

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

MICROSOFT CORPORATION,
Petitioner,

v.

DIALECT, LLC,
Patent Owner.

IPR2025-00656
U.S. Patent No. 8,447,607

**PETITION FOR *INTER PARTES* REVIEW OF
U.S. PATENT NO. 8,447,607**

TABLE OF CONTENTS

	Page
EXHIBIT LIST	vii
TABLE OF ABBREVIATIONS AND CONVENTIONS	ix
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 <i>INTER PARTES</i> REVIEW	2
A. Grounds for Standing	2
B. Identification of Challenged Claims	3
V. THE 607 PATENT	3
A. Person of Ordinary Skill in the Art	3
B. Overview of the 607 Patent	4
VI. CLAIM CONSTRUCTION	5
VII. PRINCIPAL PRIOR ART	5
A. <u>Maes</u>	5
B. The Combination of Maes’s Teachings with those of Ittycheriah and Coffman	8
VIII. THE CHALLENGED CLAIMS ARE UNPATENTABLE	12
A. Ground 1: <u>Maes</u> renders obvious claims 12 and 13	12
1. Claim 12	12
a. [12.0] <i>A method for processing natural language inputs, comprising:</i>	12
b. [12.1] receiving, by one or more processors, a multi- modal natural language input from a user, the multi-	

	modal natural language input including a natural language utterance and a non-speech input;.....	16
c.	[12.2] generating, by the one or more processors, a non-speech transcription from the non-speech input;	22
d.	[12.3] identifying, by the one or more processors, the user who provided the multi-modal natural language input;.....	28
e.	[12.4] generating, by the one or more processors, a speech-based transcription based on a cognitive model associated with the user, wherein the cognitive model includes information on one or more prior interactions between the user and the device;	33
f.	[12.5] generating, by the one or more processors, a merged transcription from the speech-based transcription and the non-speech transcription;.....	46
g.	[12.6] identifying, by the one or more processors, an entry in a context stack, from among a plurality of entries that are in the context stack and that are each indicative of context, wherein the identified entry matches information in the merged transcription;.....	55
h.	[12.7] identifying, by the one or more processors, a domain agent associated with the entry in the context stack;	65
i.	[12.8] determining, by the one or more processors, a request based on the merged transcription; and	68
j.	[12.9] communicating, by the one or more processors, the request to the domain agent, wherein the domain agent is configured to generate a response to the user.....	70
2.	Claim 13	72
a.	[13.0] The method of claim 12, wherein the cognitive model is a first cognitive model, and wherein the generating the speech-based transcription is based on a second cognitive model that includes information on one or more	

	interactions between the device and a plurality of users of the device.	72
B.	Ground 2: Maes, Coffman and Ittycheriah Render Obvious claims 12 and 13.....	74
IX.	The Board should not exercise its discretion and deny institution.....	75
A.	The Board Should Not Deny Institution Under 35 U.S.C. § 325(d)	75
B.	The Board Should Not Deny Institution Under 35 U.S.C. § 314(a).....	77
1.	Factor 1: Stay	77
2.	Factor 2: Trial Date	77
3.	Factor 3: Parallel Proceeding	78
4.	Factor 4: Overlapping Issues	78
5.	Factor 5: Overlapping Parties	78
6.	Factor 6: Strength/Other Considerations	79
X.	CONCLUSION.....	79
	CERTIFICATE OF COMPLIANCE.....	81
	CERTIFICATE OF SERVICE	82
	CLAIM LISTING	84

TABLE OF AUTHORITIES

	Page(s)
Cases	
<i>Abbot Vascular, Inc. v. Flexstent</i> , IPR2019-00882, Paper 48, 28-29 (Oct. 2, 2020)	9
<i>Advanced Display Sys., Inc. v. Kent State Univ.</i> , 212 F.3d 1272 (Fed. Cir. 2000)	6
<i>In re Apple Inc.</i> , 979 F.3d 1332 (Fed. Cir. 2020)	80
<i>Apple Inc. v. Fintiv, Inc.</i> , IPR2020-00019, Paper 15, 12 (May 13, 2020).....	79, 81
<i>Arbutus Biopharma Corp. v. ModernaTX, Inc.</i> , 65 F.4th 656, 662-63 (Fed. Cir. Apr. 11, 2023).....	6
<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)	9
<i>Bowtech, Inc. v. MCP IP, LLC</i> , IPR2019-00379, Paper 14 (PTAB Jul. 3, 2019)	78
<i>Comcast Cable Communs., LLC v. Promptu Sys. Corp.</i> , Case No. IPR2018-00342, Paper 13 (PTAB July 19, 2018)	79
<i>CommScope Techs. LLC v. Dali Wireless, Inc.</i> , IPR2022-01242, Paper No. 23 (PTAB Feb. 27, 2023).....	81
<i>CrowdStrike, Inc. v. Webroot Inc.</i> , IPR2023-00126, Paper 9 (PTAB May 5, 2023)	77
<i>Dialect, LLC v. Bank of America, N.A.</i> , 2:24-cv-00207-JRG (EDTX)	1
<i>Dialect, LLC v. Google LLC</i> , 3-24-cv-04388 (NDCA).....	1

<i>Dialect, LLC v. Microsoft Corporation,</i> 2-24-cv-01067 (EDTX)	1
<i>Dish Network v. Broadband iTV,</i> IPR2020-01280, Paper 17 (PTAB Feb. 4, 2021).....	80
<i>Fintiv,</i> IPR2020-00019, Paper 11 (PTAB Mar. 20, 2020)	81
<i>Google LLC v. Dialect, LLC,</i> IPR2024-00750	1, 81
<i>Google LLC v. Dialect, LLC,</i> IPR2024-00752	1
<i>Google LLC v. Dialect,</i> LLC, IPR2024-00753	1
<i>Google, LLC v. Parus Holdings Inc.,</i> IPR2020-00847, Paper 9 (PTAB Oct. 21, 2020).....	79
<i>Harari v. Lee,</i> 656 F.3d 1331 (Fed. Cir. 2011)	7, 40, 41, 6, 72
<i>NetNut Ltd. v. Bright Data Ltd.,</i> IPR2021-01492, Paper 12 (PTAB Mar. 21, 2022)	80
<i>Oticon Medical AB et al. v. Cochlear Limited,</i> IPR2019-00975, Paper 15 (PTAB Oct. 16, 2019).....	79
<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).....	78
<i>Scientific Design Co. Inc. v. Shell Oil Co.,</i> IPR2022-00158, Paper 7 (PTAB Apr. 4, 2022)	77
<i>Sotera Wireless, Inc. v. Masimo Corp.,</i> IPR2020-01019, Paper 12 (PTAB Dec. 1, 2020)	80, 81
<i>In re Stephan,</i> 868 F.3d 1342 (Fed. Cir. 2017)	8

<i>Vivid Techs., Inc. v. Am. Sci. & Eng.'g. Inc.</i> , 200 F.3d 795, 803 (Fed. Cir. 1999)	5
--	---

Statutes

35 U.S.C. § 102(e)	5
35 U.S.C. §103	1, 3, 20
35 U.S.C. § 314(a)	79
35 U.S.C. § 325(d)	77, 78

Other Authorities

37 C.F.R. § 42.6(e), I	85
37 C.F.R. § 42.24	84
37 C.F.R. § 42.100(b)	5

EXHIBIT LIST

No.	Exhibit Description
1001	U.S. Patent No. 8,447,607
1002	File History of U.S. Patent No. 8,447,607
1003	Declaration of Dr. Henry Houh
1004	CV of Dr. Henry Houh
1005	U.S. Patent No. 6,964,023 (“ <u>Maes</u> ”)
1006	U.S. Patent No. 5,937,383 (“ <u>Ittycheriah</u> ”)
1007	Second Amended Complaint
1008	U.S. Patent Application Publication No. 2002/0087315 (“ <u>Lee</u> ”)
1009	“A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” Lawrence Rabiner, 1989
1010	Visser, I. (2011), “Seven things to remember about hidden Markov models: a tutorial on Markovian models for time series.” <i>J. Math. Psychol.</i> 55, 403–415.
1011	RESERVED
1012	International Patent Application Publication No. WO 00/20962 (“Coffman”) filed as PCT/US99/22927
1013	U.S. Patent No. 7,137,126 (“ <u>Coffman126</u> ”)
1014	RESERVED
1015	RESERVED
1016	RESERVED

No.	Exhibit Description
1017	Excerpts from Dafydd Gibbon et al., “Handbook of Multimodal and Spoken Dialogue Systems : Resources, Terminology, and Product Evaluation” (2000) (“Gibbon”)
1018	RESERVED
1019	RESERVED
1020	RESERVED
1021	RESERVED
1022	RESERVED
1023	RESERVED
1024	RESERVED
1025	RESERVED
1026	RESERVED
1027	RESERVED
1028	Shdaifat, I., Grigat, R.R. and Lütgert, S., 2001. Viseme recognition using multiple feature matching. In <i>INTERSPEECH</i> (pp. 2431-2434)
1029	EDTX Calendar, Judge Gilstrap

TABLE OF ABBREVIATIONS AND CONVENTIONS

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

I. INTRODUCTION

Petitioner submits this Petition for *Inter Partes* Review of claims 12 and 13 (the “Challenged Claims”) of U.S. Patent No. 8,447,607 (the “607 Patent” (Ex.1001)), assigned to Patent Owner (“PO”). Petitioner respectfully submits that the Challenged Claims of the 607 Patent are unpatentable under 35 U.S.C. §103, and requests that the Board institute an *inter partes* review of the 607 Patent.

II. MANDATORY NOTICES

A. Real Party-in-Interest

Apart from Petitioner Microsoft Corporation, another real party-in-interest for this Petition is Bank of America, N.A (“BOA”).

B. Related Matters

The 607 Patent is asserted by PO in *Dialect, LLC v. Bank of America, N.A.*, 2:24-cv-00207-JRG (EDTX), and also in *Dialect, LLC v. Meta Platforms, Inc.* 7:25-cv-00060 (WDTX), and *Dialect, LLC v. Google LLC*, 3-24-cv-04388 (NDCA). In addition, the following cases involve PO and related patents: *Dialect, LLC v. Salesforce, Inc.*, 7:25-cv-00061, *Dialect, LLC v. Microsoft Corporation*, 2-24-cv-01067 (EDTX), *Google LLC v. Dialect, LLC*, IPR2024-00750 (PTAB), *Google LLC v. Dialect, LLC*, IPR2024-00752 (PTAB), *Google LLC v. Dialect, LLC*, IPR2024-00753 (PTAB).

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 607 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 607 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 607 Patent.

B. Identification of Challenged Claims

Ground 1: Maes (Ex.1005) renders obvious claims 12 and 13 under 35 U.S.C. § 103.

Ground 2: Maes, Coffman (Ex.1012), and Ittycheriah (Ex.1006) render obvious claims 12 and 13 under 35 U.S.C. § 103.

V. THE 607 PATENT

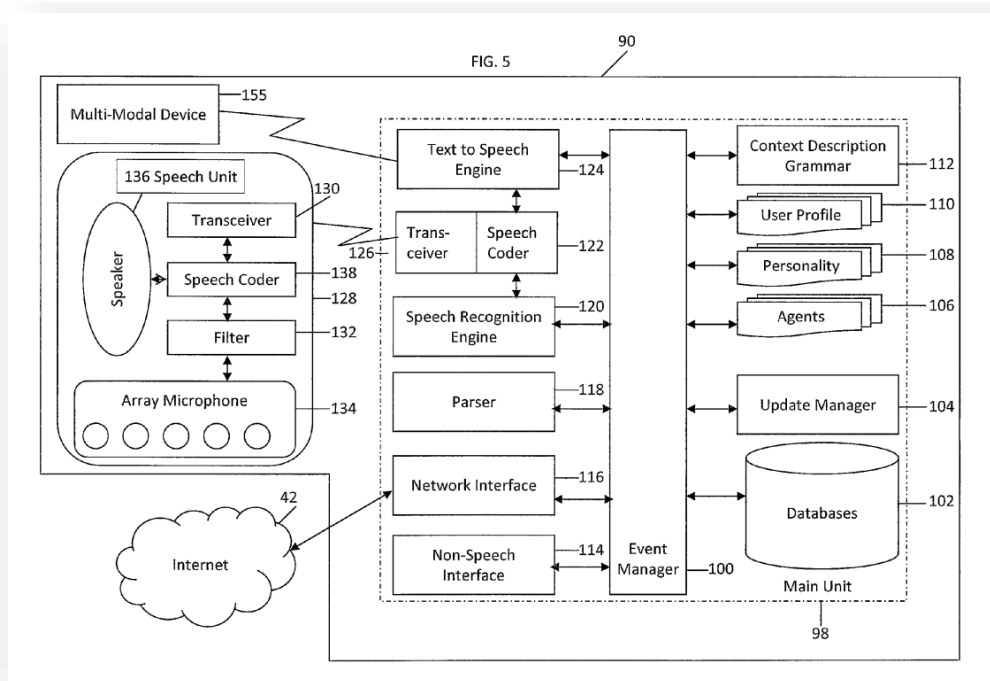
A. Person of Ordinary Skill in the Art

A POSITA with respect to the Asserted Patent as of the time of the invention, i.e., the earliest possible priority date of the 607 Patent, August 29, 2005, would have had a bachelor's degree in electrical engineering, computer science, computer engineering, or an equivalent, and two years of relevant experience involving computer science fundamentals, including natural language processing, speech recognition and transcription, non-speech recognition and transcription that is pertinent to the 607 Patent. Lack of professional experience can be remedied by additional education, and vice versa. Ex.1003, ¶24.

B. Overview of the 607 Patent

The 607 Patent was filed on June 4, 2012, and claims priority to an application filed on August 29, 2005, which Petitioner assumes is the effective filing date for this proceeding.

The 607 Patent explains that it “provides mobile devices having a speech interface and/or a combination of a speech interface and a non-speech interface to enable natural language human-machine interactions” and that it “enables mobile users to submit natural language speech and/or non- speech questions or commands in a wide range of domains.” Ex.1001, 1:20–25. Figure 5 of the 607 Patent is characterized as “an overall diagrammatic view of the interactive natural language processing system.” Ex.1001, 15:38–40.



Ex.1001, Figure 5. In Figure 5, the disclosed system comprises speech unit 128, speech recognition engine 120, context description grammar module 112, parser 118, agents 106, multi-modal device 155 and non-speech interface 114. Speech unit 128 includes a microphone to receive a spoken utterance from a user. Ex.1001, 25:48–52.

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 607 Patent, the Board need not expressly construe the claim terms. *See Vivid Techs., Inc. v. Am. Sci. & Eng. 'g. Inc.*, 200 F.3d 795, 803 (Fed. Cir. 1999).

VII. PRINCIPAL PRIOR ART

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

Maes describes “systems and methods are provided 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.

