

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

MICROSOFT CORPORATION,
Petitioner,

v.

DIALECT, LLC,
Patent Owner.

Case IPR2025-00655
Patent 7,640,160

PATENT OWNER'S PRELIMINARY RESPONSE

TABLE OF CONTENTS

I.	INTRODUCTION	1
II.	THE ’160 PATENT DISCLOSES AND CLAIMS OPERATIONS THAT FOLLOW TRANSCRIPTION OF AN UTTERANCE	2
A.	The ’160 Patent Discloses and Claims Features not Disclosed or Suggested by the Earlier Kennewick Reference	2
B.	The ’160 Patent Uses Grammar Expression Entries and Expected Contexts in a Method for Interpreting Natural Language Utterances	2
III.	THE BOARD NEED NOT REACH ANY CLAIM CONSTRUCTION ISSUES TO DENY THE PETITION.....	7
IV.	PETITIONER’S CHALLENGE FAILS ON THE MERITS	8
A.	The Petition’s Treatment of Maes and Elements [12.2] and [12.3]	9
B.	Maes Does Not Disclose the Recitations of Elements [12.2] and [12.3]	9
V.	CONCLUSION.....	15

TABLE OF AUTHORITIES

Rules

37 C.F.R. § 42.107	1
37 C.F.R. § 42.120(a).....	1

PATENT OWNER’S EXHIBIT LIST

Exhibit No.	Description
2001	First Amended Docket Control Order, <i>Dialect, LLC v. Bank of America, N.A.</i> , Civ. No. 2:24-cv-207, Dkt. 52 (E.D. Tex. Aug. 28, 2024)
2002	Comparison of Districts Within the Fifth Circuit — 12-Month Period Ending September 30, 2024 (available at https://www.uscourts.gov/sites/default/files/2025-02/fcms_na_distcomparison1231.2024.pdf)
2003	Defendant’s Invalidity Contentions, <i>Dialect, LLC v. Bank of America, N.A.</i> , Civ. No. 2:24-cv-207 (E.D. Tex. Sept. 12, 2024)
2004	Ex. A (Maes) to Defendant’s Invalidity Contentions, <i>Dialect, LLC v. Bank of America, N.A.</i> , Civ. No. 2:24-cv-207 (E.D. Tex. Sept. 12, 2024)
2005	Ex. A (Ross) to Defendant’s Invalidity Contentions, <i>Dialect, LLC v. Bank of America, N.A.</i> , Civ. No. 2:24-cv-207 (E.D. Tex. Sept. 12, 2024)
2006	Ex. A (Coffman) to Defendant’s Invalidity Contentions, <i>Dialect, LLC v. Bank of America, N.A.</i> , Civ. No. 2:24-cv-207 (E.D. Tex. Sept. 12, 2024)
2007	U.S. Pat. 7,137,126 (PCT counterpart to Coffman)
2008-2018	Reserved
2019	Return of Summons, <i>Dialect, LLC v. Bank of America, N.A.</i> , Civ. No. 2:24-cv-207 (E.D. Tex. Mar. 27, 2024)
2020	“IPBC Global 2025: Acting USPTO Director says IPR use needs to change” (June 10, 2025) (available at https://www.iam-media.com/article/ipbc-global-2025-acting-uspto-director-says-ipr-use-needs-change)
2021-2099	Reserved
2100	File History for U.S. Patent 9,263,045
2101	File History for U.S. Patent 9,865,262
2102	File History for U.S. Patent 10,002,607
2103	File History for U.S. Patent 10,424,290

Pursuant to 37 C.F.R. § 42.107, Patent Owner Dialect, LLC (“Dialect” or “Patent Owner”) files this preliminary response to the Petition, setting forth reasons why the Petition for *inter partes* review (“IPR”) of claims 12 and 13 of U.S. Patent No. 7,640,160 (the “’160 patent”), as requested by Microsoft Corporation (“Petitioner”), should be denied.¹

