

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

MICROSOFT CORPORATION

IPR2025-00659

U.S. Patent No. 9,495,957

**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	7
II. QUALIFICATIONS AND PROFESSIONAL EXPERIENCE	9
III. LEVEL OF ORDINARY SKILL IN THE ART	15
IV. RELEVANT LEGAL STANDARDS	16
V. OVERVIEW OF THE 957 PATENT	21
A. The Disclosures of the 957 Patent.....	21
VI. CLAIM CONSTRUCTION	24
VII. IDENTIFICATION OF HOW THE CLAIMS ARE UNPATENTABLE.....	24
VIII. THE PRINCIPAL PRIOR ART	25
A. Maes	25
B. The Combination of Maes’s Teachings with those of Coffman.....	27
IX. CLAIMS 7 AND 8 ARE UNPATENTABLE	31
A. Ground 1: Maes renders obvious claims 7 and 8.....	31
1. Independent Claim 7	31
a. [7.0] A computer-implemented method of processing a natural language utterance, the method comprising:	31
b. [7.1] generating a context stack comprising context information that corresponds to a plurality of prior utterances, wherein the context stack includes a plurality of context entries;.....	38
c. [7.2] receiving, by one or more computer processors, the natural language utterance, wherein the natural language utterance is associated with a command or is associated with a request;	44
d. [7.3] determining one or more words of the natural language utterance by performing speech recognition on the natural language utterance;.....	49
e. [7.4.0] identifying, from among the plurality of context entries, one or more context entries that correspond to the one or more words, wherein the	

	context information includes the one or more context entries,.....	60
f.	[7.4.1] wherein identifying the one or more context entries comprises: comparing the plurality of context entries to the one or more words;	68
g.	[7.4.2] [wherein identifying the one or more context entries comprises:] generating, based on the comparison, one or more rank scores for individual context entries of the plurality of context entries; and.....	71
h.	[7.4.3] [wherein identifying the one or more context entries comprises:] identifying, based on the one or more rank scores, the one or more context entries from among the plurality of context entries; and	74
i.	[7.5] determining based on the determined one or more words and the context information, the command or the request associated with the natural language utterance.	75
2.	Dependent Claim 8.....	80
a.	[8.0] The method of claim 7, further comprising: providing a prompt regarding the command or request, wherein the prompt is a prompt for additional information or a prompt for confirmation; and.....	80
b.	[8.1] receiving, in response to providing the prompt, a non-speech input, wherein the command or request is determined further based on the non-speech input.	84
B.	Ground 2: Maes and Coffman render obvious claims 7 and 8.....	84
C.	Ground 3: Kennewick, Weissman, and Maes renders obvious claims 7 and 8.....	85
X.	OBJECTIVE INDICIA OF NON-OBVIOUSNESS	88
XI.	CONCLUSION.....	88
	CLAIM LISTING	90

EXHIBIT LIST

No.	Exhibit Description
1001	U.S. Patent No. 9,495,957
1002	File History of U.S. Patent No. 9,495,957
1003	Declaration of Dr. Henry Houh
1004	CV of Dr. Henry Houh
1005	U.S. Patent Pub. No. 6,964,023 (“ <u>Maes</u> ”)
1006	RESERVED
1007	Second Amended Complaint
1008	RESERVED
1009	RESERVED
1010	RESERVED
1011	RESERVED
1012	International Patent Application Publication No. WO 00/20962 (“ <u>Coffman</u> ”) filed as PCT/US99/22927
1013	U.S. Patent No. 7,137,126 (“ <u>Coffman</u> 126”)
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) (“ <u>Gibbon</u> ”)
1018	RESERVED

1019	RESERVED
1020	RESERVED
1021	RESERVED
1022	RESERVED
1023	RESERVED
1024	RESERVED
1025	U.S. Patent No. 6,453,315 (“ <u>Weissman</u> ”)
1026	RESERVED
1027	RESERVED
1028	RESERVED
1029	EDTX Calendar, Judge Gilstrap
1030	RESERVED
1031	RESERVED
1032	U.S. Patent Publication No. 2004/0044516 (“ <u>Kennewick</u> ”)

TABLE OF ABBREVIATIONS AND CONVENTIONS

Abbreviation	Meaning
957 Patent	Ex.1001: U.S. Patent No. 9,495,957
IPR	<i>inter partes</i> review
Petitioner	Bank of America, N.A. (“BofA”)
Patent Owner or PO	Dialect, LLC (“Dialect”)
Second Amended Complaint	Compl.: Ex.1007
<i>xx.yy–zz</i>	column <i>xx</i> , lines <i>yy</i> to <i>zz</i>

1. I, Henry Houh, declare as follows:

I. INTRODUCTION

2. I am making this declaration in support of Petitioner’s request for an *Inter Partes* Review of U.S. Patent No. 9,495,957 (“the 957 Patent”) to Weider.

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

4. I have been asked to provide my opinions regarding whether the subject matter of claims 7 and 8 (“the Challenged Claims”) of the 957 Patent are anticipated or otherwise would have been obvious to a person having ordinary skill in the art (“POSITA”) at the time of the alleged invention, in light of the prior art. It is my opinion that the Challenged Claims would have been obvious to a POSITA.

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

- the 957 Patent, Ex.1001
- the prosecution history of the 957 Patent (“957 File History”), Ex.1002
- U.S. Patent Pub. No. 2002/0135618 (“Maes”), Ex.1005

- International Patent Application Publication No.
WO 00/20962 (“Coffman”), Ex.1012
- Second Amended Complaint, Ex.1007
- the additional references I refer to below and/or
identified in the exhibit list.

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

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

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

in this Declaration are my own, were formed by me and were documented with the assistance of counsel in the manner described. As detailed herein, it is my opinion that each of the challenged claims is rendered obvious by prior art references that predate the earliest priority date of the 957 Patent. If requested, I am prepared to testify about my opinions expressed herein.

II. QUALIFICATIONS AND PROFESSIONAL EXPERIENCE

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

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

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

12. I am currently self-employed as an independent technical consultant. Until 2022, I was also president of a company that provides supplemental science,

technology, engineering, and mathematics (“STEM”) education to children of all ages.

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

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

15. I also helped to build the web pages that initiated the above- mentioned video sessions via a web interface. Vice President Al Gore visited our group in 1996 and received a demonstration of—and remotely drove—a radio- controlled toy car with a wireless video camera mounted on it that was built by our group. This toy car device received commands transmitted over a network from a remote computer, and video data from the toy car was transmitted wirelessly then over a computer network back to the user controller. On occasion, we allowed users visiting our web site to drive the toy car from their remote computer while they watched the video on their computer. The video stream was encoded by TNS-designed hardware, streamed over the TNS-designed network, and displayed using TNS-designed software.

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

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

18. As part of my work at NBX, I designed the core audio reconstruction algorithms for the telephones, as well as the packet transmission algorithms. I also designed and validated the core packet transport protocol used by the phone system. The protocol was used for all signaling in the phone system, including for the setup of conference calls.

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

20. From 1999-2004, I was employed by Empirix or its predecessor company, Teradyne. Empirix was a leader in test tools for telecommunications protocols and systems, providing functional testing tools as well as load testing

tools. From 2000-2001, I conceived and built a test platform for testing Voice- over-IP (VoIP). The first application on this new test platform was a cloud emulator for simulating the effects of transmitting VoIP over a busy network.

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

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

- U.S. Patent No. 7,975,296, "Automated security threat testing of web pages";
- U.S. Patent No. 7,877,736, "Computer language interpretation and optimization for server testing";
- U.S. Patent No. 7,801,910, "Method and apparatus for timed tagging of media content";

- U.S. Patent 7,590,542, “Method of generating test scripts using a voice-capable markup language”;
- U.S. Patent No. 6,967,963, “Telecommunication method for ensuring on-time delivery of packets containing time-sensitive data”;
- U.S. Patent Application Publication No. 20070106685, “Method and apparatus for updating speech recognition databases and reindexing audio and video content using the same”;
- U.S. Patent Application Publication No. 20070106693, “Methods and apparatus for providing virtual media channels based on media search”;
- U.S. Patent Application Publication No. 20070106760, “Methods and apparatus for dynamic presentation of advertising, factual, and informational content using enhanced metadata in search-driven media applications”;
- U.S. Patent Application Publication No. 20070112837, “Method and apparatus for timed tagging of media content”;
- U.S. Patent Application Publication No. 20070118873, “Methods and apparatus for merging media content”; and
- U.S. Patent Application Publication No. 20090222442, “User-directed navigation of multimedia search results.”

III. LEVEL OF ORDINARY SKILL IN THE ART

23. My opinions are provided based on what a person of ordinary skill in the art (“POSITA”) in the technical field of the invention would have understood at the time of the purported invention of the 957 Patent. I have been asked to assume the date of the purported invention for the 957 patent is August 29, 2005.

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

25. Based on the materials and information I have reviewed, and on my experience in the technical areas relevant to the Asserted Patents, a POSITA with respect to the Asserted Patents as of the time of the invention, i.e., the earliest possible priority date of the 957 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 957 Patent. Lack of professional experience can be remedied by additional education, and vice versa.

26. For purposes of this Declaration, in general, and unless otherwise noted, my statements and opinions, such as those regarding my own experience and what a POSITA would have understood or known generally (and specifically related to the references I consulted herein), reflect the knowledge that existed in the relevant field as of the priority date of the 957 Patent.

IV. RELEVANT LEGAL STANDARDS

27. I am not an attorney. In preparing and expressing my opinions and considering the subject matter of the 957 Patent, I am relying on certain basic legal principles that counsel has explained to me.

28. I understand that prior art to the 957 Patent includes patents and printed publications in the relevant art that predate the priority date of the 957 Patent. For purposes of this Declaration, I am applying August 29, 2005 as the priority date of the 957 Patent, but I take no position on that issue.

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

30. A patent claim is anticipated if each limitation of the patent claim is found either expressly or inherently in a single item of prior art. While a prior art reference need not use the same words as a patent claim to anticipate the claim, the prior art reference must describe the requirements of the claim with sufficient clarity

such that POSITA would have been able to make and use the claimed invention based on the reference and his or her knowledge in the applicable technical field.

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

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

instead, the prior art must teach the claimed subject matter in the arrangement required by the claims.

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

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

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

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

37. I have been informed and understand that obviousness is based on the scope and content of the prior art, the differences between the prior art and the claim, the level of ordinary skill in the art and secondary indicia of obviousness and nonobviousness to the extent such indicia exist. I have been advised that the following evidence, when present, must be considered before forming a conclusion that a claimed invention is obvious: (1) the invention's commercial success, (2) long felt but unresolved needs, (3) the failure of others, (4) skepticism by experts, (5) praise by others, (6) teaching away by others, (7) recognition of a problem, and (8) copying of the invention by competitors. In this instance, however, I am not aware of any evidence of any of these factors. I will of course consider any evidence on these issues which is presented to me.

38. I have been informed and understand the determination of whether the asserted claims would have been obvious to a person of ordinary skill in the art and, therefore, invalid, is not governed by any rigid test or formula. A determination that a claim is obvious is, instead, based on a common-sense determination that the claimed invention is merely a combination of known limitations to achieve predictable results. Any of the following rationales are acceptable justifications to conclude that a claim would have been obvious: (1) the claimed invention is simply a combination of known prior art methods to yield predictable results; (2) the claimed invention is a simple substitution of one known element for another to

obtain predictable results; (3) the claimed invention uses known techniques to improve similar devices (methods, or products) in the same way; (4) the claimed invention applies a known technique to a known device (method, or product) ready for improvement to yield predictable results; (5) the claimed invention was “obvious to try”—choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success; (6) there is known work in one field of endeavor that may prompt variations of it for use in either the same field or a different one based on design incentives or other market forces if the variations would have been predictable to one of ordinary skill in the art; or, (7) there is some teaching, suggestion, or motivation in the prior art that would have led one of ordinary skill in the art to modify the prior art reference to combine prior art teachings to arrive at the claimed inventions.

