” and they published their first working draft in November of 2001. *See* Ex.1011; <https://www.w3.org/about/corporation/#:~:text=The%20World%20Wide%20Web%20Consortium,for%20the%20benefit%20of%20humanity>.

114. The W3C draft explains that it “defines the process of Semantic Interpretation for Speech Recognition and the syntax and semantics of semantic interpretation tags that can be added to speech recognition grammars to compute information to return to an application on the basis of rules and tokens that were matched by the speech recognizer.” Ex.1011, 1. As yet another example, Suominen explains that a semantic model “provides probabilities based on semantic relationships without regard to the particular syntax used to express those semantic relationships” and further that a semantic model “may be adapted ... for accurate identification of a present context or a user context.” Ex.1015, [0146]. Thus, well before the priority date of the 468 Patent, “*a speech recognition engine and a semantic knowledge-based model*” was known in the prior art, and for the reasons explained below, Maes teaches or suggests these claim elements.

i. Speech Recognition Engine

115. Maes discloses “*a speech recognition engine.*”

116. Referencing Figure 1, Maes discloses its system generally “performs various recognition tasks (e.g., **speech recognition** ... in accordance with the **recognition engines 16**).” *See generally* Ex.1005, 4:7–14; see also Ex.1005, 34:61–35:1 (one or more recognition engines of FIG. 1). “[D]epending on the nature of the [I/O (input/output)] event, **one or more recognition engines may be used to recognize the event**. For example, if the event is some form of spoken utterance ..., the event may be sent to an audio-visual **speech recognition engine** to have the utterance recognized” Ex.1005, 6:42–50.



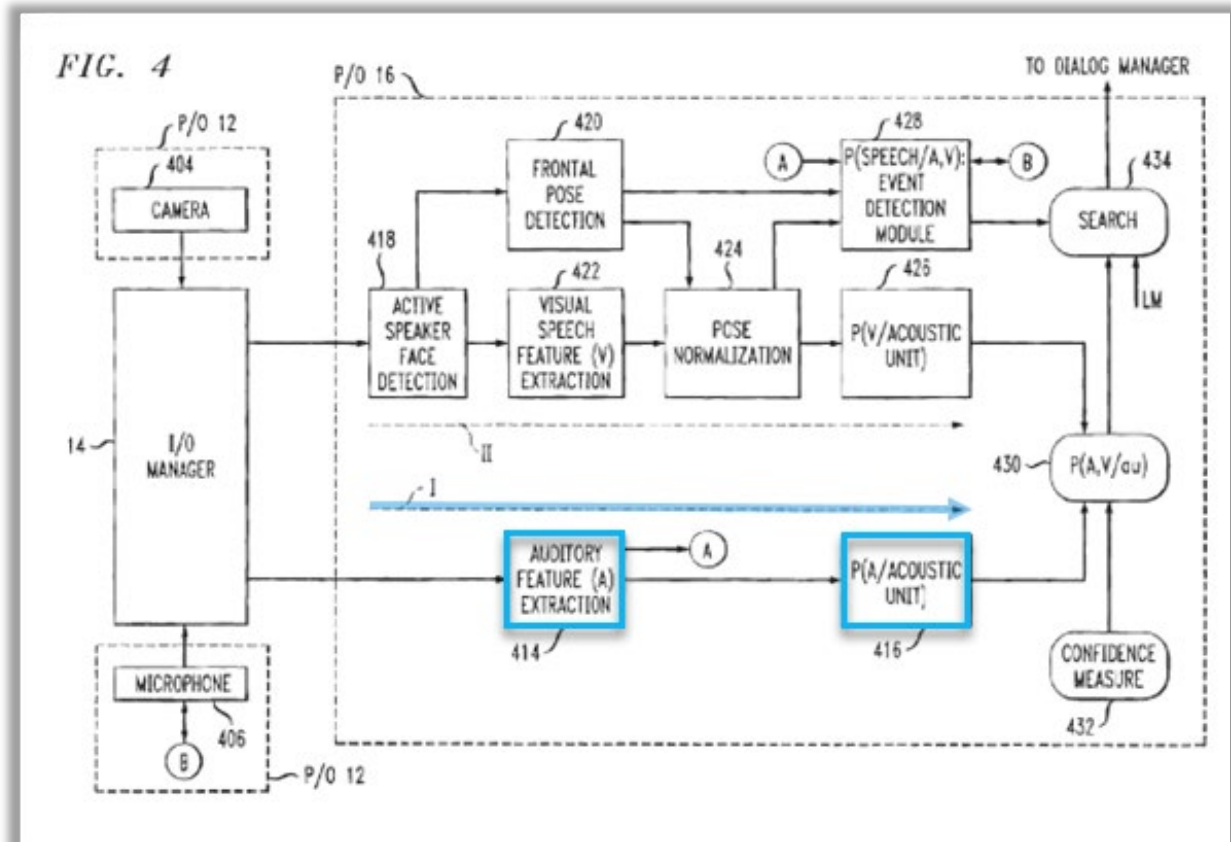
Ex.1005, Figure 1 (annotated).

117. Maes provides additional details of its “*speech recognition engine*.” Maes begins by describing Figure 4 as “an audio-visual speech recognition module that may be employed as **one of the recognition modules**¹³ of FIG. . 1 to perform **speech recognition using multi-modal input data received** in accordance with the invention.” Ex.1005, 10:50–52. Specifically, Figure 4 shows that the audio-visual speech recognition module receives multi-modal input data including an audio signal (e.g., speech provided by the speaker and/or background noise) and a video signal (e.g., the speaker’s face including lip movement and/or background objects in the environment) and then processes the multi-modal input data by passing through several processing modules. (*E.g.*, the video signal is processed by blocks/modules 418, 422, 424, 426 and the audio signal is processed by blocks/modules 414, 416).

118. Maes states that “the processing performed in blocks 414 and 416 may be accomplished via any **conventional acoustic information recognition system** [i.e., “*a speech recognition engine*”] capable of extracting and labeling acoustic feature vectors, e.g., Lawrence Rabiner, Biing-Hwang Juang, “Fundamentals of Speech Recognition,” Prentice Hall, 1993.” Ex.1005, 12:43–49. One such “acoustic information recognition system” that provides additional details on how speech is processed in Maes is described with reference to Figure 9A, as explained in detail

¹³ Maes uses the term “block” and “module” interchangeably. For example, see 12:43-38 and 17:46-50 referring to “414” as “block[] 414” and “module 414.”

below. Therefore, at least “blocks 414 and 416” disclose a “*a speech recognition engine*.”



Ex.1005, Figure 4 (showing audio information processing path (Path I) annotated in blue).

119. Referencing Figures 1 and 9A (reproduced below), Maes provides details of “*a speech recognition engine*” explaining that “via the I/O manager 14 of FIG. 1” “user-provided input data events are ... provided to” “apparatus 900”—an “apparatus for collecting data associated with a voice of a user,” including “a dialog management unit 902 for conduct[ing] a conversation with the user,” that provides

functionality including natural language understanding (NLU), natural language generation (NLG), finite state grammar (FSG), and/or text-to-speech Syntheses (TTS) for machine-prompting the user....” Ex.1005, 39:22–29, 41:13–20, and Figure 9A.

120. Maes continues: “Apparatus 900 ... includes a processing module 910” further including “**a speech recognizer 926** which ... include[s] **a speech recognition module 928**,” and “a speech prototype, language model and grammar database”. Ex.1005, 39:59, 41:35–38. “Apparatus 900 can further include **a post processor 938 ... configured to transcribe user utterances and ... perform keyword spotting thereon.** ... Post processor 938 can employ speech recognizer 926.” Ex.1005, 41:48–59. Therefore, Maes discloses “*a speech recognition engine.*”

ii. Semantic Knowledge-Based Model

121. Maes discloses “*a semantic knowledge-based model.*”

122. Initially, the 468 Patent provides no clear description of what a “*semantic knowledge-based model*” might encompass. Instead, it simply states that the disclosed invention includes “Conversational speech analyzer 804 [that] also may include a semantic knowledge-based model that analyzes the textual message and detects command components.” Ex.1001, 23:64–67. Other than the claims, the

terms “knowledge-based” and “semantic knowledge-based model” appear only once in the Specification of the 468 Patent as stated above.

123. Maes’s system (embodied in apparatus 900) includes a semantic module. Maes explains that apparatus 900 which receives “user-provided input data events” “via the I/O manager 14 of FIG. 1” (Ex.1005, 39:26–29) includes post processor 938 that further includes “a semantic module” (*e.g.*, implementing “*a semantic knowledge-based model*”) stating:

Post processor 938 can also include a **semantic module** (not shown) **to interpret meaning of phrases**. The semantic module could be used by speech recognizer 926 to indicate that some decoding candidates in a list are meaningless and should be discarded/replaced with meaningful candidates.

Ex.1005, 41:58-64. By interpreting meanings of phrases and functioning to indicate that some decoding candidates in a list are meaningless (or, otherwise irrelevant) and should be discarded/replaced with meaningful candidates, Maes’s “semantic module” comprises “*a semantic knowledge-based model*.”

124. A POSITA would have understood that the semantic module in Maes comprises “*a semantic knowledge-based model*” based, for example, on the supported teachings of several references in the art. *See e.g.*, Ex.1017, 468 (defining “*semantic[s]*” as “the study of the meaning in language”); *see also* Ex.1014, Abstract

(describing a “semantic language model” as “a technique that utilizes the semantic structure of an utterance to better rank the likelihood of words composing the sentence. When used in a conversational system, one can dynamically integrate the dialog state and domain semantics into the semantic language model to better guide the speech recognizer executing the decoding process.”). Ex.1017, 429 (describing “knowledge-based speech recognition system” as a system that “specifies explicit acoustic-phonetic rules that are robust enough to allow recognition of linguistically meaningful units and that ignore irrelevant variation in these units”). Ex.1019, 1157 (McGraw-Hill dictionary defining a “knowledge-based system” as “[a] computer system whose usefulness derives primarily from a data base containing human knowledge in a computerized format”).

125. Moreover, for the additional reasons below, the semantic module in Maes comprises “*a semantic knowledge-based model.*”

126. For example, Maes explains that “[a]ll data stored in the data warehouse 912 [which is shown in Figure 9A as coupled to postprocessor 938 that includes the semantic module] can be stored in a format to facilitate subsequent data mining thereon.” Ex.1005, 43:38–41; *see also id.* 39:63–67 (data warehouse 912 ... stores the at least one user attribute, together with at least one identifying indicia, in a form for subsequent data mining thereon.”).

127. Maes continues to explain that as part of the data mining performed on the data stored in the data warehouse 912, “dialect, socioeconomic classification, ... can be estimated based on vocabulary and word series used by the user. **Appropriate key words, sentences, or grammatical mistakes to detect can be compiled via expert linguistic knowledge.**” Ex.1005, 43:16–20; *see also* Ex.1005, 41:48–51 (“a post processor 938 [that includes a semantic module] ... coupled to the data warehouse 912 and which is configured to ... perform keyword spotting thereon”), 42:60–61 (“**a keyword spotter to detect insults**”). In view of the POSITA’s knowledge reflected in the supported teachings of the prior art references above, the functionality of detecting appropriate key words, sentences, insults, and/or grammatical mistakes fall within the purview of a semantic module (such as Maes’s semantic module).

128. Furthermore, as part of detecting key words, sentences, or grammatical mistakes, Maes teaches using “expert linguistic knowledge” (i.e., external knowledge defined by human experts in the field) which makes clear that Maes’s semantic module comprises “*a semantic knowledge-based model.*”

129. Moreover, Maes teaches “modify[ing] behavior of [its] system ... based on at least one user attribute which has been determined based on at least one acoustic feature extracted from the user’s speech” and further that the “[f]eatures can be extracted and **decisions dynamically returned by the models**”—which

demonstrate that the behavior of Maes's system is modified on-the-fly based on the knowledge (e.g., decisions) incorporated by its models (such as those included in the semantic module), further confirming that Maes's semantic module comprises “*a semantic knowledge-based model*.” Ex.1005, 43:53-54 and 44:7–11. Komissarchik, reflecting the POSITA's knowledge, proposed knowledge-based speech recognition system and methods which provide real-time analysis and transcription of spoken utterances. Ex.1018, 8:4–14.

130. Moreover, a POSITA would have understood Maes's models (including its semantic module) and techniques, which include acoustic feature extraction, labeling the extracted vectors with previously stored phonemes, natural language processing of spoken commands, use of training on word-dependent features, and language model (LM) training are—“*semantic knowledge-based*.” See e.g., Ex.1005, 1:53–57, 2:1–4, 19:7–21, and 11:64–12:28. Therefore, Maes teaches or suggests “*a semantic knowledge-based model*.”

131. A “*semantic knowledge-based model*” was well-known in the prior art. For example, Komissarchik, reflecting the POSITA's knowledge, defines “knowledge-based systems” as “involv[ing] the application of acoustic, phonetic and natural language processing (‘NLP’) information theory to the speech recognition process” and describes its invention as “knowledge-based speech recognition

apparatus and methods” that provide “semantic control of the recognition results.”
Ex.1018, 2:39–43 and 6:45–47.

132. Additionally, Maes’s system (e.g., comprising the semantic module) provides “**functions needed to adapt ... to the capabilities and constraints of the device, application and/or user preferences,**” uses “a-priori knowledge, ... information about the user,” and Maes teaches training methods based on a range of “**speaker profile features (accent, dialect, gender, age, speaking rate ...**” which are consistent with Patent Owner’s description of “*semantic knowledge-based model*” in its Complaint. Ex.1005, 36:44–48, 37:22–28, and 19:7–19; Ex.1007, ¶133 (“Erica uses a semantic knowledge-based model because ‘Erica is able to consider a range of data within Bank of America, like your cash flow, balances, transaction history and upcoming bills...’ and ‘Erica learns from the conversations with you and other Bank of America customers’”).

133. Therefore, Maes’s semantic module comprises “*a semantic knowledge-based model.*”

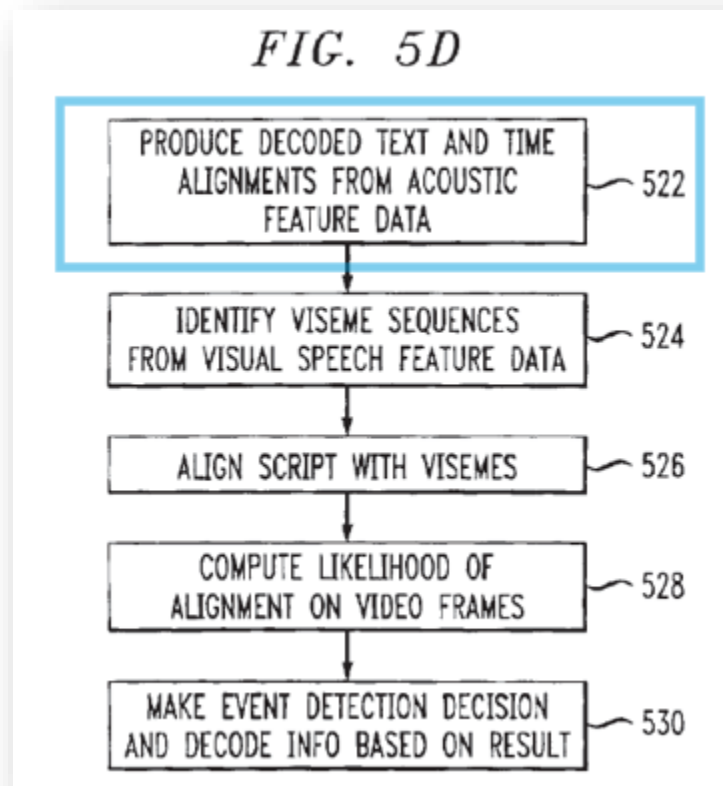
iii. Creating a Speech-Based Transcription

134. Maes expressly teaches “**one or more recognition engines of FIG. 1,**” “**need[s] data ... (e.g., grammar, acoustic models).**” Ex.1005, 34:62–66. Therefore, speech recognition module 928 (“*a speech recognition engine*”) uses data provided by a model, such as “*a semantic knowledge-based model,*” e.g.,

implemented in Maes's "semantic module." *See also* Ex.1005, 41:61–64 (explaining the semantic module used by speech recognizer 926 which includes the speech recognition module 928). Additionally, Maes discloses using "historical interaction information," which, a POSITA would have understood would be included in a model. *See generally* Ex.1005, 7:40–45 and 24:31–45.

135. Referencing Figure 4, Maes explains that "the processing performed in blocks 414 and 416"—(e.g., "*speech recognition engine*") "produces" ("*creat[es]*"), for instance, "**decoded text**" ("*a speech-based transcription*") (Ex.1005, 21:43–47), and further, referencing Figure 9, Maes explains that the "semantic module" functions to "indicate that some **decoding candidates** in a list are meaningless and should be discarded/replaced with meaningful candidates" (Ex.1005, 41:61–64). Thus, Maes explains that the "**decoded text**" produced by the processing performed in blocks 414 and 416 is the same "decoding candidates" Maes describes in connection with Figure 4, further confirming Figure 9A and its corresponding disclosure provide additional details to, among other things, the "*speech recognition engine*" described in Figure 4. For example, in Figure 9A, Maes identifies "semantic module," "speech recognizer 926," "post processor 938," "speech recognition module 928" and other components/functionalities. Therefore, Maes discloses "*creating a speech-based transcription ... using a *speech recognition engine* and a semantic knowledge-based model.*"

136. Referencing Figure 5D, Maes discloses that “in step 522, the uttered speech to be verified may be **decoded** by classical speech recognition techniques so that a decoded script and associated time alignments are available. This is accomplished using the feature data from the acoustic feature extractor 414.” Ex.1005, 21:43–47. Figure 5D (block 522) discloses that “decoded text” (or, script) (e.g., “*a speech-based transcription*”) is, therefore, “produce[d]” (“*create[ed]*”) “using the feature data from the acoustic feature extractor 414.” *See generally* Ex.1005, 21:43–47.



