

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

MICROSOFT CORPORATION,

Petitioner,

v.

DIALECT, LLC,

Patent Owner.

IPR2025-00658

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

PETITION FOR *INTER PARTES* REVIEW OF

U.S. PATENT NO. 8,195,468

TABLE OF CONTENTS

	Page
EXHIBIT LIST	VI
TABLE OF ABBREVIATIONS AND CONVENTIONS	VIII
I. INTRODUCTION	1
II. MANDATORY NOTICES	1
A. Real Party-in-Interest	1
B. Related Matters.....	1
C. Lead and Back-Up Counsel and Service Information	2
III. PAYMENT OF FEES	2
IV. REQUIREMENTS FOR INTER PARTES REVIEW.....	3
A. Grounds for Standing	3
B. Challenged Claims	3
V. THE 468 PATENT.....	3
A. Effective Filing Date	3
B. Person of Ordinary Skill in the Art.....	3
C. Overview of the 468 Patent.....	4
VI. CLAIM CONSTRUCTION	5
VII. PRINCIPAL PRIOR ART	6
A. Summary of Maes	6
B. Maes’s Teachings with those of Ittycheriah and Coffman	7
VIII. THE CHALLENGED CLAIMS ARE UNPATENTABLE	11
A. Ground 1: Maes Renders Obvious claims 19, 25-30, and 32	11

1.	Claim 19.....	11
a.	[19.0] A method for processing multi-modal natural language inputs, comprising:.....	11
b.	[19.1.0] receiving a multi-modal natural language input at a conversational voice user interface, the multi-modal input including a natural language utterance and a non-speech input provided by a user,....	13
c.	[19.1.1] wherein a transcription module coupled to the conversational voice user interface transcribes the non-speech input to create a non-speech based transcription;	17
d.	[19.2] identifying the user that provided the multi-modal input;	22
e.	[19.3.0] creating a speech-based transcription of the natural language utterance using a speech recognition engine and a semantic knowledge-based model, ...	25
f.	[19.3.1] wherein the semantic knowledge-based model includes a personalized cognitive model derived from one or more prior interactions between the identified user and the conversational voice user interface, ...	36
g.	[19.3.2] [wherein the semantic knowledge-based model includes ...] a general cognitive model derived from one or more prior interactions between a plurality of users and the conversational voice user interface, ...	41
h.	[19.3.3] [wherein the semantic knowledge-based model includes ...] an environmental model derived from an environment of the identified user and the conversational voice user interface;	43
i.	[19.4] merging the speech-based transcription and the non-speech-based transcription to create a merged transcription;	47

j.	[19.5] identifying one or more entries in a context stack matching information contained in the merged transcription;	53
k.	[19.6] determining a most likely context for the multi-modal input based on the identified entries;	62
l.	[19.7] identifying a domain agent associated with the most likely context for the multi-modal input;	64
m.	[19.8] communicating a request to the identified domain agent; and	66
n.	[19.9] generating a response to the user from content provided by the identified domain agent as a result of processing the request.	67
2.	Claim 25	68
a.	[25.0] The method of claim 19, further comprising verifying an identity of the user based on voiceprint matching, password matching, or pass-phrase matching.	68
3.	Claim 26	68
a.	[26.0] The method of claim 19, further comprising receiving one or more follow-up multi-modal inputs at the conversational voice user interface, the follow-up multi-modal inputs including at least one of a follow-up natural language utterance or a follow-up non-speech input.	68
4.	Claim 27	70
a.	[27.0] The method of claim 26, wherein the identified domain agent updates the context stack and the semantic knowledge-based model in response to processing the request.	70
5.	Claim 28	76
a.	[28.0] The method of claim 27, wherein the knowledge-enhanced speech recognition engine determines a most likely context for the follow-up multi-modal input using the updated context stack.	76

6.	Claim 29	79
a.	[29.0] The method of claim 27, wherein the speech recognition engine creates a speech-based transcription of the follow-up natural language utterance using the updated semantic knowledge-based model.	79
7.	Claim 30 and 32	79
a.	[30.0] The method of claim 19, wherein the identified domain agent processes the request by querying one or more local or network information sources.	79
b.	[32.0] The method of claim 19, wherein the identified domain agent processes the request by directing a command to one or more local or remote devices.	80
B.	Ground 2: Maes, Coffman and Ittycheriah Render Obvious claims 19, 25-30, and 32	80
IX.	BOARD DISCRETION	82
A.	35 U.S.C. § 325(d).....	82
B.	35 U.S.C. § 314(a).....	83
1.	Factor 1: Stay	83
2.	Factor 2: Trial Date	83
3.	Factor 3: Parallel Proceeding.....	84
4.	Factor 4: Overlapping Issues	84
5.	Factor 5: Overlapping Parties	85
6.	Factor 6: Strength/Other Considerations	85
X.	CONCLUSION.....	86
	CERTIFICATE OF COMPLIANCE	87
	CERTIFICATE OF SERVICE	88

CLAIM LISTING	90
---------------------	----

EXHIBIT LIST

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

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

TABLE OF ABBREVIATIONS AND CONVENTIONS

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

TABLE OF AUTHORITIES

	Page(s)
Cases	
<i>Abbot Vascular, Inc. v. Flexstent</i> , IPR2019-00882, Paper 48, 28-29 (Oct. 2, 2020).....	8
<i>In re Apple Inc.</i> , 979 F.3d 1332 (Fed. Cir. 2020)	83
<i>Apple Inc. v. Fintiv, Inc.</i> , IPR2020-00019, Paper 15, 12 (May 13, 2020).....	83, 85
<i>Arbutus Biopharma Corp. v. ModernaTX, Inc.</i> , 65 F.4th 656 (Fed. Cir. Apr. 11, 2023).....	7
<i>Black v. CE Soir Lingerie Co.</i> , No. 2:06-CV-544, 2008 WL 3852722 (E.D. Tex. Aug. 15, 2008) <i>aff'd</i> , 319 F. App'x 901 (Fed. Cir. 2009)	8
<i>CommScope Techs. LLC v. Dali Wireless, Inc.</i> , IPR2022-01242, Paper No. 23 (PTAB Feb. 27, 2023).....	85
<i>CrowdStrike, Inc. v. Webroot Inc.</i> , IPR2023- 00126 (PTAB May 5, 2023).....	82
<i>Dialect, LLC v. Bank of America, N.A.</i> , 2:24-cv-00207-JRG (EDTX)	1
<i>Dish Network v. Broadband iTV</i> , IPR2020-01280, Paper 17 (PTAB Feb. 4, 2021).....	84
<i>Fintiv</i> , IPR2020-00019, Paper 11 (PTAB Mar. 20, 2020).....	85
<i>Google LLC v. Dialect, LLC</i> , IPR2024-00750 (PTAB).....	1, 82, 84
<i>Google LLC v. Dialect, LLC</i> , IPR2024-00752 (PTAB).....	1

<i>Google LLC v. Dialect,</i> LLC, IPR2024-00753 (PTAB)	1
<i>Google, LLC v. Parus Holdings Inc.,</i> IPR2020-00847, Paper 9 (PTAB Oct. 21, 2020)	83
<i>Harari v. Lee,</i> 656 F.3d 1331 (Fed. Cir. 2011)	7
<i>NetNut Ltd. v. Bright Data Ltd.,</i> IPR2021-01492, Paper 12 (PTAB Mar. 21, 2022)	84
<i>Quasar Science LLC v. Colt International Clothing, Inc. d/b/a Colt</i> <i>LED,</i> IPR2023-00611, Paper 10 (PTAB Oct. 10, 2023)	82
<i>Scientific Design Co. Inc. v. Shell Oil Co.,</i> IPR2022-00158, Paper 7 (PTAB Apr. 4, 2022)	82
<i>Sony v. MZ Audio Scis., LLC,</i> IPR2022-01544, Paper No. 12	83
<i>Sotera Wireless, Inc. v. Masimo Corp.,</i> IPR2020-01019, Paper 12 (PTAB Dec. 1, 2020)	84
<i>Vivid Techs., Inc. v. Am. Sci. & Eng.,</i> 200 F.3d 795 (Fed. Cir. 1999)	5
Statutes	
35 U.S.C. § 102(e)	6
35 U.S.C. § 103	1, 3
35 U.S.C. § 314(a)	83
35 U.S.C. § 325(d)	82, 83
Other Authorities	
37 C.F.R. § 42.6(e)	88
37 C.F.R. § 42.24	87

37 C.F.R. § 42.100(b)	5
-----------------------------	---

I. INTRODUCTION

Petitioner submits this Petition for *Inter Partes* Review of claims 19, 25-30, and 32 (the “Challenged Claims”) of U.S. Patent No. 8,195,468 (the “468 Patent” (Ex.1001)), assigned to Patent Owner. Petitioner respectfully submits that the Challenged Claims of the 468 Patent are unpatentable under 35 U.S.C. §103 in view of the prior art references discussed herein.

II. MANDATORY NOTICES

A. Real Party-in-Interest

Apart from Petitioner Microsoft Corporation, a real party-in-interest for this Petition is Bank of America, N.A.

B. Related Matters

The 468 Patent is asserted by PO in *Dialect, LLC v. Bank of America, N.A.*, 2:24-cv-00207-JRG (EDTX) and *Microsoft Corporation*, 2-24-cv-01067 (EDTX). In addition, the following cases involve PO and related patents: *Dialect, LLC v. Salesforce, Inc.*, 7:25-cv-00061, *Google LLC v. Dialect, LLC*, IPR2024-00750 (PTAB), *Google LLC v. Dialect, LLC*, IPR2024-00752 (PTAB), *Google LLC v. Dialect, LLC*, IPR2024-00753 (PTAB). In addition, related patents are also challenged in a number of IPRs by Microsoft, including IPR2025-00655, and IPR2025-00657, and forthcoming IPRs as well, which Microsoft can identify in subsequent mandatory notices.

C. Lead and Back-Up Counsel and Service Information

Lead Counsel	Back-Up Counsel
Scott M. Border (Reg. #77,744) Winston & Strawn LLP 1901 L Street, N.W. Washington, D.C. 20036 Tel: (202) 282-5054 sborder@winston.com	Carson Swope <i>(Pro Hac Vice to be submitted)</i> Winston & Strawn LLP 255 Shoreline Dr., Ste. 520 Redwood City, CA 94065 Tel: (650) 858-6407 cswope@winston.com
	T.D. Goswami (Reg. #78,054) Winston & Strawn LLP 1901 L Street, N.W. Washington, D.C. 20036 Tel: (202) 282-5309 tdgoswami@winston.com

Petitioner consents to electronic service by email at Winston-Microsoft-Dialect-IPR@winston.com and the e-mail addresses listed above.

III. PAYMENT OF FEES

Petitioner authorizes the Office to charge the filing fee and any other necessary fee to Deposit Account No. 501814.

IV. REQUIREMENTS FOR INTER PARTES REVIEW

A. Grounds for Standing

Petitioner certifies that the 468 Patent is available for inter partes review. Petitioner is not barred or estopped from requesting an inter partes review challenging the claims on the identified grounds herein. Petitioner has not filed a civil action challenging the validity of a claim of the 468 Patent. This petition is being filed no more than 1 year after the date on which Petitioner was served with a complaint alleging infringement of the 468 Patent.

B. Challenged Claims

Ground 1: Maes (Ex.1005) renders obvious claims 19, 25–30, and 32 under 35 U.S.C. § 103.

Ground 2: Maes, Coffman (Ex.1012), and Ittycheriah (Ex.1006) render obvious claims 19, 25–30, and 32 under 35 U.S.C. § 103.

V. THE 468 PATENT

A. Effective Filing Date

Petitioner assumes for the purposes of this Petition that August 29, 2005 is the effective filing date.

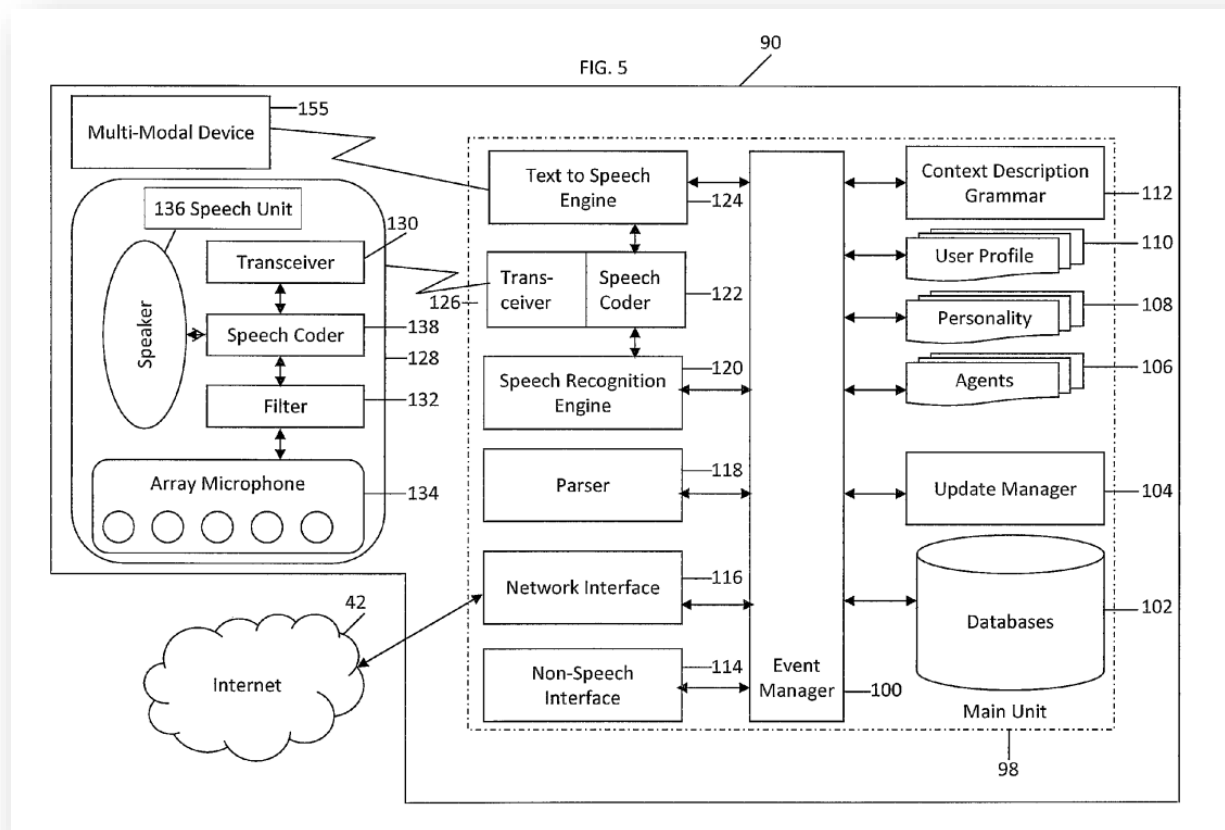
B. Person of Ordinary Skill in the Art

A POSITA with respect to the Asserted Patents as of the time of the invention, i.e., the earliest possible priority date of the 468 Patent, August 29, 2005, would have

had a bachelor's degree in electrical engineering, computer science, computer engineering, or an equivalent, and two years of relevant experience involving computer science fundamentals, including natural language processing, speech recognition and transcription, non-speech recognition and transcription that is pertinent to the 468 Patent. Lack of professional experience can be remedied by additional education, and vice versa. Ex.1003, ¶25.

C. The 468 Patent

The 468 Patent is directed to a system that “includes speech-based and non-speech-based interfaces for telematics applications,” and “identifies and uses context, prior information, domain knowledge, and user specific profile data to achieve a natural environment for users that submit requests and/or commands in multiple domains.” Ex.1001, Abstract. Figure 5 of the 468 Patent is characterized as “an overall diagrammatic view of the interactive natural language processing system according to an embodiment of the invention.” Ex.1001, 15:38–40.



Ex.1001, Figure 5.

VI. CLAIM CONSTRUCTION

Claims are given their “ordinary and customary meaning” as understood by a POSITA and the prosecution history pertaining to the patent. 37 C.F.R. § 42.100(b). Because a POSITA would find the challenged claims unpatentable under any interpretation consistent with their plain and ordinary meaning in the context of the 468 Patent, the Board need not expressly construe the claim terms. *Vivid Techs., Inc. v. Am. Sci. & Eng. ’g. Inc.*, 200 F.3d 795, 803 (Fed. Cir. 1999).

VII. PRINCIPAL PRIOR ART

A. Summary of Maes

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. Maes qualifies as prior under at least 35 U.S.C. § 102(e).

Maes describes “systems and methods ... for performing focus detection, referential ambiguity resolution and mood classification in accordance with multi-modal input data ... to provide an effective conversational computing environment for one or more users.” Ex.1005, Abstract.

Maes expressly incorporates by reference Coffman and Ittycheriah in their entireties. Maes states that “the conversational virtual machine disclosed in PCT patent application identified as PCT/US99/22927 filed on Oct. 1, 1999 and entitled “Conversational Computing Via Conversational Virtual Machine,” [*i.e.*, Coffman] the disclosure of which is incorporated by reference herein, may be employed to provide a framework for the I/O manager, recognition engines, dialog manager and context stack of the invention.” Ex.1005, 9:30–37. Maes also states “another method to detect incorrectly recognized words is disclosed in U.S. Pat. No. 5,937,383 to Ittycheria entitled “Apparatus and Methods for Speech Recognition Including Individual or Speaker Class Dependent Decoding History Caches for Fast

Word Acceptance or Rejection,” the disclosure of which is incorporated herein by reference.” Ex.1005, 19:43–49.

Coffman and Ittycheriah are, therefore, considered effectively part of Maes. *Arbutus Biopharma Corp. v. ModernaTX, Inc.*, 65 F.4th 656, 662-63 (Fed. Cir. Apr. 11, 2023) (“When a reference or material from various documents is incorporated, they are ‘effectively part of the host document as if [they] were explicitly contained therein.’”); *see also Harari v. Lee*, 656 F.3d 1331, 1335 (Fed. Cir. 2011).