Moreover, Maes expressly incorporates by reference Coffman and Ittycheriah in their entireties. In particular, Maes states that “[i]t is to be appreciated the conversational virtual machine disclosed in PCT patent application identified as PCT/US99/22927 filed on Oct. 1, 1999 and entitled “Conversational Computing Via Conversational Virtual Machine,” [*i.e.*, Coffman] the disclosure of which is incorporated by reference herein, may be employed to provide a framework for the I/O manager, recognition engines, dialog manager and context stack of the invention.” Ex.1005, 9:30–37. Maes also states “another method to detect incorrectly recognized words is disclosed in U.S. Pat. No. 5,937,383 to Ittycheria entitled “Apparatus and Methods for Speech Recognition Including Individual or Speaker Class Dependent Decoding History Caches for Fast Word Acceptance or Rejection,” the disclosure of which is incorporated herein by reference.” Ex.1005, 19:43–49.

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 Advanced Display Sys., Inc. v. Kent State Univ.*, 212 F.3d 1272,

1282 (Fed. Cir. 2000) (“Material not explicitly contained in the single, prior art document may still be considered for purposes of anticipation if that material is incorporated by reference into the document.”); *see also Harari v. Lee*, 656 F.3d 1331, 1335 (Fed. Cir. 2011).

As the entire disclosure of Coffman is directed to “Conversational Computing Via Conversational Virtual Machine,” Dr. Houh explains that a POSITA would understand Maes to be incorporating with detailed particularity the entire disclosure of Coffman. At a minimum, the disclosures related to the “conversational virtual machine” that “provide a framework for the I/O manager, recognition engines, dialog manager and context stack,” are expressly incorporated with particularity, which is the material relied upon below. Similarly, the entire disclosure of Ittycheriah is also expressly incorporated with particularity given its focus on “detect[ing] incorrectly recognized words”—the focus of the entire disclosure of Ittycheriah—the material of which is relied upon below. Ex.1003, ¶54.

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

To the extent one asserts that Coffman and Ittycheriah encompass different embodiments¹ from that of Maes, a POSITA would have been motivated to include Ittycheriah's teachings of maintaining a decoding history cache database storing keywords (vocabulary such as names and/or commands) frequently used by a user or a class of speakers, and score information generated from decoded keywords in accordance with Hidden Markov Models (HMMs) (Ex.1006, 1:55–58, 4:19–32, 4:54–63, and 5:5–19) in Maes's system that receives and processes multi-modal inputs (e.g., by producing decoded text) for the reasons below. Further, a POSITA would have been additionally motivated to combine Coffman's teachings related to searching a context stack to find entries that match transcribed user input and merging/combining speech and non-speech (e.g., keyboard input) transcriptions (Ex.1012, 41:2–5, 42:1–3, and 38:1–19) with Maes's teachings for the reasons below. Ex.1003, ¶56.

Maes expressly incorporates by reference Coffman and Ittycheriah. In addition, all three references name Stephane Maes as a common inventor, and have

¹ See *In re Stephan*, 868 F.3d 1342, 1346 n.1 (Fed. Cir. 2017) (“Whether a rejection is based on . . . combining multiple embodiments from a single reference . . . there must be a motivation to make the combination and a reasonable expectation that such a combination would be successful.”)

IBM as the common assignee. Ex.1005, Abstract, 9:30–37, 19:43–49; Ex.1012, Cover; Ex.1006, Cover. 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). Ex.1003, ¶57.

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

Dr. Houh also states that using HMMs for decoding words, as Ittycheriah teaches (Ex.1006, 4:19–32), was well-known and utilized in language understanding and pertinent for analyzing cognitive processes such as in speech recognition models at the time of the 607 Patent. Ex.1008, [0013] (“[HMMs] are described generally in ... “Robustness In Automatic Speech Recognition”, Jean Claude Junqua et al., Kluwer Academic Publishers, Norwell, Mass., 1996, pages 90-102.”); *see also* Ex.1009, 258 (“**Neither** the theory of [HMMs] nor its applications to speech recognition **is new.**”) Ex.1003, ¶59.

Furthermore, Visser identifies HMMs as “**especially suitable for studying psychological or cognitive processes in which qualitatively different cognitive states unfold over time**” and therefore, considers HMMs as “**the model of choice for analyzing cognitive processes based on time serial data.**” Ex.1010, 403, 413. Thus, these references reflect the POSITA’s knowledge as it relates to popularity of HMMs and their benefits to speech recognition techniques at the time of the claimed invention in support of Ittycheriah’s teachings. 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, 6:15–28, 6:43–50, 7:60–8:4, 8:37–42. A POSITA would be further motivated to combine the teachings of Maes and Ittycheriah because Ittycheriah’s voice-controlled navigation would

significantly benefit Maes's in-vehicle system by providing voice-based navigation assistance to drivers. 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 of the context with teachings of Maes and Ittycheriah because to do so would benefit Maes's system by providing efficient identification and retrieval of the most-recent context, allowing for faster executions and improving the responsiveness of Maes's system, and would additionally be inline with Ittycheriah's goals of quickly accepting or rejecting keywords. Ex.1012, 21, 23–28, 41:1–5, 40:15–19, and 42:7–9; Ex.1005, 7:60–8:4, 8:37–42, 37:51–62, and 5:29–34; Ex.1006, 2:58–60; Ex.1003, ¶62.

Coffman, Maes and Ittycheriah are analogous art as the 607 Patent because each are directed to functionality of natural language processing systems. Ex.1001, 1:65–2:2; Ex.1005, 6:15–28, 6:43–50; Ex.1006, 1:55–58, 2:13–19; Ex.1012, 1–4, 29:1–4. Furthermore, each of Maes and Coffman expressly disclose the use of a context stack and notably, Maes's context stack disclosures are almost verbatim as Coffman's. Ex.1012, 21, 23–28 and Ex.1005, 37:51–62. Furthermore, each of Maes and Coffman teach merging/combining speech and non-speech transcriptions or otherwise synchronize information streams. Ex.1005, 20:5–7, 11:55–63, 21:51–54,

29:30–36; Ex.1012, 38:2–19. Thus, they are reasonably pertinent to the problem addressed in Maes and the 607 Patent as well. Ex.1003, ¶63.

Dr. Houh further 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 system producing decoded text to include Ittycheriah's teachings of a cache database storing scores and words decoded in accordance with HMMs from previous interactions with a user or a class of users and to further include Coffman's teachings related to searching a context stack to find entries that match transcribed user input and merging/combining speech and non-speech transcriptions would add to the functionality of Maes's system and would have been a simple implementation and predictably resulted in a recognition system with improved accuracy, as Maes seeks to implement (Ex.1005, 11:62–63). Ex.1003, ¶64.

VIII. THE CHALLENGED CLAIMS ARE UNPATENTABLE

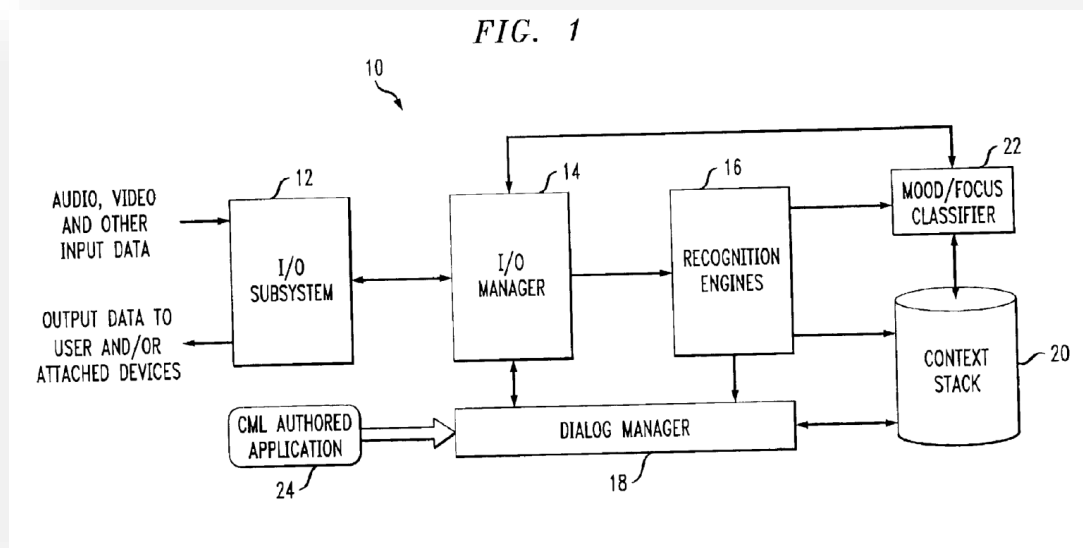
A. Ground 1: Maes renders obvious claims 12 and 13

1. Claim 12

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

Maes describes “**multi-modal data processing techniques**” that have inherent advantages over systems that use only one data input mode. Ex.1005, Abstract, 4:10–11, 1:9-23 (emphasis added²); Ex.1003, ¶66.

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



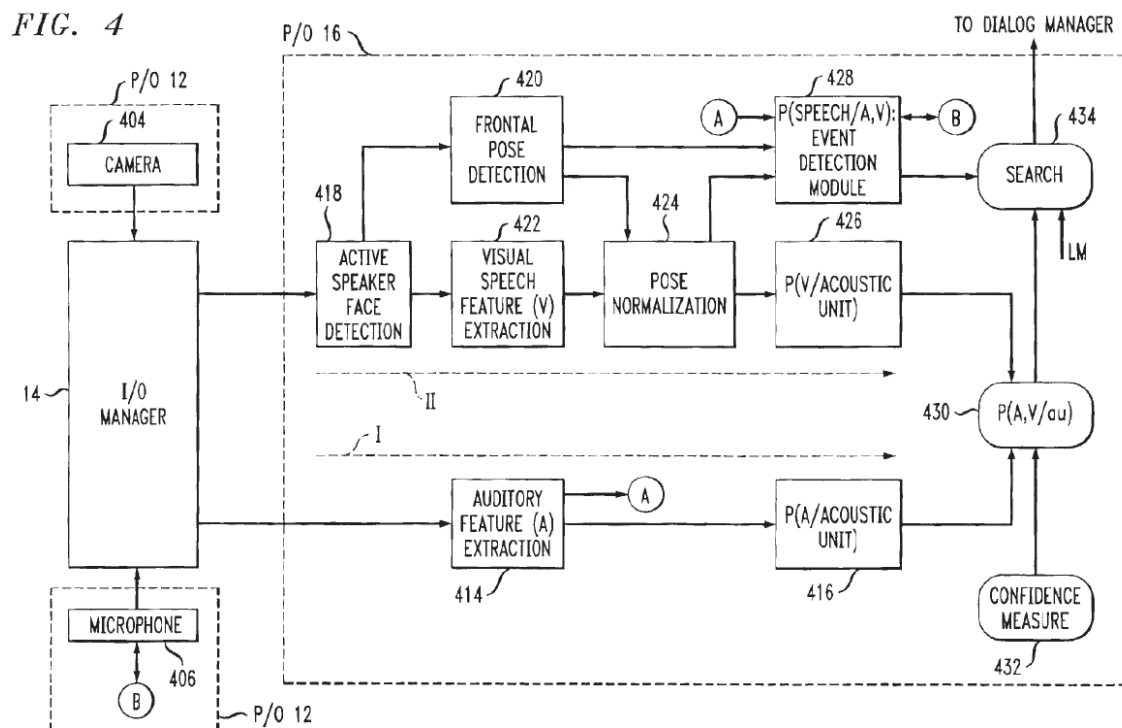
² Emphasis added throughout unless otherwise specified.

Ex.1005, Figure 1. Ex.1003, ¶67.

Referencing Figures 1 and 4, Maes describes Figure 4 (below) as “an audio-visual speech recognition module³ that may be employed as one of the recognition modules⁴ of FIG. 1 **to perform speech recognition using multi-modal input data received** in accordance with the invention.” Ex.1005, 10:50–52. Specifically, Figure 4 shows that Maes’s system receives multi-modal input data including an audio signal (e.g., speaker’s speech and/or background noise) (“*natural language inputs*”) and a video signal (e.g., speaker’s face including lip movement and/or background objects in the environment) and subsequently processes the multi-modal input data (e.g., comprising “*natural language inputs*”) by passing through several processing modules (e.g., the video signal is processed by blocks/modules 418, 422, 424, 426 and the audio signal is processed by blocks/modules 414, 416).

³ The terms “Maes’s system” and “audio-visual speech recognition module” are used interchangeably.

⁴ Maes uses the term “block” and “module” interchangeably. *Compare* 12:43-38 and 17:46-50 referring to “414” as “block[] 414” and “module 414.”



Ex.1005, Figure 4; Ex.1003, ¶68.

Moreover, Maes describes examples of spoken utterance⁵ inputs which further confirm that its method applies to “*natural language inputs*.” For example, Maes’s system, which, among others, provides functions of “natural language understanding (NLU), natural language generation (NLG)” and “NL parsing” can be used in the

⁵ As used herein, a “spoken utterance” received using Maes’s system would encompass both a speech and a non-speech portion of the utterance. Ex.1005, 6:43–47.

context of broadcast news which contain a newsperson speaking at a location where there is arbitrary activity and noise in the background, (Ex.1005, 41:14–16, 35:58, 10:63–11:15), and may be employed within a vehicle such that if a user says the spoken utterance “turn it on,” Maes’s system processes inputs relating to an I/O event representative of the user’s spoken utterance (audio portion) and the accompanying lip movement of the spoken utterance (video portion). Ex.1005, 4:30–32, 6:43–50, 10:63–11:15, and 7:63–8:29; Ex.1003, ¶¶69–72.

