

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

META PLATFORMS, INC.,
Petitioner

v.

DIALECT, LLC,
Patent Owner

Case No. IPR2025-01332
U.S. Patent No. 8,447,607

**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. INTRODUCTION	8
II. QUALIFICATIONS AND PROFESSIONAL EXPERIENCE	10
III. LEVEL OF ORDINARY SKILL IN THE ART	15
IV. RELEVANT LEGAL STANDARDS.....	17
V. OVERVIEW OF THE 607 PATENT	22
A. The Disclosures of the 607 Patent.....	22
VI. CLAIM CONSTRUCTION.....	25
VII. IDENTIFICATION OF HOW THE CLAIMS ARE UNPATENTABLE	25
VIII. THE PRINCIPAL PRIOR ART	26
A. Maes	26
IX. CLAIMS 12 and 13 ARE UNPATENTABLE	28
A. Ground 1: Maes renders obvious claims 12 and 13	28
1. The Combination of Maes’s Teachings with those of Ittycheriah and Coffman	29
2. Independent Claim 12	33
a. [12.0] A method for processing natural language inputs, comprising:	33
b. [12.1] receiving, by one or more processors, a multi- modal natural language input from a user, the multi- modal natural language input including a natural language utterance and a non-speech input;	39

- c. [12.2] generating, by the one or more processors, a non-speech transcription from the non-speech input; 50
- d. [12.3] identifying, by the one or more processors, the user who provided the multi-modal natural language input;.....58
- e. [12.4] generating, by the one or more processors, a speech-based transcription based on a cognitive model associated with the user, wherein the cognitive model includes information on one or more prior interactions between the user and the device; 64
- f. [12.5] generating, by the one or more processors, a merged transcription from the speech-based transcription and the non-speech transcription;.....79
- g. [12.6] identifying, by the one or more processors, an entry in a context stack, from among a plurality of entries that are in the context stack and that are each indicative of context, wherein the identified entry matches information in the merged transcription;.....91
- h. [12.7] identifying, by the one or more processors, a domain agent associated with the entry in the context stack; 102
- i. [12.8] determining, by the one or more processors, a request based on the merged transcription; and 105

j.	[12.9] communicating, by the one or more processors, the request to the domain agent, wherein the domain agent is configured to generate a response to the user.....	108
3.	Dependent Claim 13	110
a.	[13.0] The method of claim 12, wherein the cognitive model is a first cognitive model, and wherein the generating the speech-based transcription is based on a second cognitive model that includes information on one or more interactions between the device and a plurality of users of the device.	110
B.	Ground 2: Maes, Coffman and Ittycheriah Render Obvious claims 12 and 13.....	115
X.	OBJECTIVE INDICIA OF NON-OBVIOUSNESS.....	115
XI.	CONCLUSION	116
	CLAIM LISTING	117

EXHIBIT LIST

No.	Exhibit Description
1001	U.S. Patent No. 8,447,607
1002	File History of U.S. Patent No. 8,447,607
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	<i>Dialect LLC v. Bank of America, N.A.</i> , Case No. 2:24-cv-00207 (E.D. Tex.), Dkt. 66 (“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
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.
1011	RESERVED
1012	International Patent Application Publication No. WO 00/20962 (“Coffman”) filed as PCT/US99/22927
1013	U.S. Patent No. 7,137,126 (“ <u>Coffman126</u> ”)
1014	RESERVED
1015	RESERVED
1016	RESERVED

1017	Excerpts from Dafydd Gibbon et al., “Handbook of Multimodal and Spoken Dialogue Systems : Resources, Terminology, and Product Evaluation” (2000) (“Gibbon”)
1018	RESERVED
1019	RESERVED
1020	RESERVED
1021	RESERVED
1022	RESERVED
1023	RESERVED
1024	RESERVED
1025	RESERVED
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).
1029	EDTX Calendar, Judge Gilstrap

TABLE OF ABBREVIATIONS AND CONVENTIONS

Abbreviation	Meaning
607 Patent	Ex.1001: U.S. Patent No. 8,447,607
IPR	<i>inter partes</i> review
Petitioner	Meta Platforms, Inc. (“Meta”)
Patent Owner or PO	Dialect, LLC (“Dialect”)
Second Amended Complaint	Compl.: Ex.1007
xx:yy–zz	column xx, lines yy to zz

I, Henry Houh, declare as follows:

I. INTRODUCTION

1. I am making this declaration in support of Petitioner’s request for an *Inter Partes* Review of U.S. Patent No. 8,447,607 (“the 607 Patent”) to Weider.

2. 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.

3. I have been asked to provide my opinions regarding whether the subject matter of claims 12 and 13 (“the Challenged Claims”) of the 607 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.

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

- the 607 Patent, Ex.1001
- the prosecution history of the 607 Patent (“607 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.

5. 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 network communications and security as described below.

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

7. 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 607 Patent. If requested, I am prepared to testify about my opinions expressed herein.

II. QUALIFICATIONS AND PROFESSIONAL EXPERIENCE

8. 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.

9. 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.

10. 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.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

16. 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.

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.”

III. LEVEL OF ORDINARY SKILL IN THE ART

22. 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 607 Patent. I have been asked to assume the date of the purported invention for the 607 patent is August 29, 2005.

23. 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.

24. 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 607 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 607 Patent. Lack of professional experience can be remedied by additional education, and vice versa.

25. 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 607 Patent.

IV. RELEVANT LEGAL STANDARDS

26. I am not an attorney. In preparing and expressing my opinions and considering the subject matter of the 607 Patent, I am relying on certain basic legal principles that counsel has explained to me.

27. I understand that prior art to the 607 Patent includes patents and printed publications in the relevant art that predate the priority date of the 607 Patent. For purposes of this Declaration, I am applying August 29, 2005 as the priority date of the 607 Patent, but I take no position on that issue.

28. 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.

29. 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.

30. 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.

31. 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.

32. 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.

33. 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.

34. 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.

35. 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.

36. 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.

37. 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.

38. 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.

V. OVERVIEW OF THE 607 PATENT

39. The 607 Patent was based on U.S. Patent Appl. No. 13/488,299 filed on June 4, 2012 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.

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

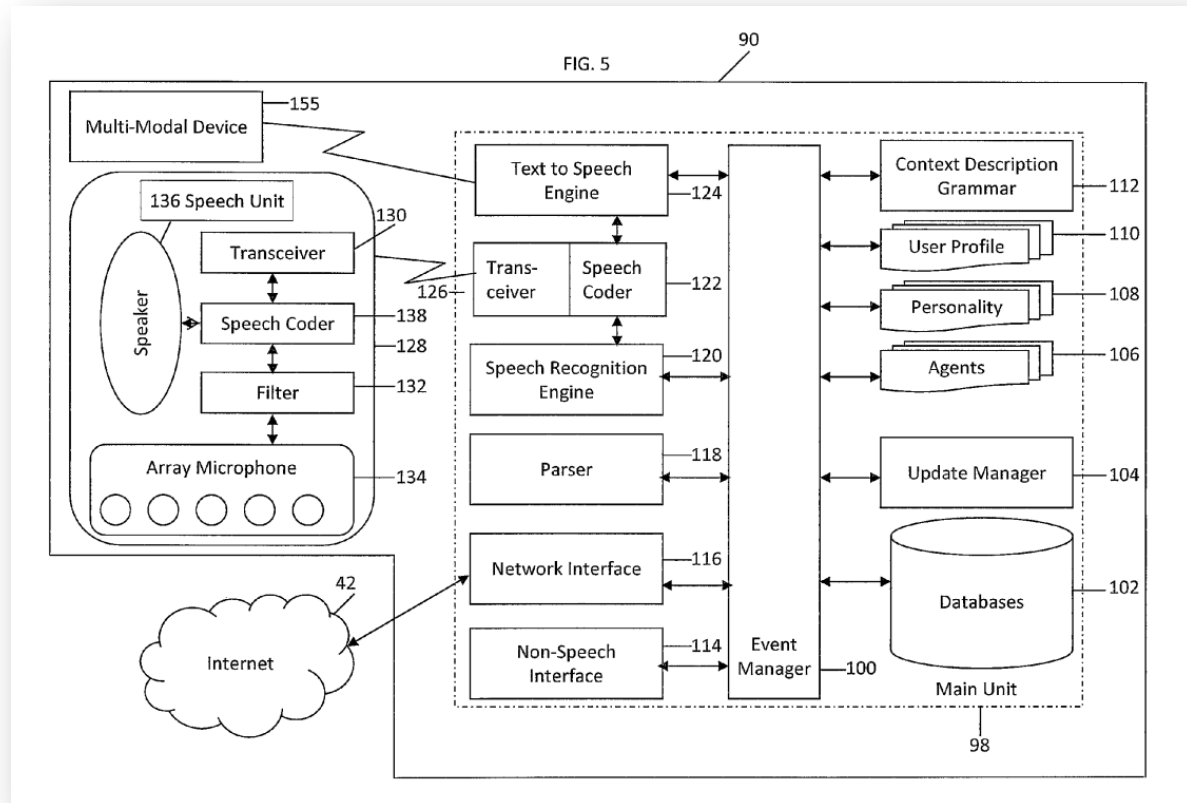
A. The Disclosures of the 607 Patent

41. The 607 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 that “[m]ore specifically, the invention enables mobile users to submit natural language speech and/or non- speech questions or commands in a wide range of domains.” Ex.1001, 1:20–25. The 607 Patent states that it “creates, stores and uses extensive personal profile information for each user, thereby improving the reliability of determining the context and presenting the expected results for a particular question or command. The invention may organize domain specific behavior and information into agents, that are distributable or updateable over a wide area network.” Ex.1001, Abstract.

42. The 607 Patent notes that “[s]peech recognition has steadily improved in accuracy and today is successfully used in a wide range of applications,” and that “[n]atural language processing has been applied to the parsing of speech queries.”

Ex.1001, 1:65–2:2. The 607 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, 2:6–8. The 607 Patent purportedly overcomes the “drawbacks” of prior art systems. Ex.1001, 2:57–58.

43. Figure 5 of the 607 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. As depicted in Figure 5, the disclosed system comprises speech unit 128, speech recognition engine 120, context description grammar module 112, parser 118, agents 106, multi-modal device 155 and non-speech interface 114.

Speech unit 128 includes a microphone to receive a spoken utterance from a user.

Ex.1001, 25:48–52.

44. After initial processing, the speech is passed to speech recognition engine 120 for processing using the context description grammar module 112. Ex.1001, 26:18–20. After that, “[a]ny recognized information may be processed by the parser 118, which transforms information into complete algorithms and questions

using data supplied by knowledge agents.” Ex.1001, 26:20–23. “The knowledge agents may then process the commands or questions” and “may return results” to the user. Ex.1001, 26:27–37. Moreover, “multimodal support may be provided to maintain the context during both voice interaction and interaction through the non-speech interface 114,” Ex.1005, 26:56–60, and “multi-modal devices 155 may be used as a substitute or complement to the speech interface.” Ex.1001, 26:51–53.

VI. CLAIM CONSTRUCTION

45. 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.

VII. IDENTIFICATION OF HOW THE CLAIMS ARE UNPATENTABLE

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

47. 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.

48. 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.

VIII. THE PRINCIPAL PRIOR ART

A. Maes

49. 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 607 Patent.

50. 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.

51. 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.

52. 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.

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

54. 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.