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

was an apparent reason to combine any known elements in the fashion claimed by the patent at issue.

V. OVERVIEW OF THE 957 PATENT

40. The 957 Patent was based on U.S. Patent Appl. No. 14/467,641 filed on August 25, 2014, which is a continuation of U.S. Patent Appl. No. 13/898,045 filed on May 20, 2013 (now U.S. Patent No. 8,849,652), which is a continuation of U.S. Patent Appl. No. 13/488,299 filed on June 4, 2012 (now U.S. Patent No. 8,447,607), which is a continuation of U.S. Patent Appl. No. 13/084,197 filed on April 11, 2011 (now U.S. Patent No. 8,195,468), which claims priority to U.S. Patent Appl. No. 11/212,693, filed on August 29, 2005 (now U.S. Patent No. 7,949,529). Ex.1001, Cover.

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

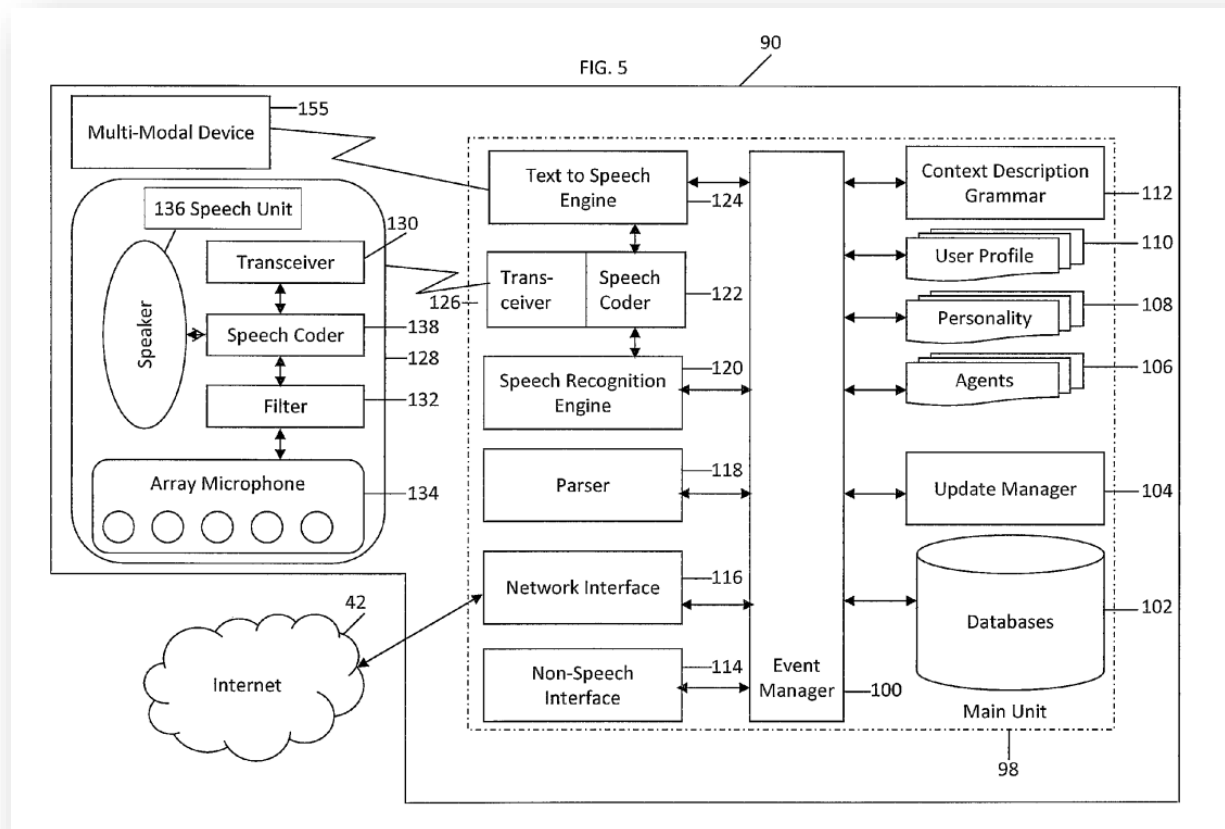
A. The Disclosures of the 957 Patent

42. The 957 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, 1:24–30. The 957 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.

43. The 957 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, 2:2–6. The 957 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, 2:10–12. The 957 Patent purportedly overcomes the “drawbacks” of prior art systems. Ex.1001, 2:61–62.

44. Figure 5 of the 957 Patent is characterized as “an overall diagrammatic view of the interactive natural language processing system according to an embodiment of the invention.” Ex.1001, 15:65–67.



45. 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, 26:27–31.

46. After initial processing, the speech is passed to speech recognition engine 120 for processing using the context description grammar module 112. Ex.1001, 26:64–66. 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, 26:67–27:2. “The knowledge agents may then process the commands or questions” and “may return results” to the user. Ex.1001, 27:2–17. Moreover, “multimodal support may be provided to maintain the context during both voice interaction and interaction through the non-speech interface 114,” Ex.1001, 27:38–41, and “multi-modal devices 155 ... may be used as a substitute or complement to the speech interface.” Ex.1001, 27:33–34.

VI. CLAIM CONSTRUCTION

47. 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

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

49. 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.

50. 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

51. 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 957 Patent.

52. 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.

53. 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.

54. Moreover, Maes expressly incorporates by reference Coffman in its entirety. 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 or Ex.1012] 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. I therefore understand that Coffman is considered effectively part of Maes.

55. As the entire disclosure of Coffman is directed to “Conversational Computing Via Conversational Virtual Machine,” I understand that Maes is 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 I rely upon below.

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

56. To the extent one asserts that Coffman discloses a different embodiment from that of Maes, a POSITA would have been motivated to combine Coffman’s teachings related to searching a context stack to find entries matching transcribed user input and merging/combining speech and non-speech transcriptions (Ex.1012, 41:2–5, 42:1–3, and 38:1–19) with Maes’s teachings that receives and processes multi-modal inputs (e.g., by producing decoded text or script) as described below.

57. As explained above, Maes expressly incorporates by reference Coffman in its entirety, names Stephane Maes as a common inventor, and has IBM as the original assignee. Ex.1005, Abstract, 9:30–37 and 19:43–49 and Ex.1012, Cover, Abstract. Accordingly, a POSITA would have been motivated to combine the teachings of Coffman with those of Maes for at least these reasons.

58. Combining Coffman's teachings of organizing/updating its context stack such that the most-recent context is at the top of the context (Ex.1012, 21, 23–28, 41:1–5, 40:15–19, and 42:7–9) with the teachings of Maes would benefit Maes's system by providing efficient identification and retrieval of the most-recent context, allowing for faster executions for matching/decoding of keywords and improving the responsiveness of Maes's system aligning with its goal of faster executions (Ex.1005, 11:64–12:30, 5:29–34). A POSITA would have understood that by keeping the context stack updated, e.g., with the most-recent context associated with a user's dialog, would allow an application to use the most-recent context, which is likely to be relevant to the user's utterance, in determining the user's intent. Ex.1005, 36:64–37:3.

59. A POSITA would be motivated to combine Maes and Coffman because they teach strikingly similar techniques. For example, Maes's technique of associating recognized events against data in a context stack (Ex.1005, 7:60–8:4 and Figures 1 and 8), and Coffman's technique of searching a context stack to find entries matching transcribed user input (Ex.1012, 42:1-9). Notably, Maes's context stack disclosures are almost verbatim as Coffman's. Ex.1012, 21, 23–28 and Ex.1005, 37:51–62; *compare also* Ex.1005, Figure 8B and Ex.1012, Figure 2.

60. Additionally, a POSITA would have understood that combining Maes's teachings and Coffman involves the application of known techniques (e.g.,

identifying contexts by performing comparisons of spoken utterances against data stored in a context stack) to improve a similar system (e.g., Maes's system) in the same way.

61. Maes and Coffman are analogous art as the 957 Patent for at least three reasons. *First*, both references are directed toward the same field of endeavor as the 957 Patent—computer-implemented systems interpreting user utterances or otherwise natural language processing systems. Ex.1001, Abstract, 1:51–61, 3:25–45; Ex.1005, Abstract, 4:7–17; Ex.1012, 1–4 and 29:1–4. *Second*, Maes and Coffman are reasonably pertinent to a problem described by the 957 Patent: an environment for **reliably processing** a user's language queries, which is a problem of conventional systems identified in the 160 Patent. Ex.1001, 1:35–44; Ex.1005, Abstract, 4:7–17 and 6:39–55; Ex.1012, 4:4–9. *Third*, similar to the 957 Patent, Maes and Coffman teach the use of **grammars** and **vocabularies** in recognizing speech inputs, which indicate compatibility of their teachings and therefore a POSITA would have had a reasonable expectation of success combining them. Ex.1001, 13:35–38; Ex.1005, 31:15–18, 34:61–67, 33:11–21, 39:59, 41:13–20, and 41:35–38; Ex.1012, 29:21–32, 25:6–9.

62. A POSITA would be motivated to combine Maes and Coffman because to do so would have been the arrangement of old elements (a speech transcription and recognition system comprising modules for speech transcription and

recognition, context stack storing current and historical data related to user interactions) with each performing the same function it has been known to perform (e.g., recognizing spoken utterances, converting spoken utterances into computer-readable format based on decoding the utterances, searching a context stack to find matches between data stored on the context stack and transcribed user input) and yielding no more than what one would expect from such an arrangement (an improved computer-implemented system interpreting user utterances), as Maes seeks to implement. Ex.1005, 11:62–63, 2:54–67.

63. Furthermore, a POSITA would have been motivated to combine Maes and Coffman because the references make clear that their systems require no specialized hardware or software. In fact, these references teach using conventional, commercially-available systems. *See* Ex.1005, 2:38–53 and 45:50–60, 12:43–49 (“processing ... may be accomplished via any **conventional acoustic information recognition system**”) (emphasis added⁸); Ex.1012, Abstract (“The conversational computing system may be built on top of a **conventional operating system and API’s ... and conventional device hardware.**”).

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

⁸ Emphasis added throughout unless otherwise specified.

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 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 enhance the functionality of Maes's system and further, as explained below, such modifications would have been simple implementations predictably resulting in a recognition system with improved accuracy and without undue experimentation, as Maes seeks to implement (Ex.1005, 11:62–63).

IX. CLAIMS 7 AND 8 ARE UNPATENTABLE

A. Ground 1: Maes renders obvious claims 7 and 8.

65. Claims 7 and 8 of the 957 Patent are rendered obvious over Maes.

1. Independent Claim 7

a. [7.0] A computer-implemented method of processing a natural language utterance, the method comprising:

66. Maes teaches or suggests “[a] computer implemented method of processing a natural language utterance.”

67. Maes describes “[s]ystems and methods ... provided for performing focus detection, referential ambiguity resolution and mood classification in accordance with multi-modal input data [in the form of audio input data, video

input data, as well as other types of input data].” Ex.1005, Abstract (emphasis added⁹), 4:10–11, 1:9–23; *see also* 1:9–23.