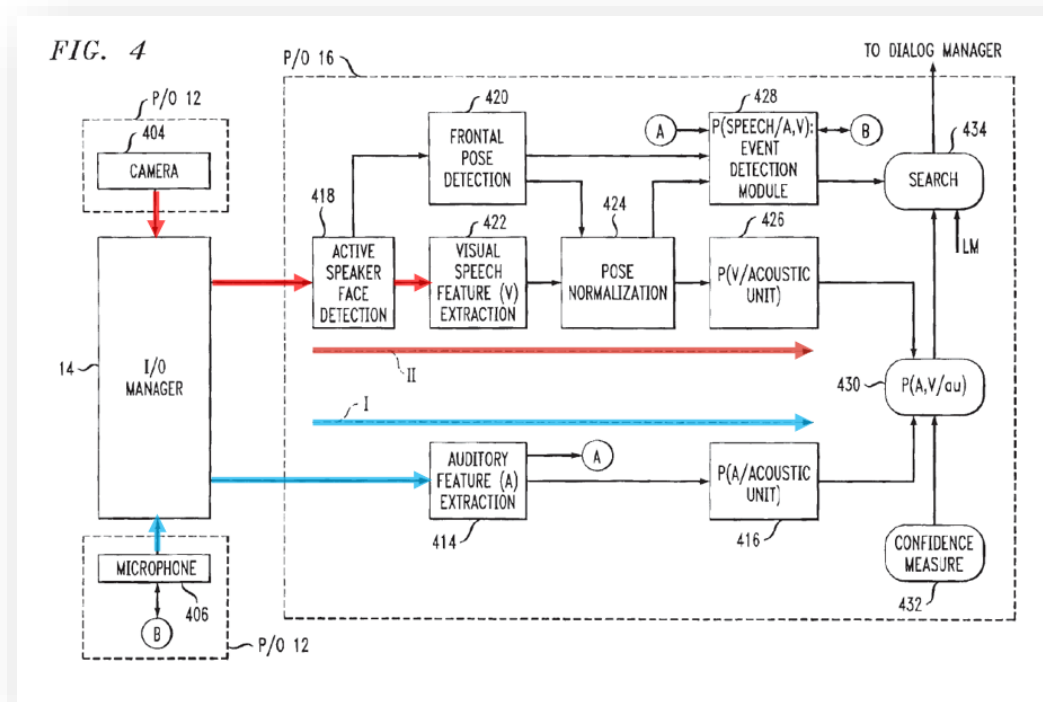
Therefore, if limiting, Maes teaches or suggests the preamble. Ex.1003, ¶73.

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

Maes describes Figure 4 as “an audio-visual speech recognition module that may be employed as one of the recognition modules of FIG. 1 **to perform speech recognition using multi-modal input data received** in accordance with the invention.” Ex.1005, 10:50–52. Specifically, Figure 4 discloses “an audio-visual speech recognition module” that “receive[s]” (“*receiving*”) multi-modal input data—comprising an audio signal and a video signal—from a speaker (“*a multi-modal natural language input from a user*”). Ex.1005, 10:47–11:23; Ex.1003, ¶75.

Maes discloses that 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/or background objects in the environment. Ex.1005, 11:8–63. Additionally, 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); Ex.1003, ¶76.

Referencing Figure 4⁶, Maes explains “**receiving real-time arbitrary content from a video camera 404 and microphone 406 via the I/O manager 14** ... While the video signals received from the camera 404 and the audio signals received from the microphone 406 are shown in FIG. 4 as not being compressed, they may be compressed and therefore need to be decompressed in accordance with the applied compression scheme.” Ex.1005, 11:23–31. Feature extractor 414 of Maes’s system “**receives an audio or speech signal**” and the active speaker face detection module 418 “**receives video input [from] camera 404**” (e.g., “the visual portion (e.g., lip movement) of the utterance” Ex.1005, 6:46-47; *see also* 11:64–12:2, 12:49–61. In Figure 4, phantom line (denoted by Roman numeral I) represents the processing path of the audio information signal and another phantom line (denoted by Roman numeral II) represents the processing path of the video information signal. Ex.1005, 11:55–63; Ex.1003, ¶¶77–78.

As Dr. Houh explains, Maes describes what constitutes the “audio signal” (“*natural language utterance*”) and the “video signal” (“*non-speech input*”). For example, the “audio signal” (“*natural language utterance*”) can include “silence,” “speech,” and “noise” and, the “video signal” (“*non-speech input*”), which “can be

⁶ A typographical error appears in Figure 4—the labels “P/O” are incorrect and should be denoted “I/O.” Ex.1005, 4:8–16 and Figure 1 confirm this.

of any particular type” comprises “the visual portion (e.g., lip movement) of the utterance.” Ex.1003, ¶79 (citing Ex.1005, 13:3–15, 6:46–47, 11:32–54). Dr. Houh further explains that 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.1003, ¶79 (citing Ex.1005, 5:60–6:15); *see also* ¶¶80–82 (providing additional details of multi-modal data).

Further, Maes discloses “*one or more processors*” “*receiving [the] multi-modal natural language input.*” Ex.1003, ¶83.

Maes’s “**system ... comprises at least one processor**, [“*one or more processors*”] operatively coupled to the user interface subsystem, and configured to receive at least a portion of the multi-modal input data from the user interface subsystem.” Ex.1005, 2:38–54; Ex.1003, ¶83.

Maes explains that “the elements ... in FIGS. 1 through 9C may be implemented in ... hardware, software, or combinations thereof, e.g., one or more digital signal processors with associated memory, application specific integrated circuit(s), functional circuitry, one or more appropriately programmed general purpose digital computers with associated memory” Ex.1005, 45:50–60; Ex.1003, ¶84.

The 607 Patent describes an embodiment in which the inventive system (mobile structure 10) is “used in vehicles” and comprises “speech interface and non-speech interface.” Specifically, Figure 1 shows the “non-speech interface” of the inventive system including video camera 16 “for receiving multimedia communications or conferences” and “speech interface” / “speech-based interface” including “one or more speech units 128” that further “include one or more microphones ... to receive the utterances from the user.” Ex.1001, 15:58–64, 16:19–20, 25:48–51 and Figure 1. Analogously, as explained below, Maes provides an example of an in-vehicle system. Ex.1005, 6:15–28, 6:43–50; Ex.1003, ¶85.

Maes describes the “*receiving ...*” step referencing the example of an in-vehicle system. For example, a user in the vehicle may say “turn it on,” while pointing at the vehicle radio. Maes’s system receives multi-modal inputs relating to an I/O event representative of the user’s spoken utterance (“*a natural language utterance*”) and lip movement (“*a non-speech input*”) corresponding to the spoken utterance 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. In Maes’s system, “cameras and microphones may be strategically placed throughout the vehicle ... to fully capture all visual activity and audio activity.” *See generally* Ex.1005, 6:15–28; Ex.1003, ¶86.

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

Additional Mapping

Maes additionally explains that “multi-modality ... may comprise a combination of other modalities other than voice and video [such as] ... **keyboard**/pointer/mouse (or telephone keypad) and other sensors.” *See generally* Ex.1005, 3:1–12. Maes’s system “may also comprise other typical input devices for obtaining user input, e.g., GUI-based devices such as a **keyboard**, a mouse, etc., and/or other devices such as a stylus and digitizer pad for capturing electronic handwriting.” Ex.1005, 6:22–26. Maes’s system therefore complements conventional interfaces and user input/output rather than replacing them, and as Maes describes “[t]his is the notion of ‘multi-modality’” whereby “speech, and video ... may be used in parallel with a mouse, **keyboard**, and other input devices such as a pen.” Ex.1005, 31:61–66. Thus, in Maes, keyboard-typed input comprised in the multi-modal input data constitutes an additional example of “*a non-speech input*” included in the “*multi-modal natural language input*.” Ex.1003, ¶88.

Maes’s above disclosures identifying system inputs received via keyboard/pointer/mouse/keypad are the same non-speech inputs described in the 607 Patent: “keypads 14 for receiving textual data input; [a] pointing device or stylus, or other user interface peripherals” and “users ... interact with the mobile device 36 through the speech unit[,], the keypad[,], or keyboard[,], or other peripheral.” Ex.1001,

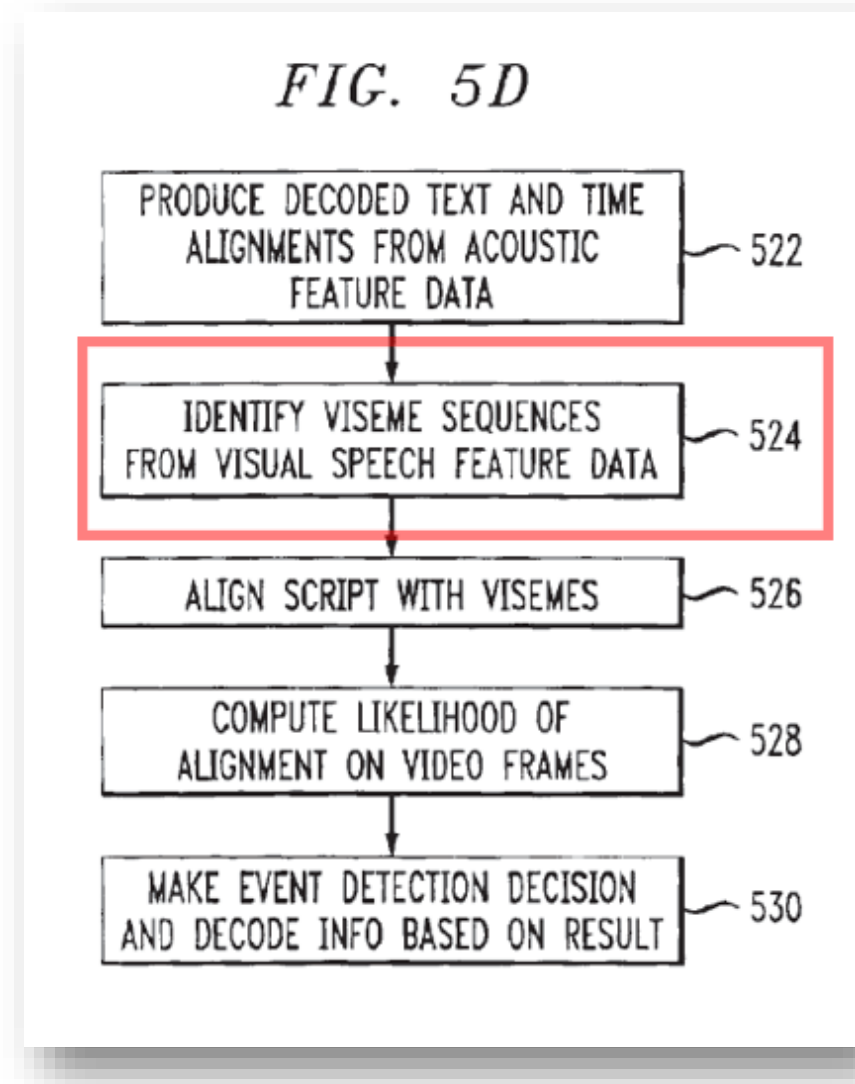
16:18–19 and 19:1–4; *see also* Dialect’s Complaint, ¶147 (keyboard-entered natural language input (“From my checking account”)); Ex.1003, ¶89.

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

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

Visual speech feature extractor 422 included in Maes’s system shown in Figure 4 “extracts visual speech feature vectors (e.g., mouth or lip-related parameters)” used to “produce” (“*generat[e]*”) “a visual phonemes (visemes) sequence” (“*a non-speech transcription*”) “*from*” the video signal (“*the non-speech input*”), e.g., the face detected in the video frame including the user’s lip movement. Ex.1005, 17:51–55, 21:47–50. Ex.1003, ¶92.

Maes discloses, for example, the “*generating . . .*” as a processing step in block 524 in Figure 5D (reproduced below).



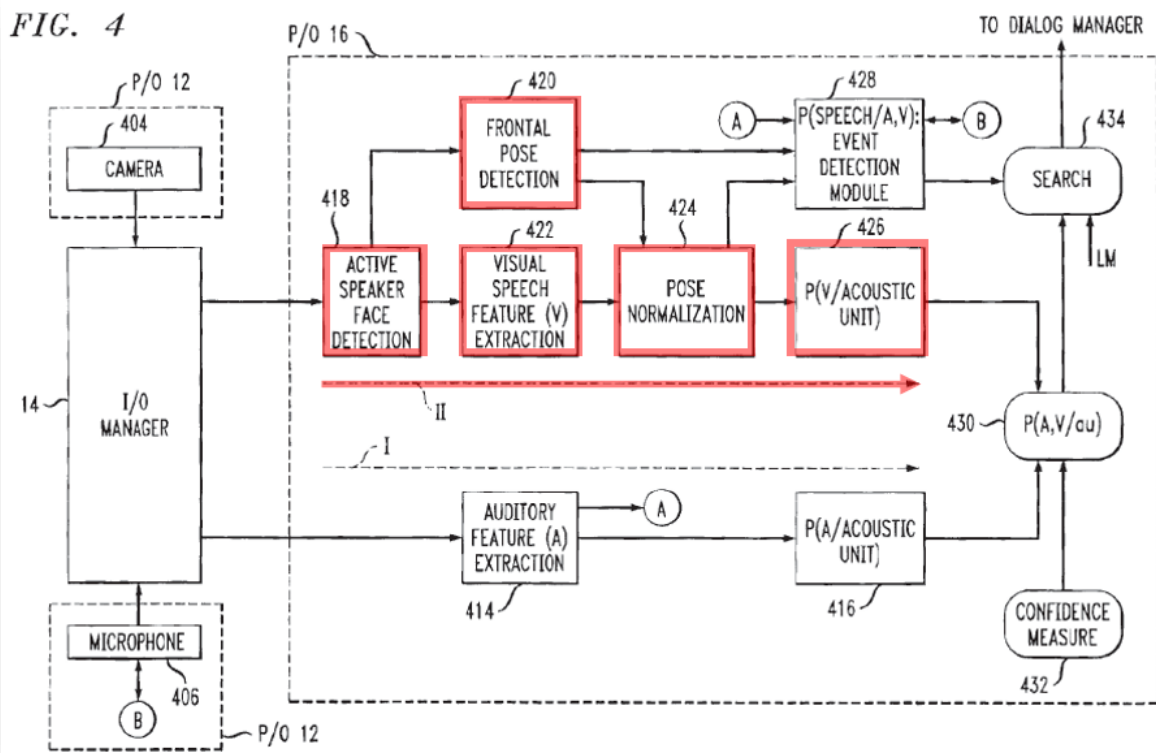
Ex.1005, Figure 5D (annotated); Ex.1003, ¶92.

Referencing Figure 4 (below), Maes explains that modules 422, 424 and 426 as extracting, normalizing, and labeling (“*generating*”) visemes in a visemes sequence (“*a non-speech transcription*”).

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. 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 a visemes sequence (“*a non-speech transcription*”) generated in Maes is essentially a string of visemes that is the tangible result of a visual recognition system’s interpretation of a user’s lip movement accompanying a spoken utterance and provides a visual representation of the non-speech content. Ex.1003, ¶93 (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* 23:1–5; Ex.1003, ¶94.

Maes provides several examples of visual speech features: “grey scale parameters of the mouth region; geometric/model based parameters such as area, height, width of mouth region; lip contours arrived at by curve fitting, spline parameters of inner/outer contour; and motion parameters obtained by three dimensional tracking. Still another feature set that may be extracted via module 422 takes into account the above factors.” Ex.1005, 17:55–63; Ex.1003, ¶95.

Maes describes multiple ways of extracting visual speech feature vectors:

Principal Component Analysis (PCA) ... may be used to extract a vector of smaller dimension from this vector of grey-scale values.

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

Ex.1005, 18:1–20. The visual speech feature vectors in Maes are used to “produce a visual phonemes (visemes) sequence” (*“generat[e] ... a non-speech transcription”*). Ex.1005, 21:43–50. Ex.1003, ¶96.