I. INTRODUCTION

Petitioner’s challenge to the ’160 patent should be denied because it rests on a fundamental misunderstanding of the prior art. The ’160 patent claims methods for interpreting natural language utterances by receiving a transcription and then scoring contextual matches against that completed transcription to determine meaning. Petitioner incorrectly argues that the Maes reference discloses this claimed context-scoring process, when Maes actually describes an entirely different technology—a multi-modal speech recognition system that uses audio and video inputs to generate accurate transcriptions in the first place. By conflating the preliminary process of creating transcripts with the claimed process of analyzing completed transcripts for contextual meaning, Petitioner has failed to present a compelling unpatentability challenge. This fundamental error in understanding the prior art renders the Petition facially deficient.

¹ By submitting this Preliminary Response, no waiver of any argument is intended by Patent Owner. Patent Owner will have a right to file a complete response to the Petition, should the Board institute *inter partes* review. 37 C.F.R. § 42.120(a).

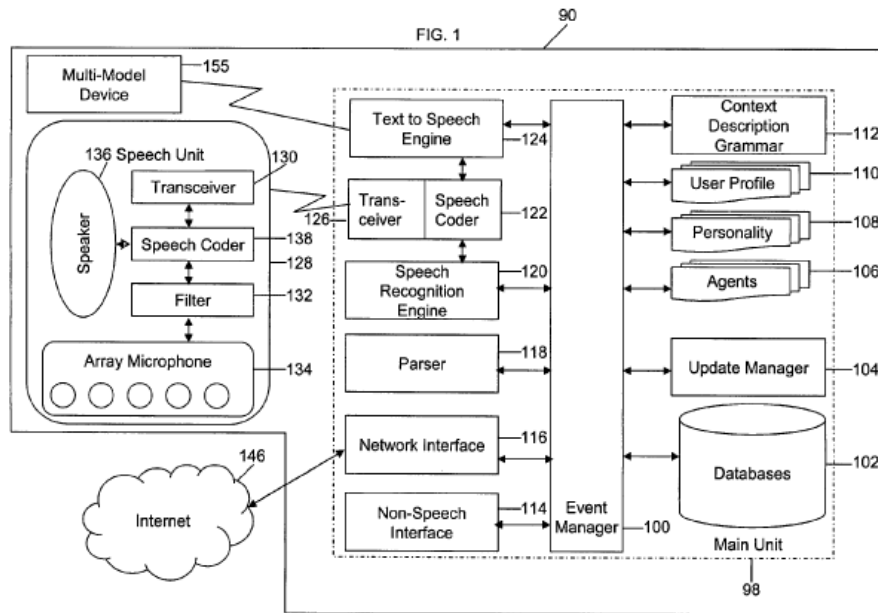
II. THE '160 PATENT DISCLOSES AND CLAIMS OPERATIONS THAT FOLLOW TRANSCRIPTION OF AN UTTERANCE

A. The '160 Patent Discloses and Claims Features not Disclosed or Suggested by the Earlier Kennewick Reference

The '160 patent is entitled, “Systems and Methods for Responding to Natural Language Speech Utterance,” and is related to speech interface processing that determines an intent and corrects false recognitions of natural language utterances. Contrary to Petitioner's theories, the '160 patent discloses an efficient and reliable way to identify a context for the natural language utterances that is not disclosed in any of the asserted art relied upon by Petitioner.

B. The '160 Patent Uses Grammar Expression Entries and Expected Contexts in a Method for Interpreting Natural Language Utterances

FIG. 1 of the '160 patent represents an “overall diagrammatic view according to one embodiment of the invention”:



'160 patent, Fig. 1, 9:64-65. The system includes speech unit 128, speech recognition engine 120, context description grammar module 112, parser 118, and agents 106. Speech unit 128 includes a microphone to receive a spoken utterance from a user. '160 patent, 11:55-57. After initial processing, the speech is passed to speech recognition engine 120 for processing using the context description grammar module 112. '160 patent, 12:23-25.