68. Further, Maes’s systems and methods are directed at “*processing a natural language utterance.*” While Maes discusses processing multi-modal inputs, the multi-modal input may comprise, at least in part, “audio input data” (e.g., “*a natural language utterance*”). Referencing Figure 1 (below), Maes discloses:

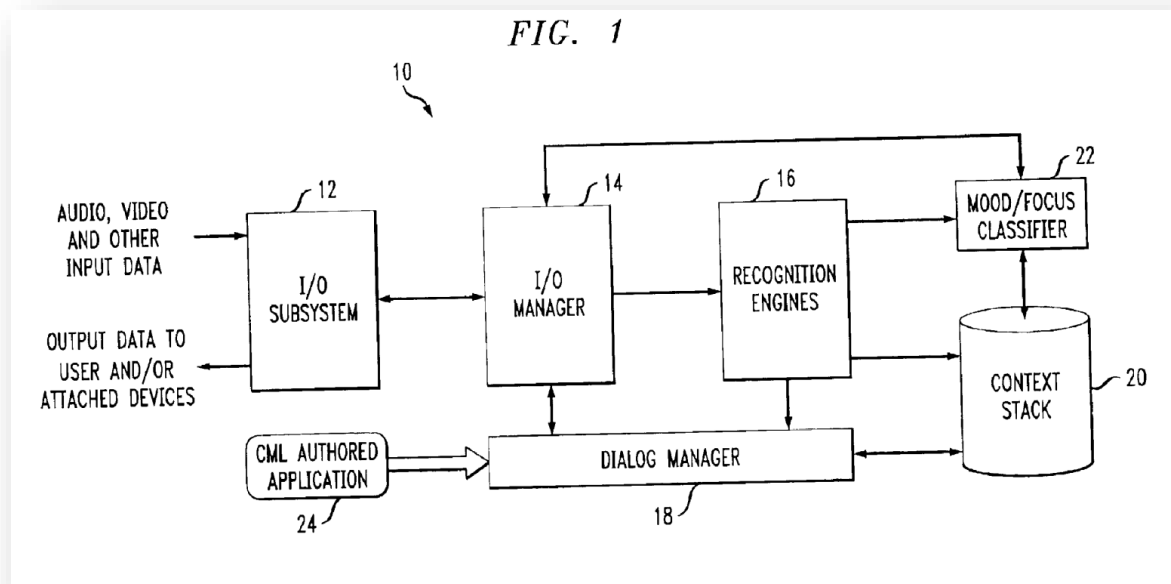
[A] multi-modal conversational computing system ...
compris[ing] an input/output (I/O) subsystem 12, an I/O
manager module 14, one or more recognition engines 16,
a dialog manager module 18, a context stack 20 and a
mood/focus classifier 22.

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 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

⁹ Emphasis added throughout unless otherwise specified.

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).

See generally Ex.1005, 3:66–4:21.



Ex.1005, Figure 1.

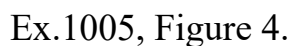
69. With reference to Figures 1 and 4, Maes describes Figure 4 (reproduced below) as “an audio-visual speech recognition module¹⁰ ... as one of the recognition

¹⁰ As used herein, the terms “Maes’s system” and “audio-visual speech recognition module” are used interchangeably.

modules¹¹ of FIG. 1 **to perform speech recognition using multi-modal input data received** in accordance with the invention.” Ex.1005, 10:50–52. Maes notes that the “microphone 406” of Figure 4 is “part of the I/O subsystem 12” (Ex.1005, 11:24–26) “for capturing audio input data from the environment in which the system is deployed.” Ex.1005, 6:5–8; *see also* 6:15–21.

70. Figure 4 shows that the audio-visual speech recognition module receives multi-modal input data—notably, including an audio signal (e.g., speech provided by the speaker and/or background noise) (“*a natural language utterance*”) and a video signal (e.g., the speaker’s face including lip movement and/or background objects in the environment) and processes the audio signal and the video signal 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 of processing a natural language utterance.*”

¹¹ 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.1003
Microsoft Corporation / Page 35

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.” Ex.1005, 45:45–60.

72. Maes describes examples in which its methods are applicable to a broadcast news system which uses spoken utterance¹² inputs which further confirms that its method applies to “*processing a natural language utterance.*” Maes’s system provides, among others, functions of “natural language understanding (NLU), natural language generation (NLG)” and “NL parsing.” Ex.1005, 41:14–16, 35:58, 10:63–11:15). 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. *See* Gibbon defining “*natural language*” as “a language used by humans for communication.” Ex.1017, 444.

73. As another example, Maes states that “the multi-modal conversational computing system 10 may be employed within a vehicle” in which a user may say

¹² As used herein, it will be understood that a “spoken utterance” received using Maes’s system, would encompass both an audio portion and a video (*i.e.*, non-speech) portion (e.g., the accompanying lip movement) of the utterance. Ex.1005, 6:43–47.

the “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, 4:46–53, 5:16, 6:43–50, 7:63–8:29.

74. Maes describes an example of how the disclosed system works. For example, Maes discloses 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 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, 6:35–38. “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.

75. Therefore, if limiting, Maes discloses the preamble.

b. [7.1] generating a context stack comprising context information that corresponds to a plurality of prior utterances, wherein the context stack includes a plurality of context entries;

76. Maes teaches or suggests this claim element.

77. First, Maes describes “the context stack 817 may be implemented ...” (“*generating a context stack*”) as part of the context stack 20 of Figure 1. Ex.1005, 37:53–55. Referencing Figure 1 (below), Maes describes “the multi-modal conversational computing system 10 comprises ... a context stack 20.” Ex.1005, 3:66–4:6.

78. Maes further describes its “*context stack*” storing “historical information” e.g., past input/output (I/O) events” or “*a plurality of context entries*” generated previously in the context of a dialog (e.g., “*context information that corresponds to a plurality of prior utterances*”).

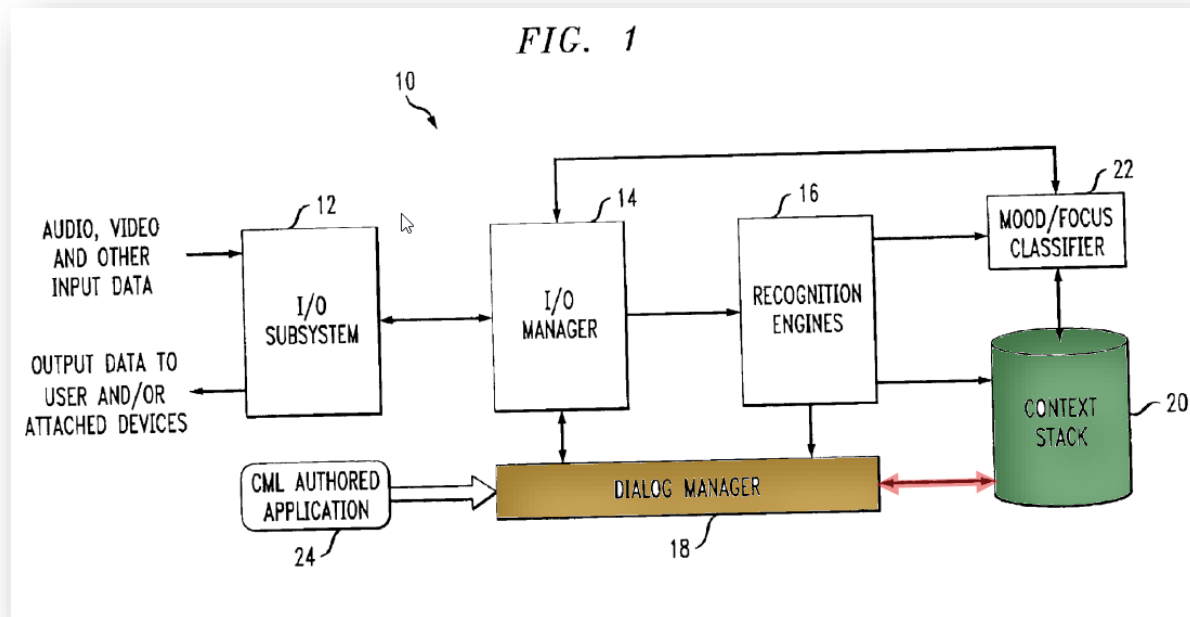
[H]istorical information (e.g., past events) stored in the context stack [such that] the context stack ... is associated with the organized/sorted context corresponding to each active dialog.

Ex.1005, 7:62–63, 37:60–61. Indeed,

79. Maes's context stack "compris[es] context information" and "includes a plurality of context entries" because Maes teaches

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; Figure 1 (reproduced below); *see also* Ex.1005, 7:40–45 (disclosing "the system ... attempts to determine the **user intent based on the current event** and the **historical interaction information stored in the context stack.**"); 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, Fig. 1 (annotated).

80. Referencing the in-vehicle example in which a user may say “turn it on,” 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 [such as dialog manager 18 in Figure 1] 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,’** [*“context information that corresponds to a plurality of prior utterances”*] or portions thereof [stored on the context stack], with which such an utterance and gesture could be associated.

Ex.1005, 7:65–8:4.

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

Ex.1005, 8:37–42. In the above example, “this event” corresponds to the current event of the user pointing to the radio, which occurred after “the previous spoken utterance event” of the user saying “turn it on. Data for both events (“*context information*”) are stored on the context stack because Maes discloses “**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**” are stored on the context stack. Ex.1005, 37:55–62. Maes clarifies that “the dialog” includes “**conversational dialog comprising speech** and other multi-modal I/O such as GUI keyboard, pointer, mouse, as well as video input, etc.” Ex.1005, 36:49–54.

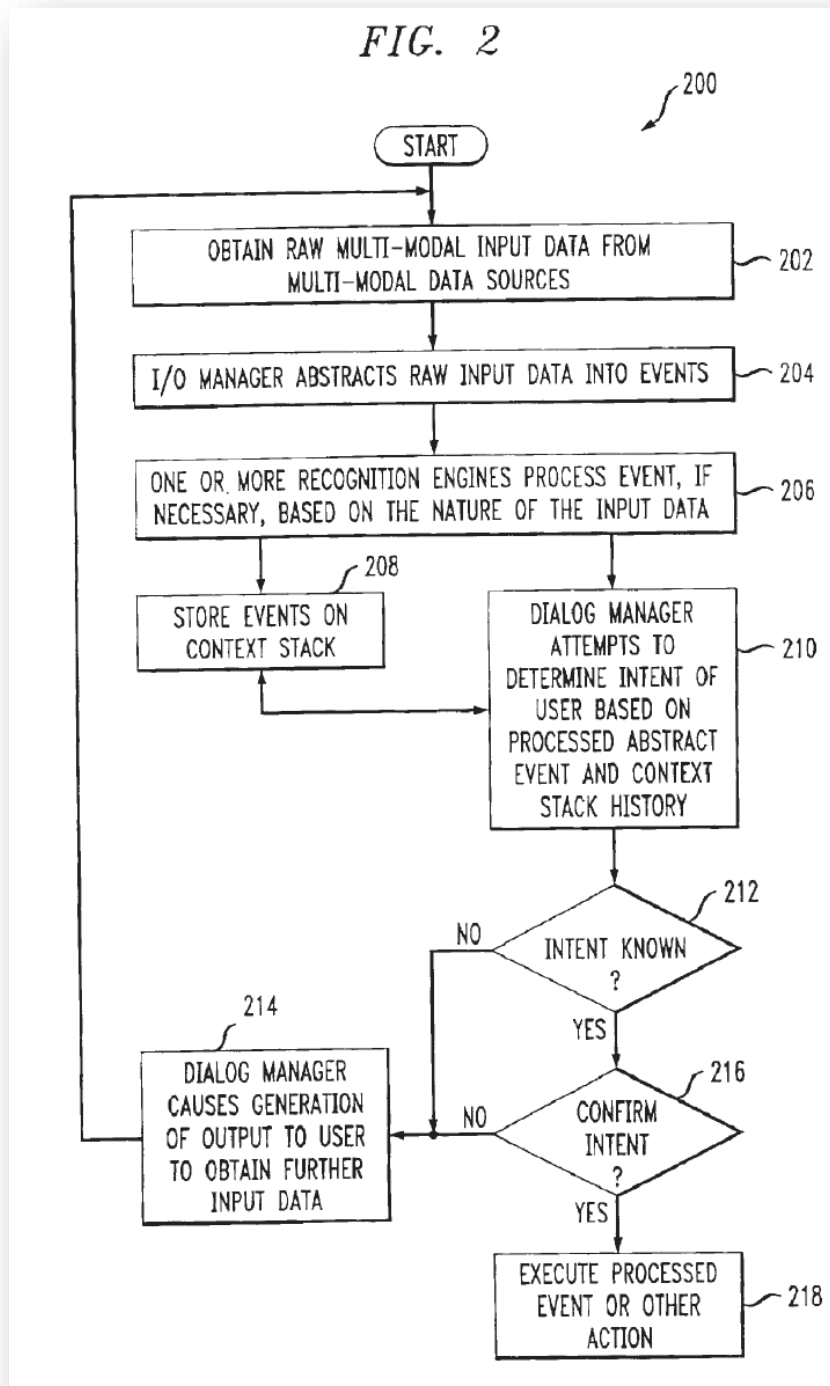
81. Therefore, Maes teaches or suggests this claim element at least for this reason.