Because the entirety of Coffman is directed to “Conversational Computing Via Conversational Virtual Machine,” Dr. Houh explains that a POSITA would have understood Maes incorporates 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 incorporated with particularity, which is the material relied upon below. Similarly, the entire disclosure of Ittycheriah is also expressly incorporated with particularity given its focus on “detect[ing] incorrectly recognized words”—the focus of the entire disclosure of Ittycheriah—the material of which is relied upon below. Ex.1003, ¶55.

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

To the extent one asserts that Coffman and Ittycheriah encompass different embodiments from that of Maes, a POSITA would have been motivated to include

Ittycheriah's teachings of maintaining a decoding history cache database storing keywords (vocabulary such as names and/or commands) frequently used by a user or a class of speakers, and score information generated from decoded keywords in accordance with Hidden Markov Models (HMMs) (Ex.1006, 1:55–58, 4:19–32, 4:54–63, 5:5–19) in Maes's semantic module for the reasons below. Further, as also explained below, a POSITA would have been additionally motivated to combine Coffman's teachings related to searching a context stack to find entries that match transcribed user input and merging/combining speech and non-speech (e.g., keyboard input) transcriptions (Ex.1012, 41:2–5, 42:1–3, 38:1–19) with Maes's teachings of processing multi-modal inputs. Ex.1003, ¶56.

Maes expressly incorporates by reference Coffman and Ittycheriah. In addition, each name Stephane Maes as a common inventor, and have IBM as the common assignee. Accordingly, a POSITA would have been motivated to combine the teachings of Coffman and Ittycheriah with those of Maes for at least these reasons. *Black v. CE Soir Lingerie Co.*, No. 2:06-CV-544, 2008 WL 3852722, at *14 (E.D. Tex. Aug. 15, 2008) *aff'd*, 319 F. App'x 901 (Fed. Cir. 2009) (“strong motivation to combine” prior art “patented by the same inventor” and that “reference each other.”); *Abbot Vascular, Inc. v. Flexstent*, IPR2019-00882, Paper 48, 28-29 (Oct. 2, 2020) (a POSITA would be motivated to combine references by the same author and from the same company).

Additionally, a POSITA would have found Ittycheriah provides a motivation to combine its teachings with Maes—a POSITA would have found Ittycheriah’s cache database (“*cognitive model[s]*”) storing words/keywords (vocabulary such as names and/or commands) frequently used by a user/class of users (Ex.1006, 1:55–58, 4:54–63, 5:5–19) would reduce the amount of hypotheses (limiting the set of possible hypotheses) to consider in Maes’s decoding process, thereby enabling faster matching/decoding of keywords, and improving the responsiveness of Maes’s system, which is consistent with Maes’s goal of faster executions. Ex.1005, 11:64–12:30, 5:29–34; Ex.1006, 2:13–19, 2:58–3:21, 1:37–43; Ex.1003, ¶58.

Dr. Houh states that using HMMs for decoding words, as Ittycheriah teaches (see Ex.1006, 4:19–32), was well-known and utilized in language understanding and determined pertinent for analyzing cognitive processes such as in speech recognition models at the time of the 468 Patent. Ex.1003, ¶59 (citing Ex.1008, [0013]; Ex.1009, 258; Ex.1010, 403 and 413). Dr. Houh explains that the evidence supporting his opinion reflect the POSITA’s knowledge as it relates to popularity of HMMs and their benefits to speech recognition techniques at the time of the claimed invention in support of Ittycheriah’s teachings. Ex.1003, ¶60.

Furthermore, Ittycheriah discloses an example of a voice-controlled navigation application and Maes discloses an example implementation of an in-vehicle system. Ex.1006, 2:49–53, 4:54–5:4; Ex.1005, 6:15–50, 6:43–50, 7:60–

8:42. A POSITA would be further motivated to combine the teachings of Maes and Ittycheriah because Ittycheriah's voice-controlled navigation would significantly benefit Maes's in-vehicle system by providing voice-based navigation assistance to drivers. Ex.1003, ¶61.

In addition, a POSITA would have been motivated to combine Coffman's teachings of organizing its context stack and updating it such that the most-recent context is at the top with the teachings of Maes and Ittycheriah to benefit Maes's system by providing efficient identification and retrieval of the most-recent context, allowing for faster executions and improving the responsiveness of Maes's system, and would be consistent with Ittycheriah's goals of quickly accepting or rejecting keywords. Ex.1012, 41:1–5, 40:15–19, 42:7–9; Ex.1005, 7:60–8:4, 8:37–42, 37:51–62, 5:29–34; Ex.1006, 2:58–60; Ex.1003, ¶62.

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

synchronize information streams. Ex.1005, 20:5–7, 11:55–63, 21:51–54, 29:30–36; Ex.1012, 38:2–19; Ex.1003, ¶63.

Dr. Houh explains that a POSITA would have had a reasonable expectation of success in combining the teachings of these references because to do so would be a simple combination of analogous teachings that would be well within the skill of a POSITA. For example, modifying Maes's semantic module to include Ittycheriah's teachings of a cache database storing scores and words decoded in accordance with HMMs from previous interactions with a user or a class of users and to further modify Maes's system to include Coffman's teachings related to searching a context stack to find entries that match transcribed user input and merging/combining speech and non-speech-based transcriptions would add to the functionality of Maes's system and would have been a simple implementation without undue experimentation and predictably resulted in a recognition system with improved accuracy, as Maes seeks to implement (Ex.1005, 11:62–63). Ex.1003, ¶64.

VIII. THE CHALLENGED CLAIMS ARE UNPATENTABLE

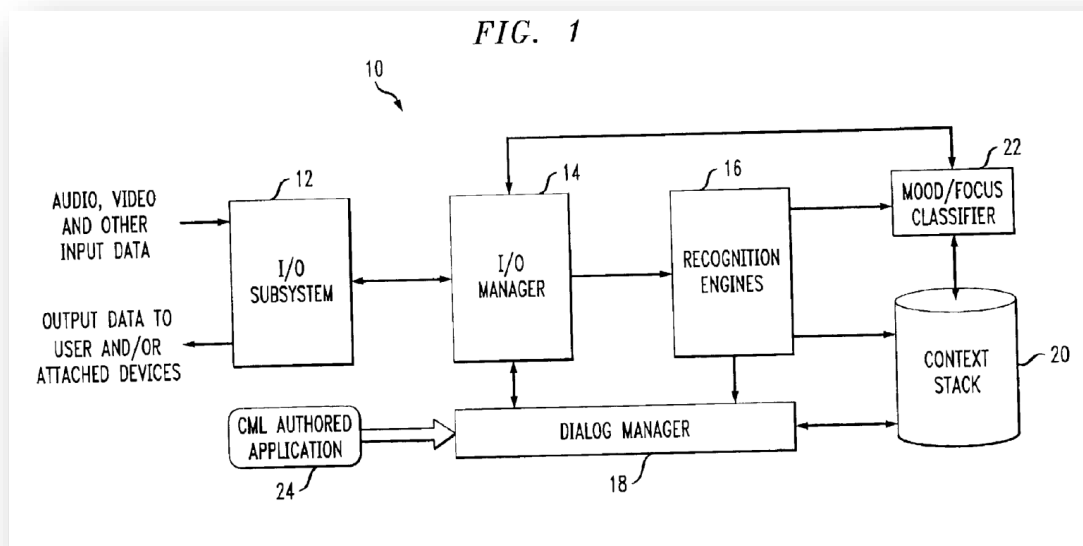
A. Ground 1: Maes Renders Obvious claims 19, 25-30, and 32

1. Claim 19

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

Maes describes “**multi-modal data processing** techniques.” Ex.1005, Abstract, 4:10–11, 1:9–23 (emphasis added¹); Ex.1003, ¶67.

Further, Maes is directed to “*processing multi-modal natural language inputs.*” Referencing Figure 1, Maes describes “**one or more recognition engines**” that “receive[] **multi-modal input** in the form of **audio** input data, **video** input data, as well as **other types** of input data,” (“*multi-modal natural language inputs*”) and “performs various recognition tasks (e.g., **speech recognition**, ... lip reading, face recognition ... in accordance with the recognition engines)” Ex.1005, 4:1–4; 3:66–4:17.



¹ Emphasis added throughout unless otherwise specified.

Ex.1005, Figure 1; Figure 4 (describing “an audio-visual speech recognition module”²). Ex.1003, ¶¶68-69.

Maes explains that “multi-modality” “may comprise a combination of other modalities other than voice and video using the same “methodology,” including keyboard/pointer/mouse (or telephone keypad) and other sensors.” Ex.1005, 3:4–11; Ex.1003, ¶¶70-71; *see also* ¶73 (describing examples of spoken utterance inputs and natural language processing).

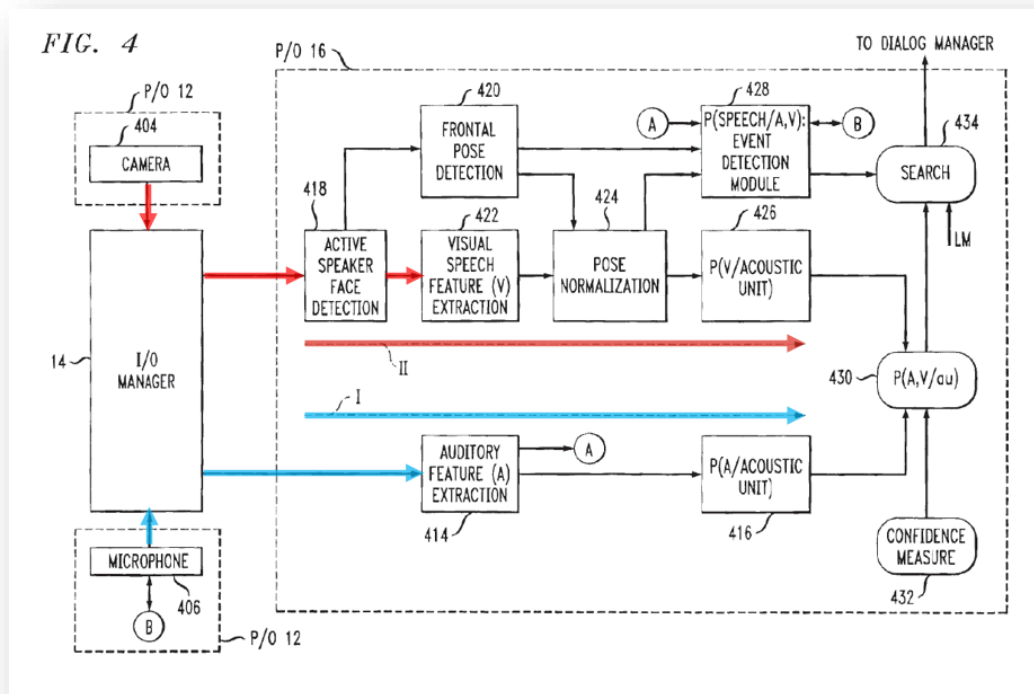
Therefore, if limiting, Maes teaches or suggests the preamble. Ex.1003, ¶74.

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

Describing Figure 4, Maes discloses “an audio-visual speech recognition module” receiving “multi-modal input data”—such as an audio signal and a video signal—from a speaker (“*receiving a multi-modal natural language input ... provided by a user*”), (Ex.1005, 10:47–11:23), at an “input/output (I/O) subsystem ... compris[ing] one or more microphones for capturing audio input data,” as well as other input devices (“*at a conversational voice user interface*”). Ex.1005, 6:5–8, 6:21–29; Ex.1003, ¶76.

² The terms “Maes’s system” and “audio-visual speech recognition module” are used interchangeably.

Maes discloses the “multi-modal input data” includes “an audio signal” (speech provided by the speaker and/or background noise) (“*a natural language utterance*”) and “a video signal” (“*a non-speech input*”), including the speaker’s face, lip movement and background objects in the environment. Ex.1005, 11:8–63. Maes’s system “receives multi-modal input in the form of **audio input data, video input data, as well as other types of input data**” Ex.1005, 4:7–15; Ex.1005, 4:67–5:5 (“system 10 ... **taking both audio input data and image data input and processing it**”):



Ex.1005, Figure 4 (annotated) and 11:55–63; Ex.1003, ¶77.

Referencing Figure 4³, Maes explains “receiving real-time arbitrary content from a video camera 404 and microphone 406 via the I/O manager 14” Ex.1005, 11:23–31. Feature extractor 414 in Maes “**receives an audio or speech signal**” and the active speaker face detection module 418 “**receives video input [from] camera 404**” (“the visual portion (e.g., lip movement) of the utterance.” Ex.1005, 6:46-47; 11:64–12:2, 12:49–61. Ex.1003, ¶¶78-79.

Maes explains that **audio signal** can include “silence,” “speech,” and “noise” and, the **video signal** can include “the visual portion (e.g., lip movement) of the utterance.” Ex.1005, 6:46–47. In Maes, cameras/sensors capture video input data comprising visible image, IR, and/or RF image data, and microphones capture audio input data from the environment in which Maes’s system is deployed. Ex.1005, 11:32–54; Ex.1003, ¶80; *see also* ¶¶81-84 (describing multi-modality) .

The 468 Patent describes an embodiment which the disclosed system (mobile structure 10) is “used in vehicles” and comprises “speech interface and non-speech interface,” (Ex.1001, 15:58–64) and includes “video camera[s]” and “microphones . . . to receive the utterances from the user.” Ex.1001, 16:19–20, 25:48–52, Figure 1. Ex.1003, ¶85.

³ A typographical error appears in Figure 4—the labels “P/O” are incorrect and should be denoted “I/O.”

Maes provides a similar vehicle example. Ex.1005, 6:15–28, 6:43–50. For example, Maes’s system receives multi-modal inputs (“*receiving a multi-modal natural language input*”) relating to an I/O event representative of the user’s spoken utterance (“*a natural language utterance*”) and lip movement corresponding to the spoken utterance (“*a non-speech input*”) such 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.” Ex.1005, 6:43–50. *See generally* Ex.1005, 6:15–28, 6:35–38, 7:63–67; Ex.1003, ¶86.

Therefore, Maes teaches or suggests this claim element. Ex.1003, ¶87.

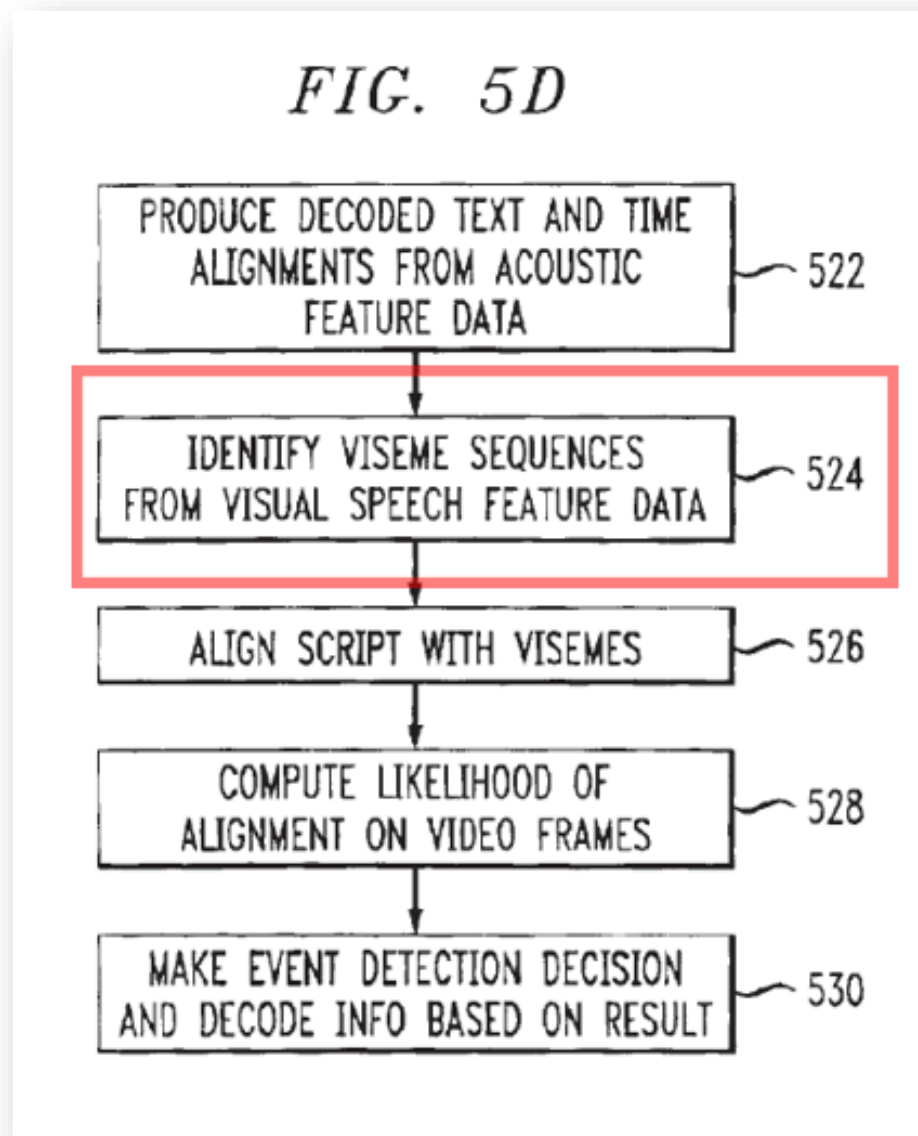
Additional Mapping

As Dr. Houh explains, in Maes, keyboard-typed input and touch screen input constitute additional examples of “*a non-speech input*” and consistent with the description in PO’s Complaint. Ex.1005, 3:1–12, 6:22–26, 16:20–21, 31:61–66, 32:29–32; Ex.1003, ¶¶88–89 (citing Ex.1007, ¶111 (keyboard-entered natural language input (“From my checking account” and “non-speech, icon-selection natural language input (by tapping one of the provided selection boxes containing the natural language input text)”).