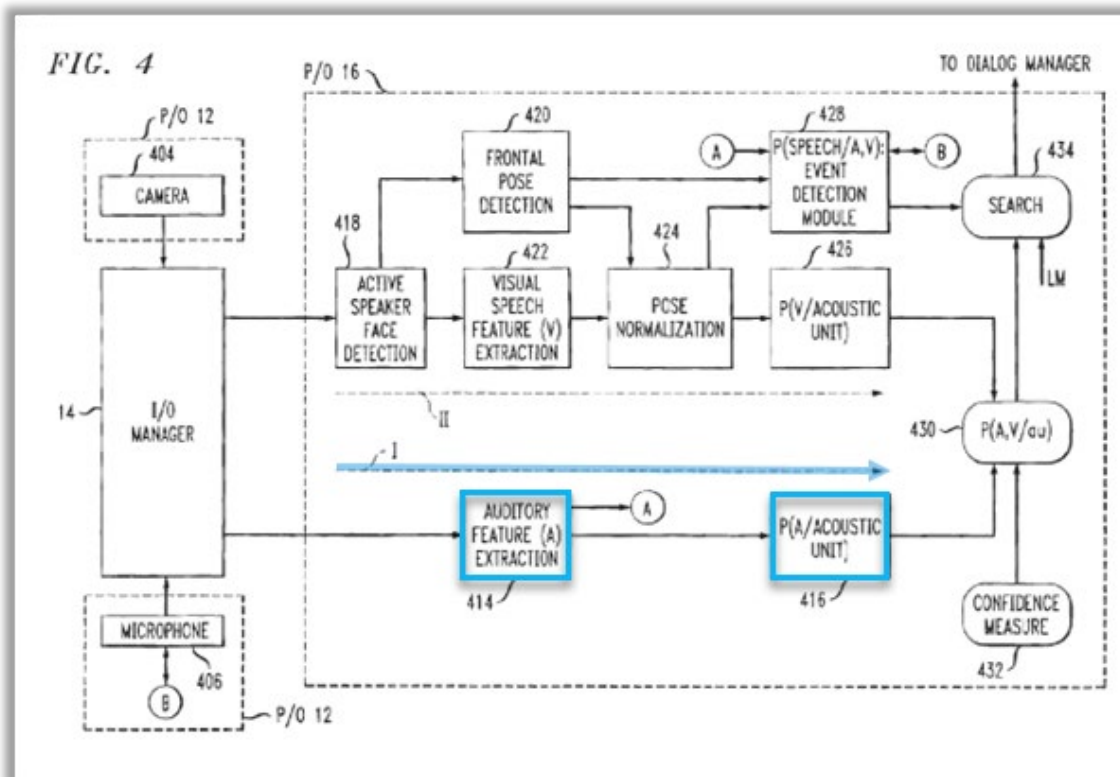
Ex.1005, Figure 5D (block 522). A POSITA would have understood that the “decoded script” produced in step 522 is essentially the tangible result of a speech recognition system’s interpretation of spoken language and provides a text representation of the multi-modal content. *E.g.*, Ex.1005, 20:61-64 (“If the buffered data is tagged as speech, in step 518, the buffered data is sent on through the acoustic path so that the buffered data may be recognized, in step 520, so as to yield a **decoded output**.”); *see also* 20:23-33.

137. With reference to Figures 1 and 4 (reproduced below), Maes provides additional details explaining the step of “*creating a speech-based transcription*”: Maes begins by describing Figure 4 as “an audio-visual speech recognition module that may be employed as **one of the recognition modules¹⁴ of FIG. 1 to perform speech recognition using multi-modal input data received** in accordance with the invention.” Ex.1005, 10:50–52.

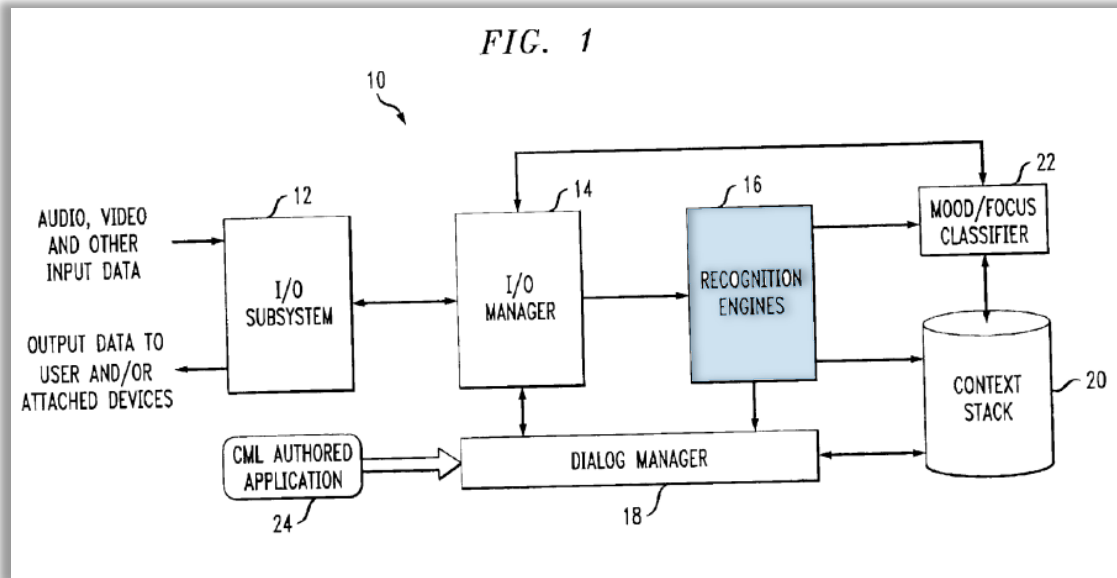
138. Specifically, Figure 4 shows that the audio-visual speech recognition module receives multi-modal input data including an audio signal (e.g., speech provided by the speaker and/or background noise) and a video signal (e.g., the speaker’s face including lip movement and/or background objects in the environment). For example, the auditory feature extractor 414 “receives an audio or

¹⁴ Maes uses the term “block” and “module” interchangeably. For example, see 12:43-38 and 17:46-50 referring to “414” as “block[] 414” and “module 414.”

speech signal and, as is known in the art, extracts spectral features from the signal at regular intervals. The spectral features are in the form of acoustic feature vectors (signals) which are then passed on to a probability module 416.” Ex.1005, 11:64–12:2. In the audio-visual speech recognition module shown in Figure 4, the acoustic feature vectors are denoted by the letter “A” and “[a] phantom line denoted by Roman numeral I represents the processing path the audio information signal takes within the module.” Ex.1005, 11:55–64.



Ex.1005, Figure 4 (showing audio information processing path (Path I) annotated in blue).



Ex.1005, Figure 1 (annotated).

139. Maes explains how acoustic features are extracted:

magnitudes of discrete Fourier transforms of samples of speech data in a frame are considered in a logarithmically warped frequency scale. Next, these amplitude values themselves are transformed to a logarithmic scale. The latter two steps are motivated by a logarithmic sensitivity of human hearing to frequency and amplitude. Subsequently, a rotation in the form of discrete cosine transform is applied. One way to capture the dynamics is to use the delta (first-difference) and the delta-delta (second-order differences) information.

Ex.1005, 12:12-28.

140. Maes further describes sampling the signal prior to extracting the acoustic feature vectors. For example, Maes explains that:

Before acoustic vectors are extracted, the speech signal may be sampled at a rate of 16 kilohertz (kHz). A frame may consist of a segment of speech having a 25 millisecond (msec) duration. In such an arrangement, the **extraction process preferably produces 24 dimensional acoustic cepstral vectors** via the process described below. Frames are advanced every 10 msec to obtain succeeding acoustic vectors. Note that other acoustic front-ends with other frame sizes and sampling rates/signal bandwidths can also be employed.

Ex.1005, 12:2–12.

141. Maes provides several examples of acoustic feature vectors, or alternatively cepstral vectors in connection with the step of extracting acoustic feature vectors. Maes describes “[i]t is to be understood that other variations on [acoustic] features may be used, e.g., [linear prediction coefficients] LPC cepstra, [Perceptual Linear Prediction] PLP and that the invention is not limited to any particular type.” As yet another example of an acoustic feature vector, Maes describes “MEL cepstra.” Ex.1005, 23:40–43, 39:55–58.

142. Referring to the example in which Maes’s multi-modal conversational computing system is employed within a vehicle, Maes explains the “*creating ...*” step as follows. The raw multi-modal input data relating to an I/O event (e.g., representative of the user’s spoken utterance and lip movement accompanying the spoken utterance) received by the multi-modal conversational computing system is abstracted into a recognizable form. For example, “[t]he I/O manager [of the multi-

modal conversational computing system] receives the raw multi-modal data and **abstracts the data into a form that represents ... a spoken utterance ...** As is known, **a data abstraction operation may involve generalizing details associated with all or portions of the input data so as to yield a more generalized representation of the data for use in further operations.**” Ex.1005, 6:21–38. Thus, as a result of processing by several modules, the system produces decoded text/script corresponding to the user’s spoken utterance, for example, of saying “turn it on.”

143. Therefore, Maes teaches or suggests this claim element.

- f. **[19.3.1] wherein the semantic knowledge-based model includes a personalized cognitive model derived from one or more prior interactions between the identified user and the conversational voice user interface,**

144. Maes teaches or suggests this claim element.

145. The “*identified user*” in this claim element is the outcome of the “*identifying ...*” step explained in [19.2], and as explained in [19.3.0], Maes describes a semantic module comprising “*a semantic knowledge-based model*”.

146. Initially, the 468 Patent describes that a “*cognitive model*” as “a model derived from a user’s interaction pattern with the system and may be used to predict what actions the user may take next in time [and] assisting with speech recognition and/or question or command recognition.” Ex.1001, 24:1–11.

147. As explained below, cognitive models using information derived from user (or users’) historical interactions with a system to assist in speech recognition were well-known at the time of the 468 Patent.

148. For example, U.S. Patent Application Publication No. 2002/0087315 to Lee (“Lee” or Ex.1008) describes techniques for assisting speech recognition accuracy using information collected about multiple users. Ex.1008, Abstract, [0002], [0016]. Lee describes a “multi-scan control unit” that may use “recognition assisting databases” to aid in recognition of speech input, including a database that “maintains a personalized profile that have already been recognized for a particular user or for users who have previously submitted requests which are similar to the request at hand.” Ex.1008, [0016]. Lee also describes a “popularity engine” that “contains data compiled from multiple users’ histories that has been calculated for the prediction of likely user requests.” Ex.1008, [0016]. Lee explains that “[t]his increases the accuracy of word recognition” because it enables the system to “scan[] for the most probable match for words in the user utterance based on past requests.” Ex.1008, [0024]–[0026].

149. Maes teaches its models utilize user attributes. For example, Maes teaches that its system can include a “a dynamic classification module ... which performs dynamic classification of the user” “based on at least one user attribute ... determined based on at least one acoustic feature extracted from the user’s speech.”

Ex.1005, 44:5–11. The “user attribute can include ...: gender **of the user**, age **of the user**, accent **of the user**, native language **of the user**, dialect **of the user**, socioeconomic classification **of the user**, educational level **of the user**, and emotional state **of the user**.” Ex.1005, 39:50–55.

150. Considering, for instance, native language as a user attribute, Maes teaches that “information about the native language of the speaker can be added to define each accent/native language model” and further that “**at the creation of the models for each native language, one employs a speaker or speakers who are tagged with that language as their native language.**” Ex.1005, 40:18–23. Thus, Maes’s system classifies the user’s native language based on a native language model created using a speaker who is tagged with that language as his/her native language. Maes’s teachings of employing a speaker for creating a model of a native language of that user, therefore, serve as describing “*a personalized cognitive model derived from one or more prior interactions between the identified user and*” “the (I/O) subsystem 12 ... compris[ing] one or more microphones for capturing audio input data” (“*the conversational voice user interface*”). Ex.1005, 6:5–8.

151. Furthermore, Maes explains that “[r]ecently, efforts have been initiated to develop appropriate confidence measures” comprising methods using “decision trees trained on word-dependent features (amount of **training utterances**, minimum and average triphone occurrences, **occurrence in language model training**, ... as

well as **speaker profile features (accent, dialect, gender, age, speaking rate, identity).**” Ex.1005, 19:3–20. Thus, in Maes, examples of word-dependent features include amount of **training utterances, occurrence in language model training, as well as speaker profile features (such as accent, dialect, gender, age, speaking rate, identity).** From the above-mentioned passage, a POSITA would have understood that in the context of speech recognition systems, training a language model involves prior interactions between a user and a system and furthermore, that utterances from a single user can be used for the training, and therefore Maes’s teachings related to training utterances and occurrence of certain utterances in language model training, and using speaker profile features (e.g., accent, dialect, gender, age, speaking rate, identity—which are speaker-specific) serve as describing “*a personalized cognitive model derived from one or more prior interactions between the identified user and*” (I/O) subsystem 12 comprising microphones (“*the conversational voice user interface*”). Ex.1005, 19:3–20 and 6:5–8.

152. Maes teaches using “*a personalized ... model derived from one or more prior interactions between the identified user and*” Maes’s system comprising I/O subsystem 12 (“*the conversational voice user interface*”). In Maes’s system, “the input acoustic data is compared to determine if the data matches closely enough with stored models. If the comparison yields a close enough match, the person uttering the speech is verified. The match is accepted or rejected by comparing the match

with competing models. These models can be selected to be similar to the claimant speaker or be speaker independent (i.e., a single or a set of speaker independent models). If the claimant wins and wins with enough margin (computed at the level of the likelihood or the distance to the models), we accept the claimant. Otherwise, the claimant is rejected. ... [A]t enrollment, **the input speech is collected for a speaker to build the mixture gaussian model M_i that characterize each speaker.**” Ex.1005, 24:31–45. “Thus, the above approach finds the closest matching model and the person whose model that represents is determined to be the person whose utterance is being processed.” Ex.1005, 24:28–30.

153. Therefore, in view of the above-mentioned disclosures, Maes teaches or suggests “*a personalized cognitive model derived from one or more prior interactions between the identified user and the conversational voice user interface.*”

154. To the extent Patent Owner contends that Maes does not teach or suggest that “*the semantic knowledge-based model includes a personalized cognitive model ...*,” a POSITA would have found it obvious to modify Maes’s “semantic module” (“*semantic knowledge-based model*”) to include Maes’s teachings of using user-specific data for making personalized interpretations of a user’s speech/audio input based on using such data as a model within Maes’s “semantic module.” Ex.1005, 6:5–15, 19:3–20, 24:28–45, 40:18–23, 44:5–11, and 41:59–64. A POSITA would be motivated to make this modification because it would improve the

capabilities of the semantic module functioning to indicate meaningless/irrelevant decoding candidates in a given user’s speech by adding user-specific keywords, variables, parameters, sentences, words, and phrases, thereby enhancing the quality of the transcribed speech output. Ex.1005, 41:61–64 and 11:42–45. Further, because the modification is based on Maes’s own teachings, a POSITA would have had a reasonable expectation of success in doing so and without undue experimentation.

155. Therefore, in view of the above disclosures, Maes teaches or suggests that “*the semantic knowledge-based model includes a personalized cognitive model.*”

156. To the extent Patent Owner contends that Maes does not teach or suggest “*a cognitive model ...*,” Ittycheriah (incorporated by reference in Maes) does. Ex.1005, 19:43–49 (incorporating by reference specific features of Ittycheriah identified as U.S. Pat. No. 5,937,383); *see* § IX.B.

157. Ittycheriah teaches making use of user-specific data (e.g., scores and words decoded in accordance with HMMs from previous interactions with a user) stored in a user-specific partition within a database as hypothesis for decoding a word in a current interaction with that user and therefore—such database used as a model provides functionality of “*a personalized cognitive model derived from one or more prior interactions between the ... user and [a voice user interface]*,” such as I/O subsystem 12 comprised in Maes’s system. Ittycheriah teaches a cache

database “storing keywords comprising vocabulary such as names and/or commands frequently used by a user” such that:

When the user first inputs a name or a command, the acoustic front-end ... generates a set of acoustic features which are then provided to the classical speech recognition engine The word is decoded in accordance with the Hidden Markov models and the decoded word, along with the different scores (discussed above) associated with the word, are stored in the cache database ... preferably partitioned for that particular user. [R]ather than a large cache database partitioned for each system user, individual cache databases may be formed for each user.

Next, when the same user enters another speech segment, the following operation is performed. First, each decoded keyword in the cache database ... corresponding to the same user is considered as a hypothesis with respect to the newly entered speech segment.