Therefore, Maes discloses “*generating ... a non-speech transcription from the non-speech input.*” Ex.1003, ¶96.

Referring to the in-vehicle example, Maes explains the “*generating ...*” step as follows: “[t]he I/O manager [of Maes’s system] receives the raw multi-modal data and **abstracts the data into a form that represents ... a spoken utterance [by] generalizing details associated with all or portions of the input data ... to yield a more generalized representation of the data for use in further operations.**” Ex.1005, 6:21–38. Thus, as a result of processing by several modules in Maes’s system, the system produces a visemes sequence corresponding to the lip movement

accompanying the spoken utterance of the user. *See also* Ex.1005, 6:43–50; Ex.1003, ¶97.

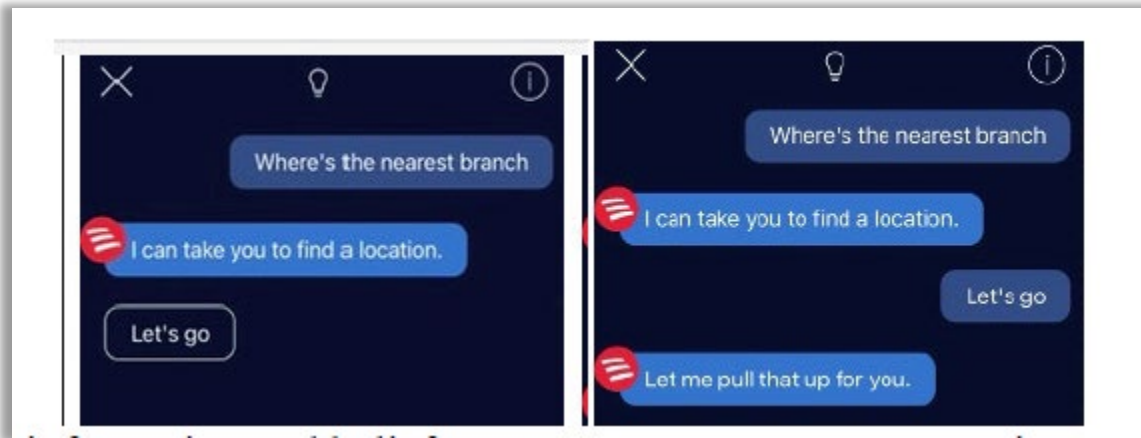
Additionally, as explained above in connection with [12.1], Maes discloses “*one or more processors.*” Ex.1003, ¶98.

Accordingly, Maes teaches or suggests this element. Ex.1003, ¶99.

Additional Mapping

As explained in [12.1], in Maes, keyboard-typed input constitutes an additional example of “*a non-speech input*” and further Maes teaches for “**typical GUI-based input (e.g., keyboard, mouse, etc.), the data may not necessarily need to be recognized since the data is already identifiable without recognition operations.**” Ex.1005, 6:62–65. Therefore, a POSITA would have understood that the keyboard-typed input inherently serves as “*a non-speech transcription.*” Ex.1005, 6:62–65. Ex.1003, ¶100.

Dialect’s Complaint, consistent with Maes’s teachings above, identifies: “*Erica* transcribed a non-speech input (tapping on box provided by *Erica* with the words “Let’s go”) into a non-speech based transcription (displayed/represented to the user in the form of “Let’s go” in the user’s dialogue stream).” Ex.1007, ¶149.



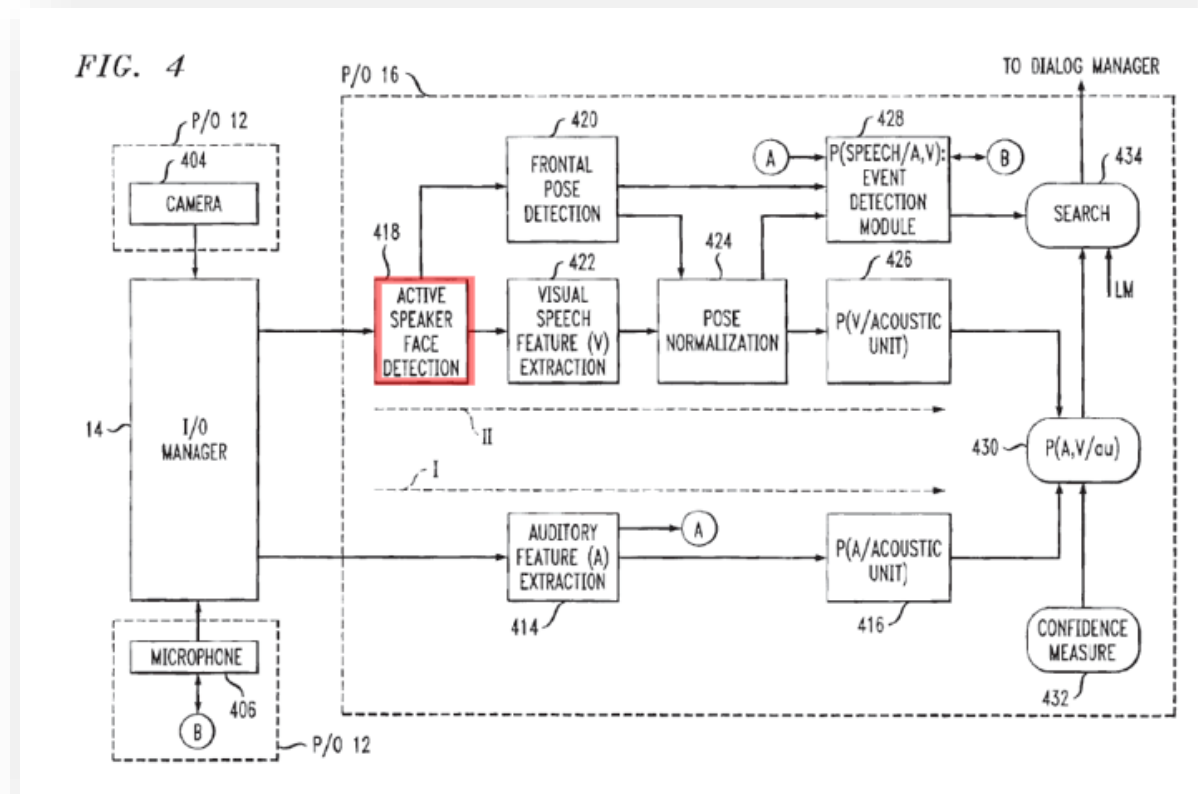
Id. Ex.1003, ¶101.

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

Ex.1003, ¶102.

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

Maes's system shown in Figure 4 (below) includes an active speaker face detection module 418 for “detecting” (“*identifying*”) a face of the “speaker” (“*the user*”) “*who provided*” “the multi-modal input data” (“*the multi-modal natural language input*”). Ex.1005, 10:47–63, 12:49–62.



Ex.1005, Figure 4 (annotated); 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 “*identiffies] ... the user who*

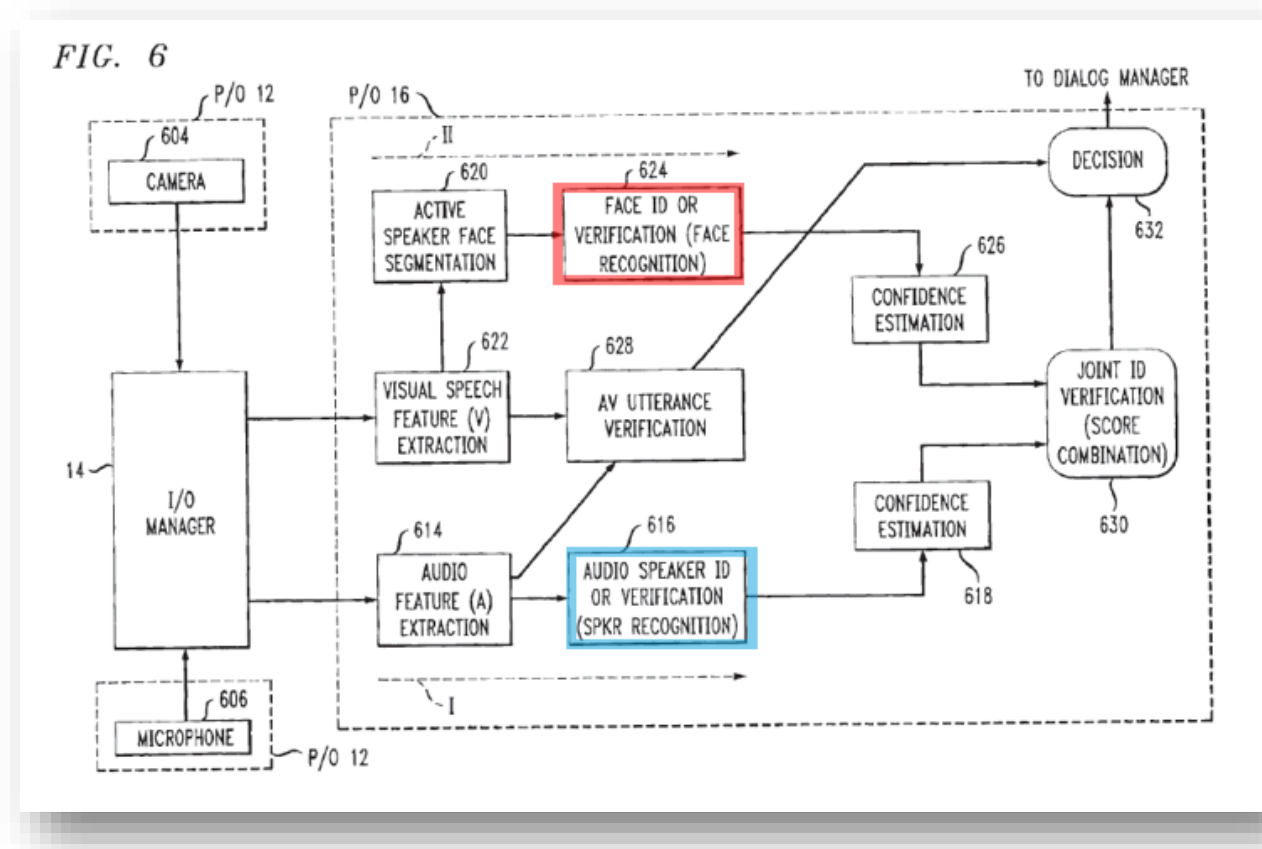
*provided the multi-modal natural language input.” See Ex.1005, 6:50–55 (“[I]n addition, the event may be sent to an audio-visual speaker recognition engine to **have the speaker of the utterance identified, verified and/or authenticated ...**”). Ex.1003, ¶105.*

Referencing Figure 6⁷, Maes explains “**a preferred embodiment** of an audio-visual speaker recognition module that may be employed as one of the recognition modules of FIG. 1 **to perform speaker recognition using multi-modal input data received in accordance with the invention.**” Ex.1005, 22:41–46. “Like the audio-visual speech recognition module of FIG. 4, the speaker recognition module of FIG. 6 may receive the same types of arbitrary content video from the camera 604 and audio from the microphone 606 via the I/O manager 14.” Ex.1005, 22:60–64. For example, “**the [audio speaker recognition] module 616 may perform speaker identification and/or speaker verification using the extracted acoustic feature vectors.** The processes of speaker identification and verification may be accomplished via any **conventional acoustic information speaker recognition system.**” Ex.1005, 23:44–24:3. Furthermore, Dr. Houh explains that Maes describes a speaker identification system that may use two techniques: a model-

⁷ There appears to be a typographical error in Figure 6—the labels “P/O” are incorrect and should have been denoted as “I/O,” e.g., “I/O subsystem 12.” Ex.1005, 4:8–16 and Figure 1 confirm this.

based approach and a frame-based approach. Maes further explains in detail how its system uses the frame-based approach for speaker identification based on audio. Ex.1003, ¶106; *see also* ¶107 (citing Ex.1005, 24:2–30 for details of frame-based approach).

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



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

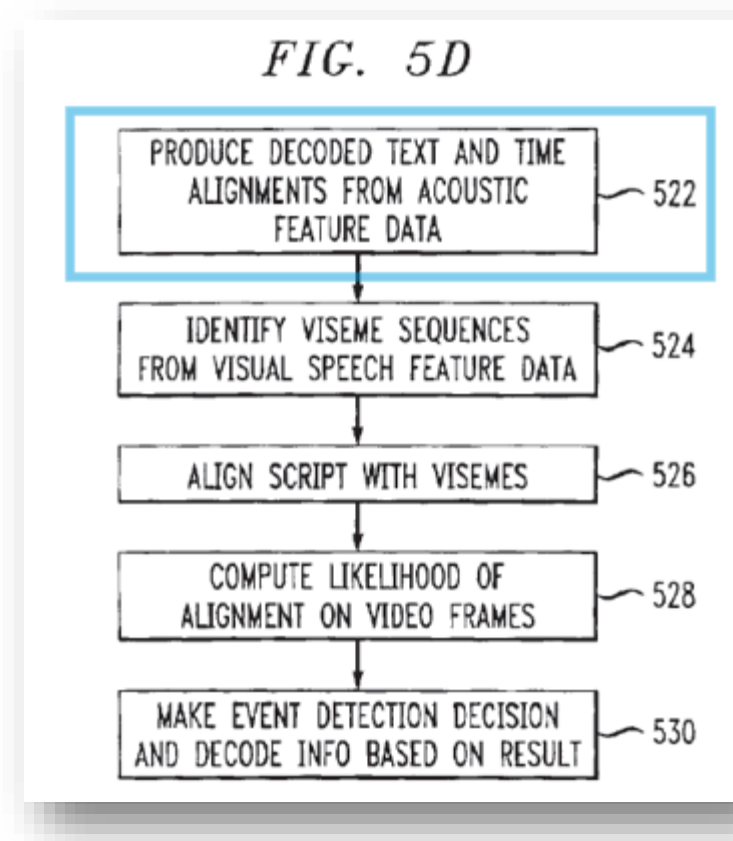
Therefore, Maes teaches or suggests “*identifying ... the user who provided the multi-modal natural language input.*” Ex.1003, ¶111.

Referring to the in-vehicle example in Maes, the “*identifying ...*” step would result in the system recognizing/identifying the specific user who provided the multi-modal natural language inputs comprising an audible portion (e.g., picked up by a microphone) of the spoken utterance of “turn it on” and a visual portion (e.g., picked up by a camera) of lip movement. Ex.1005, 6:43–47, 7:64; Ex.1003, ¶109.

Additionally, as explained above in connection with [12.1], Maes discloses “*one or more processors.*” Ex.1003, ¶110.

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

Referencing Figure 5D (reproduced below), Maes discloses classical speech recognition techniques “produc[ing]” (“*generating*”) “decoded text” (or, script) processed from the spoken utterance (“*a speech-based transcription*”) “using the feature data from the acoustic feature extractor 414.” *See generally* Ex.1005, 21:43–47.