IX. CLAIMS 12 AND 13 ARE UNPATENTABLE

A. Ground 1: Maes renders obvious claims 12 and 13

55. Claims 12 and 13 of the 607 Patent are rendered obvious over Maes.

1. 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 system that receives and processes multi-modal inputs (e.g., by producing decoded text) 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.

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*”) 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 607 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 607 Patent because each are directed to functionality of natural language processing systems. Ex.1001, 1:65–2:2; 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 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. Thus, they are reasonably pertinent to the problem addressed in Maes and the 607 Patent as well.

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 system producing decoded text 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 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 transcriptions would add to the functionality of Maes's system and would have been a simple implementation and predictably resulted in a recognition system with improved accuracy, as Maes seeks to implement (Ex.1005, 11:62–63).

2. Independent Claim 12

a. [12.0] A method for processing natural language inputs, comprising:

65. Maes teaches or suggests “[a] *method for processing natural language inputs, comprising:*”

66. 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 added¹).

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.

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

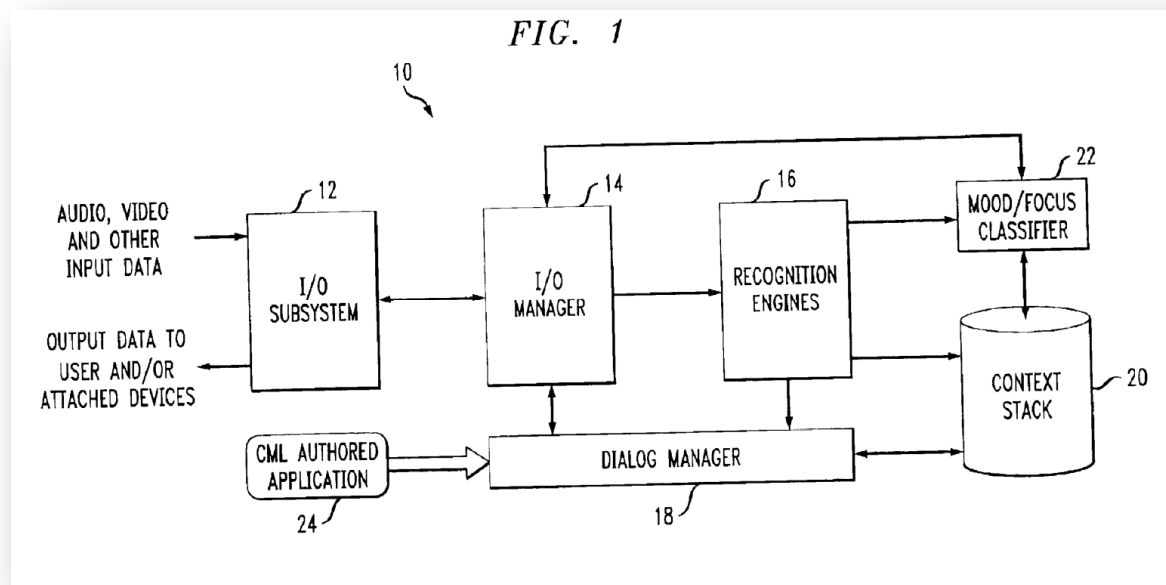
[A] multi-modal conversational computing system ...
compris[ing] an input/output (I/O) subsystem 12, an I/O manager module 14, one or more recognition engines 16, a dialog manager module 18, a context stack 20 and a mood/focus classifier 22.

Generally, the multi-modal conversational computing system 10 of the present invention receives multi-modal input in the form of audio input data, video input data, as well as other types of input data (in accordance with the I/O subsystem 12), processes the

¹ Emphasis added throughout unless otherwise specified.

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.

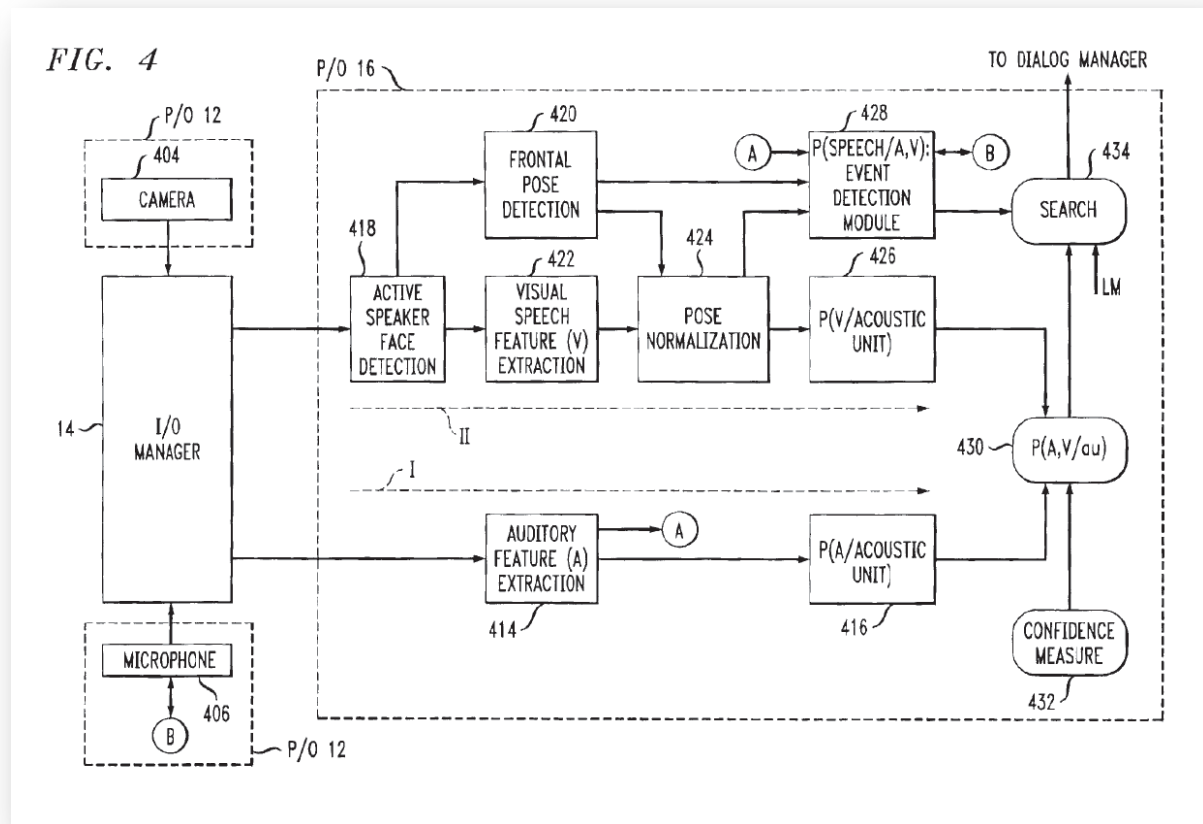


Ex.1005, Figure 1.

68. With reference to Figures 1 and 4, Maes describes Figure 4 (reproduced below) as “an audio-visual speech recognition module² that may be employed as one of the recognition modules³ 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 “*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 natural language inputs.*”

² As used herein, the terms “Maes’s system” and “audio-visual speech recognition module” are used interchangeably.

³ 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.

69. 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.

70. 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.

71. Maes describes examples in which its methods are applicable to a broadcast news system which uses spoken utterance⁴ inputs which further confirms

⁴ 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.

that its method applies to “*natural language inputs*.” For example, Maes states that “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. Maes’s system, which processes spoken utterance inputs, therefore, provides functionality for “*processing natural language inputs*. These examples comprise human communication and therefore Maes’s system and methods handle “*processing natural language inputs*.” See Gibbon defining “*natural language*” as “a language used by humans for communication.” Ex.1017, 444.

72. 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.

73. Therefore, Maes teaches or suggests “[a] *method for processing natural language inputs, comprising:*”

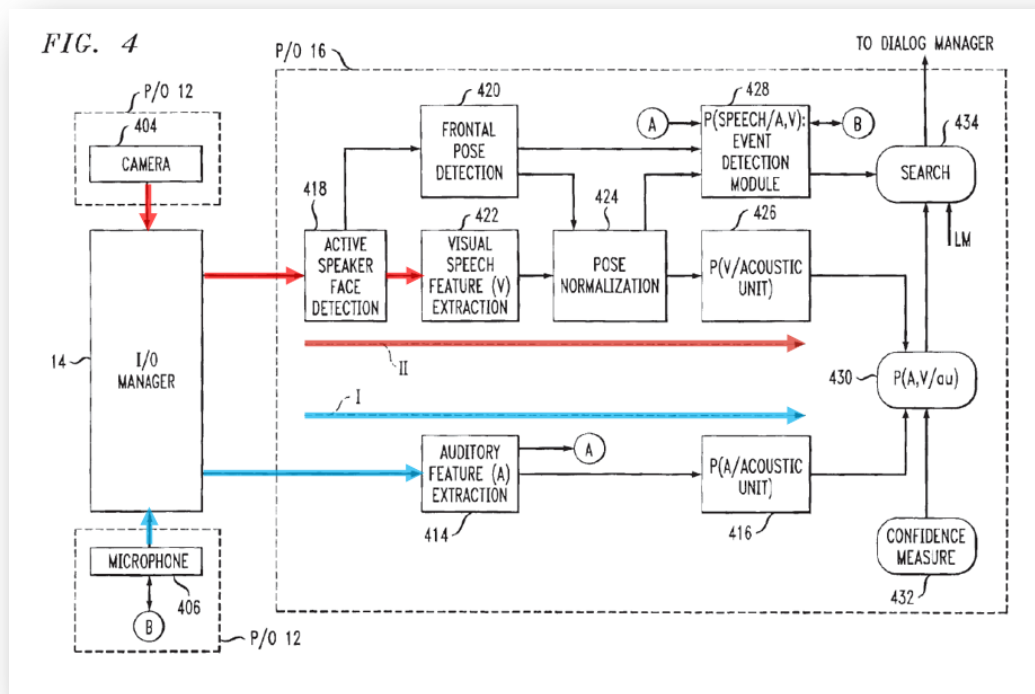
- b. **[12.1] receiving, by one or more processors, a multi-modal natural language input from a user, the multi-modal natural language input including a natural language utterance and a non-speech input;**

74. Maes teaches or suggests “*receiving ... a multi-modal natural language input from a user, the multi-modal natural language input including a natural language utterance and a non-speech input.*”

75. 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 from a user*”). Ex.1005, 10:47–11:23.

76. 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.

77. With reference to Figure 4⁵, 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 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 (reproduced below).



Ex.1005, Figure 4 (annotated).

⁵ 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.

78. 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.

79. Maes explains in detail what constitutes the “audio signal” (“*natural language utterance*”) and the “video signal” (“*non-speech input*”). For example, the “audio signal” (e.g., “*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.

80. 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 607 Patent’s description of its system coupled to “keypads 14 for receiving textual data input; ... a pointing device or stylus, or other user interface peripherals” 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.” Ex.1001, 16:18–19 and 19:1–4; *see also* Compl., ¶147 (keyboard-entered natural language input (“From my checking account”)).

81. Additionally, with reference to the Figure 1⁶, 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

⁶ 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.

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.

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

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.