Ex.1006, 4:19–32 and 1:55–58. By describing that the cache database is partitioned for that particular user and further that for the newly-entered speech previously decoded keywords (in the cache database) of the same user are considered, Ittycheriah, therefore, expressly teaches that its cache database (“*a personalized cognitive model*”) is “*derived from one or more prior interactions [with] the ... user.*”

158. Ittycheriah further explains how the user-specific data from prior interactions with a user is used to test a hypothesis for decoding a word in a current interaction with that user. Ittycheriah’s method “align[s] the baseforms of the words stored in the cache and get a score for all these comparisons. The scores are

compared with what is stored in the cache, to decide if one hypothesis is acceptable or not.” Ex.1006, 3:6–9. A POSITA would have found it obvious to modify Maes’s system (e.g., the semantic module) to include Ittycheriah’s teachings of a cache database storing scores and words decoded in accordance with HMMs from previous interactions with a user (“*a personalized cognitive model*”) and Maes’s input/output (I/O) subsystem, such that the database is used as a model within the semantic module. Ex.1005, 12:12–28; Ex.1006, 3:56–60 and 4:19–53. A POSITA would be motivated to make this modification because of the reasons above (*see* § IX.B).

159. Therefore, Maes teaches or suggests this element.

- g. **[19.3.2] [wherein the semantic knowledge-based model includes ...] a general cognitive model derived from one or more prior interactions between a plurality of users and the conversational voice user interface,**

160. Maes teaches or suggests this claim element.

161. As explained above in connection with [19.3.0] and [19.3.1], Maes describes a “*semantic knowledge-based model*.” And, as explained in [19.1.0], Maes’s input/output (I/O) subsystem comprising one or more microphones for capturing audio input data constitutes “*the conversational voice user interface*.”

162. First, Maes describes features such as acoustic feature vectors included in processed audio information that are used to “*create[] a speech-based transcription*” “**can be extracted and decisions dynamically returned by the**

models,” which explicitly discloses that Maes uses more than one “*model*.”
Ex.1005, 43:53–54.

163. Maes provides an example of a multi-speaker model, as explained below. For example, Maes’s system includes “a dynamic classification module ... which performs dynamic classification of the user” such that the system is “configured to modify behavior of the voice system ... **based on at least one user attribute** ... determined based on at least one acoustic feature extracted from the user’s speech.” Ex.1005, 44:5–11. Maes provides several examples of user attributes such as gender, age, accent, native language, and others. *See* Ex.1005, 39:50–55.

164. Considering native language as an attribute, Maes teaches that that “**at the creation of the models for each native language, one employs a speaker or speakers who are tagged with that language as their native language**” which explicitly discloses that Maes uses “*a ... model derived from one or more prior interactions ... [with] a plurality of users.*” Ex.1005, 40:18–23. Thus, in Maes, the semantic module classifies the user’s native language based on a native language model created using multiple speakers who are tagged with that language as their native language, and therefore, the native language model serves as “*a general cognitive model derived from one or more prior interactions between a plurality of users and the conversational voice user interface.*”

165. For the additional reasons below, Maes teaches using “*a ... model derived from one or more prior interactions ... [with] a plurality of users.*” In Maes’s system, “[s]peaker verification may be performed in a similar manner, ... the input acoustic data is compared to determine if the data matches closely enough with stored models. ... These models can be selected to be similar to the claimant speaker or be speaker independent (i.e., a single or a set of speaker independent models). ... [A]t enrollment, the input speech is collected for a speaker to build the mixture gaussian model M_i that characterize each speaker.” Ex.1005, 24:31–45. “Thus, the above approach finds the closest matching model and the person whose model that represents is determined to be the person whose utterance is being processed.” Ex.1005, 24:28–30. Therefore, Maes provides another example of a multi-speaker model.

166. To the extent Patent Owner contends that Maes does not suggest or teach that “*the semantic knowledge-based model includes ... a general cognitive model ...*,” a POSITA would have found it obvious to modify Maes’s “semantic module” to include Maes’s teachings of multi-speaker models. Ex.1005, 24:28–45, 40:18–23, 44:5–11, and 39:50–55. A POSITA would be motivated to make this modification because it would improve the capabilities of the semantic module functioning to indicate meaningless/irrelevant decoding candidates in a given user’s speech by adding keywords, variables, parameters, sentences, words, and phrases

collected from a pool of users, thereby improving the quality of the transcribed speech output. Ex.1005, 41:61–64 and 11:42–45. Further, because the modification is based on Maes’s own teachings, a POSITA would have had a reasonable expectation of success in doing so.

167. Additionally, Maes, based on Ittycheriah (see Ex.1005, 19:43–49) teaches or suggests “*a general cognitive model derived from one or more prior interactions ... [with] a plurality of users and ... [a device].*”

168. For example, Ittycheriah teaches a speaker classification approach (“*a general cognitive model*”) that involves a multi-user database storing keywords and scores from past history of a class of speakers (*derived from one or more prior interactions ... [with] a plurality of users and ... [a device]*”). Ex.1006, 5:5–19 (describing the speaker classification approach, including techniques based on “past history ... associated with ... speakers in a class” such that “the acceptance or rejection decision is based on past history and an average previous likelihood associated with other speakers within that class”). A POSITA would have found it obvious to modify Maes’s system (e.g., the semantic module) to include Ittycheriah’s teachings of a speaker classification approach (e.g., “*a general cognitive model*”) that involves storing in a cache database keywords and scores from past history of a class of speakers interacting with Maes’s input/output (I/O) subsystem (“*the conversational voice user interface*”), such that the database is used as a model

within the semantic module. Ex.1006, 5:5–19; Ex.1005, 41:59–64, 6:21–29, 5–8.

A POSITA would be motivated to make this modification because of the reasons above (*see* §IX.B).

169. Therefore, Maes teaches or suggests this claim element.

h. **[19.3.3] [wherein the semantic knowledge-based model includes ...] an environmental model derived from an environment of the identified user and the conversational voice user interface;**

170. Maes teaches or suggests this claim element.

171. As explained above in connection with elements [19.3.0], [19.3.1], and [19.3.2], Maes describes a “*semantic knowledge-based model*.” And, as explained in [19.1.0], Maes’s input/output (I/O) subsystem comprising one or more microphones for capturing audio input data constitutes “*the conversational voice user interface*.”

172. Initially, the 468 Patent describes an “*environmental model*” as “includ[ing] information associated with the user’s environment and surroundings. [E.g.,] [t]he information may include the type of environment that a user is in (e.g., quiet or noisy).” *See generally* Ex.1001, 24:22–36. As explained below, Maes teaches or suggests “*an environmental model*” analogous to the 468 Patent.

173. First, as explained in connection with element [19.3.2], Maes describes features such as acoustic feature vectors included in processed audio information that are used to “*creat[e] the speech-based transcription*” that “**can be extracted**

and decisions dynamically returned by the models,” which explicitly discloses that Maes uses more than one “*model*” (e.g., “*a personalized cognitive model ... a general cognitive model ... and an environmental model*”). Ex.1005, 43:53–54.

174. Maes teaches or suggests “*an environmental model derived from an environment of the identified user*” because Maes’s methods “**provide[] better modeling of background noise, that is, when no speech is detected, silence is detected.**” Maes, 22:9–12. “For instance, if the **surrounding environmental noise level** is particularly high, ... there is more of a chance that ... the acoustic decoding ... contain errors.” Ex.1005, 19:62–66. Maes notes that one problem of conventional speech recognition systems is their “inability to discriminate between extraneous audible activity, e.g., background noise or background speech not intended to be decoded, and speech that is indeed intended to be decoded [which] causes ... misfiring of the system and ‘junk’ recognition.” Ex.1005, 20:23–29.

175. Further, Maes’s system can be deployed in different environments. Specifically, Maes explains that its system “is not limited to a particular application” (i.e., usage, or setting in which it is employed). Ex.1005, 4:23–25. Referencing the in-vehicle system example, Maes states that the system “may be used to **detect a distracted or sleepy operator based on detection of ... speech that indicates distraction or sleepiness (by audio input)**” Ex.1005, 4:29–36. Maes’s system includes functionality for detecting/sensing visual activity and audio activity

associated with “*an environment of the ... user*” for making enhanced interpretations of a user’s commands/speech input. For instance “cameras and microphones may be strategically placed throughout the vehicle in order to attempt to **fully capture all visual activity and audio activity** that may be necessary for the system to make ambiguity resolution determinations.” Ex.1005, 6:16–21. Maes discloses that its in-vehicle system “detects that the user is warm or cold” and accordingly “cause the temperature to be adjusted for each passenger. If the passenger is tired, the system may cause the adjustment of the seat, increase the music volume, ...” Ex.1005, 5:24–28.

176. As a second illustrative example, Maes’s audio-visual speech recognition module is used in “in the context of broadcast news [with] a newsperson speaking at a location where there is arbitrary activity and noise in the background.” Ex.1005, 11:12–15. In the broadcast news context, Maes explains that its audio-visual speech recognition module “process[es] arbitrary content video which may not only contain faces but may also contain **arbitrary background objects in a noisy environment.**” Ex.1005, 11:8–12. By “**capturing**” environmental data comprising “**video input data representing the environment in which [the system] ... is deployed,**” such data enables Maes’s system to “locate and track a face and, more particularly, a mouth, to determine what is relevant visual information” within an arbitrary background. Ex.1005, 5:60–64, 11:15–18. Indeed,

Maes teaches the use of different types of sensors/cameras to capture data representing the environment, e.g., infra red (IR) sensors, radio frequency (RF) sensors, and/or a combination (mix) of them. “[A] general principle of the present invention of the combination of **modality through at least two different sensors** (and actuators for outputs) to disambiguate the input, and guess the mood or focus, can be generalized to any such combination” *See* Ex.1005, 6:1–4. *See also* Gibbon defining “*an environment*” as “the total context in which a system such as a recognition ... is located” and further that “*an environment*” is described by “factors which might affect the performance of the system (such as high background noise).” Ex.1017, 406. And, thus the disclosures in Maes that demonstrate “modeling” of a background and/or noise in “*an environment of the identified user*,” therefore serve as “*an environmental model*”

177. Maes further describes that “conventional speech recognition with, for example, silence detection, speech activity detection and/or noise compensation can be used. Thus, **background noise is not recognized as though it were speech**, and the starts of utterances are not mistakenly discarded.” Ex.1005, 13:3–8.

178. Thus, Maes describes multiple “conventional” speech recognition techniques which were well known at the time of the 468 Patent, all assisting the system of Maes in making enhanced interpretations of a user’s speech/audio input based on the environment of the user, as, for example, “[t]he user may not be

speaking to the system, but to another person present or on the telephone,” which constitutes “information associated with the user’s environment and surroundings” that “enhance[s] the interpretation of questions and/or commands” made by an “*identified user*” in the process of “*creating a speech-based transcription of [a] natural language utterance.*”

179. To the extent that Patent Owner contends Maes does not teach or suggest an “*environmental model*,” Maes’s description of using data or information representing the user’s environment and surroundings (*e.g.*, to “enhance the interpretation of questions and/or commands”) is sufficient to inform a POSITA that such methodology was well known at the time of the 468 Patent and taught or suggested by Maes.

180. Further and to the extent Patent Owner contends that Maes does not teach or suggest that “*the semantic knowledge-based model includes ... an environmental model ...*,” a POSITA would have found it obvious to modify Maes’s “semantic module” (comprising a “*semantic knowledge-based model*”) to include Maes’s teachings of using data captured from the user’s environment data (*e.g.*, in making enhanced interpretations of a user’s speech/audio input) as a model within Maes’s “semantic module.” Ex.1005, 6:21–29, 13:3–8, 4:29–36, 41:59–64. A POSITA would be motivated to make this modification because it would improve the capabilities of the semantic module functioning to indicate

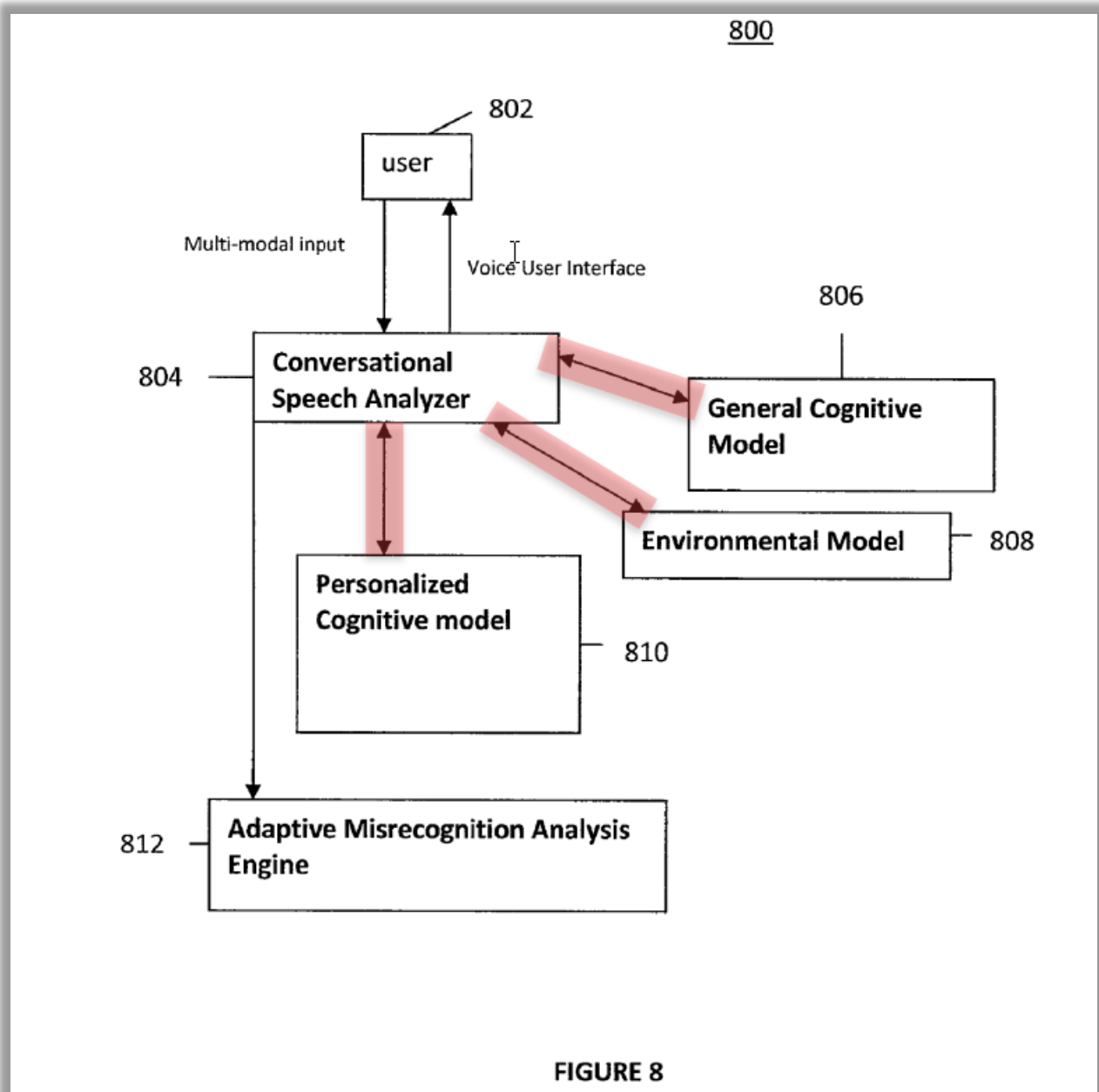
meaningless/irrelevant decoding candidates in a given user’s speech by adding keywords, variables, parameters, sentences, words, and phrases collected from a pool of users, thereby improving the quality of the transcribed speech output. Ex.1005, 41:61–64 and 11:42–45. Further, because the modification is based on Maes’s own teachings, a POSITA would have had a reasonable expectation of success in doing so.

181. Moreover, the teachings or suggestions in Maes of “*an environmental model*” are consistent with Patent Owner’s apparent interpretation of this claim limitation in its Complaint in which Patent Owner has mapped “*an environmental model*” to “information concerning the environment of the user. For example, ... when a user asks Erica “Where’s the nearest branch”, the semantic knowledge-based model can access information concerning the environment of the user (e.g., the user’s current location).” Ex.1007, ¶119.

182. Therefore, Maes teaches or suggests “[*wherein the semantic knowledge-based model includes ...] an environmental model derived from an environment of the identified user and the conversational voice user interface.*”

183. Claim 19 requires that “*the semantic knowledge-based model includes a personalized cognitive model ..., a general cognitive model ..., and an environmental model,*” and for the reasons explained in [19.3.0], [19.3.1], [19.3.2], and [19.3.3], collectively, Maes teaches or suggests this element.

184. The 468 Patent provides no description of what a “*semantic knowledge-based model*” might encompass. See analysis in [19.3.0]. While the 468 Patent separately mentions of “*a general cognitive model*,” “*an environmental model*,” “*a personalized cognitive model*,” nowhere does it state that “*a semantic knowledge-based model*”—“**includes**” “*a general cognitive model*,” “*an environmental model*,” “*a personalized cognitive model*.” At best, the 468 Patent states that “*a general cognitive model*,” “*an environmental model*,” “*a personalized cognitive model*” “**coupled to**” conversational speech analyzer 804. Figure 8 of the 468 Patent (below) further confirms the “coupled to” description in the Specification. Neither is there any description supporting that conversational speech analyzer 804 is same as “*semantic knowledge-based model*,” nor is there any mention of conversational speech analyzer 804 “**including**” “*a general cognitive model*,” “*an environmental model*,” “*a personalized cognitive model*.” Ex.1001, 23:48–55 and Figure 8.