Therefore, Maes teaches or suggests this claim element for this additional reason. Ex.1003, ¶90.

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

The visual speech feature extractor 422 in Figure 4 “extracts visual speech feature vectors (e.g., mouth or lip-related parameters)” from the video signal (“*transcribes the non-speech input*”) to produce “a visual phonemes (visemes) sequence” (“*to create a non-speech based transcription*”), including the face detected in the video frame and associated lip movement. Ex.1005, 17:51–55, 21:47–50. The “*transcri[ption]*” process in Maes is disclosed as a processing step in block 524 in Figure 5D:



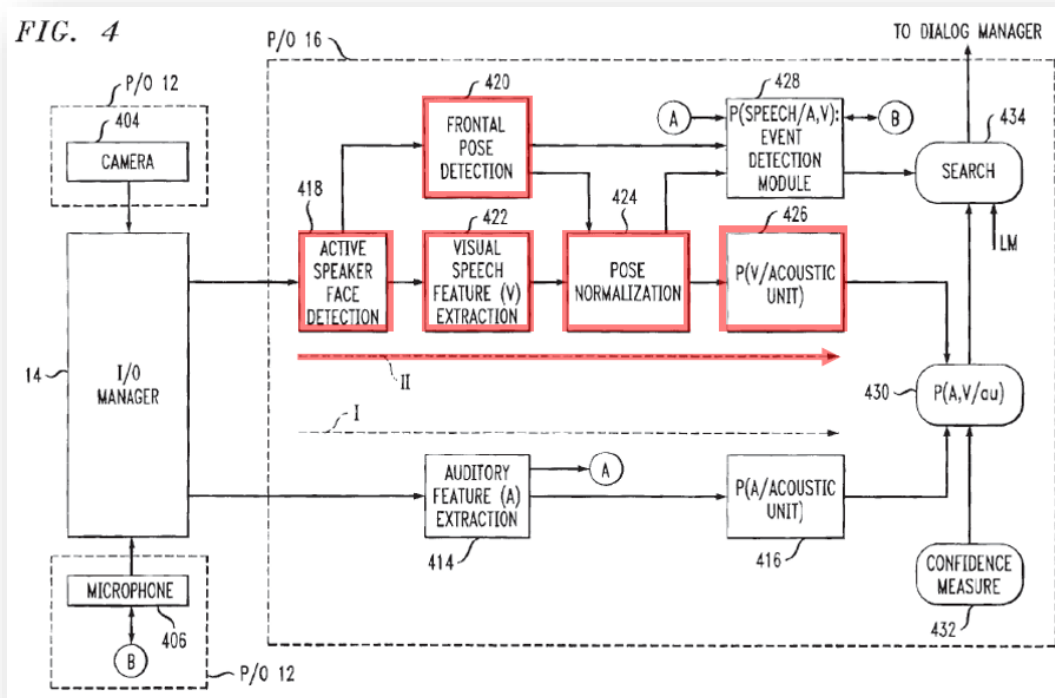
Ex.1005, Figure 5D (annotated); Ex.1003, ¶Ex.1003, ¶92.

For example, Maes explains that modules 422, 424 and 426 extract, normalize, and label visemes (“*transcribe*”) into a visemes sequence (“*a non-speech based transcription*”) using the extracted visual speech vectors.

The extracted visual speech feature vectors [produced by visual speech feature extractor 422] **are then normalized in block 424 with respect to the frontal pose estimates generated by the detection module 420. The normalized visual speech feature vectors are then provided to a probability module 426 [which] ... labels the extracted visual speech vectors with one or more previously stored phonemes. ... Alternatively, the visual speech feature vectors may be labeled with visemes which, as previously mentioned, are visual phonemes or canonical mouth shapes that accompany speech utterances.**

Ex.1005, 18:42–63; Ex.1003, ¶93.

In Maes, “visemes, or visual phonemes, are generally canonical mouth shapes that accompany speech utterances which are categorized and pre-stored similar to acoustic phonemes.” Ex.1005, 29:27–30. As Dr. Houh explains, a POSITA would have understood that a visemes sequence (“*a non-speech-based transcription*”) comprises a string of visemes that is the tangible result of a visual recognition system’s interpretation of a user’s lip movement accompanying a spoken utterance and provides a visual representation of the non-speech content. Ex.1003, ¶94 (citing Ex.1028, 1–2, Figures 1, 5 (describing visemes composed of features and represented by images)).



Ex.1005, Figure 4 (showing visual speech feature vectors denoted by the letter “V” and video information processing path (Path II) annotated in red), *see also* Ex.1005, 23:1–5. The red annotated blocks in Figure 4 collectively comprise “a transcription module,” “coupled to” “the (I/O) subsystem 12”, which includes “the conversational voice user interface.” Ex.1005, 6:5–8; 6:21–29; *see also* [19.1.1]; Ex.1003, ¶95; *see also* ¶¶96–97 (describing “visual speech” techniques in Maes).

Maes describes an exemplary in-vehicle “*transcri[ption]*” process as follows: “[t]he I/O manager receives the raw multi-modal data and **abstracts the data into a form that represents ... a spoken utterance [by] generalizing details associated with all or portions of the input data ... to yield a more generalized**

representation of the data for use in further operations.” Ex.1005, 6:21–38.

Thus, as a result of processing by several modules, the system produces a visemes sequence corresponding to the lip movement accompanying the spoken utterance of the user. *See also* Ex.1005, 6:43–50; Ex.1003, ¶98.

Accordingly, Maes teaches or suggests this element for at least this reason. Ex.1003, ¶99.

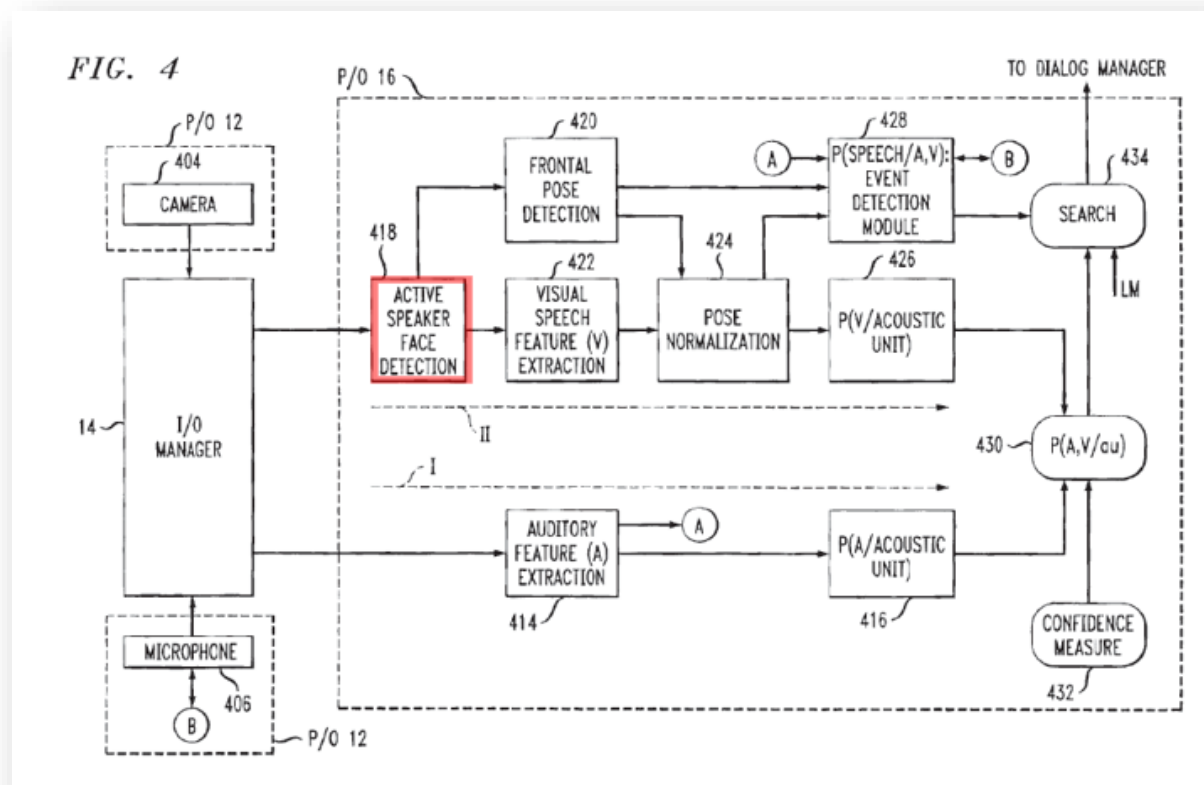
Additional Mapping

As Dr. Houh explains, in Maes, keyboard-typed input and touch screen input constitute additional examples of “*a non-speech input*” and further Maes teaches for “typical GUI-based input (e.g., keyboard, mouse, etc.), the data may not necessarily need to be recognized since the data is already identifiable [*“non-speech-based transcription”*] without recognition operations.” Ex.1003, ¶100 (citing Ex.1005, 6:62–65). PO’s Complaint is consistent with Maes’s teachings above. Ex.1003, ¶101 (citing Ex.1007, ¶114 (“Erica transcribed a non-speech input (tapping on box provided by Erica with the words “Let’s go”) into a non-speech-based transcription (displayed/represented to the user in the form of “Let’s go” in the user’s dialogue stream)).”).

Accordingly, Maes teaches or suggests this element for this additional reason. Ex.1003, ¶102.

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

Figure 4 includes an active speaker face detection module 418 for detecting (“*identifying*”) a face of the “speaker” that provided the “multi-modal input data” (“*the user that provided the multi-modal input*”). Ex.1005, 10:47–63, 12:49–62.



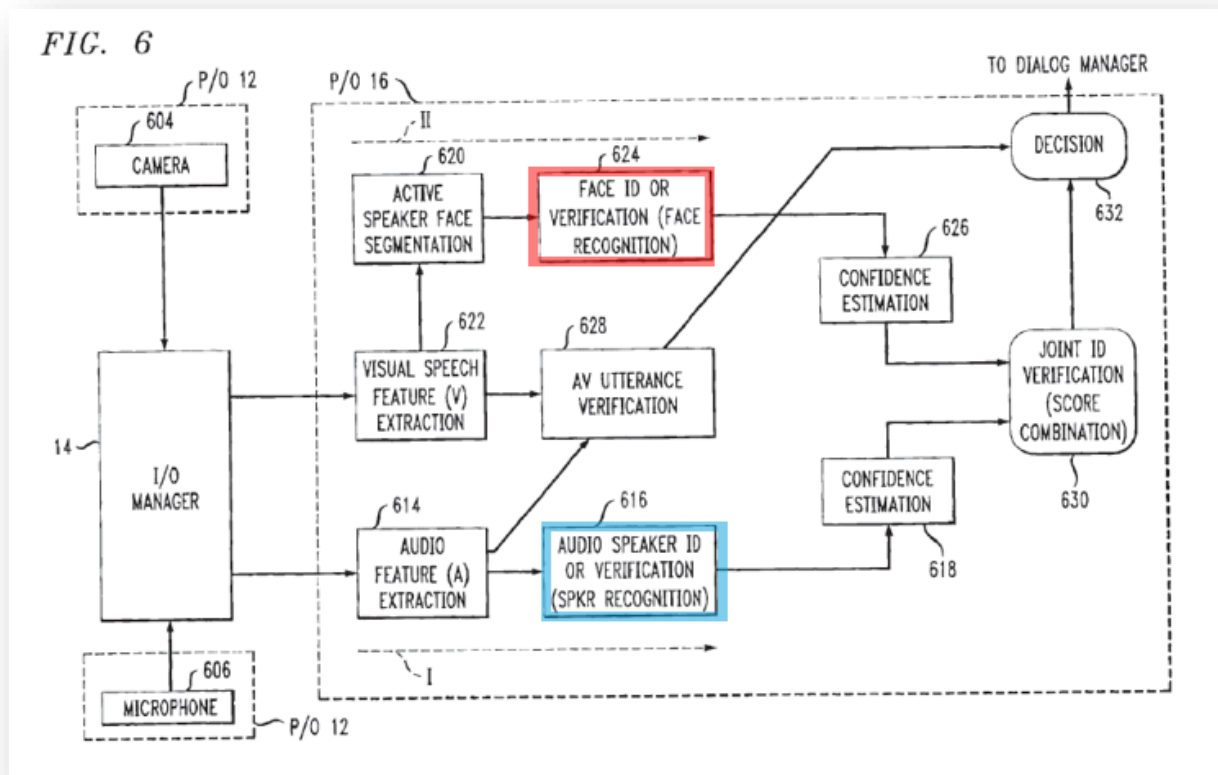
Ex.1005, Figure 4 (annotated); Ex.1003, ¶104.

Maes explains that “[t]he active speaker face detection module 418 receives video [from] input camera 404” and further that “speaker face detection can also be performed directly in the compressed data domain and/or from audio and video information rather than just from video information. [M]odule 418 generally locates

and tracks the speaker's face and facial features within the arbitrary video background.” Ex.1005, 12:49–62. Therefore, by detecting a speaker's face in the video signal comprising the speaker's face including lip movement and/or background objects in the environment, Maes's system “*identif[ies] ... the user that provided the multi-modal input.*” Ex.1005, 6:50–55 (“[I]n addition, the event may be sent to an audio-visual speaker recognition engine to **have the speaker of the utterance identified, verified and/or authenticated ...**”). Ex.1003, ¶105; *see also* ¶106 (referencing Figure 6 audio-visual speaker recognition module as one of the recognition modules of FIG. 1 **to perform speaker recognition using multi-modal input data received in accordance with the invention;**” Ex.1005, 22:41–46.)

Maes describes a speaker identification system may use two techniques: a model-based approach and a frame-based approach, and provides details of the frame-based approach. Ex.1003, ¶106; *see also* ¶107 (citing Ex.1005, 24:2–30 for details of frame-based approach).

Figure 6 of Maes shows a face recognition module 624 that uses video signals to “*identify[] the user that provided the multi-modal input.*” The face recognition module 624 “*identif[ies] and/or verif[ies] the face of the person assumed to be the speaker in the video.*” Ex.1005, 24:44–25:3.



Ex.1005, Figure 6 (annotated). “**Speaker verification may be performed in a similar manner**, ... the input acoustic data is compared to determine if the data matches closely enough with stored models. If the comparison yields a close enough match, the person uttering the speech is verified. The match is accepted or rejected by comparing the match with competing models.” Ex.1005, 24:31–36; Ex.1003, ¶108.

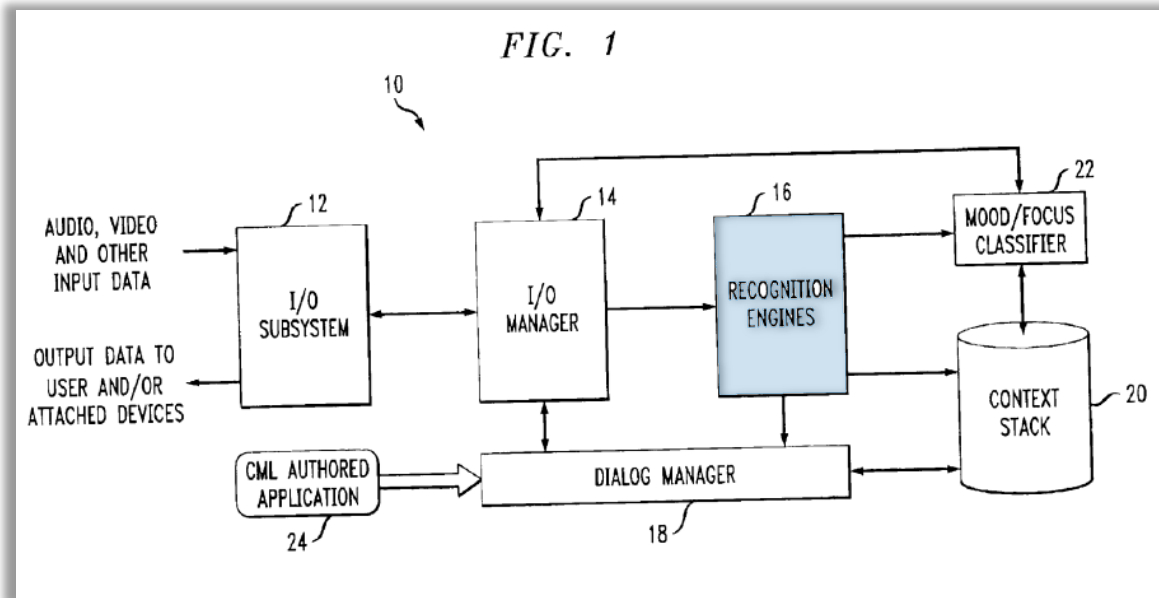
Maes’s exemplary in-vehicle system performs the “*identifying*” step by recognizing/identifying the specific user who provided the multi-modal natural language inputs. Ex.1005, 6:43–47, 7:64; Ex.1003, ¶109.

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

As Dr. Houh explains, before the priority date of the 468 Patent, “*a speech recognition engine and a semantic knowledge-based model*” were well-known in the prior art. Ex.1003, ¶112 (citing Ex.1018, 6:45–47, 8:4–7, 2:60–3:62, ¶¶113–114 (citing Ex.1011, Ex.1015, [0146])).

- i. **Speech Recognition Engine**

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.” Ex.1005, 6:42–50.



