

UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

GOOGLE LLC,
SAMSUNG ELECTRONICS CO., LTD., and
SAMSUNG ELECTRONICS AMERICA, INC.,

Petitioners

v.

CERENCE OPERATING COMPANY,

Patent Owner

Case IPR2024-01464

U.S. Patent No. 11,393,461

**DECLARATION OF CHRISTOPHER M. SCHMANDT IN SUPPORT OF
PETITION FOR *INTER PARTES* REVIEW OF
U.S. PATENT NO. 11,393,461**

TABLE OF CONTENTS

I.	INTRODUCTION	1
A.	BACKGROUND AND QUALIFICATIONS	2
B.	INFORMATION CONSIDERED	6
II.	LEGAL STANDARDS	8
A.	Anticipation	8
B.	Obviousness.....	9
C.	Level of Ordinary Skill in the Art.....	12
D.	Claim Construction	13
III.	SUMMARY OF OPINIONS.....	15
IV.	THE '461 PATENT.....	15
A.	Overview of the '461 Patent.....	15
B.	Prosecution History	18
V.	TECHNOLOGY BACKGROUND.....	19
A.	It was well known to use contextual cues in speech recognition systems for power saving benefits.	19
B.	Noise reduction and/or cancellation techniques were well known in order to better process speech input and speech recognition.	29
C.	Voice activity detection was well known to detect human speech.	35
VI.	BRIEF DESCRIPTION OF THE APPLIED PRIOR ART	37
A.	Sarin (Ex-1005)	37
B.	Hwang (Ex-1006).....	41
C.	Garner (Ex-1007)	45

D.	Gruber (Ex-1008)	47
E.	Lebeau (Ex-1016).....	50
VII.	CLAIM CONSTRUCTION	50
VIII.	DETAILED EXPLANATION OF GROUNDS.....	51
A.	Ground 1: Sarin in Combination with Hwang Renders Obvious Claims 1-4, 6, 7, 11, 14, 16, and 18	51
1.	A POSITA would have been motivated to combine Sarin’s and Hwang’s teachings and would have had a reasonable expectation of success.....	51
2.	Independent Claim 1	58
a.	1[pre]: A method of monitoring an acoustic environment of a mobile device for voice commands when the mobile device is operating in a low power mode, the method comprising:	58
b.	1[a]: receiving acoustic input from the environment of the mobile device while the mobile device is operating in idle mode; and.....	60
c.	1[b]: in response to receiving the acoustic input, determining whether the received acoustic input includes a voice command by processing the received acoustic input, ...	63
d.	1[c]: the processing employing at least one contextual cue, the at least one contextual cue comprising information other than content of the acoustic input and comprising a time the acoustic input was received, motion of the mobile device, location of th mobile device, occurrence of an event on the mobile device, recent activity on the mobile device, a stored sample of the acoustic environment of the mobile device and/or a power source of the mobile device,.....	66

- e. 1[d]: the processing of the received acoustic input comprising a plurality of processing stages, at least one of the processing stages being performed while the mobile device is operating in the idle mode.....67
- 3. Dependent Claims 2-4, 6, 7, 11, 14, and 16.....69
 - a. Claim 2: The method of claim 1, further comprising transitioning the mobile device from the idle mode to an active mode in response to detecting that the acoustic input includes a voice command. 69
 - b. Claim 3: The method of claim 1, wherein the at least one contextual cue includes the time the acoustic input was received, and the method comprises influencing evaluation of at least one of the plurality of processing stages based on the time the acoustic input was received. 70
 - c. Claim 4: The method of claim 1, wherein the at least one contextual cue includes recent activity on the mobile device, the method further comprising evaluation of at least one of the plurality of processing stages based on at least one of how recently the user activated the mobile device, how recently the user issued a voice command and how recently the user performed at least one action on the mobile device. 72
 - d. Claim 6: The method of claim 1, wherein the at least one contextual cue includes motion of the mobile device received as an input from an accelerometer of the mobile device, the method comprises influencing evaluation of at least one of the plurality of processing stages based on an assessment of whether the user moved the mobile device proximate in time with receiving the acoustic input based, at least in part, on the input from the accelerometer. 75
 - e. Claim 7: The method of claim 1, wherein the at least one contextual cue includes the location of the mobile device, and the method comprises influencing evaluation of at

	least one of the plurality of processing stages based on determining whether the mobile device is located in at least one of a vehicle, a home or a place of work of the user of the mobile device.....	82
f.	Claim 11: The method of claim 1, wherein the mobile device includes a primary processor and a secondary processor, and wherein at least one first processing stage of the plurality of processing stages is performed by the secondary processor prior to engaging the primary processor to process the acoustic input.	83
g.	Claim 14: The method of claim 11, wherein performing the at least one first processing stage comprises performing limited vocabulary automatic speech recognition on the acoustic input to detect whether the acoustic input includes an explicit voice trigger.....	85
h.	Claim 16: The method of claim 1, wherein the mobile device is capable of connecting to at least one server via a network, and the method comprises transmitting at least a portion of the acoustic input to the at least one server via the network for processing by the at least one server.	86
4.	Independent Claim 18	88
a.	18[pre]: A mobile device comprising:.....	88
b.	18[a]: at least one processor configured to:.....	88
	(i) 18[a(i)]: in response to the mobile device receiving acoustic input from the environment while operating in an idle mode, determine whether the received acoustic input includes a voice command by processing the received acoustic input, ...	89
	(ii) 18[a(ii)]: the processing employing at least one contextual cue, the at least one contextual cue comprising information other than content of the received acoustic input and comprising a time the acoustic input was	

received, motion of the mobile device, location of the mobile device, occurrence of an event on the mobile device, recent activity on the mobile device, a stored sample of the acoustic environment of the mobile device and/or a power source of the mobile device,.....89

(iii) 18[a(iii)]: the processing of the received acoustic input comprising a plurality of processing stages, at least one of the processing stages being performed while the mobile device is operating in the idle mode.....89

B. Ground 2: Sarin in Combination With Hwang and Garner Renders Obvious Claim 13.....90

1. A POSITA would have been motivated to combine Sarin's, Hwang's, and Garner's teachings and would have had a reasonable expectation of success.....90

2. Dependent Claim 13 92

a. Claim 13: The method of claim 11, wherein performing the at least one first processing stage includes performing at least one voice activity detection processing stage including performing at least one of spectral analysis on the acoustic input to evaluate whether a spectrum of the acoustic input is indicative of voice activity, periodicity analysis to evaluate whether signal periodicity of the acoustic input is indicative of voice activity, and phone loop analysis to evaluate whether the acoustic input includes speech. 92

C. Ground 3: Sarin in Combination With Hwang and Gruber Renders Obvious Claims 3, 5, 7, 9, 10, 12, 15, and 17 100

1. A POSITA would have been motivated to combine Sarin's, Hwang's, and Gruber's teachings and would have had a reasonable expectation of success.....100

2. Dependent Claims 3, 5, 7, 9, 10, 12, 15, and 17 104

- a. Claim 3: The method of claim 1, wherein the at least one contextual cue includes the time the acoustic input was received, and the method comprises influencing evaluation of at least one of the plurality of processing stages based on the time the acoustic input was received.
.....104
- b. Claim 5: The method of claim 1, wherein the at least one contextual cue includes an occurrence of an event on the mobile device, and the method comprises influencing evaluation of at least one of the plurality of processing stages based on whether the acoustic input is received proximate in time with the occurrence of the event. 107
- c. Claim 7: The method of claim 1, wherein the at least one contextual cue includes the location of the mobile device, and the method comprises influencing evaluation of at least one of the plurality of processing stages based on determining whether the mobile device is located in at least one of a vehicle, a home or a place of work of the user of the mobile device.....113
- d. Claim 9: The method of claim 1, wherein the at least one contextual cue includes the sample of the acoustic environment obtained by the mobile device by storing acoustic input previously received by the mobile device and determined not to include a voicecommand, and the method comprises subtracting the sample of the acoustic environment from the acoustic input.....116
- e. Claim 10: The method of claim 1, wherein a sequence of performing the plurality of processing stages is determined, at least in part, by the at least one contextual cue.....120
- f. Claim 12: The method of claim 11, wherein at least one second processing stage of the plurality of processing stages is performed by the primary processor only after the at least one first processing stage is performed and neither concludes that the acoustic input includes a voice

	command nor concludes that the acoustic input is the result of spurious acoustic activity.	124
g.	Claim 15: The method of claim 11, wherein performing the at least one second processing stage comprises performing at least one speech processing stage including at least one of automatic speech recognition on the acoustic input and at least one natural language processing stage to understand the meaning of the voice command.....	127
h.	Claim 17: The method of claim 1, comprising performing one or more default actions when a specific task could not be ascertained from the acoustic input.	130
D.	Ground 4: Sarin in Combination with Hwang and Lebeau Renders Obvious Claim 8.....	132
1.	A POSITA would have been motivated to combine Sarin’s, Hwang’s, and Lebeau’s teachings and would have had a reasonable expectation of success.....	132
2.	Dependent Claim 8	134
a.	Claim 8: The method of claim 1, wherein the at least one contextual cue includes the power source of the mobile device, and the method comprises influencing evaluation of at least one of the plurality of processing stages based on at least one of determining whether the mobile device is being powered using a battery or wall power, and a state of the battery if the mobile device is being powered using the battery.....	134
IX.	SECONDARY CONSIDERATIONS	137
X.	CONCLUSION.....	137

I. INTRODUCTION

1. I have been retained by the Samsung Electronics Co., Ltd. and Samsung Electronics America, Inc. (“Samsung” or “Petitioners”) as an independent expert consultant in this proceeding before the United States Patent and Trademark Office (“PTO”).

2. I am being compensated at a rate of \$550/hour for my services in this proceeding, which is my regular and customary rate.

3. My compensation is in no way contingent on the nature of my finding, the presentation of my findings in testimony or this declaration, or the outcome of this or any other proceeding. I have no other interest in this proceeding.

4. I have been asked to consider whether certain references disclose and/or suggest the features recited in Claims 1-18 of U.S. Patent No. 11,393,461 (“the ’461 Patent”) (Ex. 1001).¹ My opinions are set forth below.

5. I have been asked to provide my independent analysis of the ’461 Patent in light of the prior art publications cited below. I have also been asked to consider the state of the art and prior art available as of March 12, 2013. Based on the prior art discussed in this declaration, it is my opinion that Claims 1-18 of the ’461 Patent are unpatentable for the reasons provided below.

¹ Where appropriate, I refer to exhibits which I understand will be attached to the petition for inter partes review of the ’461 patent (the “Petition”).

A. BACKGROUND AND QUALIFICATIONS

6. I retired several years ago after a 40-year career at the Massachusetts Institute of Technology (“MIT”); for most of that time I was employed as a Principal Research Scientist at the Media Laboratory. In that role I also served as faculty for the MIT Media Arts and Sciences academic program. I was a founder of the Media Laboratory, a research lab which now spans two buildings.

7. I received my B.S. degree in Electrical Engineering and Computer Science from MIT in 1978, and my M.S. in Visual Studies (Computer Graphics) also from MIT. I was employed at MIT since 1980, initially at the Architecture Machine Group which was an early computer graphics and interactive systems research lab. In 1985, I helped found the Media Laboratory and continued to work there until retirement. I was director of a research group titled “Living Mobile.” My research spanned distributed communication and collaborative systems, with an emphasis on multi-media and user interfaces, with a strong focus on speech-based systems. I have over 70 published conference and journal papers and one book in the field of speech technology and user interaction.

8. For the first fifteen years of my career, my research emphasized speech recognition and speech user interfaces. I built the first conversational computer system utilizing speech recognition and synthesis (“Put That There”) starting in 1980. Put That There parsed spoken input into a semantic structure

representing a command or intent, and could ask questions of the user to evoke responses which would be further parsed and melded into the context of the current command. Some information could be gleaned from non-speech channels, specifically where a user was pointing. I continued to innovate speech user interfaces using recognition, text-to-speech synthesis, and recorded audio in a wide variety of projects. I built one of the first graphical user interfaces for audio editing, employing keyword recognition on voice memos in 1982 (“Intelligent Ear”). I built the first research-grade unified messaging system, which combined text and voice messages into a single inbox, with speech recognition over the phone for remote access, and a graphical user interface for desktop access in 1983 (“Phone Slave”). Again, Phone Slave included a parser which could accept fragmentary input and build up a complete representation of the user’s utterance with additional input. Especially in its follow up multi-user system, Conversational Desktop (1985), a number of contextual cues could also aid the dialog understanding, including date, time, calendar entries, and with whom a user was speaking over a teleconference channel. Along with my students we built the first system for real time spoken driving directions, including speech-accessible maps of Cambridge, Massachusetts in 1987 (“Back Seat Driver”). Back Seat Driver had to translate a route from a map form in which each block was a separate street segment with intersection information contained in nodes, into spoken directions,

using words like we use with each other to tell the driver where to go. We built some of the earliest speech-based personal assistants for managing messages, calendar, contacts, etc. (Conversational Desktop 1985, Chatter 1993, MailCall 1996). We built quite a few systems employing speech recognition in handheld mobile devices (ComMotion 1999, Nomadic Radio 2000, Impromptu 2001, and Symphony 2004, for example). We applied speech recognition to large bodies of everyday conversations captured with a wearable device and utilized as a memory aid (Memory Prosthesis 2004). We used speech recognition on radio newscasts to build a personalized version of audio newscasts (Synthetic News Radio, 1999) and also investigated adding speech recognition to a mouse-based window system a few years earlier.

9. In addition to this work running my academic research lab, I helped found and then design speech user interfaces for InTouch Systems, which built natural language based personal digital assistants accessible over the telephone. In this context I helped solve a number of problems in obtaining adequate speech recognition performance in the face of personal information, such as contact lists, containing words which were difficult for a system to understand. The system also maintained dialog context, such that while reading email, for example, commands related to email were anticipated, such as finding important messages or messages from some particular person. When a user was accessing a calendar, a different

grammar was invoked, emphasizing creation and retrieval of calendar entries by date or time.

10. I was later awarded the prestigious Association for Computing Machinery (ACM) Computer Human Interface (CHI) Academy membership specifically for those years of work pioneering speech user interfaces.

11. In the course of my research, I built a number of speech recognition client/server distributed systems, with the first being in 1985. Much of the initial motivation for a server architecture was that speech recognition required expensive digital signal processing hardware that we could not afford to put on each computer, so a central server with the required hardware was used. Later versions of the speech recognition server architecture allowed certain computers to perform specialized tasks serving a number of client computers providing voice user interfaces, either on screens or over telephone connections. Because of my early work with distributed speech systems, I served for several years in the mid-1990s with a working group on the impact of multimedia systems on the Internet reporting to the Internet Engineering Task Force (IETF) and later the Internet Activities Board (IAB). This work impacted emerging standards such as Session Initiation Protocol (SIP).

12. In my faculty position I taught graduate level courses in speech technology and user interaction design, and directly supervised student research

U.S. Patent No. 11,393,461

Declaration of Mr. Christopher M. Schmandt

and theses at the Bachelors, Masters, and PhD level. I oversaw the Masters and PhD thesis programs for the entire Media Arts and Sciences academic program during my more senior years. I also served on the Media Laboratory intellectual property committee for many years.

13. My Curriculum Vitae is attached as Ex-1003.

14. I have been retained by counsel for Petitioner to provide my expert opinion in connection with the above-captioned proceeding as set forth herein.

B. INFORMATION CONSIDERED

15. In forming my opinions, I reviewed the following documents, in addition to others I have cited in my declaration. I also relied on my own knowledge of and experience in the field of speech recognition.

Exhibit No.	DESCRIPTION
1001	U.S. Patent 11,393,461 (“’461 Patent”)
1004	File History of U.S. Patent No. 11,393,461
1005	U.S. Patent No. 9,031,847 (“Sarin”)
1006	U.S. Patent Application Publication No. 2013/0110521 (“Hwang”)
1007	U.S. Patent No. 7,475,012 (“Garner”)
1008	U.S. Patent Application Publication No. 2012/0265528 (“Gruber”)
1009	U.S. Patent No. 8,396,463 (“Marcellino”)
1010	U.S. Patent No. 8,331,315 (“Purkayastha”)
1011	U.S. Patent No. 8,694,316 (“Pickering”)

U.S. Patent No. 11,393,461
Declaration of Mr. Christopher M. Schmandt

Exhibit No.	DESCRIPTION
1012	U.S. Patent Application Publication No. 2013/0173269 (“Adler”)
1013	Intentionally Left Blank
1014	U.S. Patent Application Publication No. 2020/0279576 (“Binder”)
1015	U.S. Patent Application Publication No. 2014/0006825 (“Shenhav”)
1016	U.S. Patent Application Publication No. 2012/0034904 (“Lebeau”)
1017	U.S. Patent Application Publication No. 2010/0129068 (“Binda”)
1018	U.S. Patent Application Publication No. 2012/0296655 (“Kristjansson”)
1019	U.S. Patent Application Publication No. 2014/0163978 (“Basye”)
1020	Claim Mapping Table
1021	Plaintiff Cerence Operating Company’s Disclosure of Asserted Claims and Infringement Contentions, <i>Cerence Operating Company v. Samsung Electronics Co., Ltd. et al.</i> , No. 2:23-cv-0482-JRG-RSP (E.D. Tex. Mar. 11 2024)
1022	Docket Control Order, <i>Cerence Operating Company v. Samsung Electronics Co., Ltd. et al.</i> , No. 2:23-cv-0482-JRG-RSP (E.D. Tex. Apr. 16, 2014)
1023	Intentionally Left Blank
1024	U.S. Patent Application Publication No. 2013/0332165 (“Beckley”)
1025	U.S. Patent Application Publication No. 2014/0350927 (“Yamabe”)
1026	U.S. Patent Application Publication No. 2003/0179888 (“Burnett”)
1027	Intentionally Left Blank
1028	Intentionally Left Blank
1029	U.S. Patent Application Publication No. 2013/0006638 (“Lindahl”)

II. LEGAL STANDARDS

16. In forming my opinions and considering the subject matter of the '461 Patent and its claims in light of the prior art, I am relying on certain legal principles that counsel in this case explained to me. My understanding of these concepts is summarized below.

17. I understand that earlier publications and patents may act to render a patent unpatentable for one of two reasons: (1) anticipation, and (2) obviousness.

A. Anticipation

18. As explained to me by counsel, the claims of a patent are anticipated by a prior art reference if each and every element of the claim is found either explicitly or inherently in the reference. I understand that inherency requires a showing that the missing descriptive matter in the claim is necessarily present in the allegedly anticipating reference, and that it would have been so recognized by a person of ordinary skill in the art ("POSITA").

19. I understand that when a challenged claim covers several structures, either generically or as alternatives, the claim is deemed anticipated if any of the structures within the scope of the claim is found in the prior art reference.

20. Although anticipation typically involves the analysis of a single prior art reference, I understand that additional references may be used to show that the prior art reference has enabling disclosure (i.e., allows a POSITA to make the

invention without undue experimentation), to explain the meaning of a term used in the prior art reference, and/or to show that a characteristic is inherent in the prior art reference.

B. Obviousness

21. As explained to me by counsel, a claim is invalid as obvious if it would have been obvious to a person of ordinary skill in the art at the time the alleged invention was made. This means that even if all of the elements of the claim cannot be found in a single prior art reference that would anticipate the claim, a person of ordinary skill in the art who was aware of the prior art would have been able to come up with the claimed invention. This may be the case, for example, where the missing element represents only an insubstantial different over the prior art or a reconfiguration of a known system. I understand that in an obviousness determination, the person of ordinary skill in the art is presumed to have knowledge of all material prior art.

22. I understand that an obviousness analysis requires an understanding of the scope and content of the prior art, any differences between the alleged invention and the prior art, and the level of ordinary skill in evaluating the pertinent art.

23. I understand that when a product is available, design incentives and other market forces can prompt variations of it, either in the same field or a

different one. If a person of ordinary skill in the art can implement a predictable variation, obviousness likely bars its patentability. For the same reason, if a technique has been used to improve one device and a person of ordinary skill in the art would recognize that it would improve similar devices in the same way, using the technique would have been obvious.

24. I understand that whether a prior art reference renders a patent claim unpatentable as obvious is determined from the perspective of a person of ordinary skill in the art at the time of the alleged invention. I have been told that there is not requirement that the prior art contain an express suggestion to combine known elements to achieve the claimed invention, but a suggestion to combine known elements to achieve the claimed invention may come from the prior art, as filtered through the knowledge of one skilled in the art. In addition, I have been told that the inferences and creative steps a person of ordinary skill in the art would employ are relevant to the determination of obviousness.

25. I understand that one may consider, e.g., whether (1) the change was merely the predictable result of using prior art elements according to their known functions, or whether it was the result of true inventiveness; (2) there is some teaching or suggestion in the prior art to make the modification or combination of elements claimed in the patent; (3) the claimed innovation applies a known technique that had been used to improve a similar device or method in a similar

way; (4) the claimed invention would have been obvious to try, meaning that the claimed innovation was one of a relatively small number of possible approaches to the problem with a reasonable expectation of success by those skilled in the art; (5) the invention merely substituted one known element for another known element in order to obtain predictable results; (6) the invention merely applies a known technique to a known device, method, or product to yield predictable results; or (7) known work in one field of endeavor may have prompted variations of it for use in either the same field or a different one based on design incentives or other market forces that would have been predictable to a person of ordinary skill in the art.

26. I further understand that certain factors may support or rebut the obviousness of a claim. I understand that such secondary considerations include, among other things, commercial success of the patented invention, skepticism of those having ordinary skill in the art at the time of the invention, unexpected results of the invention, any long-felt but unsolved need in the art that was satisfied by the alleged invention, the failure of others to make the alleged invention, praise of the alleged invention by those having ordinary skill in the art, and copying of the alleged invention by others in the field. I understand that there must be a nexus—that is, a connection—between any such secondary considerations and the alleged invention. I also understand that contemporaneous and independent invention by others is a secondary consideration tending to show obviousness.

27. I am not aware of any allegations by the named inventor of the '461 Patent or any assignee of the '461 Patent that any secondary considerations tend to rebut the obviousness of the '461 Patent.

28. Additionally, I understand that in considering obviousness, it is important not to use the benefit of hindsight derived from the patent under consideration.

C. Level of Ordinary Skill in the Art

29. When interpreting a patent, I understand that it is important to identify the relevant art pertaining to that patent, as well as the level of ordinary skill in that art at the time of the claimed invention. The “art” is the field of technology to which the patent is related.

30. I am informed and understand that the POSITA is a hypothetical person who is presumed to know the relevant prior art. I understand that the actual inventor’s skill is not determinative of the level of ordinary skill. I further understand that factors that may be considered in determining level of skill include: (i) the types of problems encountered in the art; (ii) prior art solutions to those problems; (iii) the rapidity with which innovations are made; (iv) the sophistication of the technology; and (v) the educational level of active workers in the field. I understand that not all such factors may be present in every case, and one or more of them may predominate.

31. I understand that a POSITA is one who is presumed to be aware of all pertinent art, thinks along conventional wisdom in the art, and is a person of ordinary creativity.

32. A person of ordinary skill in the art at the relevant time, which I have been told to consider as March 12, 2013 (“POSITA”) would have had at least a four-year Bachelor of Science (or equivalent) in electrical engineering or computer science, and two years of work experience in the field of speech and natural language interfaces for interacting with and directing computer systems and software, including speech recognition and processing. A higher level of education or skill might make up for less experience, and vice versa.

33. Based on my knowledge, skill, and experience, I have an understanding of the capabilities of a person of ordinary skill in the relevant art. For example, from my industry experience, I am familiar with what an engineer designing speech and natural language interfaces would have knowledge of. From teaching and supervising my post-graduate students, I also have an understanding of the knowledge that a person with this academic experience possesses. Furthermore, I possessed the knowledge of a POSITA myself at least as of 1980.

D. Claim Construction

34. I understand that the United States Patent and Trademark Office interprets claim terms in an *inter partes* review proceeding under the same claim

construction standard that is used in a United States federal court. I understand that under this standard, the meaning of claim terms is considered from the viewpoint of a POSITA at the time of the alleged invention.

35. I have been informed that claim terms are generally given their ordinary and customary meaning as understood by one of ordinary skill in the art in light of the specification and the prosecution history pertaining to the patent. I understand, however, that claims terms are generally not limited by the embodiments described in the specification.

36. I understand that in addition to the claims, specification, and prosecution history, other evidence may be considered to ascertain the meaning of claim terms, including textbooks, encyclopedias, articles, and dictionaries. I have been informed that this other evidence is often less significant and less reliable than the claims, specification, and prosecution history.

37. I have given all the claim terms their plain and ordinary meaning, as would have been understood by a person having ordinary skill in the art at the time of the alleged invention, which I understand is March 12, 2013, having taken into consideration the language of the claims, the specification, and the prosecution history of record.

