

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

AMAZON.COM SERVICES LLC,

Petitioner,

v.

VB ASSETS, LLC,

Patent Owner.

Case: IPR2025-01365

U.S. Patent No. 10,297,249

DECLARATION OF PADHRAIC SMYTH, PH.D.

TABLE OF CONTENTS

	Page
I. INTRODUCTION	1
II. BACKGROUND AND QUALIFICATIONS	1
III. COMPENSATION AND RELATIONSHIP WITH PARTIES.....	5
IV. MATERIALS CONSIDERED	5
V. BASIS OF OPINIONS	8
A. Legal Standard for Anticipation	8
B. Legal Standard for Obviousness.....	8
C. Level of Ordinary Skill in the Art	10
D. Claim Construction.....	11
VI. OVERVIEW OF THE '249 PATENT	12
A. Filing and Specification.....	12
B. Prosecution History	15
C. Related Litigation and Related Patents	18
VII. THE PRIOR ART RENDERS OBVIOUS THE CLAIMED FEATURES OF THE '249 PATENT	19
A. Technological Background at Time of Filing	20
B. Summary of the Prior Art.....	27
1. Kennewick (EX1004)	27
2. Oviatt (EX1005).....	30
3. Fujimoto (EX1007, EX1008).....	33
4. Shimomura (EX1014).....	35

5.	Matsuda (EX1015, EX1016)	36
6.	Cooper (EX1017).....	37
VIII.	SUMMARY OF INVALIDITY GROUNDS.....	38
IX.	INVALIDITY GROUNDS.....	38
A.	Ground 1: Kennewick in combination with Oviatt and Fujimoto render obvious claims 1, 6, 9, 12-16, 21, and 25-28	38
1.	Claim 1	38
	[1.pre] “A computer-implemented method of facilitating natural language system responses using short-term knowledge generated based on one or more prior multi-modal device interactions, the method being implemented by a computer system that includes one or more physical processors executing one or more computer program instructions which, when executed, perform the method, the method comprising:”	38
	[1.a] “receiving, by the computer system during a first conversation, a first voice input via a first input device, the first voice input comprising a first natural language utterance;”	52
	[1.b] “receiving, by the computer system, a second voice input comprising the first natural language utterance via a second input device;”	54
	[1.c] “comparing, by the computer system, the first voice input with the second voice input;”	57
	[1.d] “filtering, by the computer system, sound from the first voice input and the second voice input based on the comparison;”	61

[1.e] “obtaining, by the computer system during the first conversation, a user interface state related to one or more non-voice inputs associated with the first voice input, the one or more non-voice inputs comprising at least a first non-voice input;”	65
[1.f] “generating, by the computer system, the short-term knowledge based on at least the first voice input and the first non-voice input;”	71
[1.g] “determining, by the computer system, based on the short-term knowledge, a first context for the first natural language utterance;”	74
[1.h] “determining, by the computer system, based on the first context, an interpretation of the first natural language utterance; and”	75
[1.i] “generating, by the computer system, based on the interpretation of the first natural language utterance, a first response to the first natural language utterance.”	77
2. Claim 6: “The method of claim 1, wherein determining the interpretation of the first natural language utterance comprises determining, based on the first context, an interpretation of one or more recognized words of the first natural language utterance.”	79
3. Claim 9	80
[9.pre] “The method of claim 1, the method further comprising:”	80

	[9.a] “accumulating, by the computer system, long-term knowledge, wherein the long-term knowledge is accumulated based on one or more natural language utterances received during conversations occurring prior to the first conversation, wherein the first context is determined based on the short-term knowledge and the long-term knowledge.”	80
4.	Claim 12: “The method of claim 1, wherein the second input device is included with the first input device in a single device.”	83
5.	Claim 13: “The method of claim 1, wherein generating the first response comprises generating, by the computer system, a non-voice response in a graphical user interface and/or a voice response via a speaker.”	85
6.	Claim 14	88
	[14.pre] “The method of claim 1, wherein the first natural language utterance relates to a request and the first response is intended to be responsive to the request, the method further comprising:”	88
	[14.a] “receiving, by the computer system, a second natural language utterance in response to the first response, wherein the second natural language utterance comprises a clarification related to the interpretation of the first natural language utterance; and”	89
	[14.b] “responsive to receipt of the second natural language utterance, generating, by the computer system, a response to the second natural language utterance based on the interpretation of the first natural language utterance and the clarification related to the interpretation of the first natural language utterance.”	90

7.	Claim 15	91
	[15.pre] “The method of claim 1, wherein the first non-voice input is received via a non-voice input device comprising a touch screen device, the method further comprising:”	91
	[15.a] “receiving, by the computer system, information indicating a location on the touch screen device that received the first non-voice input.”	93
8.	Claim 16	94
	[16.pre] “A system for facilitating natural language system responses via short-term knowledge generated based on one or more prior multi-modal device interactions, the system comprising:”	94
	[16.a] “one or more physical processors programmed with one or more computer program instructions which, when executed, cause the one or more physical processors to:”	94
9.	Claims 21 and 25-28	95
B.	Ground 2: Kennewick-Oviatt-Fujimoto in combination with Shimomura renders obvious claims 2-5 and 17-20	96
1.	Claim 2	96
	[2.pre] “The method of claim 1, the method further comprising:”	96
	[2.a] “receiving, at the computer system, a second natural language utterance via the first input device;”	96
	[2.b] “determining, by the computer system, based on the short-term knowledge, whether the second natural language utterance corresponds to the first context;”	99

	[2.c] “determining, by the computer system, responsive to the second natural language utterance not corresponding to the first context, a second context for the second natural language utterance based on the short-term knowledge;”	104
	[2.d] “determining, by the computer system, based on the second context, an interpretation of the second natural language utterance; and”	106
	[2.e] “generating, by the computer system, based on the interpretation of the second natural language utterance, a response to the second natural language utterance.”	106
2.	Claim 3	107
	[3.pre] “The method of claim 1, the method further comprising:”	107
	[3.a] “tracking, by the computer system, contexts identified for multiple consecutive natural language user utterances to generate a context stack, the context stack including the contexts identified for the multiple consecutive natural language user utterances;”	107
	[3.b] “receiving, at the computer system, a second natural language utterance via the first input device;”	112
	[3.c] “determining, by the computer system, based on the short-term knowledge, whether the second natural language utterance corresponds to one or more individual ones of the contexts included in the context stack.”	112
3.	Claim 4	113

	[4.pre] “The method of claim 3, wherein determining whether the second natural language utterance corresponds to one or more individual ones of the contexts included in the context stack includes:”	113
	[4.a] “determining whether the second natural language utterance corresponds to a most recent context included in the context stack; and”	113
	[4.b] “responsive to the second natural language utterance not corresponding to the most recent context, determining whether the second natural language utterance corresponds to a second most recent context included in the context stack.”	114
4.	Claim 5	115
	[5.pre] “The method of claim 3, further comprising:”	115
	[5.a] “determining, by the computer system, responsive to the second natural language utterance not corresponding to one or more individual ones of the contexts included in the context stack, a second context for the second natural language utterance based on the short-term knowledge; and”	115
	[5.b] “determining, by the computer system, based on the second context, an interpretation of the second natural language utterance.”	116
5.	Claims 17-20	116
C.	Ground 3: Kennewick-Oviatt-Fujimoto in combination with Matsuda renders obvious claims 7, 8, and 22	117
1.	Claim 7	117
	[7.pre] “The method of claim 1, the method further comprising:”	117

	[7.a] “expiring, by the computer system, one or more items of short-term knowledge that are based on one or more natural language utterances received during conversations occurring prior to the first conversation.”	117
2.	Claim 8	122
	[8.pre] “The method of claim 7, further comprising:”	122
	[8.a] “accumulating, by the computer system, long-term knowledge based on the one or more items of short-term knowledge that are expired.”	122
3.	Claim 22	123
	[22.pre] “The system of claim 16, wherein the one or more physical processors are caused to:”	123
	[22.a] “expire one or more items of accumulated short-term knowledge that are based on one or more natural language utterances received during conversations occurring prior to the first conversation; and”	123
	[22.b] “accumulate long-term knowledge based on the one or more items of accumulated short-term knowledge that are expired, wherein the first context is determined based on short-term knowledge and the long-term knowledge.”	124
D.	Ground 4: Kennewick-Oviatt-Fujimoto in combination with Cooper renders obvious claims 10, 11, 23, and 24	125
1.	Claim 10	125
	[10.pre] “The method of claim 1, further comprising:”	125

	[10.a] “determining, by the computer system, based on the short-term knowledge, a manner in which the first natural language utterance is spoken, wherein the first response is generated based on the determined manner and the interpretation of the first natural language utterance.”	125
2.	Claim 11: “The method of claim 10, wherein the manner in which the first natural language utterance is spoken includes at least one of a tone, a pace, an inflection, or a timing.”	130
3.	Claims 23 and 24.....	131
X.	CONCLUSION.....	131

I, Padhraic Smyth, Ph.D., declare as follows:

I. INTRODUCTION

1. I have been asked by the party requesting this review, Amazon.com Services LLC (“Petitioner”), to provide my expert opinions in support of the petition for Inter Partes review of U.S. Patent No. 10,297,249 (the “’249 patent” or “challenged patent”) challenging the patentability of claims 1-28 (the “challenged claims”).

2. In summary, it is my opinion that the challenged claims are disclosed and rendered obvious by the prior art. I set forth below my detailed opinions.

II. BACKGROUND AND QUALIFICATIONS

3. The details of my education, work experience, research, patents, and publications are summarized in my CV, attached as EX1018. I provide here a summary of some of the relevant information from my CV, as well as other relevant background information.

4. I have a Ph.D. and Master’s degree in Electrical Engineering from the California Institute of Technology. I earned a Bachelor of Engineering (Electronic) from the National University of Ireland, University College Galway in 1984. My senior undergraduate thesis was on speech recognition.

5. At present, I am a Distinguished Professor and Endowed Chair in Artificial Intelligence in the Department of Computer Science at the University of

California, Irvine. Prior to that, I was a Member of Technical Staff and Technical Group Leader at the Jet Propulsion Laboratory.

6. As a professor at UC Irvine since 1996, I conduct research in the areas of machine learning, artificial intelligence, signal processing, pattern recognition, and natural language processing, and I regularly teach courses related to these topics. I have published over 230 peer-reviewed papers and two books on these topics and my work has been cited over 70,000 times in the research literature. My research work is funded by federal agencies such as NSF, NIH, NIST, IARPA, DARPA, ONR, and NASA.

7. During my career I have worked extensively on research related to key aspects of automated speech recognition, natural language processing, user modeling, and human dialog. As an example, since 1992 I have conducted a significant body of research and published a series of papers on the topic of hidden Markov models (HMMs). HMMs are a general statistical framework for pattern recognition in sequential data (such as audio waveforms or text) and were a key component in automated speech recognition systems and natural language understanding since the 1980's. My research on HMMs produced significant advances in our understanding of the statistical and algorithmic foundations of speech recognition and natural language understanding. These advances included development of new HMM methods for detecting anomalous patterns in time-series

signals, new algorithms for automated clustering of waveforms using HMMs, new classes of HMMs with applications to speech modeling, and novel HMM methods for automated segmentation of waveforms. My papers on these topics have been collectively cited over 1,800 times in the literature.

8. I have also conducted extensive research in the field of natural language understanding since 1999. My work included co-authoring a textbook, *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*, published in 2003 with chapters devoted to algorithms for natural language understanding. I have also written papers on topics like analysis of character sequences in non-English languages, automated algorithms for extracting human-interpretable summaries of content from large text corpora, methods for automated characterization of authors from the text they generate, automated techniques for creating network and graph models of relationships between documents and entities mentioned in the documents, algorithms for analyzing text with the assistance of human-generated semantic ontologies, and analysis of large language models. As with my work on HMMs, my work in natural language understanding meant that I was, and am, broadly familiar with the research literature in this area.

9. I have also conducted research in the area of user modeling, focusing on statistical and machine learning techniques for automated generation of user models from observational data. In addition I have published papers on techniques

for automated analysis of human dialog, such as automatically classifying utterances into specific topics. As I discuss below, user modeling is an important function of many interactive conversational systems, which enables such systems to manage a dialogue with the user and infer information about the user's requests.

10. In addition to my research and publications I have taught and lectured extensively on these topics. Since my appointment as a professor at UC Irvine in 1996, I have taught both undergraduate and graduate courses there in machine learning, pattern recognition, natural language understanding, artificial intelligence, and user modeling. In addition, I have taught professional courses on these topics at UC Irvine and have given keynote lectures and tutorials on these topics at a variety of international conferences since the late 1990s.

11. In my professional capacity I have also served in editorial roles and conference chair positions for major journals and conferences in my research areas, including being a founding editorial board member for the Journal of Machine Learning Research, an associate editor for the Journal of the American Statistical Association, program chair for the ACM SIGKDD conference on Knowledge Discovery and Data Mining, co-program chair for the International Joint Conference on Artificial Intelligence, and have served as Senior Area Chair for the NeurIPS conference and the International Machine Learning Conference on multiple occasions.

12. I am also a technical consultant with private industry in areas including machine learning, user modeling, signal processing, pattern recognition, and natural language understanding. Over the past 25 years I have consulted in these areas for companies including Toshiba, Microsoft, Google, SAP, Netflix, eBay, AT&T, Samsung, Nokia, as well as several startup companies. Accordingly, I have knowledge of and experience with the practical application, in terms of operational use and commercial products, of methods from the fields of machine learning, user modeling, signal processing, pattern recognition, and natural language understanding.

III. COMPENSATION AND RELATIONSHIP WITH PARTIES

13. I am being compensated for my time at the rate of \$800 per hour. This compensation is not contingent upon my performance, the conclusions I reach in my analysis, the outcome of this matter, or any issues involved in or related to this matter.

14. I have no financial interest in Petitioner or any related parties.

IV. MATERIALS CONSIDERED

15. I have reviewed and considered, in preparing this declaration, the '249 patent (EX1001), its prosecution history (EX1003), and the following prior art references:

- U.S. Patent App. Pub. No. 2004/0193420, published September 30, 2004 and titled “Mobile Systems and Methods for Responding to Natural Language Speech Utterance” (“Kennewick”) (EX1004);
- S. Oviatt et al., “Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions,” Human-Computer Interaction, 15:4, 263-322 (2000) (“Oviatt”) (EX1005);
- Japanese Laid Open Patent Publication No. JP2004-145025, published May 20, 2004 and titled “Voice Matching Device” (“Fujimoto”) (EX1007, EX1008);
- U.S. Patent App. Pub. No. 2001/0021909, published September 13, 2001 and titled “Conversation Processing Apparatus and Method, and Recording Medium Therefor” (“Shimomura”) (EX1014);
- Japanese Laid Open Patent Publication No. JP2004-334591, published November 25, 2004 and titled “Conversation System and Conversation Processing Program” (“Matsuda”) (EX1015, EX1016);
- U.S. Patent No. 6,757,362, issued June 29, 2004 and titled “Personal Virtual Assistant” (“Cooper”) (EX1017).

16. I am also aware of and considered in forming my opinions presented in this declaration, other information available to a person of ordinary skill in the art at

the time of the effective filing date of the challenged patent (“POSITA”), such as other patents, reference materials, articles, and standards, and knowledge of a POSITA.¹ These include:

- Asthana, A. & Krzyzanowski, P., “A Small Domain Communications System for Personalized Shopping Assistance,” Proceedings of 1994 International Conference on Personal Wireless Communications, IEEE Press, 199-203 (1994) (“Asthana”) (EX1009);
- Lucente, M., “Conversational interfaces for e-commerce applications,” Communications of the ACM 43(9), 59-61 (Sept. 2000) (“Lucente”) (EX1010);
- Huang et al., “Spoken Language Processing: A Guide to Theory, Algorithm, and System Development,” Prentice Hall (2001) (“Huang”) (EX1011);
- U.S. Patent No. 6,859,776, issued February 22, 2005 and titled “Method and Apparatus for Optimizing a Spoken Dialog Between a Person and a Machine” (“Cohen”) (EX1012).
- Microsoft Computer Dictionary, Fifth Edition (2002) (EX1020).

¹ I understand that the challenged patent has an effective filing date of October 16, 2006.

17. Should new information be disclosed or discovered during this proceeding, I may supplement and elaborate on the opinions I express in this declaration as necessary. Nothing in this declaration should be interpreted as an opinion that no grounds of invalidity exist other than the ones in the petition. I do not express such an opinion and have not reached such an opinion.

V. BASIS OF OPINIONS

A. Legal Standard for Anticipation

18. I understand that a claimed invention is unpatentable as “anticipated” by a single prior art reference if it discloses each and every limitation of the claim. I also understand that, even if a claim is not explicitly disclosed in the prior art reference, the prior art reference may nevertheless anticipate the claim if the subject matter of the claim is inherently present from the perspective of a POSITA. Moreover, I understand that even if the prior art reference does not disclose the limitations of a claim in the same arrangement or combination of the limitations of the claim, it may nevertheless anticipate if a POSITA would envisage the claimed arrangement or combination.

B. Legal Standard for Obviousness

19. I understand that a claimed invention is unpatentable as “obvious” if the differences between the invention and the prior art are such that the subject matter of the claim as a whole would have been obvious to a POSITA.

20. It is my understanding that “obviousness” is a question of law based on underlying factual issues including (1) the scope and content of the prior art, (2) the differences between the prior art and the asserted claims, (3) the level of ordinary skill in the pertinent art, and (4) the existence of secondary considerations such as commercial success, long-felt but unresolved needs, and failure of others.

21. I understand that a single reference can render a patent claim obvious if any differences between that reference and the claims would have been obvious to a POSITA. Alternatively, the teachings of two or more references may be combined in the same way as disclosed in the claims, if such a combination would have been obvious to a POSITA. In determining whether a combination based on either a single reference or multiple references would have been obvious, I understand that it is appropriate to consider the following factors:

- Whether the teachings of the prior art references disclose known concepts combined in familiar ways, and when combined, would yield predictable results;
- Whether a POSITA could implement a predictable variation, and would see the benefit of doing so;
- Whether a POSITA would have recognized a reason to combine known elements in the manner described in the claim;

- Whether there is some teaching or suggestion in the prior art to make the modification or combination of elements claimed in the patent; and
- Whether the innovation applies a known technique that had been used to improve a similar device or method in a similar way.

22. In an obviousness evaluation based on a combination of multiple prior art references, I understand that the prior art references themselves may provide a suggestion, motivation, or reason to combine, but other times the nexus linking two or more prior art references is simple common sense. I further understand that obviousness analysis recognizes that market demand, rather than scientific literature, often drives innovation, and that a motivation to combine references may be supplied by the direction of the marketplace.

C. Level of Ordinary Skill in the Art

23. I understand that the '249 patent should be interpreted from the perspective of a person having "ordinary skill in the art" as of the effective filing date of the patent application (a POSITA). For the purposes of this declaration, I assume that the effective filing date of the patent application leading to the '249 patent is no earlier than October 16, 2006, and I understand that the Patent Owner has not claimed an earlier priority date in the parallel litigation. I am familiar with the technological field at issue and the state of the art as of that date.

24. I understand that the education level of those in the technological field, sophistication of the technology, types of problems encountered, prior art solutions to those problems, and the speed at which innovations are made in the field may factor into establishing what the appropriate level of ordinary skill should be.

25. In my opinion, a POSITA at the priority date of the '249 patent would have had a Bachelor's degree in computer science, computer engineering, electrical engineering, or a related field in computing technology, and two years of experience with automatic speech recognition and natural language understanding, or equivalent education, research experience, or knowledge. Additional experience can substitute for the level of education, and vice-versa.

26. I had at least the requisite level of skill with respect to the subject matter of the challenged patent at the time of the priority date as demonstrated by my qualifications listed above and my CV (EX1018).

D. Claim Construction

27. I understand that, in *inter partes* review ("IPR") proceedings, claims are construed using the same claim construction standard that district courts use, which I understand is called the "*Phillips*" claim construction standard, including construing the claim in accordance with the plain and ordinary meaning of such claim as understood by one of ordinary skill in the art and the prosecution history pertaining to the patent.

28. I applied plain and ordinary meaning to all terms of the challenged claims for my analysis in this proceeding.

29. I understand the district court in *VB Assets, LLC, v. Amazon.com, Inc., et al.*, C.A. No. 19-1410 (MN) (D. Del.) issued a claim construction order involving U.S. Patent Nos. 8,073,681 (“the ’681 Patent”) and 9,015,049 (“the ’049 Patent”), parents of the challenged patent that share the same specification. EX1019. The order indicates that the parties agreed the term “short-term knowledge” did not require construction and the district court found that the terms “long-term knowledge” and “context” did not require construction. EX1019, 1-2.

30. I am not an attorney and my opinions are limited to what I believe a POSITA would have understood the meaning of certain claim terms to be, based on the patent specifications and prosecution histories in light of the POSITA’s knowledge.

VI. OVERVIEW OF THE ’249 PATENT

A. Filing and Specification

31. The ’249 patent was filed as Application No. 14/691,445 (the “’445 application”) on April 20, 2015. The ’445 application is a continuation of Application No. 13/987,645 (the “’645 application”) filed August 19, 2013, now Patent No. 9,015,049. The ’645 application is a division of Application No. 13/251,712 (the “’712 application”) filed October 3, 2011, now Patent No.

8,515,765. The '712 application is a continuation of Application No. 11/580,926 filed October 16, 2006 (the “effective filing date”), now Patent No. 8,073,681.

32. The '249 patent describes an interface that “may understand free form human utterances.” EX1001, 2:13-17. An utterance may be part of a multimodal input. *Id.*, 7:38-40. For example, an “input device may include a combination of a microphone and a touch-screen device,” where an input to the device includes “an utterance that includes a request relating to a portion of a display on the touch-screen device that the user is touching.” *Id.*, 7:38-44.

33. The utterance may be received by multiple microphones. “For example, a single device may include multiple microphones, or multiple devices may each include one or more microphones.” *Id.*, 11:7-9. If there are multiple microphones, the interface “may collate inputs and cooperatively filter out sound by comparing a speech signal from the various microphones.” *Id.*, 11:9-11. It “may also filter out non-human environmental noise within the range of the microphones.” *Id.*, 11:12-13.