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

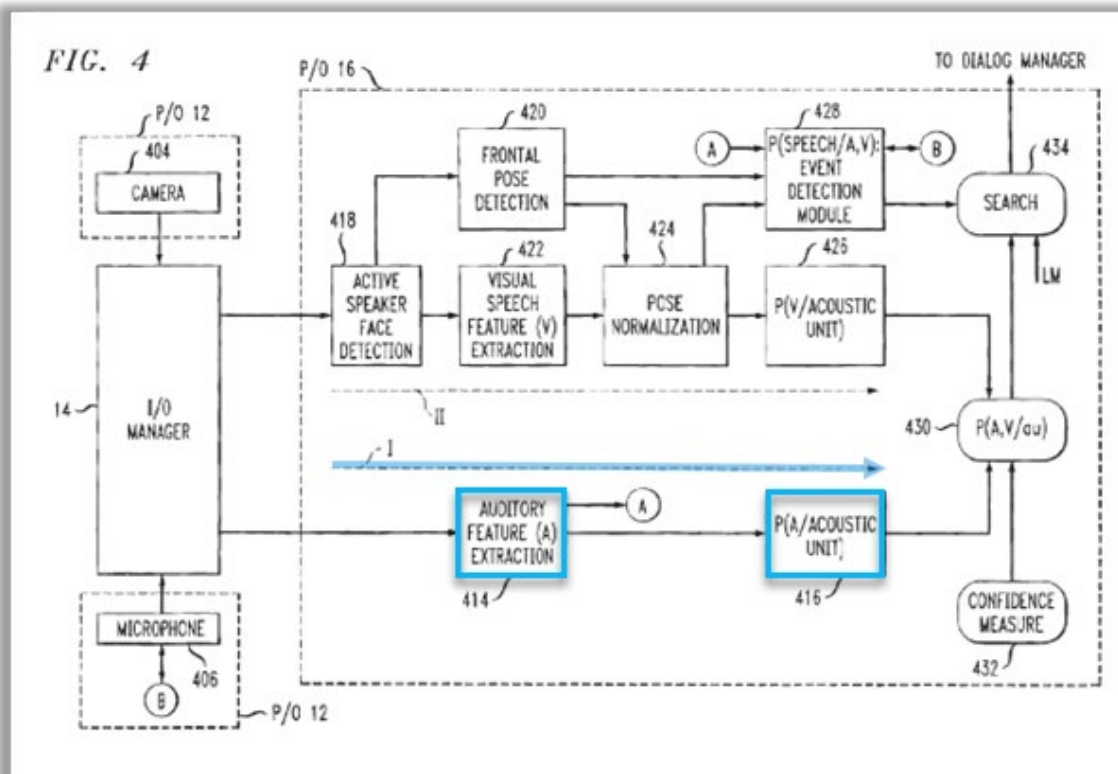
through the acoustic path so that the buffered data may be recognized, in step 520, ... to yield a **decoded output**.”); *see also* 20:23-33; Ex.1003, ¶113.

Referencing Figures 1 and 4 (below), Maes provides additional details explaining the “*generating ...*” step: Maes begins by describing Figure 4 as “an audio-visual speech recognition module that may be employed as **one of the recognition modules⁸ of FIG. 1 to perform speech recognition using multi-modal input data received** in accordance with the invention.” Ex.1005, 10:50–52. Ex.1003, ¶114.

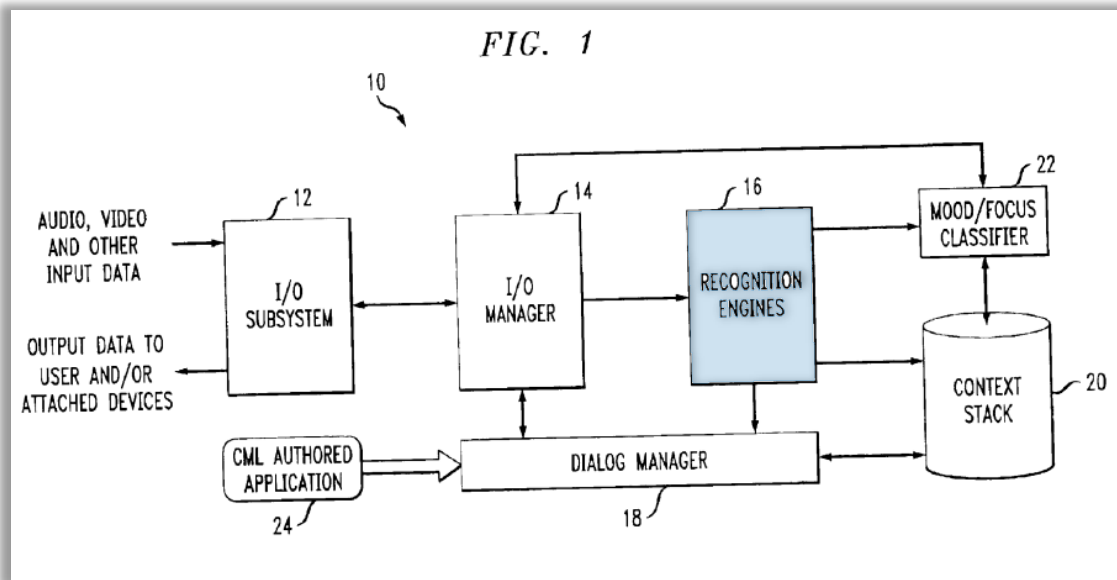
Specifically, Figure 4 shows that the audio-visual speech recognition module receives multi-modal input data including an audio signal (e.g., speech provided by the speaker and/or background noise) and a video signal (e.g., the speaker’s face including lip movement and/or background objects in the environment). For example, the auditory feature extractor 414 “receives an audio or speech signal and, ... extracts spectral features from the signal at regular intervals[,] ... in the form of acoustic feature vectors (signals) which are then passed on to a probability module 416.” Ex.1005, 11:64–12:2. In Figure 4, the acoustic feature vectors are denoted by

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

the letter “A” and a phantom line denoted by Roman numeral I represents the processing path of the audio information signal. Ex.1005, 11:55–64.



Ex.1005, Figure 4 (showing acoustic feature vectors denoted by the letter “A” and audio information processing path (Path I) annotated in blue); *see also* 23:1–5.



Ex.1005, Figure 1 (annotated); Ex.1003, ¶114.

Maes explains how acoustic features are extracted:

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

Ex.1005, 12:12–28; Ex.1003, ¶115.

Maes 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.1003, ¶ (citing Ex.1005, 12:2–12). Indeed, Maes notes that “the processing performed in blocks 414 and 416 may be accomplished via **any conventional acoustic information recognition system capable of extracting and labeling acoustic feature vectors**, e.g., Lawrence Rabiner, Biing-Hwang Juang, “Fundamentals of Speech Recognition,” Prentice Hall, 1993.” Ex.1005, 12:43–49; Ex.1003, ¶116.

As Dr. Houh explains, 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.1003, ¶117 (citing Ex.1005, 23:40–43, 39:55–58).

Additionally, Maes discloses that the step of “producing decoded text” (or, script) (“*generating ... a speech-based transcription*”) by classical speech recognition techniques is “*based on a cognitive model associated with the user [that] includes information on one or more prior interactions between the user and the device.*” Ex.1003, ¶118.

Initially, the 607 Patent describes a “*cognitive model*” as information derived from a user’s interactions with the system that may be used in determining a user’s

future actions by “assisting with speech recognition and/or question or command recognition” stating “personalized cognitive model ... is a model derived from a user’s interaction pattern with the system and may be used to predict what actions the user may take next in time, thus assisting with speech recognition and/or question or command recognition.” Ex.1001, 24:1–11. As explained below, “*cognitive model[s]*” based on information derived from user (or users’) historical interactions with a system to assist in speech recognition were well-known in the art. Ex.1003, ¶119.

For example, Lee describes techniques for assisting speech recognition accuracy using information collected about multiple users and describes using “recognition assisting databases” to aid in recognition of speech input, including a database that “**maintains a personalized profile that have already been recognized for a particular user or for users** who have previously submitted requests which are similar to the request at hand.” Ex.1008, Abstract, [0002], and [0016]. Lee also describes a “popularity engine” that “contains data compiled **from multiple users’ histories** that has been calculated for the prediction of likely user requests.” Ex.1008, [0016]. Lee explains that “[t]his increases the accuracy of word recognition” because it enables the system to “scan[] for the most probable match for words in the user utterance based on past requests.” Ex.1008, [0024]–[0026]; Ex.1003, ¶¶120–121.

As Dr. Houh explains, Maes discloses cognitive models that were well-known to POSITAs in the art, as demonstrated by Lee. For example, Maes describes “features,” e.g.,—acoustic feature vectors used for “produc[ing] decoded text” (“*speech-based transcription[s]*”)—“can be extracted and decisions dynamically returned by **the models**.” Ex.1005, 21:42–47, 43:53–54. With respect to Maes’s system providing dynamic decision making by models, Maes teaches that its system (“*the device*”) can include a “a dynamic classification module ... which performs dynamic classification of the user” such that the system is “configured to modify behavior of the voice system ... based on at least one user attribute ... determined based on at least one acoustic feature extracted from the user’s speech.” Ex.1005, 44:5–11. The “user attribute can include ...: gender **of the user**, age **of the user**, accent **of the user**, native language **of the user**, dialect **of the user**, socioeconomic classification **of the user**, educational level **of the user**, and emotional state **of the user**.” Ex.1005, 39:50–55. Ex.1003, ¶122.

Considering native language as a user attribute, Maes teaches that “[m]eta information about the native language of the speaker can be added to define each accent/native language model” and further that “**at the creation of the models for each native language, one employs a speaker or speakers who are tagged with that language as their native language.**” Ex.1005, 40:18–23. Thus, in Maes, “produc[ing] decoded text” (“*speech-based transcription[s]*”) can involve, for

instance, classifying the user's native language based on a native language model created using a speaker who is tagged with that language as his/her native language. Maes's teachings of employing a speaker for creating a model of a native language of that user, therefore, serve as describing "*a cognitive model associated with the user [and] includes information on one or more prior interactions between the user and*" Maes's system ("*the device*"). Ex.1003, ¶122.

Furthermore, and as Dr. Houh explains, Maes describes that with regard to "recognized speech" or "decoded text" ("*speech-based transcription*"), "[r]ecently, efforts have been initiated to develop appropriate confidence measures" comprising methods using "decision trees trained on word-dependent features (amount of **training utterances**, minimum and average triphone occurrences, **occurrence in language model training**, ... as well as **speaker profile features (accent, dialect, gender, age, speaking rate, identity)**)." Ex.1005, 19:3–20; Ex.1003, ¶123. From the above-mentioned passages, Dr. Houh explains that a POSITA would have understood that in the context of speech recognition systems, training a language model involves prior interactions between a user and a system, and furthermore, that utterances from a single user can be used for the training, and Maes's teachings related to training utterances, occurrence of certain utterances in language model training, and training models based on speaker profile features (e.g., accent, dialect, gender, age, speaking rate, identity—which are speaker-specific), therefore, serve as

describing “*a cognitive model associated with the user [and] includes information on one or more prior interactions between the user and*” Maes’s system (“*the device*”). Ex.1005, 19:3–20; Ex.1003, ¶123.

Additionally, as explained below, Maes teaches building models that characterize each speaker, which therefore teach or suggest “*a ... model associated with the user, wherein the ... model includes information on one or more prior interactions between the user and the device.*” Ex.1003, ¶124.

In Maes’s system (“*the device*”) “**the input acoustic data is compared to determine if the data matches closely enough with stored models. If the comparison yields a close enough match, the person uttering the speech is verified. The match is accepted or rejected by comparing the match with competing models. These models can be selected to be similar to the claimant speaker or be speaker independent (i.e., a single or a set of speaker independent models). If the claimant wins and wins with enough margin (computed at the level of the likelihood or the distance to the models), we accept the claimant. Otherwise, the claimant is rejected.**” 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; Ex.1003, ¶124.

To the extent Patent Owner contends that Maes does not disclose or suggest “*a cognitive model ...*,” Ittycheriah (incorporated by reference in Maes) does. Ex.1005, 19:43–49 (incorporating by reference specific features of Ittycheriah identified as U.S. Pat. No. 5,937,383); Ex.1003, ¶126.

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 cognitive model associated with the user [that] includes information on one or more prior interactions between the user and [a] device*,” such as Maes’s system. Ittycheriah teaches a cache database “storing keywords comprising vocabulary such as names and/or commands frequently used by a user” such that:

When the user first inputs a name or a command, the acoustic front-end ... generates a set of acoustic features which are then provided to the classical speech recognition engine The word is decoded in accordance with the Hidden Markov models and the decoded word, along with the different scores ... 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 cognitive model*”) is “*associated with the user.*” *Id*; Ex.1003, ¶126.

Ittycheriah further explains how the user-specific data from prior interactions with a user is used to test a hypothesis for decoding a word in a current interaction with that user. Ittycheriah’s method “align[s] the baseforms of the words stored in the cache and get a score for all these comparisons. The scores are compared with what is stored in the cache, to decide if one hypothesis is acceptable or not.” Ex.1006, 3:6–9. A POSITA would have found it obvious to modify Maes’s system (e.g., producing decoded text) to additionally include Ittycheriah’s teachings of a cache database storing scores and words decoded in accordance with HMMs from

previous interactions with a user (“*a cognitive model associated with the user*”) resulting in the decoded text (or, script) (“*a speech-based transcription*”) “*including information on one or more prior interactions between the user*” “*and*” Maes’s system (“*the device*”) Ex.1005, 12:12–28; Ex.1006, 3:56–60 and 4:19–53; Ex.1003, ¶127. *See* §VII.B.

Therefore, Maes teaches or suggests that “*generating ... a speech-based transcription*” is “*based on a cognitive model associated with the user [that] includes information on one or more prior interactions between the user and the device.*” Ex.1003, ¶128.

Referring to the in-vehicle example, Maes explains the “*generating ...*” step as follows: Maes’s system receives the raw multi-modal data relating to an I/O event (e.g., representative of the user’s spoken utterance and lip movement accompanying the spoken utterance) 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. As a result of processing by several modules in Maes’s system (and/or including Ittycheariah’s teachings related to a cache database), the system produces decoded text/script corresponding to the user’s spoken utterance, for example, of saying “turn it on.” Ex.1003, ¶129.

Additionally, as explained in connection with [12.1], Maes discloses “*one or more processors.*” Ex.1003, ¶130.