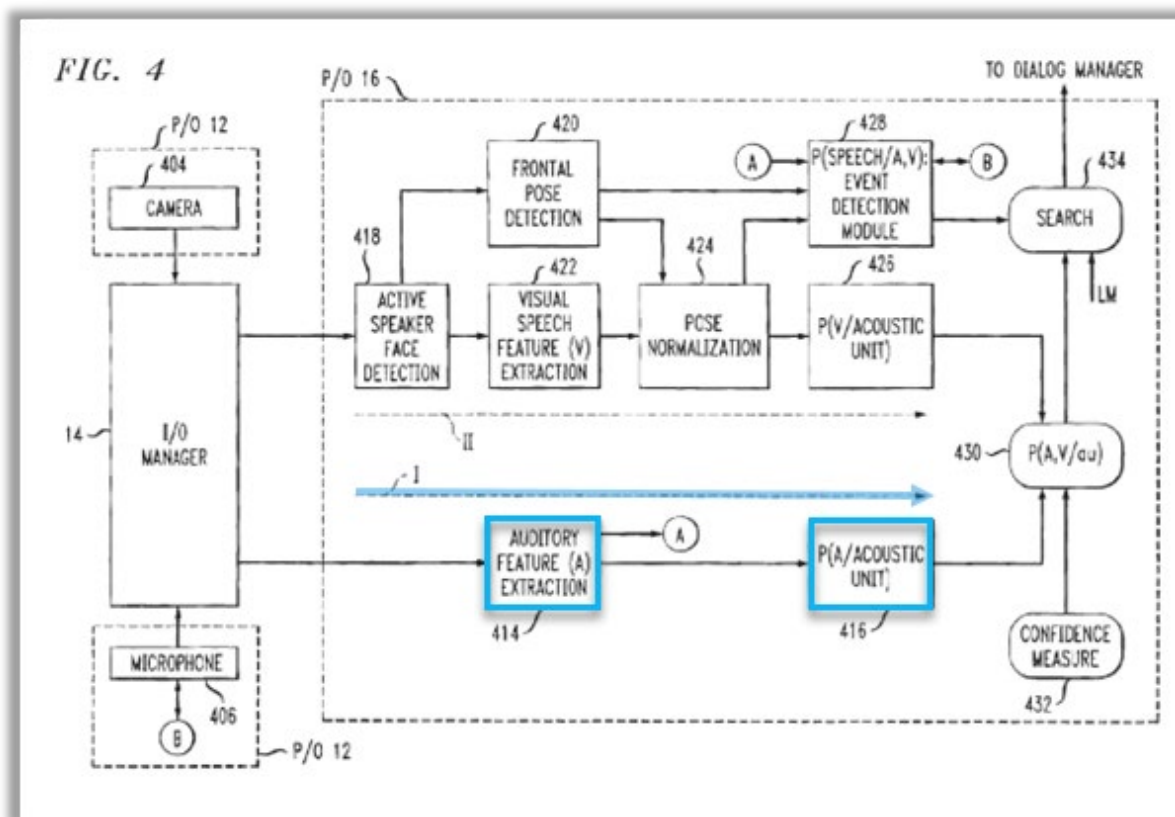
Ex.1005, Figure 1 (annotated); Ex.1003, ¶116.

Additional details of Maes's "*speech recognition engine*" are provided. Figure 4 is described as "an audio-visual **speech recognition module** that may be employed as **one of the recognition modules**⁴ of FIG. 1 to perform **speech recognition using [received] multi-modal input data.**" Ex.1005, 10:50–52. Video signals are processed by blocks/modules 418, 422, 424, 426 and audio signals are processed by blocks/modules 414, 416, in Figure 4. Ex.1003, ¶117.

Maes states that "the processing performed in blocks 414 and 416 may be accomplished via any **conventional acoustic information recognition system** [*a*

⁴ Maes uses the term "block" and "module" interchangeably. See 12:43-38, 17:46-50.

speech recognition engine”] capable of extracting and labeling acoustic feature vectors.” Ex.1005, 12:43–49. Dr. Houh explains that one such “acoustic information recognition system” that provides additional details on how speech is processed in Maes is described with reference to Figure 9A, explained below. Therefore, at least “blocks 414 and 416” disclose a “*a speech recognition engine.*”



Ex.1005, Figure 4 (showing audio information processing path (Path I) annotated in blue); Ex.1003, ¶118.

Referencing Figures 1 and 9A, Maes explains 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, Figure 9A; Ex.1003, ¶119.

Maes continues: “Apparatus 900 ... includes a processing module 910” further including “**a speech recognizer 926** which ... include[s] **a speech recognition module 928**,” and “a speech prototype, language model and grammar database”. Ex.1005, 39:59, 41:35–38. Apparatus 900 can further include **a post processor 938 ... configured to transcribe user utterances and ... perform keyword spotting thereon.** ... Post processor 938 can employ speech recognizer 926.” Ex.1005, 41:48–59. Therefore, Maes discloses “*a speech recognition engine.*” Ex.1003, ¶120.

ii. Semantic Knowledge-Based Model

Maes’s system (embodied in apparatus 900) includes a semantic module. Maes explains that apparatus 900 which receives “user-provided input data events” “via the I/O manager 14 of FIG. 1” (Ex.1005, 39:26–29) includes “a semantic module:”

Post processor 938 can also include **a semantic module (not shown) to interpret meaning of phrases.** The

semantic module could be used by speech recognizer 926 to indicate that some decoding candidates in a list are meaningless and should be discarded/replaced with meaningful candidates.

Ex.1005, 41:58-64; Ex.1003, ¶123.

Dr. Houh explains that a POSITA would have understood that Maes's semantic module comprises “*a semantic knowledge-based model.*” Ex.1003, ¶¶123–124 (citing Ex.1017, 429, 468; Ex.1014, Abstract; Ex.1019, 1157). Additional disclosures demonstrate that Maes discloses “*a semantic knowledge-based model.*” Ex.1003, ¶125.

For example, Maes describes “[a]ll data stored in the data warehouse 912 [shown in Figure 9A] can be stored in a format to facilitate subsequent data mining thereon” and accordingly, “[a]ppropriate key words, sentences, or grammatical mistakes to detect can be compiled via expert linguistic knowledge.” Ex.1003, ¶126–127 (citing Ex.1005, 43:38–41, 39:63–67, 43:16–20, 41:48–51, 42:60–61 (“a keyword spotter to detect insults”); Ex.1017, 468; Ex.1019, 1157). Dr. Houh explains that detecting appropriate key words, sentences, insults, and/or grammatical mistakes fall within the purview of Maes's semantic module. Maes's description of using “expert linguistic knowledge” and additionally, teachings of modifying the behavior of Maes's system on-the-fly further demonstrate that the semantic module

comprises “*a semantic knowledge-based model*.” Ex.1003, ¶¶127–128; Ex.1003, ¶129 (disclosing “dynamic decision” making by the models and citing Ex.1018, 8:4–14); *see also* Ex.1003, ¶130 (demonstrating disclosures that further support Maes teaching or suggesting a “*semantic knowledge-based model*” and citing Ex.1005, 1:53–57, 2:1–4, 11:64–12:28).

Other evidence demonstrate that a “*semantic knowledge-based model*” was well-known in the prior art. For example, Komissarchik, reflecting a POSITA’s knowledge, describes its invention as “knowledge-based speech recognition apparatus and methods” that provide “semantic control of the recognition results.” Ex.1018, 2:39–43, 6:45–47, Abstract; Ex.1003, ¶131.

Additionally, Maes’s “semantic module” is consistent with PO’s description of “*semantic knowledge-based model*” in its Complaint. Ex.1003, ¶132 (Ex.1007, ¶133).

iii. Creating a Speech-Based Transcription

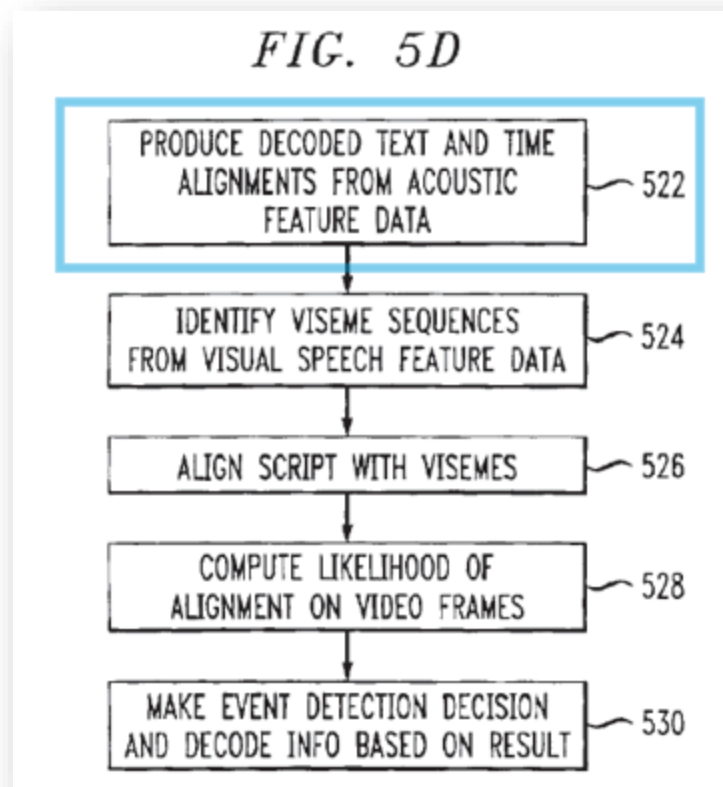
Maes expressly teaches “**one or more recognition engines of FIG. 1 . . . “need[s] data ... (e.g., grammar, acoustic models).”** Ex.1005, 34:62–66. Therefore, Maes’s “*a speech recognition engine*” uses data provided by a model, such the “*semantic knowledge-based model*” implemented in Maes’s “semantic module.” *See also* Ex.1005, 41:61–64. Additionally, Maes discloses using

“historical interaction information,” which, a POSITA would have understood would be included in such a model. Ex.1005, 7:40–45 and 24:31–45; Ex.1003, ¶134.

Referencing Figure 4, Maes explains that “the processing performed in blocks 414 and 416”—(e.g., “*speech recognition engine*”) “produces” (“*creat[es]*”), for instance, “**decoded text**” (“*a speech-based transcription*”) (Ex.1005, 21:43–47), and further, referencing Figure 9, Maes explains that the “semantic module” functions to “indicate that some **decoding candidates** in a list are meaningless and should be discarded/replaced with meaningful candidates” (Ex.1005, 41:61–64). Thus, Maes explains that the “**decoded text**” produced by the processing performed in blocks 414 and 416 is the same “decoding candidates” Maes describes in connection with Figure 4, further confirming Figure 9A and its corresponding disclosure provide additional details to, among other things, the “*speech recognition engine*” described in Figure 4. Referencing Fig 9A, Maes explains that “[a]pparatus 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; Ex.1003, ¶135.

Therefore, Maes discloses “*creating a speech-based transcription ... using a speech recognition engine and a semantic knowledge-based model.*” Ex.1003, ¶135.

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) (“*a speech-based transcription*”)” is, therefore, “produce[d]” (“*creat[ed]*”) “using the feature data from the acoustic feature extractor 414.” Ex.1005, 21:43–47.



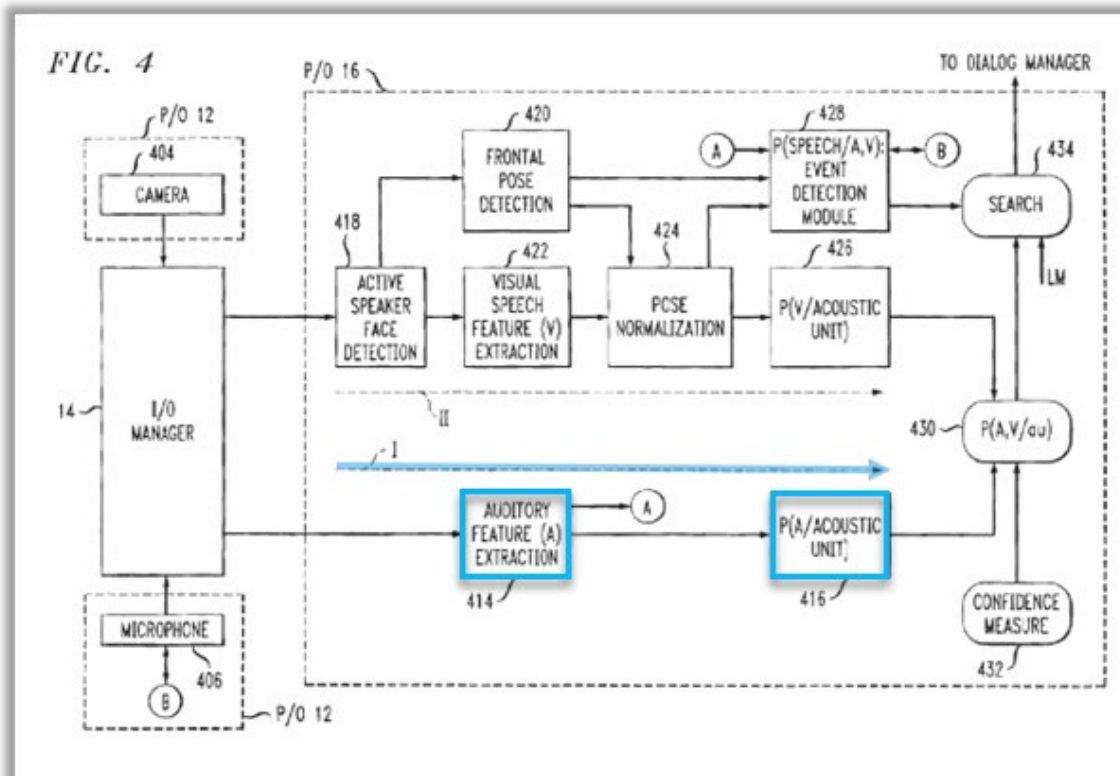
Ex.1005, Figure 5D (block 522). Dr. Houh states that a POSITA would have understood that the “decoded script” described in Maes is the tangible result of the

speech recognition system's interpretation of spoken language and provides a text representation of the multi-modal content. 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 ... to yield a **decoded output.**”); *see also* 20:23-33; Ex.1003, ¶136.

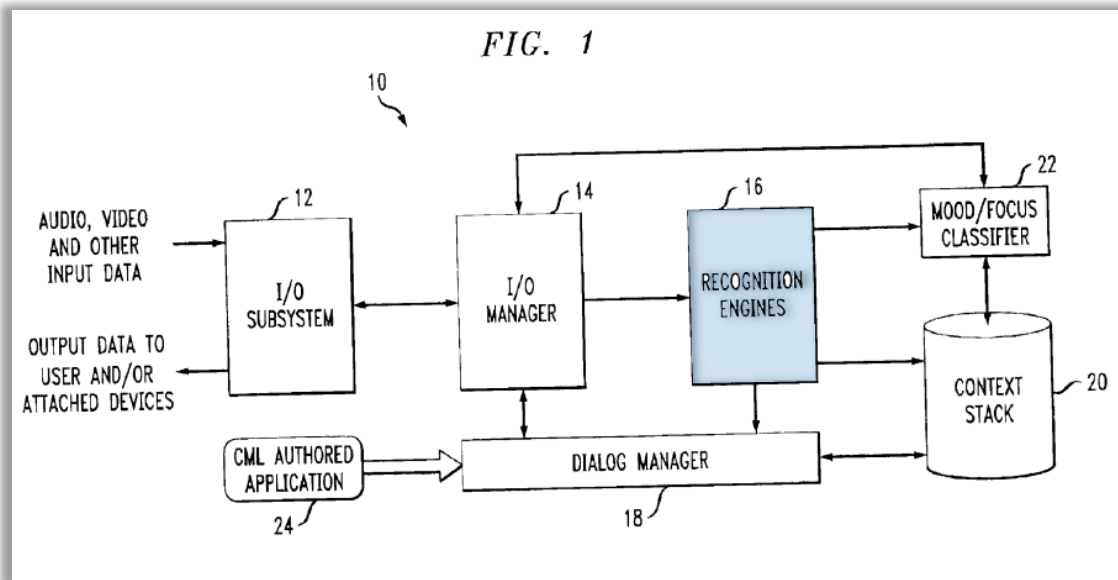
Referencing Figures 1 and 4, Maes provides additional details explaining the “*creating*” step: Maes describes Figure 4 as “an audio-visual speech recognition module that may be employed as **one of the recognition modules⁵ of FIG. 1 to perform speech recognition using multi-modal input data received** in accordance with the invention.” Ex.1005, 10:50–52; Ex.1003, ¶137.

In Figure 4, the auditory feature extractor 414 “receives an audio or speech signal and, ... extracts spectral features from the signal at regular intervals[,] ... in the form of acoustic feature vectors (signals)” Ex.1005, 11:64–12:2.

⁵ Maes uses the term “block” and “module” interchangeably.



Ex.1005, 11:55–64 and Figure 4 (showing audio information processing path (Path I) annotated in blue and acoustic feature vectors denoted by the letter “A”).



Ex.1005, Figure 1 (annotated); Ex.1003, ¶138.

Maes explains in detail how acoustic features are extracted. Ex.1005, 12:12–28; Ex.1003, ¶139. Maes also describes sampling the signal prior to extracting the acoustic feature vectors: “Before acoustic vectors are extracted, the speech signal may be sampled at a rate of 16 kilohertz (kHz).” Ex.1005, 12:2–12; Ex.1003, ¶140.

Maes provides several examples of acoustic feature vectors (or alternatively cepstral vectors) that may be extracted: e.g., [linear prediction coefficients] LPC cepstra, [Perceptual Linear Prediction] PLP, MEL cepstra, and that the invention is not limited to any particular type. Ex.1005, 23:40–43, 39:55–58; Ex.1003, ¶141.

Referencing the in-vehicle example, Maes explains the “*creating*” step: Maes’s system receives the raw multi-modal data relating to an I/O event and **“abstracts the data into a form that represents ... a spoken utterance ... [by]**

generalizing details associated with all or portions of the input data ... to yield a more generalized representation of the data for use in further operations.”