Then, “[a]ny recognized information may be processed by the parser 118, which transforms information into complete algorithms and questions using data supplied by knowledge agents.” '160 patent, 12:25-28. “The knowledge agents may then process the commands or questions.” '160 patent, 12:32-36. The agents may update a context stack, which includes an ordered list of command contexts, to enable follow-up requests. '160 patent, 13:57-59.

To select the correct agent to process the command or question, “[t]he parser 118 uses a scoring system to determine the most likely context or domain for a user’s question and/or command. ’160 patent, 21:10-12. To accomplish the scoring, the context description module 112 uses context description grammar, which may be derived from content packages in the system. ’160 patent, 16:46-54. An initial or default context is typically supplied for each command or question, and the command or question may include a grammar for the management and evaluation of the context stack. ’160 patent, 16:51-54. As noted above, the context stack includes an ordered list of command contexts. ’160 patent, 13:57-59.

A user’s command or request may be presented or converted into a text message, which may be searched for text combinations. ’160 patent, 13:29-35, 13:44-47. The text combinations may be compared against entries in a context description grammar associated with each domain agent 106. ’160 patent, 13:47-49. “If a match is identified between an active grammar in the context description grammar and the command and/or request, then the match may be scored.” ’160 patent, 13:49-51. “The score is determined from weighing a number of factors, including the user profile 110, the domain agent’s data content and previous context. Based on this scoring, the system 90 invokes the correct agent.” ’160 patent, 21:10-15.

In addition to using “grammar expression entries” in a context description grammar to accomplish the scoring, the ’160 patent also uses “expected contexts” stored in a context stack. If the process of comparing the text combinations against active grammar in the context description grammar yields no match or only a partial match, the system can semantically broaden the search. ’160 patent, 13:60-64. “The knowledge enhanced speech recognition may access a set of expected contexts that are stored in a context stack to determine a most likely context.” ’160 patent, 13:66-14:2. “The knowledge-enhanced speech recognition may use context specific matchers that are able to identify context such as time, location, numbers, dates, categories (e.g., music, movies, television, addresses, etc.).” ’160 patent, 14:2-6.

Thus, after a transcription of a natural language utterance is received, the ’160 patent discloses comparing text combinations contained in the transcription of a natural language utterance against both the grammar expression entries and one or more expected contexts to identify matching contexts. Once identified, the ’160 patent discloses scoring each identified matching context.

Claim 12 of the ’160 patent includes these requirements, claiming (with emphasis added):

12. A method for interpreting natural language utterances using knowledge-enhanced speech recognition engine, wherein the knowledge-enhanced speech recognition engine is configured to

determine an intent and correct false recognitions of the natural language utterances, comprising:

receiving a transcription of a natural language utterance at a computer comprising the knowledge-enhanced speech recognition engine;

identifying one or more contexts that completely or partially match *one or more text combinations contained in the transcription*, wherein identifying the matching contexts includes *comparing the text combinations* against the grammar expression entries in the context description grammar and against one or more expected contexts stored in a context stack;

scoring each of the identified matching contexts;

selecting the matching context having a highest score to determine a most likely context for the utterance; and

communicating a request to a domain agent configured to process requests in the most likely context for the utterance, the request formulated using at least one grammar expression entry in the context description grammar.

'160 patent, claim 12 (emphasis added).

Similarly, claim 13 of the '160 patent requires:

13. A method for processing natural language utterances, comprising:
receiving a natural language utterance at a computer comprising a multi-pass speech recognition module;

transcribing the utterance using the multi-pass speech recognition module, the multi-pass speech recognition module configured to transcribe the utterance into text;

identifying one or more contexts that completely or partially match *one or more text combinations contained in the text of the transcribed utterance*, wherein identifying the matching contexts includes *comparing the text combinations* against the grammar expression entries in the context description grammar and against one or more expected contexts stored in a context stack;

scoring each of the identified matching contexts;

selecting the matching context having a highest score to determine a most likely context for the utterance; and

communicating a request to a domain agent configured to process requests in the most likely context for the utterance, the request formulated using at least one grammar expression entry in the context description grammar.