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

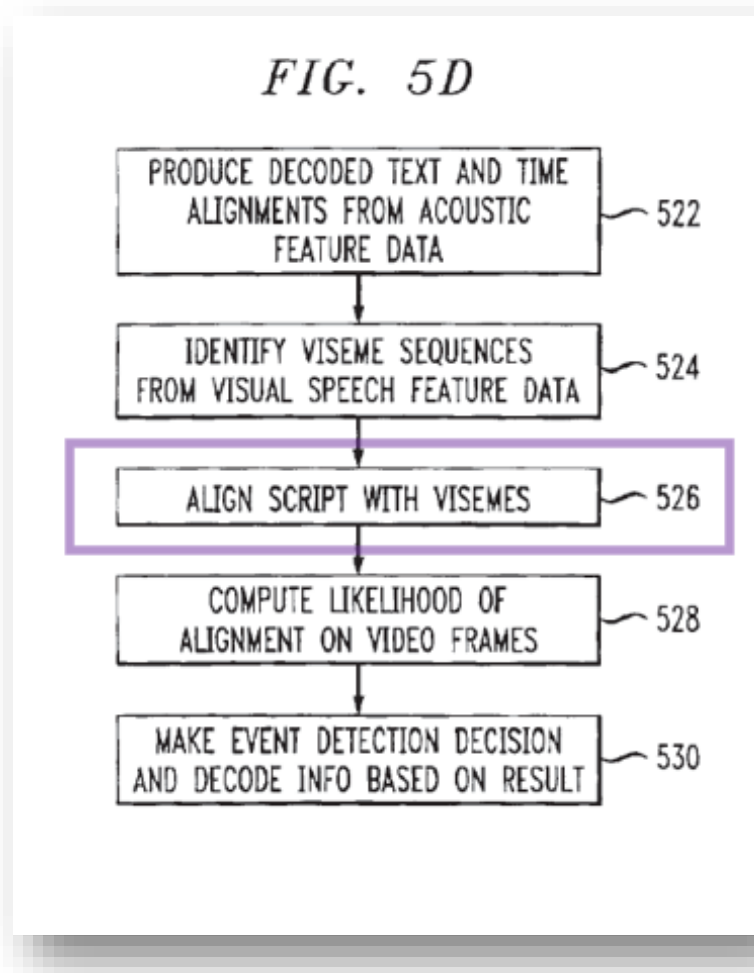
Maes teaches or suggests “*generating ... a merged transcription from the speech-based transcription and the non-speech transcription.*” As explained in connection with [12.2] and [12.4], Maes discloses “the visual phonemes (visemes) sequence” (“*the non-speech transcription*”) and “decoded text” (or, script) (“*the speech-based transcription*”). Ex.1003, ¶133.

Maes discloses that “[t]he I/O manager receives the raw multi-modal data [in the form of audio input data, video input data, as well as other types of input data] and abstracts the data into a form that represents ... a spoken utterance [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, 6:32–38. Referencing Figure 4 (below), Maes teaches the “multi-modal input data” (“*multi-modal natural language input*”) “including” “an audio signal” (e.g., speech provided by the speaker and/or background noise) (“*a natural language utterance*”) “and” “a video signal” (“*a non-speech input*”) (e.g., which can include the speaker’s face including lip movement and/or background objects in the environment) goes through several processing modules. Ex.1005, 23:1–5. Consequently, as outcome

of such operations or processing, “the visual phonemes (visemes) sequence” (“*the non-speech transcription*”) and “the decoded text” (or, script) (“*the speech-based transcription*”) are generated, which is described in Figure 5D (step 526) as:

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, Figure 5D (block 526); Ex.1003, ¶134.

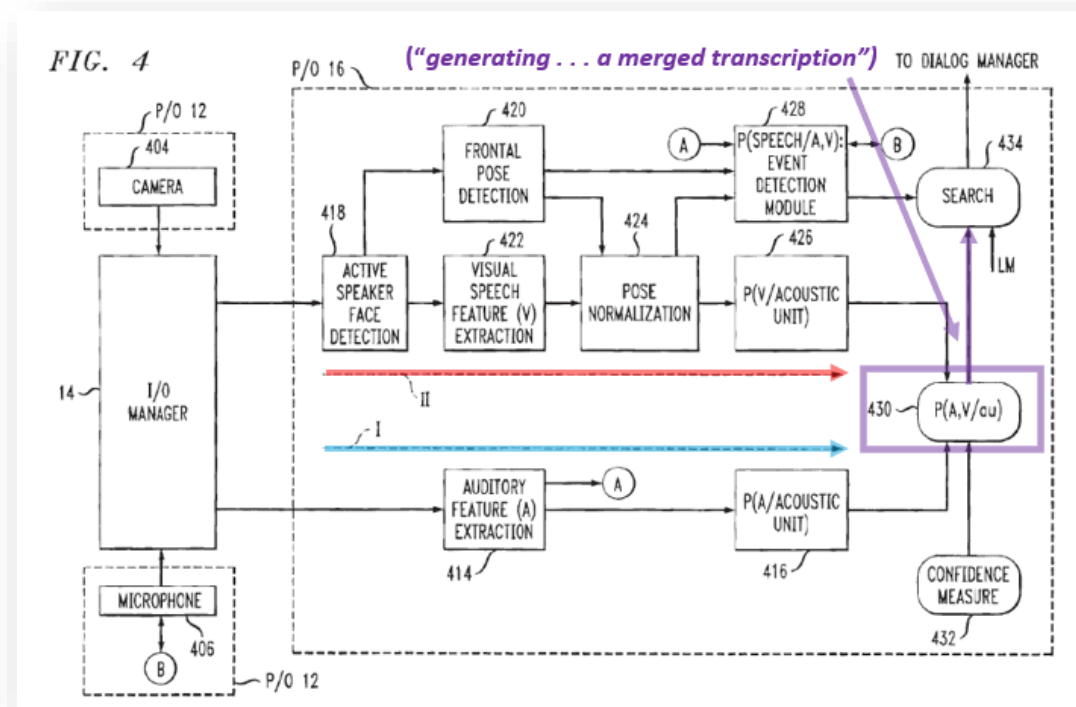
As Dr. Houh explains, Maes provides the purpose and advantages of aligning/synchronizing the audio and video information streams. “A goal associated with utterance verification is to make **a determination that the speech used to verify the speaker in the audio path I and the visual cues used to verify the speaker in the video path II correlate or align. This allows the system to be confident that the speech data that is being used to recognize the speaker is actually what the speaker uttered.**” Ex.1005, 29:30–36. Ex.1003, ¶135 (citing Ex.1005, 29:37–43).

Therefore, by teaching synchronization of the two information streams (i.e, the audio and video information streams/signals) or, more specifically, alignment of “the decoded text” (or, script) (“*the speech-based transcription*”) with “the visual phonemes (visemes) sequence” (“*the non-speech transcription*”), Maes discloses “*generating ... a merged transcription from the speech-based transcription and the non-speech transcription.*” Ex.1003, ¶136.

Referencing Maes’s in-vehicle example (Ex.1005, 7:60–8:4 and 8:37–42), as Dr. Houh explains, the aligned decoded script and visemes sequence associated with the spoken utterance “turn it on” with the accompanying lip movement, therefore, constitutes “*the merged transcription.*” Ex.1003, ¶137.

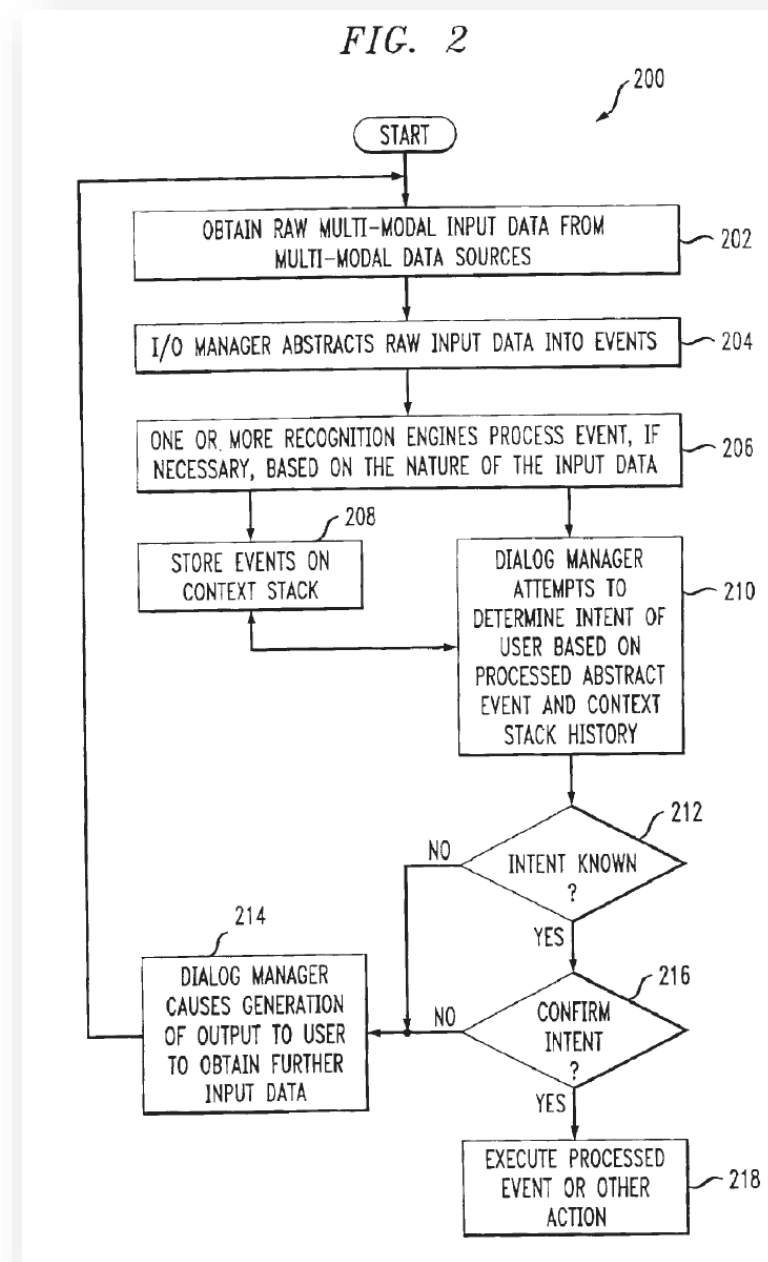
Indeed, with regards to module 430 in Figure 4, Maes identifies “**the first joint use of the visual information and audio information in module 430**” and

“an explanation of how the two types of information [i.e, the audio information included in the decoded text/script and video information included in the viseme sequence] are combined to provide improved recognition accuracy.” Ex.1005, 20:5–7, 11:55–63. Therefore, a POSITA would have understood that module 430 provides the “*generating ... a merged transcription*” step such that the outcome of the “*generating ...*” step produces the aligned decoded script and visemes sequence (i.e., “*a merged transcription*”), e.g., associated with the spoken utterance “turn it on” with the accompanying lip movement.



Ex.1005, Figure 4 (annotated); Ex.1003, ¶138.

Maes provides a detailed explanation of 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 then abstracted into a form that represents one or more events, e.g., “a spoken utterance.” Ex.1005, 6:28–35. And as Dr. Houh explains, in Maes, “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, 6:38–50; Ex.1003, ¶¶139–140 (citing Ex.1005, 6:38–50).

In Maes, one of the ways in which the **audio-visual speech recognition engines** processes or otherwise, recognizes event data is by generating an outcome (*i.e.*, “*merged transcription*”) in which the transcribed audio information (e.g., the decoded text/script) is aligned or synchronized with the transcribed video information (e.g., the visemes sequence). *See also* Ex.1005, 21:50–54 (“**[T]he script is aligned with the visemes.** A rapid (or other) alignment may be performed in a conventional manner ... 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. patent application identified as Ser. No. 09/507,526); Ex.1003, ¶¶141–142.

Therefore, Maes teaches or suggests “*generating ... a merged transcription from the speech-based transcription and the non-speech transcription.*” Ex.1003, ¶143.

Referring to the in-vehicle example, as Dr. Houh explains, the “*generating ...*” step would result in the system aligning/synchronizing the abstracted data (e.g., decoded text/script) processed from the spoken utterance of saying “turn it on” with the abstracted data (e.g., visemes sequence) processed from the lip movement of the user. Ex.1003, ¶144.

Additionally, as explained above in connection with [12.1], Maes discloses “*one or more processors.*” Ex.1003, ¶145.

Therefore, Maes teaches or suggests this claim element. Ex.1003, ¶146.

Additional Mapping

As explained in [12.1], [12.2], and [12.3], in Maes, keyboard-typed input constitutes an additional example of “*a non-speech input*” and further, it also serves as “*the non-speech transcription.*” Additionally, as explained in [12.4], in Maes’s system, “decoded text” (or, script) processed from the spoken utterance constitutes “*the speech-based transcription.*” Ex.1003, ¶147.

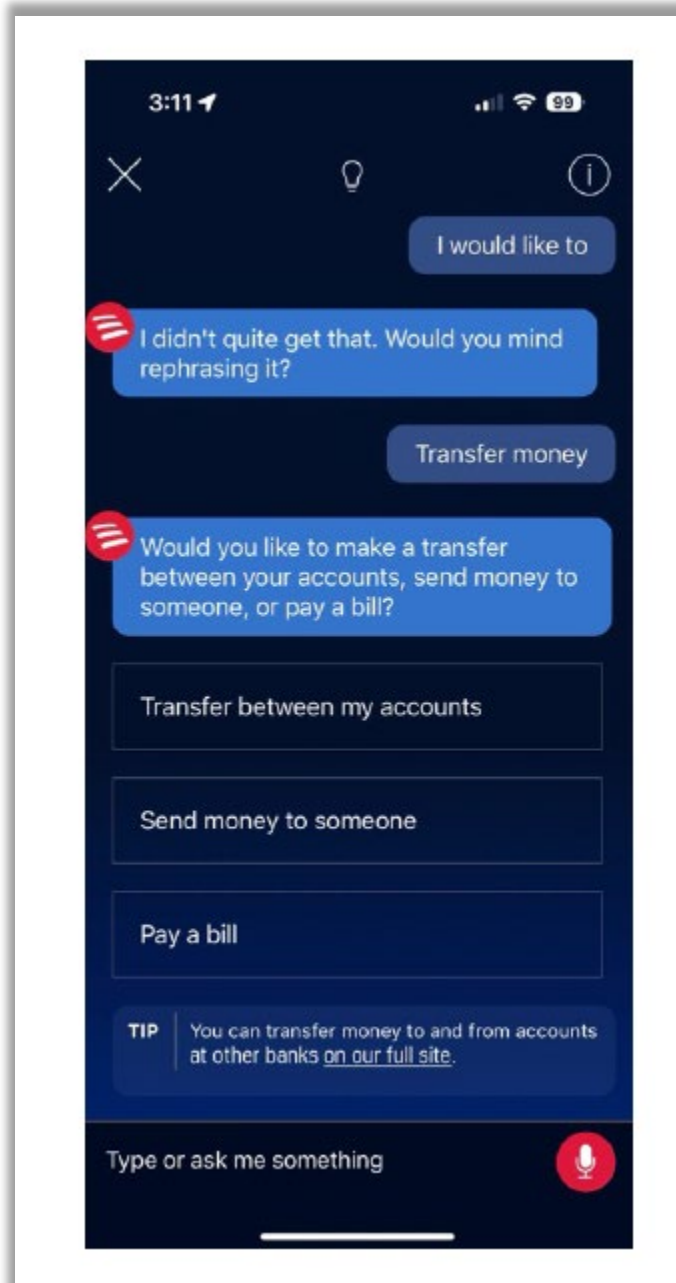
Coffman teaches “*generating ... a merged transcription*” because Coffman discloses “*merg[ing]*” input streams. For example: “**combine input keyed on the keyboard, handwritten input and speech input.**” Ex.1012, 38:2–4. Coffman

continues:

merg[ing] ... the output of multiple decoders[] [or] ASCII transcription. Each input stream is converted into its ASCII transcription and aligned with input time- marks via speech recognition processing When different input stream are associated to the same task, the transcripts are merged as follows. First, commands and queries are sorted based on the time marks and appended to a single data stream. Command formulation can be checked against [Finite State Grammar] FSG rules and re-sorted to satisfy the grammar rules.