83. Further, Maes discloses “*one or more processors*” “*receiving [the] multi-modal natural language input.*” Maes’s “**system ... comprises at least one processor, operatively coupled to the user interface subsystem**, and configured to receive at least a portion of the multi-modal input data from the user interface subsystem. The processor is further configured to then make a determination of at least one of an intent, a focus and a mood of at least one of the one or more users based on at least a portion of the received multi-modal input data. The processor is still further configured to then cause execution of one or more actions to occur in the environment based on at least one of the determined intent, the determined focus and the determined mood. The system further comprises a memory, operatively coupled to the at least one processor, which stores at least a portion of results associated with the intent, focus and mood determinations made by the processor for possible use in a subsequent determination or action.” Ex.1005, 2:38–54.

84. Maes explains that “[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, 45:50–60.

85. The 607 Patent describes an embodiment illustrated in Figure 1 in which the inventive system (mobile structure 10) is “used in vehicles” and comprises “speech interface and non-speech interface.” 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, 15:58–64, 16:19–20, 25:48–51 and Figure 1. This description of the 607 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 an example of “receiving [the] multi-modal natural language input ... [that] include[es] a natural language utterance and a non-speech

input” 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 “*receiving, by one or more processors, a multi-modal natural language input from a user, the multi-modal*

natural language input including a natural language utterance and a non-speech input.”

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. 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. Thus, in Maes, keyboard-typed input comprised in the multi-modal input data constitutes an additional example 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 are the same non-speech inputs described in the 607 Patent: “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

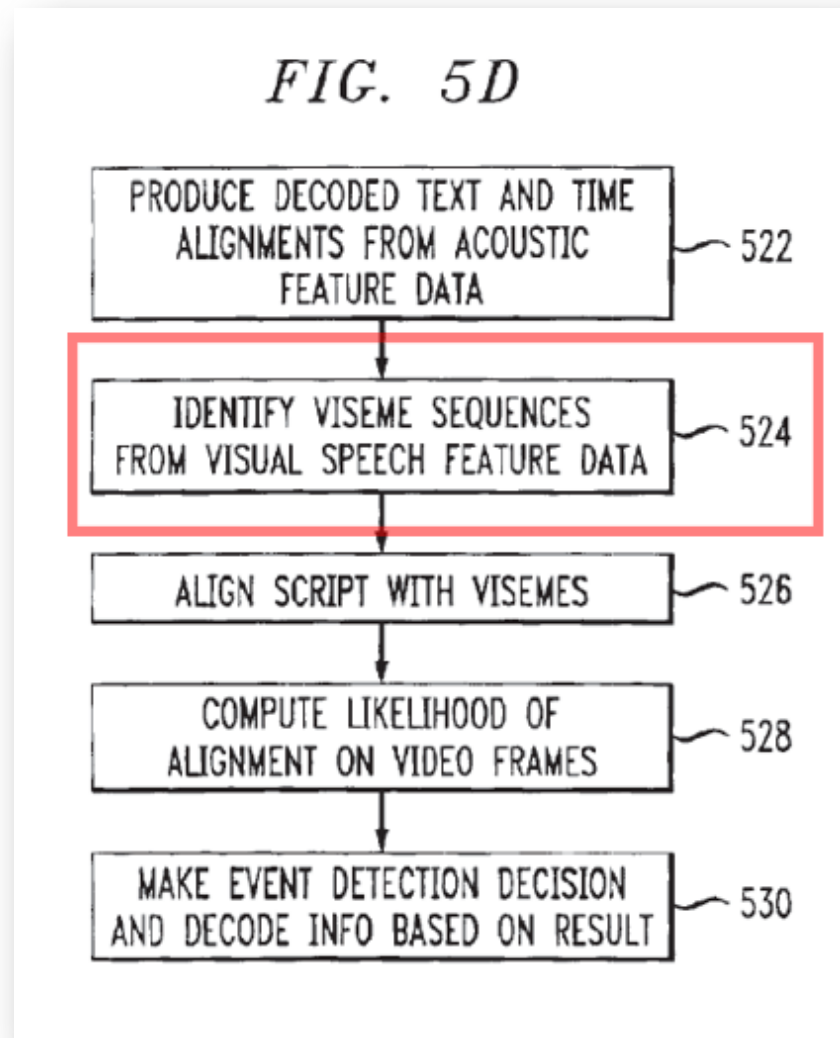
through the speech unit ..., the keypad ... or keyboard, ... or other peripheral.”
Ex.1001, 16:18–19 and 19:1–4; *see also* Dialect’s Complaint, ¶147 (keyboard-
entered natural language input (“From my checking account”).

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

**c. [12.2] generating, by the one or more processors, a non-
speech transcription from the non-speech input;**

91. Maes teaches or suggests “*generating ... a non-speech transcription
from the non-speech input.*”

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)” used to “produce”
 (“*generate*”) “a visual phonemes (visemes) sequence” (“*a non-speech
transcription*”) “*from*” the video signal (“*the non-speech input*”), e.g., the face
detected in the video frame including the user’s lip movement. Ex.1005, 17:51–55,
21:47–50. Maes discloses the “*generating ...*” step as a processing step in block 524
in Figure 5D.



Ex.1005, Figure 5D (annotated).

93. Maes further explains modules 422, 424 and 426 in Figure 4 extracting, normalizing, and labeling (e.g., “*generating*”) visemes, e.g., in a visemes sequence (“*a non-speech 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. “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 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.

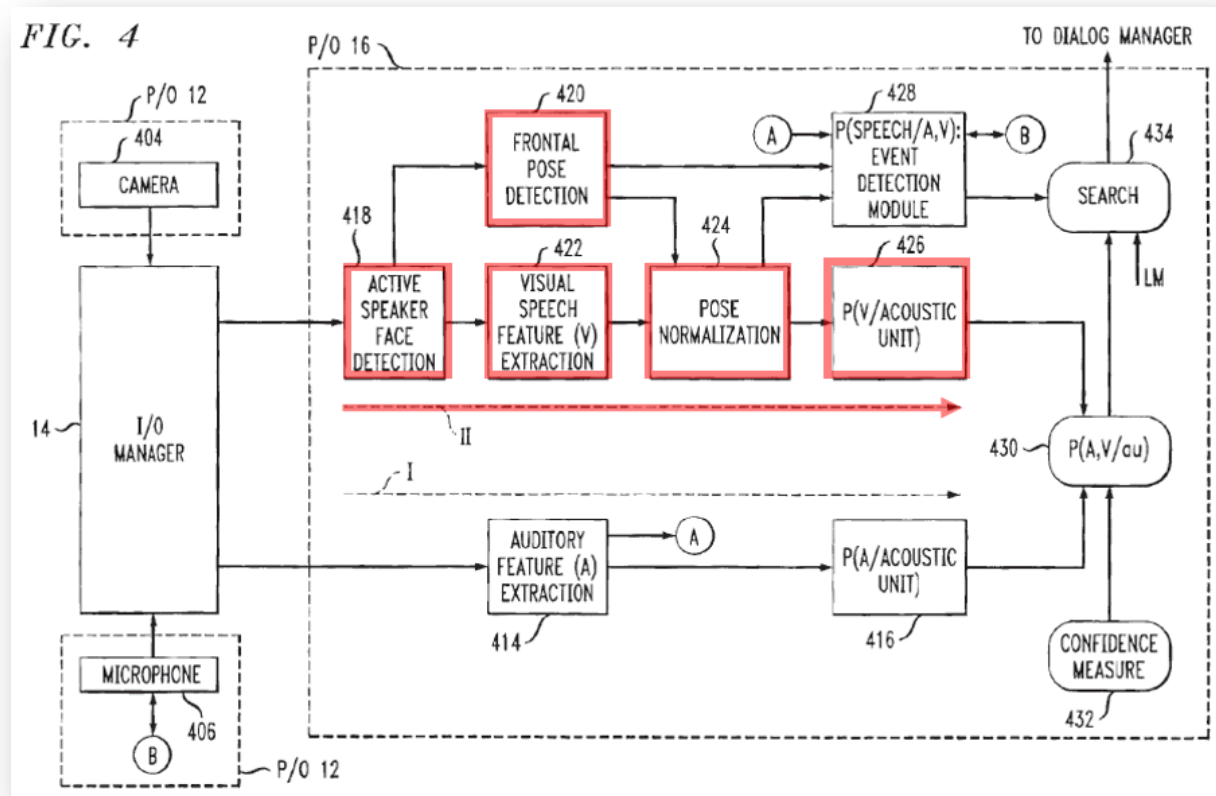
Ex.1028, Figure 5. Shdaifat further describes that visemes can be represented by images:



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.

94. In Figure 4, the visual speech feature vectors are denoted by the letter “V” and “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.



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

95. 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.

96. 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) are used to “produce a visual phonemes (visemes) sequence”

(“*generat[e] ... a non-speech transcription*”). Ex.1005, 21:43–50. Therefore, Maes discloses “*generating ... a non-speech transcription from the non-speech input.*”

97. Referring to the example in which Maes’s multi-modal conversational computing system is employed within a vehicle, Maes explains the “*generating ...*” 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.

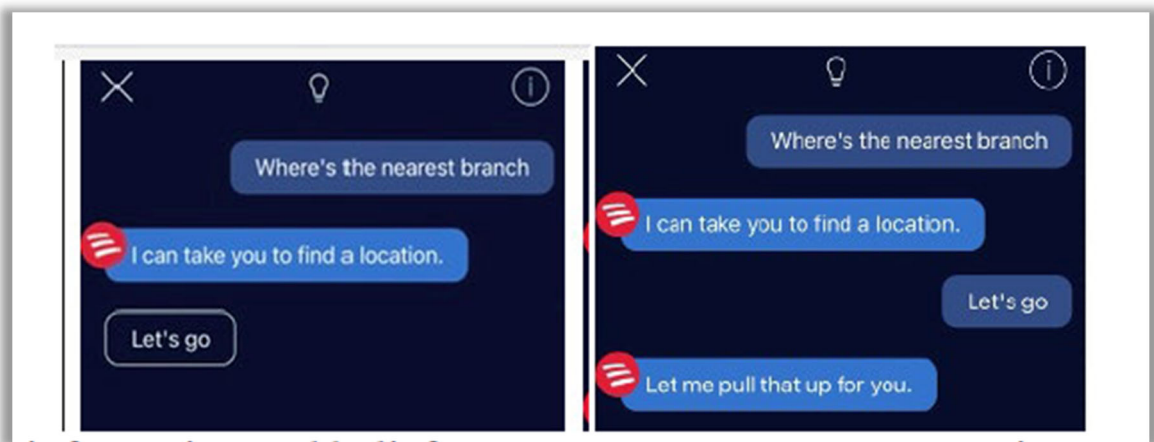
98. Additionally, as explained above in connection with element [12.1], Maes discloses “*one or more processors*” performing the methodologies of its invention.

99. Accordingly, Maes teaches or suggests this element.

Additional Mapping

100. As explained in [12.1], in Maes, keyboard-typed input constitutes an additional example 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. Therefore, a POSITA would have understood that the keyboard-typed input inherently serves as “*a non-speech 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, ¶149.



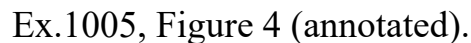
Id.

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

- d. **[12.3] identifying, by the one or more processors, the user who provided the multi-modal natural language input;**

103. Maes teaches or suggests “*identifying ... the user who provided the multi-modal natural language 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.