34. The interface includes “a context determination process that determines one or more contexts for a request to establish meaning within a conversation.” *Id.*, 3:61-63, 11:23-26. The “context determination process 255 may infer intended operations and/or context based on previous utterances and/or requests.” *Id.*, 11:56-58. “For example, if a given interpretation turns out to be incorrect, the incorrect

interpretation may be removed as a potential interpretation from one or more grammars...and/or from possible subsequent interpretations determined by context determination process 255.” *Id.*, 11:60-65.

35. The context may be used within a “cooperative conversation model...to implement a conversational Human-to-Machine interface that reflects how humans interact with each other and their normal behavior in conversation.” *Id.*, 13:49-56. An Intelligent Hypothesis Builder “generate[s] one or more explicit hypotheses of a user’s intent when an utterance contains all information...needed to complete a request or task.” *Id.*, 16:34-38. It uses “short-term shared knowledge...to proactively build and evaluate interaction with a user as a conversation progresses, as well as long-term shared knowledge to proactively build and evaluate interaction with the user over time.” *Id.*, 16:16-21. “Short-term knowledge may accumulate during a single conversation, where input received during a single conversation may be retained.” *Id.*, 5:6-8. The long-term knowledge “may generally be user-centric, rather than session-based, where inputs may be accumulated over time to build user, environmental, cognitive, historical, or other long-term knowledge models.” *Id.*, 5:26-30.

36. An Adaptive Response Builder uses the output of the Intelligent Hypothesis Builder “to formulate [a] personalized and conversational response” to the user. *Id.*, 16:63-66, 17:41-46. The response “may include a verbal or audible

reply played through an output device...and/or an action performed by a device, computer, or machine.” *Id.*, 18:9-14.

B. Prosecution History

37. I reviewed the prosecution history of the '249 patent. EX1003. The Applicant filed an information disclosure statement (“IDS”) citing several hundred references with the initial application. EX1003, 1433-1463. This IDS listed the patent (U.S. Patent No. 7,693,720) issued from the Kennewick reference (EX1004) asserted in this petition. It also listed many other references naming Kennewick as an inventor. EX1003, 1451. The Applicant subsequently filed additional IDS’s listing dozens more references. EX1003, 1374-1376, 1351-1353, 1090-1091, 1042-1044, 1034-1036, 799-802, 650-653, 367-368, 361-363, 349, 344-345, 186-187, 146.

38. The Examiner issued an Office Action making a substantive rejection based on a different Kennewick reference, U.S. Patent App. Pub. No. 2004/0044516 (“Kennewick516”). *Id.*, 661-702. Applicant distinguished Kennewick516 from the pending claims, arguing that Kennewick516 and other cited references do not teach or suggest “using both voice and non-voice inputs to determine a context for a natural language utterance of the voice input.” *Id.*, 642, 334. The Examiner agreed, noting that Kennewick516 “does not...suggest[] wherein the knowledge is generated based on one or more prior multi-modal device interactions.” *Id.*, 90-91.

39. The Examiner also made a §112 rejection alleging that the original specification “does not describe where the short-term knowledge is based on both voice and non-voice inputs.” *Id.*, 557. Applicant argued in response that the support is found in at least paragraphs 25, 42, and 43 of the as-filed specification. The relevant portion of paragraph 25 is found in the ’249 patent (EX1001) at 7:38-44 and describes an input device that includes “a combination of a microphone and a touch-screen device” where the input “may include an utterance that includes a request relating to a portion of a display on the touch-screen device that the user is touching.” EX1003, 332-333, 1473-1474. The relevant portion of paragraph 42 is found at 14:5-6 of the ’249 patent and states that “[s]hort-term knowledge may accumulate during a single conversation.” EX1003, 332-333, 1483. The relevant portion of paragraph 43 is found at 14:10-23 of the ’249 patent and states in part that “the Session Input Accumulator may accumulate a current user interface state relating to other modal inputs to further build shared knowledge models and more accurate adaptive responses (e.g., when a user utters a request relating to a portion of a touchscreen device.” EX1003, 332-333, 1483-1484. The Examiner withdrew the rejection in the next Office Action. EX1003, 273-274.

40. Applicant introduced the “comparing” and “filtering” limitations that ultimately issued as limitations [1.c] and [1.d] of claim 1 of the ’249 patent in claim 21 of an Office Action Response submitted July 11, 2018. EX1003, 330-331. In an

Office Action dated December 4, 2018, the Examiner found that the combination of Byers (US 6,219,645), Kennewick516, and Perkins (US 7,072,888) “suggests where multiple microphones receive the same utterance and comparing the signals from the different microphones are compared to each other, and selecting the best/cleanest signal out of the different microphone signals based on the comparison.” EX1003, 121.

41. The Examiner found that “Byers does not, however, describe where ‘sound from the first voice input and the second voice input’ (i.e. sound that is based on both voice inputs from the first and second input devices) are filtered, where the filtering is based on the comparison.” EX1003, 120-121 (emphases in original²). The Examiner also stated that “[r]egardless of whether the prior art does or does not teach or suggest filtering sound based on multiple microphone signals based on comparison of those multiple microphone signals to each other, at least one more reference would be required to address all of the limitations of claims 1 and 21 together, such that the prior art does not reasonably suggest the combination of all limitations in claims 21 and 1 together due to the quantity of references required to address every limitation (at least five) and due to the complexity of the manner in which those references would need to be combined.” EX1003, 121.

² All annotations and emphases are added unless otherwise noted.

42. Applicant amended the claims to incorporate claim 21. EX1003, 52-64. The Examiner then allowed the application. EX1003, 32-42. The Examiner made a Statement of Reasons for Allowance mirroring those in the December 4, 2018 Office Action. EX1003, 40-41. The Applicant filed Comments responding to the Examiner's statement submitting "that no one element or limitation in particular should be deemed to impart to or be required for patentability of the claims." EX1003, 25.

C. Related Litigation and Related Patents

43. I understand the '249 patent is the subject of a pending litigation involving Petitioner and Patent Owner: *VB Assets, LLC v. Amazon.com Services, LLC*, No. 1:24-cv-00839-MN (District of Delaware, July 18, 2024) (the "parallel litigation"). I understand Patent Owner sued Petitioner for alleged infringement of the '249 patent and four other patents. *See* EX1013.

44. I understand the '249 patent has been involved or is currently involved in other litigations, including: *VB Assets, LLC v. SoundHound AI Inc.*, No. 1:24-cv-01279-MN (D. Del.) and *VB Assets, LLC v. Apple Inc.*, No. 1:24-cv-01368-MN (D. Del.).

45. I understand U.S. Patent No. 9,015,049, a parent of the challenged patent that shares the same specification as the '249 patent, was involved in a prior

litigation: *VB Assets, LLC v. Amazon.com, Inc., et al.*, No. 1:19-cv-1410-MN (D. Del.). This patent was also involved in IPR2020-01346.

46. I understand that U.S. Patent No. 8,073,681, a parent of the challenged patent that shares the same specification, was also involved in *VB Assets, LLC v. Amazon.com, Inc., et al.*, No. 1:19-cv-01410-MN (D. Del.). U.S. Patent No. 8,073,681 is also involved in other litigations including: *VB Assets, LLC v. Samsung Electronics Am. Inc.*, No. 2:24-cv-00828-JRG (E.D. Tex.), *VB Assets, LLC v. SoundHound AI Inc.*, No. 1:24-cv-01279-MN (D. Del.), *VB Assets, LLC v. Apple Inc.*, No. 1:24-cv-01368 (D. Del.). In addition, U.S. Patent No. 8,073,681 is involved in IPR2020-01367 and IPR2025-00866.

47. I understand that U.S. Patent No. 8,515,765, a parent of the challenged patent that shares the same specification, is involved in other litigations including: *VB Assets, LLC v. Apple Inc.*, No. 1:24-cv-01368-MN (D. Del.) and *VB Assets, LLC v. Samsung Electronics Am. Inc.*, No. 2:24-cv-00828-JRG (E.D. Tex.). This patent is also involved in IPR2025-00868.

VII. THE PRIOR ART RENDERS OBVIOUS THE CLAIMED FEATURES OF THE '249 PATENT

48. I provide an overview of the state of the art at the time of the effective filing date and discuss the specific references underlying the grounds of invalidity set forth in this Petition.

A. Technological Background at Time of Filing

49. The '249 patent relates to (i) a human user interacting with a system via multi-modal input including speech, and (ii) the system engaging with the user in a conversational dialog. EX1001, 2:21-47. These ideas were well-known in the prior art at the time of the '249 patent.

50. In interactive conversational systems involving a human user and a system, a dialog (or conversation) can be represented as a series of utterances, where each utterance is alternately generated by the human and by the system. A simple example of a dialog might be the following:

User: "I would like to fly to New York."

System: "When would you like to travel?"

User: "This Friday please."

System: "What is your departure airport?"

And so on. In this example there are two utterances by the user and two by the system. Technologies and systems to broadly support such interactive conversations had been under development and were well-known prior to 2006, both in academic research laboratories and in industry. *See, e.g.*, EX1009; EX1010; EX1011, 879-982 (Chapters 17-18).

51. These types of interactive conversational systems, as developed and implemented by 2006, relied on a number of related technologies that I will briefly

describe below. The primary technologies are automated speech recognition (“ASR”), natural language understanding (“NLU”), dialog management (“DM”), and user modeling (“UM”). Broadly speaking, ASR involves recognizing words from spoken audio for each user utterance, NLU is the process of determining the meaning of the recognized words in the utterance, DM involves generation of appropriate response utterances by the system and the management of the conversation with the user across multiple utterances, and UM is the process by which the system keeps track of information about a user, e.g., what their typical goals and preferences are.

52. These terms are not the only ones used to describe these technologies. Other terms were (and are still) used by researchers and practitioners in this general context. One such term is “Spoken Language Understanding” (“SLU”) which typically encompasses aspects of ASR, NLU, and DM. Nonetheless, the terms ASR, NLU, DM, and UM were in broad use in 2006 and would have been well-known to a POSITA at that time. Thus, I focus on these terms in my discussion since these terms are generally useful for describing the technical background for the challenged patent.

53. In 2006, ASR involved a sequence of steps consisting of (i) recording and digitizing speech waveforms spoken by a user, (ii) converting the digitized speech waveforms into numerical features that (typically) represent the frequency

content of very short segments of the speech waveform, (iii) automatically classifying these features into corresponding phonemes (basic units of sounds within words), and (iv) recognizing specific words based on sequences of phonemes. This process has been the subject of extensive research and commercial activity since the 1960s. For example, the well-known text “SLP” devotes Chapters 5 through 13 (485 pages) just to ASR. EX1011, 227-711.

54. The output of the ASR component of a conversational interface system consists of the ASR’s best estimates of what words were spoken by a user in an utterance. The recognized words alone are typically not sufficient for an algorithm to understand the user’s goals and intent. Human language is very flexible and there are many ways to express the same meaning or intent. The technology of NLU consists of a broad range of algorithms and methods that allow a computer system to interpret the meaning of the recognized words from the ASR. *Id.*, 893-912. For example, consider the simple phrase “I would like to fly to New York” from the dialog example described earlier. And assume that the system is a travel reservation system that can find flights, make flight reservations, pay for the flights, suggest or advertise additional travel services, and so on. Such a system would have an internal representation (e.g., a database) of items such as names of airports, locations of airports, lists of flights, actions the system can take (e.g., book flights), and so on. Thus, it would need to interpret the user’s words “like to fly” as an implicit request

for the system to find a flight. It would also need to interpret the word “I” to mean that the flight is for the individual who is speaking. And it would need to understand that “New York” is the name of a location, and that it is not the name of a specific airport but is a city with multiple airports associated with it. There are different approaches associated with developing NLU technologies that can robustly handle such situations and by 2006, multiple such systems had been developed that demonstrated the necessary NLU capabilities for interpreting the meaning of spoken words in conversational dialog. *See, e.g., id.*, 945-982 (Chapter 18); EX1010; EX1012, 18:39–19:37.

55. As I describe above, the ASR and NLU components of a conversational interface system focus on recognizing words and interpreting these words on a per-utterance basis. DM (Dialog Management) technology addresses the problem of tracking different aspects of a dialog across multiple utterances. For example, for a travel reservation system interacting with a user via speech, the DM component of the system gradually accumulates the relevant pieces of information from the user (dates, times, arrival/departure airports, seating preferences, confirmation of booking, purchase information) over the course of a dialog, i.e., over a sequence of utterances generated the user. Thus, DM technologies involve creating internal representations of concepts such as goals for a user and methods to satisfy those goals. One such method includes directly asking the user for specific information,

which in turn involves automatically generating the appropriate words for the system's response (utterance) to the user, given what the system knows already (e.g., "What time would you like to leave for your flight to New York?") and then converting that into a spoken response that the user can hear. Another type of DM method involves inferring the relevant information from knowledge sources available to the system. For example, the DM might assume from information in the user's profile that the departure airport is likely to be the one closest to where the user lives. And the DM could infer the likely destination airport if non-stop flights from the user's home airport only go to a specific airport in the New York area. DM technology involves various software components that can make decisions about what type of information to request from the user, how to present this information to the user, in what sequence to request this information, when this information might be inferred from other sources, and so on. DM technology that exhibited these characteristics was well-known by 2006 and had been developed in research laboratories and in commercial systems before that time. *See, e.g.,* EX1011, 912-927 (Chapter 17.6-17.7), 945-982 (Chapter 18); EX1010; EX1004.

56. It was also well-known by 2006 that there were significant benefits to be gained if a conversational system had access to specific knowledge about a user (such as their travel preferences). Such access can make the interaction with the user more efficient from the system's perspective and more pleasant from the user's point

of view, by reducing the amount of information the system needs to request from the user during the dialog. The technology of user modeling (UM) consists of different methods the system uses to automatically build an internal representation of a user. These methods can include simple (but useful) approaches such as recording information about a user's full name, age, home address, etc., explicitly provided by a user during (say) a registration process. More complex UM methods extract information that is implicitly provided by the user, for example, inferences about a user's travel preferences from words in past conversations that a user has had with the system, or inferences from words in earlier utterances in the current conversation. By 2006, the benefits of UM in conversational modeling were widely acknowledged and had been developed in the context of both research and commercial systems. *See, e.g., EX1004.*

57. By 2006, it was also well-understood that the types of conversational systems that I described above, involving combinations of ASR, NLU, DM, and UM, could be directly applied to a variety of applications involving online commercial activities, such as advertising and purchasing of goods and services, in areas such as travel reservations, financial management, e-commerce purchases, and so on. *See, e.g., EX1011, 927-932 (Chapter 17.8); EX1010; EX1012; EX1004.*

58. By 2006, systems and methods for a conversational voice user interfaces were well-known. Indeed, as Kennewick recognizes:

Much work covering multiple methods has been done in the fields of natural language processing and speech recognition. Speech recognition has steadily improved in accuracy and today is successfully used in a wide range of applications. Natural language processing has previously been applied to the parsing of speech queries.

EX1004, ¶6. Speech recognition software was also commercially available. *See, e.g.,* EX1011, 952-955 (describing commercial speech recognition software).

59. Multimodal speech systems were also well-known by 2006. Oviatt, published in 2000, recognizes there is a “growing interest in multimodal interface design...inspired largely by the goal of supporting more transparent, flexible, efficient, and powerfully expressive means of human–computer interaction.” EX1005, 5. Oviatt observes that “major advances in new input technologies and algorithms, hardware speed, distributed computing, and spoken-language technology (SLT) in particular all have supported the emergence of more transparent and natural communication with” multimodal systems. *Id.*, 6. A multimodal interface beneficially “offers the user freedom to use a combination of modalities or switch to a better-suited modality, depending on the specifics of the task or environment.” *Id.*, 7. For example, “[d]uring multimodal interaction, people...prefer to use pen input to point...because it generally is a more effective and precise way to convey locations.” *Id.*, 9.

60. By 2006, it was also known to filter multiple voice inputs to an interactive conversational system. Huang dedicates an entire chapter “to ways of increasing the environmental robustness...in a speech recognition system.” EX1011, 503; *see id.*, 503-570 (Chapter 10). One technique Huang describes uses microphone arrays to “improv[e] the SNR [signal-to-noise ratio] of the received signal.” *Id.*, 531. Huang notes that “[s]everal companies are commercializing microphone arrays for...speech recognition applications.” *Id.*, 535. Consistent with Huang, Kennewick describes a “system [that] can use either one-dimensional or two-dimensional array microphones (or other devices) to receive human speech,” where “[s]peech received at the microphones may then be processed with analog or digital filters to optimize the bandwidth, cancel echoes, and notch out narrow band noise sources.” EX1004, ¶38.

61. Thus, all the foundational technologies referenced in the claims were already known to a POSITA at the time of the effective filing date of the '249 patent.

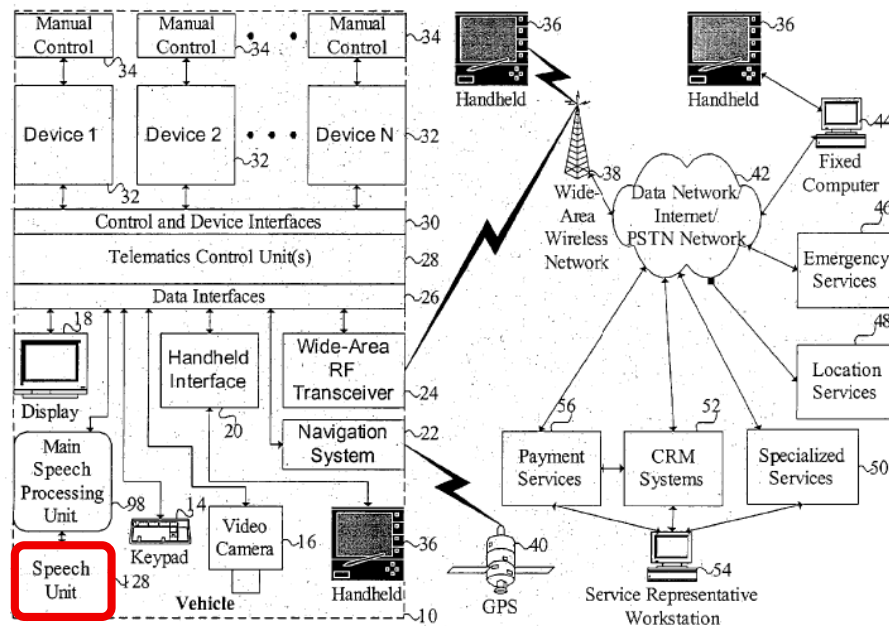
B. Summary of the Prior Art

1. Kennewick (EX1004)

62. Kennewick, U.S. Patent App. Pub. No. 2004/0193420, is entitled “Mobile Systems and Methods for Responding to Natural Language Speech Utterance.” Kennewick was filed July 15, 2003 and published September 30, 2004. EX1004. I understand Kennewick is prior art at least because its publication date

predates the October 16, 2006 effective filing date of the challenged patent. I form no opinion as to whether Kennewick is entitled to an earlier priority date.

63. Kennewick describes a system that allows “users to ask natural language speech questions” and “present[s] results in a natural manner.” *Id.*, ¶2. The described system is “a complete speech-based information query, retrieval, presentation and command environment” that “makes significant use of context, prior information, domain knowledge, and user specific profile data to achieve a natural environment for one or more users making queries or commands.” *Id.*, ¶11. In particular, Kennewick describes “a mobile interactive natural language speech system...that includes a **speech unit**.” *Id.*, ¶12. “[A] computer device or system...receives input from the **speech unit** and processes the input (e.g., retrieves information responsive to the query, takes action consistent with the command and performs other functions as detailed herein), and responds to the user with a natural language speech response.” *Id.*, ¶14.



Id., Figure 2, ¶108.

64. “The **speech unit** 128 includes one or more microphones, for example array microphone 134, to receive the utterances of the user.” *Id.*, ¶121. “The speech received at the microphone 134 may be processed by filters 132.” *Id.* For example, the system may use a filter between the array microphone and a speech coder “to optimize the signal to noise ratio at the input to the speech recognition engine.” *Id.*, ¶148.

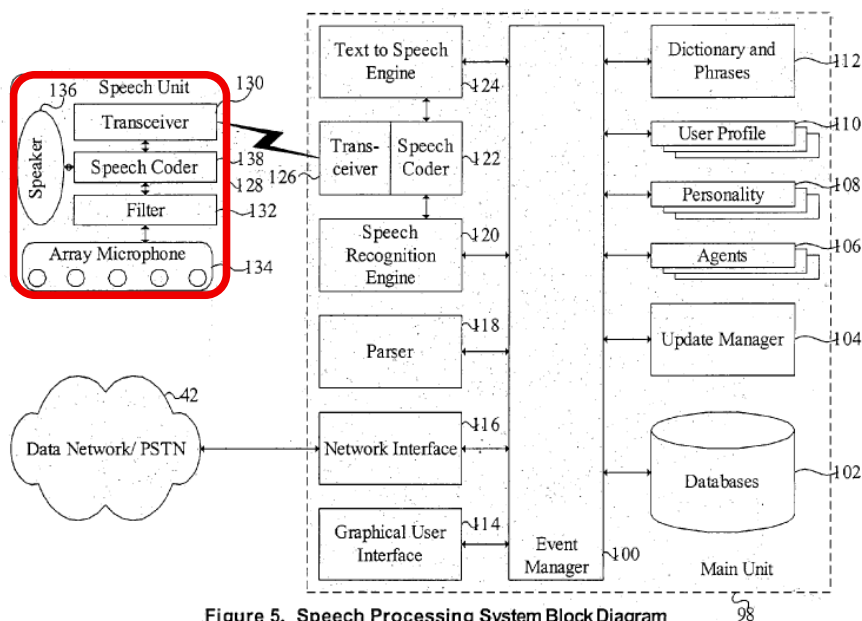


Figure 5. Speech Processing System Block Diagram

Id., Figure 5. “Algorithms comparing the background noise to the signal received from the users speech may be used to optimize the band-shaping parameters of the...filter.” *Id.*, ¶148.