'160 patent, claim 13 (emphasis added).

III. THE BOARD NEED NOT REACH ANY CLAIM CONSTRUCTION ISSUES TO DENY THE PETITION

The Petitioner identifies no terms as requiring construction. *See* Pet., 12.

Without waiving its right to submit constructions in a Patent Owner Response, Patent Owner submits that the Board need not engage in any pre-institution claim construction. For the reasons presented in Section IV below, Petitioner's grounds include fundamental flaws and fail to establish the disclosure of every claimed

element in each independent claim. These findings do not require the Board to conduct any construction of the above-identified terms.

IV. PETITIONER'S CHALLENGE FAILS ON THE MERITS

In this IPR, Petitioner asserts that claims 12 and 13 are obvious in view of Maes² and Ross³ (Ground 1) and in view of Maes, Coffman,⁴ and Ross (Ground 2).

As introduced above, the '160 patent provides techniques that mitigate against the risk of false recognitions when processing natural language utterances. As part of the claimed method for interpreting natural language utterances, challenged claim 12 of the '160 patent includes the first step of “receiving a transcription of a natural language utterance,” followed by, in element [12.2], the step of “comparing” text combinations *contained in a transcription* “against the grammar expression entries in the context description grammar and against one or more expected contexts stored in a context stack.” Those contexts that match the text combinations from the received transcription are then scored, and a most likely context is determined based on the scoring. Element [12.3] requires “scoring each of the identified matching contexts,” which are determined based on comparisons of text combinations in a transcription.

² Ex. 1005, U.S. Patent No. 6,964,023 (“Maes”).

³ Ex. 1022, U.S. Patent App. Pub. No. 2002/0133354 (“Ross”).

⁴ Ex. 1012, International Patent Application Publication No. WO 00/20962 (“Coffman”).

A. The Petition's Treatment of Maes and Elements [12.2] and [12.3]

Looking first to the Petition's consideration of element [12.2], Petitioner asserts that Maes processes a spoken utterance to generate a "decoded text (or, script)" and therefore discloses a "transcription of a natural language utterance" as claimed. Pet., 35. The Petition also relies on the Maes reference to disclose element [12.3]. Pet., 54-57. In particular, Petitioner asserts that Maes "teaches scoring one or more candidate contexts such as phonemes, which Maes describes as 'sub-phonetic or acoustic units of speech' (Ex.1005, 12:29–33) or equivalently as 'portions [of dialogs]' (Ex.1005, 7:60–8:4)." Pet., 55. Petitioner then discusses this process of scoring phonemes with respect to Maes's Figure 4 and disclosures in columns 12 and 18. Pet., 55-56. Petitioner concludes that calculating likelihood scores "for each considered phoneme in a given context in the form of probabilities, therefore, serves as" the claimed feature of [12.3]. Pet., 56. This is completely wrong.

B. Maes Does Not Disclose the Recitations of Elements [12.2] and [12.3]

Petitioner's error stems from failing to properly interpret Maes and its disclosures regarding phonemes, which is performed to generate an accurate transcript, not to later compare transcript text combinations with context entries. As identified both in the title of the invention and the Abstract, Maes is directed to receiving input via multiple modes, such as both audio and video inputs, to resolve

ambiguities in the inputs. This multi-modal input is also reflected in Maes's Figure 1, where in the input data includes "AUDIO, VIDEO AND OTHER INPUT DATA." Maes, FIG. 1. Figure 4, relied upon by Petitioner, is not directed to scoring contexts that match the text combinations from the received transcription, such as claimed in element [12.3], but rather to the use of both audio and video data to more accurately determine what was said by a speaker in the first place. Thus, the relied upon portions of Maes do not involve the step of scoring contexts that match with text combinations in a transcript, which necessarily comes after preparation of a transcript.