Meta Platforms, Inc. – Exhibit 1003
Page 59 of 118

module “*identiffies] ... the user who provided the multi-modal natural language 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 6⁷, 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.** For example, speaker

⁷ 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.

recognition module 616 may implement the recognition techniques described in the U.S. patent application identified by Ser. No. 08/788,471, filed on Jan. 28, 1997, and entitled: “Text Independent Speaker Recognition for Transparent Command Ambiguity Resolution and Continuous Access Control,” the disclosure of which is incorporated herein by reference. An illustrative speaker identification process for use in module **616** will now be described. The illustrative system is disclosed in H. Beigi, S. H. Maes, U. V. Chaudari and J. S. Sorenson, “IBM model-based and frame-by-frame speaker recognition,” *Speaker Recognition and its Commercial and Forensic Applications*, Avignon, France 1998. 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:

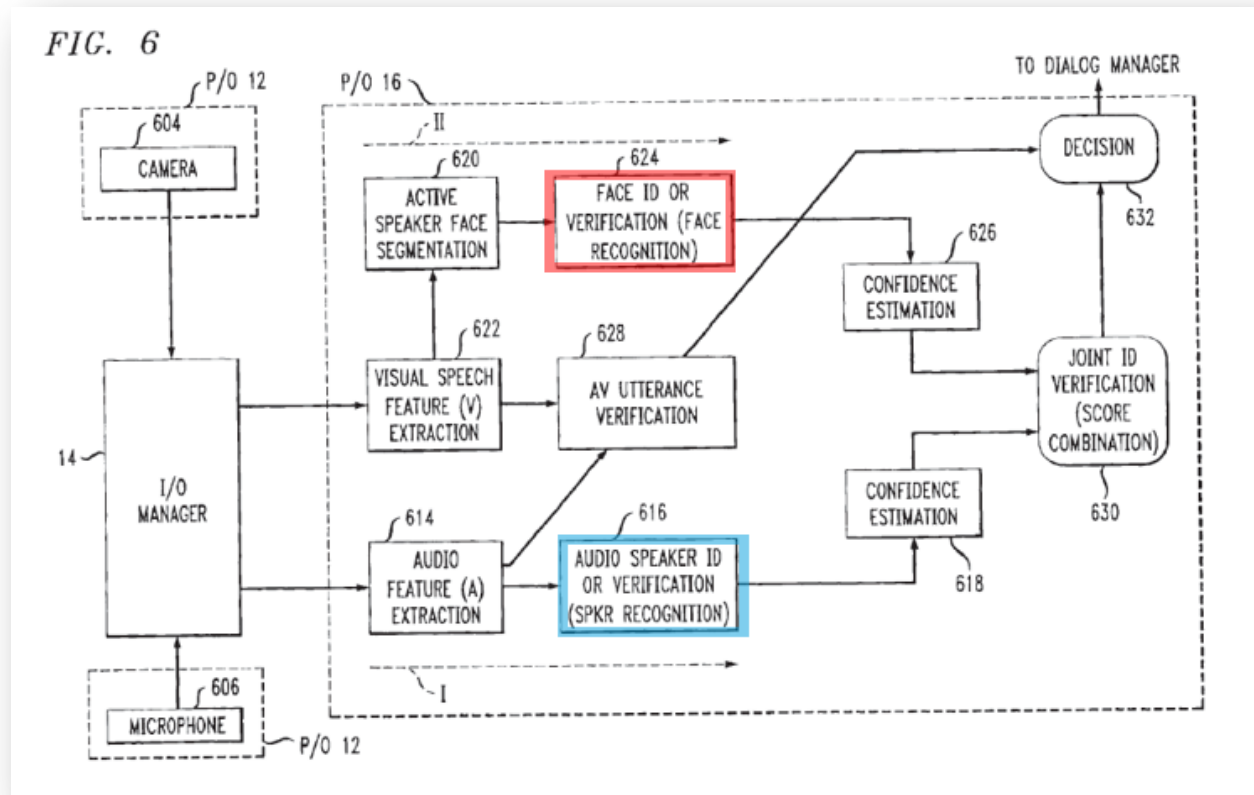
$$D_i = \sum_{n=1}^N d_{i,n}$$

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 for "*identifying the user who provided the multi-modal natural language 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**, however, 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 who provided the multi-modal natural language 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. Additionally, as explained above in connection with element [12.1], Maes discloses “*one or more processors*” performing the methodologies of its invention.

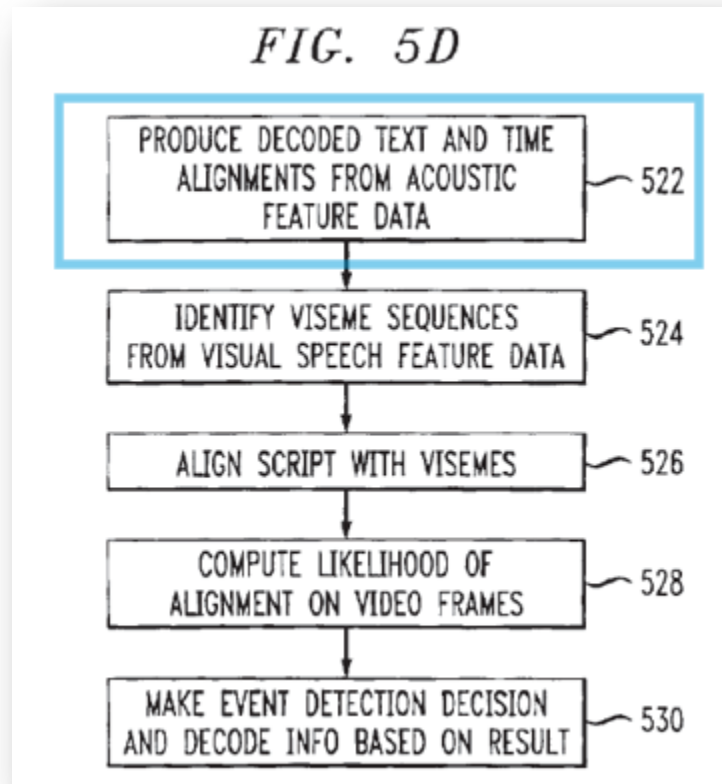
111. Accordingly, Maes teaches or suggests “*identifying, by the one or more processors, the user who provided the multi-modal natural language input.*”

- e. **[12.4] generating, by the one or more processors, a speech-based transcription based on a cognitive model associated with the user, wherein the cognitive model includes information on one or more prior interactions between the user and the device;**

112. Maes teaches or suggests “*generating ... a speech-based transcription based on a cognitive model associated with the user.*”

113. With reference to Figure 5D (reproduced below), Maes discloses classical speech recognition techniques “produc[ing]” (“*generating*”) “decoded text” (or, script) processed from the spoken utterance (“*a speech-based*

transcription”) “using the feature data from the acoustic feature extractor 414.” See generally Ex.1005, 21:43–47.



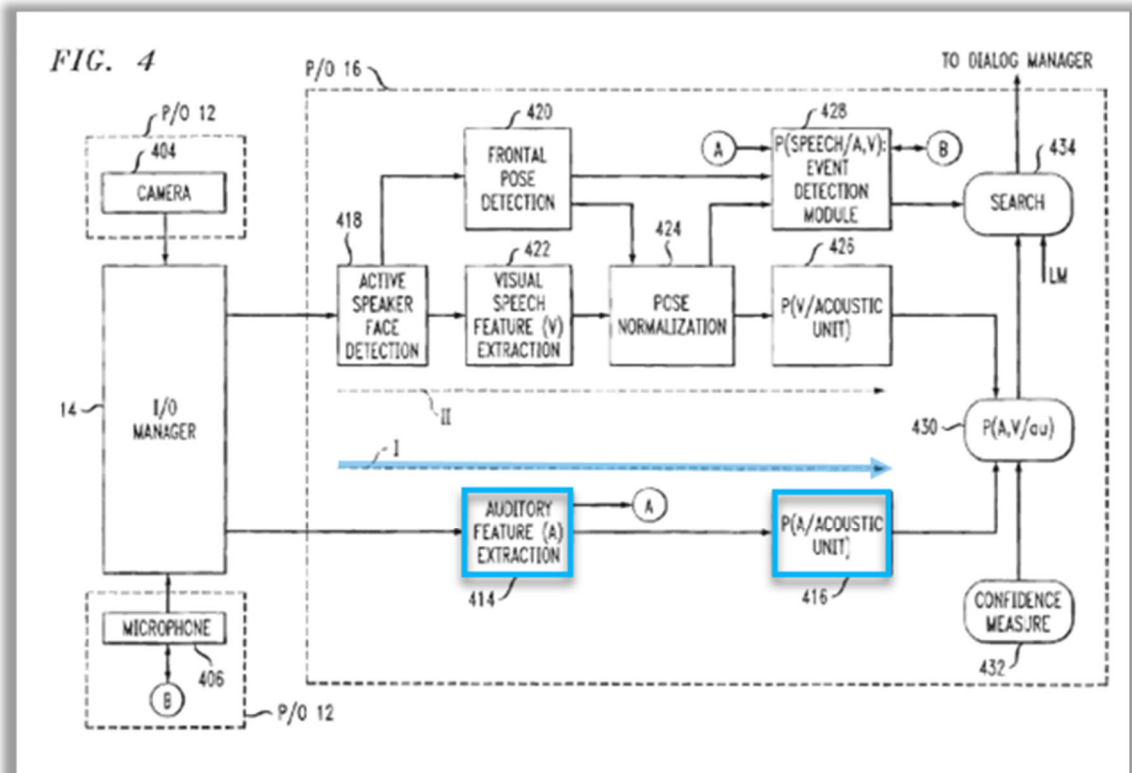
Ex.1005, Figure 5D (block 522). “[I]n step 522, the uttered speech to be verified may be **decoded** by classical speech recognition techniques so that a decoded script and associated time alignments are available. This is accomplished using the feature data from the acoustic feature extractor 414.” Ex.1005, 21:43–47. A POSITA would have understood that the “decoded script” described in Maes 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.

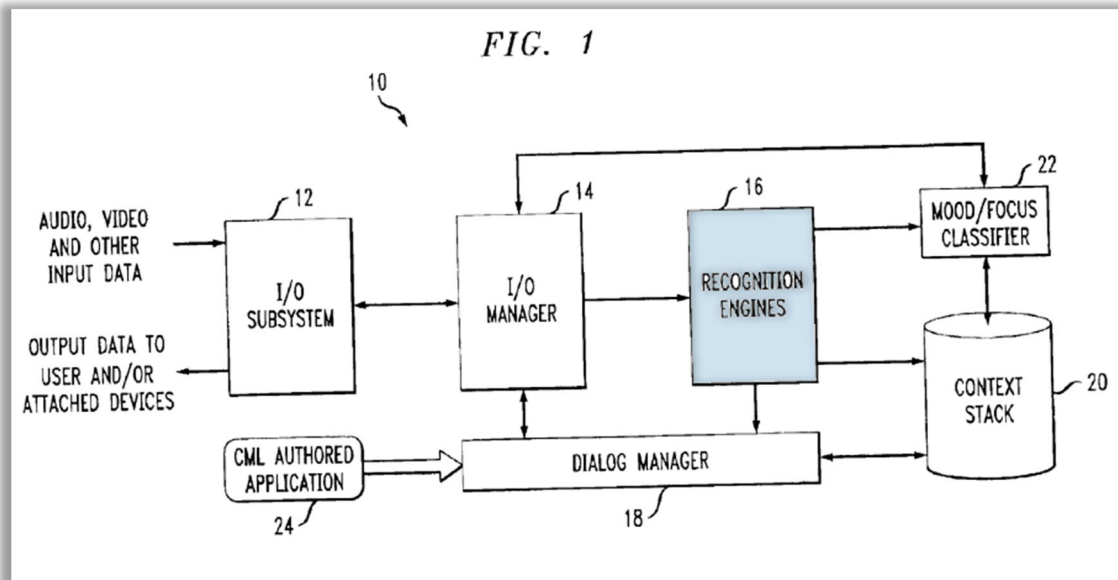
114. With reference to Figures 1 and 4 (reproduced below), Maes provides additional details explaining the process of “*generating ... a speech-based transcription*” as follows: Maes begins by describing Figure 4 as “an audio-visual speech recognition module that may be employed as **one of the recognition 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) 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 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

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.”