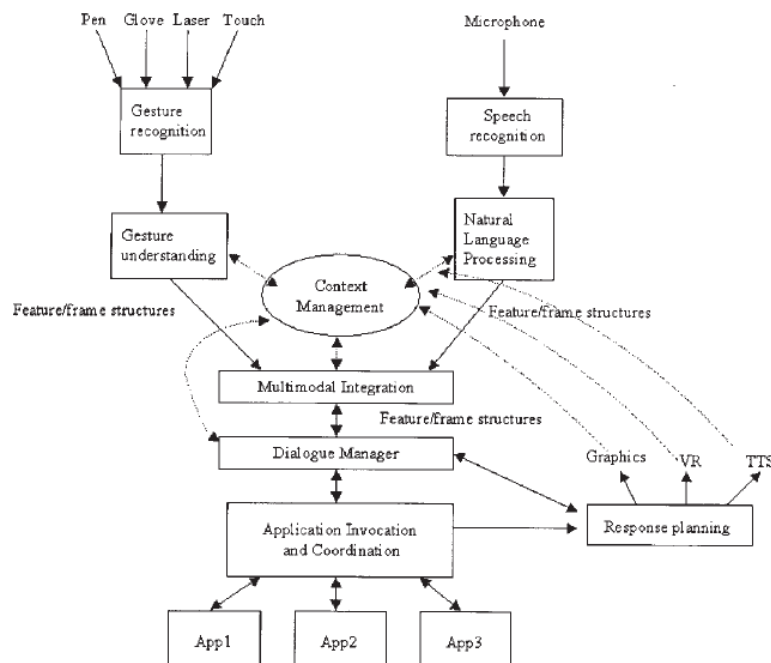
65. In connection with processing utterances, Kennewick describes “determin[ing] a context for an utterance by applying prior probabilities or fuzzy possibilities to...dialog history,” among other things. *See id.*, ¶¶160, 164. As Kennewick explains, “some questions can only be interpreted in the context of previous questions.” *Id.*, ¶7.

2. Oviatt (EX1005)

66. Oviatt is a paper titled “Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions” by Sharon Oviatt et al. EX1005. Oviatt was published in the

journal *Human-Computer Interaction* in 2000, in volume 15, pp. 263-322. *Id.*, 3. I understand Oviatt is prior art at least because its publication date predates the October 16, 2006 effective filing date of the challenged patent. I form no opinion as to whether Oviatt is entitled to an earlier priority date.

67. Oviatt “summarize[s] the emerging architectural approaches for interpreting speech and pen-based gestural input.” *Id.*, 5. It also “describe[s] a diverse collection of state-of-the-art multimodal systems that process users’ spoken and gestural input.” *Id.* Oviatt discloses a basic information flow where “speech and pen-based gestures are recognized in parallel, and each is processed by an understanding component.” *Id.*, 14-15.



Id., 15. The processing results in “meaning representations that are fused by the multimodal integration component, which also is influenced by the system’s dialogue management and interpretation of current context.” *Id.*

68. Oviatt notes that “[t]here are numerous ways to realize this information-processing flow as an architecture.” *Id.* One architecture Oviatt describes is a Portable Voice Assistant (PVA), which “is a pen-voice multimodal interface.” *Id.*, 38. “The PVA can interpret simultaneous pen-voice input so, for example, the user can say ‘Show me that part’ while pointing at a screw on the display.” *Id.*, 39. One potential application of the PVA is within “in-car navigation systems.” *Id.*, 40.

69. Oviatt is a printed publication and was publicly-accessible before the October 16, 2006 effective filing date of the challenged patent. Oviatt bears a copyright notice and other indicia showing it was published by *Human-Computer Interaction* in 2000. *Id.*, 3.

70. The declaration of Rachel J. Watters (the “Watters declaration”) (EX1006) demonstrates that a printed issue of *Human-Computer Interaction* containing Oviatt was received by the University of Wisconsin-Madison Library on or before April 25, 2001. EX1006, 2. The Watters declaration includes a copy of Oviatt as it appeared in the printed issue. *Id.*, 6-66. The printed version bears a date stamp of “April 25, 2001,” a 2000 copyright notice, and an ISSN. *Id.*, 5-6.

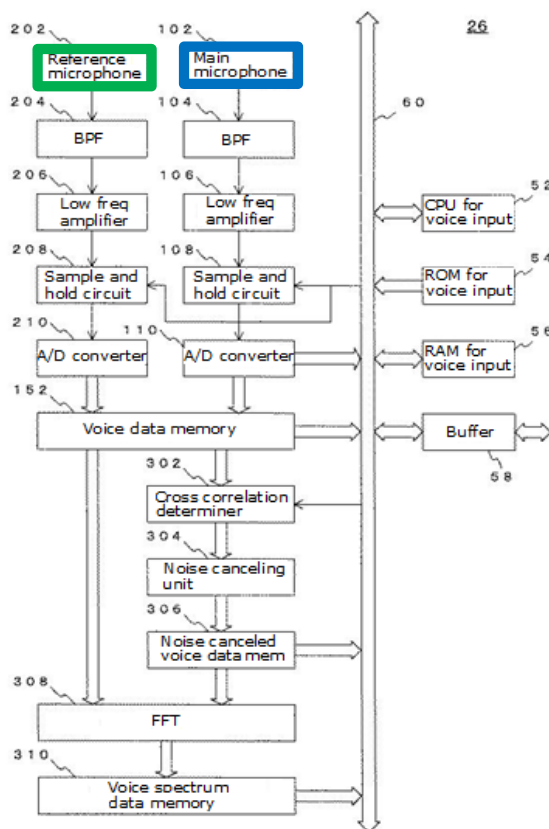
71. The Watters declaration demonstrates that Oviatt was publicly-accessible because this issue was “catalogued and available to library patrons within a few days or at most 2 to 3 weeks after April 25, 2001.” *Id.*, 2. A POSITA could then access Oviatt “by searching the public library catalog or requesting a search through WTS [Wisconsin TechSearch, an interlibrary loan department at the university].” *Id.* For example, a POSITA could have found Oviatt by searching “by title and/or subject key words” such as “multimodal” or “speech” and then accessing the publication on the library’s shelves. *Id.*, 2-3.

3. Fujimoto (EX1007, EX1008)

72. Fujimoto is Japanese Laid Open Patent Publication No. JP2004-145025. Fujimoto was filed October 24, 2002 and published May 20, 2004. EX1007. Exhibit 1008 is a certified English translation of Fujimoto. I understand Fujimoto is prior art at least because its publication date predates the October 16, 2006 effective filing date of the challenged patent. I form no opinion as to whether Fujimoto is entitled to an earlier priority date.

73. Fujimoto is titled “Voice Matching Device.” EX1008, 1. Fujimoto discloses “a voice matching device that compares voice signals of registered speakers with those of a speaker to be matched.” *Id.*, ¶9. The voice signals are

received by a “voice input device...equipped with a **main microphone 102** and a **reference microphone 202**.” *Id.*, ¶42.³



Id., Fig. 4.

74. The signal from the **reference microphone 202** is “used to correct the main voice signals supplied by the **main microphone 102**.” *Id.*, ¶44. A cross correlation determiner 302 “cross compares the main voice data with the reference data” (*id.*, ¶62) and “extracts that which is most similar to the main voice data from the reference data based on the results of the comparison” (*id.*, ¶63). The noise

³ All annotations and emphases are added unless otherwise noted.

canceling unit receives the main voice data and extracted reference voice data and, “by comparing them, corrects the main voice data and removes noise.” *Id.*, ¶¶64; *see also id.*, ¶¶83-84.

4. Shimomura (EX1014)

75. Shimomura, U.S. Patent App. Pub. No. 2001/0021909, is entitled “Conversation Processing Apparatus and Method, and Recording Medium Therefor.” Shimomura was filed December 27, 2000 and published September 13, 2001. EX1014. I understand Shimomura is prior art at least because its publication date predates the October 16, 2006 effective filing date of the challenged patent. I form no opinion as to whether Shimomura is entitled to an earlier priority date.

76. Shimomura discloses a robot with “a conversation processing apparatus for holding a conversation with a user.” *Id.*, ¶¶9, 18. Shimomura notes that when a person has a conversation, “it is not common for them to continue to discuss only one topic” because “[i]n general, people change the topic at an appropriate point.” *Id.*, ¶95. Accordingly, Shimomura’s robot “has a function for changing the topic at an appropriate circumstance when having a conversation with a user” (*id.*, ¶96) “so that the conversation with the user continues naturally and fluently” (*id.*, ¶111).

77. The robot includes a microphone that “collects ambient speech (sounds) including the speech of a user.” *Id.*, ¶56. A speech recognition unit “performs speech recognition by using the speech signal supplied from the

microphone.” *Id.*, ¶66. A “language processor 71...extracts the meaning and the intention of the input speech.” *Id.*, ¶79.

78. “Based on the information extracted by the language processor 71, a topic manager 74 manages and updates the present topic in a present topic memory 77.” *Id.*, ¶82. A conversation history memory 75 “stores data used to examine topics which were brought up prior to the present topic, which is stored in the present topic memory 77, and to control the change of topic.” *Id.*, ¶83. The robot can switch to a new topic introduced by the user during the conversation. *Id.*, ¶¶184-93.

5. Matsuda (EX1015, EX1016)

79. Matsuda is Japanese Laid Open Patent Publication No. JP2004-334591. Matsuda was filed May 8, 2003 and published November 25, 2004. EX1015. Exhibit 1016 is a certified English translation of Matsuda. I understand Matsuda is prior art at least because its publication date predates the October 16, 2006 effective filing date of the challenged patent. I form no opinion as to whether Matsuda is entitled to an earlier priority date.

80. Matsuda is titled “Conversation System and Conversation Processing Program.” EX1016, 1. Matsuda describes a conversation system that uses a “subject person number” to track the individual currently being discussed in a conversation. *Id.*, ¶26. This subject person number is reset if “at least a prescribed amount of time (for example, 10 minutes) has elapsed without a name of that person being

mentioned.” *Id.*, ¶¶26, 38. In that case, “the CPU 11 sets the ‘subject person number’ to 0...and clears the ‘subject person number latest setting date/time.’” *Id.*, ¶38. Matsuda explains that this enables “appropriate statements corresponding to the conversation having moved away from the subject of that person.” *Id.*, ¶100.

6. Cooper (EX1017)

81. Cooper, U.S. Patent No. 6,757,362, is entitled “Personal Virtual Assistant.” Cooper issued June 29, 2004. EX1017. I understand Cooper is prior art at least because its issue date predates the October 16, 2006 effective filing date of the challenged patent. I form no opinion as to whether Cooper is entitled to an earlier priority date.

82. Cooper describes a “virtual assistant [that] adapts to the user in many different ways based on the input the virtual assistant receives.” *Id.*, 2:56-64, Abstract. Cooper determines the user’s manner of speech, e.g., by identifying when “words associated with polite discourse are included in input received from the user.” *Id.*, 43:55-59. “If, such words of polite discourse are included in the input received from the user, the virtual assistant could automatically enable the politeness setting.” *Id.*, 43:59-61. A “politeness setting...when enabled, would cause the virtual assistant to include words or phrases associated with polite discourse in the output from the virtual assistant.” *Id.*, 43:48-51. Thus, Cooper’s automatically-

enabled politeness setting ensures that the responses from the virtual assistant will match the manner of speech of the user's utterances.

VIII. SUMMARY OF INVALIDITY GROUNDS

83. In my opinion, the challenged claims are invalid based on the following grounds:

- Ground 1: Kennewick in combination with Oviatt and Fujimoto renders obvious claims 1, 6, 9, 12-16, 21, and 25-28;
- Ground 2: Kennewick-Oviatt-Fujimoto in combination with Shimomura renders obvious claims 2-5 and 17-20;
- Ground 3: Kennewick-Oviatt-Fujimoto in combination with Matsuda renders obvious claims 7, 8, and 22;
- Ground 4: Kennewick-Oviatt-Fujimoto in combination with Cooper renders obvious claims 10, 11, 23, and 24.

IX. INVALIDITY GROUNDS

A. Ground 1: Kennewick in combination with Oviatt and Fujimoto render obvious claims 1, 6, 9, 12-16, 21, and 25-28

1. Claim 1

[1.pre] "A computer-implemented method of facilitating natural language system responses using short-term knowledge generated based on one or more prior multi-modal device interactions, the method being implemented by a computer system that includes one or more physical processors executing one or more

computer program instructions which, when executed, perform the method, the method comprising:”

84. Kennewick in combination with Oviatt discloses preamble [1.pre]. Kennewick relates to an “integrated environment allowing mobile users to ask natural language speech questions or give natural language commands in a wide range of domains.” EX1004, ¶2. The environment “present[s] results in a natural manner even in cases where the question asked or the responses received are incomplete, ambiguous or subjective.” *Id.*, ¶¶2, 23-24 (explaining “the system may simulate some aspects of human ‘personality’”), 101, 155-56 (disclosing a “dialog with the system”).

85. A POSITA would have understood that Kennewick’s environment performs a “***method of facilitating natural language system responses.***” Kennewick discloses that the environment is provided by “systems and methods.” *Id.*, ¶11. A POSITA would have understood Kennewick to disclose that the systems it describes perform the described methods, including a method of facilitating natural language system responses that receives natural language utterances and responds in kind. *See, e.g., id.*, ¶¶14 (the system “responds to the user with a natural language speech response”), 27 (“[t]he system, according to one aspect of the invention, can process and respond to questions, queries and commands”), 28 (“The invention can be used for generalized local or network information query, retrieval and presentation in a mobile environment. For each user utterance including a question

or query or set of questions or queries, the system may perform multiple steps”), 118 (“[t]he natural language interactive speech processing system may make maximum use of context, prior information, location information, domain information and user specific profile data to achieve a natural environment for one or more users making queries or stating commands across multiple domains”), claim 28 (“[a] method responsive to a user generated natural language speech utterance”).

86. Kennewick also discloses the method “*facilitating natural language system responses using short-term knowledge*.” The ’249 patent describes “short-term knowledge” as knowledge that “may accumulate during a single conversation, where input received during a single conversation may be retained.” EX1001, 5:6-8, 14:5-6, 16:15-19 (“Intelligent Hypothesis Builder 310 may use short-term shared knowledge from the Session Input Accumulator to proactively build and evaluate interaction with a user as a conversation progresses”), 16:50-53 (“when a conversation begins, short-term knowledge stored in the Session Input Accumulator may be empty, and as the conversation progresses, the Session Input Accumulator may build a history of the conversation”), 17:17-20 (“short-term shared knowledge may be limited to an existing conversation (where no short-term shared knowledge would be available when a new conversation is initiated)”). The short-term knowledge “may include cross-modality awareness, where in addition to accumulating input relating to user utterances, requests, locations, etc., the shared

knowledge may accumulate a current user interface state relating to other modal inputs to further build shared knowledge models.” *Id.*, 5:8-13.

87. Kennewick’s method likewise accumulates “***short-term knowledge***” during a single conversation. Kennewick describes that during a dialog (a conversation) with the user, the method adds “tags...to each utterance [in the dialog] ...to tie that utterance to the correct...dialog” and user. EX1004, ¶¶155-56. To accomplish such tagging, the method first “recognize[s] words and phrases, using for example, information in the dictionary and phrase tables 112” and “determine[s] the user’s identity by voice and name for each utterance.” *Id.* The method then tags “[r]ecognized words and phrases” “with this identity in all further processing” of the dialog. *Id.* A POSITA would have understood that tagging recognized words and phrases for each utterance to tie that utterance to the correct dialog results in building a dialog history over time containing all recognized words and phrases that are so tagged and associated with the same dialog and user.

88. A POSITA would have understood that Kennewick’s dialog history is “***short-term knowledge***” because it reflects details about a user’s utterances during a conversation. Kennewick’s description of building of a dialog history by systematically tagging each utterance of the dialog (e.g., *id.*, ¶155) is consistent with the ’249 patent’s own disclosure of “short-term knowledge” including a “history of the conversation,” which is built “as the conversation progresses.” See EX1001,

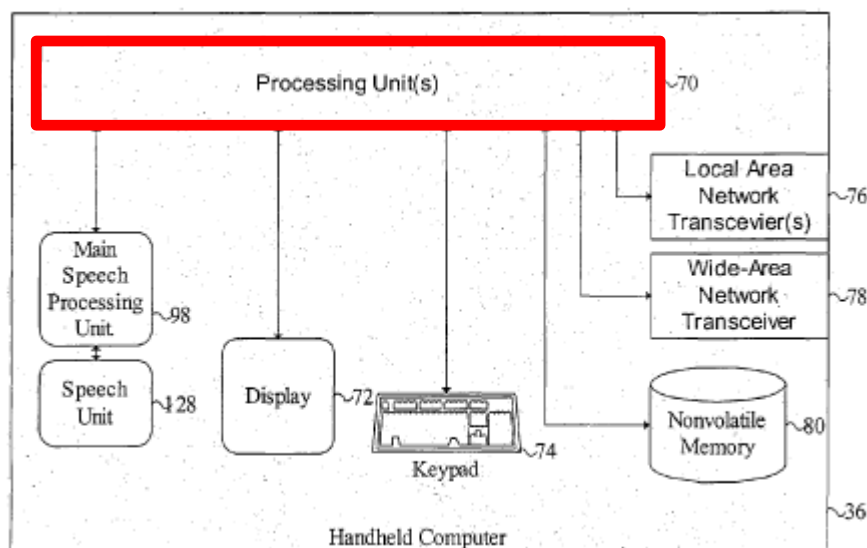
16:50-53 (“[f]or instance, when a conversation begins, short-term knowledge stored in the Session Input Accumulator may be empty, and as the conversation progresses, the Session Input Accumulator may build a history of the conversation”).

89. Kennewick also discloses the method “*facilitating natural language system responses using short-term knowledge*.” Kennewick discloses that “fuzzy set possibilities or prior probabilities” values may be derived from the short-term knowledge. *See* EX1004, ¶¶162 (referring to “the history of the dialog” as a “source” and updating these values “based on user responses...as the system learns the desired behavior.”), 157 (referring to using “user dialog history” to “dynamically update[]” these values). These values are used to “*facilitate[e] natural language system responses*” to the user. *See, e.g., id.*, ¶30 (“[t]he invention may use probabilistic and fuzzy set decision and matching methods to first identify the subjective nature of the question and to evaluate a range of possible answers”). For example, Kennewick’s method uses “the probabilities or possibilities” values based on the dialog history to score possible contexts for an utterance. *Id.*, ¶157. The method also uses the dialog history to “score the relevance of the results” of a query executed in response to a user utterance. *Id.*, ¶170. The relevance score is used to facilitate a response to the user. *See id.*, ¶¶171 (“Based on the on-going scoring processes, the domain agent 156 may determine if a single best answer can be

extracted.”), 175 (“In any case, the domain agent 156 may continue to make queries and evaluate results until a satisfactory response is constructed”).

90. Kennewick also discloses that the method is “*computer-implemented...by a computer system.*” Kennewick’s disclosed method may be implemented on numerous “computer alternatives” that Kennewick collectively refers to as “a computer.” *Id.*, ¶13. The computer may be a “fixed computer” or “handheld device.”⁴ *Id.* Kennewick’s Figure 3 illustrates a possible embodiment of the handheld computer.

⁴ Kennewick uses the terms “handheld,” “handheld device,” and “handheld computer” interchangeably, often with the same reference numeral 36. *See, e.g.*, EX1004, ¶¶22 (using terms “handheld device,” “handheld computer,” and “handheld”), 101 (“handheld devices 36”), 102 (“handheld computers 36”), 110 (“a handheld computer 36 can be used as a component of the invention”).



EX1004, Figure 3, ¶90. Kennewick’s Figure 4 is a near-identical figure illustrating an embodiment of fixed computer. *Id.*, Figure 4, ¶91.

91. A POSITA would have understood that these embodiments of the computer are “computer systems” because they are formed of various computing components working together. *See id.*, ¶¶3 (“[t]eleomatics systems are systems that bring human-computer interfaces to vehicular environments”), 96 (“[a] speech unit 128 can be permanently attached to the vehicle 10 or can be part of a handheld device 36 or a fixed home or office computer system 44”), 107 (“the main speech processing unit 98 and speech unit 128 can be embedded in one or more of the TCUs 28, handheld computers 36 and fixed computer systems 44”).

92. The “*computer system...includes one or more physical processors.*” Figure 3, shown above, illustrates that the computer has “processing units.” These “may include one or more central processing units, one or more data and address

busses, data interfaces and volatile memory.” *Id.*, ¶¶111, 115 (similar), 101, 122, 107-09. A POSITA would have understood that the “central processing units” are one or more physical processors because physical processors acting as central processing units are conventional.

93. A POSITA likewise would have understood that the physical processors “*execut[e] one or more computer program instructions which, when executed, perform the method.*” Kennewick discloses that “[o]ne aspect of the invention includes software that is installed onto the computer,” where the software includes a “a speech recognition module for capturing the user input,” “a parser for parsing the input,” and “a text to speech engine module for converting text to speech.” *Id.*, ¶13. Kennewick also discloses that “domain specific behavior and information are organized into agents” which “are executables that receive, process and respond to user questions, queries and commands.” A POSITA would have understood that Kennewick’s modules and agents are computer program instructions which Kennewick’s physical processors execute to perform the method because this is how computers conventionally operate. *See, e.g., id.*, ¶¶17 (“[a]gents can be packages of executable code”), 107 (“the main speech processing unit 98 and speech unit 128 can be embedded in...the...handheld computers 36 and fixed computer systems 44”).

94. For these reasons, a POSITA would have understood Kennewick to disclose “[a] computer-implemented method of facilitating natural language system responses using short-term knowledge..., the method being implemented by a computer system that includes one or more physical processors executing one or more computer program instructions which, when executed, perform the method.”

95. Kennewick also discloses that the computer system is a “**multi-modal device**.” In addition to a **speech unit**, the computer system may include “**displays** 18 including touch screen displays for text, graphics and video, keypads 14 for data input..., and a pointing device or stylus.” *Id.*, ¶97.