Additional Mapping

82. Moreover, for the additional reasons below, Maes teaches or suggests this claim element. For example, Figure 2 (below) described as “a flow diagram illustrating a referential ambiguity resolution methodology performed by a multi-

modal conversational computing system” expressly discloses the use of context stack for storing event-related data and used in determining the user’s intent.

Ex.1005, 3:24–28.



Ex.1005, Figure 2.

83. Referencing Figure 2, Maes explains that in step 208, “**stor[ing] (“generating”) [recognized events, as well as the events that do not need to be recognized] in a storage unit referred to as the context stack 20 ... used to create a history of interaction between the user and the system** [*“a context stack comprising context information that corresponds to a plurality of prior utterances, wherein the context stack includes a plurality of context entries”*] 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 and the historical interaction information stored in the context stack.” Ex.1005, 7:33–45.

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

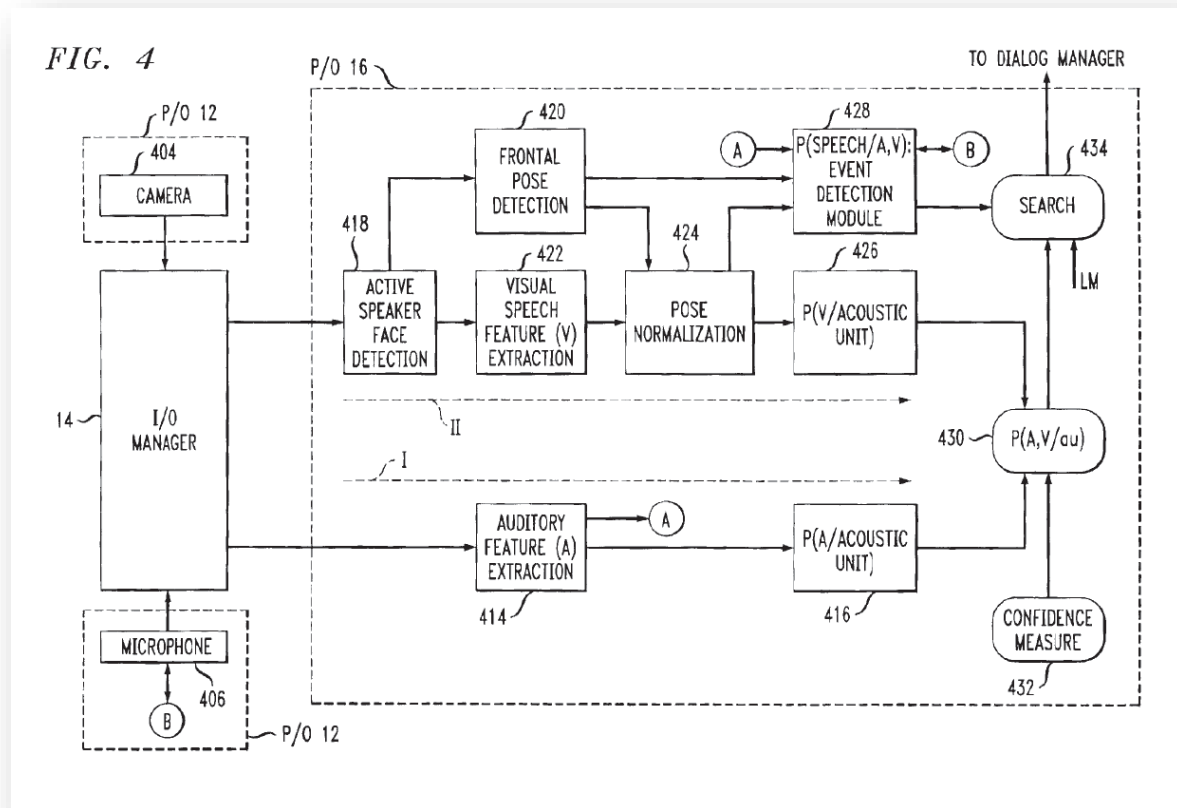
- c. **[7.2] receiving, by one or more computer processors, the natural language utterance, wherein the natural language utterance is associated with a command or is associated with a request;**

85. Maes teaches or suggests this claim element.

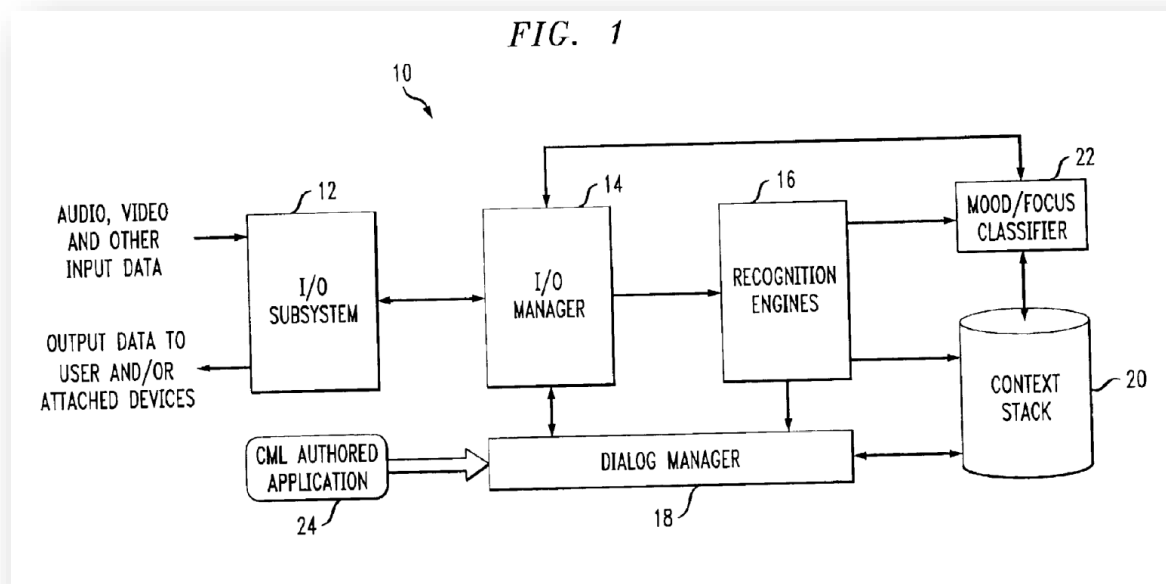
- i. **Receiving the natural language utterance**

86. First, Maes discloses “*receiving ... the natural language utterance.*” 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 audio and visual data—from a speaker (e.g., comprising “*the natural language utterance*”) at an “input/output (I/O) subsystem ... compris[ing] one or more microphones for capturing audio input data,” which “one of ordinary skill in the art will realize that other user interfaces and devices ... may be included **for capturing user activity**.” Ex.1005, 6:21–29,5–8. Maes notes that the “microphone 406” of Figure 4 is “part of the I/O subsystem 12” depicted in Figure 1 (Ex.1005, 11:24–26). Maes also teaches its system “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.” Ex.1005, 5:60–64. Thus, Maes provides functionality for capturing audio signal, video signal, or both, as part of receiving input data associated with the natural language utterance.



Ex.1005, Figure 4.



Ex.1005, Figure 1.

ii. one or more computer processors

87. Further, Maes discloses “*one or more processors*” “*receiving [the] natural language utterance.*” Maes’s “**system ... comprises at least one processor, operatively coupled to the user interface subsystem, and configured to receive ... 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.

88. 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.

iii. The natural language utterance is associated with a command/request

89. Maes discloses “*wherein the natural language utterance is associated with a command or is associated with a request.*” For 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’” (which Maes alternatively refers as “the ‘turn it on’ command”) (“*the natural language utterance ... associated with a command or ... associated with a request*”). Ex.1005, 8:30–33. 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, *see generally* 7:63–8:29.

90. Specifically, Maes explains that “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 natural language 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.

91. Maes also 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. Thus, Maes’s “system independently recognizes voice-based **commands** [“*command/s*”] and independently recognizes gesture-based commands.” Ex. 1005, 2:1–4; *see also* Ex.1005, 4:58–59 (disclosing “user input commands received and processed by the system”).

92. Therefore, Maes discloses this claim element.

d. [7.3] determining one or more words of the natural language utterance by performing speech recognition on the natural language utterance;

93. Maes teaches or suggests “*determining one or more words of the natural language utterance by performing speech recognition on the natural language utterance.*”

i. Performing Speech Recognition

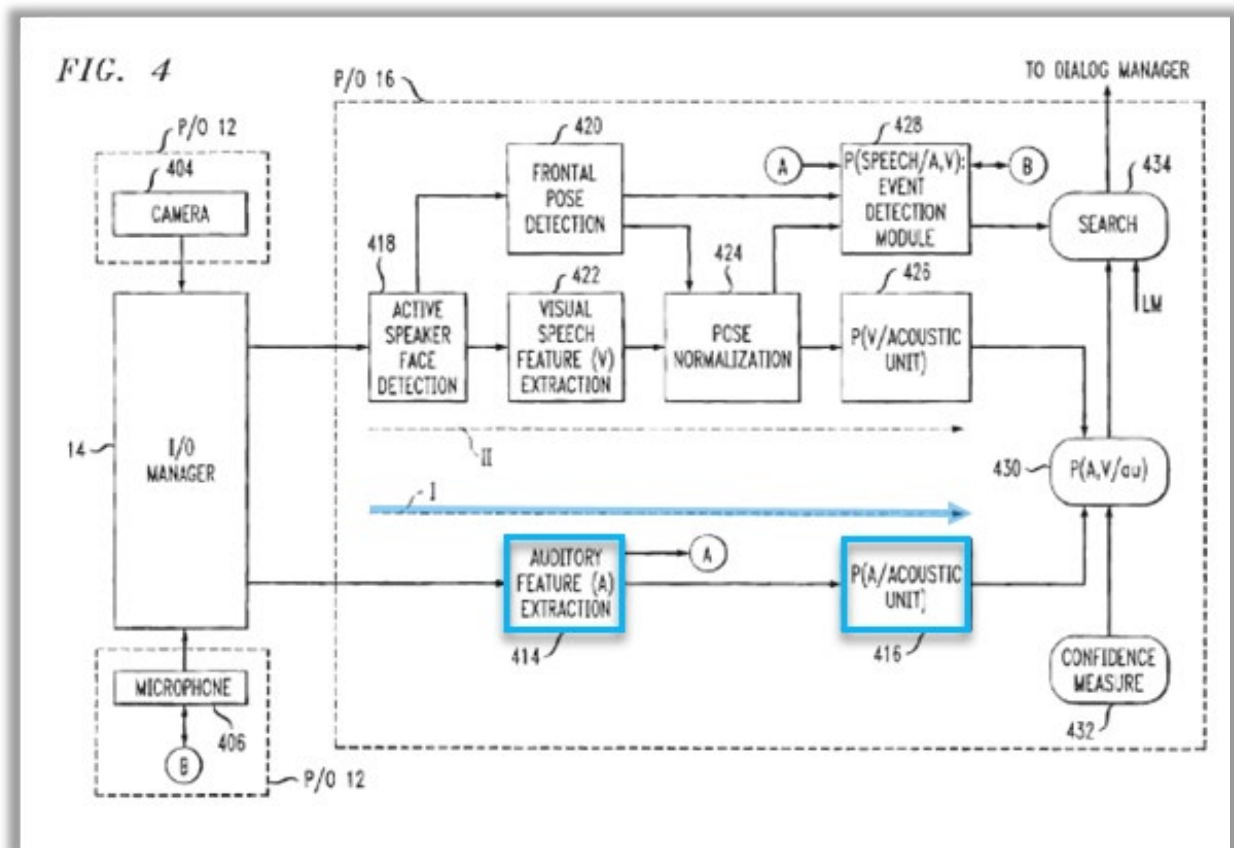
94. Maes's system **“performs various recognition tasks (e.g., speech recognition ... in accordance with the recognition engines 16).”** Ex.1005, 4:7–14; see also 34:61–35:1 (one or more recognition engines of FIG. 1). “[D]epending on the nature of the [I/O (input/output)] event, one or more recognition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