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).

115. 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.

116. 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. Indeed, Maes states that “the processing performed in blocks 414 and 416 may be accomplished via any conventional acoustic information recognition system 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.

117. 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.

118. Additionally, Maes discloses that the step of “producing decoded text” (or, script) (“*generating ... a speech-based transcription*”) by classical speech recognition techniques is “*based on a cognitive model associated with the user [that] includes information on one or more prior interactions between the user and the device.*”

119. Initially, the 607 Patent describes a “*cognitive model*” as information derived from a user’s interactions with the system that may be used in determining a user’s future actions by “assisting with speech recognition and/or question or command recognition”:

According to one embodiment of the invention, personalized cognitive model 810 is 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**, thus assisting with speech recognition and/or question or command recognition. Personalized cognitive model 810 may track actions performed by the user. When the system is attempting to predict user behavior, the personalized cognitive model may be consulted first. The system may have multiple personalized cognitive models, wherein one may correspond to each user.

Ex.1001, 24:1–11.

120. 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 607 Patent.

121. 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 32" that may use "recognition assisting databases 48" 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 144" 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].

122. Maes discloses cognitive models as were well-known to POSITAs in the art, as demonstrated by Lee. For example, Maes describes "features," e.g.,—acoustic feature vectors used for "produc[ing] decoded text" ("*speech-based transcription[s]*")—"can be extracted and decisions dynamically returned by **the**

models.” Ex.1005, 21:42–47, 43:53–54. With respect to Maes’s system providing dynamic decision making by models, Maes teaches that its system (“*the device*”) can include a “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. 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. Considering, for instance, native language as a user attribute, Maes teaches that “[m]eta 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, in Maes, “produc[ing] decoded text” (“*speech-based transcription[s]*”) can involve, for instance, classifying 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

cognitive model associated with the user [and] includes information on one or more prior interactions between the user and” Maes’s system (*“the device”*).

123. Furthermore, Maes explains that with regard to “recognized speech” or “decoded text” (*“speech-based transcription”*), “[r]ecently, efforts have been initiated to develop appropriate confidence measures:”

A first method uses decision trees trained on word-dependent features (amount of **training utterances**, minimum and average triphone occurrences, **occurrence in language model training**, number of phonemes/lefemes, duration, acoustic score (fast match and detailed match), speech or non-speech), sentence-dependent features (signal-to-noise ratio, estimates of speaking rates: number of words or of lefemes or of vowels per second, sentence likelihood provided by the language model, trigram occurrence in the language model), word in a context features (trigram occurrence in language model) as well as **speaker profile features** (accent, dialect, gender, age, speaking rate, identity, audio quality, SNR).

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 cognitive model associated with the user [and] includes information on one or more prior interactions between the user and*” Maes’s system (“*the device*”). Ex.1005, 19:3–20.

124. Maes teaches using “*a ... model associated with the user, wherein the ... model includes information on one or more prior interactions between the user and the device.*” In Maes’s system (“*the device*”) “[s]peaker verification may be performed in a similar manner, however, **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. It should be understood that, at 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.

125. Therefore, in view of the above-mentioned disclosures, Maes teaches or suggests that the “*generating ... a speech-based transcription*” is “*based on a cognitive model associated with the user [that] includes information on one or more prior interactions between the user and the device.*”

126. To the extent that Patent Owner contends that Maes does not disclose or suggest “*a cognitive model associated with the user [that] includes information on one or more prior interactions between the user and the device,*” U.S. Pat. No. 5,937,383 (“Ittycheriah”) incorporated in Maes teaches or suggests this claim element. See §IX.A.1. Maes states “[a]nother method to detect incorrectly recognized words is disclosed in U.S. Pat. No. 5,937,383 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. As explained below, U.S. Pat. No. 5,937,383 to “Ittycheriah” (Ex.1006) 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 cognitive model associated with the user [that] includes information on one or more prior interactions between the user and [a] device,*” such as Maes’s system. Specifically, Ittycheriah teaches:

It is to be understood that initially the cache database 18 does not have any keywords or corresponding scores stored therein, that is, until a particular user inputs certain names or commands. When the user first inputs a name or a command, the acoustic front-end 12 generates a set of acoustic features which are then provided to the classical speech recognition engine 16. 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 18 preferably partitioned for that particular user. It is to be understood that rather 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 18 [storing keywords comprising vocabulary such as names and/or commands frequently used by a user] corresponding to the same user is considered as a hypothesis with respect to the newly entered speech segment. This evaluation is done by the Viterbi alignment unit 14. The resulting scores from such comparison are then presented to the score comparator 22.

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 cognitive model*”) is “*associated with the user*”. *Id.*

127. 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., producing decoded text) 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 cognitive model associated with the user*”) resulting in

the decoded text (or, script) (“*a speech-based transcription*”) “*including information on one or more prior interactions between the user*” “*and*” Maes’s system (“*the device*”). Ex.1005, 12:12–28; Ex.1006, 3:56–60 and 4:19–53. See §IX.A.1.

128. Therefore, Maes teaches or suggests that the step of “*generating ... a speech-based transcription*” is “*based on a cognitive model associated with the user [that] includes information on one or more prior interactions between the user and the device*.”

129. Referring to the example in which Maes’s multi-modal conversational computing system is employed within a vehicle, Maes explains the “*generating ...*” step in that example 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 in Maes’s system (and/or including Ittycheariah’s teachings related to decoding history cache

database), the system produces decoded text/script corresponding to the user's spoken utterance, for example, of saying "turn it on."

130. Additionally, as explained above in connection with element [12.1], Maes discloses "*one or more processors*" performing the methodologies of its invention.

131. Therefore, Maes teaches or suggests this element.

f. [12.5] generating, by the one or more processors, a merged transcription from the speech-based transcription and the non-speech transcription;

132. Maes teaches or suggests "*generating ... a merged transcription from the speech-based transcription and the non-speech transcription.*"

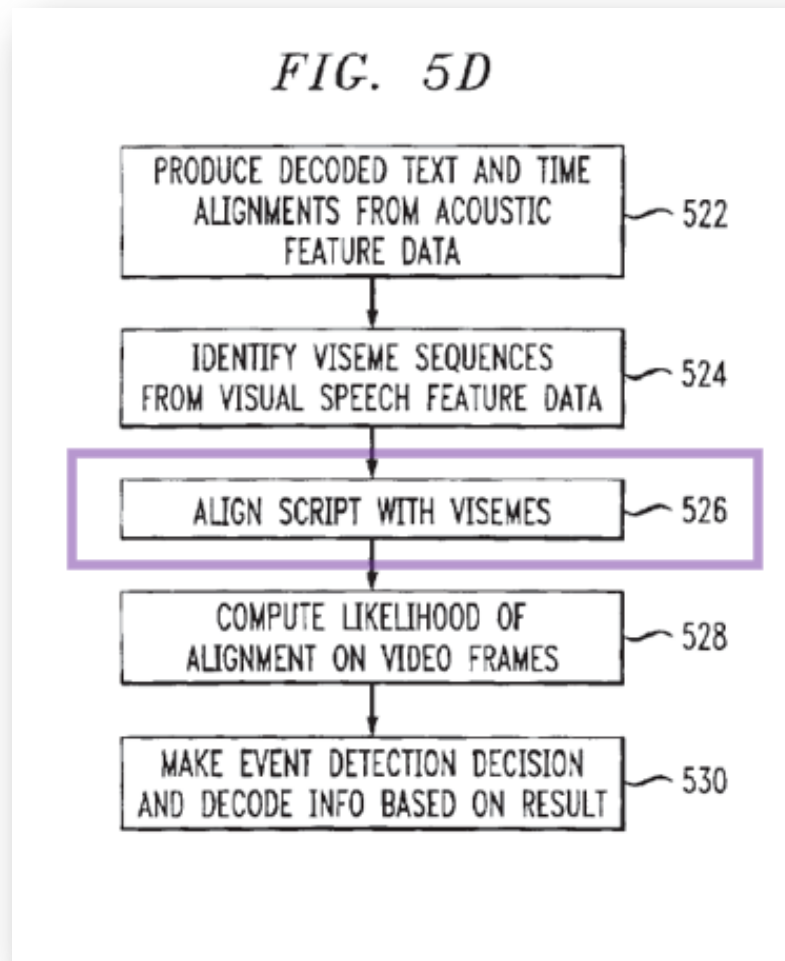
133. As explained earlier in connection with elements [12.2] and [12.4], Maes discloses "the visual phonemes (visemes) sequence" ("*the non-speech transcription*") and "decoded text" (or, script) ("*the speech-based transcription*").

134. 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. 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*”) (e.g., 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. Maes also 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 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, Figure 5D (block 526).

135. 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.

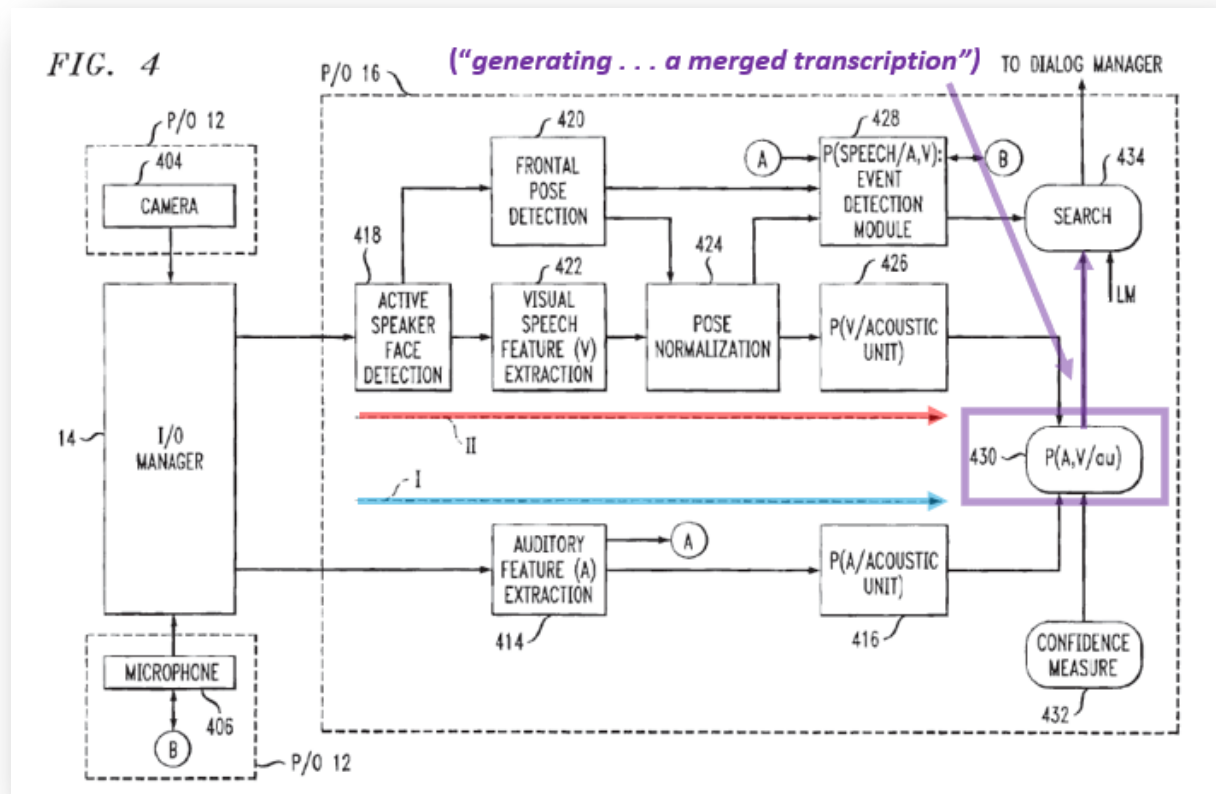
136. Therefore, by teaching synchronization of the two information streams (i.e, the audio and video information streams/signals) or, more specifically, alignment of “the decoded text” (or, script) (“*the speech-based transcription*”) with “the visual phonemes (visemes) sequence” (“*the non-speech transcription*”),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

137. 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.