Ex.1001, Figure 8 (annotated).

185. Additionally, other than parroting the claim language, Dialect’s Complaint does not make clear what a “*semantic knowledge-based model*” encompasses. See Ex.1007, ¶¶116–119.

i. **[19.4] merging the speech-based transcription and the non-speech-based transcription to create a merged transcription;**

186. Maes teaches or suggests this claim element.

187. As explained earlier in connection with elements [19.1.1] and [19.3.0], Maes discloses “the visual phonemes (visemes) sequence” (“*the non-speech-based transcription*”) and “decoded text” (or, script) (“*the speech-based transcription*”).

188. Further, with reference to Figure 4 (reproduced below), Maes discloses that “[a] phantom line denoted by Roman numeral I represents the processing path the audio information signal takes within the module, while a phantom line denoted by Roman numeral II represents the processing path the video information signal takes within the module.” Ex. 1005, 23:1–5.

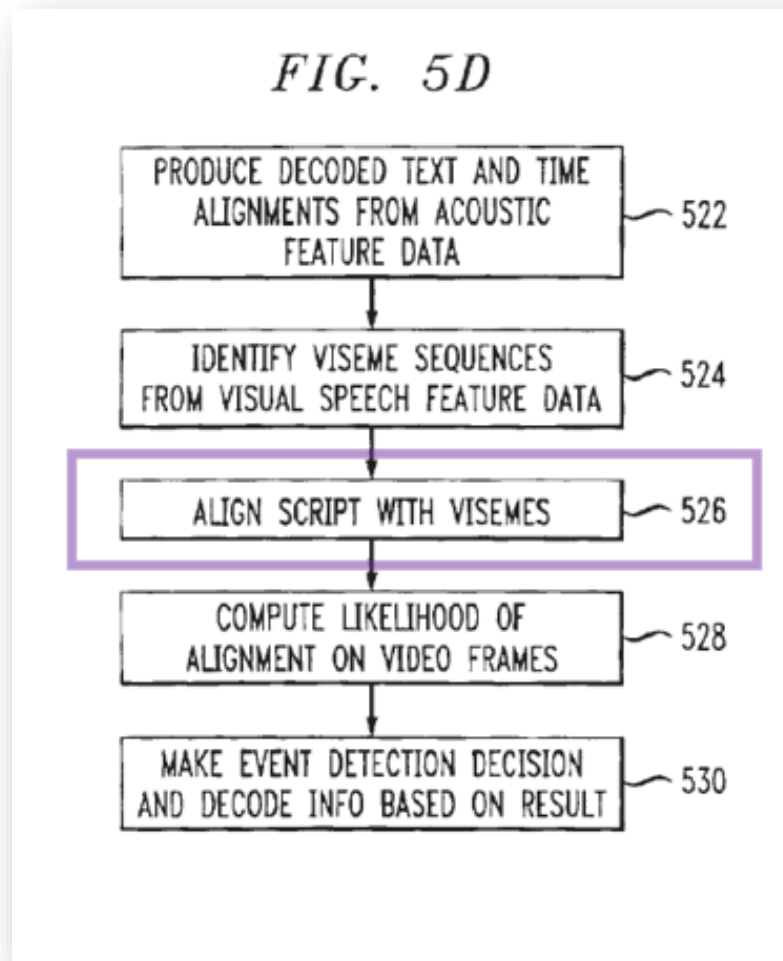
189. Maes teaches that the “multi-modal input data” (“*multi-modal natural language input*”) “including” “an audio signal” (e.g., speech provided by the speaker and/or background noise) (“*a natural language utterance*”) “and” “a video signal” (“*a non-speech input*”), (which can include the speaker’s face including lip movement and/or background objects in the environment) goes through several processing modules. The video signal is processed by modules 418, 422, 424, 426 and the audio signal is processed by modules 414, 416.

190. Maes discloses “[t]he I/O manager receives the raw multi-modal data [in the form of audio input data, video input data, as well as other types of input data]

and abstracts the data into a form that represents ... a spoken utterance [A] data abstraction operation may involve generalizing details associated with all or portions of the input data so as to yield a more generalized representation of the data for use in further operations.” Ex.1005, 6:32–38. Consequently, as outcome of such operations or processing, “the visual phonemes (visemes) sequence” (“*the non-speech-based transcription*”) and “the decoded text” (or, script) (“*the speech-based transcription*”) are generated. And, with reference to processing step 526 of Figure 5D, Maes discloses:

the script is aligned with the visemes. A rapid (or other) alignment may be performed in a conventional manner in order to attempt to synchronize the two information streams [i.e, the audio and video information streams].

Ex.1005, 21:51–54.



Ex.1005, Fig. 5D (block 526).

191. Maes explains the purpose of aligning/synchronizing the audio and video information streams. “A goal associated with utterance verification is to make a determination that the speech used to verify the speaker in the audio path I and the visual cues used to verify the speaker in the video path II correlate or align. This allows the system to be confident that the speech data that is being used to recognize the speaker is actually what the speaker uttered.” Ex.1005,

29:30–36. Further, Maes explains that aligning/synchronizing the audio and video information streams has “many advantages.” “For example, from the utterance verification, it can be determined whether the user is lip synching to a pre-recorded tape playback to attempt to fool the system. Also, from utterance verification, errors in the audio decoding path may be detected. Depending on the number of errors, a confidence measure may be produced and used by the system.” Ex.1005, 29:37–43.

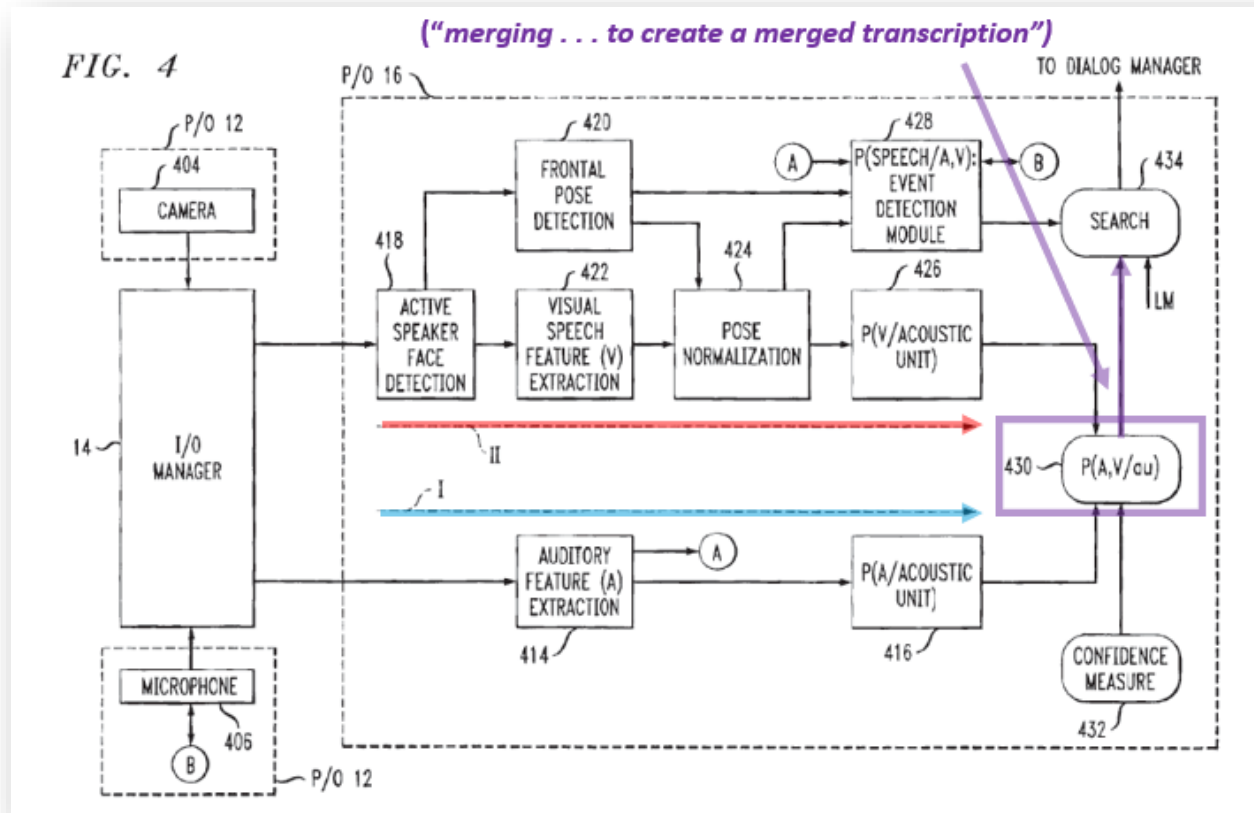
192. Therefore, by teaching synchronization of the two information streams (i.e, the audio and video information streams) or, more specifically, alignment of “the decoded text” (or, script) (“*the speech-based transcription*”) with “the visual phonemes (visemes) sequence” (“*the non-speech-based transcription*”), Maes discloses “*merging the speech-based transcription and the non-speech-based transcription to create a merged transcription.*”

193. Referencing the example of an in-vehicle system, Maes discloses:

The dialog manager would ... receive **the results of the recognized events associated with the spoken utterance “turn it on” and the gesture of pointing to the radio.** Based on these events, the dialog manager does a search of the existing applications, transactions or “dialogs,” or portions thereof [stored on the context stack], with which such an utterance [including the accompanying lip movement] and gesture could be associated.

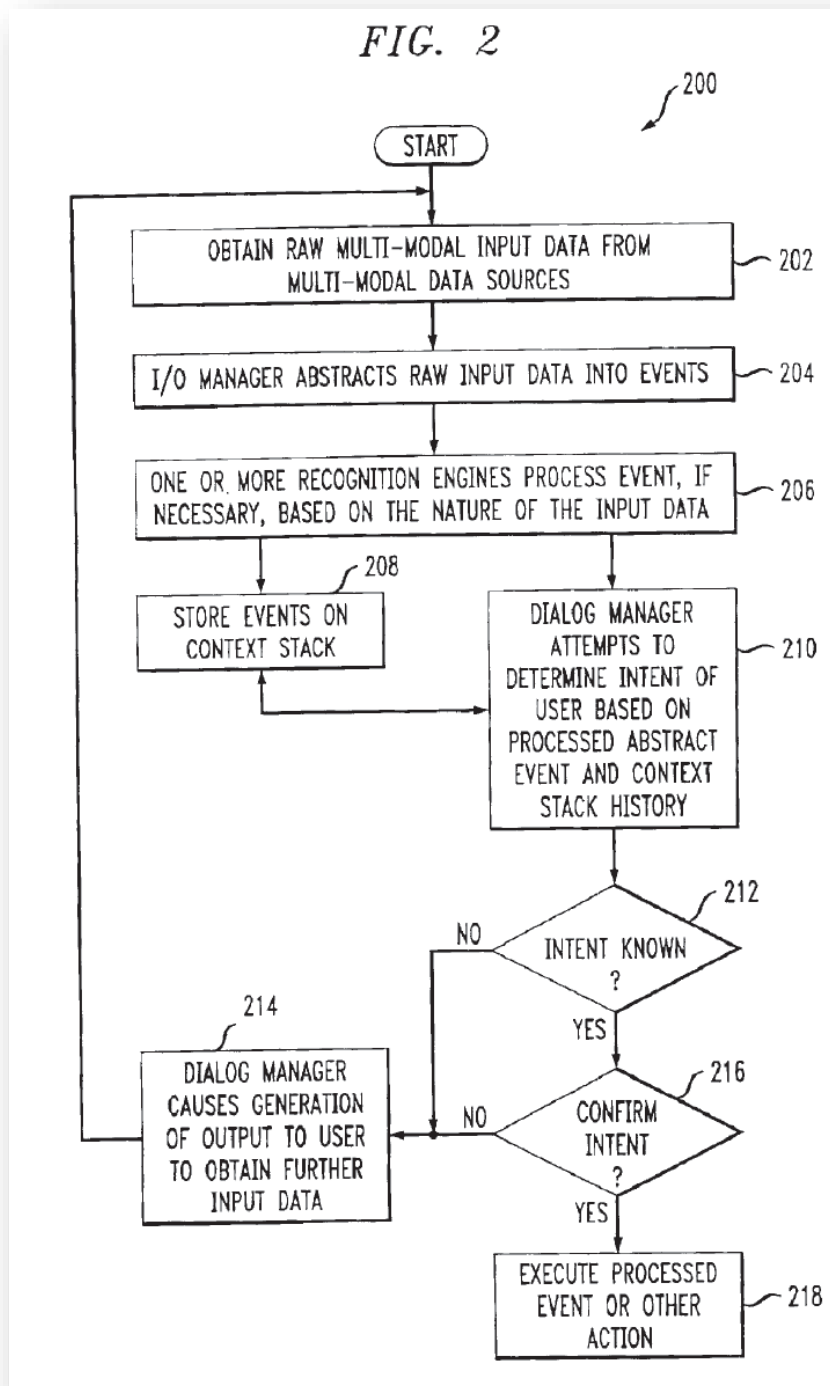
Ex.1005, 7:65–8:4, 8:37–42. In this example, the aligned decoded script and visemes sequence created from merging the transcriptions includes the results of the recognized events associated with the spoken utterance “turn it on” with the accompanying lip movement, therefore, constitutes “*a merged transcription*.”

194. Indeed, with regards to module 430 in Figure 4, Maes identifies “**the first joint use of the visual information and audio information** in module 430” and “an explanation of how the two types of information [i.e, the audio information included in the decoded text/script and video information included in the viseme sequence] are combined to provide improved recognition accuracy.” Ex.1005, 20:5–7, 11:55–63. In view of these teachings, a POSITA would have understood that module 430 provides the “*merging ... to create a merged transcription*” step such that the outcome of the “*merging ...*” step produces the aligned decoded script and visemes sequence (i.e., “*a merged transcription*”) associated with the spoken utterance, e.g., “turn it on” with the accompanying lip movement.



Ex.1005, Figure 4 (annotated).

195. Maes provides detailed explanation of how the audio signal and video signal are synchronized to generate “a merged transcription.” For example, Figure 2, below, is described in Maes as depicting “a flow diagram illustrating a referential ambiguity resolution methodology performed by a multi-modal conversational computing system.” Ex.1005, 3:24–28.



Maes, Figure 2. Referencing Figure 2, Maes explains that “in step 202, raw multi-modal input data is obtained from multi-modal data sources associated with the

system [T]he data input portion of the subsystem may comprise one or more cameras or sensors for capturing video input data ... one or more microphones for capturing audio input data from the environment in which the System is deployed.” Ex.1005, 5:57–6:8. Thus, in the car example in Maes, “cameras and microphones may be strategically places throughout the vehicle in order to attempt to fully capture all visual activity and audio activity.” Ex.1005, 6:15–22. The data is then abstracted into a form that represents one or more events, such as “a spoken utterance.” Ex.1005, 6:28–35.

196. In step 206, “the abstracted data or event is then sent by the I/O manager 14 to one or more recognition engines 16 in order to have the event recognized, if necessary. That is, depending on the nature of the event, one or more recognition engines may be used to recognize the event. For example, **if the event is some form of spoken utterance wherein the microphone picks up the audible portion of the utterance and a camera picks up the visual portion (e.g., lip movement) of the utterance, the event may be sent to an audio-visual speech recognition engine to have the utterance recognized using both the audio input and the video input associated with the speech.**” Ex.1005, 6:38–50.

197. In Maes, one of the ways in which the **audio-visual speech recognition engines** processes or otherwise, recognizes event data is by generating an outcome in which the transcribed audio information such as the decoded text/script is aligned

or synchronized with the transcribed video information such as the visemes sequence, such outcome constituting “*the merged transcription.*” Ex.1005, 21:50–54 (“[T]he **script is aligned with the visemes.** A rapid (or other) alignment may be performed in a conventional manner in order to attempt to **synchronize the two information streams.**”), 21:59–61 (“a likelihood on the alignment is computed to determine how well **the script aligns to the visual data**”).

198. In addition, with respect to other modalities of input which may comprise the I/O subsystem (e.g., “GUI-based devices such as a keyboard, a mouse, etc., and/or other devices such as a stylus and digitizer pad for capturing electronic handwriting,”) (Ex.1005, 6:22–26), Maes discloses:

It is to be appreciated that two, more or even all of the input modes described herein may be synchronized via the techniques disclosed in U.S. patent application identified as Ser. No. 09/507,526 filed on Feb. 18, 2000 and entitled “Systems and Method for Synchronizing Multi-modal Interactions,” which claims priority to U.S. provisional patent application identified as U.S. Ser. No. 60/128,081 filed on Apr. 7, 1999 and U.S. provisional patent application identified by Ser. No. 60/158,777 filed on Oct. 12, 1999, the disclosures of which are incorporated by reference herein.

Ex.1005, 7:22–32.