III. SUMMARY OF OPINIONS

38. I understand that Petitioner challenges the validity of Claims 1-18 of the '461 patent. For the reasons I discuss below, it is my opinion that Claims 1-18 of the '461 patent are rendered obvious by the prior art, according to the following grounds:

Ground	Summary
1	Claims 1-4, 6, 7, 11, 14, 16, 18 are obvious over Sarin in view of Hwang
2	Claim 13 is obvious over Sarin in view of Hwang and Garner
3	Claims 3, 5, 7, 9, 10, 12, 15, and 17 are obvious over Sarin in view of Hwang and Gruber
4	Claim 8 is obvious over Sarin in view of Hwang and Lebeau

IV. THE '461 PATENT

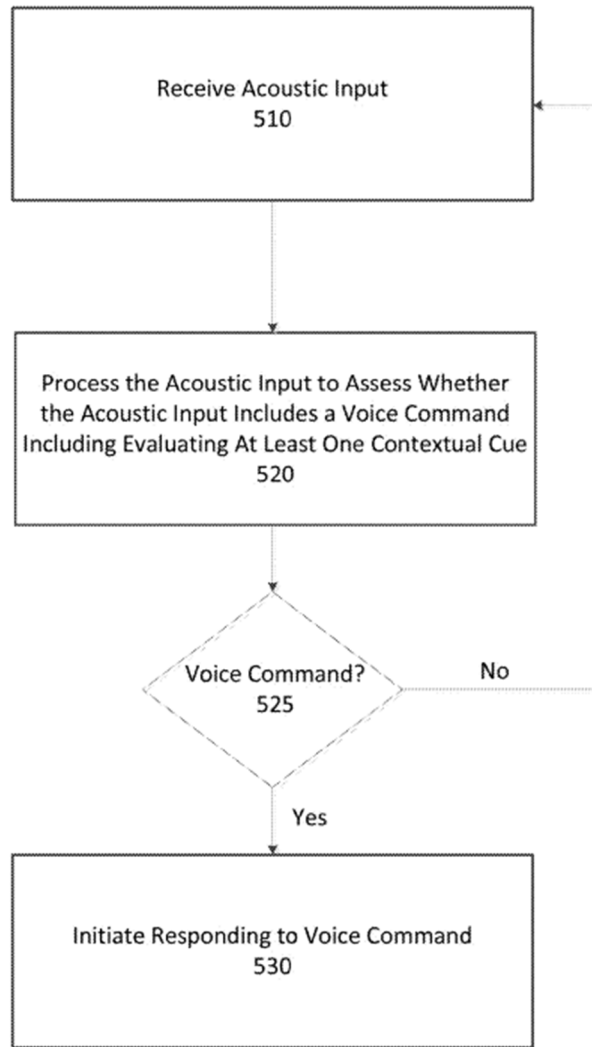
A. Overview of the '461 Patent

39. The '461 patent is generally directed to techniques used by mobile devices to detect voice commands in received audio. Ex-1001, Abstract. The techniques disclosed and claimed by the '461 patent allegedly enable mobile devices to detect voice commands in a way that minimizes the device's power consumption. *Id.*, 6:17-20. To this end, the patent discloses “a hierarchical

approach to monitoring and processing acoustic input that applies relatively low power processing to received acoustic information first and generally higher power processing as needed to evaluate the likelihood that the acoustic information includes a voice command.” *Id.*, 6:21-26.

40. Embodiments disclosed by the ’461 patent use what the patent refers to as “contextual cues” to help the mobile device detect whether received acoustic input includes a voice command. *See id.*, 2:29-31, 2:42-44, 2:53-55. The patent explains that a “contextual cue” “refers generally to information other than the properties or characteristics of the specific acoustic input being evaluated as containing a voice command.” *Id.*, 6:52-55.

41. Figure 5 of the ’461 patent illustrates a method in which one or more contextual cues are used to assist in determining whether a received acoustic input includes a voice command. *Id.*, 20:35-42.



Ex-1001, FIG. 5

42. As indicated in the figure, after acoustic input is obtained at step 510, it is processed to evaluate whether it includes a voice command. Ex-1001, 20:37-42. As part of the evaluation, “at least one contextual cue is used to assist in determining whether the acoustic input includes a voice command.” *Id.*, 20:42-44. Example contextual cues include “the time of day, the location of the mobile device, motion of the mobile device, recent activity on the mobile device,

biometric information, status of the power source, information about the acoustic environment, etc.” *Id.*, 20:45-49.

43. As one example of how a contextual cue may be used, the patent discloses that, if a mobile device user is preparing to speak a voice command, then that user may pick up the mobile device or bring it closer to their mouth. *Id.*, 6:55-57. Thus, in this instance, “motion may be used as a contextual cue to influence the evaluation of whether acoustic input likely includes a voice command.” *Id.*, 6:57-59.

B. Prosecution History

44. The ’461 patent was filed on 3/12/2013 as U.S. Application 13/795,933 (“the ’933 App.”). The ’933 App. remained pending for over nine years. This Petition presents arguments not considered during prosecution. *See* Ex-1004.

45. During prosecution, the claims of the ’933 App. were rejected ten times and the applicant filed four Requests for Continued Examination. Of note, on October 30, 2015, the applicant amended the independent claims to replace the term “low power mode” with the term “idle mode,” and to recite that the at least one contextual cue is used “while the mobile device is operating in the idle mode.” Ex-1004, 478-482. On April 4, 2016, the applicant further amended the

U.S. Patent No. 11,393,461
Declaration of Mr. Christopher M. Schmandt
independent claims to recite specific types of contextual cues used by the claimed mobile device. *Id.*, 535-539.

46. On March 3, 2022, the applicant overcame a double patenting rejection by filing a terminal disclaimer over U.S. Patent No. 9,361,885 and canceled one of the independent claims to obviate a §101 rejection. *Id.*, 1457-1471. The Office subsequently issued a Notice of Allowance on March 16, 2022. *Id.*, 1463-1471.

47. I understand, as informed by counsel, that the Lebeau reference (Ex-1016) was considered by the Examiner during the prosecution of the '461 Patent. However, the examiner did not apply Lebeau's teaching with Sarin and/or Hwang.

V. TECHNOLOGY BACKGROUND

A. It was well known to use contextual cues in speech recognition systems for power saving benefits.

48. In view of the various applications of speech recognition, there are various considerations that systems use beyond the acoustic input itself, that a POSITA would likely consider when designing a speech recognition system to run on a mobile battery powered device. I have summarized only a few considerations of the electronic device arrangements below.

49. For example, U.S. Patent No. 9,031,847 to Sarin ("Sarin", Ex-1005) describes how a voice-controlled application can use contextual cues to assist in

recognizing commands and using resources efficiently. Sarin teaches that contextual clues include “which applications or functions are active on a device, time of day, and location,” such that if the time and day and location indicate that the user is at work, “voice recognition relating to image capture operations can be turned off” so as to avoid unintended camera operations. Ex-1005, 11:32-47. Sarin also describes how the user’s recent activity, such as an active phone call, can be considered when determining whether voice recognition relating to image capture operations can be turned off, so as to avoid unintended camera operation. *Id.*, 11:47-52. This is beneficial because it saves battery by preventing voice recognition, which may require significant computation and therefore power, relating to image capture operations when it is highly unlikely that the user would invoke its mobile device’s camera since the user is at work or otherwise using their mobile device.

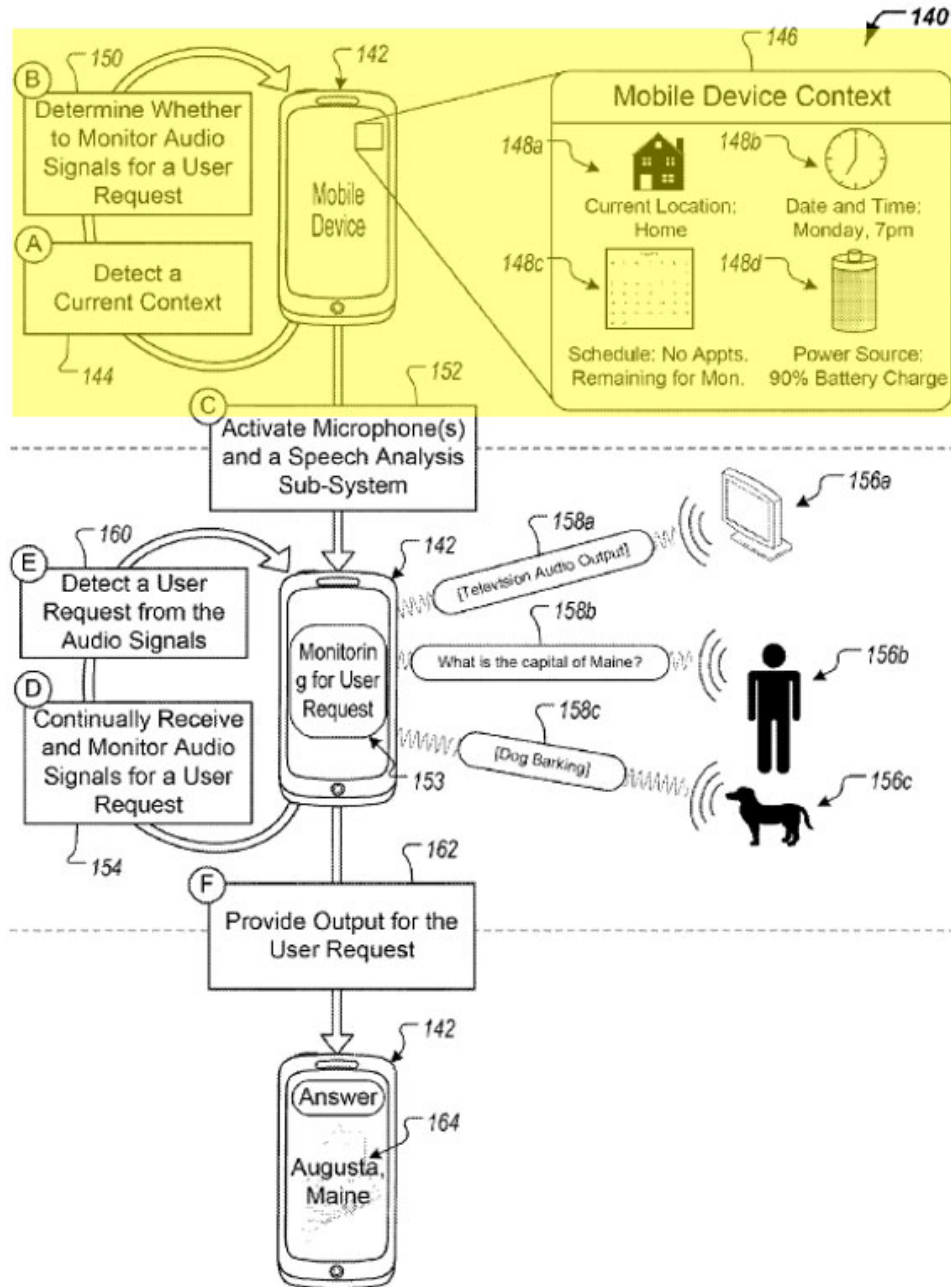
50. Similarly, U.S. Patent Application Publication No. 2012/0034904 to Lebeau (“Lebeau,” Ex-1016) titled “Automatically Monitoring for Voice Input Based on Context,” explains using contextual cues so as to evaluate the likelihood that a user would speak a verbal search request. Ex-1016 ¶¶3, 22 (“the mobile computing device can attempt to infer whether the current context indicates that the user has at least a threshold likelihood of providing voice input and, if so, monitoring for voice input in response.”), 43 (“threshold likelihood of [the user]

providing a voice-based request to the mobile computing device”). Lebeau explains:

current context associated with a mobile computing device (and/or with a user of the mobile computing device) can include a context external to the device, such as that represents an environment around the device, or a context internal to the device such as historical information about the device that is stored in the device. Context external to the device can include, for example, the physical location where the mobile computing device is located (e.g., home, work, car, etc., as determined by GPS in the device or other techniques), and motion of the mobile computing device (e.g., accelerating, stationary, etc.). Context that is internal to the device can include recent activity on the mobile computing device (e.g., social network activity, emails sent/received, telephone calls made/received, etc.). The current context for a mobile computing device (and/or its user) is separate from user input itself that would direct the device to listen for spoken input.

Ex-1016 ¶3.

51. Specifically, Lebeau’s system includes a context determination unit “that is configured to detect a current context associated with the mobile computing device,” indicating a current state of the surrounding environment. Ex-1016 ¶8. As shown in Figure 1B below, at step A, Lebeau detects contextual cues (such as the user’s location, the time of day, calendar, and battery status) and, at step B, uses the contextual cues to determine whether to monitor for voice input autonomously. Ex-1014 ¶42.



52. At step B, Lebeau explains that its device will “determine[] to proceed with monitoring ambient audio signals for a user request based on an inferred likelihood that the user will provide a user request and convenience to both the

user and the mobile computing device 142, as indicated by the context 146”. Ex-

1016 ¶43. The likelihood that the user will provide a user request,

can be inferred from, at least, the time (7 pm) and the user's schedule. Although it is evening, the user is likely to have not gone to bed yet (it is only 7 pm) and the user does not have any appointments for the remainder of the evening—the user's anticipated free time over the next several hours can indicate at least a threshold likelihood of providing a voice-based request to the mobile computing device 142. Monitoring for voice input can be convenient for the user based on, at least, the mobile computing device 142 being located at the user's home where the user may be more than an arm's length away from the mobile computing device 142 (e.g., the user may be moving around the house such that it may be more convenient for a user to simply speak his/her requests instead of having to locate the mobile computing device 142 to manually prompt the computing device 142 for each request). Additionally, monitoring the voice input can be convenient for the mobile computing device based on, at least, the battery having at least a threshold charge and based on a projection that monitoring will only last for a limited period of time (e.g., the mobile computing device 142 can forecast that the user will likely go to bed within a few hours).

Ex-1016 ¶43. Therefore, Lebeau uses the user's time, location, schedule, and the mobile device's battery status as contextual cues of when to monitor for the user's speech, which “help[s] conserve the amount of energy consumed by a mobile computing device, especially when the mobile computing device is using a portable power source, such as a battery.” Ex-1016 ¶10.

53. Another example that Lebeau provides is in the car-context, where the computing device “may determine that it is currently travelling in a car based on a detected high rate of speed at which the device is travelling (e.g., using any of a

variety of motion sensors that are standard components of the device).” Ex-1016

¶22. Using the user’s location based on “motion sensors,” Lebeau can determine whether to monitor for voice input based on this current context. Ex-1016 ¶22. Like above, this “help[s] conserve the amount of energy consumed by a mobile computing device, especially when the mobile computing device is using a portable power source, such as a battery” by not monitoring for voice when the motions indicate that it would or would not be convenient for the user to speak a user request. Ex-1016 ¶10.

54. Sarin also notes that “other states and state transitions can apply where other device components (e.g., a gyroscope, an accelerometer) provide additional contextual cues (e.g., detecting the orientation of the device as being ready perform an operation such as capturing an image or recording video) for recognizing voice commands.” Ex-1005, 12:16-24.

55. Using accelerometers in speech recognition systems is well-known. Like Sarin, Binder teaches a mobile device with a motion sensor such as an accelerometer. Ex-1014 ¶52. Binder further describes that “the context of the electronic device is used to help determine whether and how to operate the voice trigger,” and lists a number of devices useful for detecting such context, including accelerometers. Ex-1014 ¶128. Binder explains that the context of the device can be used to adjust how and whether the voice trigger operates. Ex-1014 ¶129

(further explaining that the voice trigger is deactivated when the phone is in a predetermined orientation, during predetermined time periods, when the phone is in a silent or do not disturb mode, in an enclosed space, or near other devices). The accelerometer can be used to determine that the device is face-up or face down so that the system knows to activate the voice trigger when the device is placed “face-up.” Ex-1014 ¶131; *see also* Ex-1014 ¶¶166-167 (“electronic device detects its orientation using light sensors (including cameras), microphones, proximity sensors, magnetic sensors, accelerometers, gyroscopes, tilt sensors, and the like.”). Like Lebeau’s vehicle example, Binder uses the accelerometer to determine that the device is in a vehicle by monitoring the motion and if it is determined that device is likely in a vehicle because it is traveling above 20 miles per hour, then it “is determined to be travelling along a road” and “then the voice trigger remains active and/or in a high-power or more sensitive state.” Ex-1014 ¶133. Specifically, Binder’s accelerometer detects “the motion of the electronic device and/or the external audio system 416,” which is then used to:

determine whether the voice trigger system 400 should monitor ambient sound using the on-device microphone 230 or an external microphone 418. Specifically, the difference between the motion of the device and the external audio system 416 provides information about whether the external audio system 416 is actually in use. For example, if both the device and a wireless headset are moving (or not moving) substantially identically, it may be determined that the headset is not in use or is not being worn. This may occur, for example, because both devices are near to each other and idle (e.g., sitting on a table or stored in a pocket, bag, purse, drawer, etc.).

Accordingly, under these conditions, the voice trigger system 400 monitors the on-device microphone, because it is unlikely that the headset is actually being used. If there is a difference in motion between the wireless headset and the device, however, it is determined that the headset is being worn by a user. These conditions may occur, for example, because the device has been set down (e.g., on a surface or in a bag), while the headset is being worn on the user's head (which will likely move at least a small amount, even when the wearer is relatively still). Under these conditions, because it is likely that the headset is being worn, the voice trigger system 400 maintains an active communication link and monitors the microphone 418 of the headset instead of (or in addition to) the on-device microphone 230. And because this technique focuses on the difference in the motion of the device and the headset, motion that is common to both devices can be canceled out. This may be useful, for example, when a user is using a headset in a moving vehicle, where the device (e.g., a cellular phone) is resting in a cup holder, empty seat, or in the user's pocket, and the headset is worn on the user's head. Once the motion that is common to both devices is cancelled out (e.g., the vehicle's motion), the relative motion of the headset as compared to the device (if any) can be determined in order to determine whether the headset is likely in use (or, whether the headset is not being worn).

Ex-1014 ¶139.

56. Using contextual cues of battery power or power status are frequently another contextual cue or consideration for power management in a speech recognition system for a mobile device. Lebeau is a reference that discusses considering the mobile device's battery. For example, in the car context explained above, Lebeau may determine that the user is currently in a car "based on the device being docked in a car-adapted mobile device dock (e.g., detecting a pin arrangement for a physical electronic connection between the mobile computing device and the dock). The mobile computing device can determine whether to

monitor for voice input based on this current context.” Ex-1016 ¶22; *see also id.*

¶28. Similarly, Lebeau teaches that the mobile computing device will continue monitoring for voice input until the condition of the “battery for the mobile computing device running low on stored power (e.g., below 25% charge remaining in the battery).” *Id.* ¶24. Lebeau also teaches that the mobile device can be receiving power from a power cord that is plugged into an outlet. Ex-1016 ¶30. This contextual cue allows the device to determine whether the “potential convenience to the user of monitoring for voice input is outweighed by the inconvenience of the mobile computing device ... potentially draining the battery in a relatively short period of time.” Ex-1016 ¶31.

57. Battery status is also discussed as consideration in a speech recognition system disclosed in U.S. Patent Application Publication No. 2014/0163978 (“Basye,” Ex-1019). Basye explains that by keeping the computing device in a persistent active state, the first computing device may consume unacceptable amounts of energy to maintain the persistently active state. Ex-1019 ¶9. “Such energy demands may prove especially problematic for mobile computing devices that rely on battery power.” *Id.* Basye is thus directed to power management for speech recognition. *Id.* ¶10. Specifically Basye notes that a device may be provided power wirelessly or via a portable power supply, or may be powered using a wired power supply. *Id.* ¶43 The power supply unit communicates

“the level of power that it can supply to the power management subsystem 100 (e.g., a percentage of battery life remaining, whether the power supply 218 is plugged into an electrical outlet, etc.).” *Id.* Basye’s “on-device recognition selection element 514 may be automatically deselected (and on-device speech recognition capabilities automatically disabled) if the power supply 218 drops below a threshold power supply level (e.g., a battery charge percentage), as on-device speech recognition capabilities as implemented by the processing unit 202 and/or the application processing module 112 may require a relatively large power draw.” *Id.* ¶79. Basye teaches that its implementation is “particularly advantageous for computing devices that rely on battery power, all computing devices for which power management may be desirable can benefit from the principles of the present disclosure.” *Id.* ¶14.

58. U.S. Patent Application Publication No. 20130006638 to Lindalh (“Lindalh,” Ex-1029) is another reference that emphasizes the usefulness of context cues in speech processing. Lindalh recognizes that portable electronic devices generally have limited processing power and are not always activity connected to remote databases and services of interest and are often not “contextually aware,” which make it difficult to use for voice-based control functions. Ex-1029 ¶5. Lindalh’s user device is used to “receive voice commands” and if the device “has sufficient processing power, the voice commands may be

partly or fully processed by the user device 12 (e.g., using a speech recognition engine...)." Ex-1029 ¶26. If the device does not have enough processing power, the voice commands are transmitted to a computer for further processing. *Id.* For example, Lindalh provides the voice command of "find Italian restaurants near me," and the device may transmit the voice command and contextual information to a computer due to "an inability to perform adequate speech processing due to a lack of available processing power." Ex-1029 ¶36.

59. Thus, using context cues associated with the user or the device, allows for the device to improve the processing of the voice input, which leads to power savings.

B. Noise reduction and/or cancellation techniques were well known in order to better process speech input and speech recognition.

60. A POSITA in speech processing and recognition would have understood and applied noise reduction techniques well before 2013. For example, U.S. Patent Application Publication No. 20030179888A1 to Burnett et al. ("Burnett," Ex-1026), explains that many noise suppression algorithms and techniques have been developed over the years and most of the noise suppression systems in use today for speech communication systems are based on a single-microphone spectral subtraction technique first developed in the 1970's and described in Ex-1026 ¶4 (citing S. F. Boll in "Suppression of Acoustic Noise in

Speech using Spectral Subtraction,” IEEE Trans. on ASSP, pp. 113-120, 1979).

Burnett further explains a “Pathfinder system,” which is a digital signal processing (DSP) based acoustic noise suppression system, which can use voice activity detection information and received acoustic information to reduce or eliminate noise in desired acoustic signals “by estimating the noise waveform and subtracting it from a signal including both speech and noise.” Ex-1026 ¶¶41; *see also id.*, ¶¶ 175, 193, Cl. 18.

61. Another noise suppression method that subtracts the background noise from a speech input is U.S. Patent Application Publication No. 20130332165 to Beckley et al. (“Beckley,” Ex-1024), explains how POSITAs were focused on improving speech recognition systems by reducing the background noise. Specifically, Beckley discusses how speech recognition systems, especially those with “always-on” speech recognition (referred to by Beckley as “ASR”), “have difficulty handling ambient noise, such as background conversation or other undesired noise.” Ex-1024 ¶6. Beckley explains that the “background conversation, for instance, may cause the system to recognize a command that was not intended by the user of the system, thus leading to a number of false positives and misrecognitions. For example, in a car environment, while the driver might be the issuer of voice commands to control the ASR system, the presence of passengers and ensuing conversation between them can greatly reduce the

performance of the ASR system as their conversations may lead to commands that are false positives or misrecognitions.” *Id.*

62. Beckley seeks to improve speech recognition by: (i) obtaining a voice sample from a speaker; (ii) storing the voice sample of the speaker as a voice model in a voice model database; (iii) identifying an area from which sound matching the voice model for the speaker is coming; and (iv) providing one or more audio signals corresponding to sound received from the identified area to the speech recognition system for processing. *Id.*, ¶7. Beckley further includes separating ambient noise from speech in the one or more audio signals and providing the resulting speech signal to the speech recognition system for processing. *Id.*, ¶11. Specifically, Beckley receives a digital audio signal and may generate a “voice/speech reference signal” and a “noise reference signal.” *Id.*, ¶88. Beckley’s noise reference refiner 220a may “reduce or effectively eliminate the residual desired audio from the noise reference signal 218 in order to generate a refined noise reference signal 222 a.” *Id.* Beckley provides an example that “the refined noise reference signal 222 a may be filtered and **subtracted** from the desired audio reference signal 216 in order to reduce noise in the desired audio.” *Id.* Beckley further explains that “sound from other areas than that of the area in which the identified speaker is located may be subtracted from sounds of the speaker’s area.” *Id.* ¶95. The resulting speech signal is then provided to the speech

recognition system. *Id.* Thus, Beckley improves the quality of the speech input by removing the undesired audio, such as ambient noise.