95. Maes provides additional details of *“performing speech recognition.”* Maes begins by describing Figure 4 as “an audio-visual speech recognition module that may be employed as **one of the recognition modules¹³ of FIG. . 1 to perform speech recognition using multi-modal input data received** in accordance with the invention.” Ex.1005, 10:50–52. Specifically, Figure 4 shows that the audio-visual speech recognition module receives multi-modal input data including an audio signal (e.g., speech provided by the speaker and/or background noise) and a video signal (e.g., the speaker's face including lip movement and/or background objects in the environment) and then processes the multi-modal input data by passing through

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

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).

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



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

97. Referencing Figures 1 and 9A (reproduced below), Maes provides details of “performing speech recognition” explaining that “via the I/O manager 14 of FIG. 1” “user-provided input data events are ... provided to” “apparatus 900”—an “apparatus for collecting data associated with a voice of a user,” including “a dialog management unit 902 for conduct[ing] a conversation with the user,” that provides functionality including natural language understanding (NLU), natural language generation (NLG), finite state grammar (FSG), and/or text-

to-speech Syntheses (TTS) for machine-prompting the user....” Ex.1005, 39:22–29, 41:13–20, and Figure 9A.

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

ii. Determining one or more words of the natural language utterance

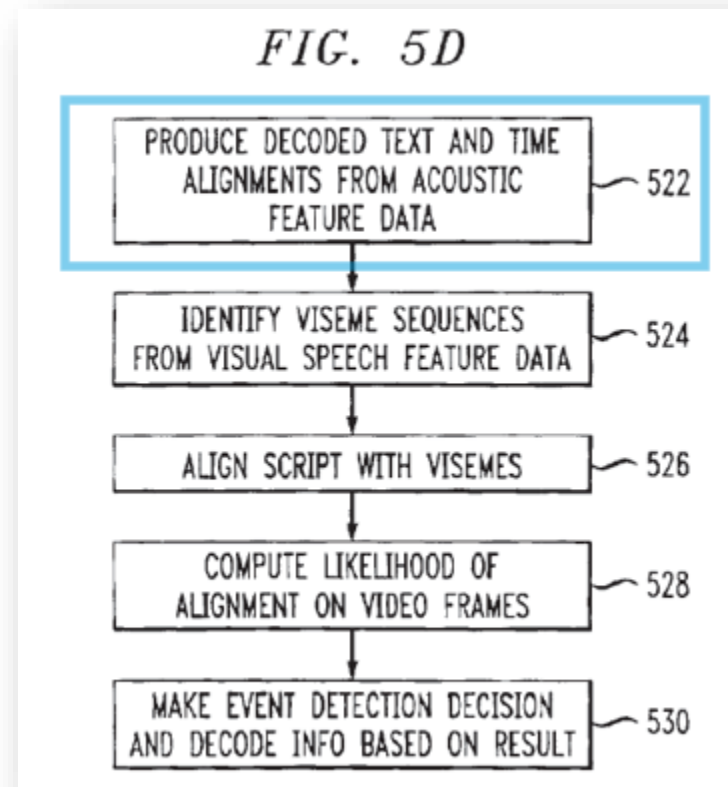
99. Referencing Figure 4, Maes explains “**produc[ing]**” (“*determining*”), for instance, “**decoded text**” (“*one or more words*”) “*by*” “the processing performed in blocks 414 and 416”—(e.g., “*performing speech recognition*”) on “the uttered speech” (“*the natural language utterance*”). Ex.1005, 21:43–47. And, referencing Figure 9, Maes explains that post processor 938 includes a semantic module “to interpret meanings of phrases” and “could be used by speech recognizer 926 indicate that some **decoding candidates** in a list are meaningless and

should be discarded/replaced with meaningful candidates” (Ex.1005, 41:58–64). Thus, Maes explains that the “**decoded text**” produced by the processing performed in blocks 414 and 416 is the same “decoding candidates” Maes describes in connection with Figure 4, further confirming Figure 9A and its corresponding disclosure provide additional details to, among other things, the “*performing speech recognition*” step described in Figure 4. For example, in Figure 9A, Maes identifies “semantic module,” “speech recognizer 926,” “post processor 938,” “speech recognition module 928” and other components/functionalities. *See e.g.*, Ex.1005, 41:48–59 (**post processor 938 ... transcribe[s] user utterances and ... perform[s] keyword spotting thereon**”).

100. Therefore, Maes discloses the “*determining one or more words of the natural language utterance by performing speech recognition on the natural language utterance.*”

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

the acoustic feature extractor 414” (“*by performing speech recognition*”) “on” “the uttered speech” (“*the natural language utterance*”). *Id.*



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

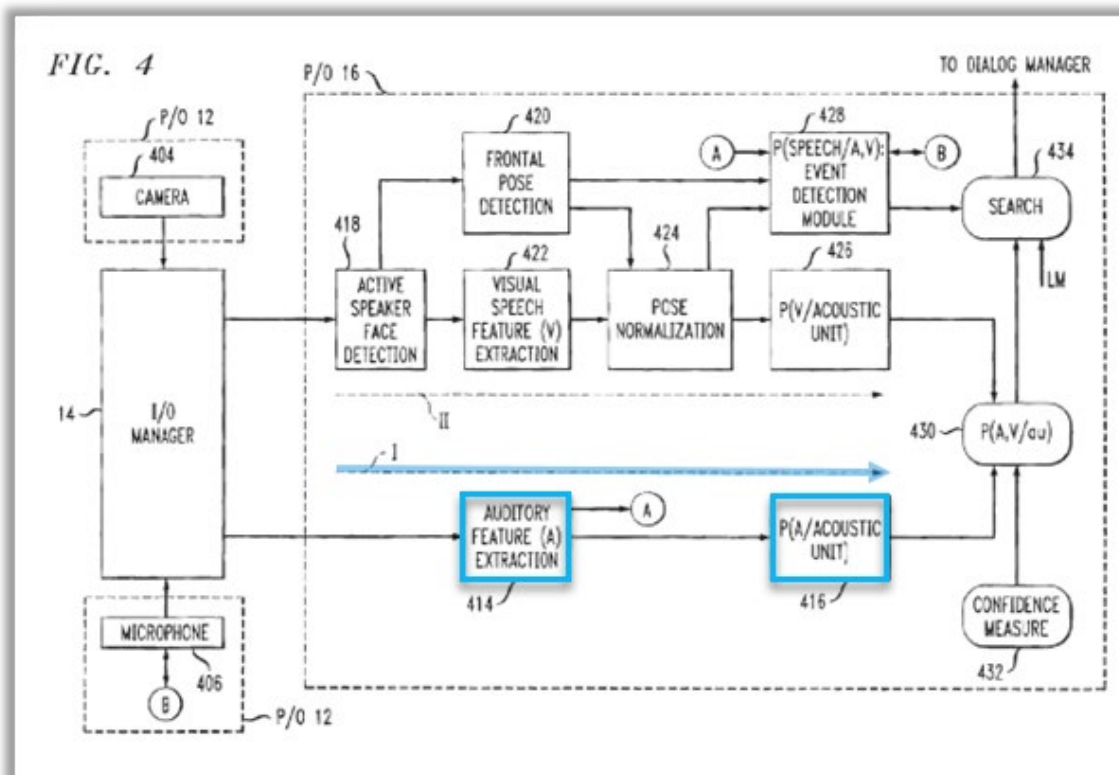
path so that the buffered data may be recognized, in step 520, so as to yield a **decoded output.**”); *see also* 20:23-33.

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

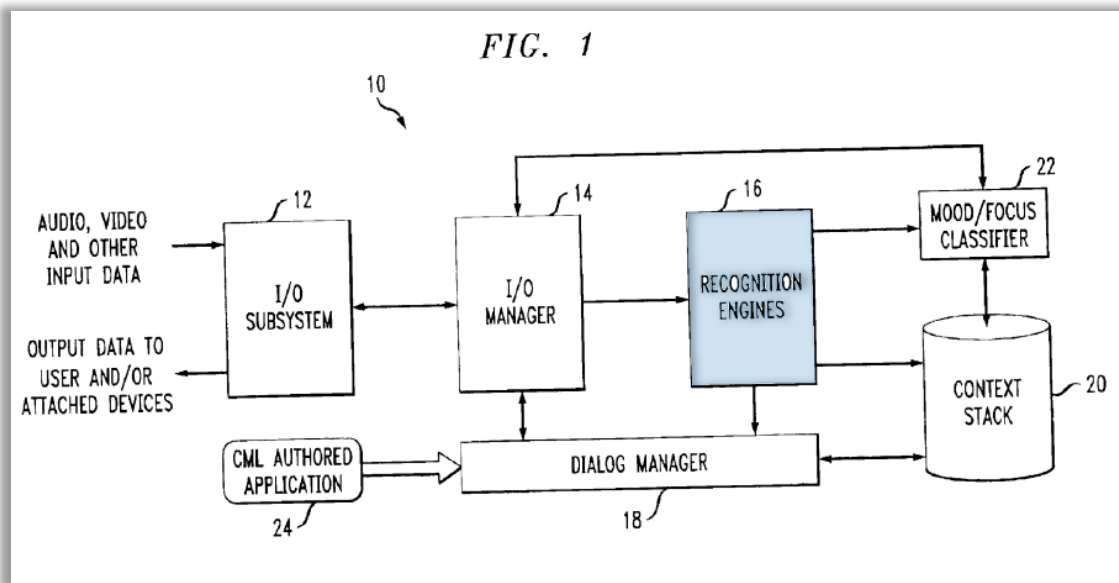
103. Specifically, Figure 4 shows that the audio-visual speech recognition module receives multi-modal input data including an audio signal (e.g., speech provided by the speaker and/or background noise) and a video signal (e.g., the speaker’s face including lip movement and/or background objects in the 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 recognition module shown in Figure 4, the acoustic

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

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).

104. 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.

105. 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.

106. 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.

107. Referring to the example in which Maes’s multi-modal conversational computing system is employed within a vehicle, Maes explains the “*determining ...*” step as follows. The raw multi-modal input data relating to an I/O event (e.g., representative of the user’s spoken utterance and lip movement accompanying the

spoken utterance) received by the multi-modal conversational computing system is abstracted into a recognizable form. For example, “[t]he I/O manager [of the multi-modal conversational computing system] receives the raw multi-modal data and **abstracts the data into a form that represents ... a spoken utterance ...** As is known, a **data abstraction operation may involve generalizing details associated with all or portions of the input data so as to yield a more generalized representation of the data for use in further operations.**” Ex.1005, 6:21–38. Thus, as a result of processing by several modules, the system produces decoded text/script corresponding to the user’s spoken utterance, for example, of saying “turn it on.”