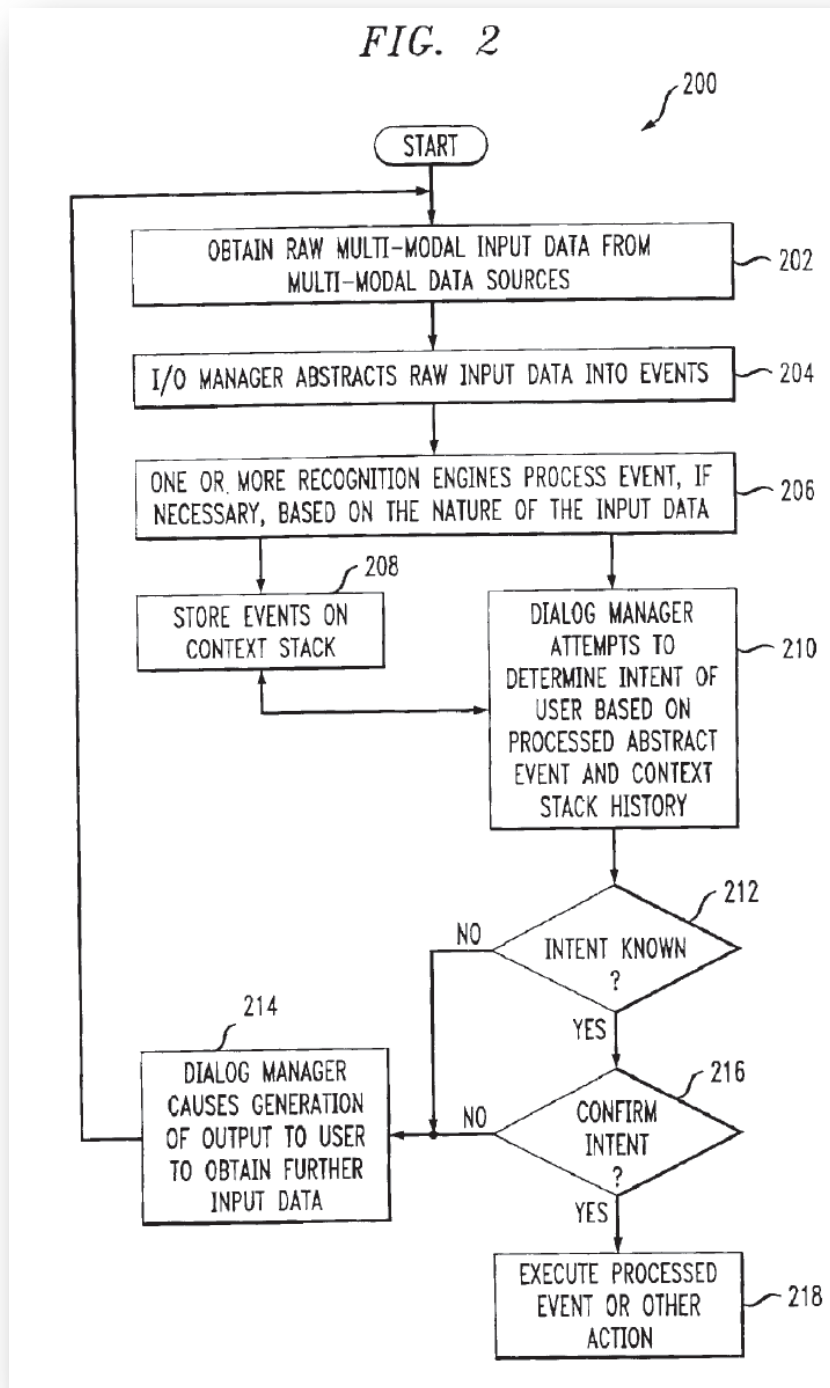
Ex.1005, 7:65–8:4, 8:37–42. In this example, the aligned decoded script and visemes sequence associated with the spoken utterance “turn it on” with the accompanying lip movement, therefore, constitutes “*the merged transcription.*”

138. 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 “*generating ... a merged transcription*” step such that the outcome of the “*generating ...*” step produces the aligned decoded script and visemes sequence (i.e., “*a merged transcription*”), e.g., associated with the spoken utterance “turn it on” with the accompanying lip movement.



Ex.1005, Figure 4 (annotated).

139. 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.



Ex.1005, 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.

140. 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.

141. 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 (*i.e.*, “*merged transcription*”) 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. *See also* 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**”).

142. 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.

143. Therefore, Maes teaches or suggests “*generating ... a merged transcription from the speech-based transcription and the non-speech transcription.*”

144. Referring to the example in which Maes’s multi-modal conversational computing system is employed within a vehicle, the “*generating ...*” step would result in the system aligning/synchronizing the abstracted data (e.g., decoded text/script) processed from the spoken utterance of saying “turn it on” with the abstracted data (e.g., visemes sequence) processed from the lip movement of the user. And, further, Maes identifies the goal and advantages of aligning/synchronizing the audio and video information streams. See Ex.1005, 29:30–43.

145. Additionally, as explained above in connection with element [12.1], Maes discloses “*one or more processors*” performing the methodologies of its invention.

146. Therefore, Maes teaches or suggests “*generating, by the one or more processors, a merged transcription from the speech-based transcription and the non-speech transcription.*”

Additional Mapping

147. As explained in [12.1], [12.2], and [12.3], in Maes, keyboard-typed input constitutes an additional example of “*a non-speech input*” and further, it also

serves as “*the non-speech transcription.*” Additionally, as explained in [12.4], in Maes’s system, “decoded text” (or, script) processed from the spoken utterance constitutes “*the speech-based transcription.*”

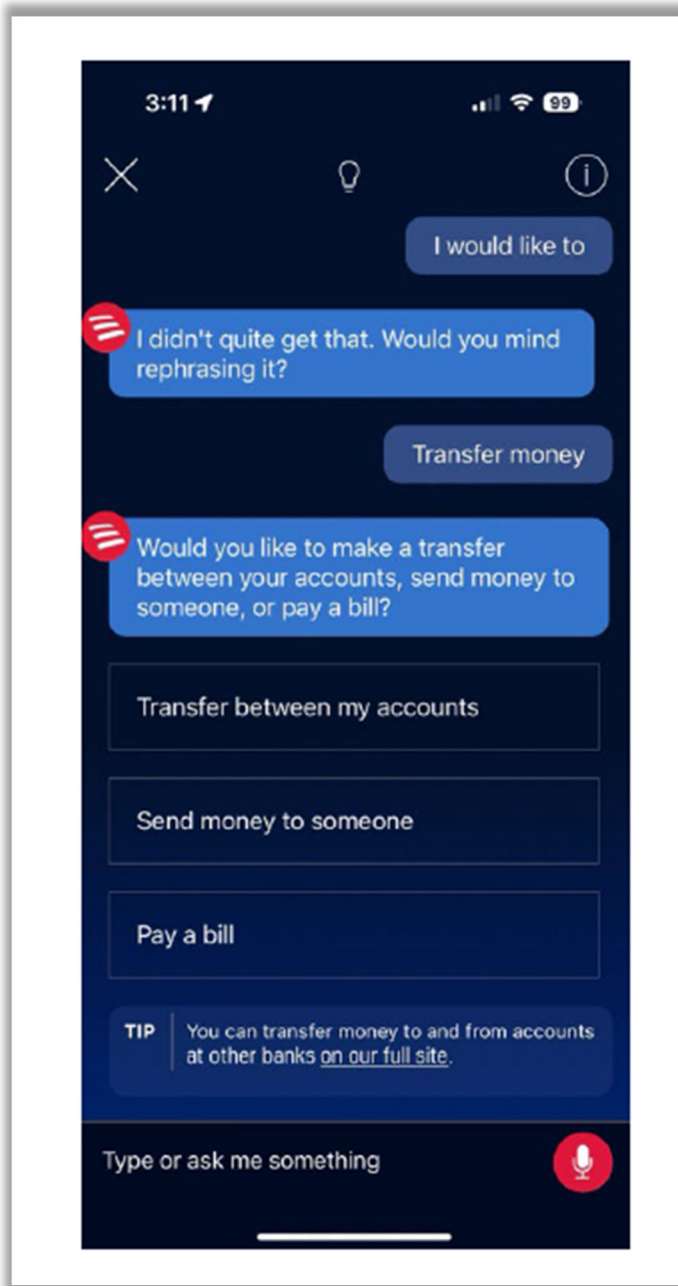
148. Coffman teaches “*generating ... a merged transcription*” because Coffman discloses “merg[ing]” input streams. For example: “**combine input keyed on the keyboard, handwritten input and speech input.**” Ex.1012, 38:2–4. Coffman continues:

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.

149. 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 communication (the words “Transfer money” which were input as speech via the microphone interface,) creating a merged transcription.” Ex.1007, ¶153.



Id.