63. Similarly, U.S. Patent Application Publication No. 20120296655A1 to Kristjansson et al. (“Kristjansson”) discusses the advantages of using “stored noise profiles.” Specifically, Kristjansson teaches that:

“speech recognition is used to transcribe utterances that may have been input in a wide variety of noisy environments, such as when the user is in a crowded room (e.g., a cafeteria), walking down the street, or in the presence of a radio or television. To filter this noise, some noise reduction algorithms require a sample of audio from the environment, without the user's speech. ***By predictively pre-recording audio before the use initiates the voice input (e.g., before the user presses a button), such a background recording becomes available for use by the noise reduction algorithm, thereby improving recognition accuracy.***

Ex-1018 ¶42.

64. Kristjansson further explains that by “invoking the functionalities of the mobile device may include commencing a background audio recording of audio prior to a user's button press. The background audio may be used as input to the speech recognition engine, ***to estimate noise models for noise suppression, and/or estimate noise models and/or levels for speech endpoint detection.***” Ex-1018 ¶57.

65. Similarly, U.S. Patent Application Publication No. 20140350927 to Yamabe et al. (“Yamabe,” Ex-1025) explains how in the field of speech recognition, noise-signal suppression processing is performed prior to speech

recognition processing, thereby suppressing unnecessary noise other than speech signals and improving the accuracy of speech recognition. Ex-1025 ¶5. Yamabe notes that typical noise reduction systems use noise spectrum detection, which involves estimating noise spectra and “subtract[ing] a noise component from an input signal including noise, thereby reducing the noise.” Ex-1025 ¶6. Yamabe recognizes that such spectrum subtraction methods uses “stationary energy, such as white noise and air conditioners, as noise with relative ease,” and have a high suppression effect but a signal that is closer to a tone signal might be difficult to classify. Ex-1025 ¶7. Similarly, for sounds that have “a harmonic structure which is a characteristic of speech” it is difficult to estimate a noise spectrum. Ex-1025 ¶7. Yamabe provides an improvement to spectrum subtraction methods, by “extracting a peak spectrum” to be suppressed from the noise signal by “subtracting a signal in the time domain obtained by the frequency/time converting unit 207 converting the extracted peak spectrum having independence from the sound signals in the time domain collected by the sound collecting unit 201.” Ex-1025 ¶132. Yamabe thus discusses how noise reduction systems commonly used “spectrum subtraction” to reduce the noise in a speech signal.

66. U.S. Patent Application Publication 2020/0279576 to Binder et al. (“Binder,” Ex-1014) is another example of noise cancellation. For example, if a voice trigger includes noise, Binder implements a “sound cancellation

functionality.” Ex-1014 ¶124. Binder explains “noise cancellation” that is performed by the audio digital signal processor, that “removes unwanted noise or sounds from the sound input prior to it being processed by the sound detectors.” *Id.* Binder gives an example that sometimes the unwanted noise is background noise from the user’s environment or just any “sounds above, below, or at predetermined amplitudes or frequencies.” *Id.* For example, sounds above the typical human vocal range (e.g., 3,000 Hz) is filtered out or removed from the signal. *Id.* Binder explains that this process allows the sound input to be cleaner and more accurate. *Id.*, ¶126.

67. U.S. Patent No. 8,694,316 to Pickering et al. (“Pickering,” Ex-1011) describes a system with “an acoustic front end (AFE) for receiving an audio signal from an input device, removing noise, identifying the parts of the audio signal containing speech, and parameterizing sections of the identified speech.” Ex-1011, 4:3-16. Specifically, the AFE is “responsible for: 1. pre-processing the audio signal to remove noise; 2. establishing which part of the audio signal contains valid speech; and 3. parameterizing successive sections of the signal to be matched by the recognizer.” *Id.*, 6:6-10.

68. Noise cancellation and noise filtering is also noted in U.S. Patent Application Publication No. 2014/0006825 (“Shenhav,” Ex-1015). Shenhav teaches a mobile device with an audio sensor module that may include a “variety

of noise reducing or noise cancellation filters.” Ex-1015 ¶40. Shenhav provides “removing background noise” as a filter used to generate a “filtered sound include[s] processing the sound input to only include those portions of the sound input that match audio frequencies associated with human speech.” Ex-1015 ¶49.

69. Thus, as shown by at least these references, POSITAs were implementing noise reduction and/or cancelling techniques before 2013, in order to focus the sound input on the speech component.

C. Voice activity detection was well known to detect human speech.

70. As I explained above, voice activity detection (VAD) methods were well known to use in “signal processing systems to denoise acoustic signals.” Ex-1026, Abstract. Burnett describes a system that provides that “the VAD signal is processed independently of the noise suppression system, so that the receipt and processing of VAD information is independent from the processing associated with the noise suppression, but the embodiments are not so limited.” Ex-1026 ¶32. Burnett generally defines “VAD” as “a vector or array signal, data, or information that in some manner represents the occurrence of speech in the digital or analog domain.” Ex-1026 ¶34. Burnett’s invention is used with noise suppression systems like the Pathfinder Noise Suppression System, as described above. Burnett

describes a VAD technique which includes evaluating “spectral information.” Ex-1026 ¶55.

71. Shenhav also describes using “spectral analysis of the sound signals.” Cl. 8. Specifically, Shenhav teaches comparing a spectral representation of the received sound to spectral representation of a predetermined wake-up phrase. Ex-1015 ¶¶38, 45. To do so, Shenhav applies a filter that may include “spectral equalization, or the like” to the sound signal, isolating a filtered “sound input that matches audio frequencies associated with human speech.” Ex-1015 ¶¶48-49. Shenhav therefore teaches using spectral analysis on the acoustic input to evaluate whether a spectrum of the acoustic input is indicative of voice activity (i.e. “human speech”), as opposed to other non-speech activity. After “only those portions of the sound input ... associated with human speech” are prepared, the filtered signal is further processed, to determine whether there is “a threshold probability that there is a match of the sound to a wake-up phrase.” Ex-1015 ¶50.

72. Basye also teaches determining whether audio input includes speech. Basye’s speech detection module applies “voice activity detection...based on various quantitative aspects of the audio input, such as the spectral slope between one or more frames of the audio input; the energy levels of the audio input in one or more spectral bands; the signal-to-noise ratios of the audio input in one or more spectral bands; or other quantitative aspects.” Ex-1019 ¶21.

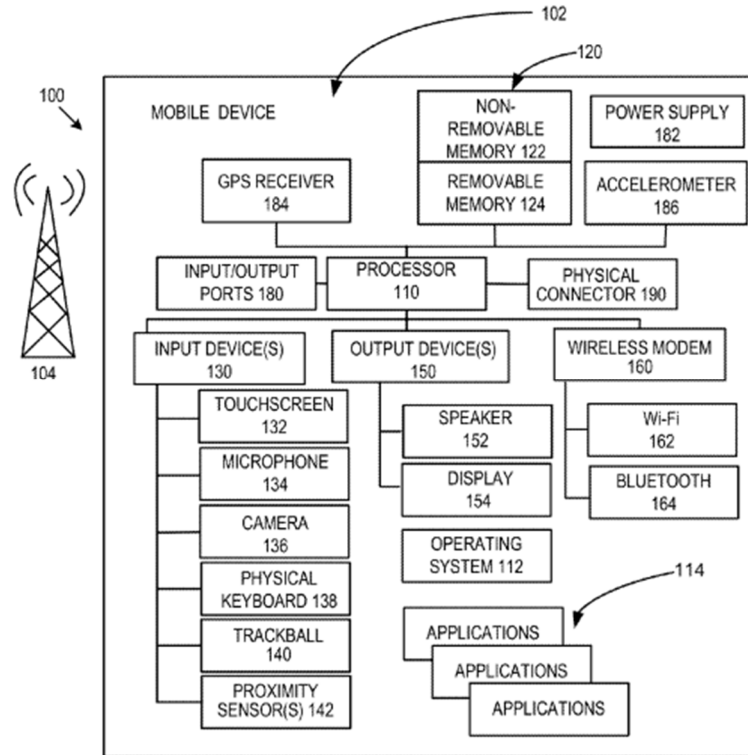
73. Thus, voice activity detection techniques were well known to use spectral analysis to detect human speech before 2013.

VI. BRIEF DESCRIPTION OF THE APPLIED PRIOR ART

74. The prior art references analyzed in the grounds below are analogous to the '461 patent because they are within the same field of endeavor as the '461 patent and reasonably pertinent to one or more problems addressed by the '461 patent. For example, like the '461 patent, Sarin discloses a mobile device that detects and recognizes voice commands using contextual cues. Hwang discloses a mobile device implementation that receives and analyzes audio signals while the device is in an idle mode. Gruber discloses a virtual assistant that may be implemented on a mobile device, such as the ones disclosed by Sarin and Hwang, and which uses contextual information to supplement speech and natural language processing. Lebeau teaches uses battery power as contextual cue, such as the contextual cues used in Sarin.

A. Sarin (Ex-1005)

75. Sarin discloses a mobile device with a camera that is capable of capturing images based on voice commands.



Sarin, FIG. 1

76. Sarin’s mobile device is capable of operating in a “low-power” state and an “active” state. Ex-1005, 1:43-44. According to Sarin, “in a low-power state, the display is typically dimmed or powered off, and other components (e.g., hardware buttons, keypads, trackballs) can be deactivated, as well.” *Id.*, 9:63-65. In an active state, by contrast, the mobile device’s display is typically powered on, and the device’s operating system may be started. *Id.*, 9:46-49; 10:53-55.

77. Sarin’s mobile device receives and recognizes image capture-related voice commands while the device is operating in a low-power state. Ex-1005, 13:38-46. Specifically, if the device determines that a received voice command matches an expected command associated with an image capture application, the

device transitions from a low-power state to an active state. *Id.*, 13:53-57, 13:57-

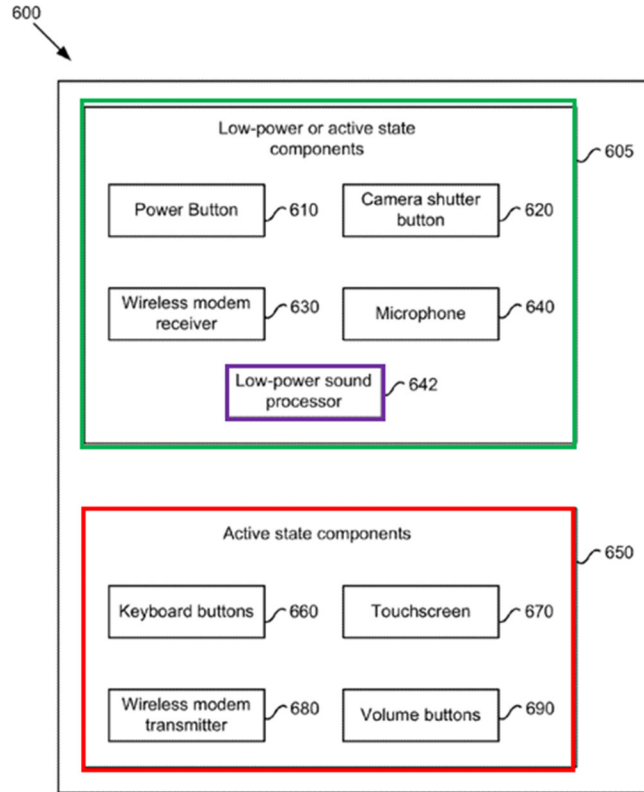
62. Once in the active state, the mobile device may activate an image capture application and cause the mobile device's camera to capture digital images. *Id.*, 13:62-14:6; FIG. 10.

78. As shown in figure 1, Sarin's mobile device includes a primary processor 110,² which performs "such tasks as signal coding, data processing, input/output processing, power control, and/or other functions." *Id.*, 4:15-20. To enable voice command recognition while in a low-power state, Sarin's mobile device utilizes a low-power sound processor, which is functional whether the mobile device is in the low-power state or the active state. *Id.*, 11:10-15; 11:22-25

79. Sarin's figure 6 depicts components of Sarin's mobile device that are functional in both low-power and active states (highlighted in **green**) and those components that are only functional in the active state (highlighted in **red**). The low-power sound processor is highlighted in **purple**.³

² Sarin also teaches that its mobile device may have more than one main processor. Ex-1005, 6:19-22.

³ I have added all emphases/annotations/color, unless noted otherwise.



Ex-1005, FIG. 6 (annotated)

80. Sarin also discloses that its mobile device can recognize voice commands when it is in an active state, such as when a voice controlled image capture tool is already running in the foreground. *Id.*, 10:3-7; 10:22-26. In this scenario, voice command recognition can be performed “at a remote server; on a general purpose processor, or on a dedicated special-purpose processor.” *Id.*, 13:6-8.

81. To facilitate the recognition of image capture-related voice commands in a power efficient manner, Sarin, like the '461 patent, uses contextual cues. *Id.*, 11:32-34. According to Sarin, contextual cues can be used to disambiguate voice

commands that are applicable to different functions which may be invoked with the same word or phrase. Ex-1005, 9:9-13. Exemplary contextual cues include “which applications or functions are active on a device, time of day, and location.” *Id.*, 9:42-43.

82. Sarin’s mobile device is not limited to recognizing image capture-related voice commands only. Sarin discloses that, “in addition to image capture operations, components for speech recognition also can be used for other applications and functionality such as voice calls ... and search functions (e.g., to perform searches on a local device, a remote server, or the Internet).” *Id.*, 6:39-44. As such, Sarin discloses that contextual cues can be used to recognize voice commands relating to different functions. *Id.*, 9:9-13. Furthermore, Sarin discloses that its speech recognition techniques can be implemented as stand-alone applications, plug-ins, or using third party services, such as Microsoft’s Bing®, or Apple Inc.’s Siri.” *Id.*, 6:44-49.

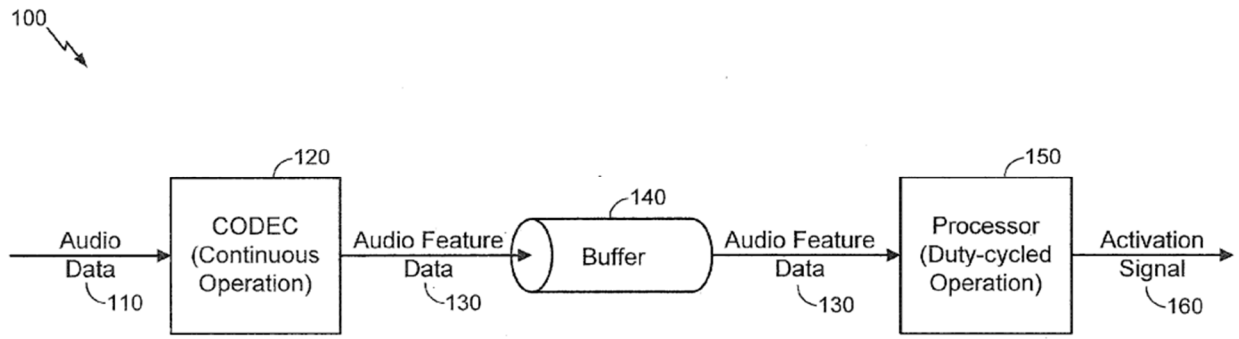
B. Hwang (Ex-1006)

83. Hwang’s disclosure focuses on techniques for processing audio in an electronic device (such as a wireless telephone) with reduced power consumption. Ex-1006 ¶¶5, 26. As such, Hwang discloses a low-power system for receiving audio which extracts and analyzes feature data from the received audio, and

activates components of the electronic device based on the analysis. Ex-1006 ¶5.

One such component may be an application processor of a mobile device. *Id.* ¶26.

84. An example of Hwang's system is shown below in figure 1.



Hwang, FIG. 1

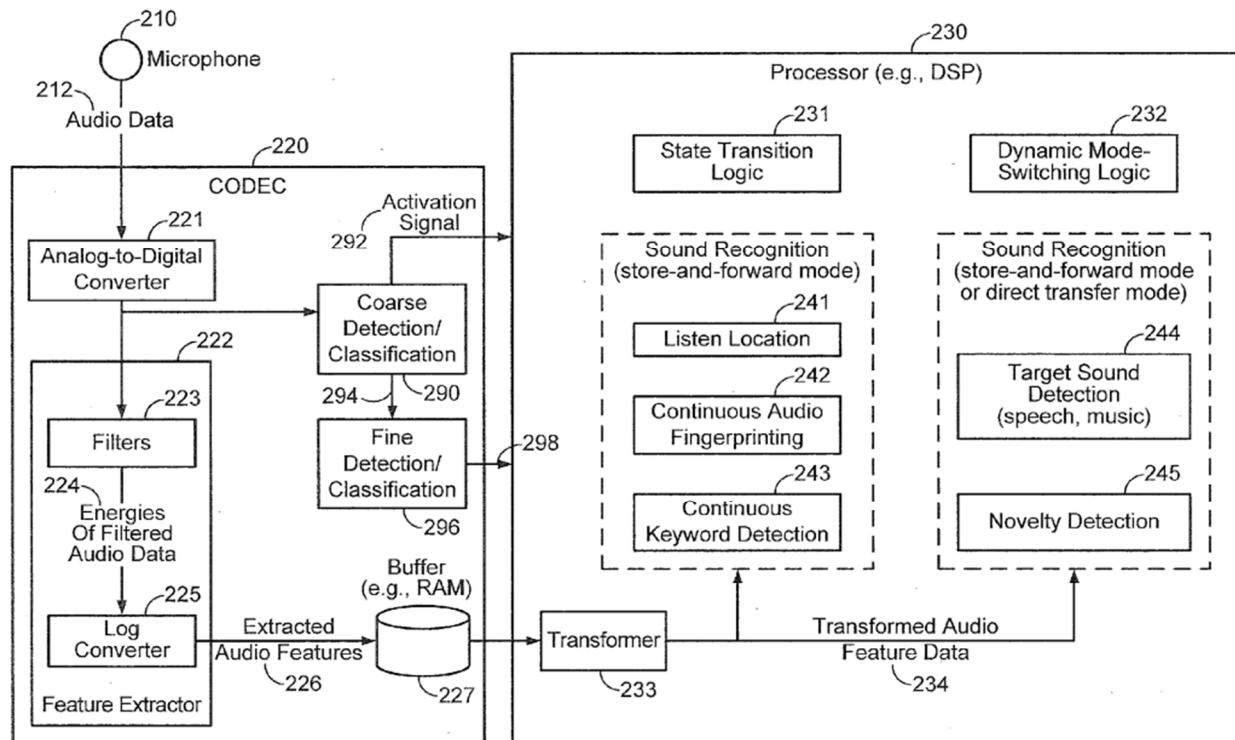
As shown, the main components of Hwang's system are a coder/decoder (CODEC) 120 and a digital signal processor (DSP) 150.

85. CODEC 120 receives "raw" audio data 110 and extracts audio features from the audio data, which it stores in buffer 140. Ex-1006 ¶24. The extracted audio feature data may indicate particular audio characteristics, "such as pitch, tone, volume, and/or rhythmic characteristics." *Id.*

86. DSP 150 retrieves and analyzes the audio feature data stored buffer 140. Ex-1006 ¶26. Based on its analysis of the audio feature data, the DSP may generate an activation signal 160 to activate various components of an electronic device. *Id.* ¶31. For example, DSP 150 may generate the activation signal when

“the analysis of the audio feature data 130 identifies a particular voice input command (e.g., ‘wake up’).” *Id.*

87. Figure 2 is a more detailed illustration of Hwang’s CODEC/DSP system.



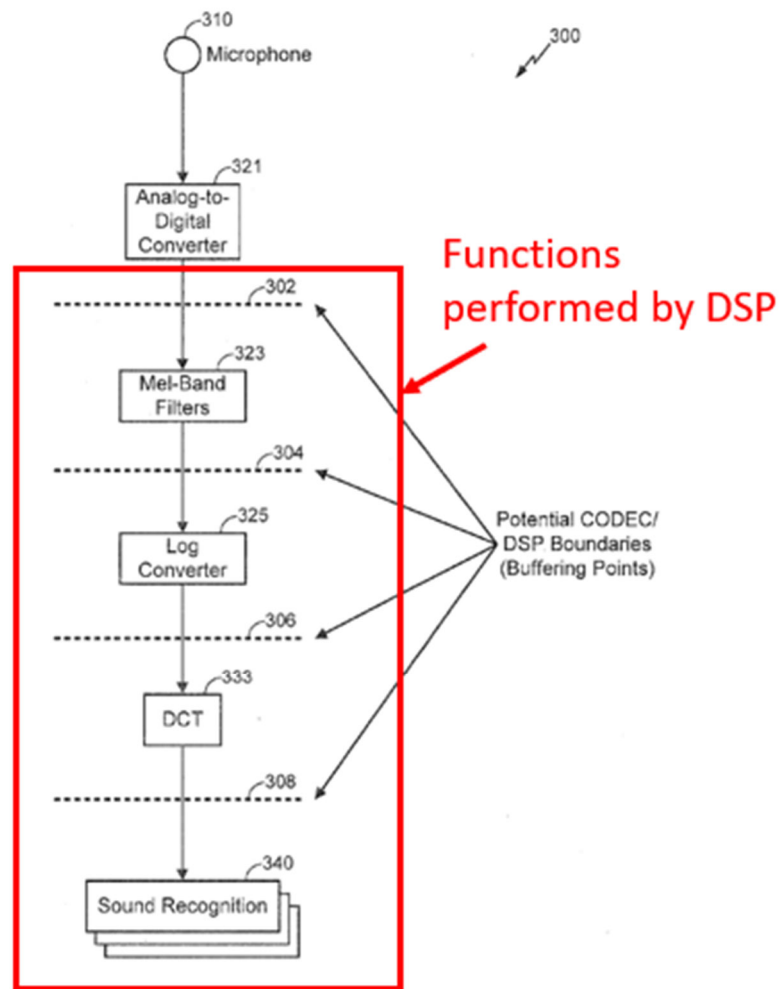
Hwang, FIG. 2

88. CODEC 220 and DSP 230 in figure 2 correspond to CODEC 120 and DSP 150 shown in figure 1. *Id.* ¶33. DSP 230 includes sound recognition modules (241-245) that analyze the audio feature data extracted by CODEC 220. *Id.* ¶38. For example, continuous keyword detection module 243 may “detect the existence of pre-enrolled (e.g., predetermined) keyword sets,” thereby enabling

“[c]ontinuous keyword detection ... in a relatively low-power state,” and which “may activate predefined applications based on detected keywords.” *Id.* ¶41.

89. Further, Hwang’s DSP “may determine whether to activate an application processor and/or component(s) of a mobile station modem (MSM) or other component based on the analysis” of its sound recognition modules. *Id.* ¶44. Hwang’s DSP may also provide “higher-level applications” with the results of its sound recognition operations via a shared memory or an application programming interface. *Id.* ¶45.

90. Hwang also discloses that the audio processing functions performed by its CODEC (e.g., audio feature extraction) may be consolidated into and performed by the DSP, as illustrated below in figure 3. *Id.* ¶52.



C. Garner (Ex-1007)

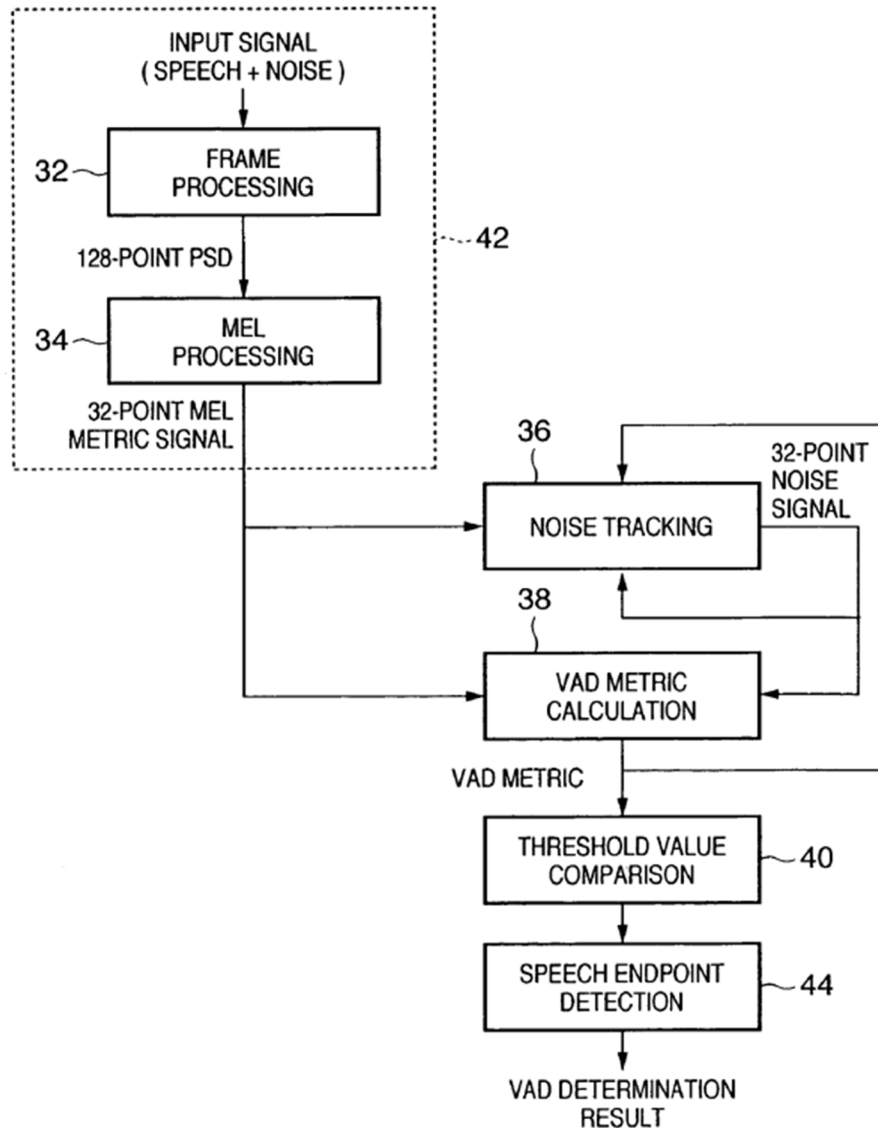
91. Garner discloses a signal detection technique that is designed to be robust against various types of background noise. Ex-1007, 2:65-67. To achieve this, *Garner* discloses a process that receives an acoustic signal that contains both speech and background noise, samples the signal using an A/D converter, and divides the signal into frames containing a number of samples. *Id.*, 5:55-60. Each frame undergoes additional processing to form a corresponding PSD (Power

Spectral Density). *Id.*, 5:63-6:1. For example, for a frame of 256 samples, the PSD has 129 bins. *Id.*, 6:2.