Ex.1005, 6:21–38. Maes’s system produces decoded text/script corresponding to the user’s spoken utterance, for example, of saying “turn it on.” Ex.1003, ¶142.

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

The “*identified user*” is the result of the “*identifying ...*” step explained in [19.2]. Ex.1003, ¶145.

Initially, the 468 Patent describes that a “*cognitive model*” as “a model derived from a user’s interaction pattern with the system[,] ... may be used to predict what actions the user may take next in time [and] assisting with speech recognition and/or question or command recognition.” Ex.1001, 24:1–11; Ex.1003, ¶146.

As Dr. Houh explains, cognitive models using information derived from user (or users’) historical interactions with a speech recognition system were well-known in the art. Ex.1003, ¶¶147–149 (citing Ex.1008, Abstract, [0002], [0016]; *see also* [0024]–[0026]).

Maes teaches models utilizing user attributes. For example, Maes teaches “a dynamic classification module ... which performs dynamic **classification of the user**” “based on at least one user attribute ... determined ... from the user’s speech.”

Ex.1005, 44:5–11. Examples of attributes for the user can be gender, age, accent, native language, dialect, socioeconomic classification, educational level, and emotional state. Ex.1005, 39:50–55; Ex.1003, ¶149.

Considering, for instance, native language as a user attribute, Maes explains that “information about the native language of the speaker can be added to define each accent/native language model” and further that “at the creation of the models for each native language, one employs a speaker or speakers who are tagged with that language as their native language.” Ex.1005, 40:18–23. Maes teaches classifying the user’s native language based on a **native language model** created using a speaker who is tagged with that language as his/her native language. Maes’s native language model, therefore, serves as an example of “*a personalized cognitive model derived from one or more prior interactions between the identified user and*” “the (I/O) subsystem 12 ... compris[ing] one or more microphones for capturing audio input data” (“*the conversational voice user interface*”). Ex.1005, 6:5–8; Ex.1003, ¶150.

Furthermore, Maes explains using methods based on “decision trees trained on word-dependent features (amount of **training utterances**, minimum and average triphone occurrences, **occurrence in language model training**, ... as well as **speaker profile features (accent, dialect, gender, age, speaking rate, identity)**.” Ex.1005, 19:3–20. As Dr. Houh explains, Maes’s teachings related to training

utterances, occurrence of certain utterances in language model training, and using speaker profile or—speaker-specific features (e.g., accent, dialect, gender, age, speaking rate, identity) further confirm Maes teaches or suggests “*a personalized cognitive model ...*” Ex.1005, 19:3–20 and 6:5–8; Ex.1003, ¶151.

In Maes’s system, “the input acoustic data is compared to determine if the data matches closely enough with stored models. If the comparison yields a close enough match, the person uttering the speech is verified. The match is accepted or rejected by comparing the match with competing models. These models can be selected to be similar to the claimant speaker or be speaker independent (i.e., a single or a set of speaker independent models). ... [A]t enrollment, **the input speech is collected for a speaker to build the mixture gaussian model M_i that characterize each speaker.**” Ex.1005, 24:31–45. Maes explains that “the above approach finds the closest matching model and the person whose model that represents is determined to be the person whose utterance is being processed.” Ex.1005, 24:28–30; Ex.1003, ¶152.

To the extent PO contends that Maes does not teach or suggest that “*the semantic knowledge-based model includes a personalized cognitive model ...*,” a POSITA would have found it obvious to modify Maes’s “semantic module” (comprising a “*semantic knowledge-based model*”) to include Maes’s teachings of using user-specific data for making personalized interpretations of a user’s

speech/audio input based on using such data as a model within Maes's "semantic module." Ex.1005, 6:5–15, 19:3–20, 24:28–45, 40:18–23, 44:5–11, 41:59–64. A POSITA would be motivated to make this modification because it would improve the capabilities of the semantic module functioning to indicate meaningless/irrelevant decoding candidates in a given user's speech by adding user-specific keywords, variables, parameters, sentences, words, and phrases, thereby enhancing the quality of the transcribed speech output. Ex.1005, 41:61–64 and 11:42–45. Further, because the modification is based on Maes's own teachings, a POSITA would have had a reasonable expectation of success in doing so and without undue experimentation. Ex.1003, ¶154.

To the extent PO contends that Maes does not teach or suggest "*a cognitive model*," Ittycheriah (incorporated in Maes) does. See §VII.B; Ex.1003, ¶156.

Ittycheriah teaches making use of user-specific data (e.g., scores and words decoded in accordance with HMMs from previous interactions with a user) stored in a user-specific partition within a database as hypothesis for decoding a word in a current interaction with that user and therefore—such database used as a model provides functionality of "*a personalized cognitive model*." Ittycheriah teaches its cache database as follows:

When the user first inputs a name or a command, the acoustic front-end ... generates a set of acoustic features which are then provided to the classical speech

recognition engine The word is decoded in accordance with the Hidden Markov models and the decoded word, along with the different scores (discussed above) associated with the word, are stored in the cache database ... preferably partitioned for that particular user. [R]ather than a large cache database partitioned for each system user, individual cache databases may be formed for each user.

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

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

Ex.1003, ¶157.

Dr. Houh explains, based on Ittycheriah’s teachings, that a POSITA would have found it obvious to modify Maes’s system (e.g., the semantic module) to include Ittycheriah’s teachings of a cache database storing scores and words decoded in accordance with HMMs from previous interactions with a user (“*a personalized cognitive model*”) and Maes’s input/output (I/O) subsystem, such that the database

is used as a model within the semantic module. Ex.1003, ¶158 (citing Ex.1005, 12:12–28; Ex.1006, 3:56–60, 4:19–53). *See* §VII.B).

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

Maes teaches “decisions dynamically returned by **the models.**” Ex.1005, 43:53–54; Ex.1003, ¶162.

Maes’s system includes “a dynamic classification module ... which performs dynamic classification of the user” such that the system is “configured to modify behavior of the voice system ... based on at least one user attribute ... based on at least one acoustic feature extracted from the user’s speech.” Ex.1005, 44:5–11. Maes provides several examples of user attributes such as gender, age, accent, native language, and others. Ex.1005, 39:50–55; Ex.1003, ¶163.

Considering native language as an attribute, Maes teaches that “at the creation of the models for each native language, one employs a speaker or speakers who are tagged with that language as their native language.” Ex.1005, 40:18–23. Thus, the semantic module classifies the user’s native language based on a native language model created using multiple speakers who are tagged with that language as their native language, and thus, the native language model serves as “*a general cognitive model*” Ex.1003, ¶164.

In addition, “[s]peaker verification may be performed in a similar manner, ... the input acoustic data is compared to determine if the data matches closely enough with stored models. ... These models can be selected to be similar to the claimant speaker or be speaker independent (i.e., a single or a set of speaker independent models).” Ex.1005, 24:31–45. “Thus, the above approach finds the closest matching model and the person whose model that represents is determined to be the person whose utterance is being processed.” Ex.1005, 24:28–30. Therefore, Maes provides another example of a multi-speaker model. Ex.1003, ¶165.

To the extent one contends that Maes does not disclose that “*the semantic knowledge-based model includes ... a general cognitive model ...*,” a POSITA would have found it obvious to modify Maes’s “semantic module” to include Maes’s teachings of multi-speaker models. Ex.1005, 24:28–45, 40:18–23, 44:5–11, and 39:50–55. A POSITA would be motivated to make these modifications because it would improve the capabilities of the semantic module functioning to indicate meaningless/irrelevant decoding candidates in a given user’s speech by adding keywords, variables, parameters, sentences, words, and phrases collected from a pool of users, thereby improving the quality of the transcribed speech output. Ex.1005, 41:61–64 and 11:42–45. Further, because these modifications are based on Maes’s own teachings, a POSITA would have had a reasonable expectation of success in doing so. Ex.1003, ¶166.

Additionally, Ittycheriah teaches a speaker classification approach (“*a general cognitive model*”) that involves multi-user database storing keywords and scores from past history of a class of speakers (*derived from one or more prior interactions ... [with] a plurality of users and ... [a device]*”). Ex.1006, 5:5–19. A POSITA would have found it obvious to modify Maes’s system (e.g., the semantic module) to include Ittycheriah’s teachings such that its database is used as a model within the semantic module. See §VII.B. Ex.1003, ¶168.

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

Maes teaches “decisions dynamically returned by the **models**,” which explicitly discloses that Maes uses more than one “*model*.” Ex.1005, 43:53–54; Ex.1003, ¶173.

Maes teaches or suggests “*an environmental model derived from an environment of the identified user*” because Maes teaches “**modeling of background noise ... when no speech is detected, silence is detected**” because “if the surrounding environmental noise level is particularly high, ... there is ... a chance that ... the acoustic decoding ... contain errors.” Ex.1005, 22:9–12 and 19:62–66. This disclosure is analogous to the 468 Patent, which describes an “*environmental model*” as “includ[ing] information associated with the user’s environment and

surroundings. [E.g.,] [t]he information may include the type of environment that a user is in (e.g., quiet or noisy).” *See generally* Ex.1001, 24:22–36; Ex.1003, ¶172.

Dr. Houh explains that Maes notes a problem of conventional speech recognition systems in their “inability to discriminate between ... background noise or background speech not intended to be decoded, and speech that is intended to be decoded.” Ex.1003, ¶174 (citing Ex.1005, 20:23–29).

As Dr. Houh explains, Maes’s system detects/senses activity associated with “*an environment of the ... user.*” Ex.1003, ¶175. Referencing the in-vehicle system example, Maes’s system “detect[s] a distracted or sleepy operator based on ... speech that indicates distraction or sleepiness (by audio input) ...” or “detect[] that the user is warm or cold” and accordingly “cause the temperature to be adjusted for each passenger. If the passenger is tired, the system may cause the adjustment of the seat, increase the music volume, ...” Ex.1005, 5:24–28 and 4:29–36; Ex.1003, ¶175.

As Dr. Houh further explains, Maes’s disclosures related to modeling of a background and/or noise in “*an environment of the identified user*” also serve as describing “*an environmental model.*” For example, in broadcast news, Maes’s system processes arbitrary content video including arbitrary background objects in a noisy environment. Ex.1005, 11:8–12. By “capturing” environmental data comprising “video input data representing the environment in which [the system] ...

is deployed,” such data enables Maes’s system to “locate and track a face and, more particularly, a mouth, to determine what is relevant visual information” within an arbitrary background. Ex.1005, 5:60–64, 11:15–18. Ex.1003, ¶¶176–178 (citing Ex.1005, 5:60–64, 6:1–4, 11:8–18 and 13:3–8; Ex.1017, 406).

To the extent PO contends that Maes does not teach or suggest that “*the semantic knowledge-based model includes ... an environmental model ...*,” Maes’s description of using data representing the user’s environment and surroundings was well-known, as Dr. Houh explains. A POSITA would have found it obvious to modify Maes’s “semantic module” (comprising a “*semantic knowledge-based model*”) to include Maes’s teachings of using such data in making enhanced interpretations of a user’s speech/audio input, as a model within Maes’s semantic module. Ex.1005, 6:21–29, 13:3–8, 4:29–36, 41:59–64. A POSITA would be motivated to make this modification because it would improve the capabilities of the semantic module functioning to indicate meaningless/irrelevant decoding candidates in a given user’s speech by adding user-specific keywords, variables, parameters, sentences, words, and phrases collected from a user’s own environment (which is likely to be pertinent in determining a user’s intent), thereby improving the quality of the transcribed speech output. Ex.1005, 41:61–64 and 11:42–45. Further, because the modification is based on Maes’s own teachings, a POSITA would have had a reasonable expectation of success in doing so. Ex.1003, ¶¶179–180.

Maes's afore-mentioned teachings are consistent with PO's description of "*an environmental model*" to "information concerning the environment of the user" in its Complaint. Ex.1003, ¶181 (citing Ex.1007, ¶119).

Claim 19 requires that "*the semantic knowledge-based model includes a personalized cognitive model ..., a general cognitive model ..., and an environmental model,*" and for the reasons explained in [19.3.0], [19.3.1], [19.3.2], and [19.3.3], collectively, Maes teaches or suggests this element. Ex.1003, ¶183.

Moreover, the 468 Patent provides no description of what a "*semantic knowledge-based model*" might encompass and also does not describe "a *semantic knowledge-based model [that]*"—"includes" "*a general cognitive model,*" "*an environmental model,*" and "*a personalized cognitive model.*" At best, the 468 Patent states that "*a general cognitive model,*" "*an environmental model,*" "*a personalized cognitive model*" "**coupled to**" conversational speech analyzer 804. Ex.1003, ¶184 (citing Ex.1001, 23:48–55 and Figure 8).

Additionally, other than parroting the claim language, PO's Complaint does not make clear what a "*semantic knowledge-based model*" encompasses. Ex.1003, ¶185 (citing Ex.1007, ¶¶116–119)

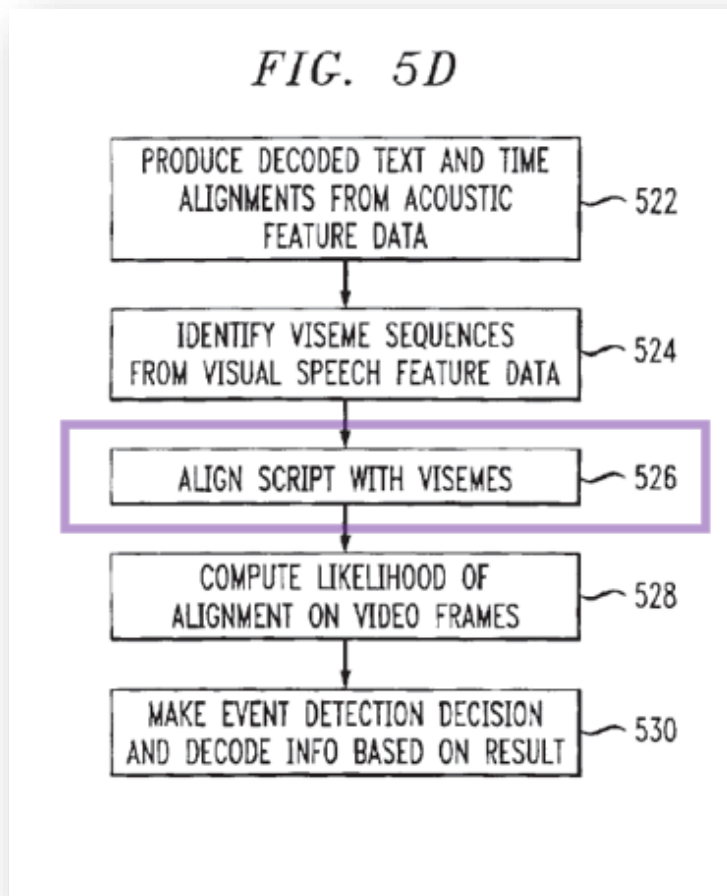
- i. **[19.4] merging the speech-based transcription and the non-speech-based transcription to create a merged transcription;**

As explained in [19.1.1] and [19.3.0], in Maes, “the visual phonemes (visemes) sequence” constitutes “*the non-speech-based transcription*” and “the decoded text” (or, script) constitutes “*the speech-based transcription.*” Ex.1003, ¶187.

Referencing step 526 of Figure 5D, Maes discloses:

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

Ex.1005, 21:51–54.

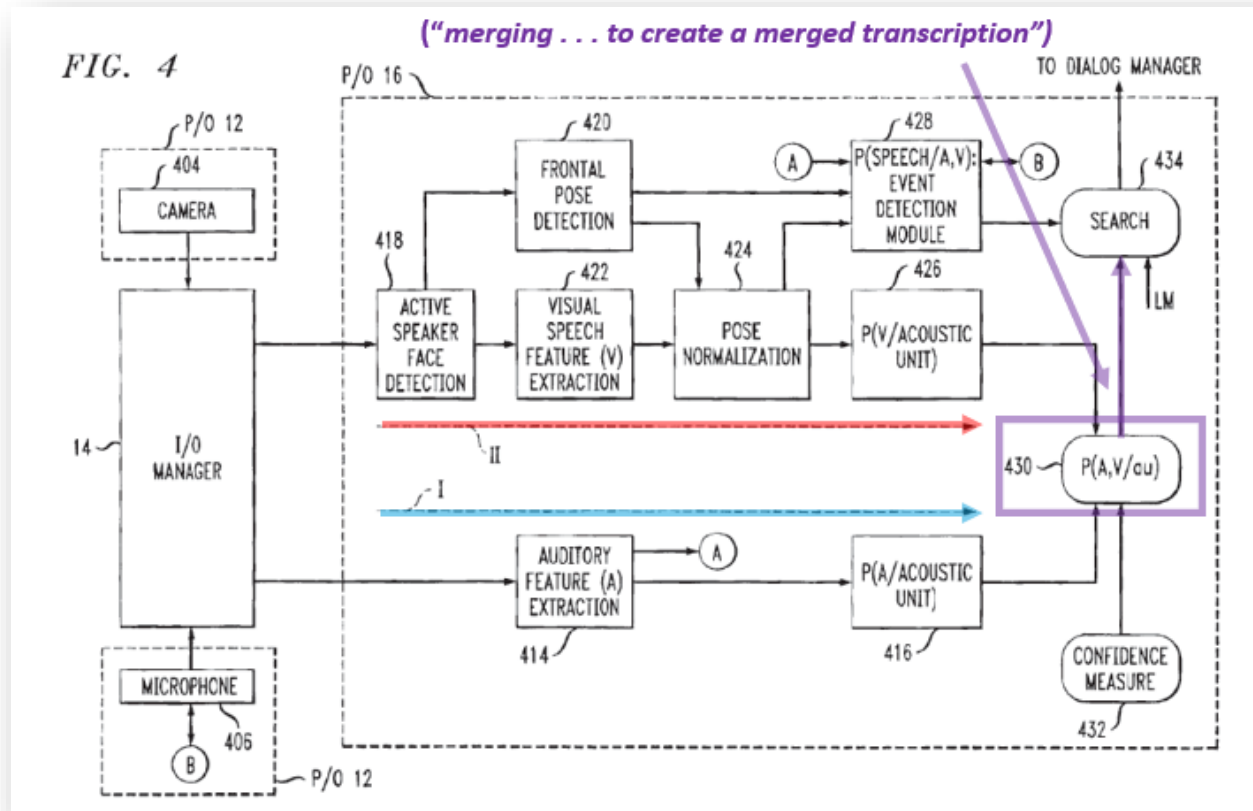


Ex.1005, Fig. 5D (block 526); Ex.1003, ¶190.