Ex.1012, 38:8–14. Therefore, in Coffman, the combined keyboard-typed input and the transcribed speech input taken together serve as “*a merged transcription.*” As Dr. Houh explains, Coffman provides further details of merging input streams. Ex.1003, ¶148 (citing Ex.1012, 38:8–19).

Dialect’s Complaint, consistent with Coffman’s teachings identifies: “*Erica* transcribed a non-speech communication (the words “I would like to” which were input via the keyboard interface) and the speech communication (the words “Transfer money” which were input as speech via the microphone interface,) creating a merged transcription.” Ex.1007, ¶153.



Id; Ex.1003, ¶149.

Therefore, Maes (based on Coffman⁹) renders this claim element obvious for

⁹ See §VII.B.

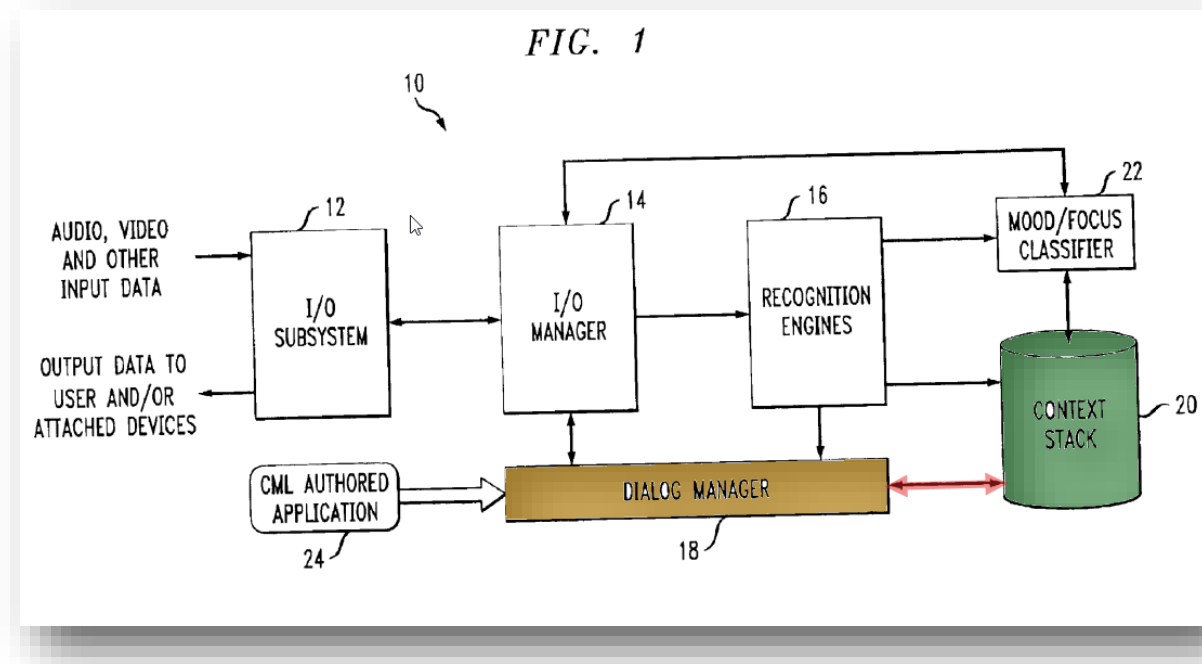
this additional reason. Ex.1003, ¶150.

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

Maes explains storing data relating to events (“*a plurality of ... entries*”) in a context stack (such as context stack 817 in Figure 8 or context stack 20 in Figure 1) (“*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* Ex.1005, 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.1005, Figure 1 (annotated); Ex.1003, ¶152.

Referring to 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.

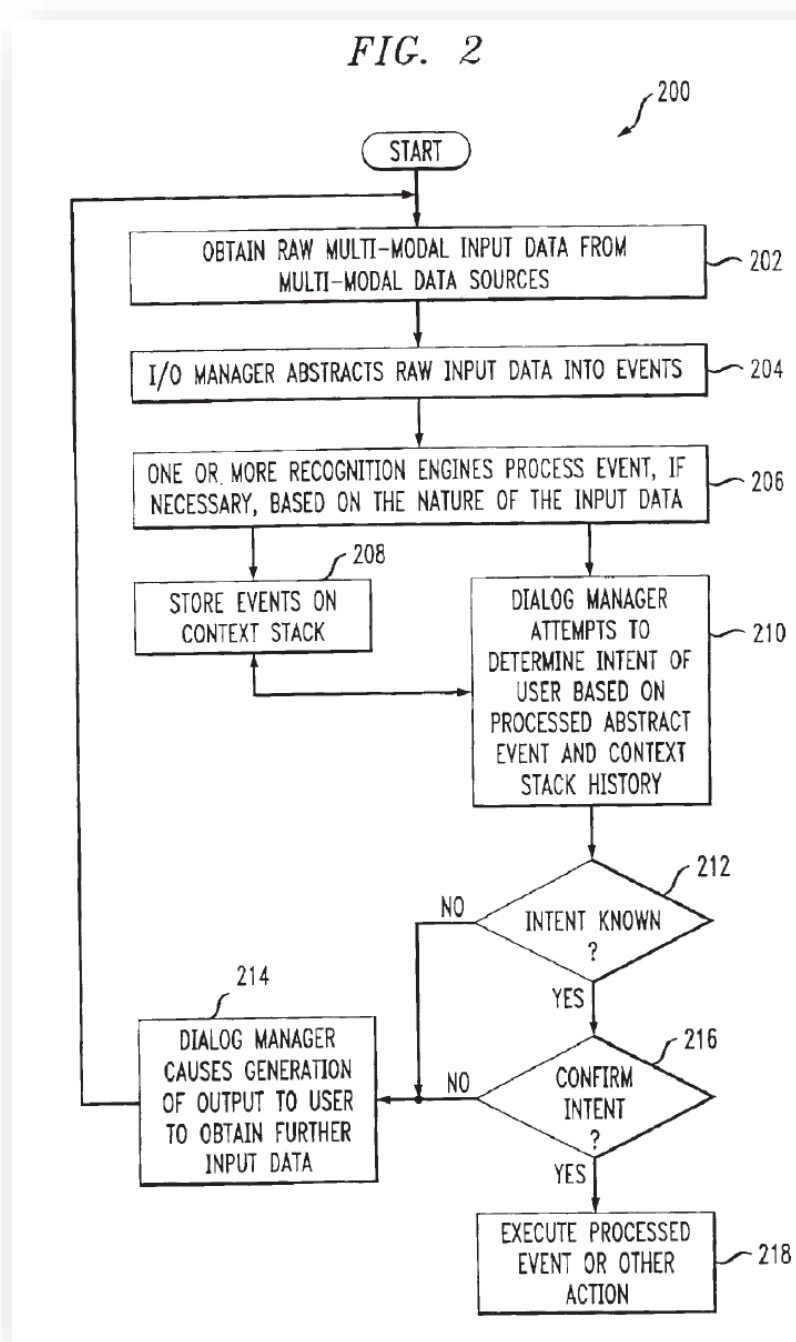
[T]his recognized spoken utterance event is stored on the context stack. Then, when the recognized gesture event (e.g., pointing to the radio) is received, the dialog manager takes this event and the previous spoken utterance event stored on the context stack and makes a determination that the user intended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

Ex.1005, 8:37–42; Ex.1003, ¶153.

In this example, the results of the recognized events constitutes “*information*,” which is included in the aligned decoded script and visemes sequence associated with the spoken utterance “turn it on” with the accompanying lip movement (“*in the merged transcription*”). Maes clarifies that “the dialog” includes “conversational dialog comprising speech and other multi-modal I/O such as GUI keyboard, pointer, mouse, as well as video input ...” Ex.1005, 36:49–54. Thus, “a dialog” in Maes comprises speech and video data. Therefore, Maes discloses the dialog manager associating the results of the recognized events (“*information*”) included in the aligned decoded script and visemes sequence produced as a result of merging the transcriptions (“*in the merged transcription*”) with an organized/sorted context in a context stack such as context stack 817 in Figure 8 or context stack 20 in Figure 1 corresponding to an active dialog which serves as “*identifying an entry in a context stack*” and consequently, the associated context(s) stored in the context stack

corresponding to the user's intent in the active dialog constitutes "*the identified entry*," which "*matches*" the results of the recognized events ("*information*") included in the aligned decoded script and visemes sequence produced as a result of merging the transcriptions) ("*in the merged transcription*"). See also Ex.1005, 37:53–55. And as explained in connection with [12.5], the "*merged transcription*" corresponds to the outcome in which the transcribed audio information such as the decoded text/script is aligned or synchronized with the transcribed video information such as the visemes sequence. Ex.1003, ¶153.

For example, Maes describes Figure 2 (below) as "a flow diagram illustrating a referential ambiguity resolution methodology performed by a multi-modal conversational computing system." Ex.1005, 3:24–28.



Ex.1005, Figure 2; Ex.1003, ¶154.

Referencing Figure 2, Maes continues: “[i]n step 208, the recognized events ... are stored in a storage unit referred to as the context stack 20 [which] ... 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** [*“wherein the identified entry matches information in the merged transcription.”*] **and the historical interaction information stored in the context stack** and then determine and execute one or more application programs that effectuate the user's intention and/or react to the user activity” (*“identifying ... an entry in a context stack, from among a plurality of entries that are in the context stack”*). Ex.1005, 7:33–45; Ex.1003, ¶158; see also Ex.1003, ¶¶155–157.

Moreover, Maes explains that *“each [entry in a context stack is] indicative of context.”* “The context stack 817 comprises all the information associated with an application. Such information includes **all the variable, states, input, output and queries to the backend that are performed in the context of the dialog** and any extraneous event that occurs during the dialog. The context stack is associated with the organized/sorted context corresponding to each active dialog.” Ex.1005, 37:55–62; Ex.1003, ¶159.

Accordingly, Maes teaches or suggests *“identifying ... an entry in a context stack, from among a plurality of entries that are in the context stack and that are each indicative of context.”* Ex.1003, ¶160.

In view of the above-mentioned teachings, Maes discloses that upon searching the context stack, the associated context(s) stored in the context stack corresponding to the user's intent in the active dialog (“*the identified entry*”) is provided by the dialog manager when such context “could be associated” (“*matches*”) the results of the recognized events (“*information*”) included in the aligned decoded script and visemes sequence (“*in the merged transcription*”), created as a result of aligning/synchronizing the abstracted data (e.g., decoded text/script) processed from the spoken utterance with the abstracted data (e.g., visemes sequence) processed from the lip movement of the user accompanying the spoken utterance. *See* Ex.1005, 7:60–8:4; *see also* Ex.1005, 21:50–54 (“**[T]he script is aligned with the visemes.** A rapid (or other) alignment may be performed in a conventional manner ... to **synchronize the two information streams.**” (*i.e.*, “*merged transcription*”); *see also* explanation for “*merged transcription*” in connection with [12.5] and [12.6]. Ex.1003, ¶161.

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

“organized/sorted contexts corresponding to multiple dialogs” (“*a plurality of entries*”) “*that are each indicative of context*” of a dialog. Therefore, Maes teaches or suggest “*a plurality of entries ... in the context stack and that are each indicative of context.*” Ex.1003, ¶162.

Additionally, as explained above in connection with [12.1], Maes discloses “*one or more processors.*” Ex.1003, ¶163.

To the extent Patent Owner contends that Maes does not disclose or suggest “*identifying... an entry in the context stack that matches information in the merged transcription,*” Coffman¹⁰ (incorporated by reference in Maes) discloses this element. Ex.1005, 9:30–37 (incorporating by reference specific features of Coffman identified as PCT/US99/22927); Ex.1003, ¶164.

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

¹⁰ Petitioner notes that the Board did not provide analysis addressing Coffman disclosing this element in IPR2024-00751. Institution Decision re IPR 2024-00751, 13–15.

that searching for the context includes comparing transcribed user inputs against information stored in the context stack:

Based on the modality (pointer, keyboard, file, speech), the task dispatcher 402 redirects the stream to the appropriate conversational subsystems or conventional subsystem with **speech inputs being transcribed/understood. The output of these subsystems is run down the context stack 405 to extract the active query and complete it.**

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

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

Ex.1012, 42:1–3. Thus, Coffman's teachings of searching the context stack until an acceptable match is found between the transcribed user inputs and data stored on the context stack serve as "*identifying...an entry in the context stack that matches information in the merged transcription.*" Accordingly, based on Maes and the incorporation of Coffman, a POSITA would have recognized that the results of the

recognized events included in the aligned decoded text/script and visemes sequence (“*the merged transcription*”) produced as a result of processing the user’s multi-modal input in Maes would be compared against data stored on the context stack, per Coffman’s teachings, to “*identify[]...an entry in the context stack that matches information in the merged transcription.*” Ex.1003, ¶165.

Therefore, Maes teaches or suggests this element for this reason. Ex.1003, ¶166.

Additional Mapping

As explained in [12.2], [12.3], [12.4], and [12.5], in Maes, keyboard-typed input constitutes an additional example of “*a non-speech input*” and further, as explained in [12.5], Coffman teaches the “*generating ... a merged transcription*” in which the combined keyboard-typed input and the transcribed speech input together constitute “*a merged transcription.*” As further explained in [12.6], Coffman’s teachings of searching the context stack until an acceptable match is found between the transcribed user inputs and data stored on the context stack serve as the “*identifying...*” step. Ex.1012, 42:1–3. A POSITA would have understood from the above-mentioned teachings of Coffman that the combined keyboard-typed input and the transcribed speech input taken together (“*the merged transcription*”) would be compared against data stored on the context stack to “*identify[]...an entry in the context stack that matches information in the merged transcription.*” Ex.1003, ¶167.