92. Each PSD is then processed by a mel processing module, which generates mel metric signals that are used as feature amounts for voice activity detection (VAD). *Id.*, 6:3-12. The mel metric signal is input to both a noise tracking module and a VAD metric calculation module. *Id.*, 6:25-26.

93. The noise tracking module tracks background noise that gradually varies with the mel metric signal. *Id.*, 6:26-28. The VAD metric calculation module computes a likelihood ratio, in which “the numerator represents a likelihood which indicates probability that [a] frame *f* is detected as speech, and the denominator represents a likelihood which indicates probability that the frame *f* is detected as noise.” *Id.*, 6:42-46. The likelihood ratio is compared with a predetermined threshold value to generate binary data, which is a value indicating speech or noise. *Id.*, 11:24-26.

94. Garner’s signal detection process is depicted in figure 4:



Ex-1007, Fig. 4

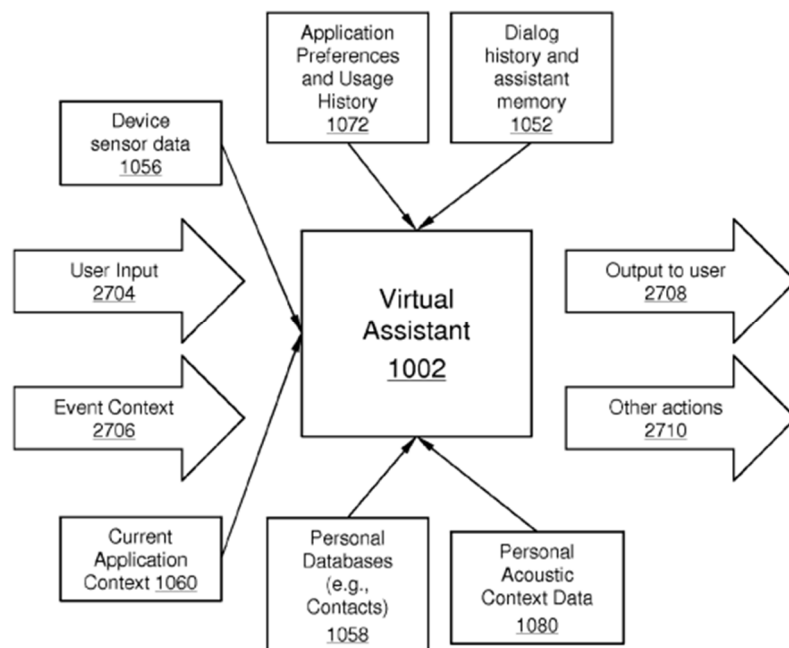
D. Gruber (Ex-1008)

95. Gruber is directed to a “virtual assistant” which uses “context information” in order to supplement natural language input from a user. Ex-1008, Abstract. According to Gruber, its virtual assistant “allows users to interact with a device or system using natural language, in spoken and/or text forms.” *Id.* ¶11. The

virtual assistant “interprets user inputs, operationalizes the user’s intent into tasks and parameters to those tasks, executes services to support those tasks, and produces output that is intelligible to the user.” *Id.*

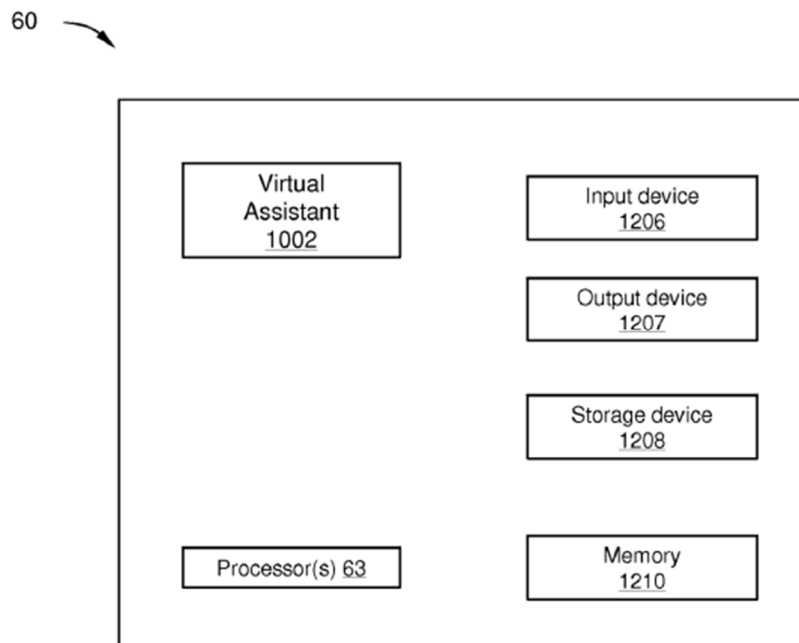
96. Gruber’s use of context information “helps to clarify the user’s intent and to reduce the number of candidate interpretations of the user’s input, ... reduc[ing] the need for the user to provide excessive clarification input.” *Id.* ¶13. Gruber discloses that “[c]ontext can include any available information that is usable by the assistant to supplement explicit user input to constrain an information-processing problem and/or to personalize results.” *Id.*

97. Figure 1 depicts a block diagram of the virtual assistant, along with exemplary sources of context that can influence that assistant’s operation. *Id.* ¶178.



98. Different categories of context information that Gruber's assistant uses in processing speech input include: applications currently in use on a device (*id.* ¶180); information stored on a device (*id.* ¶193); a user's dialog history with the virtual assistant (*id.* ¶203); sensor data that reports the user's current location, the local time, the position, orientation, and motion of the device, and environmental data (*id.* ¶218); and events, such as incoming texts, emails, phone calls, and alerts (*id.* ¶238).

99. As shown in figure 29, Gruber's virtual assistant can be implemented on a standalone system, such as a mobile phone or smartphone. *Id.* ¶74.



Gruber, FIG. 29

The virtual assistant can also be implemented in a distributed fashion, in which “any number of clients 1304 are provided; each client 1304 may run software for implementing client-side portions of the present invention” and “any number of servers 1340 can be provided for handling requests received from clients 1304.” *Id.* ¶85.

E. Lebeau (Ex-1016)

100. Lebeau teaches a mobile device that operates in two modes of operation. Ex-1016, Abstract. Lebeau determines to “switch” the device to allow for voice input that indicates a request to perform an operation. Ex-1016 ¶9. Lebeau’s device attempts to infer whether, based on the current context, monitoring for voice input would provide at least a threshold level of convenience for the user and, if so, monitoring for voice input. Ex-1016 ¶23. Lebeau explains that a context includes whether the mobile device’s battery is “running low on stored power (e.g., below 25% charge remaining in the battery)” (Ex-1016 ¶24) or whether the mobile device is receiving power from a power cord. Ex-1016 ¶30.

VII. CLAIM CONSTRUCTION

101. I have given all the claim terms their plain and ordinary meaning, as would have been understood by a person having ordinary skill in the art at the time of the alleged invention, which I understand is March 12, 2013, having taken into

consideration the language of the claims, the specification, and the prosecution history of record.

VIII. DETAILED EXPLANATION OF GROUNDS

A. Ground 1: Sarin in Combination with Hwang Renders Obvious Claims 1-4, 6, 7, 11, 14, 16, and 18

1. A POSITA would have been motivated to combine Sarin's and Hwang's teachings and would have had a reasonable expectation of success.

102. A POSITA would have been motivated to combine Sarin and Hwang (“Sarin-Hwang”) in the manner discussed in the Grounds presented here, for the reasons discussed in this section, and for the additional reasons specifically discussed below.

103. A POSITA also would have been motivated to combine Sarin with Hwang because they are analogous art and teach improvements to the same type of devices. Sarin teaches a mobile device that uses contextual cues to assist in recognizing voice commands (such as image capture commands), which reduces power consumption and can extend battery life. Ex-1005, 11:32-34. Another way that Sarin reduces power consumption is through its use of a low-power sound processor. *See id.*, 10:65-11:1. Sarin's low-power sound processor is functional when the mobile device is in a low-power state (*id.*, 11:9-19), which Sarin discloses is a state in which “the display is typically dimmed or powered off, and

other components (e.g., hardware buttons, keypads, trackballs) can be deactivated[.]” *Id.*, 9:62-65. Sarin’s disclosure of its low-power sound processor, however, is high-level, and lacks details as to how the processor is implemented and how it functions. As such, a POSITA would have been motivated to look to references that include details for implementing a low-power sound processor that meets Sarin’s goal of reducing power consumption. In doing so, a POSITA would have readily found Hwang.

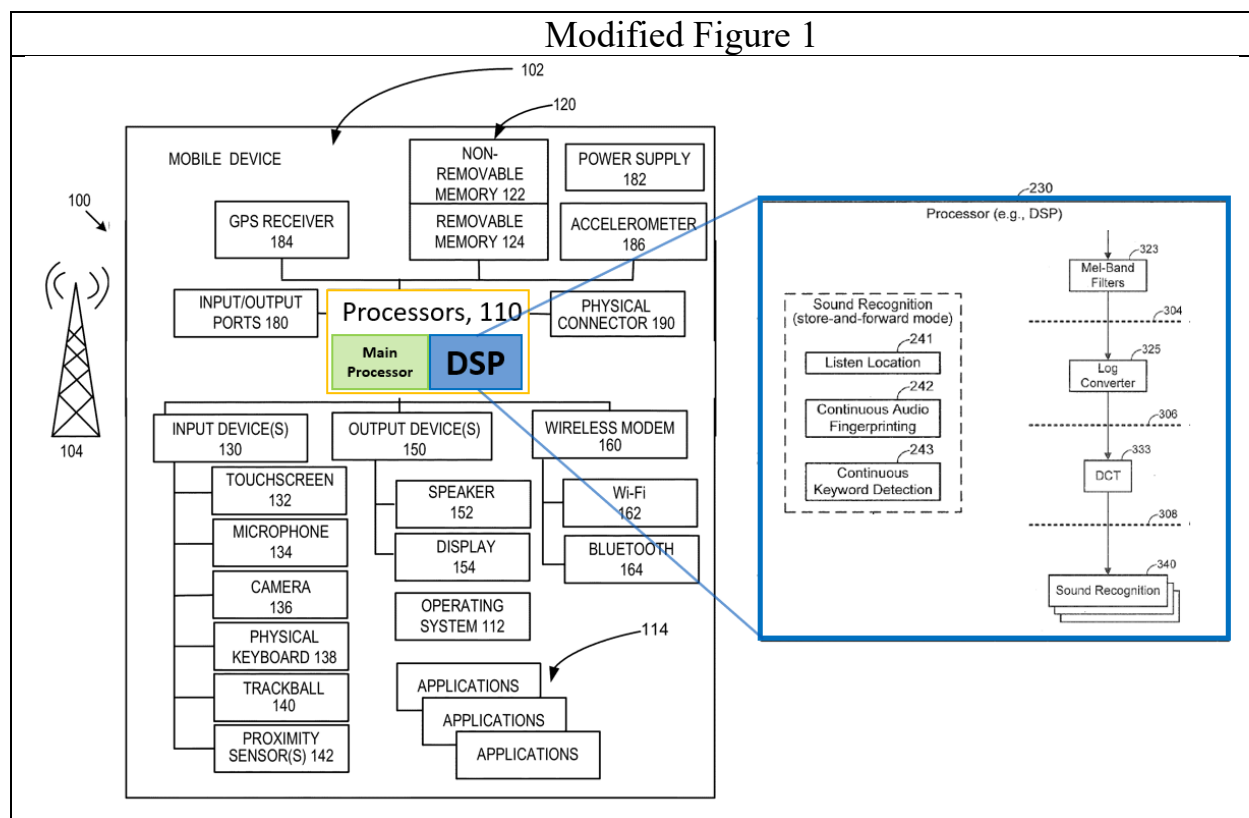
104. Hwang recognizes that conventional techniques for detecting voice commands on mobile devices require continuous power consumption, resulting in decreased battery life. Ex-1006 ¶4. To address this problem, Hwang discloses a system for detecting voice commands in audio in a way that reduces power consumption. *Id.* ¶5. Hwang achieves reduced power consumption by incorporating a DSP into a mobile device, where the DSP receives and processes audio input while other mobile device components (such as an application processor) are not activated. *Id.* ¶26. Thus, Hwang teaches techniques for processing audio in an electronic device (such as a wireless telephone) with reduced power consumption. Ex-1006 ¶¶5, 26. For example, Hwang discloses a low-power system for receiving audio which extracts and analyzes feature data from the received audio, and activates components of the electronic device based on the analysis. Ex-1006 ¶5. Moreover, Hwang discloses details of how audio

feature data is extracted from input audio data. Specifically, Hwang discloses that its feature extractor “may include a plurality of filters 223 that filter the audio data 212 to generate energies 224 (e.g., mel-band energies) of filtered audio data.” Ex-1006 ¶35. Therefore, Sarin and Hwang both listen and perform sound recognition with a goal of reducing power consumption.

105. Hwang further offers additional benefits to Sarin’s mobile device. First, while Sarin’s sound processor is functional when the mobile device is in a low-power state (Ex-1005, 11:9-19), where “the display is typically dimmed or powered off, and other components (e.g., hardware buttons, keypads, trackballs) can be deactivated” (*id.*, 9:62-65), Sarin leaves the implementation details of its low-power sound processor to a POSITA. A POSITA would have been motivated to look to references that include details for implementing a low-power sound processor that meets Sarin’s goal of reducing power consumption, such as Hwang. Hwang recognizes that conventional techniques for detecting voice commands on mobile devices require continuous power consumption, resulting in decreased battery life. Ex-1006 ¶4. To address this problem, Hwang detects voice commands in audio in a way that reduces power consumption (Ex-1006 ¶5) by using a DSP

that receives and processes audio input while other mobile device components (such as an application processor) are not activated. *Id.* ¶26.⁴

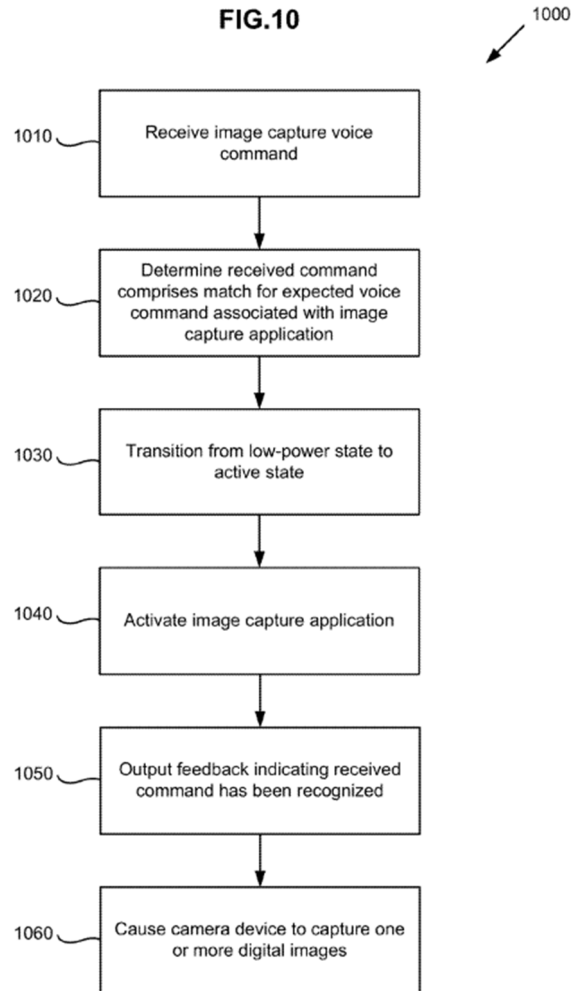
106. A POSITA would have thus been motivated to implement Sarin's low-power sound processor in processor 110 using Hwang's DSP, which can include more than one additional processor such as Sarin's main processor and low-power sound processor, as shown in modified Figure 1 of Sarin:



107. A POSITA would have recognized that Hwang's DSP is compatible with and meets the functional requirements of Sarin's low-power sound processor.

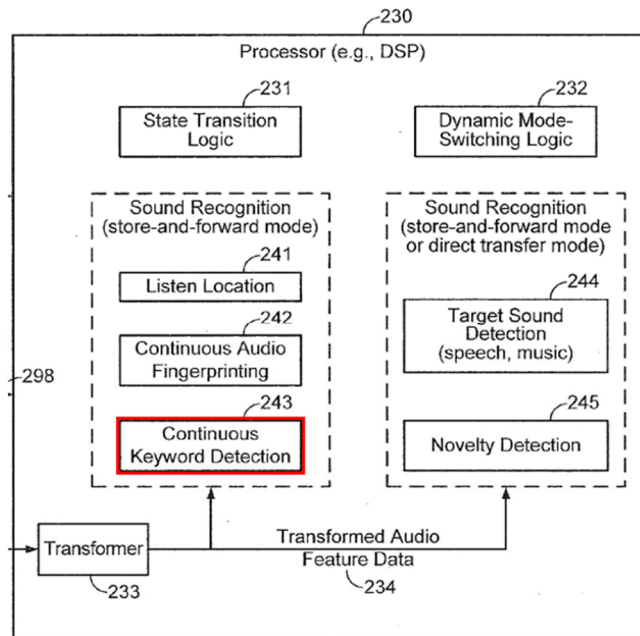
⁴As discussed in §V.B, Hwang discloses a system that includes a low-power CODEC and DSP. As noted, however, Hwang also discloses that the audio processing functions of the CODEC and DSP can be consolidated in one DSP. Ex-1006 ¶52, FIG. 3.

Sarin discloses a technique for low-power state voice command recognition, focusing on voice commands associated with image capture:



108. A POSITA would have understood that Hwang’s DSP is capable of receiving and recognizing the type of image capture-related commands that Sarin’s low-power sound processor would need to recognize. For example, as highlighted below, Hwang’s DSP includes a “continuous keyword detection module,” which operates “in a relatively low-power state and may activate predefined applications” based on “predetermined keywords.” Ex-1006 ¶41. A POSITA would have

recognized that the “keyword sets” that Hwang’s continuous keyword detection module detects are similar to Sarin’s image capture-related voice commands and that Sarin’s image capture application is an example of a predefined application that Hwang’s DSP would be capable of activating after detecting image capture related commands. Ex-1005, 13:62-63 (after transitioning to an active state, at step 1040 “the computing device activates the image capture application.”).



Hwang, FIG. 2 (excerpted and annotated)

109. Based on Hwang’s teachings, a POSITA would have been motivated to incorporate Hwang’s teachings of its DSP into Sarin’s mobile device because Hwang furthers Sarin’s goal of conserving power in a mobile device. Sarin recognizes that, aside from the display, *other components* may be deactivated in a low-power state, and that transitioning out of a low-power state may entail *starting*

the operating system. A POSITA would have recognized that the operating system typically runs on the primary processors and, thus, would have understood that Sarin teaches changing the power state and/or usage of the primary processor. For this additional reason, a POSITA would have been motivated to apply Hwang's express teachings and perform low-power voice command detection while Sarin's main processor is deactivated. A POSITA would have recognized that receiving and processing audio using the low-power DSP while the primary processor is inactive is desirable in a mobile device that runs on a battery, as maintaining the primary processor in a deactivated state until the processor is needed would tend to conserve power usage and extend the battery's life, which is a goal of Sarin. Ex-1005, 11:32-34.

110. Further, a POSITA would have had a reasonable expectation of success in combining Sarin and Hwang because a POSITA would have recognized that Hwang's DSP meets the functional requirements of Sarin's low-power sound processor and could be readily incorporated into Sarin's mobile device. A POSITA would have understood, therefore, that these references disclose interrelated teachings based on routine technologies that would have been amenable to various well-understood and predictable combinations to provide quick and efficient speech recognition and processing.

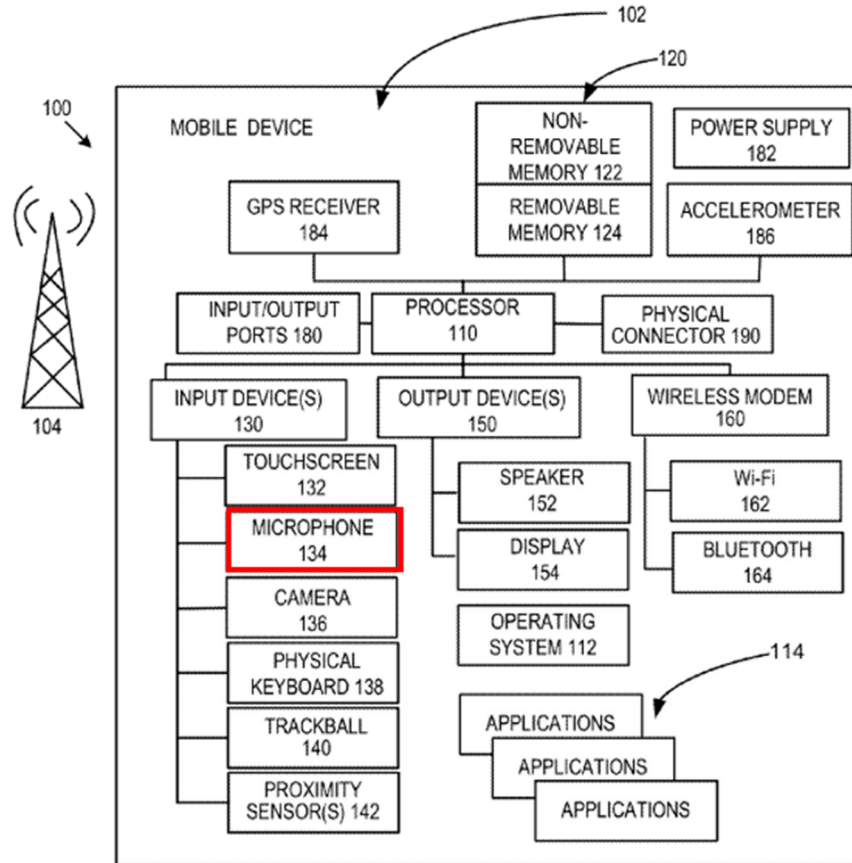
2. Independent Claim 1

a. 1[pre]: A method of monitoring an acoustic environment of a mobile device for voice commands when the mobile device is operating in a low power mode, the method comprising:

111. To the extent the preamble is limiting, Sarin teaches or suggests this element.⁵

112. Sarin teaches monitoring an acoustic environment of a mobile device for voice commands. As shown in Figure 1 below, Sarin teaches a mobile device, such as a cell phone, with a microphone 134. Ex-1005, 4:10-14, 4:50-54. Sarin's microphone allows the user to speak a voice command that cause the mobile device to capture an image. Ex-1005, 7:49-52.

⁵ I understand from counsel, that the term "teaches" includes both express teaching or those fairly suggested to a person of ordinary skill in the art.



Ex-1005, FIG. 1 (annotated)

113. Sarin’s mobile device monitors an acoustic environment for voice commands when the mobile device is operating in a low power mode. For example, Sarin teaches “a mobile phone,” such as Figure 1’s mobile device, that *“receives the image capture voice command while in a low-power state.”* *Id.*, 13:48-53.