By teaching synchronization of the two information streams (the audio and video information streams) or, specifically, alignment of “the decoded text” (or, script) (“*the speech-based transcription*”) with “the visual phonemes (visemes) sequence” (“*the non-speech-based transcription*”), Maes discloses the “*merging ...*” step. Maes explains that aligning/synchronizing the audio and video information streams has “many advantages,” e.g., detecting errors in decoding. Ex.1005, 29:30–43; Ex.1003, ¶¶191–192.

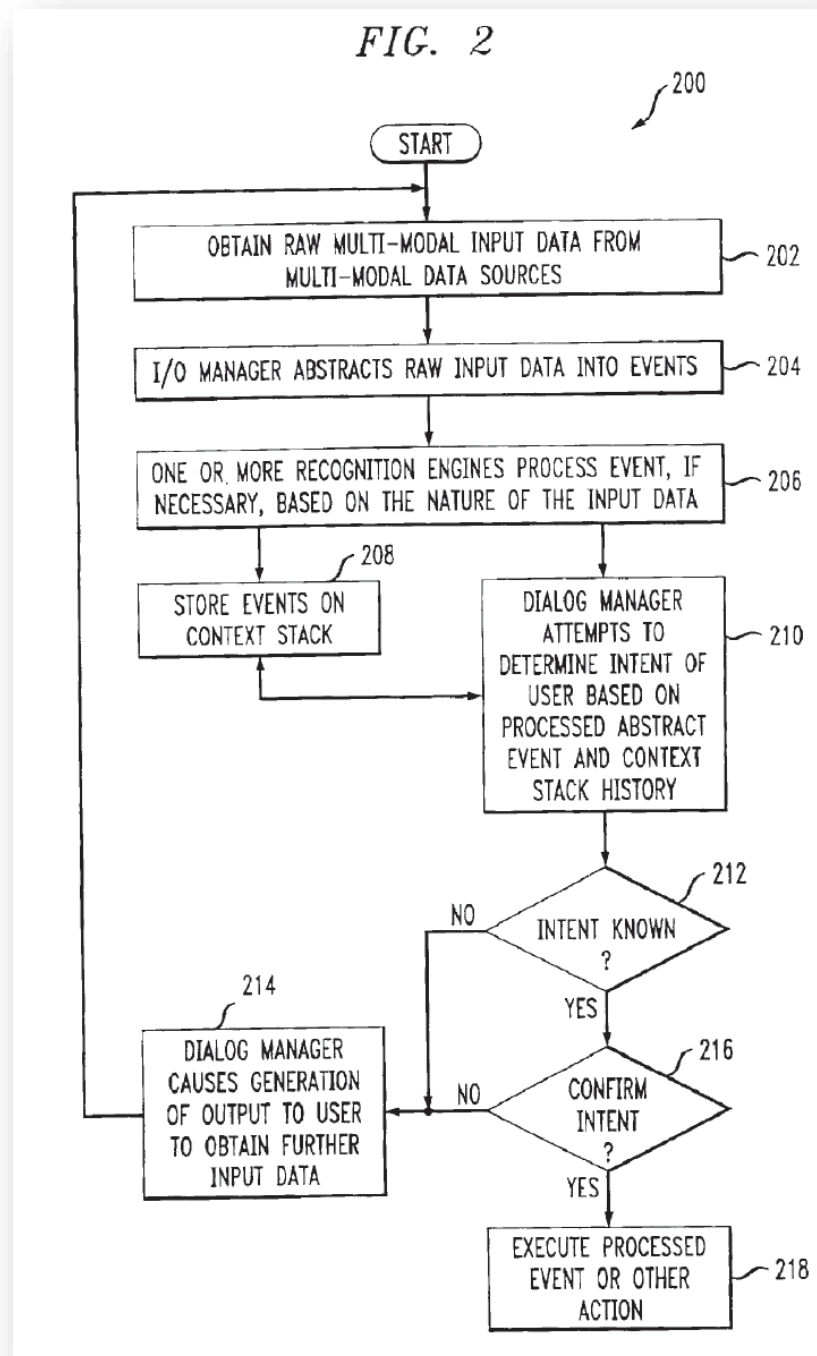
Referencing the in-vehicle system example, the aligned decoded script and visemes sequence created from merging the transcriptions includes the results of the recognized events associated with the spoken utterance “turn it on” with the accompanying lip movement, therefore, constitutes “*a merged transcription*.” Ex.1003, ¶193 (citing Ex.1005, 7:65–8:4, 8:37–42).

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



Ex.1005, Figure 4 (annotated); Ex.1003, ¶194.

Maes explains how the audio signal and video signal are synchronized to generate “*a merged transcription.*” For example, Maes describes Figure 2 (below) as “a flow diagram illustrating a referential ambiguity resolution methodology performed by ...[its] system.” Ex.1005, 3:24–28.



Ex.1005, Figure 2. Referencing Figure 2, Maes explains that “in step 202, raw multi-modal input data ... obtained from multi-modal data sources associated with the

system is ... abstracted into a form that represents one or more events, e.g., “a spoken utterance,” and “the abstracted data or event is then sent ... to one or more recognition engines ... to have the event recognized For example, ... **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, 6:28–50; Ex.1003, ¶¶195–196 (citing Ex.1005, 6:38–50).

One of the ways in which the **audio-visual speech recognition engines** processes or otherwise, recognizes event data is by generating an outcome (*i.e.*, “*merged transcription*”) in which the transcribed audio information (e.g., the decoded text/script) is aligned or synchronized with the transcribed video information (e.g., the visemes sequence). Ex.1005, 21:50–54 (“**[T]he script is aligned with the visemes.** A rapid (or other) alignment may be performed in a conventional manner ... to attempt to **synchronize the two information streams.**”), 21:59–61 (“a likelihood on the alignment is computed to determine how well **the script aligns to the visual data**”), 7:22–32 (**two, more or even all of the input modes described herein may be synchronized**” via the techniques disclosed in U.S. App. No. 09/507,526); Ex.1003, ¶¶197–198 (citing Ex.1005, 7:22–32).

Referencing the in-vehicle example, as Dr. Houh explains, the “*merging ...*” step would result in the system aligning/synchronizing decoded text/script processed

from the spoken utterance of saying “turn it on” with the visemes sequence processed from a video of the lip movement of the user. Ex.1003, ¶200.

Therefore, Maes teaches or suggests this claim element at least for this reason. Ex.1003, ¶201.

Additional Mapping

As Dr. Houh explains, in Maes, keyboard-typed inputs and GUI-based input constitute additional examples of “*a non-speech input*” and further, each of these examples serves as “*the non-speech-based transcription.*” Ex.1003, ¶202. As Dr. Houh further explains, Coffman teaches the “*merging ...*” step because Coffman discloses “merg[ing]” input streams. For example: “combine input keyed on the keyboard, handwritten input and speech input” which is consistent with the description in PO’s Complaint. Ex.1003, ¶¶202–204 (citing Ex.1012, 38:2–4 and 38:8–19; Ex.1007, ¶120).

Therefore, Maes renders obvious this claim element for this additional reason. Ex.1003, ¶205.

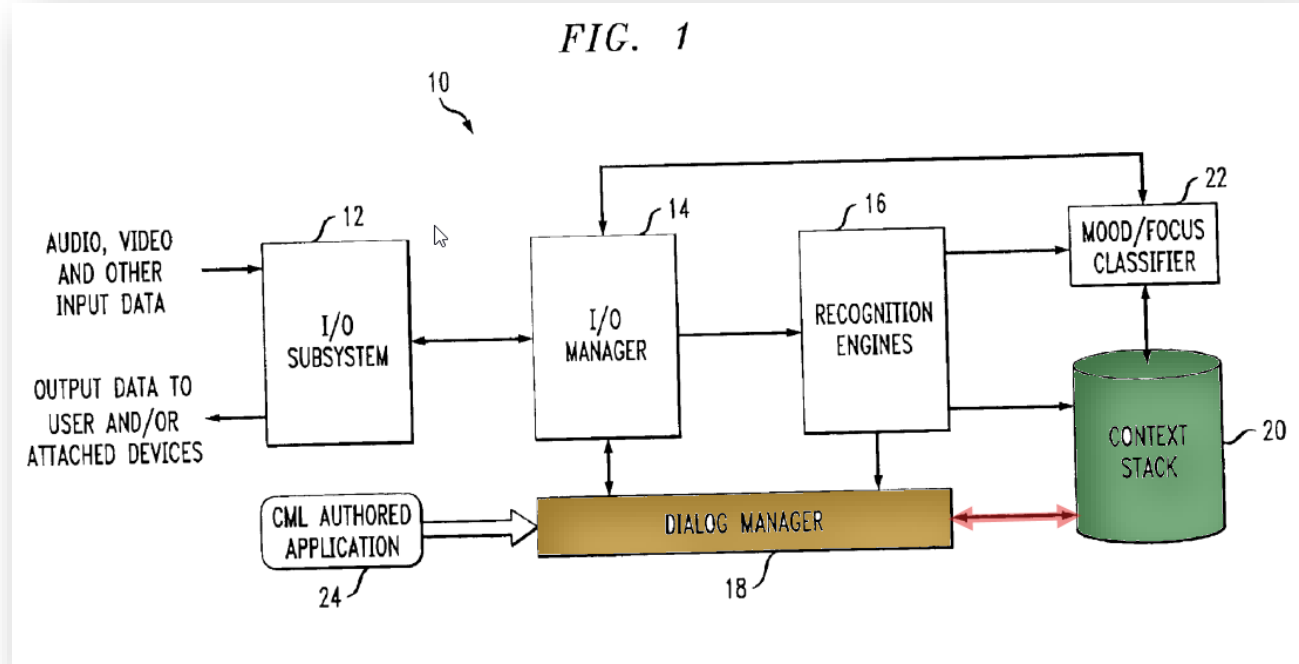
j. [19.5] identifying one or more entries in a context stack matching information contained in the merged transcription;

Maes explains storing data relating to “historical information” e.g., past input/output (I/O) events (“*one or more entries*”) in a context stack (such as context stack 817 in Figure 8 or context stack 20 in Figure 1) (“*in a context stack*”):

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

Ex.1005, 7:62–63, 37:60–61; Figure 1 (below); *see also* 7:40–45 (“the system ... determine[s] the user intent based on the current event and the historical interaction information **stored in the context stack**”); 5:16–19 (“**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.1003, ¶208.

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



Ex.1005, Figure 1 (annotated); Ex.1003, ¶209.

Referencing the in-vehicle example, Maes describes its system identifying words in a recognized utterance from contexts—such as by matching the utterance “turn it on” to the context “radio”:

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

Ex.1005, 7:65–8:4.

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

Ex.1005, 8:37–42; Ex.1003, ¶210.

In this example, the results of the recognized events constitute “*information*,” which is included in the aligned decoded script and visemes sequence associated with the spoken utterance “turn it on” with the accompanying lip movement (“*in the merged transcription*”). Further, in the above example, “this event” corresponds to the current event of the user pointing to the radio, which occurred few seconds after “the previous spoken utterance event” of the user saying “turn it on.” Data for both events (“*one or more entries*”) are stored on the context stack (“*in a context stack*”) because Maes discloses “**all the 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.

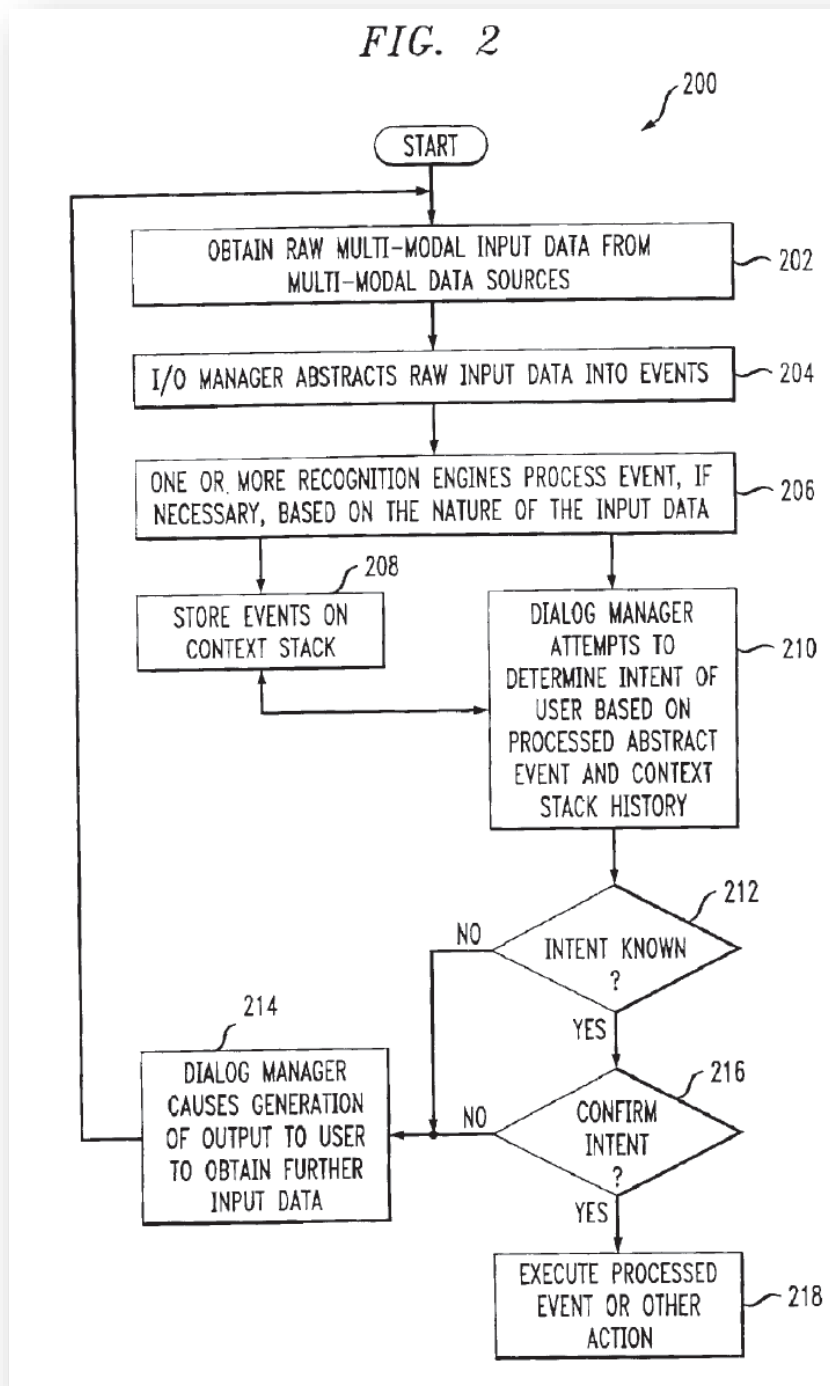
Maes clarifies that “the dialog” includes “**conversational dialog comprising speech** and other multi-modal I/O such as GUI keyboard, pointer, mouse, **as well as video input.**” Ex.1005, 36:49–54; Ex.1003, ¶211.

Therefore, Maes discloses the dialog manager associating the results of the recognized events (“*information*”) included in the aligned decoded script and visemes sequence produced as a result of merging the transcriptions (“*contained in the merged transcription*”) with an organized/sorted context in a context stack such as context stack 817 in Figure 8 or context stack 20 in Figure 1 corresponding to an active dialog which serves as the “*identifying ...*” step. Ex.1003, ¶211.

And, the associated context(s) stored in the context stack corresponding to the user’s intent in the active dialog constitutes the “*one or more entries in a context stack*” that “*match[]*” (“*matching*”) the results of the recognized events (“*information*”) included in the aligned decoded script and visemes sequence produced as result of merging the transcriptions (“*contained in the merged transcription*”) associated with the spoken utterance, e.g., “turn it on” with the accompanying lip movement. Ex.1003, ¶211.

Dr. Houh explains Maes teaches identifying matches between associated context(s) stored in the context stack corresponding to the user’s intent in the active dialog (or, simply contexts stored on the context stack) and results of recognized events (or, simply transcribed user inputs). Ex.1003, ¶212.

Further, Maes Figure 2 depicts “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. 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; Ex.1003, ¶213.

To the extent PO contends that Maes does not teach or suggest the "*identifying ... ,*" step, Coffman does. See § VII.B; Ex.1003, ¶215.

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

Based on the modality (pointer, keyboard, file, speech), the task dispatcher 402 redirects the stream to the appropriate conversational subsystems or conventional subsystem with **speech inputs being transcribed/understood. The output of these**

subsystems is run down the context stack 405 to extract the active query and complete it.

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

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

Ex.1012, 42:1–3. Thus, Coffman's teachings of searching the context stack until an acceptable match is found between the transcribed user inputs and data stored on the context stack serve as the “*identifying* ...” functionality. Indeed, Coffman teaches updating the context stack in a manner such that “[w]hen a message is returned to the user, the context is updated and then pushed to the top of the context stack ... under the active context. The active context is updated to inform of the existence of a returned value.” Ex.1012, 42:7–9. This additionally confirms that Coffman provides the “*identifying*” functionality. Ex.1003, ¶216.