199. Therefore, Maes teaches or suggests “*merging the speech-based transcription and the non-speech-based transcription to create a merged transcription.*”

200. Referring to the example in which Maes's multi-modal conversational computing system is employed within a vehicle, the “*merging ...*” step would result in the system aligning/synchronizing decoded text/script processed from the spoken utterance of saying “turn it on” with the visemes sequence processed from a video of the lip movement of the user.

201. Therefore, Maes teaches or suggests this claim element for at least this reason.

Additional Mapping

202. As explained in [19.1.0] and [19.1.1], in Maes, each of keyboard-typed inputs and GUI-based input constitute additional examples of “*a non-speech input*” and further, each of these examples serves as “*the non-speech-based transcription.*” Additionally, as explained in [19.3.0], in Maes's system, “decoded text” (or, script) processed from the spoken utterance constitutes “*the speech-based transcription.*”

203. Coffman teaches the “*merging ...*” step because Coffman discloses “merg[ing]” input streams. For example: “combine input keyed on the keyboard, handwritten input and speech input.” Ex.1012, 38:2–4. Coffman continues:

merg[ing] ... the output of multiple decoders[] [or] ASCII transcription. Each input stream is converted into its ASCII transcription and aligned with input time- marks via speech recognition processing When different input stream are associated to the same task, the transcripts

are merged as follows. First, commands and queries are sorted based on the time marks and appended to a single data stream. Command formulation can be checked against [Finite State Grammar] FSG rules and re-sorted to satisfy the grammar rules.

Ex.1012, 38:8–14. Therefore, in Coffman, the combined keyboard-typed input and the transcribed speech input taken together serve as “*a merged transcription*.” Coffman provides further details of merging input streams, *see* Ex.1012, 38:8–19.

204. Dialect’s Complaint, consistent with Coffman’s teachings identifies: “Erica transcribed a non-speech communication (the words ‘I would like to’ which were input via the keyboard interface) and the speech communications (the words ‘Transfer money’ which were input as speech via the microphone interface), creating corresponding textual messages.”) Ex.1007, ¶120.

205. Therefore, Maes (which incorporates Coffman) renders this claim element obvious for this additional reason.

j. [19.5] identifying one or more entries in a context stack matching information contained in the merged transcription;

206. Maes teaches or suggests this claim element.

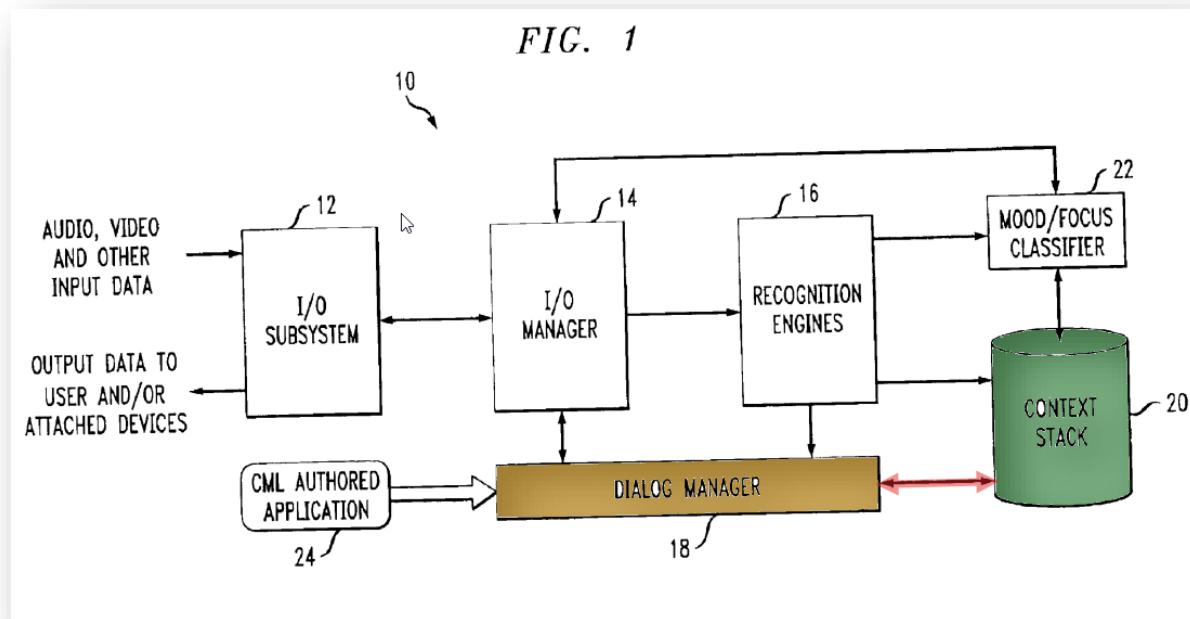
207. As explained in [19.4], in Maes, the aligned decoded text/script and visemes sequence constitutes “*the merged transcription*.”

208. Maes describes its context stack storing “historical information” e.g., past input/output (I/O) events or “*one or more entries*” generated previously in the context of a dialog.

[H]istorical information (e.g., past events) stored in the context stack [such that] the context stack ... is associated with the organized/sorted context corresponding to each active dialog.

Ex.1005, 7:62–63, 37:60–61; Figure 1 (reproduced below); *see also* Ex.1005, 7:40–45 (disclosing “the system 10 attempts to determine the **user intent based on the current event** and the **historical interaction information stored in the context stack** and then determine and execute one or more application programs that **effectuate the user’s intention and/or react to the user activity.**”); 5:16–19 (disclosing “an I/O (input/output) event generated previously and stored in a **context manager/history stack** (e.g., if a beeper rang and the user asked ‘turn it off’”).

209. Maes describes that “the context stack 817 may be implemented ...” as part of the context stack 20 of Figure 1. Ex.1005, 37:53–55; *see also* 3:66–4:6 (“the ... system ... comprises ... a context stack 20).



Ex.1005, Fig. 1 (annotated).

210. Specifically, Maes provides an example of a vehicular system in which a user may say “turn it on,” while pointing or gesturing at the vehicle radio. Maes describes its system identifying words in a recognized utterance from contexts—such as by matching the utterance “turn it on” to the context “radio”:

The dialog manager would ... receive the results of the recognized events associated with the spoken utterance “turn it on” and the gesture of pointing to the radio. **Based on these events, the dialog manager does a search of the existing applications, transactions or “dialogs,” or portions thereof [stored on the context stack], with which such an utterance and gesture could be associated.**

Ex.1005, 7:65–8:4.

[T]his recognized spoken utterance event is stored on the context stack. Then, when the recognized gesture event (e.g., pointing to the radio) is received, the dialog manager takes this event and **the previous spoken utterance event stored on the context stack** and makes a determination that the user intended to have the radio turned on.

Ex.1005, 8:37–42.

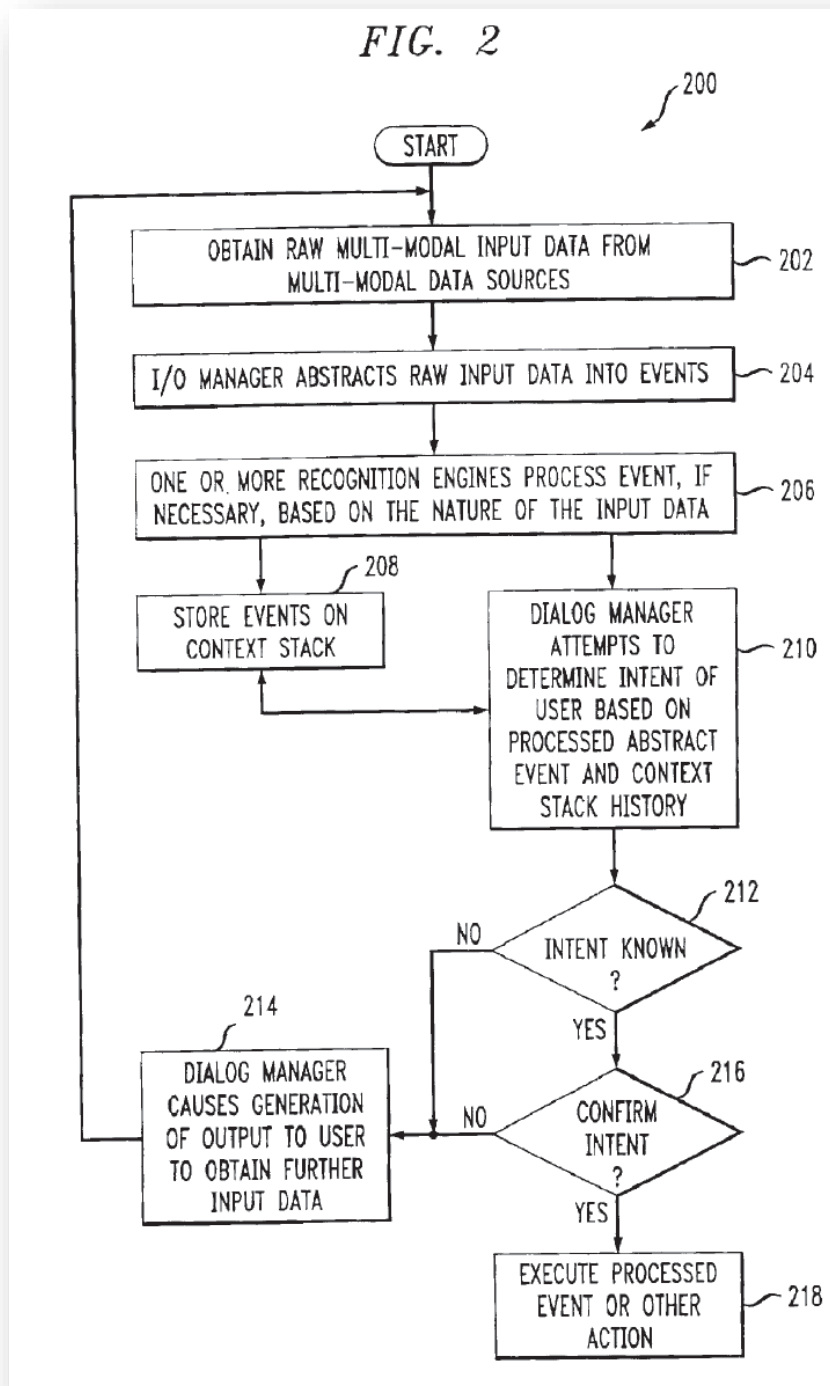
211. In this example, the results of the recognized events constitutes “*information*,” which is included in the aligned decoded script and visemes sequence associated with the spoken utterance “turn it on” with the accompanying lip movement (“*in the merged transcription*”). Further, in the above example, “this event” corresponds to the current event of the user pointing to the radio, which occurred few seconds after “the previous spoken utterance event” of the user saying “turn it on. Data for both events [“*one or more entries*”] are stored on the context stack (“*in a context stack*”) because Maes discloses “**all the variables, states, input, output and queries to the backend that are performed in the context of the dialog and any extraneous event that occurs during the dialog**” are stored on the context stack. Ex.1005, 37:55–62. Maes clarifies that “the dialog” includes “**conversational dialog comprising speech** and other multi-modal I/O such as GUI keyboard, pointer, mouse, **as well as video input**, etc.” Ex.1005, 36:49–54. Therefore, Maes discloses the dialog manager associating the results of the recognized events (“*information*”) included in the aligned decoded script and

visemes sequence produced as a result of merging the transcriptions (“*contained in the merged transcription*”) with an organized/sorted context in a context stack such as context stack 817 in Figure 8 or context stack 20 in Figure 1 corresponding to an active dialog which serves as “*identifying one or more entries in a context stack matching information contained in the merged transcription*” and consequently, the associated context(s) stored in the context stack corresponding to the user’s intent in the active dialog constitutes the “*one or more entries in a context stack*” that “*match[]*” (“*matching*”) the results of the recognized events (“*information*”) included in the aligned decoded script and visemes sequence produced as result of merging the transcriptions (“*contained in the merged transcription*”) associated with the spoken utterance, e.g., “turn it on” with the accompanying lip movement.

212. As explained above, Maes teaches identifying matches between associated context(s) stored in the context stack corresponding to the user’s intent in the active dialog (or, simply contexts stored on the context stack) and results of recognized events (or, simply transcribed user inputs). A POSITA would have understood that in speech recognition systems, it may be the case that—the associated context(s) stored in the context stack corresponding to the user’s intent in the active dialog may not have a full or 100% (i.e., “complete”) match with the results of the recognized events (included in the decoded text/script produced as a result of the transcription) but instead have a partial match—and Maes provides for

such a possibility describing it as “could be associated.” *See generally* Ex.1005, 7:60–8:4 and 37:53–55. Therefore, a POSITA would have understood that the outcome of the “*identifying ...*” step would generate one or more candidate contexts selected from the context stack that potentially (e.g., completely or partially) match the results of the recognized events (e.g., included in the decoded text/script produced as a result of the transcription).

213. Further, Figure 2, below, described in Maes as depicting “a flow diagram illustrating a referential ambiguity resolution methodology performed by a multi-modal conversational computing system” expressly discloses the use of context stack for storing event-related data. Ex.1005, 3:24–28.



Maes, Figure 2. Referencing Figure 2, Maes explains that “[i]n step 208, the recognized events ... are stored in a storage unit referred to as the context stack 20.

The **context stack** is used to create a history of interaction between the user and the system ... **to assist the dialog manager 18 in making referential ambiguity resolution determinations** when determining the user’s intent. Next, in step 210, the system 10 attempts to **determine the user intent based on the current event and the historical interaction information stored in the context stack** and then determine and execute one or more application programs that effectuate the user’s intention and/or react to the user activity.” Ex.1005, 7:33–45.

214. Therefore, Maes teaches or suggests this claim element.

215. To the extent that Patent Owner contends that Maes does not teach or suggest the “*identifying ...*,” step, Coffman does. First, Maes incorporates Coffman by reference. Ex.1005, 9:30–37; *see* § IX.B.

216. Coffman explains that context stack entries are searched to identify a context matching transcribed user input. “**Completion/search for the active context is performed from context to context down the stack. That is, new queries or arguments are compared by the dialog engine by going down the stack until an acceptable match is obtained ...**.” Ex.1012, 42:1–3. Coffman further explains that searching for the context includes comparing transcribed user inputs against information stored in the context stack:

Based on the modality (pointer, keyboard, file, speech),
the task dispatcher 402 redirects the stream to the

appropriate conversational subsystems or conventional subsystem with **speech inputs being transcribed/understood. The output of these subsystems is run down the context stack 405 to extract the active query and complete it.**

Ex.1012, 41:2–5. Thus, in Coffman, the outputs of conversational subsystems or conventional subsystems which constitute transcribed user input, are compared against data stored in the context stack until a match is found.

Completion/search for the active context is performed from context to context down the stack. That is, new queries or arguments are compared by the dialog engine by going down the stack until an acceptable match is obtained

Ex.1012, 42:1–3. Thus, Coffman's teachings of searching the context stack until an acceptable match is found between the transcribed user inputs and data stored on the context stack serve as the “*identifying ...*” functionality. Indeed, Coffman teaches updating the context stack in a manner such that “[w]hen a message is returned to the user, the context is updated and then pushed to the top of the context stack ... under the active context. The active context is updated to inform of the existence of a returned value.” Ex.1012, 42:7–9. This additionally confirms that Coffman provides the “*identifying ...*” functionality. Accordingly, the results of the recognized events (“*information*”) included in the aligned decoded script and

visemes sequence produced as a result of merging the transcriptions (“*contained in the merged transcription*”) produced as a result of processing the user’s spoken utterance, as Maes teaches, would be compared against data stored on the context stack, per Coffman’s teachings, to “*identify[] one or more contexts that completely or partially match one or more text combinations contained in the transcription.*” Ex.1012, 41:1–5, 40:15–19; Ex.1005, 7:60–8:4, 8:37–42, and 37:53–55; *see also* § IX.B.

217. Therefore, Maes teaches or suggests this claim element.

k. [19.6] determining a most likely context for the multi-modal input based on the identified entries;

218. As explained in [19.5], Maes teaches “*identifying ... entries in a context stack matching information contained in the merged transcription,*” and as such, in Maes, the associated context(s) stored in the context stack corresponding to the user’s intent in the active dialog constitutes “*the identified entries,*” which equivalently are one or more candidate contexts selected from the context stack that potentially (e.g., completely or partially) match the results of the recognized events (e.g., included in the aligned decoded script and visemes sequence generated from merging the speech and non-speech-based transcriptions). And, further, as explained below, Maes teaches or suggests this claim element.

219. Maes teaches scoring one or more candidate contexts such as phonemes, which Maes describes as “sub-phonetic or acoustic units of speech”

(Ex.1005, 12:29–33) (or, equivalently as “portions [of dialogs]” (Ex.1005, 7:60–8:4)). Specifically, Maes discloses generating, for each considered phoneme in a given context, likelihood scores (indicating the likelihood that it was that particular phoneme that was spoken) based, for example, on the audio information and/or the video information received as input to the system. Referencing Figure 4, Maes discloses:

[T]he probability module 416 in the audio information path ... labels the acoustic feature vectors with one or more [one or more previously stored] phonemes [which ... are sub-phonetic or acoustic units of speech]. [Each phoneme associated with one or more feature vectors has a **probability associated therewith indicating the likelihood that it was that particular acoustic unit that was spoken.**]

Ex.1005, 18:46–48 and 12:29–37. Specifically, Maes teaches calculating likelihood scores for acoustic feature vectors and visual speech feature vectors extracted from the multi-modal input:

[T]he probability module yields **likelihood scores for each considered phoneme in the form of the probability** that, given a particular phoneme or acoustic unit (au), the acoustic unit represents the uttered speech **characterized by one or more acoustic feature vectors A** or, in other words, **$P(A|\text{acoustic unit})$** .