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

- e. **[7.4.0] identifying, from among the plurality of context entries, one or more context entries that correspond to the one or more words, wherein the context information includes the one or more context entries,**

109. Maes teaches or suggests “*identifying, from among the plurality of context entries, one or more context entries that correspond to the one or more words, wherein the context information includes the one or more context entries.*”

- i. **context information includes context entries**

110. First, Maes discloses “*wherein the context information includes the one or more context entries*” of this claim element because, as explained in connection with element [7.1], Maes discloses “[*a*] *context stack comprising context information ... and includ[ing] a plurality of context entries.*”

ii. identifying context entries corresponding to the words

111. Furthermore, as explained in connection with element [7.3], Maes discloses the results of the recognized events producing “decoded text” (or, script) (“*one or more words*”) processed from the spoken utterance (“*the natural language utterance*”).

112. Further, as explained below, Maes teaches or suggests “*identifying, from among the plurality of context entries, one or more context entries that correspond to the one or more words.*”

113. 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:

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

stack], with which such an utterance and gesture could be associated.

Ex.1005, 7:65–8:4.

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

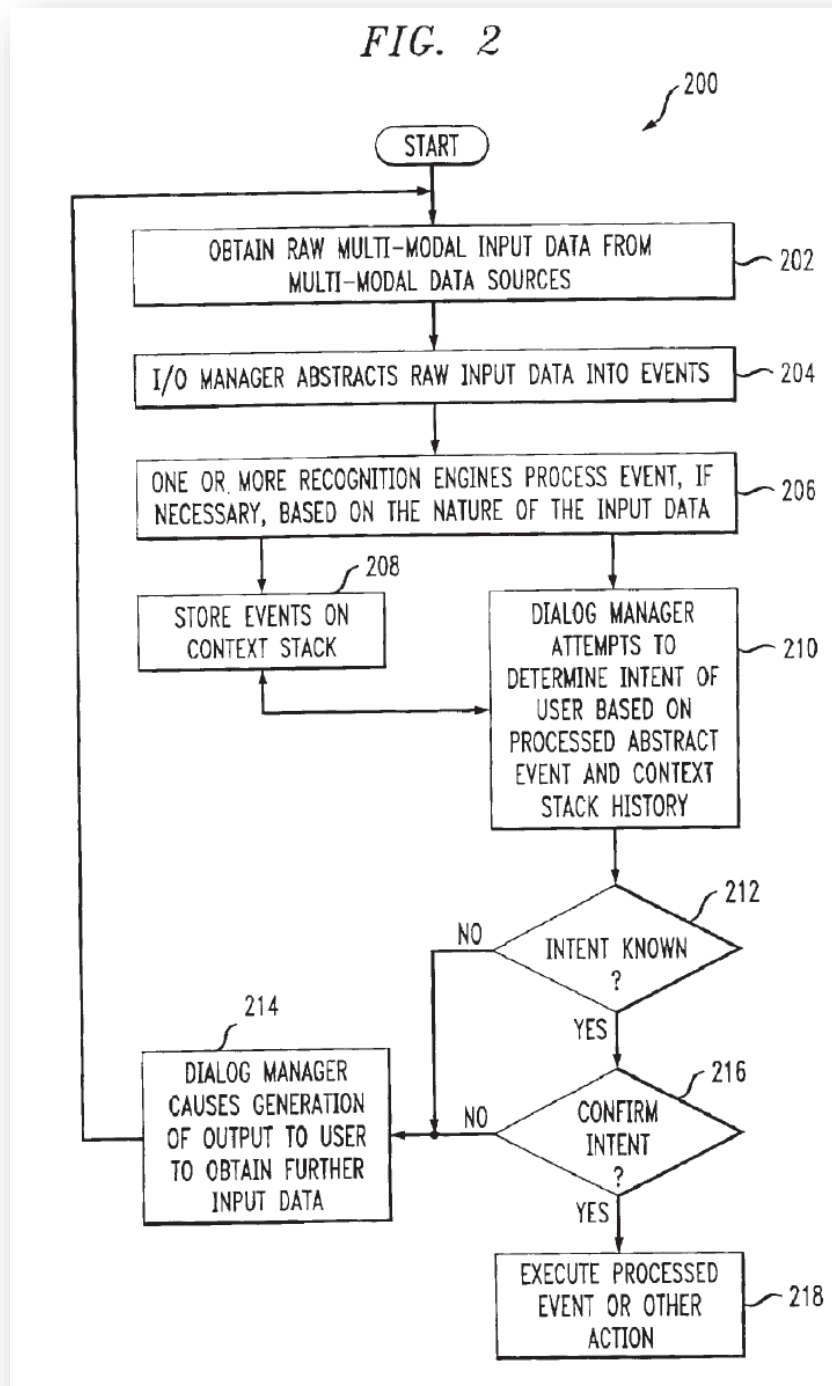
Ex.1005, 8:37–42.

114. In this example, the results of the recognized events comprise decoded text (or, script) (“*the one or more words*”) associated with the spoken utterance “turn it on” with the accompanying lip movement. Further, in the above example, “this event” corresponds to the current event of the user pointing to the radio, which occurred few seconds after “the previous spoken utterance event” of the user saying “turn it on.” Data for both events (“*the plurality of context entries*”) are stored on the context stack because Maes discloses “**all** the variable, states, input, output and queries to the backend ... performed in the context of the dialog and any extraneous event that occurs during the dialog” are stored on the context stack. Ex.1005, 37:55–62. Maes clarifies that “the dialog” includes “**conversational dialog comprising speech** and other multi-modal I/O.” Ex.1005, 36:49–54.

115. Therefore, Maes discloses the dialog manager associating the results of the recognized events included in the decoded text or script (i.e., “*the one or more words*”) 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 the user’s intent in an active dialog which serves as the “*identifying ...*” step and consequently, the associated context stored in the context stack corresponding to the user’s intent in the active dialog constitutes the “*one or more context entries that correspond to the one or more words.*” See also Ex.1005, 37:53–55.

116. Therefore, Maes teaches or suggests the “*identifying ...*” step for at least this reason.

117. Further, Maes Figure 2 depicts “a flow diagram illustrating a referential ambiguity resolution methodology performed by a multi-modal conversational computing system” used in determining the user’s intent teaches or suggests the “*identifying ...*” step. Ex.1005, 3:24–28.



Ex.1005, Figure 2. Maes explains that “[i]n step 208, the recognized events ... are stored in a storage unit referred to as the context stack 20. The **context stack** is

used to create a history of interaction between the user and the system ... **to assist the dialog manager 18 in making referential ambiguity resolution determinations** when determining the user's intent. Next, in step 210, the system 10 attempts to **determine the user intent based on the current event and the historical interaction information stored in the context stack ...**.” Ex.1005, 7:33–45.

118. Therefore, Maes teaches or suggests the “*identifying ...*” step for this additional reason.

119. For the above reasons, the associated context(s) stored in the context stack corresponding to the user's intent in the active dialog constitutes “*the plurality of context entries*” that “match[]” (“*correspond to*”) the decoded text (or, script) (“*the one or more words*”) associated with the spoken utterance, e.g., “turn it on” with the accompanying lip movement.

120. As further explained above, Maes teaches identifying matches between associated context(s) stored in the context stack corresponding to the user's intent in the active dialog (or, simply contexts stored on the context stack) and results of recognized events (or, simply transcribed user utterances). A POSITA would have understood that in speech recognition systems, it may be the case that—the associated context(s) stored in the context stack corresponding to the user's intent in the active dialog may not have a full or 100% (i.e., “complete”) match with the

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

121. To the extent that Patent Owner contends that Maes does not adequately disclose the “*identifying ...*,” step, Coffman discloses this element. See § [CITE].

122. Coffman explains that context stack entries are searched to identify a context matching one or more words transcribed/determined from the user’s spoken utterance. “**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 words of the user’s spoken utterance 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 one or more words transcribed/determined from the user’s spoken utterance, 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 one or more words transcribed/determined from the user’s spoken utterance and data stored on the context stack serve as the “*identifying ...*” functionality. Accordingly, based on Maes and the incorporation of Coffman, a POSITA would have found it obvious that, for instance, “decoded text” (or, script) (“*one or more words*”) processed from the spoken utterance (“*of the natural language utterance*”) produced as a result of processing the user’s

spoken utterance in Maes would be compared against data stored on the context stack, per Coffman's teachings, to “*identify[] ... one or more context entries that correspond to the one or more words.*” Ex.1012, 41:1–5, 40:15–19; Ex.1005, 7:60–8:4, 8:37–42, and 37:53–55. *See* § VIII.B.

123. Therefore, Maes teaches or suggests this element.

f. [7.4.1] wherein identifying the one or more context entries comprises: comparing the plurality of context entries to the one or more words;

124. Maes teaches or suggests “*wherein identifying the one or more context entries comprises: comparing the plurality of context entries to the one or more words.*”

125. As part of the “*identifying ...*” step explained in [7.4.0], Maes discloses that upon searching the context stack, the associated context stored in the context stack corresponding to the active dialog is provided by the dialog manager when such context “could be associated” with the “[*natural language*] utterance,” which therefore constitutes “*comparing*” data stored on the context stack (“*the plurality of context entries*”) “*to*” “*decoded text*” (or, script) (“*the one or more words*”) processed from the spoken utterance.

126. For example, Maes discloses:

Thus, the dialog manager must first determine the user's intent based on the current event and, if available, the historical information (e.g., past events) stored in the context stack. For instance, returning to the vehicle example, the user may say "turn it on," while pointing at the vehicle radio. The dialog manager would therefore receive the results of the recognized events associated with the spoken utterance "turn it on" and the gesture of pointing to the radio. **Based on these events, the dialog manager does a search of the existing applications, transactions or "dialogs," or portions thereof [stored on the context stack], with which such an utterance and gesture could be associated.**

Ex.1005, 7:60–8:4, *see also* 8:37–42. Therefore, Maes discloses "*wherein identifying the one or more context entries comprises: comparing the plurality of context entries to the one or more words.*"

127. To the extent that Patent Owner contends that Maes does not disclose the step of "*wherein identifying the one or more context entries comprises: comparing ...*," Coffman (incorporated by Maes) discloses this step. Ex.1005, 9:30–37.

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 one or more words transcribed/determined from the user’s spoken utterance, are compared against data stored in the context stack until a match is found. Thus, Coffman’s teachings of searching the context stack until an acceptable match is found between one or more words transcribed/determined from the user’s spoken utterance and data stored on the context stack serve as the “*wherein identifying the one or more context entries comprises: comparing the plurality of context entries to the one or more words.*”

128. Coffman explains that identifying context entries stored on the context stack includes “*comparing*” data stored on the context stack to one or more words transcribed/determined from the user’s spoken utterance. “**Completion/search for the active context is performed from context to context down the stack. That is, new queries or arguments [“*the one or more words*” processed from the utterance] are compared [“*comparing*”] by the dialog engine by going down the stack [“*the plurality of context entries*”] until an acceptable match is obtained**” Ex.1012, 42:1–3; *see also* § [CITE].

129. Therefore, Maes teaches or suggests this element.

- g. **[7.4.2] [wherein identifying the one or more context entries comprises:] generating, based on the comparison, one or more rank scores for individual context entries of the plurality of context entries; and**

130. Maes teaches or suggests “[*wherein identifying the one or more context entries comprises:] generating, based on the comparison, one or more rank scores for individual context entries of the plurality of context entries.*”

131. As part of the “*identifying ...*” step in [7.4.0] and based on the outcome of the “*comparing ...*” step in [7.4.1], Maes teaches “*generating ... scores for individual context entries*” or otherwise scoring one or more candidate contexts such as phonemes. Maes describes “phonemes” as “sub-phonetic or acoustic units of speech” (Ex.1005, 12:29–33) or equivalently as “portions [of dialogs]” (Ex.1005, 7:60–8:4).