Accordingly, the results of the recognized events (“*information*”) included in the aligned decoded script and visemes sequence produced as a result of merging the transcriptions (“*contained in the merged transcription*”) produced as a result of

processing the user's spoken utterance, as Maes teaches, would be compared against data stored on the context stack, per Coffman's teachings, to “*identify[] one or more contexts that completely or partially match one or more text combinations contained in the transcription.*” Ex.1012, 41:1–5, 40:15–19; Ex.1005, 7:60–8:4, 8:37–42, and 37:53–55; Ex.1003, ¶216. A POSITA would be motivated to make this modification for the reasons above (*see* § VII.B).

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

As explained in [19.5], Maes teaches “*identifying ... entries in a context stack matching information contained in the merged transcription,*” and as such, in Maes, the associated context(s) stored in the context stack corresponding to the user's intent in the active dialog constitutes “*the identified entries.*” Ex.1003, ¶218.

Dr. Houh explains Maes teaches scoring one or more candidate contexts such as phonemes, which Maes describes as “sub-phonetic or acoustic units of speech” (or, equivalently as “portions [of dialogs].” Ex.1003, ¶219 (citing Ex.1005, 7:60–8:4 and 12:29–33).

Specifically, Maes teaches calculating likelihood scores for acoustic feature vectors and visual speech feature vectors extracted from the multi-modal input. Ex.1003, ¶219 (citing Ex.1005, 12:38–43 and 18:46–60).

Maes then describes using the likelihood scores for “identify[ing]” (“*determining*”) “**the acoustic units** [i.e., “*a most likely context*” identified from amongst the entries in the context stack as explained in [19.5]] ... as having the highest probabilities of representing what was uttered” “characterized by acoustic feature vectors A and visual speech feature vectors V extracted from the multi-modal input (“*for the multi-modal input*”). Thus, the candidate context (such as a phoneme or an acoustic unit) that has the highest probability of representing what was uttered corresponds to “*a most likely context for the utterance.*” Ex.1005, 18:64–65 and 20:13–15. The highest-probability-candidate-context (“*a most likely context*”) is identified from (“*determin[ed] ... based on*”) one or more candidate contexts, which, as explained in [19.5], are equivalently the associated context(s) stored in the context stack corresponding to the user’s intent in the active dialog (“*the identified entries*”). Ex.1003, ¶220.

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

Initially, the 468 Patent describes a “domain agent” is used for organizing “generic and domain specific behavior and information The domain specific agents may provide complete, convenient and re-distributable packages or modules for each application area.” Ex.1001, 27:14–20. Thus, “*domain agent[s]*” in the 468 Patent broadly refer to software modules that are specific to each application area. Ex.1003, ¶223.

In Maes, the acoustic units/phonemes having the highest probabilities of representing what was uttered constitutes “*the most likely context for the multi-modal input*,” which as explained in [19.5] and [19.6], is “*based on the identified entries*” in the context stack. Ex.1003, ¶224.

To determine the user’s intent based on the multi-modal input, Maes’s system:

determine[s] and execute[s] one or more application programs that effectuate the user’s intention and/or react to the user activity. The application depends on the environment that the system is deployed in.

Ex.1005, 7:40–46. Therefore, in Maes, application programs (effectuating the user’s intention and depending on the environment that the system is deployed) are analogous to the 468 Patent’s description of “*domain agent*” stated above.

Therefore, “application programs” serve as “domain agent[s].” Maes explains this further in greater detail. Ex.1003, ¶224.

Maes describes that “the conversational dialog manager 819 determines what information the user has, which inputs the user presents, and **which application(s) should handle the user inputs**. ... The DMA function [of the dialog manager] handles multi-modal I/O events to: (1) **determine the target application or dialog (or portion of it)** and (2) use past history and context” and further, in Maes, “the historical information (e.g., past events) [are] stored in the context stack.” See Ex.1005, 36:49–37:3, 37:51–38:2; *see also* 7:60–8:19. In addition, “[t]he context stack [] comprises all the information associated with an application.” Ex.1005, 37:55-56; Ex.1003, ¶225.

By teaching that Maes’s system determines which application(s) should handle the user inputs, and further that the system determines the target application or dialog (or portion of it), Maes expressly discloses the step of “*identifying ... a domain agent*.” Ex.1003, ¶226.

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

“*associated with*” the acoustic units having the highest probabilities of representing what was uttered (“*the most likely context for the multi-modal input*”). Ex.1003, ¶226.

Moreover, for the additional reasons below, Maes (which incorporates Coffman) discloses this claim element. As explained above, Maes teaches determining the application or dialog (or portion of it) (“*domain agent*”) and furthermore, as explained in connection with [19.7], Coffman teaches determining an acceptable match (“*most likely context*”) found between the transcribed user inputs and data stored on the context stack. As Dr. Houh explains, the target application in Maes is “*associated with*” the context corresponding to the acceptable match, as taught by Coffman, to “*identify[] a domain agent associated with the most likely context for the multi-modal input,*” such as the user’s multi-modal input received by Maes’s system. Ex.1003, ¶227.

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

As explained in [19.7], “applications” in Maes are “*domain agent[s]*” and “target application[s]” constitute “*the identified domain agent.*” Ex.1005, 36:64–66; Ex.1003, ¶230.

Referencing the in-vehicle example, Maes’s system “selects the appropriate device (e.g., radio) activation routine [“*the identified domain agent*”] and instructs

the I/O manager to output a command [*“communicat[e] a request”*] to activate the radio.” Ex.1005, 8:16–19; Ex.1003, ¶231.

Further, as Dr. Houh explains, Maes explains that based on the multi-modal input, if the system determines that the application it selects will effectuate the users desire, it **launches the execution of an action ... of that application**” which additionally serves as *“communicating a request to the identified domain agent.”* Ex.1005, 8:8–16; Ex.1003, ¶232.

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

Referencing the in-vehicle example, Maes’s system “selects the appropriate device (e.g., radio) activation routine and instructs the I/O manager to output a command to activate the radio.” Ex.1005, 8:16–19. Thus, in Maes, the application “radio activation routine” (*“the identified domain agent”*) would be activated, and Dr. Houh explains that upon processing the command (*“as a result of processing the request”*), the system would “adapt some behavior/service/appliances accordingly” to *“generat[e] a response to the user”* (e.g., cause the radio to be turned on and play, for instance, entertainment, news, sports, or weather content) (*“content provided by the identified domain agent”*). Ex.1003, ¶236 (citing Ex.1005, 7:60–8:29, 2:54–64).

In Maes, an application has “associated business logic ... with which the user interacts.” Ex.1005, 2:65–67. The reaction of the business logic of an application is “*a response to the user*.” Ex.1003, ¶237.

2. Claim 25

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

As explained in [19.2], Maes teaches “*identifying the user*.”

Furthermore, Maes discloses “speaker[/user] verification using the extracted acoustic feature vectors” (“*verifying an identity of the user based on voiceprint matching*.”) Ex.1005, 23:46–48. Dr. Houh explains that a POSITA would have understood that Maes’s “acoustic feature vectors” provide the “*voiceprint*” of a user which can be matched/verified by any conventional system. Ex.1003, ¶240 (citing Ex.1030, 7:4–7, 1:6–13; Ex.1005, 23:46–58).

3. Claim 26

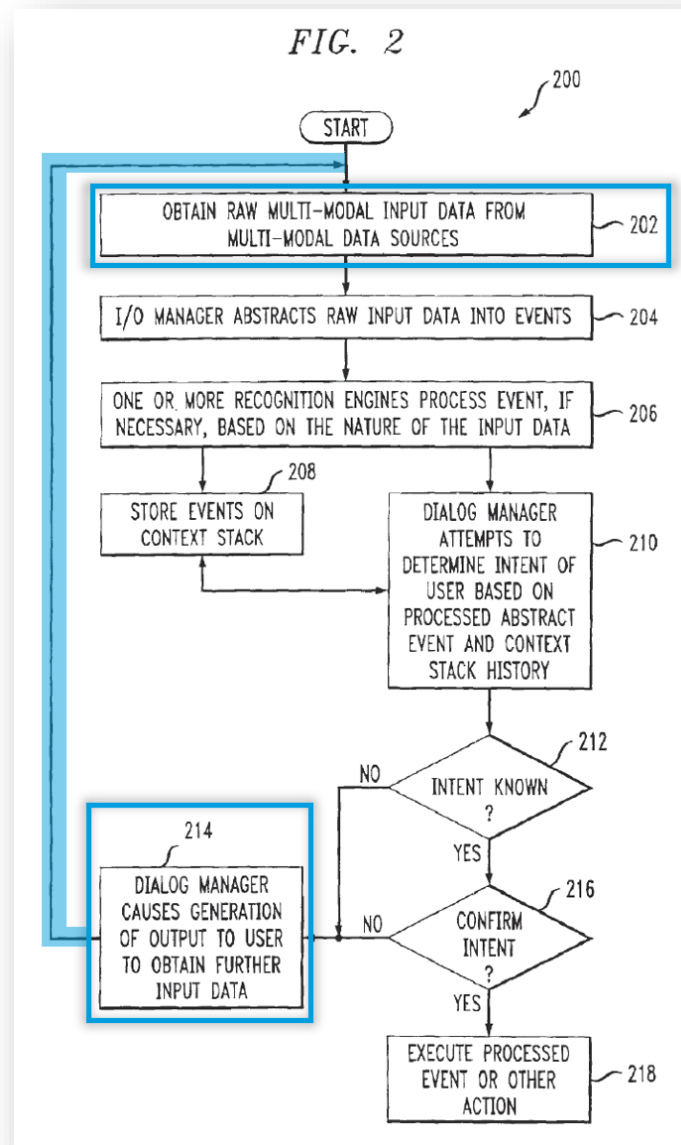
- a. **[26.0] The method of claim 19, further comprising receiving one or more follow-up multi-modal inputs at the conversational voice user interface, the follow-up multi-modal inputs including at least one of a follow-up natural language utterance or a follow-up non-speech input.**

As explained in [19.1.1], the (I/O) subsystem 12 discloses “*the conversational voice user interface*.” Ex.1003, ¶243.

Referencing Figure 2, Maes explains: “in step 214,” the system “generat[es] ... an output to the user **requesting further input data** [*“one or more follow-up multi-modal inputs”*] so that the user’s intent can be disambiguated,” “the system ... obtains the raw input data, again in step 202, and the process ... iterates based on the new data. Such iteration can continue as long as necessary ... to determine the user’s intent.” Ex.1005, 8:47–50, 8:58–65. The system may “seek confirmation in step 216 from the user in the same manner as the request for more information (step 214).” Ex.1005, 8:66–9:1. Ex.1003, ¶243.

Maes’s system, which “receiv[es] a multi-modal natural language input ...” (see [19.1.0]), therefore, applies the same receiving technique of the “further input data” (*“the one or more follow-up multi-modal inputs”*) as the originally-received/raw input data. Ex.1003, ¶243.

Referencing the in-vehicle example in which the system does not have enough original input to determine the user intent, Dr. Houh explains that the user’s feedback saying the spoken utterance “the radio” (including the accompanying lip movement) constitutes the *“one or more follow-up multi-modal inputs ... including ... a follow-up natural language utterance or a follow-up non-speech input.”* Ex.1005, 8:43–65; 36:59–37:2 (“follow[ing] up with a dialog,” “confirmation, disambiguation, correction, more details” are needed).



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

Ex.1003, ¶¶244–245.

4. Claim 27

- a. [27.0] The method of claim 26, wherein the identified domain agent updates the context stack and the

semantic knowledge-based model in response to processing the request.

iv. Domain Agent Updating the Context Stack

Maes explains use of the context stack “compris[ing] **all the information** associated with an application ... includ[ing] all the variable, states, input, output, and **queries to the backend** ... in the context of the dialog and any extraneous event that occurs during the dialog” “to create a history of interaction between the user and the system ...” Ex.1005, 7:33–45, 37:55–60. Creating a history of interaction in the context stack with **all** the information related to a dialog would require accumulating/adding data or otherwise updating the context stack over time, and therefore, Maes teaches or suggests “*updat[ing] the context stack.*” Ex.1003, ¶248.

Dr. Houh explains an example in which multiple queries to the backend are generated, which further confirms that Maes’s context stack is updated. Ex.1003, ¶249 (citing Ex.1005, 37:55–60).

Furthermore, as explained in [26.0], Maes’s Figure 2 process, working iteratively, discloses storing the processed results related to the follow-up/further multi-modal inputs (block 214) (in addition to the originally-received input) as events on the context stack (block 208). Ex.1005, 7:33–39. Storing the processed results related to the follow-up/further multi-modal inputs encompasses “*updat[ing] the context stack.*” Ex.1003, ¶250 (citing Ex.1005, 37:55–60).

Maes teaches or suggests that “*the context stack*” is “*update[d]*” by “*the identified domain agent*.” Ex.1003, ¶251.

Maes teaches “allow[ing] [multiple] access[es] to backend business logic” and further that information describing “**such accesses to be added to the context stack.**” Ex.1005, 38:3–7. Adding information to the context stack discloses “*updat[ing] the context stack.*” Additionally, because Maes teaches “an application” as having an “associated business logic [*e.g.*, backend business logic in Maes’s system] ... with which the user interacts” (Ex.1005, 2:65–67 and 38:3–7), it would have been obvious to a POSITA that the target application (corresponding to the user’s intent) updates its context stack. Ex.1003, ¶252.

Additionally, Maes describes “[t]he context stack ... compris[ing] **all the information associated with an application**” and further that “applications(s) ... handle the user inputs” by “interact[ing]” with users. Ex.1005, 2:65–67, 36:56–57, 37:55–56, and 38:3–7. Accordingly, a POSITA would have understood that “*the context stack*” comprising **all** the information (including all user inputs) would be updated by the respective/corresponding application (“*the identified domain agent*”) which handles the user inputs or otherwise interacts with the user. Ex.1005, 2:65–67, 37:56, and 38:3–7. Ex.1003, ¶253.

To the extent PO asserts Maes does not expressly disclose that “*the identified domain agent updates the context stack,*” it would have been obvious to a POSITA

to modify Maes's teachings such that applications (which have first-hand/direct knowledge of user inputs) update the context stack. Dr. Houh explains how such a modification would be an obvious design choice, provide an efficient updating method, and there would have been a reasonable expectation of success in making this modification. Ex.1003, ¶254-.

Furthermore, referencing the in-vehicle system example (Ex.1005, 8:16–19), a POSITA would have understood the radio activation routine updates the context stack of the in-vehicle system upon processing the command (“*in response to processing the request*”) to activate the radio. See [19.8], [19.9] (explaining “*as a result of processing the request*” / “*in response to processing the request*”). Ex.1003, ¶¶255–257.

v. Domain Agent Updating the semantic knowledge-based model

Dr. Houh explains that Maes teaches or suggests **training models**, such as a semantic module (comprising “*a semantic knowledge-based model*”), which serve as describing “*updat[ing]*” the semantic module (comprising “*the semantic knowledge-based model*”); see also [19.3.0]. Ex.1003, ¶259 (citing Ex.1005, 19:7–21).

Dr. Houh further explains that Maes discloses application programs having a number of “input parameters or **variables**” and referencing the in-vehicle system,

Maes provides a suggestion of a radio activation routine/application updating the semantic module with the **variables** “turn” “it” “on” so that in future interactions with the user, the system determines that the context of the user’s utterance is turning on a device in the vehicle, such as the radio. Ex.1003, ¶260 (citing Ex.1005, 7:43–46, 8:20–21, 7:60–8:22, 5:18–19); *see also* [19.8] (explaining “*the request*”) and [19.9] (explaining “*as a result of processing the request*”).

Maes teaches that its system provides “real-time feedback to the business logic unit [associated with an application (*e.g.*, “*the identified domain agent*”) with which the user interacts],” from which a POSITA would have understood that the feedback is used to update a model (*e.g.*, comprised in a semantic module) and further that the application would be performing such updates for future interactions with the user. Ex.1003, ¶261 (citing Ex.1005, 38:3–7, 44:22–25, 41:60–64, 2:65–67); *see also* Ex.1005, 36:49–37:3.

Therefore, Maes teaches or suggests “*wherein the identified domain agent updates ... the semantic knowledge-based model in response to processing the request.*” Ex.1003, ¶262.

To the extent one asserts that Maes does not expressly or implicitly disclose that “*the identified domain agent updates ... the semantic knowledge-based model,*” Dr. Houh provides an example of IBM’s ViaVoice speech recognition software describing that such a step was well within the POSITA’s knowledge, and it would

have been obvious to implement such a step in Maes's system. Ex.1003, ¶263 (citing Ex.1031, 1)

As another example, Dr. Houh explains Weissman reflecting the POSITA's knowledge, provides an example in which a search engine application (*"the identified domain agent"*) allows updating a meaning-based lexicon (*"a semantic knowledge-based model"*). Ex.1003, ¶264 (citing Ex.1025, 3:10–34, Figure 1).

Dr. Houh provides additional examples reflecting the POSITA's knowledge demonstrating that a POSITA would have been motivated to implement the step of *"the identified domain agent updat[ing] ... the semantic knowledge-based model"* in Maes. Ex.1003, ¶265 (citing Ex.1015, [0146] (describing that a semantic model "may be adapted"); Ex.1016, [0048] (explaining the need to update semantic models; Ex.1014, 1).

Dr. Houh further explains that such a modification would expand the ability of Maes's semantic module to detect relevant keywords in a user's input (such as a spoken utterance) to a given speech-enabled application, thereby improving the quality of the transcribed speech output, which is consistent with Maes's goals of "improved speech recognition." Ex.1003, ¶266 citing (Ex.1005, 11:42–45 and 41:58–64).