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

- g. **[12.6] identifying, by the one or more processors, an entry in a context stack, from among a plurality of entries that are in the context stack and that are each indicative of context, wherein the identified entry matches information in the merged transcription;**

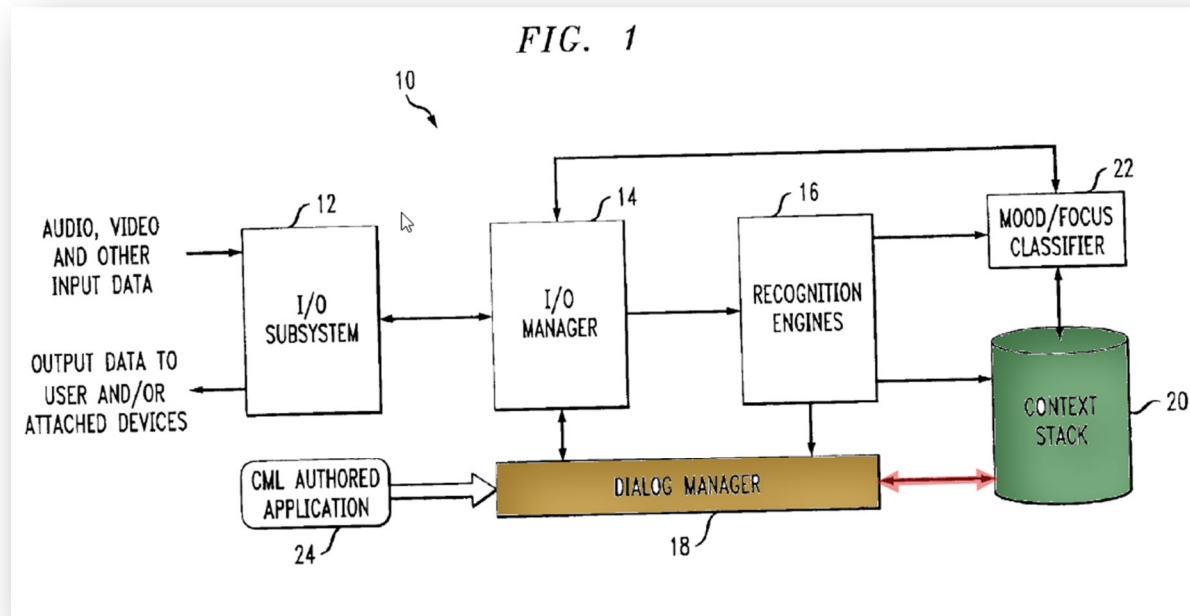
151. Maes teaches or suggests this claim element.

152. Maes teaches or suggests “*identifying ... an entry in a context stack, from among a plurality of entries that are in the context stack and that are each indicative of context.*” Maes explains storing data relating to events (“*a plurality of ... entries*”) in “a context stack” (such as context stack 817 in Figure 8 or context stack 20 in Figure 1):

[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’”).



Ex.1005, Figure 1 (annotated).

153. 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 [including the accompanying lip movement] and gesture could be associated.

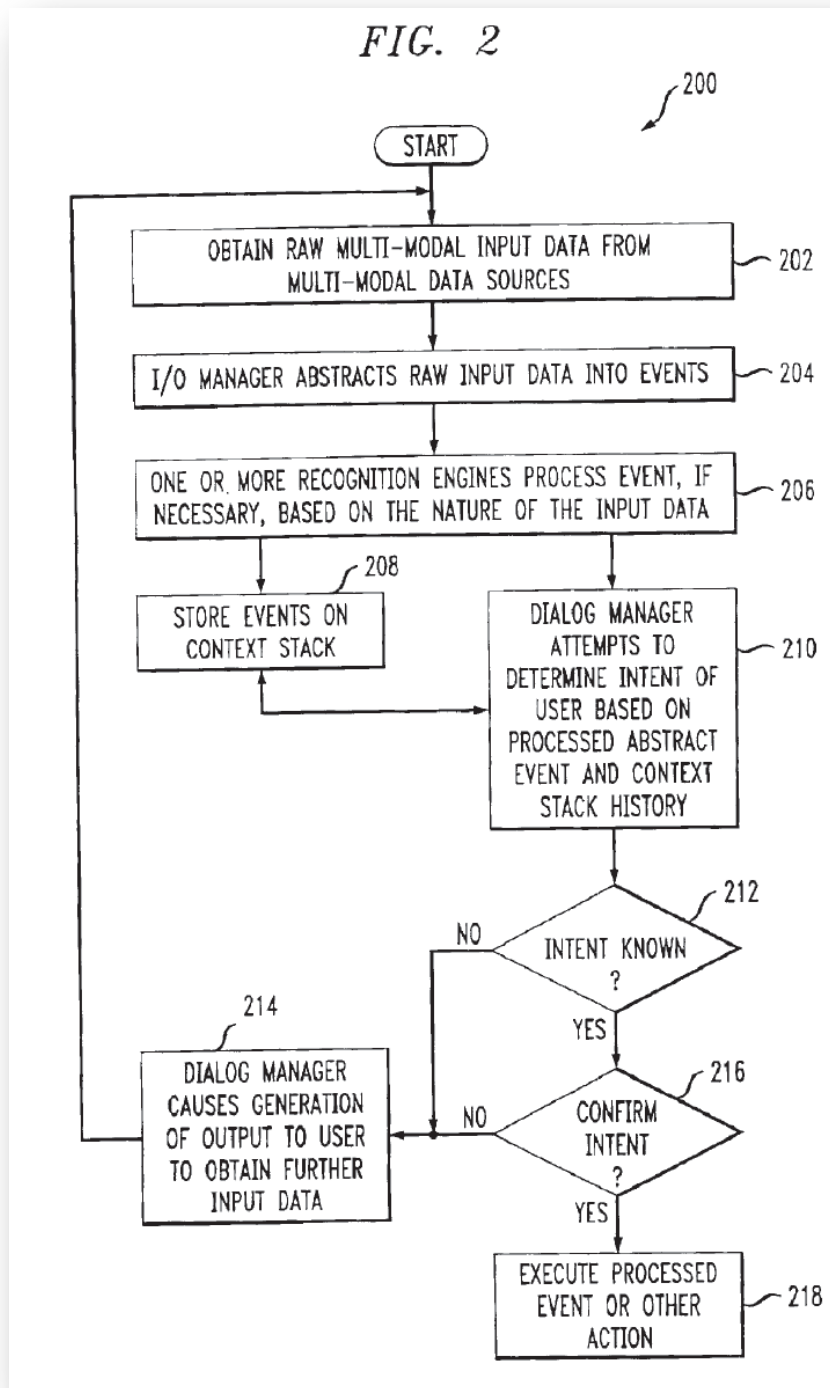
Ex.1005, 7:65–8:4.

[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. 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*”). 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. Thus, “a dialog” in Maes comprises speech and video data. 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 (“*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 an entry in a context stack*” and consequently, the associated context(s) stored in the context stack corresponding to the user’s intent in the active dialog constitutes “*the identified entry,*” which “*matches*” the results of the recognized events (“*information*”) included in the aligned decoded script and visemes sequence produced as a result of merging the transcriptions) (“*in the merged transcription*”). See also Ex.1005, 37:53–55.

154. 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.



Ex.1005, Figure 2.

155. 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.

156. 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.

157. 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*”. *See also* 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**”); *see also* detailed explanation for “*merged transcription*” provided earlier in connection with element [12.5]. Maes thus teaches or suggests “*wherein the identified entry matches information in the merged transcription.*”

158. Maes continues: “[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 **so as 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** [*i.e.*, “*wherein the identified entry matches information in the merged transcription.*”] **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" (*"identifying ... an entry in a context stack, from among a plurality of entries that are in the context stack"*). Ex.1005, 7:33–45.

159. Moreover, Maes explains that *"each* [entry in a context stack is] *indicative of context"*. "The context stack 817 comprises all the information associated with an application. Such information includes **all the variable, 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. The context stack is associated with the organized/sorted context corresponding to each active dialog." Ex.1005, 37:55–62.

160. Accordingly, Maes teaches or suggests *"identifying ... an entry in a context stack, from among a plurality of entries that are in the context stack and that are each indicative of context."*

161. In view of the above-mentioned teachings, Maes discloses that upon searching the context stack, the associated context(s) stored in the context stack corresponding to the user's intent in the active dialog (*"the identified entry"*) is provided by the dialog manager when such context "could be associated" (*"matches"*) the results of the recognized events (*"information"*) included in the aligned decoded script and visemes sequence (*"in the merged transcription"*), created as a result of aligning/synchronizing the abstracted data (e.g., decoded

text/script) processed from the spoken utterance with the abstracted data (e.g., visemes sequence) processed from the lip movement of the user accompanying the spoken utterance. *See* Ex.1005, 7:60–8:4; *see also* 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.**” (i.e., “*merged transcription*”); *see also* detailed explanation for “*merged transcription*” provided earlier in connection with elements [12.5] and [12.6].

162. Maes further discloses “*a plurality of entries ... in the context stack and that are each indicative of context.*” In the above-mentioned teachings, Maes discloses that the context stack is associated with the organized/sorted context corresponding to each active dialog such that the information stored in the context stack (such as variable, states, input, output and queries) is in the context of a dialog—and therefore, when storing contexts of multiple dialogs, Maes’s context stack stores “organized/sorted contexts corresponding to multiple dialogs” (“*a plurality of entries*”) “*that are each indicative of context*” of a dialog. Therefore, Maes teaches or suggest “*a plurality of entries ... in the context stack and that are each indicative of context.*”

163. Additionally, as explained above in connection with element [12.1], Maes discloses “*one or more processors*” performing the methodologies of its invention.

164. To the extent that Patent Owner contends that Maes does not disclose or suggest “*identifying... an entry in the context stack that matches information in the merged transcription,*” Coffman discloses this element. See §IX.A.1. First, Maes incorporates Coffman by reference: “It 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,’ 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.

165. 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 *“identifying...an entry in the context stack that matches information in the merged transcription.”* Accordingly, based on Maes and the incorporation of Coffman, a POSITA would have recognized that the results of the recognized events included in the aligned decoded text/script and visemes sequence (*“the merged transcription”*) produced as a result of processing the user's multi-modal input in Maes would be compared against data stored on the context stack, per Coffman's teachings, to *“identify[]...an entry in the context stack that matches information in the merged transcription.”*

166. Therefore, Maes teaches or suggests this element.

Additional Mapping

167. As explained in [12.2], [12.3], [12.4], and [12.5], in Maes, keyboard-typed input constitutes an additional example of “*a non-speech input*” and further, as explained in [12.5], Coffman teaches the “*generating ... a merged transcription*” in which the combined keyboard-typed input and the transcribed speech input together constitute “*a merged transcription.*” As further explained in [12.6], 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...*” step. Ex.1012, 42:1–3. A POSITA would have understood from the above-mentioned teachings of Coffman that the combined keyboard-typed input and the transcribed speech input taken together (“*the merged transcription*”) would be compared against data stored on the context stack to “*identify[]...an entry in the context stack that matches information in the merged transcription.*”

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

h. [12.7] identifying, by the one or more processors, a domain agent associated with the entry in the context stack;

169. Maes teaches or suggests this claim element.

170. Initially, the 607 Patent describes a “*domain agent*” as follows: “In some embodiments ... generic and domain specific behavior and information may

be organized into domain agents 156. ... 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 607 Patent broadly refer to software modules that are specific to each application area.

171. In Maes, the associated context(s) stored in the context stack corresponding to the user’s intent in the active dialog constitutes “*the entry in the context stack*” and which is the outcome of the “*identifying ...*” step, as explained in connection with element [12.6]. Furthermore, for the reasons below, Maes discloses “*identifying ... a domain agent associated with the entry in the context stack.*” For example, Maes discloses:

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. The application depends on the environment that the system is deployed in.**

Ex.1005, 7:40–46.

172. As such, Maes’s teachings of application programs (effectuating the user’s intention and depending on the environment that the system is deployed) are analogous to the 607 Patent’s description of “*domain agent*,” and therefore “application programs” in Maes serve as “*domain agent[s]*.” Maes explains this

further in greater detail. 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 [associated with the organized/sorted context corresponding to
each active dialog (or deferred dialog-agents/memorization)].**” *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,
38:51-56.