132. Specifically, Maes discloses generating, for each considered phoneme in a given context, likelihood scores (indicating the likelihood that it was that particular phoneme that was spoken) based, for example, on the audio information and/or the video information received as input to the system.

133. And referencing Figure 4, Maes discloses:

After the acoustic feature vectors, denoted in FIG. 4. by the letter A, are extracted, the probability module labels

the extracted vectors with one or more **previously stored phonemes which ... are sub-phonetic or acoustic units of speech. The module may also work with lefemes, which are portions of phones [sic: phonemes] in a given context. Each phoneme associated with one or more feature vectors has a probability associated therewith indicating the likelihood that it was that particular acoustic unit that was spoken.**

Ex.1005, 12:29–37. Ex.1005, 12:29–37.

[T]he probability module 416 in the audio information path ... labels the acoustic feature vectors with one or more phonemes.

Ex.1005, 18:46–48.

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

Ex.1005, 12:38–43.

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. Again, each phoneme associated with one or more visual speech feature vectors has a probability associated therewith indicating the likelihood that it was that particular acoustic unit that was spoken in the video segment being considered. Thus, the probability module yields likelihood scores**

for each considered phoneme in the form of the probability that, given a particular phoneme or acoustic unit (au), the acoustic unit represents the uttered speech characterized by one or more visual speech feature vectors V or, in other words, $P(V|\text{acoustic unit})$. 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, 18:46–60.

134. As explained earlier the outcome of the “*identifying ...*” step (*see* [7.4.0]) would generate one or more candidate contexts from the context stack (Ex.1003, ¶120), and as such, the candidate contexts constitute “*individual context entries of the plurality of context entries*” in the context stack.

135. In Maes, “likelihood scores” (“*one or more rank scores*”) generated for “each ... phoneme” / candidate context (“*individual context entries of the plurality of context entries*”) is “*based on*” “*associat[ing]*” [i.e., “*the comparison*”] words of the user’s spoken utterance against information stored in the context stack. Ex.1005, 7:60–8:4 (disclosing “*associat[ing]*”), 8:37–42, 12:35–37; *see also* [7.4.1] explaining “*the comparison*” / the “*comparing ...*” step). Calculation of likelihood scores for each considered phoneme in a given context in the form of probabilities, accordingly, serves as the “*generating ...*” step in this claim element. Accordingly, Maes teaches calculation of probabilities/likelihood scores (e.g., “*generating ... one*

or more rank scores”) for each considered phoneme in a given context (e.g., *“individual context entries of the plurality of context entries”*).

136. Therefore, Maes teaches or suggests this element.

- h. [7.4.3] [wherein identifying the one or more context entries comprises:] identifying, based on the one or more rank scores, the one or more context entries from among the plurality of context entries; and**

137. Maes teaches or suggests *“[wherein identifying the one or more context entries comprises:] identifying, based on the one or more rank scores, the one or more context entries from among the plurality of context entries.”*

138. As explained in detail in connection with element [7.4.0], Maes teaches or suggests *“identifying, from among the plurality of context entries, one or more context entries”* and as explained in connection with element [7.4.2], Maes teaches or suggests calculating “likelihood scores” (*“one or more rank scores”*) for each considered phoneme.

139. Further, the *“identifying ...”* step is *“based on”* “likelihood scores” (*“one or more rank scores”*) because Maes teaches “likelihood scores” (*“one or more rank scores”*) of a considered/identified phoneme **“indicating the likelihood that it was that particular acoustic unit that was spoken.”** Ex.1005, 12:35–37. That is, in Maes, a phoneme, e.g., among *“the plurality of context entries”* stored on the

context stack is deemed to be “a considered phoneme” / candidate context in view of its “likelihood score” (“*rank score[]*”).

140. Indeed, Maes discloses using a “degree of confidence” (e.g., “*one or more rank scores*”) to determine that the “the [selected] application” (e.g., *one or more context entries from among the plurality of context entries*) “is the one which will effectuate the users desire” and accordingly, “the next step of the multi-modal dialog ... of that application based on the multi-modal input” is executed. Ex.1005, 8:19–21 and 8:8–11; *see also* Ex.1005, 7:60–8:4 (disclosing applications as an example of an entry stored on the context stack). In other words, Maes teaches selecting a particular application (e.g., “*identifying the one or more context entries*”) using (“*based on*”) a “degree of confidence” (e.g., “*one or more rank scores*”).

141. Therefore, Maes teaches or suggests this element.

- i. **[7.5] determining based on the determined one or more words and the context information, the command or the request associated with the natural language utterance.**

142. Maes teaches or suggests this claim element.

143. First, as explained earlier in connection with elements [7.3] and [7.1], in Maes, “decoded text” (or, script) constitutes “*the determined one or more words*” processed from the spoken utterance and Maes’s context stack (such as context stack

817 in Figure 8 or context stack 20 in Figure 1 comprises or stores “*context information.*” And, for the reasons below, Maes discloses this claim element.

i. Determining the command/request associated with the natural language utterance

144. Maes teaches or suggests “*determining ... the command or the request associated with the natural language utterance*” because Maes’s system provides functionality to “determine[] what information the user has, which inputs the user presents, and which application(s) should handle the user inputs [e.g., “the natural language utterance”].” Ex.1005, 36:49–59.

145. Specifically:

When an abstract event occurs, the DMA determines the target of the event and, if needed, seeks confirmation, disambiguation, correction, more details, etc., until the intent is unambiguous and fully determined. **The DMA 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]. The DMA function 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 to: (a) understand the intent of the user; (b) follow up with a dialog to disambiguate, complete, correct or confirm the understanding; (c) or, dispatch a task resulting from full understanding of the intent of the user.**

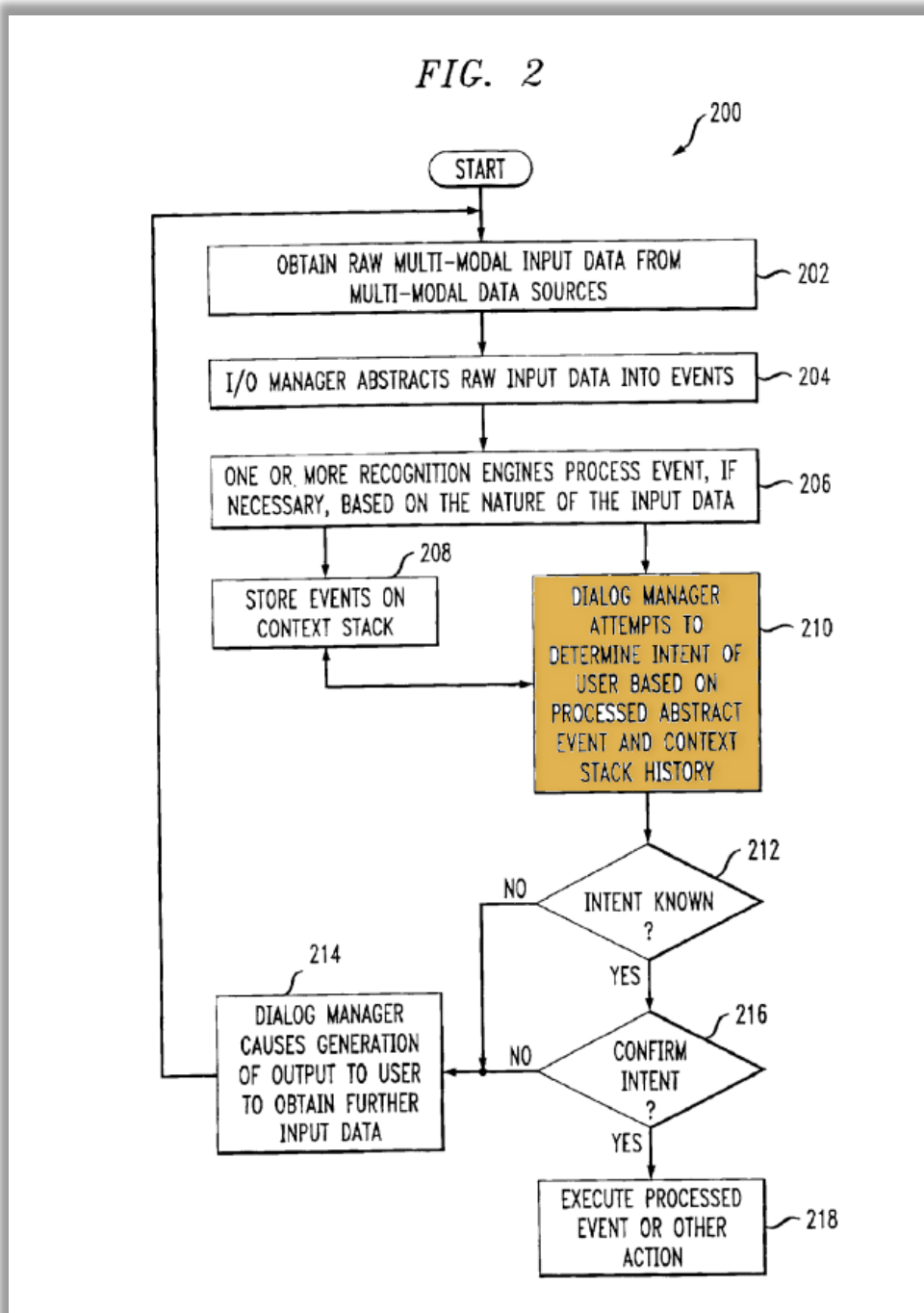
Ex.1005, 36:59–37:3; *see also* 4:6–22. Thus, Maes’s “system ... **recognizes voice-based commands** and ... recognizes gesture-based commands.” Ex. 1005, 2:1–4; *see also* Ex.1005, 4:58–59 (disclosing “user input commands received and processed by the system”). Therefore, Maes discloses “determin[ing]” (“*determining*”) “a user’s query” (“*the command or the request*”) “*associated with*” multi-modal data received in the form of audio input data, (“*the natural language utterance*”), video input data, and possibly other types of input data).

ii. Determining based on one or more words and the context information,

146. In Maes, “*determining ...*” “a user’s query” (“*the command or the request*”) is “*based on*” “decoded text” (or, script) (“*the determined one or more words*”) which is produced using context stack data (“*context information*”), explained below.

147. Referencing the Figure 2 process, Maes describes “in step 210, **the system ... attempts to determine the user intent based on the current event and the historical interaction information stored in the context stack.**” Ex.1005, 7:40–42. In step 210, the “processed abstract event” is the “decoded text” (or,

script) (“*the determined one or more words*”) and the “context stack history” constitutes “*the context information.*”



Ex.1005, Figure 2 (annotated).

148. Referencing the in-vehicle example, Maes states:

[T]he dialog manager must first determine the user's intent based on the current event and ... the historical information (e.g., past events) stored in the context stack. For instance, returning to the vehicle example, the user may say "turn it on," while pointing at the vehicle radio. The dialog manager would ... receive the results of the recognized events associated with the spoken utterance "turn it on" and the gesture of pointing to the radio. Based on these events, the dialog manager does a search of the existing applications, transactions or "dialogs," or portions thereof [stored on the context stack], with which such an utterance and gesture could be associated.

Ex.1005, 7:65–8:4, *see also* 8:37–42. In this example, determining the user's intent (*determining ... the command or the request associated with the natural language utterance*) is based on the results of the recognized events which comprise decoded text (or, script) (*"the one or more words"*) associated with the spoken utterance "turn it on" with the accompanying lip movement and the historical information (e.g., past events) stored in the context stack (*"the context information"*).