114. A POSITA would have understood that Sarin’s mobile device’s “low-power state” is similar to the ’461 patent’s “low power mode.” Sarin describes its “low-power state” as a state in which “the display is typically dimmed or powered

off, and other components (e.g., hardware buttons, keypads, trackballs) can be deactivated, as well.” Ex-1005, 9:62-65. Similarly, the ’461 patent teaches that “[e]ntering a low power mode may include turning off the display, shutting down one or more processors (e.g., disabling the clock tree associated with one or more processors) and/or deactivating components that would otherwise draw power if left operational.” Ex-1001, 5:5-9.

115. As such, in my opinion, Sarin teaches a mobile device that listens for voice commands (“a method of monitoring an acoustic environment of a mobile device for voice commands”) while the mobile device is operating in a low-power state (“when the mobile device is in a low power mode”).

b. 1[a]: receiving acoustic input from the environment of the mobile device while the mobile device is operating in idle mode; and

116. Sarin-Hwang teaches this element.

117. Sarin-Hwang’s mobile device receives acoustic input from its environment while operating in idle mode. As discussed in §VIII.A.1, the Sarin-Hwang’s DSP is functional when the primary processor is not activated.⁶

⁶ Sarin’s mobile device receives acoustic input and interprets voice commands (such as commands relating to image capture), while operating in a low-power state in which “the display is typically dimmed or powered off, and other components (e.g., hardware buttons, keypads, trackballs) can be deactivated, as well.” Ex-1005, 9:63-65. Sarin also teaches that, when its mobile device recognizes a voice command, the device can “transition to an active state, start the operating

Specifically, Hwang achieves reduced power consumption in a mobile device by incorporating a DSP, where the DSP receives and processes audio input while other mobile device components (such as an application processor) are not activated. *Id.* ¶26. Hwang’s DSP includes a “continuous keyword detection module,” which, “may receive sound input and may detect the existence of *pre-enrolled (e.g., predetermined) keyword sets. Continuous keyword detection may be performed in a relatively low-power state and may activate predefined applications based on detected keywords.*” Ex-1006 ¶41. Hwang teaches that its DSP includes sound recognition modules that “analyze ... transformed audio feature data” and, based on the analysis, “may determine whether to *activate an application processor* and/or component(s) of a mobile station modem (MSM).” Ex-1006 ¶44. In mobile devices, an “application processor” typically refers to the device’s primary (or main) processor. For example U.S. Patent No. 8,396,463 to Marcellino teaches a mobile device that includes an application processor, which is “the central processor of the mobile device 160” so that it “processes the operating system(s) of the mobile device 160... and any applications running on the mobile device 160.” Ex-1009, 5:53-62. Similarly, U.S. Patent No. 8,331,315 (“Purkayastha”) teaches how cell phones include an “application processor” that is

system, display a GUI for the tool, and generate command events to cause one or more digital images to be captured by an image capture device.” *Id.*, 10:52-57. Thus, Sarin’s sound processor is functional when the mobile device is in a low-power state. *Id.*, 11:9-19.

“configured to run an OS 106A, high level applications 150A (such as a user interface (UI) or NCC), network control components 160A (such as MIP, transmission control protocol (TCP), user datagram protocol (UDP), internet protocol (IP)), and the MIH client 130A.” Ex-1010, 3:60-4:4.

118. Accordingly, a POSITA would have understood Hwang teaches detecting voice commands while a mobile device’s primary processor is not activated, in order to reduce power consumption and extend battery life.

119. When Sarin-Hwang’s primary processor is not activated, the device consumes less power and is in an idle mode. The ’461 patent states that, “when a mobile device enters a low power mode (*e.g., powers down into a sleep or idle mode*), *the primary processor may be powered down* to a desired extent.” Ex-1001, 26:46-48. Based on this disclosure, “idle mode,” as used in the ’461 patent, encompasses using less power, including by powering down a primary processor.

120. As discussed above in Section VIII.A.1, a POSITA would have been motivated to detect audio commands while a mobile device’s primary processor is not activated, as taught in Hwang, in Sarin because it would have been desirable to reduce power consumption and extend battery life, furthering Sarin’s goals.

121. As such, in my opinion, Sarin-Hwang thus teaches a mobile device that receives audio (“receives acoustic input from the environment of the mobile

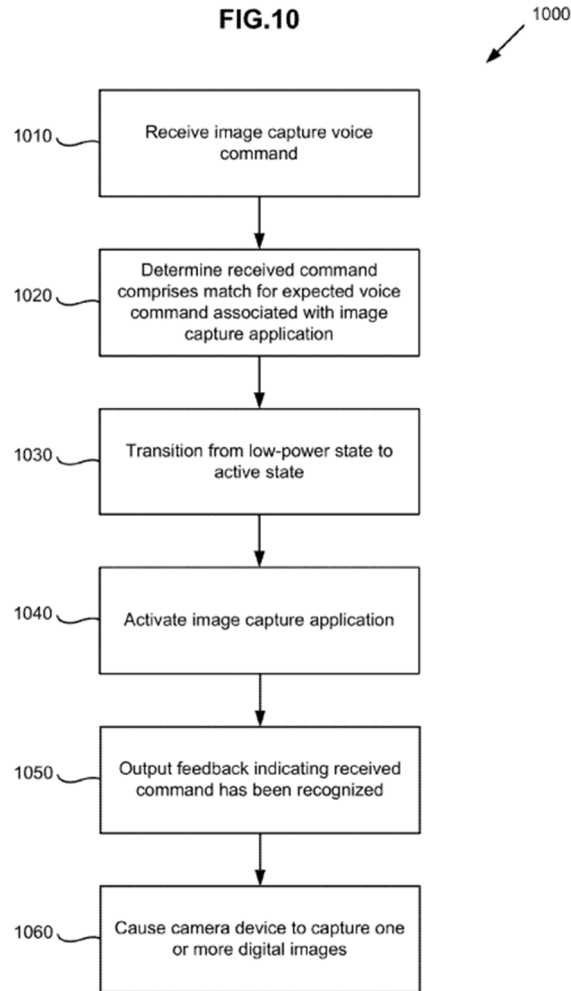
device”) while consuming less power as the primary processor of the mobile device is not activated (“operating in an idle mode”).

c. 1[b]: in response to receiving the acoustic input, determining whether the received acoustic input includes a voice command by processing the received acoustic input,

122. Sarin-Hwang teaches this element.

123. Sarin-Hwang teaches to determine whether the received acoustic input includes a voice command. As shown in Figure 10 below, Sarin “receives an image capture voice command” at step 1010, and then, in response, at step 1020, “determines that the received command comprises a match for an expected command associated with an image capture application.” Ex-1005, 13:47-55. As Sarin discloses, image capture-related voice commands are typically single words or short phrases. *See, e.g.*, Ex-1005, 8:8-16 (identifying exemplary commands, such as “capture image,” “take photo,” “cheese!,” etc.).

FIG.10



124. A POSITA would also have understood that the Sarin-Hwang mobile device “process[es] the received acoustic input.” For example, Hwang teaches that its DSP includes a “continuous keyword detection module 243” that “may receive sound input and may detect the existence of pre-enrolled (e.g., predetermined) keyword sets.” Ex-1006 ¶41. As discussed in §VIII.A.1, a POSITA would have understood that Sarin’s image capture-related voice commands would be included among the “keyword sets” recognized by Hwang’s DSP because the “keyword sets” that Hwang’s continuous keyword detection module detects are similar to

Sarin's image capture-related voice commands. Further, Hwang discloses that its keyword detection module allows for keyword detection "in a relatively low-power state" and may "activate predefined applications based on detected keywords." Ex-1006 ¶41. A POSITA would have understood that Sarin's image capture application is an example of a predefined application that Hwang's DSP would be capable of activating after detecting image capture related commands. Ex-1005, 13:62-63 (after transitioning to an active state, at step 1040 "the computing device activates the image capture application."). Detecting keywords in audio input by the continuous keyword detection module is consistent with the "processing" described in the '461 patent. *See, e.g.*, Ex-1001, 30:56-62 ("speech processing stages may include ASR ... acoustic input 705 may undergo limited vocabulary ASR to perform *keyword spotting*").

125. As such, in my opinion, Sarin-Hwang thus teaches a mobile device that recognizes predetermined keywords corresponding to image capture-related voice commands ("determining whether the received acoustic input includes a voice command") in received audio input ("in response to receiving the acoustic input"), using a continuous keyword detection module ("by processing the received acoustic input").

- d. 1[c]: the processing employing at least one contextual cue, the at least one contextual cue comprising information other than content of the acoustic input and comprising a time the acoustic input was received, motion of the mobile device, location of the mobile device, occurrence of an event on the mobile device, recent activity on the mobile device, a stored sample of the acoustic environment of the mobile device and/or a power source of the mobile device,

126. Sarin teaches this element.

127. Sarin teaches several “contextual cue[s] comprising information other than content of the acoustic input” that may assist in recognizing voice commands. As I explained in Section VIII.A.2.b, Sarin’s sound processor and/or Sarin-Hwang’s low-power DSP processes the acoustic input while the device is in an idle mode. Sarin further teaches that “voice-controlled applications” deployed on its mobile device “can use *contextual cues* to assist in recognizing commands and using resources (e.g., battery life) efficiently.” Ex-1005, 11:32-34; *see also id.*, 9:9-12 (using contextual cues is to disambiguate voice commands). For example, Sarin teaches contextual cues comprising: a “time the acoustic input was received” (Ex-1005, 11:42-46), “motion of the mobile device” (*id.*, 12:19-24 (a gyroscope or an accelerometer for additional contextual cues)), “location of the mobile device” (*id.*, 11:42-46), “occurrence of an event on the mobile device” (*id.*, 11:47-52 (if “a voice call is active on a smart phone” or “whenever the camera is not active”)), and “recent activity on the mobile device” (*id.*).

128. A POSITA would have understood that Sarin's contextual cues are similar to the contextual cues described in the '461 patent, including information other than content of the acoustic input in the same way. *See also* Ex-1001, 21:29-33 (time of day), 21:5-10 (mobile device's location in the home or office), 22:41-47 (response from an accelerometer).

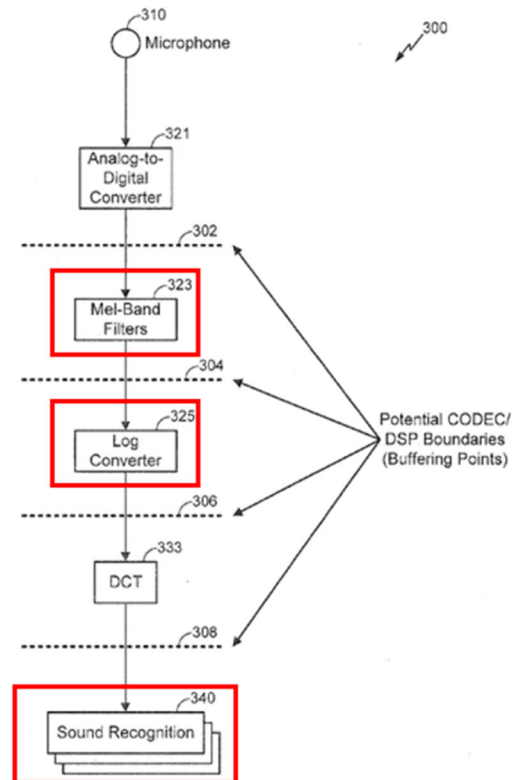
129. As such, in my opinion, Sarin-Hwang teaches a mobile device that processes audio input using a contextual cue ("the processing employing at least one contextual cue") comprising a number of the specific contextual cues required by the claim. Moreover, as the claim is drafted, only one contextual clue is required, yet Sarin discloses several.

e. 1[d]: the processing of the received acoustic input comprising a plurality of processing stages, at least one of the processing stages being performed while the mobile device is operating in the idle mode.

130. Sarin-Hwang teaches this element.

131. Sarin-Hwang teaches a plurality of processing stages with at least one being performed while the mobile device is operating in the idle mode. As shown in Figure 3, Sarin-Hwang's DSP "may perform feature extraction (e.g., via mel-band filters 323 and a log converter 325), data transformation (e.g., via a DCT transformer 333), and sound recognition (e.g., via sound recognition modules

340).” Ex-1006 ¶52. Each of these are “processing stages” because they are stages used in the processing of the received acoustic input to extract identifiable features.



132. Sarin-Hwang teaches that at least one of the above processing stages is performed when the mobile device is operating in idle mode. As discussed in §VIII.A.2.b, Sarin-Hwang’s DSP receives and processes audio while the mobile device’s primary processor is not activated and the mobile device is in an idle mode as it consumes less power. Specifically, both feature extraction and sound recognition are processing stages that are performed prior to the DSP activating the mobile device’s primary processor. Ex-1006 ¶¶44 (“the processor 230 may analyze the transformed audio feature data 234 via one or more of the sound recognition

modules 241-245, and may determine whether to activate an application processor and/or component(s) of a mobile station modem (MSM).”), 53-55 (“Data transformation and sound recognition may be performed by the DSP.”).

133. Accordingly, in my opinion, a POSITA would have understood that Sarin-Hwang teaches a mobile device that extracts audio features from audio input and performs keyword detection on the extracted audio features (“the processing of the received acoustic input comprising a plurality of processing stages”) while the primary processor of the mobile device is not activated and consuming less power (“at least one of the processing stages being performed while the mobile device is operating in the idle mode”).

3. Dependent Claims 2-4, 6, 7, 11, 14, and 16

a. Claim 2: The method of claim 1, further comprising transitioning the mobile device from the idle mode to an active mode in response to detecting that the acoustic input includes a voice command.

134. Sarin-Hwang teaches this claim.

135. Sarin-Hwang transitions the mobile device from idle to an active mode in response to a detected voice command. Specifically, Hwang teaches that its DSP, “[a]fter retrieving the audio feature data 130, ... may analyze the audio feature data 130 and may generate the activation signal 160 based on the analysis.”

Ex-1006 ¶31. When the analysis of the audio feature data 130 “identifies a

particular *voice input command* (e.g., ‘wake up’), the DSP “may generate the activation signal 160 to activate various components of an electronic device.” *Id.*; *see also id.* ¶¶41 (describing Hwang’s continuous keyword detection module as enabling “voice-activation commands without the use of a dedicated voice command button or non-verbal user input.”), 44 (DSP “may analyze the transformed audio feature data ... via one or more of the sound recognition modules 241-245, and *may determine whether to activate an application processor* ... of a mobile station modem (MSM).”).

136. In my opinion, Sarin-Hwang thus teaches a mobile device that activates its primary processor (“transitions the mobile device from the idle mode to an active mode”) when it recognizes a “wake up” voice command (“in response to detecting that the acoustic input includes a voice command”).

b. Claim 3: The method of claim 1, wherein the at least one contextual cue includes the time the acoustic input was received, and the method comprises influencing evaluation of at least one of the plurality of processing stages based on the time the acoustic input was received.

137. Sarin-Hwang teaches this claim.

138. Sarin-Hwang teaches that the contextual cue includes the time the acoustic input was received and that time influences the evaluation of a processing stage. First, Sarin teaches that “voice-controlled applications” deployed on its

mobile device “can use *contextual cues* to assist in recognizing commands and using resources (e.g., battery life) efficiently.” Ex-1005, 11:32-34. Second, Sarin teaches different “contextual cues” that may assist in recognizing voice commands corresponding to image capture functionality, such as the time the acoustic input was received. *Id.*, 11:42-46 (“Other contextual clues include ... time of day.”). A POSITA would have understood that Sarin’s consideration of the “time of day,” is a contextual cue that influences the processing stage when used to recognize voice commands corresponding to image capture functionality. For example, Sarin explains that the “time of day” cue can be used to influence the evaluation of the processing stage by, for example, turning off the processing. Ex-1005, 11:44-46 (“if the time of day and location indicate that the user is at work, voice recognition relating to image capture operations can be turned off.”). This was a common way to use contextual cues, as I discussed above in Section V.A. Specifically, time of day was a common contextual cue as shown by Lebeau., which uses the user’s time, location, schedule, and the mobile device’s battery status as contextual cues of when to monitor for the user’s speech, which “help[s] conserve the amount of energy consumed by a mobile computing device, especially when the mobile computing device is using a portable power source, such as a battery.” Ex-1016 ¶¶10, 43.

139. Accordingly, it is my opinion that Sarin-Hwang teaches a mobile device that uses time of day (“the at least one contextual cue includes the time the acoustic input was received”) to recognize voice commands corresponding to image capture functionality (“the method comprises influencing evaluation of at least one of the plurality of processing stages based on the time the acoustic input was received.”).

c. Claim 4: The method of claim 1, wherein the at least one contextual cue includes recent activity on the mobile device, the method further comprising evaluation of at least one of the plurality of processing stages based on at least one of how recently the user activated the mobile device, how recently the user issued a voice command and how recently the user performed at least one action on the mobile device.

140. Sarin-Hwang teaches this claim.

141. Sarin teaches that the contextual cues includes recent activity on the mobile device. As I explained in Section VIII.A.2.d, Sarin teaches how voice-controlled applications can use contextual cues to assist in recognizing commands. Ex-1005, 11:32-34. Sarin’s contextual cues include “which applications or functions are active on a device.” Ex-1005, 11:42-43. Sarin thus teaches a contextual cue of the recent activity on the mobile device, similar to the ’461 patent. The ’461 patent explains:

For example, if a user opens a text or email, the user may indicate the desire to speak the information by commanding the mobile device to “Take dictation.” Opening a calendar or scheduling application make be cue that the user may want to schedule something using voice (e.g., “Schedule lunch with Jane at noon tomorrow”).

Ex-1001, 23:67-24:5.

142. Sarin teaches basing the evaluation on how recently the user performed at least one action on the mobile device. Sarin provides that if a “voice call is active on a smart phone, voice recognition relating to image capture operations can be turned off” to avoid “unintended camera operation.” Ex-1005, 11:47-53. Sarin’s consideration of whether a “voice call is active” is a contextual cue that influences the processing stage when used to recognize voice commands corresponding to image capture functionality. For example, when a phone call is recently started, Sarin’s system knows that any more recent voice input will not relate to image capture operations because the user is currently on the voice call. The call being currently “active” is as recent as can be. Therefore, the user’s recently performed action, such as an activating a voice call, will be used to influence the evaluation of the processing stage by, for example, turning off the processing to detect any more recent voice commands, until the call is finished. A POSITA would have understood that Sarin consideration of a user’s active voice call is an example of considering how “recently a user interacted with a mobile device” because the user is actively interacting with the device.

143. This was a common way to use contextual cues, as I discussed above in Section V.A., specifically in view of Binder, which further describes that “the context of the electronic device is used to help determine whether and how to operate the voice trigger.” Ex-1014 ¶128. Binder explains that the context of the device can be used to adjust how and whether the voice trigger operates. Ex-1014 ¶129 (further explaining that the voice trigger is deactivated when the phone is in a silent or do not disturb mode). Gruber also teaches that pronouns in voice commands can be disambiguated based on “current” operations on a mobile device. For example, Gruber teaches that to “interpret instructions such as ‘send him a text message’, wherein the ‘him’ is interpreted according to context information derived from a *current* phone call.” Ex-1008 ¶18 Gruber also provides an example of its natural language process “determining that the user input is about making a phone call to a person referred to by the pronoun ‘her’, and finding a specific data representation for this person” by identifying “her” as the person that the user is text messaging. Ex-1008 ¶20 (“the context of a text messaging application can help constrain the interpretation of “her” to mean “the person with whom I am conversing in text.”). Thus, it was well known to consider how recent the user performed at least one action on the mobile device when processing the acoustic input and a POSITA would have understood that Sarin does the same when determining if there is an active voice call.

144. Accordingly, it is my opinion that Sarin-Hwang teaches a mobile device that uses whether a voice call is currently active (“the at least one contextual cue includes recent activity on the mobile device”) to turn off the processing of image capture-related voice commands until the call is finished (“evaluation of ... how recently the user performed at least one action on the mobile device”).

d. Claim 6: The method of claim 1, wherein the at least one contextual cue includes motion of the mobile device received as an input from an accelerometer of the mobile device, the method comprises influencing evaluation of at least one of the plurality of processing stages based on an assessment of whether the user moved the mobile device proximate in time with receiving the acoustic input based, at least in part, on the input from the accelerometer.

145. Sarin-Hwang teaches this claim.

146. Sarin teaches that the contextual cues includes motion of the mobile device received as an input from the mobile device’s accelerometer. Sarin explains that the mobile device includes an “accelerometer” to provide additional contextual cues “for recognizing voice commands.” Ex-1005, 12:19-24. Sarin provides an example of using an accelerometer and/or gyroscope to detect “the orientation of the device as being ready perform an operation such as capturing an image or recording video.” *Id.* A POSITA would have understood that Sarin thus teaches

that the contextual cue includes “motion of the mobile device received as an input from an accelerometer of the mobile device” because Sarin receives input from other device components, such as accelerometers, which sense motion.

147. Sarin teaches that the processing stage is influenced by the accelerometer in assessing whether the user moved the mobile device (e.g. into a specific orientation) proximate in time with receiving the acoustic input. Sarin’s accelerometer and gyroscope detect the “orientation of the device,” which is then used to determine that the device is “ready [to] perform an operation such as capturing an image or recording video.” Ex-1005, 12:19-24. Sarin’s consideration of device’s orientation (from the accelerometer) is a contextual cue that influences the processing stage when used to recognize voice commands corresponding to image capture functionality by indicating that the device is “ready” to capture the image because there will be a movement to change to the specific orientation. The rotation of the phone or moving it near the face both constitute “motion” and in particular both are sensed by the accelerometer as differing forces in the x, y, and z directions.

148. A POSITA would have understood that the correlation of the device’s orientation with the image capture command in Sarin is considering the device’s motion proximate in time with the received acoustic input because it was well known to monitor the device’s orientation over a time period before performing an

action. For example, U.S. Patent Application Publication No. 2010/0129068

(“Binda”) shows that it was well known for accelerometers to supply signals to generate an interrupt when the electronic device is brought from one position to another, within a time interval T. Binda teaches an “orientation detecting device 10” and a “inertial sensor”, which is “a triaxial microelectromechanical accelerometer (or MEMS accelerometer), having independent detection axes X, Y, Z, and supplies signals each indicating the effect of Earth's gravity force according to a respective one of the three detection axes X, Y, Z.” Ex-1017 ¶22. An interface converts the signals supplied by the inertial sensor 15 into numeric inertial signals AX, AY, AZ for respective detection axes X, Y, Z of the inertial sensor 15 itself. *Id.* ¶24. These signals are then transmitted to the control module and can be forwarded to the interrupt-generator module so that the device is activated. *Id.*

¶¶25-28. Specifically, Binda teaches:

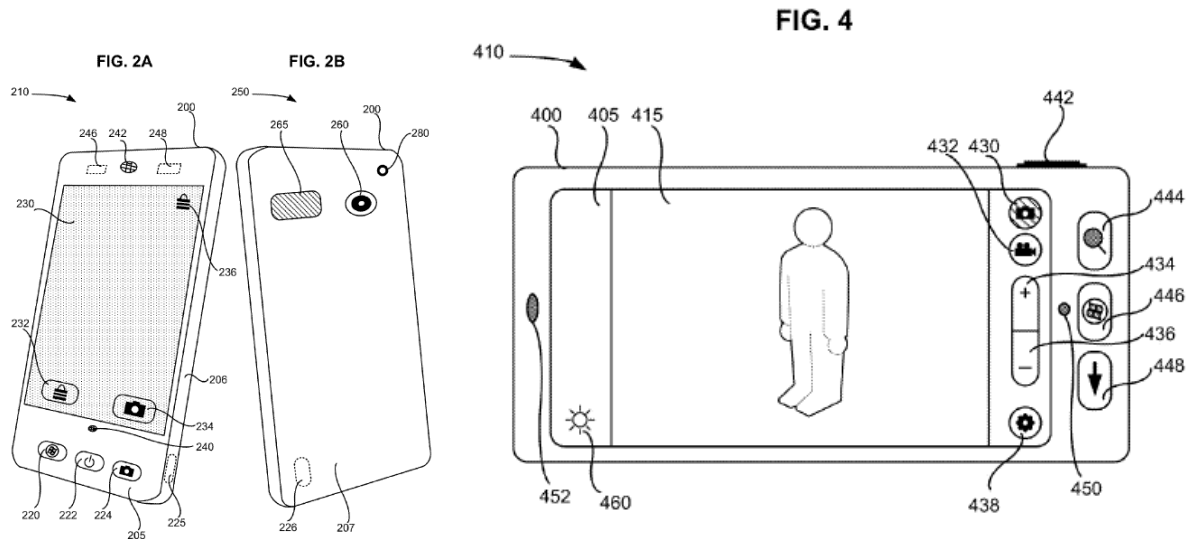
the interrupt-generator module 18 compares the modules of the inertial signals A_X , A_Y , A_Z with respective thresholds TH_X , TH_Y , TH_Z and determines that a detection axis X, Y, Z of the inertial sensor 15 is substantially horizontal if the corresponding inertial signal A_X , A_Y , A_Z remains below the respective threshold TH_X , TH_Y , TH_Z for a time interval T. When one of the detection axes is approximately horizontal and the electronic apparatus 1 is at rest, in fact, the component due to the acceleration of gravity along that detection axis is negligible, in so far as it is approximately perpendicular. In one operating mode, as already mentioned, the interrupt signal INT is generated when the electronic apparatus 1 is brought from one to another of

the first, second, and third operating positions. Assume, for example, that the electronic apparatus 1 is initially in the first operating position (“landscape”), in which the detection axis X is substantially horizontal (as likewise the detection axis Z). The electronic apparatus 1 is then rotated so as to be brought into the second operating position (“portrait”) with the detection axis Y substantially horizontal (for example, the electronic apparatus 1 is turned about the detection axis Z, which remains horizontal). During displacement, the interrupt-generator module 18 detects that the inertial signal A_X increases until it exceeds the threshold TH_X , since the effect of the acceleration of gravity along the detection axis X increases. Instead, the inertial signal A_Y decreases and goes below the respective threshold TH_Y . If the second operating position is kept stable during the time interval T, the interrupt-generator module 18 sends the interrupt signal INT to the control unit 3 and loads a new orientation string SOR into the I/O register 20.