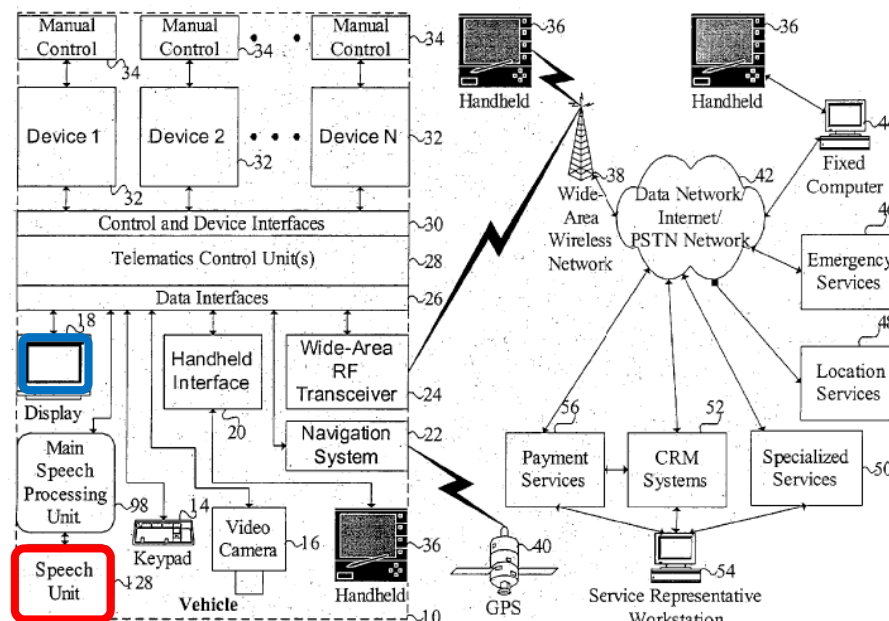
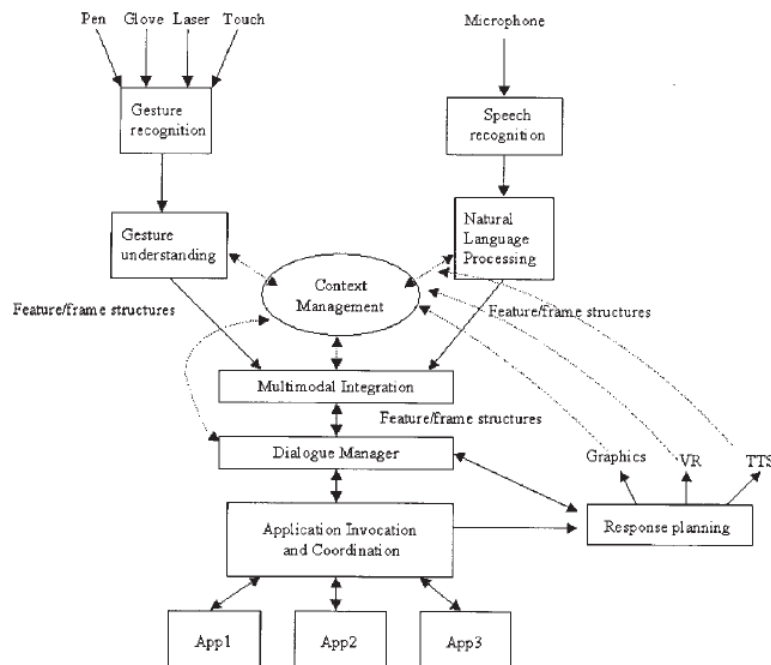


Figure 2. Second Embodiment System Block Diagram

Id., Figure 2. A POSITA would have understood that the computer system is “**multi-modal device**” because it accepts different modes of input, including voice, touch screen, and keypad input.

96. Kennewick does not explicitly disclose that short term knowledge used by the system is “**generated based on one or more prior multi-modal device interactions.**” This aspect is taught by Oviatt.

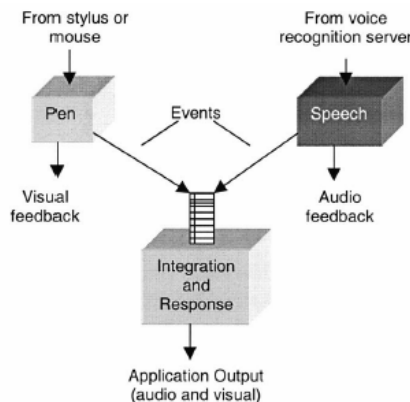
97. Oviatt discloses a “basic information flow needed to process multimodal speech and gesture input.” EX1005, 14.



Id., 15. Here, “speech and pen-based gestures are recognized in parallel, and each is processed by an understanding component.” *Id.*, 14-15. “The results are meaning representations that are fused by the multimodal integration component, which also

is influenced by the system's dialogue management and interpretation of current context.” *Id.*, 15.

98. Oviatt discloses a portable voice assistant (“PVA”), a “***multi-modal device***” that “can interpret simultaneous pen-voice input.” *Id.*, 39. For example, the user can say “show me that part” “while pointing at a screw on the display.” *Id.* “[T]he PVA time stamps both speech and pen input events and integrates them to form a frame-based description of the user’s request.” *Id.*, 39.



Id., 40. In particular:

[t]he speech thread generates events in response to output from the speech recognizer, and the pen thread generates selection events that are stored in a time-sorted buffer where they can be retrieved by the integrator thread. For this example, the integrator simply would look for a pen event that occurred closest in time to the spoken deictic event “that.” This design allows for asynchronous processing of speech and pen-based input, while still processing them as coordinated multimodal pieces.

Id., 39. A POSITA would have understood that the generated events are “**prior multi-modal device interactions**” because the events describe multimodal interactions that occurred earlier. The PVA generates the speech and thread events based on the interactions and subsequently processes the events. Thus, at processing time the events describe multimodal device interactions that already occurred, which are “**prior multi-modal device interactions.**”

99. A POSITA would have found it obvious to modify Kennewick to support multimodal inputs as taught by Oviatt, and to generate short-term knowledge from these inputs, thereby producing “**short-term knowledge generated based on one or more prior multi-modal device interactions.**” Kennewick and Oviatt are analogous references; both relate to human-computer interactions involving a speech interface. *See* EX1004, ¶2 (“[t]he present invention relates to the retrieval of online information and processing of commands through a speech interface”); EX1005, 6 (“[i]n this article, we introduce multimodal speech and pen-based gesture interfaces”). A POSITA would have been motivated to make this combination because interfaces involving multimodal input “are particularly effective for supporting mobile tasks, such as communications and personal navigation.” *Id.*, 10.

100. For example, a POSITA would have been motivated to combine Kennewick and Oviatt in order to provide navigation services that leverage both voice and touch-screen input. Kennewick discloses using the computer system

(which includes multimodal input devices) for personal navigation. EX1004, ¶¶15 (“[v]ehicle systems employed can include...navigation equipment”), 61 (“[t]he invention can provide vehicle operators interactive directions to a destination or waypoint”). Kennewick’s “speech unit 128 can be permanently attached to [a] vehicle 10 or can be part of a handheld device 36.” *Id.*, ¶96. The “speech-processing unit built into [the] handheld device may be connected with...data interfaces 26” with which “navigation system components” are also connected. *Id.*, ¶97.

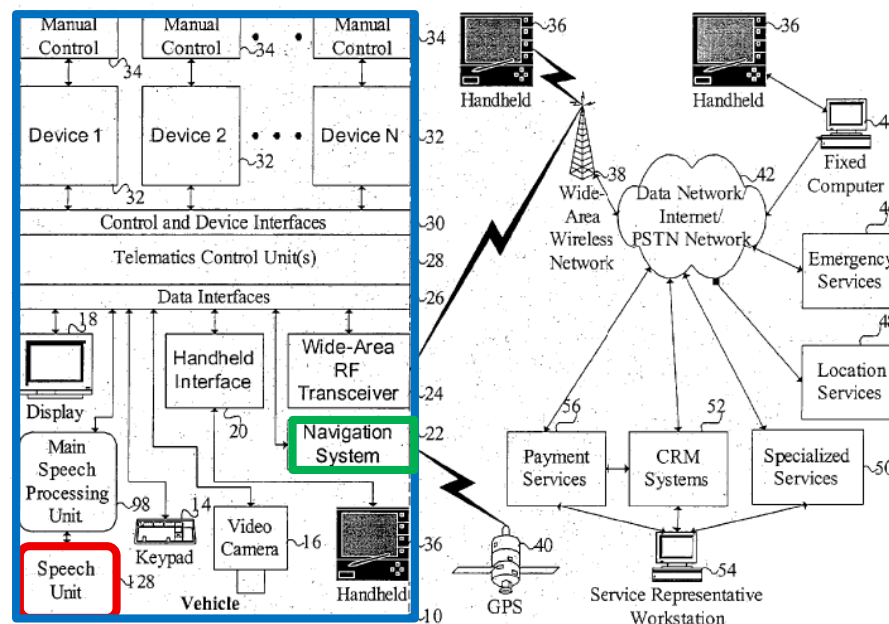


Figure 2. Second Embodiment System Block Diagram

EX1004, Figure 2. Oviatt discloses that the PVA “can interpret simultaneous pen-voice input” (EX1005, 39) and that “[f]uture field applications of this [PVA] system could include...in-car navigation systems.” EX1005, 40.

101. A POSITA would have found it obvious to incorporate Oviatt's "*prior multi-modal device interactions*" (e.g., "speech and pen-based gestures," *id.*, 14-15) into Kennewick's short-term knowledge. This combination allows the user to use voice in combination with Kennewick's existing touch-screen and stylus to control Kennewick's navigation equipment. For example, the user could say "give me directions to here," where the user indicates the place corresponding to "here" by pointing to a location on a map presented on the touch screen display. A POSITA would have been motivated to make this combination because "[a] pen stylus is a natural, precise, and effective mode of interaction for spatially oriented tasks." EX1005, 29.

102. This is a combination of prior art elements according to known methods to yield predictable results. Kennewick discloses a natural language speech interface and also describes computer interfaces using "point and click techniques and touch screen displays" as conventional. EX1004, ¶¶3-4. Oviatt discloses prior art techniques involving multimodal device interactions including speech and touch-screen interactions. A POSITA would have found it obvious to use Oviatt's techniques within Kennewick's interface, would have found the results predictable, and would have reasonably expected the results to be successful, because Oviatt explains how to use multimodal inputs to a computer.

103. The combination of Kennewick and Oviatt is also the application of a known technique – Oviatt’s multimodal interactions involving a touch screen– to a known device and method – Kennewick’s computer having a touch screen– ready for improvement. A POSITA would have found the results of this combination predictable, and would have reasonably expected the results to be successful, because Oviatt explains how to use touch screen and voice inputs together.

[1.a] “receiving, by the computer system during a first conversation, a first voice input via a first input device, the first voice input comprising a first natural language utterance;”

104. Kennewick discloses limitation [1.a]. Kennewick’s “interactive natural language speech system,” which is “***the computer system,***” includes a speech unit. EX1004, ¶12 (“a mobile interactive natural language speech system (herein “the system”) is provided that includes a speech unit”). The speech unit “***receiv[es] a first voice input via a first input device.***” *Id.*, ¶121 (“[t]he speech unit 128 includes one or more microphones, for example array microphone 134, to receive the utterances of the user”).

105. For example, the speech unit may include an array microphone that receives voice inputs, which are user utterances. EX1001, 2:26-28 (“an ‘utterance’ may be words, syllables, phonemes, or any other audible sound made by a human being”); EX1004, ¶¶38 (“[t]he system can use either one-dimensional or two-dimensional array microphones...to receive human speech”), 121 (“[t]he speech unit

128 includes one or more microphones, for example array microphone 134, to receive the utterances of the user”), 149 (“[t]he speech received by the array microphone 134 and passed through the filter 132 may be sent to the speech digitizer or coder 138”). A POSITA would have understood that a first microphone of an array of microphones of the speech unit is a “*first input device*” because it is a device that receives “*a first voice input*” from a user.

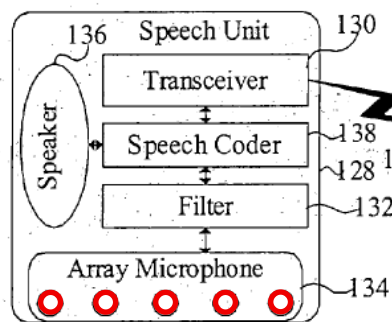
106. The “*first voice input*” received “*via a first input device*” comprises “*the first natural language utterance*” because it is a natural language utterance provided by the user. EX1004, ¶¶2 (“the invention is a fully integrated environment allowing mobile users to ask natural language speech questions or give natural language commands in a wide range of domains”), 73 (“[v]ehicle operators and other occupants can use the interactive natural language interactive speech interface of the invention to perform many types of information query, retrieval and presentation operations”), 101 (“[i]n every case, the invention may provide an interactive natural language speech user interface to the services offered”), 152 (“the invention is intended to be able to accept most any natural language question or command”), claims 1, 28, 43 (reciting variations of “receiving the user generated natural language speech utterance”). Kennewick’s “*computer system*,” by “*receiving*” the natural language voice input from a user “*via a first input device*,” therefore receives “*the first voice input comprising a first natural language utterance*.”

107. Kennewick also discloses that the “*computer system*” receives the “*first voice input*” “*during a first conversation.*” Kennewick discloses that the “dialog with the system may begin when a user first addresses it” such as “by speaking a generic word (‘computer’) and/or addressing a specific name (‘Fred’).” EX1004, ¶156. “At the end of a dialog or to interrupt a dialog, the user may speak a dismissal word (‘good by’).” *Id.* A POSITA would have understood that, in the context of Kennewick, a dialog is a conversation between the user and the computer system. Thus, a natural language utterance received by “*the computer system*” during a dialog is a “*first voice input comprising a first natural language utterance*” received “*during a first conversation.*”

[1.b] “receiving, by the computer system, a second voice input comprising the first natural language utterance via a second input device;”

108. Kennewick discloses limitation [1.b]. As I discuss for limitation [1.a], Kennewick’s computer system includes a speech unit that receives the first voice input comprising the first natural language utterance via a first input device. EX1004, ¶¶121 (“[t]he speech unit 128 includes one or more microphones...to receive the utterances of the user”), 152 (“the invention is intended to be able to accept most any natural language question or command”). The speech unit may include an array microphone. *Id.*, ¶38 (“[t]he system can use either one-dimensional or two-dimensional array microphones (or other devices) to receive human speech”), 121 (“[t]he speech unit 128 includes one or more microphones, for example array

microphone 134, to receive the utterances of the user”), 149 (“the speech received by the array microphone 134”). A POSITA would have understood that the array microphone includes multiple microphones for receiving natural language utterances. See EX1011, 530 (“[m]icrophone arrays use multiple microphones”). Kennewick illustrates the array as having **separate microphones**:



EX1004, Figure 5 (excerpt).

109. Kennewick also describes other configurations in which the speech unit has multiple microphones. For example, the speech unit can have “one or more external microphones.” *Id.*, ¶121. Alternatively, “the array microphone can be replaced by a set of physically distributed microphones or headsets worn by the users” or by distributed microphones “placed in different parts of a vehicle or room.” *Id.*, ¶151.

110. A POSITA would have understood that a second microphone of the computer system’s speech unit, such as second microphone in an array, is a “**second input device**” because it is a second device that receives a voice input from a user.

A POSITA also would have understood that the voice input is “*a second voice input*” because it is distinct from the first voice input received by the first input device of limitation [1.a]. Therefore, a POSITA would have understood that Kennewick’s “*computer system*,” receives, “*via [the] second input device*,” “*a second voice input*.”

111. The second voice input “*compris[es] the first natural language utterance*.” The first and second input devices are in close proximity. For example, the first and second input devices may be adjacent microphones in the array or microphones within the same vehicle or room. A POSITA would have understood that when the user makes an utterance near the microphones, the utterance is received by both microphones. *See, e.g.,* EX1004, ¶¶121 (“[t]he speech unit 128 includes one or more microphones, for example array microphone 134, to receive the utterances of the user”), 147 (“the invention may use one or more arrays of microphones 134 to provide better directional signal capture and noise elimination than can be achieved with a single microphone”), 149 (“the speech received by the array microphone 134”); EX1011, 530-531 (“microphone arrays use multiple microphones and knowledge of the microphone locations to predict delays and thus create a beam that focuses on the direction of the desired speaker and rejects signals coming from other angles”). Therefore, a POSITA would have understood and

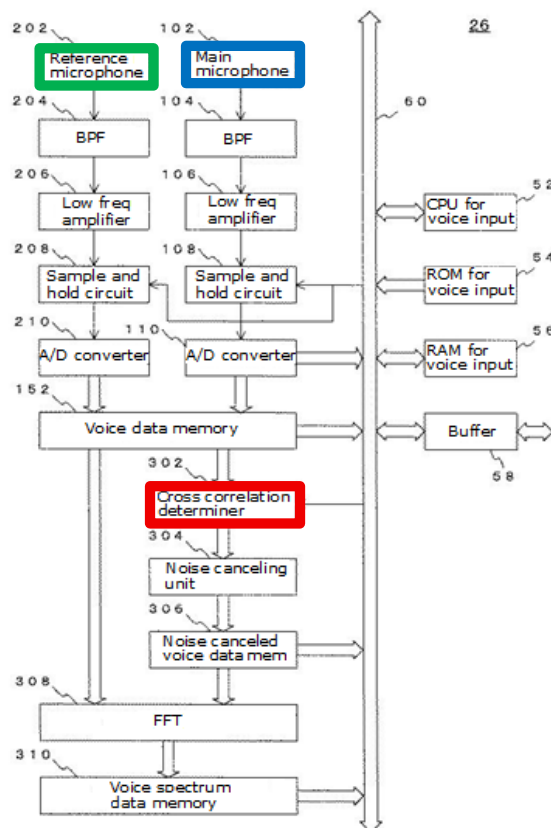
found obvious that the “*computer system*” receives “*a second voice input comprising the first natural language utterance via [the] second input device.*”

[1.c] “comparing, by the computer system, the first voice input with the second voice input;”

112. Kennewick in combination with Fujimoto discloses limitation [1.c]. Kennewick discloses “*the computer system*” as I discuss for preamble [1.pre]. Kennewick also discloses the computer system receiving “*the first voice input*” and “*the second voice input*” as I discuss for limitations [1.a]-[1.b]. Fujimoto also discloses “*comparing the first voice input with the second voice input.*”

113. Fujimoto discloses a “voice input device 26...equipped with a [main microphone 102](#) and a [reference microphone 202](#).” EX1008, ¶42. The “[main microphone 102](#) receives main voice signals.” *Id.*, ¶43. The [reference microphone 202](#), in contrast, “receives voice signals...that are used to correct the main voice signals supplied by the main microphone 102.” *Id.*, ¶44.

114. A [cross correlation determiner 302](#) compares the signal from the [main microphone](#) with the signal from the [reference microphone](#):



EX1008, Figure 4. In particular, “the cross correlation determiner 302 reads the main voice data supplied by the main microphone 102 and the reference data supplied by the multiple reference microphones, including reference microphone 202 from the voice data memory 152.” *Id.*, ¶83. It then “compares the main voice data with the reference data, and as a result of this comparison, extracts the reference data that is most similar to the main voice data from the reference data.” *Id.*; see also *id.*, ¶62 (“The cross correlation determiner 302 reads the main voice data supplied from the main microphone 102 and the reference data supplied from the multiple reference microphones including reference microphone 202..., and cross compares the main voice data with the reference data”). “The cross correlation

determiner” then “outputs the main voice data and the extracted reference voice data.” *Id.*, ¶63.

115. A POSITA would have found it obvious to use Fujimoto’s comparison within Kennewick’s computer system to “*compar[e]...the first voice input with the second voice input.*” Both references relate to receiving user utterances. *See, e.g.*, EX1004, ¶2 (“the invention is a fully integrated environment allowing mobile users to ask natural language speech questions or give natural language commands”); EX1008, ¶9 (“the present invention is a voice matching device that compares voice signals of registered speakers”). Both references also disclose receiving voice inputs via multiple input devices. *See, e.g.*, EX1004, ¶121 (“[t]he speech unit 128 includes one or more microphones...to receive the utterances of the user”); EX1008, ¶77 (“voice input processing is executed using multiple microphones”).

116. Given these similarities, a POSITA would have been motivated to modify Kennewick’s computer system to compare the first and second voice inputs as part of a process to reduce noise in order to improve speech recognition. Kennewick specifically discloses that the computer system uses multiple microphones “to provide better...noise elimination than can be achieved with a single microphone.” EX1004, ¶147. Kennewick also discloses that “noise elimination may be a key aspect of the invention” and “[t]he accurate recognition and parsing of the user’s speech requires the best possible signal to noise ratio at the

input to the speech recognition engine 120.” *Id.*, ¶146. Fujimoto discloses comparing first and second voice inputs as part of a noise reduction process. EX1008, ¶83 (“The cross correlation determiner 302 compares the main voice data with the reference data, and as a result of this comparison, extracts the reference data that is most similar to the main voice data from the reference data. The cross correlation determiner 302 then supplies the main voice data and the extracted reference voice data...to the noise canceling unit 304.”). Huang confirms that cross-correlation is used to reduce noise. EX1011, 532 (“[t]o attenuate the additive noise $v[n]$, it is not necessary to identify the absolute delays, but rather the delays relative to one reference microphone...[i]t can be shown that the maximum likelihood solution consists in maximizing the energy of...the sum of cross-correlations”).

117. This combination of Fujimoto’s comparison with Kennewick’s computer system is a combination of prior art elements according to known methods to yield the predictable result of noise reduction. It is also the application of a known technique (cross-correlation) as taught in Fujimoto to a known device (Kennewick’s computer system) ready for improvement to yield the predictable result of extracting data for use in noise reduction. A POSITA would have had a reasonable expectation of success because comparing two signals is well-understood, and Kennewick states that noise reduction can be performed with its multiple microphones and leads to more “accurate recognition and parsing of the user’s speech.” EX1004, ¶146.

[1.d] “filtering, by the computer system, sound from the first voice input and the second voice input based on the comparison;”

118. Kennewick in combination with Fujimoto discloses limitation [1.d]. Kennewick discloses “***the computer system***” as I discuss for preamble [1.pre]. Kennewick in combination with Fujimoto discloses “***the comparison***” as I discuss for limitation [1.c]. Fujimoto additionally discloses and renders obvious “***filtering...sound from the first voice input and the second voice input based on the comparison.***”

119. I construe limitation [1.d] in accordance with “the ordinary and customary meaning of such claim as understood by one of ordinary skill in the art and the prosecution history pertaining to the patent.” 37 C.F.R. § 42.100(b). Here, the prosecution history reflects that the Examiner understood “***sound from the first voice input and the second voice input***” to mean “sound that is based on both voice inputs from the first and second input devices.” EX1003, 40, 121. The prosecution history further indicates the Examiner understood limitation [1.d] to mean “filtering sound based on multiple microphone signals based on comparison of those multiple microphone signals to each other.” EX1003, 40-41, 121. I therefore apply the Examiner’s understanding as the ordinary and customary meaning of limitation [1.d].