As Maes explains it, Figure 4 illustrates "a preferred embodiment of 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." Maes, 10:47-52 (emphasis added). This audio-visual speech recognition module "is able to locate and track a face and, more particularly, a mouth, to determine what is relevant visual information to be used in more accurately recognizing the accompanying speech provided by the speaker." Maes, 11:16-20. The module's auditory feature extractor 414 "receives an audio or speech signal" to yield—following processing steps—the "acoustic feature vectors" denoted as letter "A" in FIG. 4. Maes, 11:64-12:30. Maes's probability module then "yields likelihood scores for each

considered phoneme in the form of the probability that, given a particular phoneme or acoustic unit (au), the acoustic unit represents the uttered speech characterized by one or more acoustic feature vectors A.” Maes, 12:38-42. This scoring is still about determining whether the audio-visual speech recognition module has correctly recognized what a user said based on the audio signals received from the speaker.

Moving from using audio signals to recognize a user's utterance, Maes next describes the use of video by the audio-visual speech recognition module to do the same. Maes, 12:49-61. The “the visual speech feature extractor 422 extracts visual speech feature vectors (e.g., mouth or lip-related parameters), denoted in FIG. 4 as the letter V, from the face detected in the video frame by the active speaker face detector 418.” Maes, 17:51-55. This may include extracting “geometric vectors” related to the geometry of a speaker's mouth, such as “lip contour,” “curvature of inner as well as the outer lips,” and other mouth, teeth, and tongue positions to perform what Maes describes as “speech-reading.” Maes, 18:12-20; 18:38 (“speech-reading”).

Just as Maes used audio vectors in order to try to determine what a user said, i.e. to generate a transcript, Maes also uses the video vectors to lip read in an attempt to determine what a user said. Per Maes, “each phoneme associated with one or more visual speech feature vectors has a probability associated therewith

indicating the likelihood that it was that particular acoustic unit that was spoken in the video segment being considered.” Maes, 18:50-54.

The respective audio and video probabilities are used together to determine “a likelihood or other confidence measure being determined with regard to the recognized input.” Maes, 19:1-3. Maes uses both audio and video data inputs because some environments might yield low confidence of one type of input over the other. For example, in a noisy environment, the audio data might yield a lower confidence measure of a user's speech than the video data, while a quiet environment might yield a higher acoustic confidence measure. Maes, 19:50-20:10. As Maes confirms, these steps are part of the “translation process,” directed to correctly recognize the words that were spoken. Maes, 19:39-49. This corresponds to Maes's attempt to use both video and audio data to correctly recognize what a speaker is saying and to generate an accurate transcript, not to score contexts that match the text combinations from a received transcription.

This discrepancy is clear based on how the Petition defines context.

In relation to claim element [12.3], dealing with the scoring of contexts, Petitioner defines contexts as phonemes, which are sub-phonetic or acoustic units of speech:

Maes teaches scoring one or more candidate contexts such as phonemes, which Maes describes as “sub-phonetic or acoustic units of speech” (Ex.1005,

However, in the previous claim element [12.2.1], the Petitioner characterizes context as “data for events”:

previous spoken utterance event” of the user saying “turn it on. Data for both events
[“one or more contexts”] are stored on the context stack because Maes discloses “all
the variable, states, input, output and queries to the backend that are performed in
the context of the dialog and any extraneous event that occurs during the dialog” are
stored on the context stack. Ex.1005, 37:55–62. Therefore, in making a

Petitioner never reconciles these two different definitions of the context. Doing so is in fact impossible, because it would require showing how phonemes are identified by matching them against text combinations contained in the transcription, as required in [12.2.1]. This is impossible for two reasons. First, phonemes are used to **generate** said transcription. Maes confirms this purpose of phonemes, which is discussed in 12:29-48 in relation to Fig. 4, describing a block diagram of an “audio-visual speech recognition module” (3:33-34) whose purpose is to translate acoustic and visual signal to text. Second, element [12.2.1] requires that text combinations of context are compared against grammar entries.