Id. ¶31. Therefore, it was known to consider the device’s motion proximate in time with the received input.

149. For example, Sarin explains moving the device into a particular orientation of the device as a cue that can influence the evaluation of the processing stage by ensuring that the device is “ready” to capture the picture near in time to when the motion ends, the device is oriented correctly, and the user issues the command to capture an image. At least in the common case, wherein the user takes a photo in landscape mode (such as Sarin’s Figure 4), the user would have to frame the photo correctly before taking the photo, and would not want to take the photo while the phone is still rotating (such as in Sarin’s Figure 2A and

2B). A POSITA would have understood that Sarin would not want to listen for speech while the camera is rotating (because the image would be blurry and not framed) but would want to listen as soon as the phone was moved to a stable and in the new orientation. The user is also likely to move the position of the phone so it is in front of the face to be able to see how the photo is framed, like looking through a camera view finder.



Moreover, it was well known to identify changes in location caused by motion, as taught by Hwang. Hwang teaches a “novelty detection module 245” that “may detect changes in input audio that correspond to changes in location and/or changes in activity.” Ex-1006 ¶43. This may be used in conjunction with other sound operations to identify location and sound activity, and “log the corresponding time for subsequent usage and analysis.” *Id.*

150. As I explained in Section V.A, using accelerometers in speech recognition systems, is common. Like Sarin, Binder teaches a mobile device with a motion sensor such as an accelerometer. Ex-1014 ¶52. Binder further describes that “the context of the electronic device is used to help determine whether and how to operate the voice trigger,” such as information from the accelerometer may be used as a context of the device. Ex-1014 ¶128. The accelerometer can be used to determine that the device is face-up or face down so that the system knows to activate the voice trigger when the device is placed “face-up.” Ex-1014 ¶131; *see also* Ex-1014 ¶¶166-167 (“electronic device detects its orientation using light sensors (including cameras), microphones, proximity sensors, magnetic sensors, accelerometers, gyroscopes, tilt sensors, and the like.”). Binder also provides the example that the accelerometer to determine that the device is in a vehicle by monitoring the motion and if it is determined that device is likely in a vehicle because it is traveling above 20 miles per hour, then it “is determined to be travelling along a road” and “then the voice trigger remains active and/or in a high-power or more sensitive state.” Ex-1014 ¶133. Specifically, Binder’s accelerometer detects “the motion of the electronic device and/or the external audio system 416,” which is then used in the voice trigger system. Ex-1014 ¶139.

151. LeBeau is another reference I explained above to show how POSITAs understood to use an accelerometer in a speech recognition system. Specifically,

Lebeau provides that the computing device “may determine that it is currently travelling in a car based on a detected high rate of speed at which the device is travelling (e.g., using any of a variety of motion sensors that are standard components of the device).” Ex-1016 ¶22. Using the user’s location based on “motion sensors,” Lebeau can determine whether to monitor for voice input based on this current context. Ex-1016 ¶22. Lebeau explains how this “help[s] conserve the amount of energy consumed by a mobile computing device, especially when the mobile computing device is using a portable power source, such as a battery” by not monitoring for voice when the motions indicate that it would or would not be convenient for the user to speak a user request. Ex-1016 ¶10.

152. Accordingly, it is my opinion that Sarin-Hwang teaches a mobile device that uses motion information received from an accelerometer to determine the device’s orientation (“the at least one contextual cue includes motion of the mobile device received as an input from an accelerometer of the mobile device”) to determine whether the device is “ready” to take a picture near in time to the end of the motion because a movement would have brought the device into the specific orientation (“influencing evaluation of at least one of the plurality of processing stages based on an assessment of whether the user moved the mobile device proximate in time with receiving the acoustic input based, at least in part, on the input from the accelerometer”).

- e. Claim 7: The method of claim 1, wherein the at least one contextual cue includes the location of the mobile device, and the method comprises influencing evaluation of at least one of the plurality of processing stages based on determining whether the mobile device is located in at least one of a vehicle, a home or a place of work of the user of the mobile device.**

153. Sarin-Hwang teaches this claim.

154. Sarin teaches that the user's work location is a contextual cue that influences the processing stage. Sarin teaches different "contextual cues" that may assist in recognizing voice commands corresponding to image capture functionality, such when the "location indicate[s] that the user is at *work*, voice recognition relating to image capture operations can be turned off." Ex-1005, 11:42-46. Being in this "work" location is a cue used to influence the evaluation of the processing stage by, for example, turning off image capture-related voice recognition. Ex-1005, 11:44-46 ("image capture operations can be turned off.").

155. As I explained in Section V.A, the location of the mobile device is a common contextual cue used in evaluating the acoustic input. For example, Lebeau (Ex-1016) specifically discusses using contextual cues to evaluate the likelihood that a user would speak a verbal search request. Ex-1016 ¶¶3, 22 ("the mobile computing device can attempt to infer whether the current context indicates that the user has at least a threshold likelihood of providing voice input and, if so, monitoring for voice input in response."), 43 ("threshold likelihood of [the user]

providing a voice-based request to the mobile computing device”). Lebeau explains that current context associated with a mobile computing device includes the environment around the device, or a context internal to the device such as “the physical location where the mobile computing device is located (e.g., home, work, car, etc., as determined by GPS in the device or other techniques).” Ex-1016 ¶3; *see also* Ex-1016 ¶43 (considering that the user is based at home).

f. Claim 11: The method of claim 1, wherein the mobile device includes a primary processor and a secondary processor, and wherein at least one first processing stage of the plurality of processing stages is performed by the secondary processor prior to engaging the primary processor to process the acoustic input.

156. Sarin-Hwang teaches this claim.

157. Sarin teaches that its mobile device includes a primary processor. Sarin’s mobile device includes “a controller or processor 110 ... for performing such tasks as signal coding, data processing, input/output processing, power control, and/or other functions.” Ex-1005, 4:15-20.

158. Sarin, in view of Hwang, teaches that the mobile device has a secondary processor. Sarin teaches that its mobile device may have more than one main processor. Ex-1005, 6:19-22. Moreover, as discussed in §VIII.A.1, Hwang teaches a device that includes a DSP, which processes and extracts features from audio input. *See, e.g.*, Ex-1006 ¶52. A POSITA would have understood that

Hwang's DSP is a distinct processor from the first processor because Hwang describes that the DSP may determine whether to activate a different application processor. Ex-1006 ¶44. Moreover, in systems utilizing DSPs, the DSP is generally considered to be a secondary processor working in concert with a general purpose primary processor, to perform particular tasks for which the DSP architecture is optimized.

159. Sarin-Hwang's secondary processor performs a first processing stage before engaging the primary processor to process the acoustic input. Sarin teaches to perform speech recognition to listen for and interpret voice input, on the local device. Ex-1005, 6:29-49, 13:15-20. Moreover, the DSP of the Sarin-Hwang mobile device performs feature extraction, data transformation, and sound recognition *before* engaging a primary processor to further process audio. For example, the DSP "may analyze ... transformed audio feature data 234 via one or more of the sound recognition modules 241-245, and *may determine whether to activate an application processor* and/or component(s) of a mobile station modem (MSM) or other component *based on the analysis*." Ex-1006 ¶44. Because the DSP analyzes with sound recognition modules before activating the application processor, the application processor is not engaged to process the acoustic input.

160. Accordingly, it is my opinion that Sarin-Hwang teaches a mobile device that includes audio feature extraction, data transformation, and sound

recognition (“at least one first processing stage of the plurality of processing stages”) performed by a DSP (“performed by the secondary processor”) and, based on the results of the sound recognition, may activate an application processor (“prior to engaging the primary processor to process the acoustic input”).

g. Claim 14: The method of claim 11, wherein performing the at least one first processing stage comprises performing limited vocabulary automatic speech recognition on the acoustic input to detect whether the acoustic input includes an explicit voice trigger

161. Sarin-Hwang teaches this claim.

162. Sarin-Hwang teaches performing a limited vocabulary automatic speech recognition, as part of the first processing stage. As I explained in the previous section (Claim 11), the first processing stage is performed by Sarin-Hwang’s DSP. Sarin-Hwang’s DSP includes a “continuous keyword detection module 243,” which “may receive sound input and may detect the existence of pre-enrolled (e.g., predetermined) keyword sets.” Ex-1006 ¶41. Sarin-Hwang’s continuous keyword detection using “pre-enrolled (e.g., predetermined) keyword sets,” is the claimed “limited vocabulary automatic speech recognition” because it is a restricted set of predefined words, just like the ’461 patent. Ex-1001, 18:14-16 (“using a vocabulary restricted to the words in the explicit voice trigger phrase (which may include as few as a single word”).

163. Sarin-Hwang teaches using limited vocabulary automatic speech recognition to detect whether the acoustic input includes an explicit voice trigger. For example, a keyword set recognized by the DSP may include a specific voice trigger such as “wake up,” which may cause the DSP to “generate [an] activation signal ... to activate various components of an electronic device.” Ex-1006 ¶31.

164. In my opinion, the Sarin-Hwang combination thus teaches a mobile device that includes a DSP that includes a continuous keyword detection module (“performing limited vocabulary automatic speech recognition on the acoustic input”) that recognizes keyword sets, including phrases such as “wake up” (“to detect whether the acoustic input includes an explicit voice trigger”).

h. Claim 16: The method of claim 1, wherein the mobile device is capable of connecting to at least one server via a network, and the method comprises transmitting at least a portion of the acoustic input to the at least one server via the network for processing by the at least one server.

165. Sarin-Hwang teaches this claim.

166. Sarin-Hwang teaches that the mobile device connects to a server via a network. For example, Sarin teaches that the mobile device has “[a] wireless modem [which] can be coupled to an antenna ... and can *support two-way communications between the processor 110 and external devices*, as is well understood in the art.” Ex-1005, 4:60-63. Further, the wireless modem may be “a

cellular modem for communicating with [a] *mobile communication network*.” Ex-1005, 4:63-65. Sarin also teaches that the mobile device’s speech recognition components can also be “implemented in whole or in part *on remote servers*.” *Id.*, 6:15-18.

167. Sarin-Hwang teaches transmitting a portion of the acoustic input for the server to process. For example, in performing voice command recognition for image-capture functionality, the mobile device or remote server (Ex-1005, 13:4-8) “compares ... received audio input with an expected audio command (e.g., a word or phrase) associated with an image capture operation.” *Id.*, 12:32-35. Sarin teaches that the decision “on where to perform speech recognition tasks can be based on,” among other things, “resources available at local devices and remote servers (e.g., processing power, memory or storage capacity).” *Id.*, 13:8-12. Thus, for recognizing more general voice commands, e.g. beyond those directed to image capture, a POSITA would have been motivated to transmit audio that includes such general voice commands to a remote server with sufficient processing power to process those commands. Ex-1005, 6:44-49 (“Services such as the Microsoft® Bing® service, or Apple Inc.’s Siri service can be used to implement speech recognition (e.g., in generalized voice input modes).”).

168. Accordingly, in my opinion, a POSITA would have understood that Sarin-Hwang’s mobile device “is capable of connecting to at least one server via a

network” because it has a wireless cellular modem for communicating over a communication network in order to transmit received audio input to a remote server for speech recognition (“transmitting at least a portion of the acoustic input to the at least one server via the network for processing by the at least one server”).

4. Independent Claim 18

a. 18[pre]: A mobile device comprising:

169. Sarin teaches this element for the reasons discussed in §VIII.A.2.a.

b. 18[a]: at least one processor configured to:

170. Sarin in combination with Hwang teaches this element for the reasons discussed in §VIII.A.2.b. In, particular, as discussed in §VIII.A.1, the Sarin-Hwang mobile device includes a low power DSP and an application processor. While Claim 18 does not recite the specific processor to employ the following processing steps, the low power Sarin-Hwang DSP (identified as the “processor” for Claim 1) is being applied as the claimed “at least one processor” of Claim 18, in substantially the same way it is applied with respect to Claim 1.

- (i) **18[a(i)]: in response to the mobile device receiving acoustic input from the environment while operating in an idle mode, determine whether the received acoustic input includes a voice command by processing the received acoustic input,**

171. Sarin-Hwang teaches this element for the reasons discussed in §VIII.A.2.c.

- (ii) **18[a(ii)]: the processing employing at least one contextual cue, the at least one contextual cue comprising information other than content of the received acoustic input and comprising a time the acoustic input was received, motion of the mobile device, location of the mobile device, occurrence of an event on the mobile device, recent activity on the mobile device, a stored sample of the acoustic environment of the mobile device and/or a power source of the mobile device,**

172. Sarin-Hwang teaches this element for the reasons discussed in §VIII.A.2.d.

- (iii) **18[a(iii)]: the processing of the received acoustic input comprising a plurality of processing stages, at least one of the processing stages being performed while the mobile device is operating in the idle mode.**

173. Sarin-Hwang teaches this element for the reasons discussed in §VIII.A.2.e.

**B. Ground 2: Sarin in Combination With Hwang and Garner
Renders Obvious Claim 13**

**1. A POSITA would have been motivated to combine
Sarin's, Hwang's, and Garner's teachings and would
have had a reasonable expectation of success.**

174. A POSITA would have been motivated to combine the teachings of Sarin, Hwang, and Garner (“Sarin-Hwang-Garner”) and would have had a reasonable expectation of success in doing so because each relates to the same well-known technologies.

175. A POSITA also would have been motivated to combine Garner with Sarin-Hwang because they are analogous art and teach improvements to the same type of devices. Each teach using speech processing. Ex-1005, 6:29-49, 13:15-20; Ex-1006 ¶¶52 (“feature extraction (e.g., via mel-band filters 323 and a log converter 325), data transformation (e.g., via a DCT transformer 333), and sound recognition (e.g., via sound recognition modules 340).”), 41 (“continuous key word detection”); Ex-1007, 1:45-51.

176. Garner teaches additional benefits to Sarin-Hwang's mobile device. Garner discloses that “VAD ... prevents a speech recognition process in an ASR unit ... from recognizing background noise as speech.” Ex-1007, 1:45-47, which “prevent[s] an error of converting noise into a word.” *Id.*, 1:47-49. VAD also “makes it possible to more skillfully manage the throughput of the entire [automatic speech recognition] system ... that utilizes many computer resources.”

Id., 1:49-51. Thus, a POSITA would have understood that VAD as taught by Garner allows for more efficient use of device resources because it allows the device “to concentrate on other functions while speech recognition is not in progress and concentrate on ASR while the user utters.” *Id.*, 1:56-58. Indeed, a POSITA would have understood that incorporating Garner’s VAD as a lead-in to Sarin-Hwang’s continuous keyword detection would have furthered Sarin-Hwang’s goal of conserving device power by allowing the keyword detection to focus on speech instead of noise.

177. A POSITA would have understood that Garner’s VAD is compatible with, and could successfully be implemented on, Hwang’s DSP. In particular, Hwang’s feature extractor includes a plurality of mel-band filters that generate mel-band energies from input audio data, which, similar to Garner, is divided into frames. Ex-1006 ¶¶35-36. A POSITA would have understood that the mel-band energies produced in Hwang are comparable to the mel metric signals produced in Garner.

178. Additionally, Hwang discloses that its algorithms “may be embodied directly in hardware, [or] in a software module executed by a processor.” Ex-1006 ¶92. Similarly, Garner discloses that its “invention can be implemented by supplying a software program, which implements the functions of the [disclosed] embodiments.” Ex-1007, 11:62-64. Thus, it would have been routine for a POSITA

to implement VAD as taught by Garner on Hwang's DSP to improve the functionality of Hwang's software modules (such as keyword detection module 243) such that the DSP performs keyword detection when audio input more likely includes speech rather than noise.

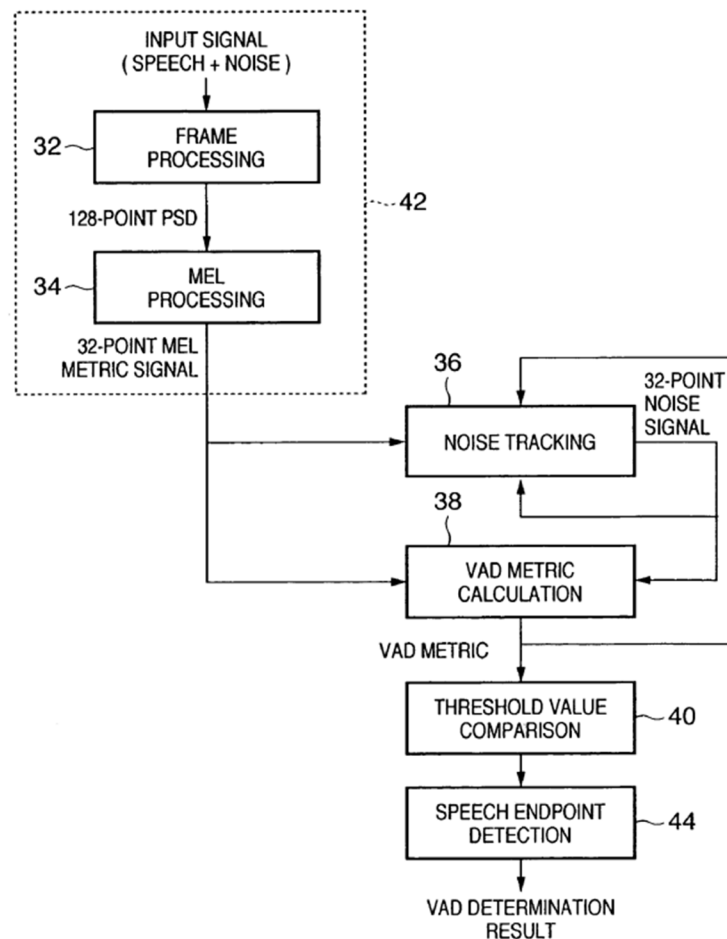
2. Dependent Claim 13

- a. Claim 13: The method of claim 11, wherein performing the at least one first processing stage includes performing at least one voice activity detection processing stage including performing at least one of spectral analysis on the acoustic input to evaluate whether a spectrum of the acoustic input is indicative of voice activity, periodicity analysis to evaluate whether signal periodicity of the acoustic input is indicative of voice activity, and phone loop analysis to evaluate whether the acoustic input includes speech.**

179. Sarin in combination with Hwang and Garner teaches this claim. Sarin-Hwang-Garner teaches that its first processing stage includes performing a voice activity detection processing stage including performing spectral analysis on the acoustic input to evaluate whether a spectrum of the acoustic input is indicative of voice activity. As explained in Section VIII.A.2.e (Element 1[d]), Sarin-Hwang's DSP "may perform feature extraction (e.g., via mel-band filters 323 and a log converter 325), data transformation (e.g., via a DCT transformer 333), and sound recognition (e.g., via sound recognition modules 340)." Ex-1006 ¶52. More

particularly, Hwang discloses a “target sound detection module 244,” which “detect[s] speech.” Ex-1006 ¶42.

180. A POSITA would have understood that the detection of speech in input audio may be an example of VAD, as taught by Garner. Garner teaches a signal detection technique that uses voice activity detection (VAD) to determine whether a signal includes speech. Garner’s figure 4 illustrates this process:



181. As shown Garner’s Figure 4, Garner’s process first receives “[a]n acoustic signal (which can contain speech and background noise).” Ex-1007, 5:55-56. After being processed into frames, the acoustic signal is input to a mel

processing module 34, which generates mel metric signals “as feature amounts for VAD.” *Id.*, 6:3-12. The mel metric signals are then input to noise tracking and VAD metric modules which, as shown in the figure, result in the generation of a VAD metric. *Id.*, 6:25-26. The VAD metric indicates the likelihood that an acoustic frame is speech as opposed to noise. *Id.*, 6:31-46.

182. In particular, Garner teaches a “feature extraction” technique, whereby “[a]n acoustic signal (which can contain speech and background noise)” is input from a microphone, and, after sampling by an A/D converter, the signal is divided into frames by a frame processing module (designated 32). Ex-1007, 5:55-60. Each frame is then processed further into a periodogram, or a Power Spectral Density (PSD), having a number of “bins.” *Id.*, 5:63-6:1. Next, each PSD is reduced in size by a “mel processing module” (34), which generates mel metric signals “as feature amounts for VAD.” *Id.*, 6:3-12. The mel metric signals are then input to noise tracking and VAD metric modules which, as shown in the figure, result in the generation of a VAD metric. *Id.*, 6:25-26.

183. Garner discloses a VAD metric that is a “likelihood ratio,” which is a ratio of the likelihood that a frame is speech to a likelihood that the frame is noise. *Id.*, 6:31-46. The likelihood ratio is compared with a predetermined threshold value to generate an indication of whether the audio input signal contains speech or noise. *Id.*, 11:24-26. For example, in describing its calculation of its VAD metric,

Garner discloses the calculation of speech variance in a given frame of audio using a “spectral subtraction method.” Ex-1007, 6:60-67. Garner also discloses the use of “spectral magnitude” or “spectral power” in computing an estimate of speech variance in the frame. *Id.*, 6:57-60.

184. In addition to the reasons I discussed in the previous section, a POSITA would have been motivated to incorporate VAD as taught by Garner into Sarin-Hwang such that Hwang’s keyword detection module 243 analyzes audio input for keyword(s) only if the VAD indicates that the audio input more likely comprises speech rather than noise.

185. As an initial matter, a POSITA would have understood that Hwang’s DSP, without modification, already performs VAD. More particularly, Hwang discloses a “target sound detection module 244,” which, “upon detecting speech, ... may cause a voice recording application to record the speech.” Ex-1006 ¶42. A POSITA would have understood the detection of speech in input audio to be synonymous with VAD. Ex-1007, 1:13-14 (“Detection of speech periods is generally referred to as VAD.”). Thus, although Hwang does not explicitly state that its VAD is implemented as part of its keyword detection process, a POSITA would have understood that Hwang at least contemplates the use of VAD by its DSP.

186. A POSITA would have been motivated to employ VAD to improve the Sarin-Hwang combination's continuous keyword detection feature for the reasons suggested by Garner. As Garner teaches, "VAD ... prevents a speech recognition process in an ASR unit ... from recognizing background noise as speech." Ex-1007, 1:45-47. As a result, "VAD ... prevent[s] an error of converting noise into a word." *Id.*, 1:47-49. VAD also "makes it possible to more skillfully manage the throughput of the entire [automatic speech recognition] system ... that utilizes many computer resources." *Id.*, 1:49-51. Thus, VAD allows for more efficient use of device resources because it allows the device "to concentrate on other functions while speech recognition is not in progress and concentrate on ASR while the user utters." *Id.*, 1:56-58.

187. Relatedly, a POSITA would have understood that Garner's VAD is compatible with, and could successfully be implemented in, Hwang's DSP. As I have stated above, Hwang already includes VAD for voice detection as software, as Hwang is implemented on a DSP. That software could trivially be also used in the acoustic processing chain for speech recognition. Although Hwang does not detail the operation of its VAD, its speech processing is directly compatible with Garner. For example, Hwang's feature extractor includes a plurality of mel-band filters that generate mel-band energies from input audio data, which, similar to Garner, is divided into frames. Ex-1006 ¶¶35-36. A POSITA would have

understood that the mel-band energies produced in Hwang are analogous to the mel metric signals produced in Garner.