[T]he probability module 426 labels the extracted visual speech vectors with one or more previously stored

phonemes. Again, each phoneme associated with one or more visual speech feature vectors has a probability associated therewith indicating the likelihood that it was that particular acoustic unit that was spoken in the video segment being considered. Thus, the probability module yields **likelihood scores for each considered phoneme in the form of the probability** that, given a particular phoneme or acoustic unit (au), the acoustic unit represents the uttered speech **characterized by one or more visual speech feature vectors V** or, in other words, **P(V|acoustic unit)**.

Ex.1005, 12:38–43 and 18:46–60. Thus, Maes's system calculates likelihood scores (e.g., probabilities P(A|acoustic unit) and P(V|acoustic unit)).

220. Maes then describes using the likelihood scores for “identify[ing]” (“*determining*”) “**the acoustic units** [i.e., “*a most likely context*” identified from amongst the entries in the context stack as explained in [19.5]] ... as having the highest probabilities of representing what was uttered” “characterized by acoustic feature vectors A and visual speech feature vectors V extracted from the multi-modal input (“*for the multi-modal input*”). Thus, the candidate context (such as a phoneme or an acoustic unit) that has the highest probability of representing what was uttered corresponds to “*a most likely context for the utterance.*” See Ex.1005, 18:64–65 and 20:13–15. The highest-probability-candidate-context (“*a most likely context*”) is identified from (“*determin[ed]* ... *based on*”) one or more candidate contexts, which, as explained in [19.5], are equivalently the associated context(s) stored in the

context stack corresponding to the user’s intent in the active dialog (“*the identified entries*”).

221. Therefore, Maes discloses this claim element.

I. [19.7] identifying a domain agent associated with the most likely context for the multi-modal input;

222. Maes discloses this element.

223. Initially, the 468 Patent describes a “*domain agent*” as used for organizing “generic and domain specific behavior and information The domain specific agents may provide complete, convenient and re-distributable packages or modules for each application area.” Ex.1001, 27:14–20. Thus, “*domain agent[s]*” in the 468 Patent broadly refer to software modules that are specific to each application area.

224. In Maes, the acoustic units having the highest probabilities of representing what was uttered constitutes “*the most likely context for the multi-modal input based on the identified entries*” and which is explained in connection with elements [19.5] and [19.6]. Furthermore, for the reasons below, Maes discloses “*identifying ... a domain agent associated with the most likely context for the multi-modal input.*” To determine the user’s intent based on the multi-modal input, Maes’s system:

determine[s] and execute[s] one or more application programs that effectuate the user’s intention and/or

**react to the user activity. The application depends on
the environment that the system is deployed in.**

Ex.1005, 7:40–46. Therefore, in Maes application programs (effectuating the user’s intention and depending on the environment that the system is deployed) are analogous to the 468 Patent’s description of “domain agent,” and therefore “application programs” serve as “domain agent[s].” Maes explains this further in greater detail.

225. Maes describes that “the conversational dialog manager 819 determines what information the user has, which inputs the user presents, and which application(s) should handle the user inputs. ... The DMA function [of the dialog manager] handles multi-modal I/O events to: (1) determine the target application or dialog (or portion of it) and (2) use past history and context” and further, in Maes, “the historical information (e.g., past events) [are] stored in the context stack.” See Ex.1005, 36:49–37:3, 37:51–38:2; *see also* Ex.1005, 7:60–8:19. In addition, “[t]he context stack [] comprises all the information associated with an application.” Ex.1005, 37:51-56.

226. By teaching that Maes’s dialog manager determines which application(s) should handle the user inputs, and further that the DMA function of the dialog manager determines the target application or dialog (or portion of it), Maes expressly discloses the step of “*identifying ... a domain agent.*” Further, as

explained in [19.6], because the context stack includes all the information associated with an application, the context stack includes the acoustic units having the highest probabilities of representing what was uttered (“*the most likely context for the multi-modal input*”). Therefore, Maes discloses “determining” (“*identifying*”) “the target application or dialog (or portion of it)” (“*a domain agent*”) “*associated with*” the acoustic units having the highest probabilities of representing what was uttered (“*the most likely context for the multi-modal input*”).

227. Moreover, for the additional reasons below, Maes (which incorporates Coffman) discloses this claim element. As explained above, Maes teaches determining the target application or dialog (or portion of it) (“*domain agent*”) and furthermore, as explained in connection with [19.7], Coffman teaches determining an acceptable match (“*most likely context*”) found between the transcribed user inputs and data stored on the context stack. Accordingly, a POSITA would have recognized that the target application in Maes is “*associated with*” the context corresponding to the acceptable match, as taught by Coffman, to “*identify[] a domain agent associated with the most likely context for the multi-modal input,*” such as the user’s multi-modal input received by Maes’s system. Indeed, Coffman teaches updating the context stack such that the active context is “pushed to the top of the context stack” (*see generally* Ex.1012, 42:7–9) and accordingly, a POSITA

would have been motivated to combine the teachings of Maes and Coffman (*see* IX.B).

228. Therefore, Maes teaches or suggests this element.

m. [19.8] communicating a request to the identified domain agent; and

229. Maes discloses this claim element.

230. As explained in [19.7], “application programs” or “applications” in Maes serve as “*domain agent[s]*” and Maes’s “target application” constitutes “*the identified domain agent*” in this claim element. Ex.1005, 36:64–66.

231. And, referencing the in-vehicle example, Maes’s system “selects the appropriate device (e.g., radio) activation routine [“*the identified domain agent*”] and instructs the I/O manager to output a command [“*communicat[e] a request*”] to activate the radio.” The target application such as a radio activation routine constitutes “*the identified domain agent*” and the command output/sent by Maes’s system to the radio activation routine (“*the identified domain agent*”) constitutes “*a request*.” Ex.1005, 8:16–19.

232. Further, Maes explains that “[i]f the dialog manager determines with some predetermined degree of confidence that the application it selects is the one which will effectuate the users desire, the dialog manager executes the next step of the multi-modal dialog[,] ... [it] launches the execution of an action associated to a fully understood multi-modal request from the user) of that application based on the

multi-modal input” which additionally serve as “*communicating a request to the identified domain agent.*” Ex.1005, 8:8–16.

233. Therefore, Maes discloses this claim element.

n. [19.9] generating a response to the user from content provided by the identified domain agent as a result of processing the request.

234. Maes teaches or suggests this claim element.

235. As explained in [19.8], Maes discloses “*a request*” “*communicated*” to *the identified domain agent.*”

236. And, referencing the in-vehicle example, Maes’s system “selects the appropriate device (e.g., radio) activation routine and instructs the I/O manager to output a command to activate the radio.” Ex.1005, 8:16–19. Thus, in Maes, the application such as a “radio activation routine” (“*the identified domain agent*”) would be activated, and a POSITA would have understood that upon processing the command (“*as a result of processing the request*”), the system would “adapt some behavior/service/appliances accordingly” to “*generat[e] a response to the user*” (e.g., cause the radio to be turned on and play, for instance, entertainment, news, sports, or weather content) (“*content provided by the identified domain agent*”). Ex.1005, 7:60–8:29 and 2:54–64.

237. Maes teaches an application having an “associated business logic ... with which the user interacts” and further that “one or more application programs ... effectuate the user's intention and/or react to the user activity.” Ex.1005, 2:65–67 and 7:43–45. The reaction of the business logic of an application (e.g. upon processing a command sent to it), therefore, constitutes serves as “*generating a response to the user from content provided by the identified domain agent as a result of processing the request.*” Ex.1005, 2:65–67.

238. Therefore, Maes teaches or suggests this claim element.

2. Dependent Claim 25

- a. **[25.0] The method of claim 19, further comprising verifying an identity of the user based on voiceprint matching, password matching, or pass-phrase matching.**

239. Maes teaches or suggests this claim.

240. As explained in [19.2], Maes’s system provides “*identifying the user*” For example, Maes discloses “speaker[/user] verification using the extracted acoustic feature vectors” which serves as describing “*verifying an identity of the user based on voiceprint matching.*” Ex.1005, 23:46–48. And, according to Maes, “speaker identification and verification may be accomplished via any conventional acoustic information speaker recognition system,” e.g., described in the U.S. patent application identified by Ser. No. 08/788,471. *See generally* Ex.1005, 23:49–58. Using voiceprint to determine the identity of users, or otherwise authenticate users

was well-known in the art. For example, Talmor, reflecting a POSITA’s knowledge, describes a “voiceprint” as “features are specific to each individual and as such can be used to characterize that individual” and further that a voiceprint “is typically generated by extracting voice features from a word or phrase uttered by an individual.” Ex.1030, 7:4–7. Talmor also describes a voice authentication method and a system which “compare a voice print of a user with each of a plurality of stored voice prints of known individuals” to determine “if the user voice print is most similar to a stored voice print of an individual the user claims to be among all other stored voice prints.” Ex.1030, 1:6–13. A POSITA would have understood that “acoustic feature vectors” in Maes provide the “*voiceprint*” of a user that is used in a matching/verification process. Ex.1005, 23:49–58.

241. Therefore, Maes teaches or suggests this claim.

3. Dependent Claim 26

- a. **[26.0] The method of claim 19, further comprising receiving one or more follow-up multi-modal inputs at the conversational voice user interface, the follow-up multi-modal inputs including at least one of a follow-up natural language utterance or a follow-up non-speech input.**

242. Maes teaches or suggests this claim.

243. As explained in [19.1.1], in Maes, the (I/O) subsystem 12 comprising one or more microphones for capturing audio input data” constitutes “*the conversational voice user interface*” of this claim. And, referencing Figure 2, Maes

explains: “in step 214,” the system “generat[es] ... an output to the user **requesting further input data** [*“one or more follow-up multi-modal inputs”*] so that the user’s intent can be disambiguated,” “the system ... obtains the raw input data, again in step 202, and the process ... iterates based on the new data. Such iteration can continue as long as necessary ... to determine the user’s intent.” Ex.1005, 8:47–50 and 8:58–65. The system may “seek confirmation in step 216 from the user in the same manner as the request for more information (step 214).” Ex.1005, 8:66–9:1. Maes’s system, which “receiv[es] a multi-modal natural language input ... including a natural language utterance and a non-speech input provided by a user” (see analysis in [19.1.0]), therefore, applies the same receiving technique and processes the further input data (*“the one or more follow-up multi-modal inputs”*) in a similar manner as the originally-received/raw input data until the user’s intent is understood, iteratively.

244. With reference to the in-vehicle example in which the system does not have enough original input to determine the user intent, referencing Figure 2, Maes discloses:

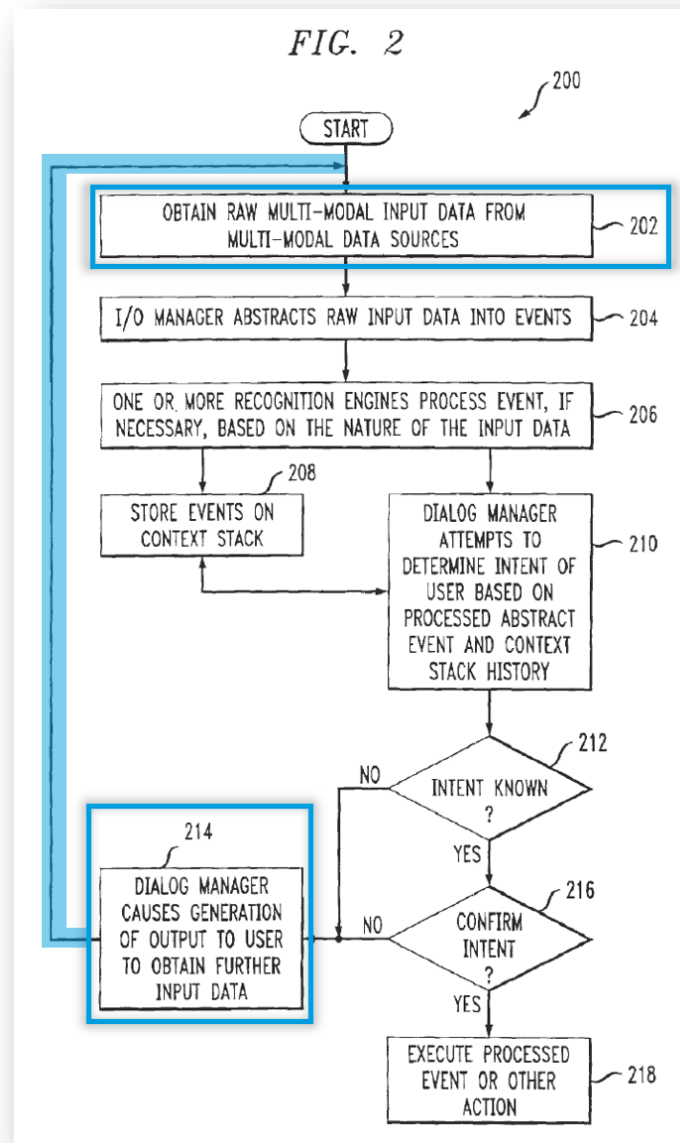
Consider the case where the user says “turn it on,” but makes no gesture and provides no other utterance. In this case, assume that the dialog manager does not have enough input to determine the user intent (step 212 in FIG. 2) and thus implement the command. The dialog manager,

in step 214, then causes the generation of an output to the user **requesting further input data** so that the user's intent can be disambiguated. This may be accomplished by the dialog manager instructing the I/O manager to have the I/O subsystem output a request for clarification. ... The user, hearing the query, could then point to the radio or say **“the radio”** thereby providing the dialog manager with the additional input data to disambiguate his request. That is, with reference to FIG. 2, the system 10 obtains the raw input data, again in step 202, and the process 200 iterates based on the new data. Such iteration can continue as long as necessary for the dialog manager to determine the user's intent.

Ex.1005, 8:43–65. In the above example, the user's feedback saying the spoken utterance “the radio” (including the accompanying lip movement) constitutes the *“one or more follow-up multi-modal inputs ... including ... a follow-up natural language utterance.”* The iterative nature of the process in Figure 2 confirms that Maes's system handles processing of the further/follow-up/additional input data (*see below*) in a manner similar to the originally-received input data.

245. Maes's system “handles multi-modal I/O events to ... **follow[ing] up with a dialog** to disambiguate, complete, correct or confirm the understanding” if “confirmation, disambiguation, correction, more details” are needed until the intent of the user is unambiguous and fully determined. See generally Ex.1005, 36:59–

37:2. For example, the follow-up event can be some form of spoken utterance wherein the microphone picks up the audible portion of the utterance and a camera picks up the visual portion (e.g., lip movement) of the utterance. See Ex.1005, 6:43–50.



Ex. 1005, Figure 2 (annotated) (depicting “further input data” received in block 214).

Ex.1003

246. Therefore, Maes teaches or suggests this claim.

4. Dependent Claim 27

a. [27.0] The method of claim 26, wherein the identified domain agent updates the context stack and the semantic knowledge-based model in response to processing the request.

i. Domain Agent Updating the Context Stack

247. Maes teaches or suggests this claim element.

248. Maes explains use of the context stack “compris[ing] **all the information** associated with an application ... includ[ing] all the variable, states, input, output, and **queries to the backend** ... in the context of the dialog and any extraneous event that occurs during the dialog” “to create a history of interaction between the user and the system ...” Ex.1005, 7:33–45 and 37:55–60. Creating a history of interaction in the context stack with **all** the information related to a dialog would require accumulating/adding data or otherwise updating the context stack over time, and therefore, Maes teaches or suggests “*updat[ing] the context stack.*”

249. Moreover, circumstances in which users provide multiple requests/commands to the system, this would result in multiple queries to the backend. Because a context stack stores **all** the information (Ex.1005, 37:55–60), including **all** the queries to the backend, the context stack, therefore, would necessarily need to be updated with each of the multiple queries. This example further confirms that Maes’s context stack is updated.

250. Furthermore, as explained in [26.0], Maes’s Figure 2 process, working iteratively, discloses shows receiving follow-up multi-modal inputs (block 214), processing them, and storing the processed results as events on the context stack (blocks 208 and 210) until the user’s intent is determined. Ex.1005, 7:33–39. For example, storing the results related to the follow-up/further multi-modal inputs (in addition to the originally-received input) on the context stack, therefore, encompasses “*updat[ing] the context stack.*” See 37:55–60 (describing all the input stored on the context stack).

251. Maes teaches or suggests that “*the context stack*” is “*update[d]*” by “*the identified domain agent.*”

252. Maes teaches “allow[ing] [multiple] access[es] to backend business logic” and further that information describing “**such accesses to be added to the context stack.**” Ex.1005, 38:3–7. Adding information to the context stack discloses “*updat[ing] the context stack.*” Further, because Maes teaches “an application” as having an “associated business logic [*e.g.*, backend business logic in Maes’s system] ... with which the user interacts” (Ex.1005, 2:65–67 and 38:3–7), it would have been obvious to a POSITA that in Maes, the target application (corresponding to the user’s intent) updates its context stack.