Phonemes are not text. They are “sub-phonetic or acoustic units of speech” and as such, there are no “text combinations” to speak of.

Thus, Petitioner is completely wrong in its analysis of element [12.3] because these relied-upon portions of Maes are directed to the predicate step of using both audio and video data to ensure that a transcript is accurate and matches

what a user actually said. *See* Pet., 55 (“Maes discloses generating, for each considered phoneme in a given context, likelihood scores (indicating the likelihood that it was that particular phoneme that was spoken) ...”).” These relied-upon disclosures of Maes are not directed to scoring contexts that match with an already-prepared and received transcription. As such, these disclosures of Maes are irrelevant to the process claimed in step [12.3], and confirm that the Petition’s challenges fail.

With regard to Ground 1, Petitioner does not contend that secondary reference Ross is provided to remedy these above-discussed deficiencies of Maes with regard to step [12.3]. *See* Pet., 54-57 (relying solely on Maes for element [12.3]). Additionally, Petitioner’s incorrect understanding of the Maes reference as discussed above also infects its challenge to claim 13, which merely incorporates Petitioner’s mistaken understanding as presented in its claim 12 analysis. *See* Pet., 69.

Finally, Petitioner’s Ground 2 adds Coffman to the Ground 1 combination of Maes and Ross, but without adding any separate analysis. Indeed, Petitioner concedes that “Maes’s context stack disclosures are almost verbatim as Coffman’s. Pet., 12. Thus, the Petition adds Coffman in Ground 2 for the sole purpose of responding to the hypothetical argument that Coffman is separate from Maes; as Petitioner explains it, “[t]o the extent one asserts that Coffman is not

expressly and/or particularly incorporated by reference into Maes as if they were effectively the same document.” Pet., 70 (presenting Ground 2 in a single paragraph). As with Ground 1, Petitioner does not contend that secondary reference Coffman is provided to remedy the above-discussed deficiencies of Maes with regard to step [12.3]. See Pet., 54-57 (relying solely on Maes for element [12.3]). Thus, the above-discussed shortcomings of Ground 1 also infect Ground 2.

Neither challenge establishes a basis for institution, and each should be rejected for failure to establish any likelihood of invalidity.

V. CONCLUSION

For the reasons presented above, the Petition should be denied, and no *inter partes* review should be instituted.

Dated: July 15, 2025

/Wayne M. Helge/
Wayne M. Helge, Reg. No. 56,905
James T. Wilson, Reg. No. 41,439
BUNSOW DE MORY LLP
277 S. Washington St., Suite 210 #1088
Alexandria, VA 22314
T: (571) 208-0186
Email: whelge@bdiplaw.com
Email : jwilson@bdiplaw.com

Counsel for Patent Owner

CERTIFICATE OF WORD COUNT

The undersigned certifies that the foregoing PATENT OWNER'S PRELIMINARY RESPONSE complies with the type-volume limitation in 37 C.F.R. § 42.24(b)(1). According to the word-processing system's word count, the brief contains 2,958 words, excluding the parts of the brief exempted by 37 C.F.R. § 42.24(a).

By: /Wayne M. Helge/
Wayne M. Helge (Reg. No. 56,905)
Counsel for Patent Owner

CERTIFICATE OF SERVICE

I hereby certify that on July 15, 2025, a true and correct copy of the foregoing document was served via email, by consent, to Petitioner by serving the correspondence email addresses of record as follows:

WINSTON & STRAWN LLP
Alias (Winston-Microsoft-Dialect-IPR@winston.com)
Scott M. Border (sborder@winston.com)
Carson Swope (cswope@winston.com)
T.D. Goswami (tdgoswami@winston.com)

/Wayne M. Helge/
Wayne M. Helge (Reg. No. 56,905)
Counsel for Patent Owner