120. As I discuss for limitation [1.c], Fujimoto’s cross-correlation determiner compares the signal from the main microphone with the signal from the

reference microphone (EX1008, ¶62) and “outputs the main voice data and the extracted reference voice data” (*id.*, ¶63). This output is “*sound from the first voice input* [the main voice data] *and the second voice input* [the extracted reference voice data].” The output is received by Fujimoto’s noise canceling unit 304, which then “corrects the main voice data and removes noise.” *Id.*, ¶64 (“This noise canceling unit 304 receives the main voice data supplied by the cross correlation determiner 302 and the extracted reference voice data, and by comparing them, corrects the main voice data and removes noise.”). Fujimoto does not provide details of how this noise-removal process happens.

121. However, it would have been clear to a POSITA that Fujimoto’s noise-removal process involves “*filtering*.” A POSITA would have understood that a “filter” is “hardware or software that selectively passes certain elements of a signal and eliminates or minimizes others.” EX1020, 3. Thus, Fujimoto’s noise canceling unit 304, which receives two signals (the main voice data and the extracted reference voice data) and outputs a signal that “corrects the main voice data and removes noise,” is “filtering” the input signals by removing noise (i.e., eliminating or minimizing noise). *See* EX1008, ¶¶64 (“This noise canceling unit 304 receives the main voice data supplied by the cross correlation determiner 302 and the extracted reference voice data, and by comparing them, corrects the main voice data and removes noise.”), 84 (“[T]he noise canceling unit 304 receives the main voice data

and the extracted reference voice data supplied by the cross correlation determiner 302 and compares them. The noise canceling unit 304 then corrects the main voice data so that the difference between the main voice data and the extracted reference voice data is reduced.”).

122. For example, a POSITA would have found it obvious for Fujimoto’s noise canceling unit to implement a filter averaging the two input signals (the main voice data and the extracted reference voice data) since it would have been well-known to a POSITA that averaging signals from multiple microphones is a way to reduce noise. *See, e.g.*, EX1011, 531-532 (describing equation 10.64 which discloses obtaining the “optimal estimate” (e.g., least noisy estimate) of a source signal x by averaging received signals y_i from N microphones each with known delays D_i , and equation 10.66 which discloses using cross-correlation to obtain the “maximum likelihood solution” for the relative delay of one microphone relative to another).

123. The filtering is “*based on the comparison*” because the noise canceling unit that performs the filtering receives and operates on the main voice data and the extracted reference voice data output by the cross correlation determiner that performs the comparison. EX1008, ¶¶63-64 (“[T]he cross correlation determiner 302...outputs the main voice data and the extracted reference voice data (“extracted reference voice data” below)...This cross correlation determiner 302 is connected to

a noise canceling unit 304 [which] receives the main voice data supplied by the cross correlation determiner 302 and the extracted reference voice data.”), 84 (“the noise canceling unit receives the main voice data and the extracted reference voice data supplied by the cross correlation determiner”).

124. Accordingly, a POSITA would have found the computer system “*filtering...sound from the first voice input and the second voice input based on the comparison*” obvious in view of Fujimoto’s description of the noise canceling unit, which receives two signals (the main voice data and the extracted reference voice data) and outputs a signal that “corrects the main voice data and removes noise.” EX1008, ¶64. The noise canceling unit filters “sound that is based on both voice inputs from the first and second input devices” (EX1003, 40, 121), the main voice data and extracted reference voice data. It also “filter[s] sound based on multiple microphone signals based on comparison of those multiple microphone signals to each other” (EX1003, 40-41, 121) because it filters the main and extracted reference voice data from the main and reference microphones based on the output of the cross correlation determiner that performs the comparison.

125. A POSITA would have found it obvious to perform Fujimoto’s noise canceling within Kennewick’s computer system to “*filter[]...sound from the first voice input and the second voice input based on the comparison*” for the same reasons I provide for limitation [1.c]. A POSITA would have been motivated to

modify Kennewick's computer system to perform Fujimoto's noise canceling in order to reduce noise in order to improve speech recognition. EX1004, ¶¶146-47; EX1008, ¶¶64, 84. In addition, Kennewick discloses that the computer system "may use an analog or digital [filter] 132 between the array microphone or conventional microphone 134 and the speech coder 138...to optimize the signal to noise ratio at the input to the speech recognition engine 120." EX1004, ¶148. A POSITA would have found it obvious to optimize the signal by using the filtering process described by Fujimoto, which "removes noise." EX1008, ¶64.

126. This combination of Fujimoto's noise reduction with Kennewick's computer system is a combination of prior art elements according to known methods to yield the predictable result of noise reduction. It is also the application of a known technique (noise reduction) as taught in Fujimoto to a known device (Kennewick's computer system) ready for improvement to yield the predictable result reducing noise in a signal. A POSITA would have had a reasonable expectation of success because noise reduction is well-understood leads to more "accurate recognition and parsing of the user's speech." EX1004, ¶146.

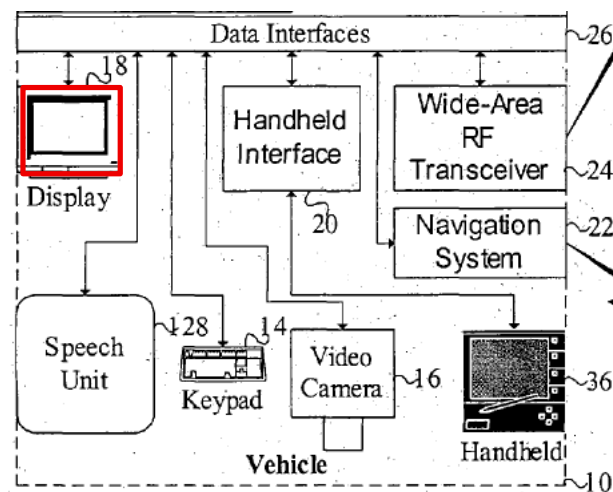
[1.e] "obtaining, by the computer system during the first conversation, a user interface state related to one or more non-voice inputs associated with the first

voice input, the one or more non-voice inputs comprising at least a first non-voice input;

127. Kennewick in combination with Oviatt discloses limitation [1.e]. Kennewick discloses “*the computer system*” receiving “*the first voice input*” “*during the first conversation*” as I discuss for limitation [1.a].

128. Kennewick also discloses that the computer system may include input devices for “*obtaining*” “*one or more non-voice inputs.*” EX1004, ¶97. For example, the computer system may include “touch screen displays [18] for text, graphics and video, keypads 14 for data input..., and a pointing device or stylus.”

Id.



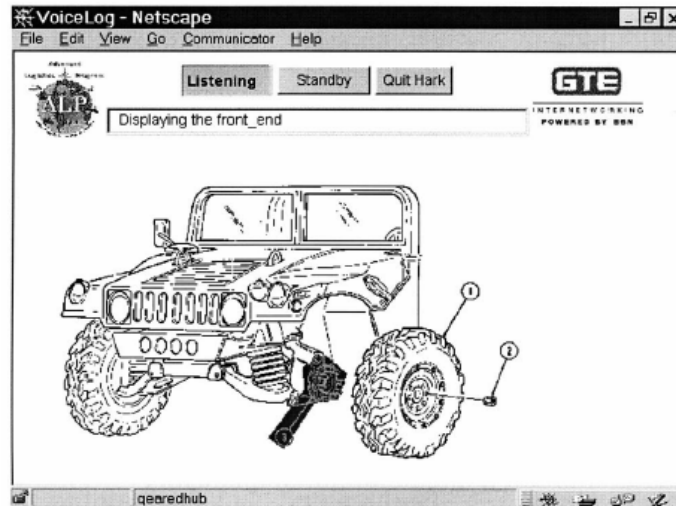
EX1004, Figure 1 (excerpt), ¶¶112 (“users interact with the handheld computer 36 through the speech unit 128, the keypad 74 or keyboard, and a display 72 for text, graphics and video. In some embodiments the display is of a touch screen type. An optional pointing device (not shown) may be used as well”), 116 (same). Kennewick recognizes that these types of non-voice input devices are conventional. *Id.*, ¶3

(“[c]onventional computer interfaces use some combination of keyboards, keypads, point and click techniques and touch screen displays”).

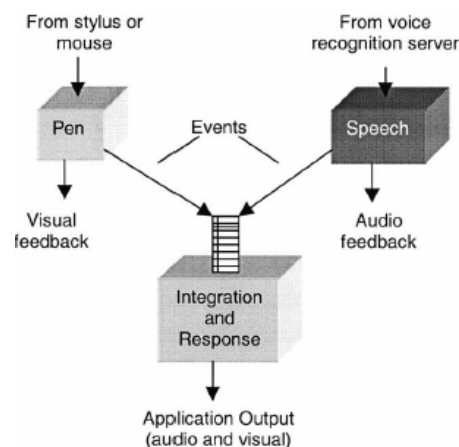
129. Kennewick also discloses that a non-voice input device has “***a user interface***” with which the user can interact. A “graphical user interface 114 can be used as a...complement to the speech interface.” *Id.*, ¶125. “For example, the graphical user interface can be used to view and interact with graphical or tabular information in a manner more easily digested by the user...[and] [t]he graphical user interface can show system state and history in a more concise manner than the speech interface.” *Id.*

130. But Kennewick does not explicitly disclose “***obtaining, by the computer system during the first conversation, a user interface state related to one or more non-voice inputs associated with the first voice input, the one or more non-voice inputs comprising at least a first non-voice input***” as recited by limitation [1.e]. This is taught by Oviatt.

131. As I discuss for preamble [1.pre], Oviatt discloses a PVA with a “a pen-voice multimodal interface...available over a wireless network using a handheld device.” EX1005, 38. Oviatt’s PVA has “***a user interface***” illustrating images the user can select using a pen and touch screen:



EX1005, 39 (Figure 13). Oviatt’s PVA “can interpret simultaneous pen-voice input so, for example, the user can say ‘Show me that part’ while pointing at a screw on the display.” *Id.* “[T]he PVA time stamps both speech and pen input events and integrates them to form a frame-based description of the user’s request, as illustrated in Figure 14.” *Id.*



Id., 40 (Figure 14).

132. A POSITA would have understood from this disclosure that Oviatt discloses “*obtaining...a user interface state related to one or more non-voice*

inputs...comprising at least a first non-voice input.” The “*first non-voice input*” is the user “pointing at a screw on the display.” *Id.*, 39. The PVA “*obtain[s]...a user interface state related to one or more non-voice inputs...comprising at least a first non-voice input*” by receiving an event indicating when the user pointed to (touched) the display and to where the user pointed on the user interface presented by the display (e.g., the part pointed to by the user). *Id.* (“the PVA time stamps...pen input events...the pen thread generates selection events that are stored in a time-sorted buffer”), 39-40 (the identity of the selected part “is included in the...pen input event[] when the...user’s pen selects the image”). This event indicates the state of the user interface when the user points to it, and the “*user interface state*” is therefore “*related to one or more non-voice inputs...comprising at least a first-non-voice input.*”

133. Oviatt also discloses that the “*non-voice input[] is associated with the first voice input.*” An integrator thread associates the voice and non-voice inputs based on time. Thus, when the user says “‘Show me that part’ while pointing at a screw on the display” “the integrator simply would look for a pen event that occurred closest in time to the spoken deictic event ‘that.’” *Id.*

134. Given the complementary disclosures of Kennewick and Oviatt, a POSITA would have found it obvious to combine the references so Kennewick’s computer system “*obtain[s]...during the first conversation, a user interface state*

related to one or more non-voice inputs associated with the first voice input, the one or more non-voice inputs comprising at least a first non-voice input during the first conversation.” For example, a POSITA would have found it obvious to combine the references in the navigation context, where a user, “*during the first conversation*” with Kennewick’s computer system utters “give me directions to here” while indicating the location corresponding to “here” by touching a location on the touch screen display. Kennewick discloses that the user converses with the computer system as I discuss for limitation [1.a] and that the computer system includes a touch screen display and “navigation system components.” EX1004, ¶97. Oviatt discloses “*obtaining...a user interface state related to [a] non-voice input[] associated with the first voice input,*” and suggests using this disclosure in a navigation system. EX1005, 40.

135. A POSITA would have been motivated to combine the references in this manner because “[i]nterfaces involving spoken or pen-based input, as well as the combination of both, are particularly effective for supporting mobile tasks, such as communications and personal navigation.” *Id.*, 10; *see also id.*, 8 (“pen input to maps and other graphic displays can easily and efficiently convey spatial information”), (“[t]ogether, the speech and pen modes easily can be used to provide flexible descriptions of objects, events, spatial layouts, and their interrelation”).

136. As I discuss for preamble [1.pre], this is a combination of prior art elements according to known methods to yield predictable results. Kennewick discloses using voice inputs in a computer system having a conventional touch screen. EX1004, ¶3. Oviatt discloses a known way to effectively combine both voice and non-voice inputs on a computer system having a touch screen. EX1005, 38-40. The combination is also the application of a known technique (Oviatt's combination of voice and non-voice inputs) to a known device ready for improvement (Kennewick's computer system with voice input and a touch screen) to yield predictable results, and use of a known technique improve a similar device (computer with voice and touch screen inputs) in the same way. A POSITA would have found the combination predictable because these are well-understood technologies, and would have reasonably expected the combination to be successful for the same reason.

[1.f] “generating, by the computer system, the short-term knowledge based on at least the first voice input and the first non-voice input;”

137. Kennewick combined with Oviatt discloses limitation [1.f]. Kennewick and Oviatt disclose a computer system implementing a method of using short term knowledge generated based on one or more prior multi-modal device interactions as I discuss for preamble [1.pre]. These references combined also disclose obtaining a user interface state related to one or more non-voice inputs comprising at least a first non-voice input as I discuss for limitation [1.e]. Given

this combination, a POSITA would have found it obvious that the “**computer system**” “**generat[es] the short-term knowledge based on at least the first voice input and the first non-voice input.**”

138. As I discuss for preamble [1.pre], Kennewick’s dialog history is short-term knowledge because it reflects details about a user’s utterances during a conversation. EX1004, ¶¶23, 25, 29, 31, 37, 155, 157, 160. Kennewick’s computer system generates the short-term knowledge by systematically tagging each utterance of the dialog. *Id.*, ¶155 (“[r]ecognized words and phrases are tagged with [the user’s] identity in all further processing”). Kennewick thus discloses the “**computer system**” “**generating...the short-term knowledge based on at least the first voice input.**”

139. Oviatt discloses “**generating...the short-term knowledge based on at least first voice input and the non-voice input**” because it receives voice and non-voice inputs and generates short-term knowledge based on these inputs. As I discuss for [1.pre] and limitation [1.e], Oviatt describes a PVA that “can interpret simultaneous pen-voice input...and integrates them to form a frame-based description of the user’s request.” EX1005, 39. The PVA includes a speech thread that “generates events in response to output from the speech recognizer” and a pen thread that “generates selection events” that “are stored in a time-sorted buffer where they can be retrieved by the integrator thread.” *Id.*

140. A POSITA would have understood that these events and description are “*generat[ed]...short-term knowledge based on at least first voice input and the first non-voice input*” at least because they include a history of the user’s voice and non-voice interactions with the PVA built as the user’s conversation progresses. See EX1001, 16:50-53 (“[f]or instance, when a conversation begins, short-term knowledge stored in the Session Input Accumulator may be empty, and as the conversation progresses, the Session Input Accumulator may build a history of the conversation”); EX1005, 15 (“[t]he results [of processing the speech and pen-based gestures] are meaning representations that are fused by the multimodal integration component, which also is influenced by the system’s dialogue management and interpretation of current context”).

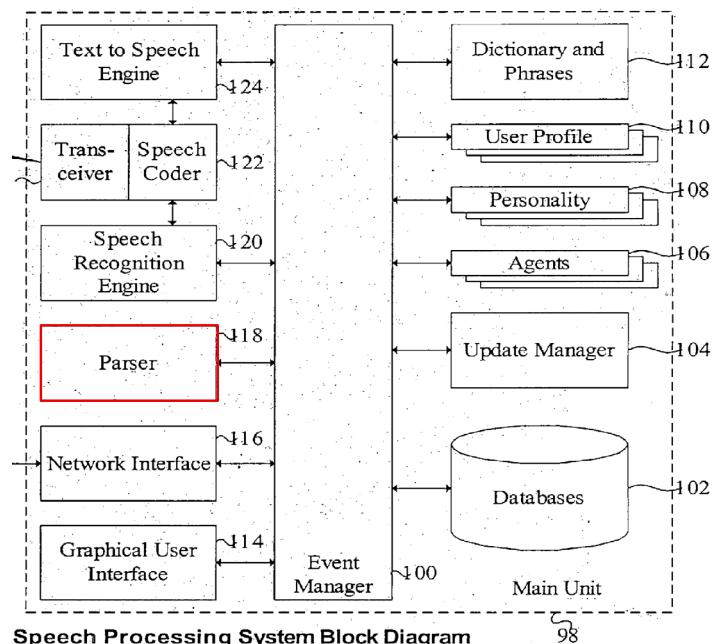
141. A POSITA would have found it obvious to combine Kennewick and Oviatt so “*the computer system*” “*generat[es] the short-term knowledge based on at least the first voice input and the first non-voice input.*” For example, a POSITA would have found it obvious for Kennewick’s computer system to incorporate Oviatt’s first non-voice input into the dialog history by tagging it along with the first voice input to generate the short-term knowledge because Kennewick teaches tagging voice inputs and Oviatt teaches using a similar approach for voice and non-voice inputs. See, e.g., EX1004, ¶155 (“tags, added by the speech recognition engine 120 to each utterance, may allow other components of the system to tie that utterance

to the correct user and dialog”); EX1005, 62 (“the PVA time stamps...speech...input events and...generates events in response to output from the speech recognizer”). In another example, a POSITA would have found it obvious to store the history information Oviatt’s PVA generates based on the user’s first voice input and the first non-voice input as part of Kennewick’s short term knowledge. A POSITA would have been motivated to combine the references in these ways to allow Kennewick’s system to use multimodal inputs, which are “are particularly effective for supporting mobile tasks, such as communications and personal navigation” (EX1005, 10) in addition to the other motivations I provide for [1.pre] and limitation [1.e].

[1.g] “determining, by the computer system, based on the short-term knowledge, a first context for the first natural language utterance;”

142. Kennewick discloses limitation [1.g]. Kennewick discloses receiving the first natural language utterance as I discuss for limitation [1.a]. Kennewick and Oviatt disclose the computer system generating and using short-term knowledge as I discuss for preamble [1.pre] and limitation [1.f]. Kennewick also discloses the “computer system” “determining...based on the short-term knowledge, a first context for the first natural language utterance.”

143. Kennewick discloses that the computer system’s speech processing unit includes a **parser** that “may determine a context for an utterance by applying prior probabilities or fuzzy possibilities to keyword matching, user profile 110, dialog history, and context stack contents.” EX1004, ¶160.



Speech Processing System Block Diagram

EX1004, Figure 5 (excerpt). Kennewick also explains that the dialog history is used to determine context for the utterance because “some questions can only be interpreted in the context of previous questions” in the dialog. *Id.*, ¶7.

144. Accordingly, a POSITA would have understood Kennewick to disclose and render obvious the “*computer system*” “*determining...based on the short-term knowledge, a first context for the first natural language utterance*” because the parser uses the dialog history, short term knowledge, to determine a context (“*a first context*”) for the utterance.

[1.h] “*determining, by the computer system, based on the first context, an interpretation of the first natural language utterance; and*”

145. Kennewick discloses limitation [1.h]. As I explain for limitation [1.f], Kennewick’s computer system determines a first context for the first natural language utterance. EX1004, ¶160.

146. Kennewick also discloses the computer system “*determining...based on the first context, an interpretation of the first natural language utterance.*” *Id.* Kennewick’s system includes a parser that interprets natural language utterances received from the user. *Id.*, ¶¶28 (“[f]or each user utterance including a question or query or set of questions or queries, the system may perform multiple steps possibly including...parsing and interpreting the question or query”), 41 (“[f]or each user command utterance, the system may execute a number of steps possibly including...parse and interpret the command”), 143 (“[c]ommands and questions are interpreted”), 155 (“[t]he speech recognition engine 120 may recognize words and phrases...and pass these to the parser 118 for interpretation”).

147. The parser determines the interpretation based on the context of the utterance. *Id.*, ¶7 (“some questions can only be interpreted in the context of previous questions”). In particular, the parser uses recognized words and phrases, context, and other criteria to determine the appropriate context-based agent to be invoked for a user utterance forming a question or command. *Id.*, ¶¶160 (“[t]he context of a question or command may determine the domain and thereby, the domain agent 156, if any, to be invoked”), 162 (“[o]nce the context for the question or command has been determined, the parser 118 can invoke the correct agent”), 165 (“[b]ased on the context, the parser can evoke the correct agent to process the question or command”). The invoked agent works with the parser to interpret the utterance based on the

context. For example, if the user asks “what about flight one hundred and twenty too?,” the parser may determine that the context is “travel planning” and invoke a travel planning domain agent specific to this context, where “[t]he parser and domain agent use flight information in the database and network information along with context to determine the most plausible interpretation among; flight 100 and flight 20 also, flight 100 and flight 22, flight 122, etc.” *Id.*, ¶163.

148. Thus, because Kennewick describes its computer system determining an interpretation of the utterance by invoking an agent based on context, a POSITA would have understood that Kennewick discloses and renders obvious ***“determining, by the computer system, based on the first context, an interpretation of the first natural language utterance.”***

[1.i] “generating, by the computer system, based on the interpretation of the first natural language utterance, a first response to the first natural language utterance.”