5. Claim 28

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

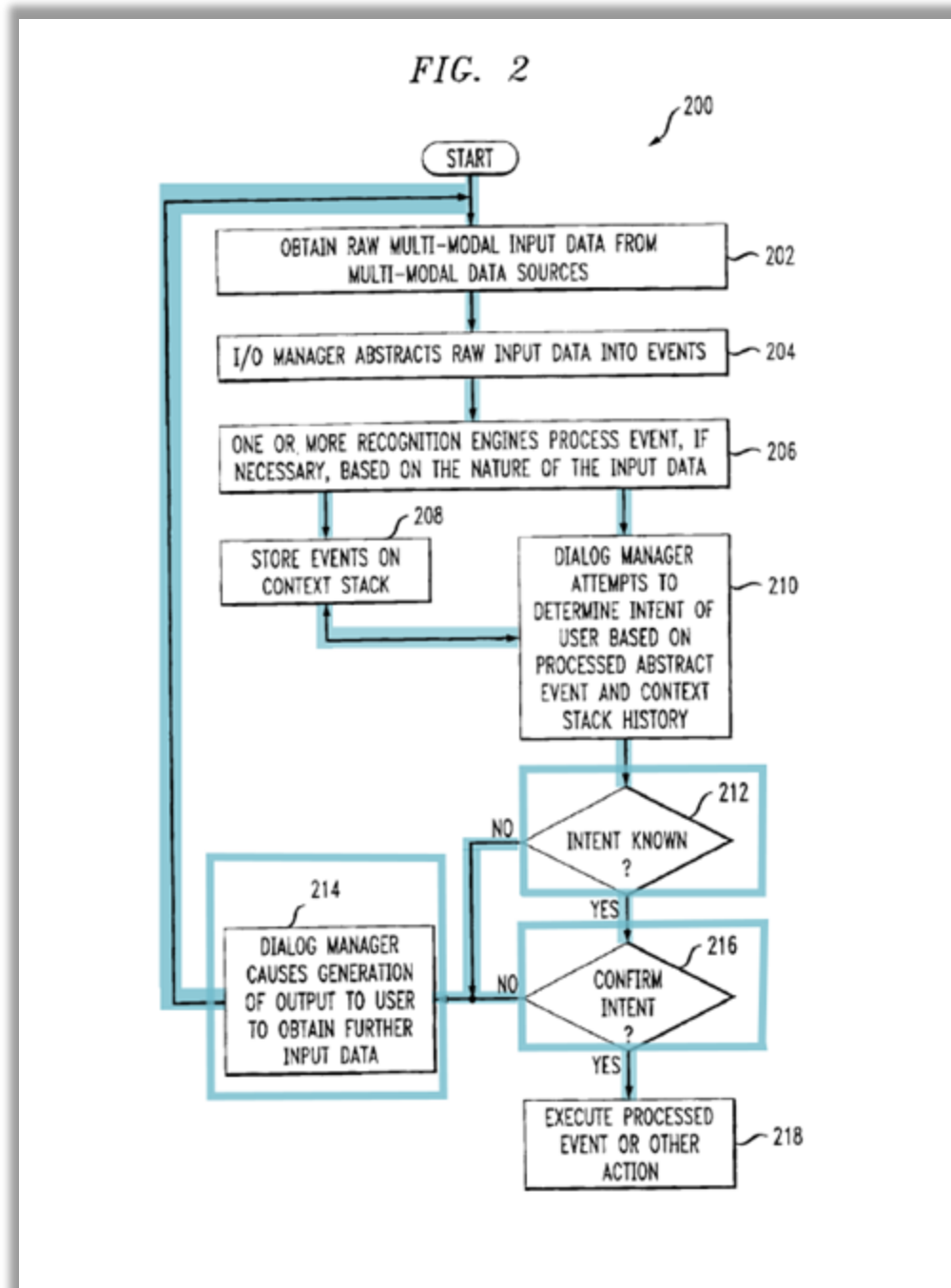
As explained in [19.3.0], Maes discloses a semantic module (comprising “a *knowledge-based semantic model*”) and speech recognition (“a *speech recognition engine*”). Maes’s semantic module “interpret[s] meaning of phrases” such that it is “used by speech recognizer 926 to indicate that some decoding candidates in a list are meaningless and should be discarded/replaced with meaningful candidates.” Ex.1005, 41:35–38 and 41:59–64. As Dr. Houh explains, Maes’s semantic module provides ambiguity resolution function or otherwise disambiguates the user’s intent through functionality that indicates certain decoding candidates are meaningless and should be discarded/replaced with meaningful candidates, which makes Maes’s “*speech recognition engine*,” “*knowledge-enhanced*.” Ex.1003, ¶¶268–269 (citing Ex.1017, 429; Ex.1005, Abstract, 4:10–11, and 36:59–63).

Therefore, Maes’s “*speech recognition engine*” and semantic module work together to provide functionality of a “*knowledge-enhanced speech recognition engine*.” Ex. 1003, ¶270.

- ii. **Determining most likely context for the follow-up input**

Referencing Figure 2, Dr. Houh explains that Maes applies the same techniques to additional input data (“*the follow-up multi-modal input*”) in the same manner as the originally-received/raw input data until the user’s intent is understood. See [19.5] and [19.6] (explaining “*determining a most likely context*” using “*entries in a context stack*”); see also [26.0] (receiving “further input data”). Ex.1003, ¶272.

Referencing the in-vehicle example, the user’s utterance of “the radio” constitutes “*the follow-up multi-modal input*” and “*a most likely context*” corresponds to the user’s intent to have the vehicle’s radio turned on. Ex.1005, 8:43–65; see also [26.0]. These techniques and features are already explained in [19.5], [19.6], [19.7], [26.0], and the iterative Figure 2 process.



Ex.1005, Figure 2 (annotated) (depicting the process iterating on the “further input data” received in step 214); Ex.1003, ¶¶273-274.

iii. Using the updated Context Stack

As explained in [27.0] and [26.0], Maes describes “*updat[ing]*” “*the context stack*” with I/O events related to “further input data” (“*the follow-up multi-modal input*”) received in step 214 of the Figure 2 process. As Dr. Houh explains, Maes’s system uses the updated context stack for storing I/O events (steps 208 and 210) during the iterative processing (Figure 2) until the user’s intent is understood. Ex.1003, ¶¶276-277; *see also* [27.0] (providing a POSITA’s motivation to use an updated context stack).

6. Claim 29

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

As Dr. Houh explains, Maes discloses the “*creating*” step (*see* [19.3.0]), updates the “*semantic knowledge-based model*” (*see* [27.0]), and it would have been obvious to a POSITA to use the updated semantic module (comprising “*updated semantic knowledge-based model*”) because the updated semantic module would have an expanded vocabulary for “*creating*” a transcription of the “follow-up ... utterance.” Ex.1003, ¶¶280–281.

7. Claim 30 and 32

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

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

Referencing the in-vehicle example, Maes teaches that the components of its system including “memory” which “may be located locally (in the vehicle)” or “remotely (e.g., connected wirelessly to the vehicle).” Ex.1005, 32:24–27; 45:30–34, Figure 10, *see generally* 45:50–60; Ex.1003, ¶285.

A target application program, such as a radio activation routine (“*the identified domain agent*”) “*processes the request*” of turning on the radio by using “instructions or code” accessing (“*querying*”) stored in **the system’s memory** (“*one or more local or network information sources.*”). Ex.1005, 8:15–19, 45:45–46; Ex.1003, ¶286.

It also “*processes the request*” of turning on the radio by “*directing a command*” (“activate the radio”) “*to*” **the radio** (“*one ... local ... device[]*”). Ex.1005, 8:15–19; Ex.1003, ¶290.

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

To the extent one asserts that Coffman and Ittycheriah are not particularly incorporated by reference into Maes as if they were effectively the same document, claims 19, 25-30, and 32 would have been obvious in view of the combination of Maes, Coffman and Ittycheriah for the same reasons articulated in Ground 1. Moreover, a POSITA would have been motivated to combine the references and had

a reasonable expectation of success for making such a combination for the reasons articulated in § VII.B. Ex.1003, ¶292.

IX. BOARD DISCRETION

A. 35 U.S.C. § 325(d)

Maes is neither cumulative nor redundant to the prosecution of the 468 Patent, and Maes was neither cited nor considered by the Examiner during prosecution of the 468 Patent. Maes incorporates by reference Coffman and Ittycheriah. Ittycheriah was not cited by the Examiner during prosecution. A “relative” of Coffman was listed on an IDS, but was also not evaluated substantively and was not the basis of a rejection. *CrowdStrike, Inc. v. Webroot Inc.*, IPR2023- 00126, Paper 9 at 14 (PTAB May 5, 2023); *Scientific Design Co. Inc. v. Shell Oil Co.*, IPR2022-00158, Paper 7 at 24-26 (PTAB Apr. 4, 2022); The first condition of *Advanced Bionics* is therefore not satisfied.

Nevertheless, the Examiner materially erred by not substantively evaluating the Maes (and/or Coffman and Ittycheriah) against the 468 Patent and further erred by failing to reject any claims as obvious over Maes as highlighted by the strength of Petitioner’s Proposed Grounds. Where a reference was never “substantively discussed by the Examiner[,]” “a petitioner’s showing that the challenged claims are unpatentable over the asserted prior art may itself be evidence of material error by the Office during prosecution.” *Quasar Science LLC v. Colt International Clothing, Inc. d/b/a Colt LED*, IPR2023-00611, Paper 10 at 14 (PTAB Oct. 10, 2023)

Moreover, because Maes was not before the Examiner, the Examiner did not have the opportunity to evaluate Maes (with or without Coffman and Ittycheriah). *Sony v. MZ Audio Scis., LLC*, IPR2022-01544, Paper No. 12, *7 (§325(d) discretion improper where “Examiner did not consider the specific combination of references asserted”). Additionally, the Examiner did not have the opportunity to review Dr. Houh’s detailed expert testimony.

B. 35 U.S.C. § 314(a)

1. Factor 1: Stay

Factor 1 is **neutral or weights against discretionary denial** because no request for stay has been filed in the litigation involving BOA or Microsoft. *Apple Inc. v. Fintiv, Inc.*, IPR2020-00019, Paper 15, 12 (May 13, 2020) (informative) (“*Fintiv*”); *Google, LLC v. Parus Holdings Inc.*, IPR2020-00847, Paper 9 at 12 (PTAB Oct. 21, 2020).

2. Factor 2: Trial Date

Microsoft is a defendant in litigation involving this patent, although has not yet been served with a Complaint so no trial date yet. Trial for RPI BOA will not occur before January 26, 2026. Any FWD can be expected before the end of September 2026, which is eight months after the currently scheduled BOA trial. “[A] court’s general ability to set a fastpaced schedule is not particularly relevant ... where, like here, the forum itself has not historically resolved cases so quickly.” *In*

re Apple Inc., 979 F.3d 1332, 1344 (Fed. Cir. 2020). Indeed, there are **13 other trials** before Judge Gilstrap currently scheduled for January 26, 2026, so it appears likely that trial will slip, Ex.1029, and much can change in the next 10 months. *Dish Network v. Broadband iTV*, IPR2020-01280, Paper 17 at 16 (PTAB Feb. 4, 2021) (“We cannot ignore the fact that the currently scheduled trial date is more than nine months away and much can change during this time.”); *see also NetNut Ltd. v. Bright Data Ltd.*, IPR2021-01492, Paper 12 (PTAB Mar. 21, 2022).

Accordingly, this factor is **neutral or weighs against** discretionary denial.

3. Factor 3: Parallel Proceeding

Factor 3 is **neutral or weighs against** discretionary denial. In the related litigation involving BOA, claim construction briefing does not start until June 17, 2025, and the Claim Construction hearing is not scheduled until July 29, 2025. The investment by the parties to date has been relatively minimal.

4. Factor 4: Overlapping Issues

Consistent with *Sotera Wireless* Petitioner and BOA stipulate that if the PTAB institutes *inter partes* review of this proceeding, neither Petitioner nor BOA will pursue in any related district court case the grounds that were raised in this Petition, nor any other grounds that could have been reasonably raised in this Petition. *Sotera Wireless, Inc. v. Masimo Corp.*, IPR2020-01019, Paper 12 at 13-14 (PTAB Dec. 1, 2020) (precedential as to § II.A).

Factor 4 **weighs strongly against** discretionary denial.

5. Factor 5: Overlapping Parties

Factor 5 is **neutral or weighs against** discretionary denial. Microsoft provides software accused of infringement of the challenged patent in related litigation involving BOA. As *Fintiv* demonstrates, this factor should play a role only where a Petitioner is unrelated to a defendant in a district court proceeding. See *Fintiv*, IPR2020-00019, Paper 11 at 13–14 (PTAB Mar. 20, 2020).

6. Factor 6: Strength/Other Considerations

“[I]n circumstances where the Board determines that the other *Fintiv* factors 1–5 do not favor discretionary denial, the Board shall decline to discretionarily deny under *Fintiv* without reaching the compelling merits analysis.” *CommScope Techs. LLC v. Dali Wireless, Inc.*, IPR2022-01242, Paper No. 23 at 4–5 (PTAB Feb. 27, 2023).

Here, the merits of the Petition are particularly compelling—for example, Maes (and/or Coffman and Ittycheriah) is a strikingly similar disclosure to that of the 468 Patent, is directed at the same problem, proposes the same solutions, and discloses nearly identical functionality as the claimed “*method for processing multi-modal natural language inputs*.” The evidence presented, if unrebutted, would lead to a conclusion that one or more claims are unpatentable by a preponderance of the evidence.

X. CONCLUSION

Claims 19, 25-30, and 32 of the 468 Patent are unpatentable for the reasons discussed above.

Dated: March 18, 2025

Respectfully Submitted,

/Scott M. Border/

Scott M. Border

Reg. No. 77,744

Winston & Strawn LLP

1901 L Street, N.W.

Washington, D.C. 20036

T: 202-282-5054

sborder@winston.com

CERTIFICATE OF COMPLIANCE

I hereby certify that this brief complies with the type-volume limitations of 37 C.F.R. § 42.24, because it contains 13,721 words (as determined by the Microsoft Word word-processing system used to prepare the brief and including annotated figures), excluding the parts of the brief exempted by 37 C.F.R. § 42.24.

Dated: March 18, 2025

Respectfully Submitted,

/Scott M. Border/

Scott M. Border

Reg. No. 77,744

Winston & Strawn LLP

1901 L Street, N.W.

Washington, D.C. 20036

T: 202-282-5054

sborder@winston.com

CERTIFICATE OF SERVICE

Pursuant to 37 C.F.R. § 42.6(e), I hereby certify that on this 18th day of March, 2025, I caused to be served a true and correct copy of the foregoing and any accompanying exhibits by Federal Express Priority Overnight on the following:

David Gerasimow
The Law Offices of David A. Gerasimow, P.C.
211 W. Wacker Dr.
Ste. 1717
Chicago, IL 60606

A courtesy copy of this Petition and supporting material was also served on litigation counsel for Patent Owner via email:

Garland Stephens (garland@bluepeak.law)
Justin Constant (justin@bluepeak.law)
Richard Koehl (richard@bluepeak.law)
Kate Falkenstien (kate@bluepeak.law)
Heng Gong (heng@bluepeak.law)
BLUE PEAK LAW GROUP LLP
3139 West Holcombe Blvd. PMB 8160
Houston, TX 77025

Dated: March 18, 2025

Respectfully Submitted,

/Scott M. Border/

Scott M. Border

Reg. No. 77,744

Winston & Strawn LLP

1901 L Street, N.W.

Washington, D.C. 20036

T: 202-282-5054

sborder@winston.com

CLAIM LISTING

Claim 19	
[19.0]	A method for processing multi-modal natural language inputs, comprising:
[19.1.0]	receiving a multi-modal natural language input at a conversational voice user interface, the multi-modal input including a natural language utterance and a non-speech input provided by a user, ...
[19.1.1]	... wherein a transcription module coupled to the conversational voice user interface transcribes the non-speech input to create a non-speech based transcription;
[19.2]	identifying the user that provided the multi-modal input;
[19.3.0]	creating a speech-based transcription of the natural language utterance using a speech recognition engine and a semantic knowledge-based model,
[19.3.1]	wherein the semantic knowledge-based model includes a personalized cognitive model derived from one or more prior interactions between the identified user and the conversational voice user interface,
[19.3.2]	[wherein the semantic knowledge-based model includes ...] a general cognitive model derived from one or more prior interactions between a plurality of users and the conversational voice user interface,
[19.3.3]	[wherein the semantic knowledge-based model includes ...] an environmental model derived from an environment of the identified user and the conversational voice user interface;
[19.4]	merging the speech-based transcription and the non-speech-based transcription to create a merged transcription;

[19.5]	identifying one or more entries in a context stack matching information contained in the merged transcription;
[19.6]	determining a most likely context for the multi-modal input based on the identified entries;
[19.7]	identifying a domain agent associated with the most likely context for the multi-modal input;
[19.8]	communicating a request to the identified domain agent; and
[19.9]	generating a response to the user from content provided by the identified domain agent as a result of processing the request.
Claim 25	
[25.0]	The method of claim 19, further comprising verifying an identity of the user based on voiceprint matching, password matching, or pass-phrase matching.
Claim 26	
[26]	The method of claim 19, further comprising receiving one or more follow-up multi-modal inputs at the conversational voice user interface, the follow-up multi-modal inputs including at least one of a follow-up natural language utterance or a follow-up non-speech input.
Claim 27	
[27]	The method of claim 26, wherein the identified domain agent updates the context stack and the semantic knowledge-based model in response to processing the request.
Claim 28	
[28]	The method of claim 27, wherein the knowledge-enhanced speech recognition engine determines a most likely context for the follow-up multi-modal input using the updated context stack.

Claim 29	
[29]	The method of claim 27, wherein the speech recognition engine creates a speech-based transcription of the follow-up natural language utterance using the updated semantic knowledge-based model.
Claim 30	
[30]	The method of claim 19, wherein the identified domain agent processes the request by querying one or more local or network information sources.
Claim 32	
[32]	The method of claim 19, wherein the identified domain agent processes the request by directing a command to one or more local or remote devices.