253. Additionally, Maes describes “[t]he context stack ... compris[ing] **all the information associated with an application,**” which Maes describes having an

“associated ... logic [*e.g.*, business logic] ... with which the user interacts,” and further that “applications(s) ... handle the user inputs.” Ex.1005, 2:65–67, 36:56–57, 37:55–56, and 38:3–7. Accordingly, a POSITA would have understood that “*the context stack*” comprising **all** the information (including all user inputs) would be updated by the respective/corresponding application (“*the identified domain agent*”) which handles the user inputs or otherwise interacts with the user. Ex.1005, 2:65–67, 37:56, and 38:3–7.

254. To the extent PO asserts Maes does not expressly disclose that “*the identified domain agent updates the context stack*,” it would have been obvious to a POSITA to modify Maes’s teachings such that applications in Maes update the context stack. A POSITA would have understood that such a modification would be an obvious design choice and provide an efficient updating method because applications have first-hand/direct knowledge of user inputs based on interacting with users. And further, because the modification is based on Maes’s own teachings, there would have been a reasonable expectation of success in making this modification. Additionally, by updating the context stack, the most-recent context (*e.g.*, highly pertinent in determining the user’s intent by Maes’s system) would be made available.

255. Maes teaches or suggests that the “*updat[ing]*” is “*in response to processing the request.*”

256. Furthermore, referencing the in-vehicle system example (Ex.1005, 8:16–19), a POSITA would have understood the radio activation routine updates the context stack of the in-vehicle system upon processing the command (“*in response to processing the request*”) to activate the radio. See [19.8] (explaining “*the request*” communicated to the target application (“*the identified domain agent*”)) and [19.9] (explaining “*as a result of processing the request*” equivalently recited in this claim as “*in response to processing the request*”).

257. Therefore, Maes teaches or suggests “*wherein the identified domain agent updates the context stack ... in response to processing the request.*”

ii. Domain Agent Updating the semantic knowledge-based model

258. Maes teaches or suggests “*wherein the identified domain agent updates ... the semantic knowledge-based model in response to processing the request.*”

259. As explained in connection with element [19.3.0], Maes discloses a semantic module (comprising “*a semantic knowledge-based model*”) and Maes teaches that models are subjected to **training** (Ex.1005, 19:7–21). A POSITA would have understood that Maes’s teachings of training or, otherwise updating a model, such as the semantic module, therefore, serve as describing “*updat[ing]*” the semantic module (comprising “*the semantic knowledge-based model*”).

260. Maes provides a suggestion of an application (“*the identified domain agent*”) updating the semantic module because Maes teaches “application programs”

function to “effectuate the user’s intention and/or react to the user activity ... depend[ing] on the environment” in which Maes’s system is deployed.” Ex.1005, 7:43–46. Maes also teaches application programs having a number of “input parameters or variables.” Ex.1005, 8:20–21. For example, referencing the in-vehicle system, upon processing a command [*“in response to processing the request”*] to activate the radio, the radio activation routine (*“the identified domain agent”*) would update the “semantic module” with the variables “turn” “it” “on” so that in future interactions with the user, the system determines that the context of the user’s utterance is turning on a device in the vehicle, such as the radio. Ex.1005, 7:60–8:22. As another example, Maes states “a beeper rang and the user asked ‘turn it off,’ the term ‘it’ can be disambiguated.” Ex.1005, 5:18–19. A POSITA would have understood that in order to provide the above-described function, the application program associated with the beeper would update its semantic module with the variables “turn” “it” “off” so that in future interactions with the user, the system determines that the context of the user’s utterance is with respect to turning off the beeper. Additionally, as explained in [19.8] and [19.9], Maes teaches or suggests the claim terms *“the request”* and *“as a result of processing the request”*.

261. Maes teaches that its system provides “real-time feedback to the business logic unit [associated with an application (*e.g.*, *“the identified domain agent”*) with which the user interacts],” from which a POSITA would have

understood that the feedback is used to update a model (e.g., comprised in a semantic module) for future interactions with the user and further that the application (e.g., “*the identified domain agent*”) would be performing such updates. For instance, if a user speaks a keyword/phrase that is not included in the semantic module, then a POSITA would have understood that an application updating the semantic module with that particular keyword/phrase would expand the capabilities of the semantic module in additionally interpreting that keyword/phrase. Ex.1005, 38:3–7, 44:22–25, 41:60–64, and 2:65–67; *see also* Ex.1005, 36:49–37:3

262. Therefore, Maes teaches or suggests “*wherein the identified domain agent updates ... the semantic knowledge-based model in response to processing the request.*”

263. To the extent one asserts that Maes does not expressly or implicitly disclose the step of “*the identified domain agent updat[ing] ... the semantic knowledge-based model,*” Dr. Houh states that such a step was well within the POSITA’s knowledge, and, based on such knowledge it would have been obvious to modify Maes to implement such a step in Maes’s system. For example, at the time of the claimed invention, IBM’s ViaVoice speech recognition software, which is an example of “*the identified domain agent*” included functionality for updating/supplementing its base vocabulary (e.g., “*the semantic knowledge-based model*”) by adding more words.

See Ex.1031 (<https://www.speechtechmag.com/Articles/Editorial/Features/IBM-Announces-ViaVoice-98-29469.aspx>)

Three Products

The ViaVoice 98 consists of three products: the executive, office and home editions.

ViaVoice 98 Executive Edition is IBM's most powerful continuous speech software and can be a productivity tool for writers, executives and experienced PC users. ViaVoice Executive offers direct dictation into most popular Windows applications, voice control for the desktop and applications and includes two topics, **ViaVoice Computer and ViaVoice Business and Finance, which supplement the basic vocabulary.** Topics are special interest vocabularies that add to the existing base vocabulary and increase accuracy. The executive edition is priced at \$149.

ViaVoice 98 Office Edition is a continuous speech dictation and voice control software tool designed to help maximize the efficiency of a small office/home office user, faculty member or advanced student. **Users can supplement their vocabulary with the included ViaVoice Business and Finance.** The price for the Office edition is \$89.

ViaVoice 98 Home Edition is IBM's entry-level continuous speech dictation product and is ideal for the family to use to create letters, reports or e-mail at home. ViaVoice 98 Home includes the ViaVoice Cuisine Topic to let users produce original recipes or make a keepsake of family recipes. The cost is \$49.95.

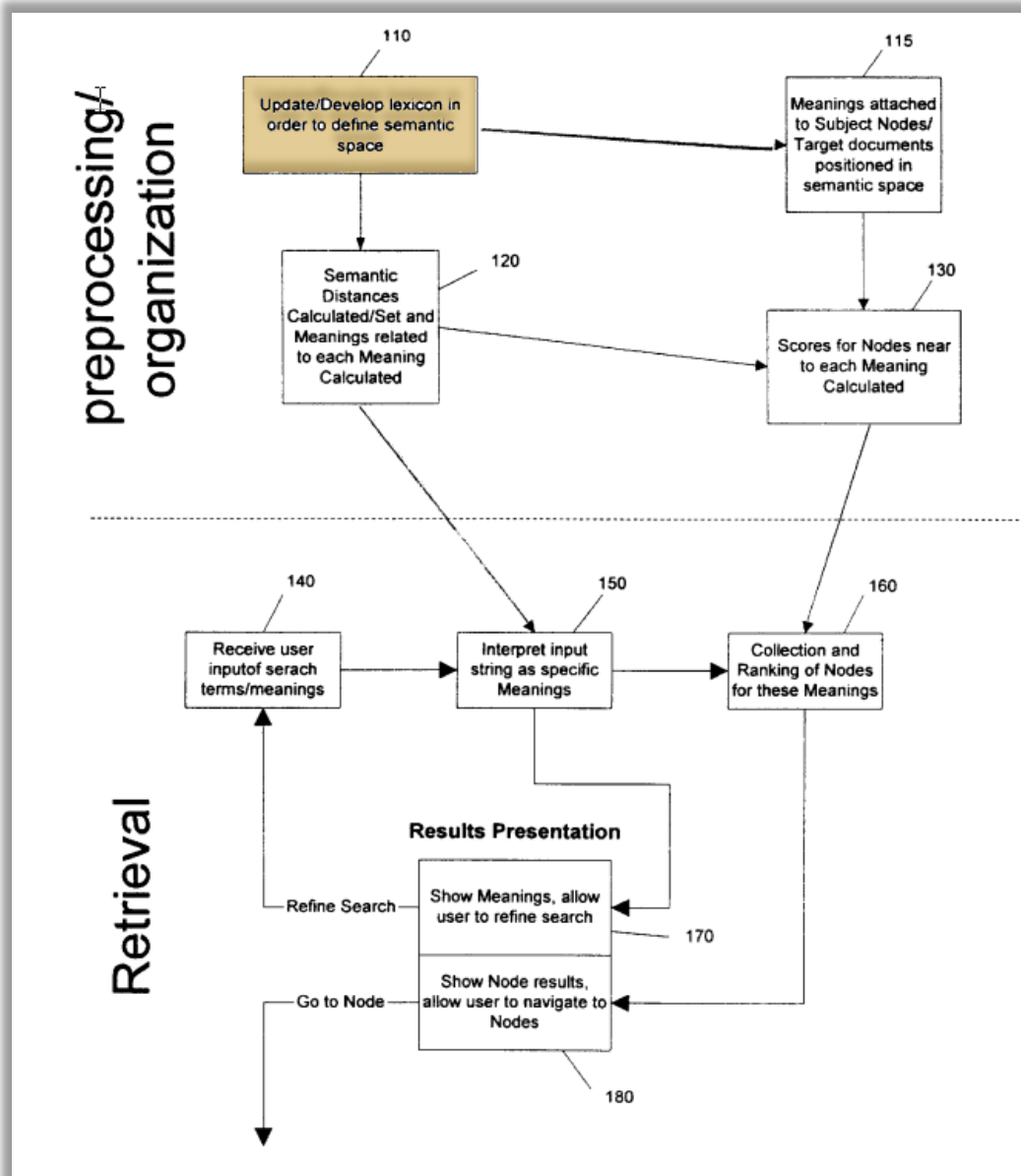
Ex. 1031, 1. See also <https://aapl.org/docs/newsletter/N232voice-recog.htm> A POSITA would have found it obvious to implement such a step in Maes's system because Maes is also an IBM patent (*see* Ex.1005, Cover), which would motivate a POSITA to incorporate teachings from IBM's ViaVoice speech recognition software into Maes.

264. As another example, Weissman reflecting the POSITA's knowledge, provides an example in which a search engine application (e.g., "*the identified domain agent*") allows updating a meaning-based lexicon (e.g., "*a semantic knowledge-based model*"). Weissman's teaches

develop[ing] a useful meaning-based lexicon ... [which]
is a network of interconnected meanings[] ... describ[ing]
the "semantic space" ... employed by the search engine.
...

[T]he lexicon can be updated [“*updat[ing]*”] and expanded to include new meanings or update connections for meanings already present.

Ex.1025, 3:10–23.



Ex.1025, Figure 1 (annotated). Weissman highlights the importance of updating its lexicon, stating that:

New meanings can be updated mainly to reflect the importance in everyday culture of a large number of proper nouns, which are not available in many lexicons such as the names of companies, people, towns, etc. In addition, **new meanings may have to be developed within the lexicon in order to cover subject areas of common interest with greater specificity**. For example, whereas an existing lexicon might define the idea of a “programming language”, it may be useful to **expand this concept within the lexicon** to designate the many hundreds of specific programming languages that may exist as “kind of” children to this meaning.

Ex.1025, 3:23–34.

265. As yet another example, Wang teaches the MiPad speech-enabled application provides functionality to “enrich the [semantic language model] SLM when suitable amount of training data become available” and because a “SLM can be updated in real-time during the course of interaction, irrelevant expressions, including the spontaneous speech, can be gracefully rejected based on what makes sense to the dialog context.” Ex.1014, 1. *See also*, Ex.1016, [0048] (explaining the need to update semantic models: “feedback dialogue data ... generated during active dialogue sessions” “provide a useful statistical basis ... to learn new models or to adapt existing semantic models” and that the “feedback dialogue data ... may ... be collected with respect to either individuals or a general population”); Ex.1015,

[0146] (describing that a semantic model “may be adapted ... for accurate identification of a present context or a user context”).

266. A POSITA would have been motivated to make this modification based on the well-known knowledge in the art related to the improved capabilities of models with training/refining/adding more words to an application’s vocabulary, as reflected in the above-mentioned references. For example, it would have been obvious to a POSITA to modify Maes’s “semantic module” (comprising a “*semantic knowledge-based model*”) which functions to “indicate that some decoding candidates in a list are meaningless and should be discarded/replaced with meaningful candidates,” in a manner such that it is updated by a corresponding/associated application (“*the identified domain agent*”). Ex.1005, 41:58–64. Such a modification would expand the ability of Maes’s semantic module to detect relevant keywords in a user’s input (such as a spoken utterance) to a given speech-enabled application, thereby improving the quality of the transcribed speech output, which is consistent with Maes’s goals of “improved speech recognition.” Ex.1005, 11:42–45. For example, the updated semantic module would have an expanded vocabulary (i.e., more number of keywords, variables, parameters, sentences, words, phrases), than a non-updated semantic module, and consequently provide improved keyword spotting capability, resulting in speech-based transcriptions that are “closer in meaning” to the user inputs.

267. Therefore, Maes teaches or suggests this claim.

5. Dependent Claim 28

a. [28.0] The method of claim 27, wherein the knowledge-enhanced speech recognition engine determines a most likely context for the follow-up multi-modal input using the updated context stack.

i. The Knowledge-enhanced Speech Recognition Engine

268. As explained in [19.3.0], Maes discloses a semantic module (comprising “*a knowledge-based semantic model*”) and speech recognition functionality/components of Maes (“*a speech recognition engine*”). Maes’s semantic module “**interpret[s] meaning of phrases**” such that it is “**used by speech recognizer 926 to indicate that some decoding candidates in a list are meaningless and should be discarded/replaced with meaningful candidates.**” Ex.1005, 41:35–38 and 41:59–64. Maes discloses: “When an abstract [input/output (I/O)] event occurs,” the system “determines the target of the event and ... seeks confirmation, disambiguation, correction, more details, ... until the intent is unambiguous and fully determined.” Ex.1005, 36:59–63. For example, Maes’s system provides “prompt or display for missing, ambiguous or confusing information, asks for confirmation or launches the execution of an action associated to a fully understood multi-modal request from the user) of that application based on the multi-modal input.” Ex.1005, 8:12–16. Therefore, Maes’s system is capable of handling multi-modal I/O events to “(a) understand the intent of the user; (b) follow

Ex.1003

up with a dialog to disambiguate, complete, correct or confirm the understanding [of the intent of the user]; (c) or, dispatch a task resulting from fall [sic:full] understanding of the intent of the user.” Ex.1005, 36:64–37:3; *see also* 5:15–19 (“disambiguating an I/O (input/output) event generated previously”). Ambiguity resolution or disambiguating the user’s intent, which is a goal of Maes (*see* Ex.1005, Abstract, 4:10–11, and 36:59–63) is therefore handled by the semantic module because the semantic module functions to indicate that some decoding candidates in a list are meaningless and should be discarded/replaced with meaningful candidates. Ex.1005, 41:61–64.

269. Maes’s afore-mentioned teachings related to ambiguity resolution are consistent with the 468 Patent’s description that “*knowledge-enhanced speech recognition system*” “may be used to ... determine the intent of the request and/or to correct false recognitions.” Ex.1001, 20:13–22. Thus, in the context of the 468 Patent, the semantic module makes Maes’s “*speech recognition engine*” “*knowledge-enhanced*.”