149. Kennewick discloses limitation [1.i]. Kennewick discloses determining an interpretation of the first natural language utterance as I discuss for limitation [1.h]. Kennewick also discloses and renders obvious the “computer system” ***“generating..., based on the interpretation of the first natural language utterance, a first response to the first natural language utterance.”*** See, e.g., EX1004, ¶14 (“the system...receives spoken natural language queries, commands

and/or other utterances from a user,...processes the input..., and responds to the user with a natural language speech response”).

150. As I explain for limitation [1.h], Kennewick’s computer system invokes an agent based on the context of the user’s first natural language utterance. *Id.*, ¶162 (“[o]nce the context for the question or command has been determined, the parser 118 can invoke the correct agent”). The agent in turn generates an appropriate response to the user’s first natural language utterance for output in a suitable presentation format. *Id.*, ¶17 (“[a]gents are executables that receive, process and respond to user questions, queries and commands”), 124 (“[t]he agents 106 may return results of questions as responses to users”), 32 (“[t]he agent may make one or more queries and may rapidly return a formatted response”), 178 (“[o]nce the domain agent 156 has created a satisfactory response to a question, or to a command, the agent may format that response for presentation”).

151. And “[t]ypically, the domain agent can format the response into the markup format used by the text to speech engine 124.” *Id.*, ¶¶178, 124 (“[t]he agent may create a response string, which may be sent to the text to speech engine 124.”). The “text to speech engine creates the required utterances,” which after further processing are “output by the speaker 136.” *Id.*, ¶¶124, 23-25, 32-33, 178-82 (describing, among other things, templates, transformations, other display formats that can influence presentation of the output).

152. Accordingly, given Kennewick's description of system components that generate responses to user utterances, a POSITA would have understood Kennewick to disclose and render obvious the "*computer system*" "*generating..., based on the interpretation of the first natural language utterance, a first response to the first natural language utterance.*" A POSITA would have understood and found obvious that the response is based on the interpretation of the first natural language utterance because the response addresses the interpreted utterance. *See, e.g., id.*, ¶124 ("[t]he agents 106 may return results of questions as responses to users"). A POSITA likewise would have understood and found obvious that the response is a first response to the first utterance because it may be the only the response or a first of multiple responses to the first utterance.

2. **Claim 6: "The method of claim 1, wherein determining the interpretation of the first natural language utterance comprises determining, based on the first context, an interpretation of one or more recognized words of the first natural language utterance."**

153. Kennewick discloses claim 6. Kennewick discloses the computer system "*determining the interpretation of the first natural language utterance*" "*based on the first context*" as I discuss for limitation [1.h]. Kennewick further discloses that the interpretation of the utterance "*comprises determining, based on the first context, an interpretation of one or more recognized words of the first natural language utterance.*"

154. Kennewick discloses that the parser receives and interprets “***one or more recognized words of the first natural language utterance.***” *Id.*, ¶160 (“[o]nce the words and phrases have been recognized by the speech recognition engine 120, the tokens and user identification may be passed to the parser 118”). Therefore, a POSITA would have understood that when the parser “***determin[es] the interpretation of the first natural language utterance***” “***based on the first context***” as I discuss for limitation [1.h], it is “determining...an interpretation of one or more recognized words of the first natural language utterance.”

3. Claim 9

[9.pre] “The method of claim 1, the method further comprising:”

155. Kennewick-Oviatt-Fujimoto teaches the method of claim 1.

[9.a] “accumulating, by the computer system, long-term knowledge, wherein the long-term knowledge is accumulated based on one or more natural language utterances received during conversations occurring prior to the first conversation, wherein the first context is determined based on the short-term knowledge and the long-term knowledge.”

156. Kennewick discloses the “***computer system***” “***accumulating...long-term knowledge.***” The ’249 patent describes long-term knowledge as “generally...user-centric, rather than session-based, where inputs may be accumulated over time to build user, environmental, cognitive, historical, or other long-term knowledge models.” EX1001, 5:26-30. In addition, “[l]ong-term shared knowledge may include explicit and/or implicit user preferences, a history of recent contexts, requests, tasks, etc., user-specific jargon related to vocabularies and/or

capabilities of a context, most often used word choices, or other information.” *Id.*, 5:32-36.

157. Kennewick’s computer system “*accumulat[es]*” information from past or historical conversations in a user profile, by “creat[ing], stor[ing] and us[ing] extensive personal profile information for each user.” EX1004, ¶31. The user profile includes, among other things, a “history of questions asked,” “session histories,” a “history of commands and queries,” and a “history of [individual] interactions with the system.” *Id.*, ¶143. The historical information in the profile is unconstrained by the current session. *See, e.g., id.*, ¶31 (“Information in the profile may be added and updated automatically as the user uses the system or can be manually added or updated by the user.”). A POSITA would have understood, therefore, that information within the user profile, which is user-centric and unconstrained from the current session, is “*accumulat[ed]...long-term knowledge.*”

158. A POSITA would have understood and found obvious that the “*the long-term knowledge is accumulated based on one or more natural language utterances received during conversations occurring prior to the first conversation.*” Kennewick discloses receiving natural language utterances during a first conversation as I discuss for preamble [1.pre] and limitation [1.a]. EX1004, ¶155 (“as multiple users engage in overlapping sessions, the tags, added by the speech recognition engine 120 to each utterance, may allow other components of the

system to tie that utterance to the correct user and dialog”). Kennewick also discloses that conversations have beginnings and ends. *Id.*, ¶156 (“A dialog with the system may begin when a user first addresses it...At the end of a dialog or to interrupt a dialog, the user may speak a dismissal word (‘good by’).”). Accordingly, a POSITA would have understood and found obvious that a user may have multiple conversations with Kennewick’s computer system, including conversations before and after the first conversation, and that the computer system receives natural language utterances from the user during these conversations. Therefore, a POSITA would have understood and found obvious that the computer system “*receive[s]*” “*natural language utterances...during conversations occurring prior to the first conversation*” and that “*the long-term knowledge is accumulated based on*” these utterances, such as the “history of questions asked,” “session histories,” “history of commands and queries,” and “history of [individual] interactions with the system.” *Id.*, ¶143.

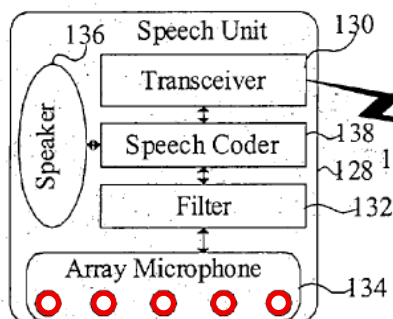
159. Kennewick also discloses that “*the first context is determined based on the short-term knowledge and the long-term knowledge.*” Kennewick discloses that “*the first context is determined based on the short-term knowledge*” as I discuss for limitation [1.g]. Kennewick further discloses that “*the first context is determined based...the long-term knowledge*” in addition to the short-term knowledge. Kennewick’s “parser may determine a context for an utterance by

applying prior probabilities or fuzzy possibilities to keyword matching, user profile 110, dialog history, and context stack contents.” *Id.*, ¶160. The dialog history is short-term knowledge as I discuss for preamble [1.pre] and the user profile includes long-term knowledge as I discuss above. Therefore, Kennewick discloses that “*the first context is determined based on the short-term knowledge and the long-term knowledge.*”

4. **Claim 12: “The method of claim 1, wherein the second input device is included with the first input device in a single device.”**

160. Kennewick-Oviatt-Fujimoto teaches the method of claim 1.

161. Kennewick discloses the “*first input device*” and “*second input device*” as I discuss for limitations [1.a] and [1.b]. Kennewick’s computer system has a “speech unit 128 [that] includes one or more microphones, for example array microphone 134, to receive the utterances of the user.” *Id.*, ¶121. The array microphone includes **multiple microphones**.



EX1004, Figure 5 (excerpt); see EX1011, 530 (“[m]icrophone arrays use multiple microphones”). First and second microphones of the array are the first and second input devices.

162. Kennewick also discloses that “*the second input device is included with the first input device in a single device.*” The speech unit in which the first and second input devices are included is a single device. EX1004, ¶12 (“The speech unit may be incorporated into a vehicle computer device or system, or may be a separate device.”). For example, the “speech unit 128...can be part of a handheld device 36 or a fixed home or office computer system 44.” *Id.*, ¶96; see also *id.*, ¶¶107 (the “speech unit 128 can be embedded in one or more of the TCUs 28, handheld computers 36 and fixed computer systems 44”), 111 (“the...speech unit 128 may be embedded into the handheld computer 36”), 115 (“the...speech unit 128 may be embedded into the fixed computer 44”). A POSITA would have understood each of these examples to demonstrate that the speech unit is a single device, e.g., a vehicle computer device, handheld device, or fixed computer system, in which the first and second input devices are included.

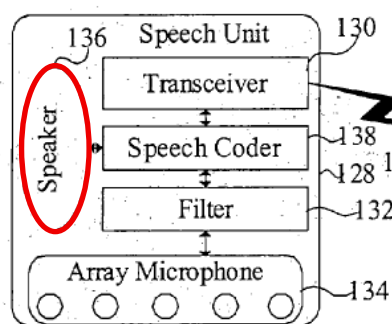
163. Moreover, a POSITA would have found it obvious to include the second input with the first input device in a single device because Kennewick discloses that the speech unit can have different configurations. *Id.*, ¶109 (“[i]t will be understood that the exact distribution and packaging of the...speech unit 128, can

be determined by the details of the deployment situation and will not change the functionality, capabilities or spirit of the invention in any way”).

5. Claim 13: “The method of claim 1, wherein generating the first response comprises generating, by the computer system, a non-voice response in a graphical user interface and/or a voice response via a speaker.”

164. Kennewick-Oviatt-Fujimoto teaches the method of claim 1.

165. Kennewick discloses that “*generating the first response comprises generating, by the computer system...a voice response via a speaker.*” Kennewick discloses the computer system generating the first response as I discuss for limitation [1.i]. The first response may be “*a voice response.*” See, e.g., EX1004, ¶¶14 (“the system...receives spoken natural language queries, commands and/or other utterances from a user,...processes the input..., and responds to the user with a natural language speech response”), 124 (“[t]he agent may create a response string, which may be sent to the text to speech engine 124.”). The voice response is “output by the **speaker 136**” of the speech unit. *Id.*, ¶124.

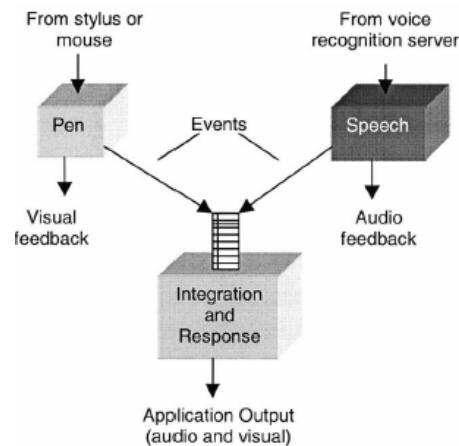


EX1004, Figure 5 (excerpt); *see also id.*, ¶121 (“[c]oded speech received from the main unit is detected by the transceiver 130, then decoded and decompressed by the speech coder 138 and annunciated by the speaker 136”). A POSITA would have understood that Kennewick’s “speech response” is a “*voice response*,” and the generated voice response is “*via a speaker*” because it is output by the speaker.

166. Kennewick also discloses that “*generating the first response comprises generating, by the computer system, a non-voice response in a graphical user interface.*” Kennewick discloses that the computer system has “*a graphical user interface*” as I discuss for limitation [1.e]. “The graphical user interface 114 can be used as a substitute or complement to the speech interface. For example, the graphical user interface can be used to view and interact with graphical or tabular information in a manner more easily digested by the user.” EX1004, ¶125. A POSITA would have found it obvious for Kennewick’s computer system to generate the first response as “*a non-voice response in a graphical user interface*” because “[s]ome information, in formats such as video, pictures and graphics, may be best presented in a displayed format.” EX1004, ¶182. Kennewick discloses how this is done: “[t]he domain agents 156 apply suitable presentation templates in these cases and present the information through the graphical user interface 114.” *Id.*

167. In addition, the computer system “*generating a non-voice response in a graphical user interface*” is obvious in view of Kennewick and Oviatt. Oviatt

discloses that the output of multimodal systems “typically is presented via a graphical, auditory (e.g., for telephony), or multimedia display.” EX1005, 15. Oviatt illustrates the multimodal PVA I discuss for limitation [1.e] producing a graphical output:



Id., 40 (Figure 14). Here, a POSITA would have understood Oviatt’s text saying “Application Output (audio and visual)” to mean the PVA may generate a non-voice response presented in a graphical user interface shown on a display. A POSITA would have been motivated to combine Kennewick and Oviatt such that Kennewick’s computer system generates and presents a non-voice output as taught by Oviatt because “[s]ome information, in formats such as video, pictures and graphics, may be best presented in a displayed format” and for the other reasons I provide for preamble [1.pre] and limitation [1.e]. EX1004, ¶182.

6. Claim 14

[14.pre] “The method of claim 1, wherein the first natural language utterance relates to a request and the first response is intended to be responsive to the request, the method further comprising:”

168. Kennewick-Oviatt-Fujimoto teaches the method of claim 1.

169. Kennewick further teaches wherein **“the first natural language utterance relates to a request.”** Kennewick discloses receiving “the first natural language utterance” as I discuss for limitation [1.a]. EX1004, ¶121 (“[t]he speech unit 128 includes one or more microphones, for example array microphone 134, to receive the utterances of the user”). Kennewick also discloses that this utterance **“relates to a request.”** Kennewick teaches that “[t]he system...can process and respond to questions, queries and commands.” *Id.*, ¶27. At POSITA would have understood that a command is a request because a command requests that the computer system perform an action. EX1004, ¶¶123 (“[c]ommands typically result in actions taken by the system itself”), 166 (“[c]ommands can be directed to the system or to an external entity”). For example, a user may request that the computer system tune a radio by uttering the command “tune in my favorite radio station.” *Id.*, ¶27.

170. Kennewick also discloses **“wherein...the first response is intended to be responsive to the request.”** Kennewick discloses the computer system generating the first response as I discuss for limitation [1.i]. The first response is **“intended”**

by the computer system to be responsive to request because the computer system makes the response in reply to the request. For example, if the computer system does not understand the user's request, "the system can ask a question of the user to verify the question or command is correctly understood," such as "[d]id I understand that you want such-and-such." *Id.*, ¶161; *see also id.*, ¶¶32 ("[t]he user can...be asked one or more questions to resolve the ambiguity"), 33 ("the system can ask a question of the user to verify the question or command is correctly understood"), 177 ("[i]f the command is ambiguous or cannot be executed for some other reason, the system may ask the user for more information or may suggest what the problem is and a likely approach to the solution").

[14.a] "receiving, by the computer system, a second natural language utterance in response to the first response, wherein the second natural language utterance comprises a clarification related to the interpretation of the first natural language utterance; and"

171. Kennewick discloses "*the computer system*" "*receiving...a second natural language utterance in response to the first response, wherein the second natural language utterance comprises a clarification related to the interpretation of the first natural language utterance.*" Kennewick discloses that the computer system may provide a first response that asks the user for "*a clarification related to the interpretation of the first natural language utterance.*" For example, when the user's command in the first natural language utterance is ambiguous, the computer system can ask the user "one or more questions to resolve the ambiguity." *Id.*, ¶¶32,

33 (“the system can ask a question of the user to verify the question or command is correctly understood”), 161 (“the system can ask one or more questions to attempt to resolve the ambiguity”), 177 (“[i]f the command is ambiguous or cannot be executed for some other reason, the system may ask the user for more information”). A POSITA would have understood and found obvious from these examples that the user provides, and the computer system receives, “*a second natural language utterance in response*” to the question from the system, and the “*utterance comprises a clarification related to the interpretation of the first natural language utterance.*” See, e.g., *id.*, ¶¶33 (“the user can rephrase the original question, perhaps adding additional information to remove ambiguity”), 161 (same).

[14.b] “responsive to receipt of the second natural language utterance, generating, by the computer system, a response to the second natural language utterance based on the interpretation of the first natural language utterance and the clarification related to the interpretation of the first natural language utterance.”

172. Kennewick discloses and renders obvious the “*computer system*” “*generating...a response to the second natural language utterance based on the interpretation of the first natural language utterance and the clarification related to the interpretation of the first natural language utterance*” “*responsive to receipt of the second natural language utterance.*” The second utterance from the user may clarify a command the computer system misunderstands or finds ambiguous as I discuss for limitation [14.a]. A POSITA would have understood and found obvious that the computer system responds to receipt of the clarification by generating a

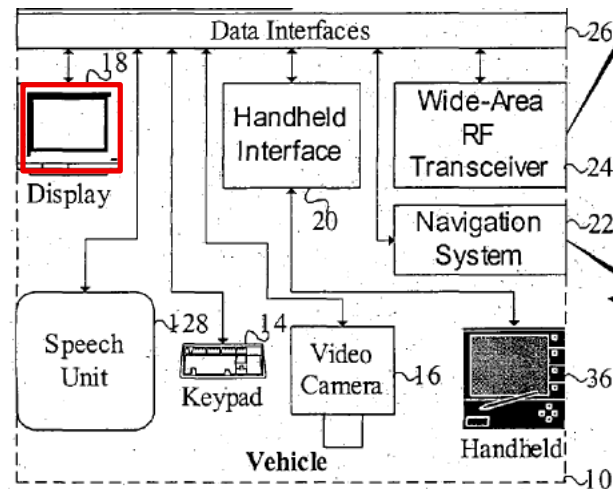
response to the command as I discuss for preamble [14.pre] and limitation [1.i], and that the response is responsive to both the interpretation of the first utterance (as I discuss for limitation [1.h]) and the user's clarification in the second utterance. *See, e.g., id.*, ¶¶32 (“The user can then be asked one or more questions to resolve the ambiguity. Additional queries may then be made before an adequate response is made.”), 33 (“If the user confirms that the question is correct then the system may proceed to produce a response.”), 161 (same).

7. Claim 15

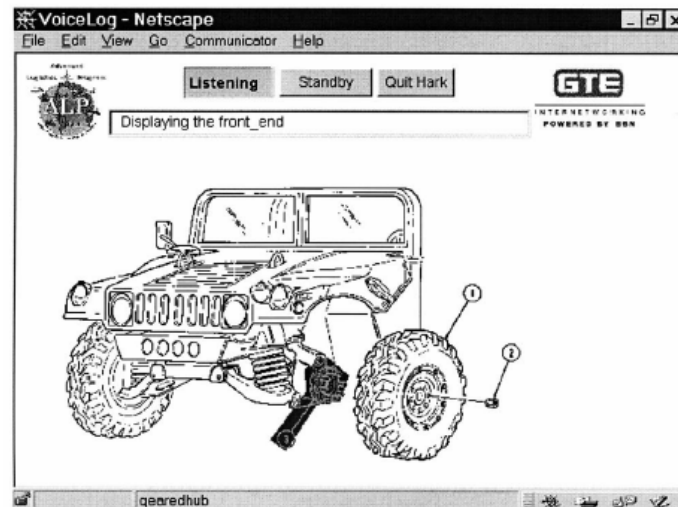
[15.pre] “The method of claim 1, wherein the first non-voice input is received via a non-voice input device comprising a touch screen device, the method further comprising:”

173. Kennewick-Oviatt-Fujimoto teaches the method of claim 1.

174. Kennewick and Oviatt disclose “*wherein the first non-voice input is received via a non-voice input device comprising a touch screen device*” for the reasons I provide for limitation [1.e]. Kennewick’s computer system has a touch screen, which a POSITA would have understood is a “*non-voice input device comprising a touch screen device.*” *Id.*, ¶97 (“[o]ther user interface peripherals...can include, **displays 18** including touch screen displays for text, graphics and video”).



EX1004, Figure 1 (excerpt), ¶¶112 (“users interact with the handheld computer 36 through the speech unit 128, the keypad 74 or keyboard, and a display 72 for text, graphics and video. In some embodiments the display is of a touch screen type. An optional pointing device (not shown) may be used as well”), 116 (same). Oviatt’s PVA also has a touch screen (“*non-voice input device comprising a touch screen device.*”).



EX1005, 39 (Figure 13).

175. Oviatt discloses that “*the first non-voice input is received via a non-voice input device comprising a touch screen device.*” Oviatt’s PVA “can interpret simultaneous pen-voice input so, for example, the user can say ‘Show me that part’ while pointing at a screw on the display.” *Id.* The user “pointing at a screw on the display” is the first non-voice input as I discuss for limitation [1.e]. A POSITA would have understood that the non-voice input is received by the non-voice input device comprising a touch screen device, i.e., the touch screen display of Kennewick’s computer system in the combination of Kennewick and Oviatt I describe for [1.pre] and limitation [1.e].

[15.a] “receiving, by the computer system, information indicating a location on the touch screen device that received the first non-voice input.”

176. Kennewick and Oviatt disclose the “*computer system*” “*receiving...information indicating a location on the touch screen device that received the first non-voice input.*” Kennewick and Oviatt disclose the computer system receiving the first non-voice input via a non-voice input device comprising a touch screen device as I discuss for preambles [1.pre] and [14.pre] and limitation [1.e].

177. Oviatt discloses and renders obvious “*receiving...information indicating a location on the touch screen device that received the first non-voice input.*” Oviatt’s PVA “can interpret simultaneous pen-voice input so, for example, the user can say ‘Show me that part’ while pointing at a screw on the display.” *Id.*,

39. A POSITA would have understood and found obvious that the PVA interprets the pen input by “*receiving...information indicating a location on the touch screen device*” to which the user is pointing. *See, e.g., id.*, 8 (“pen input to maps and other graphic displays can easily and efficiently convey spatial information about precise points, lines, and areas”), 10 (“pen input conveys rich and precise graphical information”). The touch screen device for which the location is received is same device “that received the first non-voice input.” Thus, the input is “**receiv[ed]...by the computer system...that received the first non-voice input.**”

8. Claim 16

[16.pre] “A system for facilitating natural language system responses via short-term knowledge generated based on one or more prior multi-modal device interactions, the system comprising:”

[16.a] “one or more physical processors programmed with one or more computer program instructions which, when executed, cause the one or more physical processors to:”

178. Claim 16 is unpatentable for the same reasons as claim 1. Preamble [16.pre] and limitation [16.a] are substantively identical to preamble [1.pre]. Preamble [16.pre] and limitation [16.a] recite “[a] system...comprising...one or more physical processors programmed with one or more computer program instructions which, when executed, cause the one or more physical processors to” perform recited operations. Preamble [1.pre] similarly recites “[a] computer-implemented method...implemented by a computer system that includes one or more physical processors executing one or more computer program instructions

which, when executed, perform the method” comprising recited steps. The “computer system” implementing the method of claim 1 is equivalent to the “system” of claim 16 and obvious for the same reasons.