Therefore, for this additional reason, Maes (based on Coffman¹¹) renders obvious this claim element. Ex.1003, ¶168.

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

Initially, the 607 Patent describes a “*domain agent*” as follows: “In some embodiments ... generic and domain specific behavior and information may be organized into domain agents 156 [such that they] ... may provide complete, convenient and re-distributable packages or modules for each application area.” Ex.1001, 27:14–20. Thus, “*domain agent[s]*” in the 607 Patent broadly refer to software modules that are specific to each application area. Ex.1003, ¶170.

As explained in connection with element [12.6], in Maes, the associated context(s) stored in the context stack corresponding to the user’s intent in the active dialog constitutes “*the entry in the context stack*” and which is ascertained in the “*identifying ...*” step. Furthermore, for the reasons below, Maes discloses “*identifying ... a domain agent associated with the entry in the context stack*.”¹² For example, Maes teaches:

¹¹ See §VII.B.

¹² Petitioner notes that the Board denied institution of the 607 Patent in IPR2024-00751 based on finding that there is no disclosure in Coffman relating to determining a matching application. Institution Decision re IPR2024-00751, 20-22. Petitioner relies upon Maes for this functionality.

the system ... attempts to determine the user intent based on the current event and the historical interaction information stored in the context stack and then determine and **execute one or more application programs that effectuate the user's intention and/or react to the user activity. The application depends on the environment that the system is deployed in.**

Ex.1005, 7:40–46; Ex.1003, ¶171.

As such, Maes's teachings of application programs (effectuating the user's intention and depending on the environment that the system is deployed) are analogous to the 607 Patent's description of "*domain agent*," and therefore Maes's "application programs" or "applications" serve as "*domain agent[s]*." See also Ex.1005, 30:67 ("**run applications**"), 8:1–2 ("**search of the existing applications**"). Maes explains this in greater detail describing that "the conversational dialog manager ... **determines** what information the user has, which inputs the user presents, and **which application(s) should handle the user inputs.** ... The DMA function [of the dialog manager] handles multi-modal I/O events to: (1) **determine the target application or dialog (or portion of it)** and (2) use past history and context" and further, in Maes, "**the historical information (e.g., past events) [are] stored in the context stack [associated with the organized/sorted context corresponding to each active dialog (or deferred dialog-agents/memorization)]**." See Ex.1005, 36:49–37:3, 37:51–38:2; see also Ex.1005,

7:60–8:29. Indeed, “[t]he context stack [] comprises all the information associated with an application.” Ex.1005, 38:51-56; Ex.1003, ¶172.

By teaching that Maes’s dialog manager determines which application(s) should handle the user inputs, and further that the DMA function of the dialog manager determines the target application or dialog (or portion of it), Maes expressly discloses the step of “*identifying ... a domain agent.*” Therefore, Maes discloses “determining” (“*identifying*”) “the target application or dialog (or portion of it)” (“*a domain agent*”) “*associated with*” the associated context(s) stored in the context stack corresponding to the user’s intent in the active dialog (“*the entry*”) “*in*” the context stack (such as context stack 817 in Figure 8/context stack 20 in Figure 1) (“*the context stack*”). See also Ex.1005, 37:53–55. Accordingly, Maes teaches or suggests “*identifying ... a domain agent associated with the entry in the context stack.*” Ex.1003, ¶173.

Additionally, as explained above in connection with [12.1], Maes discloses “*one or more processors.*” Ex.1003, ¶174.

Therefore, Maes teaches or suggests this element. Ex.1003, ¶175.

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

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

determines what information the user has, which inputs the user presents, and which application(s) should handle the user inputs.

Ex.1005, 36:49–59. Specifically:

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

Ex.1005, 36:59–37:3; *see also* Ex.1005, 4:6–22. Maes’s system processes inputs relating to an I/O event representative of the user’s spoken utterance (audio portion) and the accompanying lip movement of the spoken utterance (video portion). Ex.1005, 4:30–32, 6:43–50, and 7:63–8:29. Therefore, Maes discloses “determin[ing]” (“*determining*”) “a user’s query” (“*a request*”) (associated with the multi-modal data in the form of audio input data, video input data, and possibly other types of input data) “*based on*” processing the user input multi-modal data yielding “a more generalized representation of the data” (e.g., the decoded script and visemes

sequence) included in “the aligned decoded script and visemes sequence” (“*the merged transcription*”). Ex.1003, ¶177.

As Dr. Houh explains, Maes’s system advantageously:

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

Ex.1003, ¶ (citing Ex.1005, 2:54–67). Therefore, Maes discloses “*determining ... a request based on the merged transcription.*” Ex.1003, ¶177.

Additionally, as explained above in connection with [12.1], Maes discloses “*one or more processors.*” Ex.1003, ¶178.

Therefore, Maes teaches or suggests this element. Ex.1003, ¶179.

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

First, as explained in connection with [12.7], Maes's "application programs" constitutes "*the domain agent*" and further, Maes discloses that "[t]he DMA [function of the dialog manager] ... determines the target of the event and ... then launches the action associated to the user's query." Ex.1005, 36:49–37:3. For example,

[i]f the dialog manager determines with some predetermined degree of confidence that the application it selects is the one which will effectuate the users desire, the dialog manager executes the next step of the multi-modal dialog (e.g., ... launches the execution of an action associated to a fully understood multi-modal request from the user) of that application based on the multi-modal input. That is, the dialog manager selects the appropriate device (e.g., radio) activation routine and instructs the I/O manager to output a command to activate the radio.

Ex.1005, 8:8–8:19. Maes identifies that its system provides the capability to: "(i) **determine an object, application or appliance** addressed by the user; (ii) ... **determine if the user is actively focused on an appropriate application and, on that basis, to determine if an action should be taken;** ... and (iv) ... **initiate and/or adapt some behavior/service/appliances accordingly.**" Ex.1005, 2:54–64.

Therefore, launching or executing the action (e.g., selecting an application such as a radio activation routine) associated with the user's query (e.g., comprising a request for launching an application such as a radio activation routine) serves as "*communicating ... the request to the domain agent.*" Ex.1003, ¶181.

Further, because Maes discloses that "the dialog manager selects the appropriate device (e.g., radio) activation routine and instructs the I/O manager to output a command to activate the radio," (Ex.1005, 8:15–19), a POSITA would have therefore recognized that "the application" ("*the domain agent*"), such as a "radio activation routine," would be accordingly activated, or otherwise some behavior/service/appliances would be initiated and/or adapted to "*generate a response to the user*" (e.g., cause the radio to be turned on). *See generally* Ex.1005, 7:60–8:19; 2:54–64. Indeed, Maes teaches that the application (e.g., "*domain agent*") having an "associated business logic ... with which the user interacts." Ex.1005, 2:65–67. Dr. Houh explains that a POSITA would have understood that, in the process of interacting with the user, e.g., the business logic of an application would "*generate a response to the user.*" *See* Ex.1005, 2:65–67; Ex.1003, ¶182.

Therefore, Maes teaches or suggests "*wherein the domain agent is configured to generate a response to the user.*" Ex.1003, ¶182.

Additionally, as explained above in connection with [12.1], Maes discloses "*one or more processors.*" Ex.1003, ¶183.

Therefore, Maes teaches or suggests this element. Ex.1003, ¶184.

2. Claim 13

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

First, Maes describes features such as acoustic feature vectors included in processed audio information that are used to “*generat[e] the speech-based transcription*” that “**can be extracted and decisions dynamically returned by the models,**” which explicitly discloses that Maes uses more than one “*cognitive model.*” Ex.1005, 43:53–54. Ex.1003, ¶186.

Furthermore, Maes’s system (“*the device*”) can include “a dynamic classification module 1040 which performs dynamic classification of the user” such that the system is “configured to modify behavior of the voice system 1000 **based on at least one user attribute** which has been determined based on at least one acoustic feature extracted from the user’s speech.” Maes, 44:5–11. Maes provides several examples of user attributes such as gender, age, accent, native language, and others. Ex.1003, ¶187 (citing Ex.1005, 39:50–55).

Considering, e.g., native language as an attribute, Maes teaches that that “**at the creation of the models for each native language, one employs a speaker or speakers who are tagged with that language as their native language**” which

explicitly discloses that Maes uses “*a ... model that includes information on ... a plurality of users.*” Ex.1005, 40:18–23. Thus, in Maes, “produc[ing] decoded text” (“*speech-based transcription[s]*”) can involve classifying the user’s native language based on a native language model created using multiple speakers who are tagged with that language as their native language (“*a second cognitive model that includes information on one or more interactions between the device and a plurality of users of the device*”). Ex.1003, ¶187.

For the additional reasons below, Maes teaches using “*a ... model that includes information on ... interactions between the device and a plurality of users.*” In Maes’s system, “[s]peaker verification may be performed in a similar manner, however, **the input acoustic data is compared to determine if the data matches closely enough with stored models.** [A]t enrollment, **the input speech is collected for a speaker to build the mixture gaussian model M_i that characterize each speaker.**” Ex.1005, 24:31–45. “Thus, the above approach finds the closest matching model and **the person whose model that represents is determined to be the person whose utterance is being processed.**” Ex.1005, 24:28–30; Ex.1003, ¶188.

Additionally, Maes, based on Ittycheriah (see Ex.1005, 19:43–49) teaches or suggests that “*the generating ...*” is “*based on a second cognitive model that includes information on one or more interactions between [a] device and a plurality of users*”

of the device.” As explained in connection with [12.4], 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 within a user-specific partition within a database as hypothesis for decoding a word in a current interaction with that user and therefore—such database (“*a first cognitive model*”) used as a model includes information on one or more prior interactions between a single user/speaker and a device (such as Maes’s system). Ex.1003, ¶190 (citing Ex.1006, 4:19–32 and 1:55–58).

Additionally, Ittycheriah teaches a speaker classification approach (“*a second cognitive model*”) that involves storing keywords and scores from past history of a class of speakers (“*includes information on one or more interactions between [a] device and a plurality of users of the device.*”) Ex.1006, 5:5–19 (describing the speaker classification approach, including techniques based on “past history ... associated with ... speakers in a class.”) Ex.1003, ¶191.

Therefore, Maes teaches or suggests this claim. Ex.1003, ¶192.

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

To the extent one asserts that Coffman and Ittycheriah are not expressly and/or particularly incorporated by reference into Maes as if they were effectively the same document, claims 12 and 13 would have nonetheless been obvious in view of the combination of Maes, Coffman and Ittycheriah for the 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, ¶193.

IX. THE BOARD SHOULD NOT EXERCISE ITS DISCRETION AND DENY INSTITUTION

A. The Board Should Not Deny Institution Under 35 U.S.C. § 325(d)

Maes was neither cited nor considered by the Examiner during the prosecution of the 607 Patent, and no arguments related to these disclosures were considered. Coffman and Ittycheriah, which Maes incorporates by reference, are also relied upon to the extent Patent Owner contends that Maes does not teach or suggest certain claim limitations. A “relative” of Coffman was listed on an IDS (Ex.1002, 168 (describing Examiner’s initials; Ex.1013, cover), but was also not evaluated substantively. *CrowdStrike, Inc. v. Webroot Inc.*, IPR2023-00126, Paper 9 at 14 (PTAB May 5, 2023) (“[the Board] cannot determine the extent to which [cited reference] was evaluated during the examination because it was never used in a rejection” and because “[cited reference] was not the basis of a rejection, there is no overlap of arguments.”); *Scientific Design Co. Inc. v. Shell Oil Co.*, IPR2022-00158, Paper 7 at 24-26 (PTAB Apr. 4, 2022) (declining to exercise discretion under § 325(d) based on “credit[ing] Petitioner’s argument that although the references ‘were cited in an IDS, together with 111 references, none of them was used in any rejections, or otherwise addressed by the examiner’”); *Bowtech, Inc. v. MCP IP*,

LLC, IPR2019-00379, Paper 14 at 18 (PTAB Jul. 3, 2019). The first condition of *Advanced Bionics* is therefore not satisfied.

Nevertheless, the Examiner materially erred by not substantively evaluating the Maes reference (alone or in conjunction with Coffman and/or Ittycheriah) against the 607 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).

Further, the Examiner did not have the opportunity to review Dr. Houh’s detailed expert testimony supporting this petition. The Board has found the *Becton* factors favor institution where even a single noncumulative secondary reference distinguished the prior-art ground, regardless of other overlapping references. *Oticon Medical AB et al. v. Cochlear Limited*, IPR2019-00975, Paper 15 at 9-20 (PTAB Oct. 16, 2019) (precedential). The Board has declined to deny institution where Petitioners relied upon prior art cited on the face of the challenged patent but “not applied by the examiner ... in any rejection of claims.” *Comcast Cable*

Communs., LLC v. Promptu Sys. Corp., Case No. IPR2018-00342, Paper 13 at 17 (PTAB July 19, 2018).

B. The Board Should Not Deny Institution Under 35 U.S.C. § 314(a)

1. Factor 1: Stay

Factor is **neutral or weights against discretionary denial** because no request for stay has been filed in the litigation involving BOA, and Microsoft is not a defendant in that litigation. *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 not a defendant in litigation involving this patent, although BOA—an RPI—is. Trial for 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 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 that are currently scheduled for January 26, 2026, so it appears likely that the trial date will slip. Ex.1029. Moreover, the trial date is 10 months away and much can change during this time. *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) (granting institution when trial was to be held six months before the FWD).

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). For at least this reason, the Board need not reach the question on Factor 6.

Even so, the merits of the Petition are particularly compelling—for example, Maes (and, by incorporation Coffman and Ittycheriah) is a strikingly similar disclosure to that of the 607 Patent, is directed at the same problem, proposes the same solutions as the 607 Patent, and discloses nearly identical functionality as the claimed “*method for processing 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 12 and 13 of the 607 Patent are unpatentable for the reasons discussed above.

Dated: March 7, 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 13,681 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 7, 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 7th 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 7, 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 12	
[12.0]	A method for processing natural language inputs, comprising:
[12.1]	receiving, by one or more processors, a multi-modal natural language input from a user, the multi-modal natural language input including a natural language utterance and a non-speech input;
[12.2]	generating, by the one or more processors, a non-speech transcription from the non-speech input;
[12.3]	identifying, by the one or more processors, the user who provided the multi-modal natural language input;
[12.4]	generating, by the one or more processors, a speech-based transcription based on a cognitive model associated with the user, wherein the cognitive model includes information on one or more prior interactions between the user and the device;
[12.5]	generating, by the one or more processors, a merged transcription from the speech-based transcription and the non-speech transcription;
[12.6]	identifying, by the one or more processors, an entry in a context stack, from among a plurality of entries that are in the context stack and that

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