173. 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.*” Therefore,
Maes discloses “determining” (“*identifying*”) “the target application or dialog (or
portion of it)” (“*a domain agent*”) “*associated with*” the associated context(s) stored
in the context stack corresponding to the user’s intent in the active dialog (“*the
entry*”) “*in*” the context stack (such as context stack 817 in Figure 8/context stack
20 in Figure 1) (“*the context stack*”). *See also* Ex.1005, 37:53–55. Accordingly,

Maes teaches or suggests “*identifying ... a domain agent associated with the entry in the context stack.*”

174. Additionally, as explained above in connection with element [12.1], Maes discloses “*one or more processors*” performing the methodologies of its invention.

175. Therefore, Maes teaches or suggests this element.

- i. **[12.8] determining, by the one or more processors, a request based on the merged transcription; and**

176. Maes teaches or suggests “*determining, by the one or more processors, a request based on the merged transcription.*”

177. First, as explained earlier in connection with elements [12.5] and [12.6], in Maes, the aligned decoded script and visemes sequence constitutes “*the merged transcription.*” And, further, Maes discloses

[t]he conversational dialog manager 819 [comprising functions for managing the dialog (conversational dialog comprising speech and other multi-modal I/O such as GUI keyboard, pointer, mouse, as well as video input, etc.) and **arbitration (dialog manager arbitrator or DMA) across all registered applications] determines what information the user has, which inputs the user presents, and which application(s) should handle the user inputs. The DMA processes abstracted I/O events (abstracted by the I/O manager).**

Ex.1005, 36:49–59. Specifically:

[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 one or more events, e.g., a spoken utterance, a visual gesture. ... [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.

When an abstract event occurs, the DMA **determines** the target of the event and ... then launches the action associated to **the user’s query [e.g., received as multi-modal data in the form of audio input data, video input data, as well as other types of input data]**.

Ex.1005, 36:59–37:3; *see also* Ex.1005, 4:6–22. 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. Therefore, Maes discloses “determin[ing]” (“*determining*”) “a user’s query” (“*a request*”) (associated with the multi-modal data in the form of audio input data, video input data, and possibly other

types of input data) “*based on*” processing the user input multi-modal data yielding “a more generalized representation of the data” (e.g., the decoded script and visemes sequence) included in “the aligned decoded script and visemes sequence” (“*the merged transcription*”). Advantageously, Maes discloses that its multi-modal conversational computing system:

provides the capability to: (i) determine an object, application or appliance addressed by the user; (ii) determine the focus of the user and therefore determine if the user is actively focused on an appropriate application and, on that basis, to determine if an action should be taken; (iii) understand queries based on who said or did what, what was the focus of the user when he gave a multi-modal query/command and what is the history of these commands and focuses; and (iv) estimate the mood of the user and initiate and/or adapt some behavior/service/appliances accordingly. The computing system may also change the associated business logic of an application with which the user interacts.

Ex.1005, 2:54–67. Therefore, Maes discloses “*determining ... a request based on the merged transcription.*”

178. Additionally, as explained above in connection with element [12.1], Maes discloses “*one or more processors*” performing the methodologies of its invention.

179. Therefore, Maes teaches or suggests this element.

- j. **[12.9] communicating, by the one or more processors, the request to the domain agent, wherein the domain agent is configured to generate a response to the user.**

180. Maes teaches or suggests “*communicating, by the one or more processors, the request to the domain agent, wherein the domain agent is configured to generate a response to the user.*”

181. First, as explained earlier in connection with element [12.7], Maes’s teachings of application programs (effectuating the user’s intention and depending on the environment that the system is deployed) are analogous to the 607 Patent’s description of a “*domain agent.*” Further, Maes discloses that “[t]he DMA [function of the dialog manager] processes abstracted I/O events (abstracted by the I/O manager) using the context/history to understand the user intent. When an abstract event occurs, the DMA determines the target of the event and ... then launches the action associated to the user’s query.” Ex.1005, 36:49–37:3. For example,

[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 (e.g., 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. That is, the dialog manager 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, 8:8–8:19. Maes identifies that its system “provides the capability to: (i) determine an object, application or appliance addressed by the user; (ii) determine the focus of the user and therefore determine if the user is actively focused on an appropriate application and, on that basis, to determine if an action should be taken; (iii) understand queries based on who said or did what, what was the focus of the user when he gave a multi-modal query/command and what is the history of these commands and focuses; and (iv) estimate the mood of the user and **initiate and/or adapt some behavior/service/appliances accordingly.**” Ex.1005, 2:54–64. Therefore, launching or executing the action (e.g., selecting an application such as a radio activation routine) associated with the user’s query (e.g., comprising a request for launching an application such as a radio activation routine) serves as “communicating ... the request to the domain agent.”

182. Further, because Maes discloses that “the dialog manager 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, 8:15–19), a POSITA would have therefore recognized that “the application” (“*the domain agent*”), such as a “radio

activation routine,” would be accordingly activated, or otherwise some behavior/service/appliances would be initiated and/or adapted to “*generate a response to the user*” (e.g., cause the radio to be turned on) upon receiving a command to activate the radio based on Maes. *See generally* Ex.1005, 7:60–8:29; 2:54–64. Indeed, Maes teaches that the application (e.g., “*domain agent*”) having an “associated business logic ... with which the user interacts.” Ex.1005, 2:65–67. A POSITA would have understood that, in the process of interacting with the user, the business logic of an application would “*generate a response to the user.*” Ex.1005, 2:65–67. Therefore, Maes teaches or suggests “*wherein the domain agent is configured to generate a response to the user.*”

183. Additionally, as explained above in connection with element [12.1], Maes discloses “*one or more processors*” performing the methodologies of its invention.

184. Therefore, Maes teaches or suggests this element.

3. Dependent Claim 13

- a. [13.0] The method of claim 12, wherein the cognitive model is a first cognitive model, and wherein the generating the speech-based transcription is based on a second cognitive model that includes information on one or more interactions between the device and a plurality of users of the device.**

185. Maes teaches or suggests this claim element.

186. First, Maes describes features such as acoustic feature vectors included in processed audio information that are used to “*generat[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 “*cognitive model.*” Ex.1005, 43:53–54. Also, as explained in connection with element [12.4], Maes teaches or suggests “[*a*] *cognitive model [that] includes information on one or more prior interactions between the user and the device,*” and therefore teaches or suggests “*wherein the cognitive model is a first cognitive model*” of this claim.

187. Furthermore, Maes teaches or suggests “*generating the speech-based transcription is based on a second cognitive model that includes information on one or more interactions between the device and a plurality of users of the device.*” Maes teaches that its system (“*the device*”) can include a “a dynamic classification module 1040 which performs dynamic classification of the user” such that the system is “configured to modify behavior of the voice system 1000 **based on at least one user attribute** which has been determined based on at least one acoustic feature extracted from the user’s speech.” Maes, 44:5–11. Maes explains that the “at least one user attribute can include at least one of the following: 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.” Maes, 39:50–55. For example, considering native language as a user

attribute, Maes teaches that “[m]eta 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”** which explicitly discloses that Maes uses “*a ... model that includes information on ... a plurality of users*” Maes, 40:18–23. Thus, in Maes, “produc[ing] decoded text” (“*speech-based transcription[s]*”) can involve, for instance, classifying 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 (“*a second cognitive model that includes information on one or more interactions between the device and a plurality of users of the device*”).

188. Maes teaches using “*a ... model that includes information on ... interactions between the device and a plurality of users.*” In Maes’s system (“*the device*”) “[s]peaker verification may be performed in a similar manner, however, **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. It should be understood that, at 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.

189. Therefore, in view of the above disclosures, Maes teaches or suggests that the “*generating the speech-based transcription is based on a second cognitive model that includes information on one or more interactions between the device and a plurality of users of the device.*”

190. Additionally, Maes, which incorporates by reference Ittycheriah (see Ex.1005, 19:43–49), discloses “*wherein the generating the speech-based transcription is based on a second cognitive model that includes information on one or more interactions between [a] device and a plurality of users of the device.*” As explained earlier in connection with element [12.4], Ittycheriah first making use of data (e.g., scores and words decoded in accordance with Hidden Markov models (HMMs) from previous interactions) 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 first*

cognitive model” that includes information on one or more prior interactions between a single user/speaker and a device (such as Maes’s system).

191. Ittycheriah additionally teaches a speaker classification approach (e.g., “*a second cognitive model*”) that involves storing keywords and scores from past history associated with a class of speakers, i.e., “*includes information on one or more interactions between [a] device and a plurality of users of the device*”:

It is possible to extend the concept of a cache database for storing keywords and scores to unknown **speakers** who have not enrolled in the speech recognition system and, therefore, have no prior history. **This may be accomplished by vector quantization (VQ) clustering of the speakers** as is disclosed in U.S. Ser. No. 08/787,031. Also, a new speaker may be **associated with a class of speakers with similar characteristics to himself** or herself wherein the acceptance or rejection decision is based **on past history and an average previous likelihood associated with other speakers within that class**. Such a **speaker classification approach may be done completely speaker independently, in which case, the cache database may preferably be built common to all users in the class**.

Ex.1006, 5:5–19. A POSITA would have found it obvious to modify Maes’s system to include Ittycheriah’s teachings of a speaker classification approach (e.g., “*a second cognitive model*”) that involves storing in a cache database keywords and scores from past history of a class of speakers would result in decoded text (or, script) (“*a speech-based transcription*”) produced from Maes’s system “*including*

information on one or more interactions between [a] device” such as Maes’s system “and a plurality of users of the device.” Ex.1006, 5:5–19. Therefore, Maes, which incorporates by reference Ittycheriah (see Ex.1005, 19:43–49), teaches or suggests “wherein the generating the speech-based transcription is based on a second cognitive model that includes information on one or more interactions between the device and a plurality of users of the device.”

192. Therefore, Maes teaches or suggests this claim.

B. Ground 2: Maes, Coffman and Ittycheriah Render Obvious claims 12 and 13

193. 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 12 and 13 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.A.1.

X. OBJECTIVE INDICIA OF NON-OBVIOUSNESS

194. I am not aware of any objective indicia of non-obviousness that would change my opinion that the challenged claims of the 607 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.

XI. CONCLUSION

195. 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.



Henry H. Houh

Dated: July 24, 2025

CLAIM LISTING

Claim 12	
[12.0]	A method for processing natural language inputs, comprising:
[12.1]	receiving, by one or more processors, a multi-modal natural language input from a user, the multi-modal natural language input including a natural language utterance and a non-speech input;
[12.2]	generating, by the one or more processors, a non-speech transcription from the non-speech input;
[12.3]	identifying, by the one or more processors, the user who provided the multi-modal natural language input;
[12.4]	generating, by the one or more processors, a speech-based transcription based on a cognitive model associated with the user, wherein the cognitive model includes information on one or more prior interactions between the user and the device;
[12.5]	generating, by the one or more processors, a merged transcription from the speech-based transcription and the non-speech transcription;
[12.6]	identifying, by the one or more processors, an entry in a context stack, from among a plurality of entries that are in the context stack and that are each indicative of context, wherein the identified entry matches information in the merged transcription;

[12.7]	identifying, by the one or more processors, a domain agent associated with the entry in the context stack;
[12.8]	determining, by the one or more processors, a request based on the merged transcription; and
[12.9]	communicating, by the one or more processors, the request to the domain agent, wherein the domain agent is configured to generate a response to the user.
Claim 13	
[13.0]	The method of claim 12, wherein the cognitive model is a first cognitive model, and wherein the generating the speech-based transcription is based on a second cognitive model that includes information on one or more interactions between the device and a plurality of users of the device.