179. The remaining limitations of claim 16 are substantively identical to the corresponding limitations of claim 1 and obvious for the same reasons.

9. Claims 21 and 25-28

180. Each of dependent claims 21 and 25-28 is unpatentable for the same reasons as the respective corresponding claim indicated in the following table:

Claim Number	Corresponding Claim Number
21	6
25	12
26	13
27	14
28	15

181. Each of claims 21 and 25-28 is substantively identical to the corresponding claim, and contains only minor differences to reflect that independent claim 16 and any intervening claims recite a system.

B. Ground 2: Kennewick-Oviatt-Fujimoto in combination with Shimomura renders obvious claims 2-5 and 17-20

1. Claim 2

[2.pre] “The method of claim 1, the method further comprising:”

182. Kennewick-Oviatt-Fujimoto discloses the method of claim 1.

[2.a] “receiving, at the computer system, a second natural language utterance via the first input device;”

183. Kennewick discloses the computer system receiving a first natural language utterance via a first input device as I discuss for limitation [1.a]. EX1004, ¶¶14, 38, 121, 149. Kennewick alone and in combination with Shimomura discloses the “*computer system*” “*receiving...a second natural language utterance via the first input device.*”

184. Kennewick discloses “*the computer system*” “*receiving...a second natural language utterance*” as I discuss for limitation [14.a]. *See, e.g., id.*, ¶¶32, 33, 161, 177. A POSITA would have understood that Kennewick’s computer system “*receiv[es]...[the] second natural language utterance via the first input device.*”

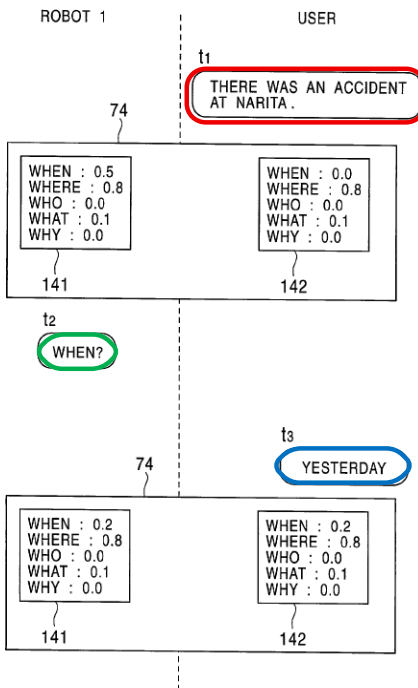
As I discuss for limitation [1.a], Kennewick’s computer includes a speech unit with an array microphone that receives user utterances. EX1004, ¶121 (“[t]he speech unit 128 includes one or more microphones, for example array microphone 134, to receive the utterances of the user”). A first microphone of the array is “*the first input device,*” and a POSITA would have understood and found obvious that the microphone receives multiple user utterances, including the second utterance, during

a dialog. For example, the first microphone of the array would receive multiple utterances of a dialog between the user and the computer system while the user is in a vehicle. *Id.*, ¶73 (“[v]ehicle operators and other occupants can use the interactive natural language interactive speech interface of the invention to perform many types of information query, retrieval and presentation operations”). Accordingly, Kennewick discloses the “*computer system*” “*receiving...a second natural language utterance via the first input device.*”

185. Kennewick in combination with Shimomura also discloses the “*computer system*” “*receiving...a second natural language utterance via the first input device.*” Shimomura discloses a robot with “a conversation processing apparatus for holding a conversation with a user.” EX1014, ¶¶9, 17. The robot includes a microphone that “collects ambient speech (sounds) including the speech of a user.” *Id.*, ¶56. A POSITA would have understood from this disclosure that Shimomura’s robot “*receiv[es]*” utterances from the user via the microphone, which is a “*first input device.*”

186. Shimomura discloses that a conversation between the user and robot includes multiple utterances. For example, Shimomura describes the user initiating a conversation with the robot by saying “[t]here was an accident at Narita.” The robot responds with “when?” and the user answers “[y]esterday.” *Id.*, ¶¶122-25.

FIG. 12



Id., Fig. 12. A POSITA would have understood that this conversation includes first and second natural language utterances from the user: 1) “**there was an accident at Narita**”; and 2) “**yesterday**.” A POSITA also would have found it obvious that these utterances are captured by the robot’s microphone (a first input device) since the microphone “collects ambient speech (sounds) including the speech of a user.” *Id.*, ¶56.

187. A POSITA would have found it obvious to modify Kennewick in view of Shimomura to support dialogs in which the “*computer system*” “*receiv[es]...a second natural language utterance via the first input device.*” Both Kennewick and Shimomura relate to natural language interactions with users. *See* EX1004, ¶2 (“the invention is a fully integrated environment allowing mobile users to ask natural

language speech questions or give natural language commands in a wide range of domains”); EX1014, ¶2 (“The present invention relates to conversation processing apparatuses and methods”). Doing so would support Kennewick’s goal of “allow[ing] a person to make natural language speech-based queries,” especially given that “that most natural language queries and commands are incomplete in their definition” and may require follow-up utterances. EX1004, ¶¶5, 7-8. This is a combination of prior art elements according to known methods to yield predictable results, and a POSITA would have reasonably expected the combination to be successful because Kennewick discloses how the computer system processes utterances.

[2.b] “determining, by the computer system, based on the short-term knowledge, whether the second natural language utterance corresponds to the first context;”

188. Kennewick discloses the computer system determining a first context based on the short-term knowledge as I discuss for limitation [1.g]. Kennewick alone and in combination with Shimomura disclose receiving the second natural language utterance as I discuss for limitation [2.a]. Kennewick in combination with Shimomura also discloses the “*computer system*” “*determining...based on the short-term knowledge, whether the second natural language utterance corresponds to the first context.*”

189. Kennewick’s computer system determines a context for each utterance it receives. EX1004, ¶¶28 (“[f]or each user utterance including a question or query

or set of questions or queries, the system may perform multiple steps possibly including...determining the...context”), 41 (“[f]or each user command utterance, the system may execute a number of steps possibly including...determin[ing] the...context”), 160 (“[t]he parser examines the tokens for the questions or commands, context and criteria. The parser may determine a context for an utterance[.]”). The computer system determines the context for an utterance “*based on short-term knowledge*” as I discuss for limitation [1.g]. Accordingly, a POSITA would have understood and found obvious that Kennewick’s computer system determines the context for the second natural language utterance “*based on the short-term knowledge.*” See EX1004, ¶160.

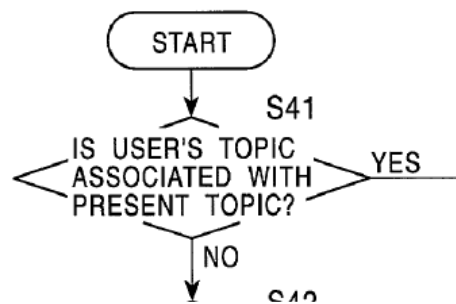
190. The computer system also determines the context for the second natural language utterance based in part on the first context for the first natural language utterance. The parser “use[s] a scoring system to determine the most[] likely context or domain for a user’s question or command.” *Id.*, ¶161. “The score can be determined from weighting a number of factors including, the...previous context.” *Id.*; see also *id.*, ¶¶32 (“[t]he invention may determine the most[] likely context or domain for a user’s question or command, for example, by using a real-time scoring system or other techniques”), 160 (“[b]ased on these probabilities or possibilities the possible contexts are scored and the top one or few are used for further processing”).

191. A POSITA would have understood and found obvious that the “computer system” “*determin[es] based on the short-term knowledge, whether the second natural language utterance corresponds to the first context*” by using the “scoring system to determine the most[] likely context or domain for a user’s question or command.” *Id.*, ¶161. A POSITA would have understood that the context for the second utterance may be different than the context of the first utterance. *See id.*, ¶¶167, 177 (“real-time selection of the correct agent allows the invention to dynamically switch contexts”). A POSITA also would have known that “some questions can only be interpreted in the context of previous questions.” *Id.*, ¶7. Therefore, a POSITA would have found it obvious for the scoring system to determine whether the context for the second utterance changed from the context of the first utterance, and would have understood that this determination involves determining “*whether the second natural language utterance corresponds to the first context.*”

192. To the extent Kennewick does not explicitly disclose “*determining...whether the second natural language utterance corresponds to the first context,*” a POSITA would have found it obvious based on Shimomura. Shimomura’s robot talks to the user about a topic. The robot detects when a user’s utterance changes the topic. EX1014, ¶¶184 (“a case is possible in which the user

changes the topic”), 195 (“[i]n actual conversations, there are cases in which topics are changed...by the user”).

FIG. 25



Id., Fig. 25 (excerpt).

193. The robot detects topic changes by comparing the topic introduced by the user with the current topic. *Id.*, ¶184 (“[i]n step S41, the topic manager 74 of the robot 1 determines whether the topic introduced by the user is associated with the present topic stored in the present topic memory 77”). It performs the comparison using a scoring system. *Id.*, ¶¶142 (“the coincidence between the speech by the robot 1 and the speech by the user is measured to obtain a score”), 143 (“the topic is changed if the score thus obtained exceeds a predetermined threshold”), 184 (determining whether the topic introduced by the user is associated with the present topic is performed “using a method similar to that for computing the degree of association between topics (keywords) when the topic is changed by the robot”), 185 (“the degree of association is computed between a group of keywords extracted from a single oral statement made by the user and the keywords of the present topic...[t]he

process determines whether the topic introduced by the user is associated with the present topic based on the computation result”). Accordingly, a POSITA would have understood and found obvious that Shimomura discloses determining whether a second natural language utterance corresponds to a topic of a first natural language utterance.

194. A POSITA would have found it obvious in view of Shimomura for Kennewick’s “*computer system*” to “*determin[e]...whether the second natural language utterance corresponds to the first context*” when determining the context for the second utterance. Kennewick’s parser “use[s] a scoring system to determine the most[] likely context or domain for a user’s question or command” where “[t]he score can be determined from...the...previous context.” EX1004, ¶161. Shimomura discloses an analogous scoring system for detecting topic changes that determines whether a second utterance corresponds to a topic of a first utterance. EX1014, ¶185 (“the degree of association is computed between a group of keywords extracted from a single oral statement made by the user and the keywords of the present topic...[t]he process determines whether the topic introduced by the user is associated with the present topic based on the computation result”). Therefore, a POSITA would have found it obvious for Kennewick’s computer system to use a process for contexts analogous to the process for topics disclosed in Shimomura to

detect context changes by “*determining...whether the second natural language utterance corresponds to the first context.*”

195. A POSITA would have been motivated to make this combination because “a person asking a question or giving a command typically relies heavily on context” and “some questions can only be interpreted in the context of previous questions” EX1004, ¶¶5, 7. Accordingly, a POSITA would have been motivated to use Shimomura’s technique within Kennewick’s computer system in order “to hold natural conversations” by determining whether the second utterance corresponds to the (first) context of the first utterance. EX1014, ¶201. This is a combination of prior art elements according to known methods to yield predictable results and use of a known technique to improve a similar device in the same way. A POSITA would have had a reasonable expectation of success because the technology at issue is conventional and well understood.

[2.c] “determining, by the computer system, responsive to the second natural language utterance not corresponding to the first context, a second context for the second natural language utterance based on the short-term knowledge;”

196. Kennewick and Shimomura disclose the computer system determining, based on the short-term knowledge, whether the second natural language utterance corresponds to the first context as I discuss for limitation [2.b]. Kennewick additionally discloses and renders obvious the “*computer system*” “*determining...responsive to the second natural language utterance not*

corresponding to the first context, a second context for the second natural language utterance based on the short-term knowledge.”

197. Kennewick discloses the computer system determining a context for each received utterance based on the short-term knowledge as I discuss for limitations [1.g] and [2.b]. EX1004, ¶¶28 (“[f]or each user utterance including a question or query or set of questions or queries, the system may perform multiple steps possibly including...determin[ing] the...context”), 41 (“[f]or each user command utterance, the system may execute a number of steps possibly including...determining the...context”), 160 (“[t]he parser examines the tokens for the questions or commands, context and criteria. The parser may determine a context for an utterance[.]”), 162 (“[p]rior probabilities or fuzzy possibilities and associated values may be received from a number of sources including, for example, the history of the dialog”).

198. A POSITA would have understood and found obvious that the computer system determines a different context for the second natural language utterance if it does not correspond to the first context. The computer system uses the “scoring system to determine the most[] likely context or domain for a user’s question or command,” with the outcome that the context is different from the first context. *Id.*, ¶161; *see id.*, ¶¶167, 177 (“real-time selection of the correct agent allows the invention to dynamically switch contexts”). Hence, Kennewick discloses

the “*computer system*” “*determining...responsive to the second natural language utterance not corresponding to the first context, a second context for the second natural language utterance based on the short-term knowledge.*”

[2.d] “determining, by the computer system, based on the second context, an interpretation of the second natural language utterance; and”

199. Kennewick’s computer system determines an interpretation of a natural language utterance based on the determined context as I discuss for limitation [1.h]. See, e.g., *id.*, ¶¶160 (“[t]he context of a question or command may determine the domain and thereby, the domain agent 156, if any, to be invoked”), 162 (“[o]nce the context for the question or command has been determined, the parser 118 can invoke the correct agent”), 165 (“[b]ased on the context, the parser can evoke the correct agent to process the question or command”). For limitation [2.d], a POSITA would have understood that the “*computer system*” “*determin[es], based on the second context, an interpretation of the second natural language utterance*” in the same manner it interprets the first utterance based on the first context.

[2.e] “generating, by the computer system, based on the interpretation of the second natural language utterance, a response to the second natural language utterance.”

200. Kennewick’s computer system generates a response to a natural language utterance based on the interpretation of the utterance as I discuss for limitation [1.i]. See, e.g., EX1004, ¶¶14 (“the system...receives spoken natural language queries, commands and/or other utterances from a user,...processes the

input... and responds to the user with a natural language speech response”), 17 (“[a]gents are executables that receive, process and respond to user questions, queries and commands”), 124 (“[t]he agents 106 may return results of questions as responses to users”), 32 (“[t]he agent may make one or more queries and may rapidly return a formatted response”), 178 (“[o]nce the domain agent 156 has created a satisfactory response to a question, or to a command, the agent may format that response for presentation”). For limitation [2.e], a POSITA would have understood that the “*computer system*” “*generat[es]...based on the interpretation of the second natural language utterance, a response to the second natural language utterance*” in the same manner it generates the first response to the first natural language utterance based on the interpretation as I discuss for limitation [1.i].

2. Claim 3

[3.pre] “The method of claim 1, the method further comprising:”

201. Kennewick-Oviatt-Fujimoto discloses the method of claim 1.

[3.a] “tracking, by the computer system, contexts identified for multiple consecutive natural language user utterances to generate a context stack, the context stack including the contexts identified for the multiple consecutive natural language user utterances;”

202. Kennewick in combination with Shimomura discloses limitation [3.a].

203. Kennewick alone and in combination with Shimomura discloses “*the computer system*” receiving “*multiple consecutive natural language user utterances.*” As I discuss for limitations [14.a] and [2.a], Kennewick discloses

examples of dialogs between the computer system and user having multiple natural language utterances. *See, e.g., id.*, ¶¶32, 33, 161, 177. A POSITA would have found it obvious for these dialogs to include “***multiple consecutive natural language user utterances.***” For example, Kennewick describes the user rephrasing a question the computer system finds ambiguous, and a POSITA would have understood that the initial question is a first natural language user utterance and the rephrased question is a second, consecutive natural language user utterance. *See id.*, ¶33 (“either the user can rephrase the original question, perhaps adding additional information to remove ambiguity, or the system can ask one or more questions to attempt to resolve the ambiguity or other actions may [be] taken”).

204. Shimomura explicitly discloses conversations including “***multiple consecutive natural language user utterances.***” For example, Shimomura describes the user initiating a conversation with the robot by saying “[t]here was an accident at Narita.” The robot responds with “when?” and the user answers “[y]esterday.” EX1014, ¶¶122-25. Here, the user’s utterances “[t]here was an accident at Narita” and “[y]esterday” are “***multiple consecutive natural language user utterances.***”

205. Kennewick discloses “***identify[ing] contexts...for...multiple ...natural language user utterances***” as I discuss for limitations [1.g], [2.b], and [2.c]. The computer system identifies a context for each utterance it receives.

EX1004, ¶¶28 (“[f]or each user utterance including a question or query or set of questions or queries, the system may perform multiple steps possibly including...determining the...context”), 41 (“[f]or each user command utterance, the system may execute a number of steps possibly including...determin[ing] the...context”), 160 (“[t]he parser examines the tokens for the questions or commands, context and criteria. The parser may determine a context for an utterance[.]”). A POSITA would have found it obvious from Kennewick’s disclosure to “*identify contexts...for...multiple consecutive natural language user utterances*” since this would beneficially allow the user to converse with the computer system by detecting potential context changes. See EX1004, ¶¶160 (“[t]he context of a question or command may determine the domain and thereby, the domain agent 156, if any, to be invoked”), 162 (“[o]nce the context for the question or command has been determined, the parser 118 can invoke the correct agent”), 165 (“[b]ased on the context, the parser can evoke the correct agent to process the question or command”), 167 (“[t]he real-time selection of the correct agent allows the invention to dynamically switch contexts”).

206. Kennewick also discloses that the computer system uses “*a context stack*.” *Id.*, ¶170 (“part of the dialog history is maintained in a context stack”); see also *id.*, ¶160 (“[t]he parser may determine a context for an utterance by applying prior probabilities or fuzzy possibilities to keyword matching, user profile 110,

dialog history, and context stack contents”). A stack is a conventional data structure for tracking and storing information. *See, e.g.,* EX1011, 642 (describing storing data “in a last-in first-out (LIFO) stack”). A POSITA therefore would have understood and found it obvious that Kennewick’s “*computer system*” “*generate[s] a context stack...including the contexts identified for the multiple consecutive natural language user utterances*” and uses it to “*track[]...contexts identified for multiple consecutive natural language user utterances.*”

207. To the extent not disclosed in Kennewick, Shimomura discloses a data structure storing and tracking information about a conversation:

The conversation history memory 75 accumulates the content of conversation or information extracted from conversation. The conversation history memory 75 also stores data used to examine topics which were brought up prior to the present topic, which is stored in the present topic memory 77, and to control the change of topic.

EX1014, ¶83. For example, the conversation history memory stores information about previous conversation topics. *Id.*, ¶176 (“the determination [of whether a topic has been brought up recently] is performed by examining information in the conversation history memory”). Given this disclosure, a POSITA would have understood that Shimomura’s robot updates the topics stored in the conversation history memory when it changes topics based on user utterances. *Id.*, ¶82 (“[i]n

preparation for the subsequent change of topic...the topic manager 74 appropriately updates information under management of a conversation history memory 75”).

208. A POSITA would have found it obvious in view of Kennewick and Shimomura for Kennewick’s “*computer system*” to “*track[]...contexts identified for multiple consecutive natural language user utterances to generate a context stack...including the contexts identified for the multiple consecutive natural language user utterances.*” Kennewick explicitly discloses a context stack and may identify the context for each utterance, including for multiple consecutive natural language user utterances. *See, e.g.*, EX1004, ¶¶160, 170. Shimomura discloses storing analogous information about a conversation in a data structure for tracking purposes. *See, e.g.*, EX1014, ¶83 (“[t]he conversation history memory 75 also stores data used to examine topics which were brought up prior to the present topic”). A POSITA would have found it obvious from these disclosures to generate a context stack and store contexts in it for tracking purposes. A POSITA would have been motivated to use the context stack to provide an “environment allowing mobile users to ask natural language speech questions or give natural language commands in a wide range of domains” because “[t]he context of a question or command may determine the domain and thereby, the domain agent 156, if any, to be invoked” and the possible contexts are scored using information in the context stack. EX1004,

¶¶2, 160. This is a combination of prior art elements according to known methods to yield predictable results.

[3.b] “receiving, at the computer system, a second natural language utterance via the first input device;”

209. Kennewick alone and in combination with Shimomura discloses “***the computer system***” “***receiving...a second natural language utterance via the first input device***” as I discuss for limitations [14.a] and [2.a].

[3.c] “determining, by the computer system, based on the short-term knowledge, whether the second natural language utterance corresponds to one or more individual ones of the contexts included in the context stack.”

210. Kennewick in combination with Shimomura discloses limitation [3.c]. As I discuss for limitation [2.b], Kennewick and Shimomura disclose the “***computer system***” determining the context for the second natural language utterance based on the short-term knowledge and determining whether the second natural language utterance corresponds to the first context for a first (prior) utterance. Kennewick and Shimomura also disclose a context stack that includes contexts for prior utterances as I discuss for limitation [3.a].

211. Given these disclosures, a POSITA would have found it obvious for the “***computer system***” to “***determin[e]..., based on the short-term knowledge, whether the second natural language utterance corresponds to one or more individual ones of the contexts included in the context stack.***” A POSITA would have found it obvious that the computer system determines whether the second natural language

utterance corresponds to the first context per limitation [2.b] by determining whether the second utterance corresponds to the first context (an individual one of the contexts), or another individual context, included in the context stack per limitation [3.a]. *See, e.g.*, EX1014, ¶¶160, 170 (“[t]he weight of each context for the scoring may be based on the relevance of one context to another and the age of the contexts”); *see also id.*, ¶¶83, 176. A POSITA would have been motivated to use the context stack this way because Kennewick’s computer system “determine[s] a context for an utterance by applying prior probabilities or fuzzy possibilities to keyword matching, user profile 110, dialog history, and context stack contents” and the context stack stores contexts of prior utterances. *Id.*, ¶160.