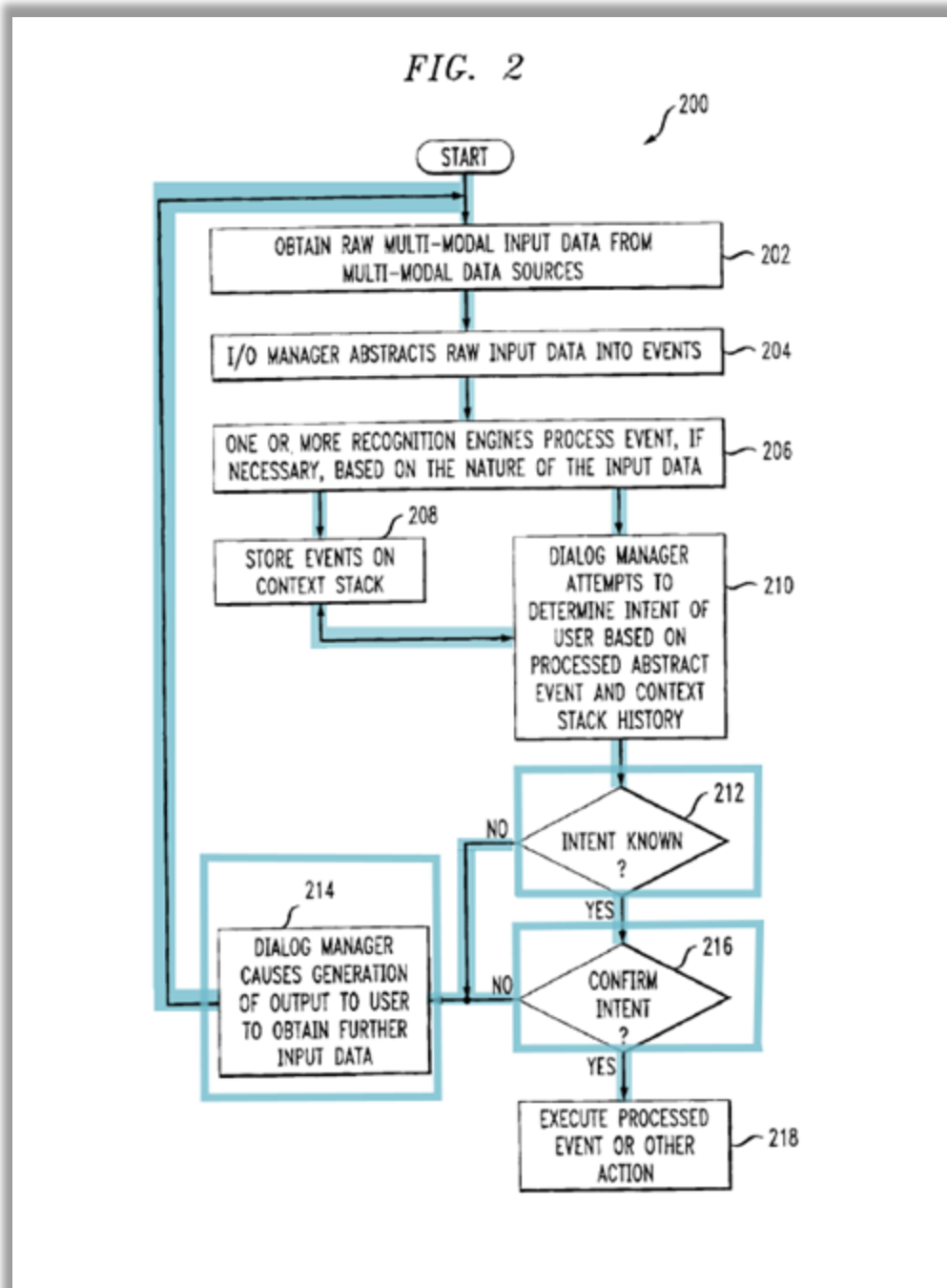
270. Therefore, taken together, Maes’s “*speech recognition engine*” and the semantic module, both comprised in Maes’s system, jointly provide functionality that disclose a “*knowledge-enhanced speech recognition engine*.”

ii. Determining a most likely context for the follow-up multi-modal input

271. Maes teaches or suggests that “*the knowledge-enhanced speech recognition engine determines a most likely context for the follow-up multi-modal input.*”

272. As explained in [19.5] and [19.6], Maes teaches or suggests “*determining a most likely context*” using “*entries in a context stack.*” Further, as explained in [26.0] earlier, Maes discloses receiving “further input data” (“*follow-up multi-modal inputs*”) in connection with step 214 of the iterative Figure 2 process. Maes, which teaches or suggests “*determin[ing] a most likely context for the multi-modal input*” (see analysis in [19.5] and [19.6]), therefore, applies the same technique and processes the further input data (“*the follow-up multi-modal input*”) in the same manner as the originally-received/raw input data until the user’s intent is understood.

273. Referencing the in-vehicle example in which the system does not have enough original input to determine the user intent, the user’s utterance of “the radio” constitutes “*the follow-up multi-modal input*” and “*a most likely context*” corresponds to the user’s intent to have the vehicle’s radio turned on. Ex.1005, 8:43–65.



Ex.1005, Figure 2 (annotated) (depicting the process iterating on the “further input data” received in step 214).

274. To the extent one argues that Maes does not expressly disclose that “*the knowledge-enhanced speech recognition engine determines a most likely context*,” it would have been obvious to a POSITA to modify Maes’s “*speech recognition engine*” with the techniques and features explained in [19.5], [19.6], [19.7], [26.0], and the iterative Figure 2 process such that Maes’s “*speech recognition engine*” “*determines a most likely context for the follow-up multi-modal input*.”

iii. Using the updated Context Stack

275. Maes teaches or suggests “*the knowledge-enhanced speech recognition engine ... using the updated context stack*” in the “*determin[ing] ...*” step.

276. As explained in [27.0] and [26.0], Maes teaches or suggests “*updat[ing]*” “*the context stack*” with I/O events related to “further input data” (“*the follow-up multi-modal input*”) received in connection with step 214 of the iterative Figure 2 process. Maes, which teaches or suggests the “*determin[ing] ...*” step explained above, therefore, uses the updated context stack for storing I/O events (step 208) during the iterative processing until the user’s intent is understood. Maes explains that “[i]n step 208, the recognized events ... are stored in a storage unit referred to as the context stack 20. The **context stack** is used to create a history of interaction between the user and the system ... **to assist the dialog manager 18 in making referential ambiguity resolution determinations** when determining the user’s intent. Next, in step 210, the system 10 attempts to **determine the user intent**

based on the current event and the historical interaction information stored in the context” Ex.1005, 7:33–45. Because the context stack stores a history of user interaction on an ongoing basis, Maes teaches (step 210) using the updated current stack. *See* also analysis in [27.0] (explaining reasons which would have motivated a POSITA to use an updated context stack).

277. To the extent one argues that Maes does not expressly disclose “*the knowledge-enhanced speech recognition engine ... using the updated context stack,*” it would have been obvious to a POSITA to modify Maes’s “*knowledge-enhanced speech recognition engine*” with the techniques and features explained in [26.0], [27.0] and the iterative Figure 2 process such that Maes’s “*knowledge-enhanced speech recognition engine*” “*us[es] the updated context stack.*”

278. Therefore, Maes teaches or suggests this claim.

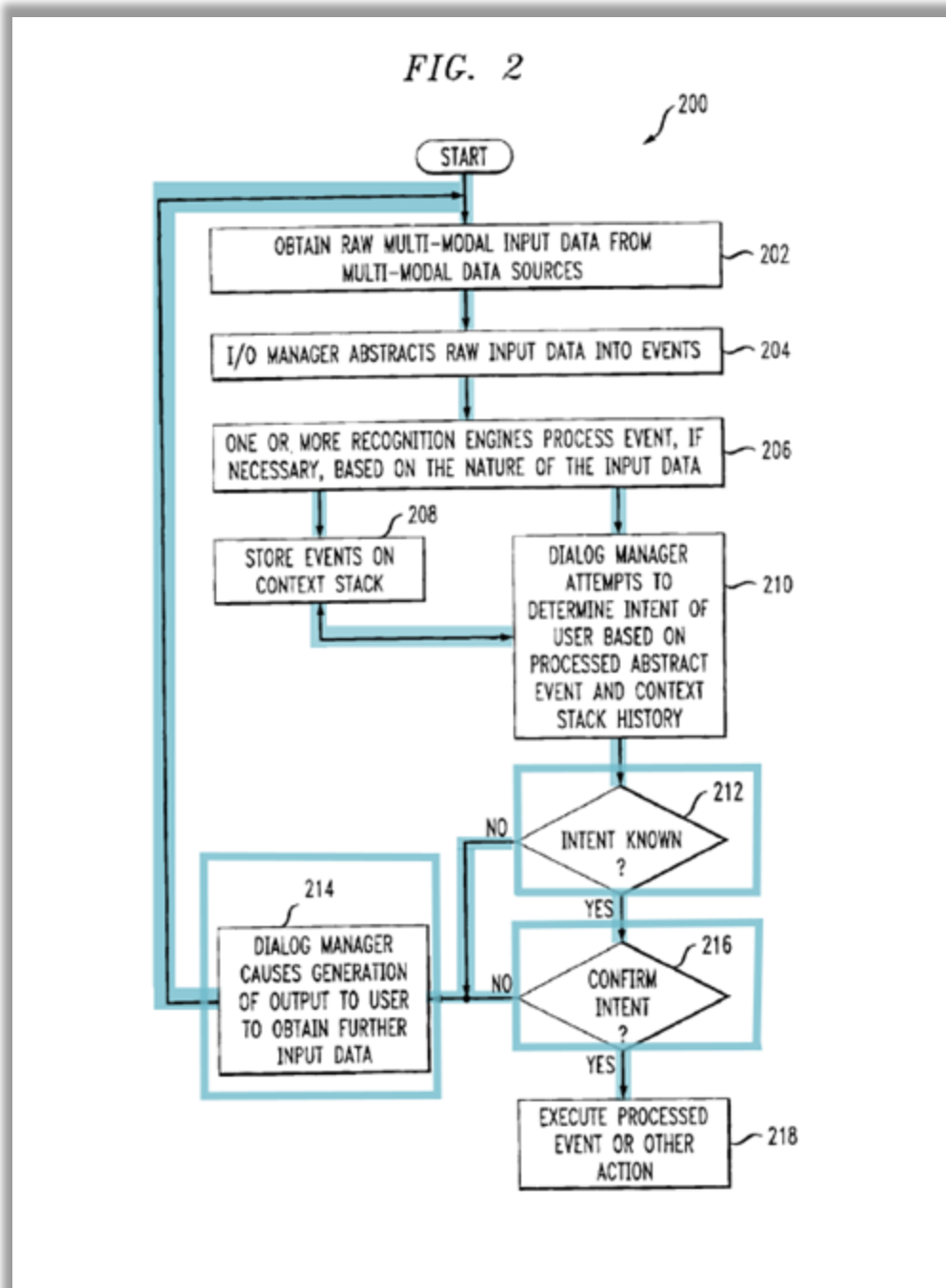
6. Dependent Claim 29

- a. [29.0] The method of claim 27, wherein the speech recognition engine creates a speech-based transcription of the follow-up natural language utterance using the updated semantic knowledge-based model.**

279. Maes teaches or suggests this claim.

280. As explained in connection with element [19.3.0], Maes discloses “*creating a speech-based transcription of the natural language utterance using a speech recognition engine and a semantic knowledge-based model.*” Furthermore,

as explained in connection with claims 26 and 27, Maes teaches or suggests receiving “further input data” (“*follow-up multi-modal inputs*” or “*follow-up ... utterance*”) in connection with step 214 of an iterative process (see Figure 2, below) and also teaches or suggests updating the semantic module (see claim 27). Maes’s process, which comprises “*creating a speech-based transcription of the natural language utterance*” (see analysis in [19.3.0]), involves applying the same technique and proceeds in the same manner as the originally-received/raw input data until the user’s intent is understood, iteratively. See generally Ex.1005, 8:66–9:6 (“The dialog manager ... may ... seek confirmation in step 216 from the user **in the same manner as the request for more information** (step 214)”).



Ex. 1005, Figure 2 (annotated).

281. Furthermore, as explained in connection with [27.0], Maes teaches or suggests “*updat[ing]*” “the semantic module” (“*the semantic knowledge-based*

model”) and it would have been obvious to a POSITA to use the updated semantic module (comprising the “*updated semantic knowledge-based model*”), because the updated semantic module would have an expanded vocabulary for “*creat[ing] a speech-based transcription of the follow-up natural language utterance.*” See [27.0].

282. Therefore, Maes teaches or suggests this claim.

7. Dependent Claim 30

- a. [30.0] The method of claim 19, wherein the identified domain agent processes the request by querying one or more local or network information sources.**

283. Maes teaches or suggests this claim.

284. As described in connection with element [19.7], “application programs” or “applications” in Maes serve as “*domain agent[s]*” and Maes’s “target application” constitutes “*the identified domain agent*” in this claim element.

285. Referencing the in-vehicle example, Maes teaches that the components of its system (e.g., implemented in software and/or hardware) including “memory” such as “RAM, ROM, a fixed memory device (e.g., hard drive), a removable memory device (e.g., diskette), a flash memory” which “may be located locally (in the vehicle)” or “remotely (e.g., connected wirelessly to the vehicle).” Ex.1005, 32:24–27; 45:30–34, Figure 10, *see generally* 45:50–60.

286. For example, a target application program, such as a “device (e.g., radio) activation routine (*“the identified domain agent”*) *“processes the request”* of turning on the radio by using “instructions or code” accessing (*“querying”*) stored in the system’s memory (*“one or more local or network information sources.”*). Ex.1005, 8:15–19, 45:45–46.

287. Therefore, Maes teaches or suggests this claim.

8. Dependent Claim 32

- a. [32.0] The method of claim 19, wherein the identified domain agent processes the request by directing a command to one or more local or remote devices.**

288. Maes teaches or suggests this claim.

289. Other than replacing *“querying”* and “network information sources” in claim 30 with *“directing a command”* and *“remote devices”* in claim 32, this claim is substantively the same as claim 30.

290. Further, referencing the in-vehicle example, Maes teaches that the components of its system (e.g., implemented in software and/or hardware) “may be located locally (in the vehicle)” or “remotely (e.g., connected wirelessly to the vehicle).” Ex.1005, 32:24–27; *see also* 45:50–60. For example, a target application program, such as a “device (e.g., radio) activation routine (*“the identified domain agent”*) *“processes the request”* of turning on the radio by *“directing a command”*

(e.g., “activate the radio”) “to” the radio (“*one ... local ... device[]*”). Ex.1005, 8:15–19.

291. Therefore, Maes teaches or suggests this claim.

B. Ground 2: Maes, Coffman and Ittycheriah Render Obvious claims 19, 25-30, and 32

292. To the extent one asserts that Coffman and Ittycheriah are not expressly and/or particularly incorporated by reference into Maes as if they were effectively the same document, claims 19, 25-30, and 32 would have nonetheless been obvious in view of the combination of Maes, Coffman, and Ittycheriah for the reasons articulated above in Ground 1. Moreover, a POSITA would have been motivated to combine the references and had a reasonable expectation of success for making such a combination for the reasons articulated in § IX.B.

XI. OBJECTIVE INDICIA OF NON-OBVIOUSNESS

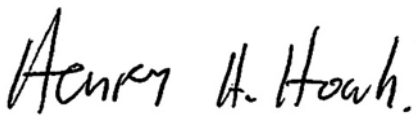
293. I am not aware of any objective indicia of non-obviousness that would change my opinion that the challenged claims of the 468 Patent are obvious in view of the prior art discussed above in this declaration. It is my opinion that any potential objective indicia of non-obviousness would not overcome the clear disclosures in the prior art showing that the subject matter of these claims would have been obvious to a POSITA as of August 29, 2005. However, I reserve the right to supplement my

opinions Patent Owner alleges or offers any purported evidence of objective indicia of non-obviousness.

XII. CONCLUSION

294. I declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and that these statements were made with knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under section 1001 of Title 18 of the United States Code.

Dated: March 18, 2025



CLAIM LISTING

Claim 19	
[19.0]	A method for processing multi-modal natural language inputs, comprising:
[19.1.0]	receiving a multi-modal natural language input at a conversational voice user interface, the multi-modal input including a natural language utterance and a non-speech input provided by a user, ...
[19.1.1]	... wherein a transcription module coupled to the conversational voice user interface transcribes the non-speech input to create a non-speech based transcription;
[19.2]	identifying the user that provided the multi-modal input;
[19.3.0]	creating a speech-based transcription of the natural language utterance using a speech recognition engine and a semantic knowledge-based model, ...
[19.3.1]	... wherein the semantic knowledge-based model includes a personalized cognitive model derived from one or more prior interactions between the identified user and the conversational voice user interface, ...
[19.3.2]	... wherein the semantic knowledge-based model includes ... a general cognitive model derived from one or more prior interactions between a plurality of users and the conversational voice user interface, ...
[19.3.3]	... wherein the semantic knowledge-based model includes ... an environmental model derived from an environment of the identified user and the conversational voice user interface;
[19.4]	merging the speech-based transcription and the non-speech-based transcription to create a merged transcription;
[19.5]	identifying one or more entries in a context stack matching information contained in the merged transcription;
[19.6]	determining a most likely context for the multi-modal input based on the identified entries;
[19.7]	identifying a domain agent associated with the most likely context for the multi-modal input;
[19.8]	communicating a request to the identified domain agent; and
[19.9]	generating a response to the user from content provided by the identified domain agent as a result of processing the request.
Claim 25	
[25.0]	The method of claim 19, further comprising verifying an identity of the user based on voiceprint matching, password matching, or pass-phrase matching.

Claim 26	
[26.0]	The method of claim 19, further comprising receiving one or more follow-up multi-modal inputs at the conversational voice user interface, the follow-up multi-modal inputs including at least one of a follow-up natural language utterance or a follow-up non-speech input.
Claim 27	
[27.0]	The method of claim 26, wherein the identified domain agent updates the context stack and the semantic knowledge-based model in response to processing the request.
Claim 28	
[28.0]	The method of claim 27, wherein the knowledge-enhanced speech recognition engine determines a most likely context for the follow-up multi-modal input using the updated context stack.
Claim 29	
[29.0]	The method of claim 27, wherein the speech recognition engine creates a speech-based transcription of the follow-up natural language utterance using the updated semantic knowledge-based model.
Claim 30	
[30.0]	The method of claim 19, wherein the identified domain agent processes the request by querying one or more local or network information sources.
Claim 32	
[32.0]	The method of claim 19, wherein the identified domain agent processes the request by directing a command to one or more local or remote devices.