188. To note, Hwang also discloses that its mel-band energies (i.e., extracted audio features) may first be provided to a log converter, which “may apply a logarithm function to the energies.” Ex-1006 ¶36. Further, Hwang’s DSP may also include “[a] transformer 233,” which “may be configured to apply a discrete cosine transform (DCT) function.” *Id.* ¶37. In this regard, Hwang’s disclosure is similar to Garner, which observes that, in conventional speech recognition settings, “[a mel] representation is typically used in a process of executing logarithmic transformation and then cosine transformation for a mel spectrum.” Ex-1007, 6:14-17. Relatedly, Garner discloses that its feature amounts can be values “having directly undergone ... mel processing” or can be based on another metric. Ex-1007, 6:18-22. Thus, a POSITA would have understood that Hwang’s log-converted and transformed mel-band energies would be compatible with Garner’s VAD technique.

189. Additionally, Hwang discloses that its algorithms “may be embodied directly in hardware, [or] in a software module executed by a processor.” Ex-1006, ¶92. Similarly, Garner discloses that its “invention can be implemented by supplying a software program, which implements the functions of the [disclosed] embodiments.” Ex-1007, 11:62-64. Thus, it would have been routine for a POSITA

to implement VAD as taught by Garner to improve the functionality of an existing software module (such as keyword detection module 243) such that it performs keyword detection when the audio input more likely includes speech rather than noise.

190. A POSITA would have also understood that Garner's VAD technique is spectral analysis. Both Garner and Hwang utilize mel-based methods of representing the input audio. Mel-based systems determine the acoustic energy of the input audio in different frequency bands, similar to those of the human ear. This is precisely spectral analysis, as the spectrum of a sound relates to which frequencies are present in it, just as the spectrum of white light can be decomposed into rainbow colors by a prism. Additionally, in describing its calculation of a maximum likelihood ratio as a VAD metric, Garner discloses the calculation of speech variance in a given frame of audio using a "spectral subtraction method." Ex-1007, 6:60-67. Garner also discloses the use of "spectral magnitude" or "spectral power" in computing an estimate of speech variance in the frame. *Id.*, 6:57-60.

191. Accordingly, A POSITA would have found it obvious to combine Garner with Sarin and Hwang because doing so would have amounted to applying a known technique (Garner's VAD that uses mel processing and spectral analysis) to a known device ready for improvement (the Sarin-Hwang mobile device, which

has a DSP that uses mel-energies as audio features for sound detection, but which does not explicitly teach the use of VAD as part of its keyword detection process) to yield predictable results (a mobile device that uses VAD to determine that the audio input on which it performs keyword detection more likely comprises speech rather than noise.).

192. A POSITA would have also found it obvious to combine Garner with Sarin and Hwang because doing so would have amounted to applying a known technique (using VAD to ensure that a signal on which ASR/keyword detection is performed more likely comprises speech rather than noise) to improve a similar device (the Sarin-Hwang combination which also performs ASR/keyword detection) in the same way (to ensure that the Sarin-Hwang combination performs ASR/keyword detection on audio that more likely comprises speech rather than noise). As I discussed in Section V.C, a POSITA would have understood that voice activity detection techniques include using spectral analysis to determine that there is speech included an audio input.

193. Accordingly, Sarin-Hwang-Garner teaches a mobile device that includes a DSP that performs voice activity detection to determine that audio input more likely comprises speech rather than noise before performing keyword detection on the audio input (“performing at least one voice activity detection processing stage”) that uses spectral subtraction (“performing spectral analysis on

the acoustic input”) to compute a likelihood ratio that indicates whether the audio input more likely includes speech (“to evaluate whether a spectrum of the acoustic input is indicative of voice activity”).

C. Ground 3: Sarin in Combination With Hwang and Gruber Renders Obvious Claims 3, 5, 7, 9, 10, 12, 15, and 17

1. A POSITA would have been motivated to combine Sarin’s, Hwang’s, and Gruber’s teachings and would have had a reasonable expectation of success.

194. A POSITA would have found it obvious to combine Gruber with Sarin and Hwang (“Sarin-Hwang-Gruber”) and would have had a reasonable expectation of success in doing so because each relates to the same well-known technologies.

195. A POSITA also would have been motivated to combine Gruber with Sarin-Hwang because they are analogous art and teach improvements to the same type of devices. Each teach using speech processing to determine a voice command. Ex-1005, 6:29-49, 13:15-20; Ex-1006 ¶¶52 (“feature extraction (e.g., via mel-band filters 323 and a log converter 325), data transformation (e.g., via a DCT transformer 333), and sound recognition (e.g., via sound recognition modules 340).”), 41 (“continuous key word detection”); Ex-1008 ¶11 (“interprets user inputs, operationalizes the user’s intent into tasks and parameters to those tasks, executes services to support those tasks, and produces output that is intelligible to

the user.”). Relatedly, Sarin and Gruber both teach using contextual cues. Ex-1005, 11:31-52; Ex-1008 ¶¶131, 354.

196. Sarin’s mobile device is not limited to recognizing only image capture-related voice commands. Rather, Sarin’s mobile device also recognizes more generalized voice commands. As such, Sarin discloses that speech recognition “for generalized voice input modes ... can be implemented in software such as stand-alone applications or plug-ins.” *Id.*, 6:44-46. Furthermore, according to Sarin, Microsoft’s Bing service and Apple’s Siri service can be used to implement speech recognition for generalized voice input modes.⁷ *Id.*, 6:46-49. However, Sarin does not describe in detail how contextual cues can be used to recognize more general, non-image capture-related voice commands using generalized voice input tools, such as Microsoft’s Bing or Apple’s Siri. Given Sarin’s stated preference for using contextual cues to disambiguate voice commands, as well as to reduce power usage, a POSITA would have been motivated to look for references that provide additional implementation details regarding use of contextual cues for generalized voice command recognition.

⁷ Sarin states that “generalized voice input tools can be used to implement generalized voice input modes.” Ex-1005, 7:25-26. Further, “[a] generalized voice input tool can accept voice input for more than one function (e.g., search functionality, voice call functionality).” *Id.*, 7:26-29. Thus a POSITA would have understood that Microsoft Bing and Apple Siri are “generalized voice input tools” as defined by Sarin.

Gruber is one such reference that a POSITA would have found. Gruber discloses a virtual assistant that receives general user input, such as spoken language, processes the input, and generates output to the user and/or performs actions on behalf of the user. Ex-1008 ¶178. The processing of Gruber’s virtual assistant includes “speech recognition, natural language processing, task flow processing, and dialog generation.” *Id.* ¶14. Gruber, moreover, discloses that its virtual assistant can be implemented on a mobile computing device. *Id.* ¶76. Thus, a POSITA would have understood that Gruber’s virtual assistant is an example of a “generalized voice input tool” capable of being implemented on Sarin’s mobile device

197. Gruber offers additional benefits to Sarin-Hwang’s mobile device, such as enhanced processing when a voice command is ambiguous. Gruber teaches software that receives general user input, such as spoken language, processes the input, and generates output to the user and/or performs actions on behalf of the user. Ex-1008 ¶178. Gruber’s processing includes “speech recognition, natural language processing, task flow processing, and dialog generation.” *Id.* ¶14. Like Sarin, Gruber recognizes that contextual cues are advantageous because they assist in disambiguating voice commands. *Compare* Ex-1005, 9:9-12 (“the same word or phrase can be used as a voice command for different functionality, with the desired functionality being determined based on contextual cues”) *with* Ex-1008 ¶15 (“a

virtual assistant of the present invention can use context when ... disambiguating among competing interpretations”). Gruber additionally teaches a number of actions such as “prompt[ing] the user for confirmation and/or request additional context information from any suitable source before calling a service ... to perform a function.” Ex-1008 ¶88.

198. Additionally, like Sarin, Gruber takes into “account the acoustic environments in which the speech is entered.” Ex-1008 ¶250. Gruber further teaches to store “noise profiles” of the user’s location since “noise profiles of a quiet office are different from those of automobiles or public places.” *Id.* Gruber, therefore, addresses a stated goal of Sarin (disambiguation), and does so for voice commands beyond image capture functionality. A POSITA would have understood that implementing Gruber’s teachings on Sarin’s mobile device would result in a device that is better equipped to disambiguate and recognize general voice commands by “reduc[ing] the need for the user to provide excessive clarification input.” Ex-1008, Abstract.

199. A POSITA would have understood that Gruber’s teachings are compatible with, and can be implemented on Sarin-Hwang’s combined system. Gruber explains that its software can be implemented on “CPU 62 may include one or more processor(s) 63 such as, for example, a processor from the Motorola or Intel family of microprocessors or the MIPS family of microprocessors.” Ex-1008

¶77; *see also id.* ¶¶75-76, Fig. 28. A POSITA would have understood that the application processor of the Sarin-Hwang combined system is similar to the processor disclosed in Gruber, and, therefore, a POSITA would have understood that the Sarin-Hwang combined system would be capable of implementing Gruber's teachings. Accordingly, a POSITA would have had a reasonable expectation of success in combining Gruber with Sarin-Hwang.

2. Dependent Claims 3, 5, 7, 9, 10, 12, 15, and 17

- a. Claim 3: The method of claim 1, wherein the at least one contextual cue includes the time the acoustic input was received, and the method comprises influencing evaluation of at least one of the plurality of processing stages based on the time the acoustic input was received.**

200. Sarin-Hwang-Gruber teaches this claim.

201. Sarin-Hwang-Gruber teaches that the contextual cue includes the time the acoustic input was received and that time influences the evaluation of a processing stage. As I explained above in Ground 1, Claim 3, Sarin's consideration of the "time of day," is a contextual cue that influences the processing stage when used to recognize image-capture related voice commands.

202. A POSITA would have been motivated to apply the teachings of Gruber to Sarin-Hwang to recognize more general voice commands based on the time acoustic input was received in order to help disambiguate those commands.

For example, Gruber teaches that “[t]ime information from clocks on client devices” can be used as contextual information for interpreting phrases within voice commands. Ex-1008 ¶132; *see also id.* ¶19 (“Contextual information sources include, for example and without limitation: the current state of a device used as an interface to the assistant, such as the current time, location, application, or data object; personal data such as a user's address book, calendar, and application usage history; and the state of the dialog between the user and the virtual assistant, such as recently mentioned people and/or places”). Gruber teaches that exemplary phrases which may be spoken in the context of user requests, such as “in an hour” and “tonight,” and for which time information can be used to ascertain the meaning of the request. Ex-1008 ¶132.

203. Additionally, in my opinion, a POSITA would have understood, or at least found it obvious, that in a conversational system such as the virtual assistant implemented in the Sarin-Hwang-Gruber mobile device (*see* Ex-1008 ¶¶17, 19), which a user interacts with in real time, that phrases such as “in an hour” or “tonight” that appear in voice commands would be interpreted based on the time the voice command was received by the device. For example, Gruber describes using natural language processing to parse text and associate the words with syntactic and semantic roles. Ex-1008 ¶20. Gruber provides an example of its natural language process “determining that the user input is about making a phone

call to a person referred to by the pronoun ‘her’, and finding a specific data representation for this person” by identifying “her” as the person that the user is text messaging. Ex-1008 ¶20. Similarly, Gruber teaches that time information from clocks on the client device may be used in the context of the user request, such as to interpret the phrases of “in an hour” and “tonight.” Ex-1008 ¶132. A POSITA would have understood that Gruber’s consideration of time information from clocks on the client device influences Gruber’s natural language processing stage in the same way that the user’s recent text message influences Gruber’s natural language processing to find specific data of what the user’s input means. For example, a user’s request of “call home in an hour” may be recognized as a command to place a phone call to the user’s home but the time to place the call would not be recognized except for the NLP sub-system that consults the clock and determines that the call should be placed an hour later.

204. This was a common way to use contextual cues, as I discussed above in Section V.A. Specifically, time of day was a common contextual cue as described by Lebeau, which uses the user’s time, location, schedule, and the mobile device’s battery status as contextual cues of when to monitor for the user’s speech, which “help[s] conserve the amount of energy consumed by a mobile computing device, especially when the mobile computing device is using a portable power source, such as a battery.” Ex-1016 ¶¶10, 43.

205. Accordingly, the Sarin-Hwang-Gruber combination to teach a mobile device that includes a conversational virtual assistant that uses time (“the at least one contextual cue includes the time the acoustic input was received”) in order to interpret certain phrases within received voice commands (“the method comprises influencing evaluation of at least one of the plurality of processing stages based on the time the acoustic input was received.”).

b. Claim 5: The method of claim 1, wherein the at least one contextual cue includes an occurrence of an event on the mobile device, and the method comprises influencing evaluation of at least one of the plurality of processing stages based on whether the acoustic input is received proximate in time with the occurrence of the event.

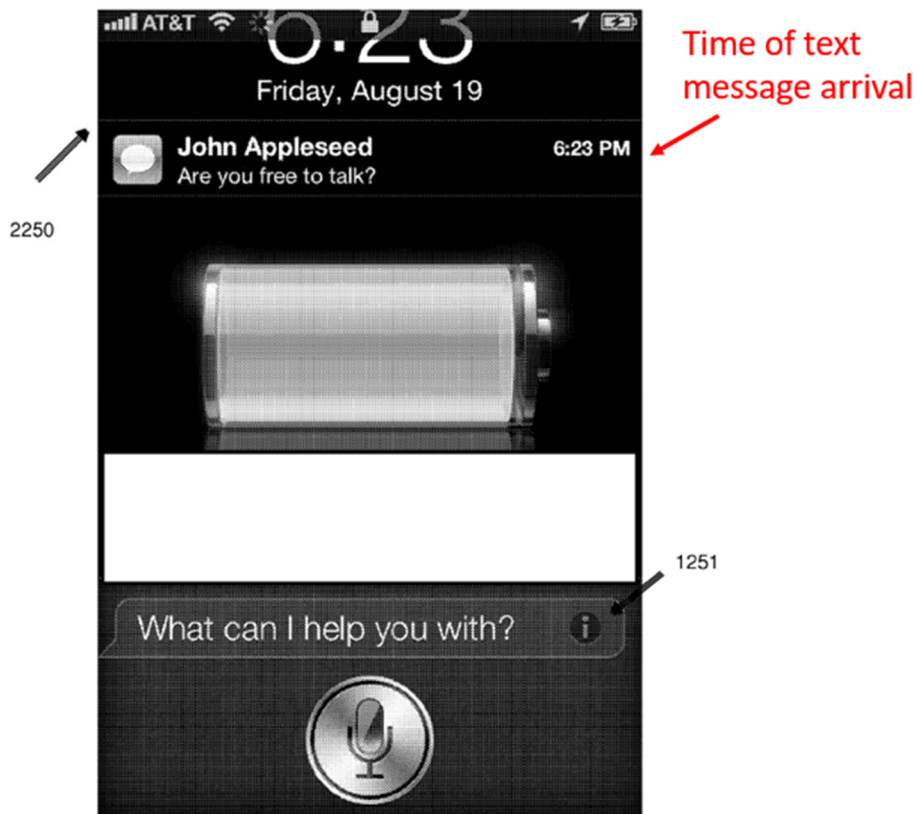
206. Sarin-Hwang-Gruber teaches this claim.

207. Sarin-Hwang teach that an occurrence of an event on the mobile device is a contextual cue, as explained above in §VIII.A.2.d. *See, e.g.*, Ex-1005, 11:47-52 (“if a voice call is active on a smart phone” or “whenever the camera is not active”).

208. A POSITA would have been motivated to use Sarin’s contextual cue and influence evaluation of at least one of the plurality of processing stages based on whether the acoustic input is received proximate in time with the occurrence of the event, in view of Gruber.

209. First, Gruber's "virtual assistant 1002 is able to use context associated with *asynchronous events* that happen independently of the user's interaction with virtual assistant 1002." Ex-1008 ¶238. Exemplary asynchronous events include "incoming text messages or pages; incoming email messages; incoming phone calls; reminder notifications or task alerts; calendar alerts; alarm clock, timers, or other time-based alerts; notifications of scores or other events from games; notifications of financial events such as stock price alerts; news flashes or other broadcast notifications; [and] push notifications from any application." *Id.* ¶¶240-249.

210. Second, Gruber teaches to consider whether the acoustic input is received proximate in time with the occurrence of various events. For example, Gruber discloses that an incoming text message is an event that is used by the system, as illustrated in figure 23. Ex-1008 ¶238. As annotated in the figure, the text message was received at 6:23 PM.



Ex-1008, FIG. 23 (annotated)

211. Next, as shown below in figure 24, an end user has spoken the command “call him.” *Id.* Gruber teaches to then “use[] the event context to disambiguate the command by interpreting ‘him’ to mean the person who sent the incoming text message.” *Id.*



Ex-1008, FIG. 24

212. As shown in figure 24, based on the time displayed on the device (6:34 PM), a POSITA would have understood that the device received the voice command (“call him”) proximate in time with the event (the receipt of the 6:23 PM text message). Furthermore, a POSITA would have been motivated to use this proximity in time between the receipt of the text message and the utterance of the command “call him” in order to determine to whom the command is directed to (i.e., to discern who the “him” refers to in the command “call him”).

213. Indeed, Gruber specifically teaches that pronouns in voice commands can be disambiguated based on “current” operations on a mobile device. For example, Gruber teaches that “[t]he virtual assistant of the present invention can ... interpret instructions such as ‘send him a text message’, wherein the ‘him’ is interpreted according to context information derived from a *current* phone call.” Ex-1008 ¶18. Another example is when Gruber’s natural language processing used to parse text and associate the words with syntactic and semantic roles. Ex-1008 ¶20. Gruber provides an example of its natural language process “determining that the user input is about making a phone call to a person referred to by the pronoun ‘her’, and finding a specific data representation for this person” by identifying “her” as the person that the user is text messaging. Ex-1008 ¶20. A POSITA would have understood that Gruber’s consideration of context information derived from an occurrence of an event like a current phone call or text message thus influences evaluation of at least one of the plurality of processing stages, such as its natural language processing stage.

214. Thus, a POSITA, in order to disambiguate the pronoun “him” in the voice command “call him,” would have been motivated to implement Gruber’s teachings to take into account whether receipt of an event (in the above Gruber example, a text message, but not necessarily limited to that) was received contemporaneously with the voice command. This is just like the ’461 Patent,

which provides an event example of the mobile device providing a “reminder to call John,” so that “the user may instruct the mobile device to ‘Call John Doe,’ to initiate a telephone call.” Ex-1001, 23:46-64 (providing certain events that trigger a voice response from the user).

215. The ’461 patent further explains that “various system resources may be activated to, for example, check for recent activity, check for an occurrence of an event, check the battery power, etc. to incrementally wake-up portions of the mobile device on an as-needed basis.” Ex-1001, 16:26-30. A POSITA would have understood that, just like the ’461 patent wakes-up portions of the mobile device to check for an occurrence of an event, Gruber’s teaches to take into account the event and timing of the event.

216. A POSITA would have been motivated to implement Gruber’s teachings to take into account whether the asynchronous events occurred proximate in time with the receipt of audio input, in Sarin-Hwang’s mobile device. For example, doing so would better address events that happen independently of the user’s interaction with the mobile device, and better evaluate the acoustic input for a precise, executable statement. Ex-1008, Abstract, ¶238.

217. Accordingly, the Sarin-Hwang-Gruber combination thus teaches a system that uses events, such as the receipt of a text message (“the at least one contextual cue includes an occurrence of an event on the mobile device”) in which

the combined system takes into account whether the reception of the text message (i.e. the event) was close in time with the receipt of audio containing a voice command (“influencing evaluation of at least one of the plurality of processing stages based on whether the acoustic input is received proximate in time with the occurrence of the event”).

c. Claim 7: The method of claim 1, wherein the at least one contextual cue includes the location of the mobile device, and the method comprises influencing evaluation of at least one of the plurality of processing stages based on determining whether the mobile device is located in at least one of a vehicle, a home or a place of work of the user of the mobile device.

218. Sarin-Hwang-Gruber teaches this claim.

219. Sarin-Hwang-Gruber teaches that the contextual cue includes the location of the mobile device and that location influences the evaluation of a processing stage. As I explained above in Section VIII.A.3.e (Ground 1, Claim 7), Sarin’s consideration of the “location of the mobile device,” is a contextual cue that influences the processing stage of image-capture related voice commands.

220. Although Sarin offers limited examples of the use of location context, a POSITA would appreciate that context is a powerful aspect of speech understanding systems and hence consult Gruber for suggestions of broader applications of location based context to better understand these commands. Thus,

a POSITA would have been motivated to use Sarin's contextual cue and influence evaluation of at least one of the plurality of processing stages based on the location of the mobile device, in view of Gruber. For example, the virtual assistant implemented in the Sarin-Hwang-Gruber mobile device uses "[l]ocation information coming from sensors or location-based systems" as contextual information. Ex-1008 ¶131. Sources of such information may "include Global Positioning System (GPS) and Assisted GPS (A-GPS) on mobile phones." *Id.*

221. Furthermore, "location information is combined with explicit user input" by the virtual assistant in order "to *detect when a user is at home*, based on known address information and current location determination." *Id.* Thus, "*certain inferences may be made about the type of information the user might be interested in when at home*" as opposed to outside the home, as well as the type of services and actions that should be invoked on behalf of the user depending on whether or not he or she is at home." *Id.*

222. Accordingly, a POSITA would have understood the Sarin-Hwang-Gruber combination teaches a virtual assistant that uses location data from a GPS of the mobile device to assist in interpreting voice commands, ("the at least one contextual cue includes the location of the mobile device"), in which the virtual assistant takes into account whether a user is at home in making certain inferences while interpreting the voice command ("influencing evaluation of at least one of

the plurality of processing stages based on determining whether the mobile device is located in .. a home ... of the user of the mobile device”).

223. As I explained in Section V.A, the location of the mobile device is a common contextual cue used in evaluating the acoustic input. For example, Lebeau (Ex-1016) specifically discusses using contextual cues so as to evaluate the likelihood that a user would speak a verbal search request. Ex-1016 ¶¶3, 22 (“the mobile computing device can attempt to infer whether the current context indicates that the user has at least a threshold likelihood of providing voice input and, if so, monitoring for voice input in response.”), 43 (“threshold likelihood of [the user] providing a voice-based request to the mobile computing device”). Lebeau explains that current context associated with a mobile computing device includes the environment around the device, or a context internal to the device such as “the physical location where the mobile computing device is located (e.g., home, work, car, etc., as determined by GPS in the device or other techniques).” Ex-1016 ¶3; *see also* Ex-1016 ¶43 (considering that the user is based at home).

224. Accordingly, it is my opinion that Sarin-Hwang-Gruber’s mobile device uses its GPS location data to assist in interpreting voice commands, (“contextual cue includes the location of the mobile device”), and considers whether a user is at home in making certain inferences while interpreting the voice

command (“influencing evaluation...based on determining whether the mobile device is located in ... a home”).

d. Claim 9: The method of claim 1, wherein the at least one contextual cue includes the sample of the acoustic environment obtained by the mobile device by storing acoustic input previously received by the mobile device and determined not to include a voice command, and the method comprises subtracting the sample of the acoustic environment from the acoustic input.

225. Sarin-Hwang-Gruber teaches this claim.

226. Sarin-Hwang teaches a contextual cue that includes the sample of the acoustic environment of the mobile device and subtracting the sample of the acoustic environment from the acoustic input. Sarin determines whether audio input is a sufficient match to a command depending on “environmental factors,” such as the background noise. Ex-1005, 8:21-25. For example, Sarin’s voice recognition engine listens for specific commands related to the image capture functionality by “quickly dismiss[ing] some sounds (such as background noise or conversation) and focus on a command grammar and vocabulary that is relevant to the current context (e.g., words or phrases that correspond to image capture commands).” Ex-1005, 11:37-41. Sarin thus samples the acoustic environment and subtracts the acoustic environment from the received acoustic input by sampling the background noise and dismissing the background noise to focus on the acoustic

input of the “image capture command.” A POSITA would have understood that “dismissing the background noise” to better focus on the acoustic input would be obvious to do so, in a number of ways, such as noise reduction, i.e. subtracting the background noise. Therefore, a POSITA would have understood that subtracting the background noise is one example of Sarin’s “dismissing the background noise” in order to focus the acoustic input on the part to be further processed.

227. A POSITA would have been motivated to store the acoustic input previously received by the mobile device and determined not to include a voice command as taught by Gruber. Gruber stores acoustic profile data, which can be used as contextual information to interpret speech. Ex-1008 ¶250. Specifically, Gruber’s virtual assistant takes into “account the acoustic environments in which the speech is entered,” by storing “noise profiles” of the user’s location. For example, “noise profiles of a quiet office are different from those of automobiles or public places.” *Id.* The noise profiles are identified and stored, so that the data can be provided as contextual information to “aid in recognition and interpretation of input.” *Id.* The noise profiles are identified and stored as the acoustic environment, which includes sounds but not a voice command.