3. Claim 4

[4.pre] “The method of claim 3, wherein determining whether the second natural language utterance corresponds to one or more individual ones of the contexts included in the context stack includes:”

212. Kennewick in combination with Shimomura discloses the method of claim 3, including “***determining whether the second natural language utterance corresponds to one or more individual ones of the contexts included in the context stack***” as I discuss for preamble [3.pre] and limitations [3.a]-[3.c].

[4.a] “determining whether the second natural language utterance corresponds to a most recent context included in the context stack; and”

213. A POSITA would have found limitation [4.a] obvious in view of Kennewick and Shimomura. These references disclose including contexts in the

context stack, and determining whether the second natural language utterance corresponds to an individual context included in the context stack as I discuss for limitation [3.c].

214. In addition, a POSITA would have found it obvious to “***determin[e] whether the second natural language utterance corresponds to a most recent context included in the context stack***” because a stack is a LIFO (last-in first-out) data structure. EX1011, 642. A POSITA would have understood that the top of the stack includes the last-in (most recent) context. In addition, Kennewick discloses weighing the influence of a context based on the age of the context. EX1014, ¶¶160, 170 (“[t]he weight of each context for the scoring may be based on the relevance of one context to another and the age of the contexts”); *see also id.*, ¶¶83, 176. Therefore, a POSITA would have found it obvious for the computer system to determine whether the second natural language utterance corresponds to a context in the stack by determining whether the second utterance corresponds to the most recent (i.e., youngest) context included at the top of the stack.

[4.b] “responsive to the second natural language utterance not corresponding to the most recent context, determining whether the second natural language utterance corresponds to a second most recent context included in the context stack.”

215. A POSITA would have found limitation [4.b] obvious in view of Kennewick and Shimomura. Following on from limitation [4.a], a POSITA would have understood and found obvious that “***responsive to the second natural language***

utterance not corresponding to the most recent context” at the top of the stack, the computer system would “*determin[e] whether the second natural language utterance corresponds to a second most recent context included in the context stack,*” which is the next context included in the context stack because the stack is a LIFO data structure. *See id.*, ¶170 (“[t]he weight of each context for the scoring may be based on the relevance of one context to another and the age of the contexts”).

4. Claim 5

[5.pre] “The method of claim 3, further comprising:”

216. Kennewick in combination with Shimomura discloses the method of claim 3 for the reasons I provide for preamble [3.pre] and limitations [3.a]-[3.c].

[5.a] “determining, by the computer system, responsive to the second natural language utterance not corresponding to one or more individual ones of the contexts included in the context stack, a second context for the second natural language utterance based on the short-term knowledge; and”

217. Kennewick and Shimomura disclose the computer system determining, based on the short-term knowledge, whether the second natural language utterance corresponds to one or more individual ones of the contexts included in the context stack as I discuss for limitation [3.c]. *See also* limitation [2.b].

218. Kennewick additionally discloses the computer system determining, responsive to the second natural language utterance not corresponding to a first context, a second context for the second natural language utterance based on the short-term knowledge as I discuss for limitation [2.c]. A POSITA would have found

it obvious for the “*computer system*” to “*determin[e]...a second context for the second natural language utterance based on the short-term knowledge*” “*responsive to the second natural language utterance not corresponding to one or more individual ones of the contexts included in the context stack*” for these same reasons. That is, if the computer system determines that the second utterance does not correspond to contexts in the context stack, a POSITA would have found it obvious for the computer system to determine a second context for that utterance based on the short-term knowledge for the reasons I provide for limitation [2.c].

[5.b] “determining, by the computer system, based on the second context, an interpretation of the second natural language utterance.”

219. Kennewick discloses limitation [5.b]. As I explain for limitation [1.h], Kennewick discloses the computer system determining an interpretation of the first natural language utterance based on the first context. Kennewick discloses the “*computer system*” “*determining...an interpretation of the second natural language utterance*” “*based on the second context*” for the same reasons.

5. Claims 17-20

220. Each of dependent claims 17-20 is unpatentable for the same reasons as the respective corresponding claim indicated in the following table:

Claim Number	Corresponding Claim Number
17	2
18	3
19	4
20	5

221. Each of claims 17-20 is substantively identical to the corresponding claim, and contains only minor differences to reflect that independent claim 16 and any intervening claims recite a system.

C. Ground 3: Kennewick-Oviatt-Fujimoto in combination with Matsuda renders obvious claims 7, 8, and 22

1. Claim 7

[7.pre] “The method of claim 1, the method further comprising:”

222. Kennewick-Oviatt-Fujimoto teaches the method of claim 1.

[7.a] “expiring, by the computer system, one or more items of short-term knowledge that are based on one or more natural language utterances received during conversations occurring prior to the first conversation.”

223. Kennewick alone and in combination with Matsuda discloses limitation [7.a]. Kennewick discloses “*one or more items of short-term knowledge that are based on one or more natural language utterances received during*” a conversation as I discuss for preamble [1.pre] and limitation [1.a]. EX1004, ¶155 (“as multiple users engage in overlapping sessions, the tags, added by the speech recognition engine 120 to each utterance, may allow other components of the system to tie that

utterance to the correct user and dialog”). Kennewick also discloses “*receiv[ing] one or more natural language utterances...during conversations occurring prior to the first conversation*” as I discuss for limitation [9.a].

224. A POSITA would have understood and found obvious from Kennewick’s disclosure for the computer system to “*expir[e]...one or more items of short-term knowledge that are based on one or more natural language utterances received during conversations occurring prior to the first conversation.*” As I discuss for limitation [1.a], Kennewick’s computer system accumulates “short-term knowledge” reflecting details about a user’s utterances during a conversation. EX1004, ¶¶23, 25, 29, 31, 37, 155, 157, 160. Since a user can have multiple conversations with the computer system, a POSITA would have found it obvious that the system accumulates short-term knowledge based on one or more natural language utterances received during prior conversations, just as it accumulates short-term knowledge during a current conversation.

225. In addition, a POSITA would have found it obvious that Kennewick’s computer system “*expir[es]*” short-term knowledge associated with prior conversations because the expired knowledge no longer reflects details about a user’s utterances during a current conversation. For example, the computer system may expire short-term knowledge from a prior conversation at the beginning of a new, first, conversation.

226. To the extent Kennewick does not explicitly disclose “*expiring, by the computer system, one or more items of short-term knowledge*,” Matsuda does. Matsuda discloses a conversation system that receives natural language utterances from a user. EX1016, ¶36 (“[h]ere techniques that are generally known are used for the ‘voice recognition processing’”). The system stores information identifying a current conversation subject in a memory, which enables a conversation “to progress naturally while returning appropriate responses in relation to the subject.” *Id.*, ¶12.

227. Matsuda uses a “subject person number” which “is a number that indicates a person that is the present subject of the conversation.” *Id.*, ¶26. Matsuda stores the subject person number in a working memory. *Id.*, ¶38 (“the CPU 11 sets the ‘subject person number’...in the information storing portion 16b of the working memory 16”). If the subject person number in the working memory is 0 there is no current subject person. *Id.*, ¶26. If the number is at least 1, then the person indicated by the number is the subject. *Id.*

228. Matsuda’s system tracks elapsed time since the subject person was last mentioned by the user. *Id.*, ¶¶27 (“[t]he ‘subject person number latest setting date/time’ is used for identifying that the prescribed time has elapsed, indicating the date/time at which the name of the person who is the current subject was last mentioned by the user that is the participant in the conversation”), 69 (recording the current date/time when the subject person number is set). The system expires the

subject person number if the elapsed time exceeds a specified amount. Specifically, “if a prescribed time has elapsed since” the subject person number was last set, Matsuda’s system sets the subject person number to 0 and clears the timer tracking the elapsed time. *Id.*, ¶38. This way, the system avoids referencing the wrong person in the conversation. *Id.*, ¶100 (“if the user has, for a prescribed period of time, not used the name of a person, the subject person number is set to 0, and no statement is made regarding that person, thus enabling appropriate statements corresponding to the conversation having moved away from the subject of that person”).

229. A POSITA would have understood that the subject person number stored in the working memory is short-term knowledge because it identifies the “present subject of the conversation.” *Id.*, ¶26. This understanding is consistent with the ’249 patent’s own disclosure of “short-term knowledge” including a “history of the conversation,” which is built “as the conversation progresses.” EX1001, 16:50-53 (“[f]or instance, when a conversation begins, short-term knowledge stored in the Session Input Accumulator may be empty, and as the conversation progresses, the Session Input Accumulator may build a history of the conversation”). Accordingly, Matsuda discloses “*expiring, by the computer system, one or more items of short-term knowledge*,” it expires the subject person number which is short-term knowledge.

230. A POSITA would have been motivated to use Matsuda’s technique for expiring short term knowledge within Kennewick’s computer system. Kennewick and Matsuda are analogous art; both relate to voice-based interactions with computers. *See, e.g.*, EX1004, ¶2 (“[t]he present invention relates to the retrieval of online information and processing of commands through a speech interface”); EX1016, ¶10 (“the object of the present invention is to provide a conversation system, and a conversation processing program, able to return an appropriate response, in accordance with the subject at that time”). Kennewick’s computer system uses short-term knowledge and seeks to “achieve a natural environment for one or more users making queries or commands.” EX1004, ¶11. A POSITA would have found it obvious to incorporate Matsuda’s teachings into Kennewick because Matsuda discloses that expiring short-term knowledge “makes it possible to achieve a natural conversation.” EX1016, ¶116.

231. For example, it would have been obvious to expire Kennewick’s short-term knowledge about a prior conversation at the beginning of a new conversation, such as by clearing Kennewick’s dialog history similar to how Matsuda clears the subject person number stored in the working memory. Doing so would allow the new conversation “to progress naturally while returning appropriate responses” as taught by Matsuda because the new conversation would not use the dialog history from the prior conversation as short-term knowledge. EX1016, 4-5. This is a

combination of prior art elements according to known methods to yield predictable results. A POSITA would have had a reasonable expectation that the combination would be successful because Matsuda teaches how to expire short-term knowledge.

2. Claim 8

[8.pre] “The method of claim 7, further comprising:”

232. Kennewick-Oviatt-Fujimoto-Matsuda teaches the method of claim 7.

[8.a] “accumulating, by the computer system, long-term knowledge based on the one or more items of short-term knowledge that are expired.”

233. Kennewick and Matsuda disclose limitation [8.a].

234. Kennewick discloses the “*computer system*” “*accumulating...long-term knowledge*” as I describe for limitation [9.a]. Kennewick and Matsuda disclose “*one or more items of short-term knowledge that are expired*” as I discuss for limitation [7.a]. For example, the computer system may expire short-term knowledge from a prior conversation at the beginning of a new conversation, thereby producing “*items of short-term knowledge that are expired.*”

235. Given this disclosure, a POSITA would have found it obvious for Kennewick’s “computer system” to “*accumulat[es]...long-term knowledge based on the one or more items of short-term knowledge that are expired.*” As I discuss for limitation [9.a], Kennewick’s computer system “*accumulat[es]*” information from past or historical conversations in a user profile, by “creat[ing], stor[ing] and us[ing] extensive personal profile information for each user.” EX1004, ¶31. The

user profile includes, among other things, a “history of questions asked,” “session histories,” a “history of commands and queries,” and a “history of [individual] interactions with the system.” *Id.*, ¶143. A POSITA therefore would have found it obvious to store the expired short-term knowledge as long-term knowledge within the user profile, e.g., as part of the “session histories” because Kennewick teaches that the user profile stores this type of information to improve the operation of the computer system. *See id.*, ¶143 (“[c]ommands and questions are interpreted, queries formulated, responses created and results presented can be based on the user’s personal or user profile 110 values”). Hence, the “*long-term knowledge [is] based on the one or more items of short-term knowledge that are expired.*”

3. Claim 22

[22.pre] “The system of claim 16, wherein the one or more physical processors are caused to:”

236. Preamble [22.pre] is unpatentable for the same reasons I provide for preamble [16.pre] and limitation [16.a].

[22.a] “expire one or more items of accumulated short-term knowledge that are based on one or more natural language utterances received during conversations occurring prior to the first conversation; and”

237. Limitation [22.a] generally corresponds to limitation [7.a]. Limitation [22.a] recites “*expir[ing] one or more items of accumulated short-term knowledge*” while limitation [7.a] recites expiring “items of short-term knowledge.” As I discuss for preamble [1.pre], Kennewick’s computer system uses accumulated short-term

knowledge to facilitate natural language system responses. *See, e.g.,* EX1004, ¶155 (“as multiple users engage in overlapping sessions, the tags, added by the speech recognition engine 120 to each utterance, may allow other components of the system to tie that utterance to the correct user and dialog”). Therefore, the expired short-term knowledge is accumulated, and limitation [22.a] is obvious for the same reasons as limitation [7.a].

[22.b] “accumulate long-term knowledge based on the one or more items of accumulated short-term knowledge that are expired, wherein the first context is determined based on short-term knowledge and the long-term knowledge.”

238. Limitation [22.b] generally corresponds to limitations [8.a] and [9.a]. The first clause of limitation [22.b], “*accumulate long-term knowledge based on the one or more items of accumulated short-term knowledge that are expired,*” is nearly identical to limitation [8.a], except limitation [22.b] recites “*accumulated short-term knowledge*” while limitation [8.a] recites “short-term knowledge.” Kennewick discloses “*accumulated short-term knowledge*” as I discuss for limitation [22.a]. Accordingly, the first clause of limitation [22.b] is obvious for the same reasons I provide for limitations [8.a] and [22.a].

239. The second clause of limitation [22.b] recites “*wherein the first context is determined based on short-term knowledge and the long-term knowledge.*” This clause is identical to the last clause of limitation [9.a] and is obvious for the same reasons.

D. Ground 4: Kennewick-Oviatt-Fujimoto in combination with Cooper renders obvious claims 10, 11, 23, and 24

1. Claim 10

[10.pre] “The method of claim 1, further comprising:”

240. Kennewick-Oviatt-Fujimoto teaches the method of claim 1.

[10.a] “determining, by the computer system, based on the short-term knowledge, a manner in which the first natural language utterance is spoken, wherein the first response is generated based on the determined manner and the interpretation of the first natural language utterance.”

241. Kennewick discloses the “*computer system*” receiving a first voice input comprising a first natural language utterance as I discuss for limitation [1.a], generating short-term knowledge based on the first voice input as I discuss for limitation [1.f], and “*generating*” “*the first response...based on...the interpretation of the first natural language utterance*” as I discuss for limitation [1.i]. Kennewick does not disclose the computer system “*determining,...based on the short-term knowledge, a manner in which the first natural language utterance is spoken*” or that “*the first response is generated based on the determined manner.*” These aspects are taught by Cooper (EX1017).

242. Cooper describes a “virtual assistant [that] adapts to the user in many different ways based on the input the virtual assistant receives.” EX10017, 2:56-64, Abstract. The input includes speech (a natural language utterance) received from a user. *Id.*, Abstract (disclosing “a voice user interface for inputting information into...the virtual assistant by speech”), 6:35-37 (“Using speech recognition and text-

to-speech technology, the VA [virtual assistant] communicates with callers in spoken English.”), 7:50-53 (“the VA Engine uses its telephony interface 104 to communicate with the user 106 and its speech interface 110 to recognize speech into text and translate text into speech”), 46:58-59 (“in making utterances to the VA, the user should speak clearly and at a moderate pace”).

243. Cooper’s virtual assistant “*determin[es],...based on...short-term knowledge, a manner in which the first natural language utterance is spoken.*”

The virtual assistant determines “whether words associated with polite discourse are included in input received from the user.” *Id.*, 43:55-59, 2:26-30 (criticizing prior art’s failure to automatically adapt when the “user’s input contains words associated with polite discourse”). Cooper also explains that “the user information input into the virtual assistant could be information about the user’s emotion, which could be based on information about the user’s voice volume, word choice and speech rate. Based on such information, the virtual assistant could automatically determine the user’s emotional state, calm or angry, for example.” *Id.*, 43:62-67. Cooper thus “*determin[es],...a manner in which the first natural language utterance is spoken*” by determining whether the utterance includes words associated with polite discourse (*id.*, 43:55-59) and by determining whether the user spoke in a manner conveying an emotional state such as being calm or angry (*id.*, 43:62-67). A POSITA would have understood and found obvious that Cooper determines the

manner “***based on...short-term knowledge***” because Cooper uses information about natural language utterances received from user during a conversation to make the determination. *Id.*, 43:55-67; *see* EX1001, 5:6-7 (“[s]hort-term knowledge may accumulate during a single conversation”), 14:5-6 (same), 16:15-19 (“Intelligent Hypothesis Builder 310 may use short-term shared knowledge from the Session Input Accumulator to proactively build and evaluate interaction with a user as a conversation progresses”).

244. Cooper also discloses that “***the first response is generated based on the determined manner.***” If “words of polite discourse are included in the input received from the user, the virtual assistant could automatically enable the politeness setting.” *Id.*, 43:59-61. The politeness setting “cause[s] the virtual assistant to include words or phrases associated with polite discourse in the output from the virtual assistant.” *Id.*, 43:49-51. For example, the politeness setting could cause the virtual assistant to generate a response including words like “‘please,’ [‘]thank you,’ ‘thanks,’ ‘excuse,’ ‘pardon,’ ‘may I,’ ‘would you mind,’ or other words and phrases associated with polite discourse.” *Id.*, 43:51-55. The virtual assistant can similarly generate a response adapted to the user’s emotional state. For example, “[i]f the user’s emotional state is angry, the output of the virtual assistant could automatically include words associated with submissive discourse, such as ‘sorry,’ ‘regret’ and ‘apologize.’” *Id.*, 44:1-4.

245. Cooper's disclosure is consistent with disclosure in the '249 patent corresponding to *"determining...a manner in which the...natural language utterance is spoken, wherein the first response is generated based on the determined manner."* The '249 patent describes creating a response "by adapting to a user's manner of speaking, framing responses appropriately, and having natural variation and/or personality (e.g., by varying tone, pace, timing, inflection, word use, jargon, and other variables in a verbal or audible response)." EX1001, 6:24-28, 18:35-40 (same); *see also id.*, 6:29-32 ("the intelligent responses may adapt to a user's manner of speaking by using contextual signifiers and grammatical rules to generate one or more sentences that may cooperate with the user"), 18:44-48 (describing "taking advantage of short-term...knowledge about how a user utters a request").

246. A POSITA would have found it obvious to combine Cooper with Kennewick such that Kennewick's *"computer system"* *"determin[es]...based on the short-term knowledge, a manner in which the first natural language utterance is spoken, wherein the first response is generated based on the determined manner and the interpretation of the first natural language utterance."* Kennewick and Cooper are analogous art; both relate to voice-based interactions with computers. *See, e.g.*, EX1004, ¶2 ("[t]he present invention relates to the retrieval of online information and processing of commands through a speech interface"); EX1017,

2:36-39 (“one aspect of the present invention is a computer-based virtual assistant...comprising a voice user interface”). Kennewick’s computer system generates the short-term knowledge from the dialog with the user. EX1004, ¶¶23, 25, 29, 31, 37, 155, 157, 160. Cooper discloses using short-term knowledge including the words of a dialog to determine the manner in which the user is speaking. EX1017, 43:55-59, 44:1-4. Accordingly, a POSITA would have found it obvious to determine the manner in which the first natural language utterance is spoken based on Kennewick’s short-term knowledge because the short-term knowledge includes the types of information Cooper’s voice assistant uses to determine the user’s manner of speaking. *See, e.g.*, EX1004, ¶155 (“tags, added by the speech recognition engine 120 to each utterance, may allow other components of the system to tie that utterance to the correct user and dialog”). A POSITA also would have found it obvious to store other types of short-term knowledge taught by Cooper within Kennewick’s short-term knowledge, e.g., information about the user’s voice volume, word choice and speech rate (EX1017, 43:62-67), to determine the user’s manner of speaking.

247. A POSITA would have been motivated to provide the user with natural interactions with the computer system, and would have been motivated to use Cooper’s teachings to further this goal by adapting the computer system’s interactions, including the generated first response, to match the manner of the user.

See, e.g., *id.*, ¶11 (“[t]his environment makes significant use of context, prior information, domain knowledge, and user specific profile data to achieve a natural environment for one or more users”); EX1017, 2:25-32 (“while the prior art systems adapt automatically to the caller, the degree of adaptation is relatively limited”), 2:44-45 (“the behavior of the virtual assistant changes responsive to user input”). This is a combination of prior art elements according to known methods to yield the predictable result of matching the response from the computer system to the user’s manner of speaking. It is also use of the known technique for improving human-computer interactions disclosed in Cooper to improve human-computer interactions disclosed in Kennewick in the same way. A POSITA would have had a reasonable expectation of success because the technologies involved are conventional.

2. Claim 11: “The method of claim 10, wherein the manner in which the first natural language utterance is spoken includes at least one of a tone, a pace, an inflection, or a timing.”

248. Kennewick-Oviatt-Fujimoto-Cooper teaches the method of claim 10. As I discuss for limitation [10.a], Cooper discloses determining “*the manner in which the first natural language utterance is spoken*” by using “information about the user’s emotion, which could be based on information about the user’s voice volume, word choice and speech rate.” EX1017, 43:62-65.

249. A POSITA would have understood and found obvious that the user’s speech rate includes information about the pace and timing of the user’s speech

because these are factors that influence the rate of speech. A POSITA also would have understood and found obvious that the user's voice volume includes information about the tone and inflection of the user's speech because a user can convey tone and inflection by varying the volume of their speech. Accordingly, Cooper discloses that the manner *"includes at least one of a tone, a pace, an inflection, or a timing"* as recited in claim 11.

3. Claims 23 and 24

250. Each of dependent claims 23 and 24 is unpatentable for the same reasons as the respective corresponding claim indicated in the following table:

Claim Number	Corresponding Claim Number
23	10
24	11

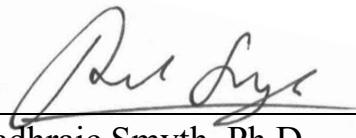
251. Each of claims 23 and 24 is substantively identical to the corresponding claim, and contains only minor differences to reflect that independent claim 16 and any intervening claims recite a system.

X. CONCLUSION

252. For the reasons I provide above, it is my opinion that claims 1-28 of the '249 patent are unpatentable.

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Dated: July 30, 2025



Padhraic Smyth, Ph.D.