149. Maes further discloses 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 [“determining ... the command or request”] based on ... said ... what [“the determined one or more words”]**, what was the focus of the user when he gave a multi-modal query/command **and what is the history of these commands [“the context information”]** and focuses

Ex.1005, 2:54–67.

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

2. Dependent Claim 8

a. **[8.0] The method of claim 7, further comprising: providing a prompt regarding the command or request, wherein the prompt is a prompt for additional information or a prompt for confirmation; and**

151. Maes teaches or suggests “*providing a prompt regarding the command or request, wherein the prompt is a prompt for additional information or a prompt for confirmation.*”

i. Providing a prompt regarding the command or request

152. Referencing the in-vehicle example and Figure 2, Maes explains:

Consider the case where the user says “turn it on,” but makes no gesture and provides no other utterance. In this case, assume that the dialog manager does not have enough input to determine the user intent (step 212 in FIG. 2) and thus implement the command. The dialog manager, in step 214, .. causes the **generation of an output to the user requesting further input data so that the user’s intent can be disambiguated.** ... [For example,] [t]he

**dialog manager ... generates a predetermined question
such as “what device do you want to have turned on?”**

....

See generally, Ex.1005, 8:43–65

153. In the above example, providing the question “what device do you want to have turned on?” constitutes “*providing a prompt*” “*regarding*” “the user’s query” (“*the command or request*”), e.g., relating to the spoken utterance of “turn it on.” *Id.*

**ii. The prompt is a prompt for additional
information or confirmation**

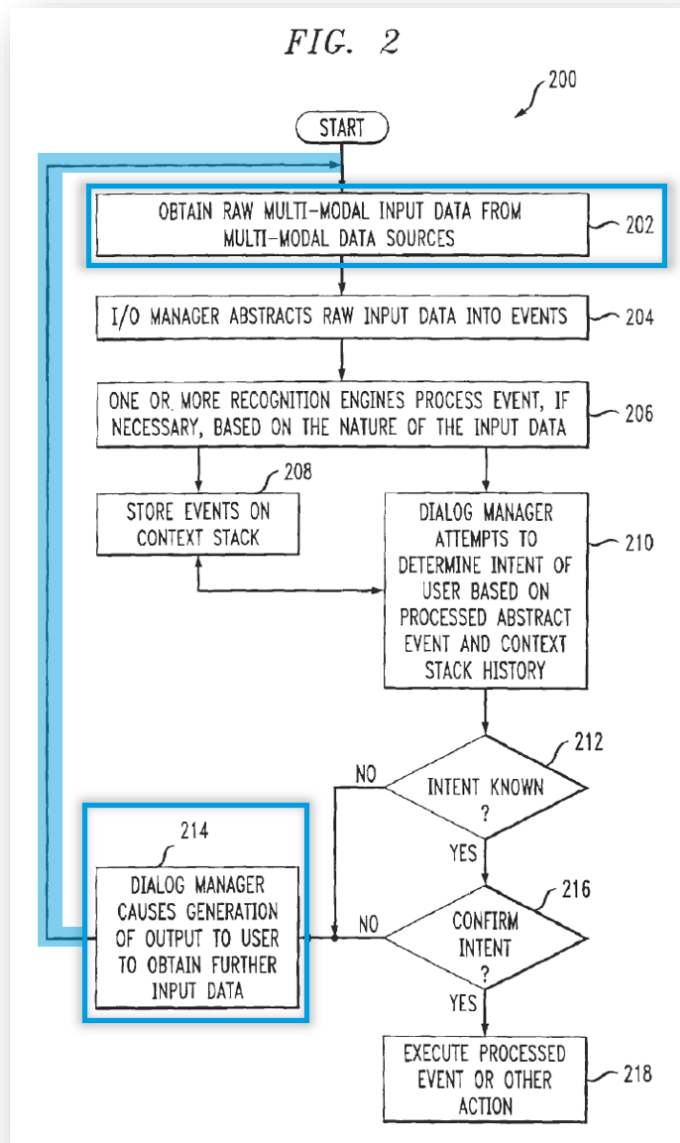
154. Additionally, continuing with the above example (Ex.1005, 8:43–65), the question (“*the prompt*”) “*is*” “output to the user requesting further input data” (“*a prompt for additional information ...*”).

155. Maes further explains that when an abstract input/output (I/O) event occurs, Maes’s system seeks “**confirmation**, disambiguation, correction, **more details**” (“*a prompt for additional information or ... confirmation*”) (“*a prompt for additional information or ... confirmation*”). Ex.1005, 36:59–63. Specifically, Maes’s system seeks a user’s follow-up input / feedback “to disambiguate, complete, correct or **confirm** the understanding.” Ex.1005, 36:64–37:2.

156. Maes discloses requesting additional information/confirmation with more disclosures:

The dialog manager 18 may also seek **confirmation** in step 216 from the user in the same manner as the request for **more information** (step 214) before executing the processed event For example, the system may output “**do you want the radio turned on?**” To which the user may respond “yes.”

Ex.1005, 8:66–9:6. In this example, Maes’s system asking the user “do you want the radio turned on?” constitutes “*a prompt for additional information ...*”.



Ex. 1005, Figure 2 (annotated) (depicting “generation of output to user to obtain further input data” received in block 214).

157. Therefore, Maes teaches or suggests this element.

- b. [8.1] receiving, in response to providing the prompt, a non-speech input, wherein the command or request is determined further based on the non-speech input.**

158. Continuing with the in-vehicle system example discussed in [8.0] (see Ex.1005, 8:43–65), Maes teaches that upon hearing the question “what device do you want to have turned on?,” the user “could then point to the radio or say “the radio” thereby providing the dialog manager with the additional input data to disambiguate his request.” 8:58–61. Thus, Maes’s system receives (“*receiv[es]*”) in response to providing the question (“*the prompt*”), the user’s gesture of pointing to the radio or the lip movement accompanying the spoken utterance “the radio,” each of which serve as examples of “*a non-speech input*.” Furthermore, Maes teaches that the user’s query (“the command or request”), e.g., relating to the spoken utterance of “turn it on” is disambiguated (“determined further”) “based on” the user’s gesture or the lip movement accompanying the spoken utterance (“non-speech input”).

159. Therefore, Maes teaches or suggests this element.

B. Ground 2: Maes and Coffman render obvious claims 7 and 8.

160. To the extent one asserts that Coffman is not particularly incorporated by reference into Maes as if they were effectively the same document, claims 7 and 8 would have nonetheless been obvious in view of Maes and Coffman for the reasons articulated above in Ground 1. Moreover, as demonstrated above, a

POSITA would have been motivated to combine the references and had a reasonable expectation of success without undue experimentation for such a combination as articulated in §VIII.B.

C. Ground 3: Kennewick, Weissman, and Maes renders obvious claims 7 and 8.

161. I understand that Kennewick and Weissman were a basis for a final rejection and cancellation of at least claim 7 of the 957 Patent from an *ex parte* reexamination. I further understand that Patent Owner is collaterally estopped from asserting that claim 7 of the 957 is valid based on the results of that earlier proceeding, and I have been asked to assume that claim 7 is invalid based on the combination of the Kennewick and Weissman references, which I understand were at issue in that *ex parte* reexam proceeding.

162. Maes teaches or suggests all the elements of claim 8 for the reasons demonstrated above in Ground 1. *See* [8.0] and [8.1] above.

163. It would have been obvious to a POSITA to combine Kennewick and Weissman with Maes's teachings regarding prompting for additional input or confirmation, including where such additional input or confirmation is received by means other than speech. The combination would have predictably resulted in Kennewick and Weissman's speech recognition processing functionalities enhanced

to include follow-up prompting, including from non-speech data per Maes, as claimed.

164. Kennewick, Weissman, and Maes are analogous art as the 957 Patent for two reasons. *First*, each are directed toward the same field of endeavor as the 957 Patent—computer-implemented and natural language processing systems that include a focus on the meaning of language (e.g., semantics), and are reasonably pertinent to a problem described by the 957 Patent: an environment for **reliably processing** a user’s language queries, which is a problem of conventional systems identified in the 957 Patent. Ex.1001, Abstract, 24:40-43; Ex.1032, [0003], [0009]; Ex.1025, Abstract, 2:17-49; Ex.1005, 41:58-64.

165. A POSITA would have been motivated to combine Kennewick, Weissman and Maes in the manner described. As discussed above, a POSITA would have already understood and found it obvious based on Kennewick’s disclosures and background knowledge that Kennewick’s system sought means to reliably process user requests, and Maes discloses a technique that seeks a prompt for additional information and/or confirmation: “**confirmation, disambiguation, correction, more details, etc., until the intent is unambiguous and fully determined**” Ex.1005, 36:59–63. Specifically, Maes discloses a technique that seeks a user’s follow-up input / feedback “to disambiguate, complete, correct or confirm the understanding.” Ex.1005, 36:64–37:2.

166. Kennewick confirms this understanding by demonstrating a similar technique—it explains that:

If the confidence level of the domain or context score is not high enough to ensure a reliable response, **the system can request that the user verify the question or command is correctly understood**. In general, the question may be phrased to indicate the context of the question including all criteria or parameters. If the user confirms that the question is correct the system may proceed to produce a response. Otherwise, either the user can rephrase the original question, **perhaps adding additional information to remove ambiguity**, or the system may ask one or more questions to attempt to resolve the ambiguity or other actions may taken.

Ex.1032, [0131]; *see also* Ex.1032, [0145]. Because of the similarity of Maes's techniques, a POSITA, seeking to further understand or implement Kennewick's disclosures, would have appreciated that Maes adds certain additional functionality, including multimodal input, so a POSITA would have turned to Maes for its enhanced functionality and more detailed teachings.

167. A person of ordinary skill in the art would have had a reasonable expectation of success, as combining and Kennewick would have involved the

straightforward application of conventional techniques and technologies without undue experimentation.

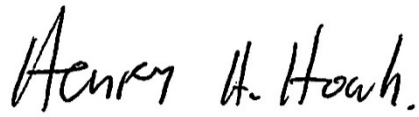
168. Accordingly, Kennewick and Weissman, in further view of Maes renders obvious claim 8.

X. OBJECTIVE INDICIA OF NON-OBVIOUSNESS

169. I am not aware of any objective indicia of non-obviousness that would change my opinion that the challenged claims of the 957 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

170. 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.

A handwritten signature in black ink that reads "Henry H. Houh." The signature is written in a cursive style with a period at the end.

Henry H. Houh

Dated: March 21, 2025

CLAIM LISTING

Claim 7	
[7.0]	A computer-implemented method of processing a natural language utterance, the method comprising:
[7.1]	generating a context stack comprising context information that corresponds to a plurality of prior utterances, wherein the context stack includes a plurality of context entries;
[7.2]	receiving, by one or more computer processors, the natural language utterance, wherein the natural language utterance is associated with a command or is associated with a request;
[7.3]	determining one or more words of the natural language utterance by performing speech recognition on the natural language utterance;
[7.4.0]	identifying, from among the plurality of context entries, one or more context entries that correspond to the one or more words, wherein the context information includes the one or more context entries,
[7.4.1]	wherein identifying the one or more context entries comprises: comparing the plurality of context entries to the one or more words;
[7.4.2]	[wherein identifying the one or more context entries comprises:] generating, based on the comparison, one or more rank scores for individual context entries of the plurality of context entries; and
[7.4.3]	[wherein identifying the one or more context entries comprises:] identifying, based on the one or more rank scores, the one or more context entries from among the plurality of context entries; and
[7.5]	determining based on the determined one or more words and the context information, the command or the request associated with the natural language utterance.
Claim 8	
[8.0]	The method of claim 8, further comprising: providing a prompt regarding the command or request, wherein the prompt is a prompt for additional information or a prompt for confirmation; and
[8.1]	receiving, in response to providing the prompt, a non-speech input, wherein the command or request is determined further based on the non-speech input