228. A POSITA would have been motivated to remove the background noise using Gruber’s stored acoustic environment (noise profiles) to “aid in recognition and interpretation of input” by being used as a contextual cue. Ex-1008

¶250. Additionally, Sarin-Hwang's recognition would have been improved by using Gruber's saved noise profiles of the user's location so that when Sarin considers the user's location and compares audio input with other audio data to determine the audio command associated an image capture operation, it can estimate noise models and dismiss the background noise commonly associated with the user's location more efficiently. Ex-1005, 12:38-41.

229. Specifically, a POSITA would have appreciated that different environments and locations a person frequents will have different background noise characteristics, and look to Gruber for the idea that you should have models for each of these environment. For example, U.S. Patent Application Publication No. 20120296655A1 to Kristjansson et al. ("Kristjansson") discusses the advantages of using "stored noise profiles," such as Gruber's. Specifically, Kristjansson teaches that:

"speech recognition is used to transcribe utterances that may have been input in a wide variety of noisy environments, such as when the user is in a crowded room (e.g., a cafeteria), walking down the street, or in the presence of a radio or television. To filter this noise, some noise reduction algorithms require a sample of audio from the environment, without the user's speech. ***By predictively pre-recording audio before the use initiates the voice input (e.g., before the user presses a button), such a background recording becomes available for use by the noise reduction algorithm, thereby improving recognition accuracy.***

Ex-1018 ¶42.

230. Kristjannson further explains that by “invoking the functionalities of the mobile device may include commencing a background audio recording of audio prior to a user's button press. The background audio may be used as input to the speech recognition engine, *to estimate noise models for noise suppression, and/or estimate noise models and/or levels for speech endpoint detection.*” Ex-1018 ¶57. Therefore, a POSITA would have recognized these benefits when implementing Gruber’s saved noise profiles of the user’s location into Sarin-Hwang’s recognition.

231. Moreover, this was well known as I explained in Section V.B. For example, Burnett explained eliminating noise in desired acoustic signals “by estimating the noise waveform and subtracting it from a signal including both speech and noise.” Ex-1026 ¶41. Beckley is another example that explains that “sound from other areas than that of the area in which the identified speaker is located may be subtracted from sounds of the speaker’s area.” Ex-1024 ¶95. Yamabe also notes that typical noise reduction systems use noise spectrum detection, which involves estimating noise spectra and “subtract[ing] a noise component from an input signal including noise, thereby reducing the noise.” Ex-1025 ¶6; *see also* Ex-1015 ¶49 (“[a]dditional filtering may include ...removing background noise”).

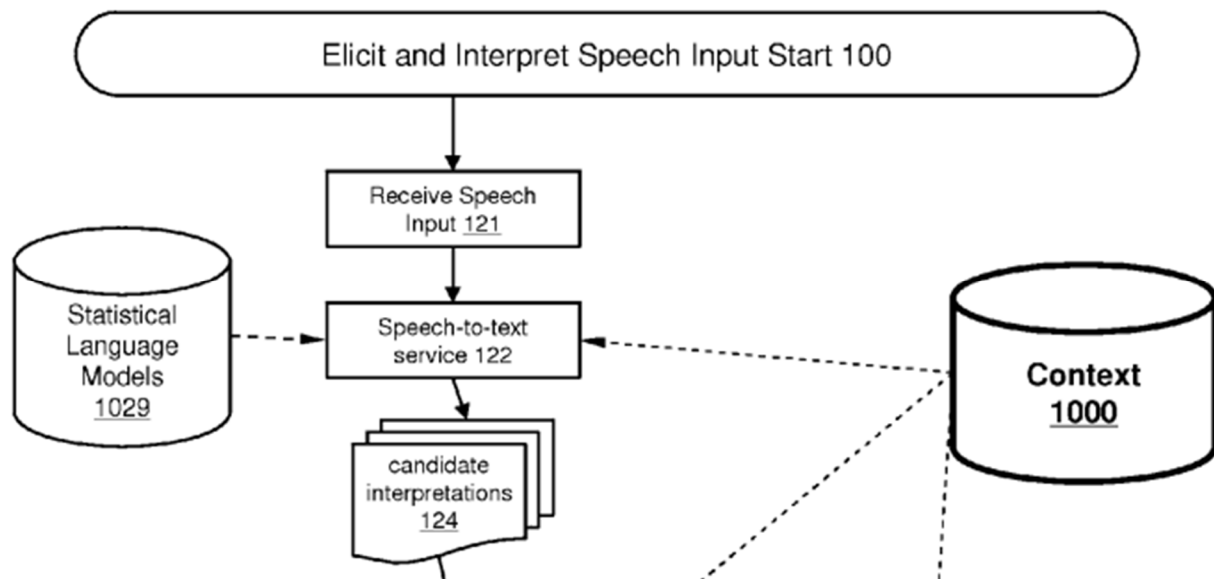
232. Accordingly, it is my opinion that Sarin-Hwang-Gruber teaches to store noise profiles of the user's location(s) which include sound but not a voice command ("the at least one contextual cue includes the sample of the acoustic environment obtained by the mobile device by storing acoustic input previously received by the mobile device and determined not to include a voice command") to estimate noise models and dismiss the background noise commonly associated with the user's location to more efficiently detect a voice command ("subtracting the sample of the acoustic environment from the acoustic input").

e. Claim 10: The method of claim 1, wherein a sequence of performing the plurality of processing stages is determined, at least in part, by the at least one contextual cue.

233. Sarin-Hwang-Gruber teaches this claim.

234. As I explained in Section VIII.A.2.d (Ground 1, Element 1[c]), Sarin-Hwang uses "contextual cues" in recognizing voice commands corresponding to image capture functionality. Moreover, Sarin-Hwang performs a plurality of processing stages. *See* Section VIII.A.2.e (element 1[d]), ("feature extraction (e.g., via mel-band filters 323 and a log converter 325), data transformation (e.g., via a DCT transformer 333), and sound recognition (e.g., via sound recognition modules 340)."). Ex-1006 ¶52.

235. A POSITA would have been motivated to perform Sarin-Hwang’s plurality of processing stages in a sequence based on a contextual cue, in view of Gruber. Gruber’s virtual assistant employs “statistical language models” (“SLMs”) to generate candidate text interpretations of input speech. Ex-1008 ¶317. As shown in Figure 3 below, contextual information (denoted as context 1000) is used by speech-to-text service 122 in order to select appropriate SLMs 1029 to generate candidate interpretations 124 of received input speech 121.



Gruber, FIG. 3 (excerpted)

Specifically, Gruber’s “Personal Acoustic Context Data 1080” is “used to *select from possible SLMs* 1029 or otherwise tune them to optimize for recognized acoustical contexts.” *Id.* ¶332. “Device Sensor Data 1056, describing properties of microphones and/or cameras in use, can be used to *select from possible SLMs* 1029 or otherwise tune them to optimize for recognized acoustical contexts.” *Id.*

¶333. Gruber’s personal acoustic context data includes the acoustic environments, which can be used to “aid in recognition and interpretation of input.” *Id.* ¶250.

Thus, Gruber’s context data is used to determine the statistical language model for further processing the voice input.

236. A POSITA would have understood that Gruber’s selection and tuning of statistical language models by speech-to-text service 122 are “processing stages” that use contextual cues, consistent with the ’461 patent because the selection of appropriate SLMs based on contextual information determines the sequence that these speech-to-text processing stages are performed. For example, after interpreting speech input, Gruber’s teaches to produce multiple candidate interpretations, where each candidate interpretation is based on the application of one or more statistical language models to the speech input. Ex-1008 ¶¶307, 317; FIG. 2 (element 190). Further, Gruber teaches to “process steps, method steps, algorithms or the like may be described in a *sequential* order.” *Id.* ¶63. Thus, selecting additional SLMs based on contextual information impacts the sequence of speech recognition processing performed by the virtual assistant. Ex-1008 ¶321 (noting that an SLM can be *added* and tuned based on contextual information).

237. A POSITA would have been motivated to implement Gruber’s teaching to apply processing sequences based on at least one contextual cue into Sarin-Hwang’s processing stages to improve the speech recognition processing to

further personalize results. Ex-1008, Abstract. For example, Gruber teaches that one benefit of its approach is to provide adaptive processing based on user or “environmental measures” (Ex-1008 ¶¶218, 222, 250), such that the contextual cue will be used to evaluate the processing steps appropriate based on the user’s location or schedule. As I explained in section V.A, using contextual cues to determine how to process the acoustic input was well known. For example, Sarin teaches that contextual clues include “which applications or functions are active on a device, time of day, and location,” such that if the time and day and location indicate that the user is at work, “voice recognition relating to image capture operations can be turned off” so as to avoid unintended camera operations. Ex-1005, 11:32-47. Sarin also describes how the user’s recent activity, such as an active phone call, can be considered when determining whether voice recognition relating to image capture operations can be turned off, so as to avoid unintended camera operation. *Id.*, 11:47-52. This is beneficial because it saves battery by preventing voice recognition relating to image capture operations to use power resources when it is highly unlikely that the user would invoke its mobile device’s camera since the user is at work or otherwise using their mobile device.

238. Similarly, U.S. Patent Application Publication No. 2012/0034904 to Lebeau (“Lebeau,” Ex-1016) titled “Automatically Monitoring for Voice Input Based on Context,” explains to use contextual cues so as to evaluate the likelihood

that a user would speak a verbal search request. Ex-1016 ¶¶3, 22 (“the mobile computing device can attempt to infer whether the current context indicates that the user has at least a threshold likelihood of providing voice input and, if so, monitoring for voice input in response.”), 43 (“threshold likelihood of [the user] providing a voice-based request to the mobile computing device”).

239. Accordingly, in my opinion, the Sarin-Hwang-Gruber combination thus teaches a virtual assistant that selects certain statistical language models during speech recognition (“a sequence of performing the plurality of processing stages is determined”) based on contextual information such as personal acoustic context data (“at least in part, by the at least one contextual cue”).

f. Claim 12: The method of claim 11, wherein at least one second processing stage of the plurality of processing stages is performed by the primary processor only after the at least one first processing stage is performed and neither concludes that the acoustic input includes a voice command nor concludes that the acoustic input is the result of spurious acoustic activity.

240. Sarin-Hwang-Gruber teaches this claim.

241. Sarin-Hwang-Gruber teaches a second processing stage performed by the primary processor only after the first processing stage. As I explained in Section VIII.A.2.e (Element 1[d]), Sarin-Hwang teaches a plurality of processing stages. Specifically, the Sarin-Hwang DSP “may perform feature extraction (e.g.,

243. Gruber teaches that when a user request is ambiguous, Gruber's virtual assistant may undertake a number of actions, "including active elicitation, paraphrasing, suggestions, and the like, and including obtaining context information, so that the correct services ... are called and the intended action taken." Ex-1008 ¶88. Gruber "may *prompt the user for confirmation and/or request additional context information* from any suitable source before calling a service ... to perform a function." *Id.*

244. It is my opinion that Sarin-Hwang-Gruber thus would have been implemented to have a second stage of processing when the first stage of processing is not sufficient to specifically identify the voice command, in order to request additional information to conclude that the acoustic input indeed includes a voice command.

245. Accordingly, it is my opinion that Sarin-Hwang-Gruber teaches a second processing stage to determine that an audio input includes a voice command ("at least one second processing stage of the plurality of processing stages is performed by the primary processor") after a first processing stage is unable to definitively identify the command ("only after the at least one first processing stage is performed and neither concludes that the acoustic input includes a voice command nor concludes that the acoustic input is the result of spurious acoustic activity").

g. Claim 15: The method of claim 11, wherein performing the at least one second processing stage comprises performing at least one speech processing stage including at least one of automatic speech recognition on the acoustic input and at least one natural language processing stage to understand the meaning of the voice command.

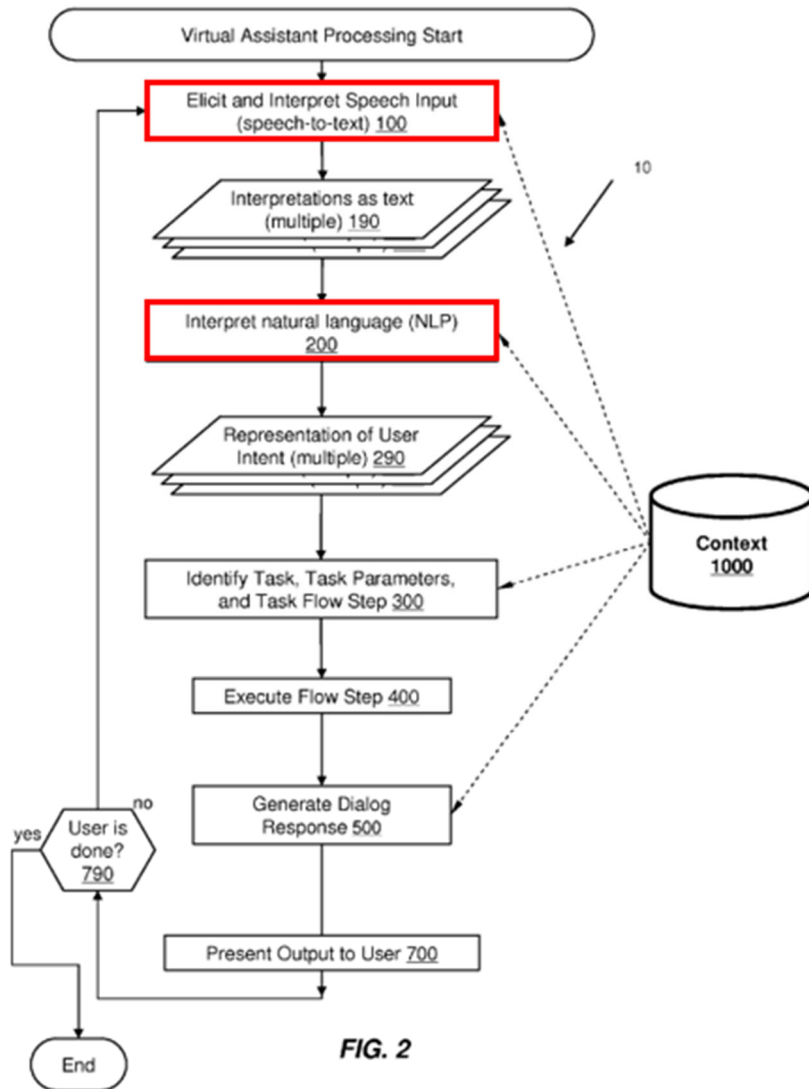
246. Sarin-Hwang-Gruber teaches this claim.

247. Sarin-Hwang teaches performing a second processing stage of automatic speech recognition on the acoustic input to understand the meaning of the voice command, as I explained in Section VIII.A.2.e (Element 1[d]). *See* Ex-1006 ¶52 (“feature extraction (e.g., via mel-band filters 323 and a log converter 325), data transformation (e.g., via a DCT transformer 333), and sound recognition (e.g., via sound recognition modules 340).”). Hwang specifically teaches a number of “sound detection modules” and in particular the “continuous keyword detection module 243.” Ex-1006 ¶41.

248. A POSITA would have been motivated to perform the second processing stage of automatic speech recognition and natural language processing to understand the meaning of the voice command, because doing so was well known, as taught by Gruber. As shown in Figure 2 below, at step 100 “[a]ssistant 1002 receives 121 voice or speech input in the form of an auditory signal. *A speech-to-text service 122 or processor generates a set of candidate text interpretations 124 of the auditory signal.*” Ex-1008 ¶316. A POSITA would have

understood that the speech-to-text processing performed by the virtual assistant on the primary processor as entailing automatic speech recognition. For example, U.S. Patent Application Publication No. 2013/0173269 (“Adler”) teaches that automatic speech recognition (ASR) transcribes speech data to text (e.g., textual words). Ex-1012 ¶92.

249. At Gruber’s step 200, “[t]he set of candidate interpretations” produced at step 100 is “processed ... by language interpreter 2770 (*also referred to as a natural language processor, or NLP*), which parses the text input and generates a set of possible interpretations of the user’s intent.” Ex-1008 ¶308.



Gruber, FIG. 2 (annotated)

250. A POSITA would have been motivated to implement Gruber's automatic speech recognition and natural language process into Sarin-Hwang's speech recognition that uses speech models in order to better parse the input and generate a set of possible interpretations of the user's intent by providing additional processing power by using Gruber's virtual assistant on Sarin-Hwang's primary processor. Moreover, a POSITA would have understood the advantages of

implementing Gruber's virtual assistant on Sarin-Hwang's more powerful primary processor because a POSITA would have understood that Sarin-Hwang's low power DSP would not have sufficient processing power to implement Gruber's speech models.

251. Accordingly, the Sarin-Hwang-Gruber combination thus teaches to perform processing in several stages, including a speech-to-text processing step ("performing at least one speech processing stage including at least one of automatic speech recognition on the acoustic input") and a natural language processing step ("and at least one natural language processing stage to understand the meaning of the voice command").

h. Claim 17: The method of claim 1, comprising performing one or more default actions when a specific task could not be ascertained from the acoustic input.

252. Sarin-Hwang-Gruber teaches this claim.

253. Sarin-Hwang-Gruber teaches performing a default action when a specific task could not be ascertained from the acoustic input. Specifically, when a user request is ambiguous or needs further clarification, Gruber teaches to undertake a number of actions, "including active elicitation, paraphrasing, suggestions, and the like, and including obtaining context information, so that the correct services ... are called and the intended action taken." Ex-1008 ¶88. In one

embodiment, the virtual assistant “may *prompt the user for confirmation and/or request additional context information* from any suitable source before calling a service ... to perform a function.” *Id.*

254. These actions are consistent with “default actions” disclosed in the ’461 patent that are performed when a detected voice command “cannot be suitably mapped to a specific task or set of tasks.” Ex-1001, 37:27-28. For example, “the default action may include *formulating a question to be rendered to the user to obtain further information* and/or clarification regarding the voice command.” *Id.*, 37:30-33.

255. A POSITA would have been motivated to implement a default action when a specific task could not be ascertained from the acoustic input, as taught by Gruber, in Sarin-Hwang’s mobile device in order to request additional information in so that the task can be performed.

256. Accordingly, the Sarin-Hwang-Gruber combination thus teaches a virtual assistant that prompts the user to confirm the request and/or requests additional contextual information (“performs one or more default actions”) when a user request is found to be ambiguous, (“when a specific task could not be ascertained from the acoustic input”).

**D. Ground 4: Sarin in Combination with Hwang and Lebeau
Renders Obvious Claim 8**

**1. A POSITA would have been motivated to combine
Sarin's, Hwang's, and Lebeau's teachings and would
have had a reasonable expectation of success.**

257. A POSITA would have been motivated to combine the teachings of Sarin, Hwang, and Lebeau ("Sarin-Hwang-Lebeau") and would have had a reasonable expectation of success in doing so because each relates to the same well-known technologies.

258. A POSITA also would have been motivated to combine Lebeau with Sarin-Hwang because they are analogous art and teach improvements to the same type of devices. Each teach an electronic device in a first mode of operation which then switches to a second mode of operation. Ex-1005, 13:38-46 (image capture-related voice commands), 13:53-62 (transitions from a low-power state to an active state); Ex-1006 ¶¶5 (low-power system for receiving audio), ¶31 (based on analysis of the audio feature data, the DSP may generate an activation signal 160 to activate various components of an electronic device); Ex-1016, Abstract, ¶¶7-9 ("switch to the second mode of operation so that the mobile computing device receives a stream of audio data"). Each teach using speech processing to determine a voice command. Ex-1005, 6:29-49, 13:15-20; Ex-1006 ¶¶52; Ex-1016 ¶¶7-9 ("monitors ambient sounds for voice input that indicates a request to perform an operation").

259. Lebeau offers additional benefits to Sarin-Hwang's mobile device.

Lebeau teaches that the mobile device will detect the mobile device's condition in order to determine whether to switch the device to monitor for voice input that indicates a request to perform an operation. Ex-1016 ¶9. Lebeau's device attempts to infer whether, based on the current context, monitoring for voice input would provide at least a threshold level of convenience for the user and, if so, monitor for voice input. Ex-1016 ¶23. Lebeau explains that a context includes whether the mobile device's battery is "running low on stored power (e.g., below 25% charge remaining in the battery)" (Ex-1016 ¶24) or whether the mobile device is receiving power from a power cord. Ex-1016 ¶30. Lebeau's device determines whether the "potential convenience to the user of monitoring for voice input is outweighed by the inconvenience to the mobile computing device...of potentially draining the battery in a relatively short period of time (short when compared to standby operation)." Ex-1016 ¶31.

260. A POSITA would have recognized that considering the mobile device's battery when processing a voice input is desirable in a mobile device that runs on a battery, as maintaining the primary processor in a deactivated state until the processor is needed would tend to conserve power usage and extend the battery's life, which is a goal of Sarin-Hwang. Ex-1005, 11:32-34.

261. A POSITA would have had a reasonable expectation of success in combining Lebeau with Sarin-Hwang. The electronic components of Lebeau with Sarin-Hwang's devices are similar and are interchangeable. A POSITA would have understood, therefore, that these references disclose interrelated teachings based on routine technologies that would have been amenable to various well-understood and predictable combinations to provide quick and efficient speech recognition and processing.

2. Dependent Claim 8

- a. Claim 8: The method of claim 1, wherein the at least one contextual cue includes the power source of the mobile device, and the method comprises influencing evaluation of at least one of the plurality of processing stages based on at least one of determining whether the mobile device is being powered using a battery or wall power, and a state of the battery if the mobile device is being powered using the battery.**

262. Sarin-Hwang-Lebeau teaches Claim 8.

263. Sarin-Hwang-Lebeau teaches that the mobile device's power source is a contextual cue that influences the processing stage's evaluation based on the mobile device being powered using the battery. As discussed above in §VIII.A.2.d, Sarin-Hwang uses contextual cues. A POSITA would have been motivated to use Sarin-Hwang's contextual cue and influence evaluation of at least one of the plurality of processing stages based whether the mobile device is being powered

using a battery or wall power, and a state of the battery if the mobile device is being powered using the battery, in view of Lebeau.

264. Lebeau teaches to consider “a battery for the mobile computing device running low on stored power (e.g., below 25% charge remaining in the battery” when determining whether “to switch the mobile computing device from a current mode of operation to a second mode of operation” based on a voice input indicating a request. Ex-1016, Abstract, ¶24. Lebeau also teaches that the mobile device can be receiving power from a power cord that is plugged into an outlet. Ex-1016 ¶30. Therefore, Lebeau considers the mobile device’s power source and allows the battery’s status to influence the evaluation performed in the processing stage by not performing voice recognition if the battery is low but performing voice recognition when the device is powered using wall power.

265. A POSITA would have been motivated to consider the mobile device’s battery status, as taught by Lebeau, as one of Sarin-Hwang’s contextual cues used during to process the acoustic input because doing so would allow the device to determine whether the “potential convenience to the user of monitoring for voice input is outweighed by the inconvenience of the mobile computing device ... potentially draining the battery in a relatively short period of time.” Ex-1016 ¶31. For example, as I explained in Section V.A, battery status is a well known consideration in a speech recognition system and is “particularly

advantageous for computing devices that rely on battery power, all computing devices for which power management may be desirable can benefit from the principles of the present disclosure.” Ex-1019 ¶14. Basye also notes that by keeping the computing device in a persistent active state, the first computing device may consume unacceptable amounts of energy to maintain the persistently active state. Ex-1019 ¶9. “Such energy demands may prove especially problematic for mobile computing devices that rely on battery power.” *Id.* Thus, Basye shows how it was well known to consider the battery status, such that “if the power supply 218 drops below a threshold power supply level (e.g., a battery charge percentage), as on-device speech recognition capabilities as implemented by the processing unit 202 and/or the application processing module 112 may require a relatively large power draw.” *Id.* ¶79.

266. Accordingly, it is my opinion that Sarin-Hwang-Lebeau teaches considering whether the device is being powered by a battery or wall power (“the at least one contextual cue includes the power source of the mobile device”) and not performing voice recognition if the device is being powered by the battery and it is running low, but performing voice recognition when the device is powered using wall power (“influencing evaluation of at least one of the plurality of processing stages based on at least one of determining whether the mobile device is

being powered using a battery or wall power, and a state of the battery if the mobile device is being powered using the battery”).

IX. SECONDARY CONSIDERATIONS

267. I am not aware of any evidence in the '461 patent's prosecution history or elsewhere supporting any secondary considerations arguments, or evidence of nexus of such alleged evidence to the Challenged Claims. *See generally* Ex. 1004. To the extent Patent Owner asserts the existence of any secondary considerations in its responses, I reserve the right to address any such evidence.

X. CONCLUSION

268. I declare that all statements made herein of my knowledge are true, that all statements made on information and belief are believed to be true, and 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.

U.S. Patent No. 11,393,461
Declaration of Mr. Christopher M. Schmandt
Respectfully submitted,

Date: 4 October 2024

A handwritten signature in black ink, appearing to read "Chris Schmandt", written in a cursive style.
