UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

AMAZON.COM, INC.,
AMAZON.COM SERVICES LLC, and
AMAZON WEB SERVICES, INC.,
Petitioners,

v.

SOUNDCLEAR TECHNOLOGIES LLC,
Patent Owner.

Case No. IPR2025-00673
U.S. Patent No. 11,244,675

REPLY DECLARATION OF RICHARD STERN, PH.D.

Amazon v. SoundClear
US Patent 11,244,675
Amazon EX-1202

TABLE OF CONTENTS

I. CLAIMS 1-7 WOULD HAVE BEEN OBVIOUS ----------1

A. Claim Construction----------1

B. Obviousness Generally ----------3

1. The References Disclose Classifying a Voice Based on Calculated Distance----------3

C. Shin and Aoyama----------4

1. Shin Discloses Classifying Voice Based on Distance ----------4

2. Shin and Aoyama Disclose Replacing Words ----------7

a. Shin ----------7

b. Motivation to Combine Shin and Aoyama --------7

3. Shin Discloses Classifying Voice When the User Whispers ----------9

D. Shimomura and Aoyama ----------9

1. Shimomura Discloses Voice Classification ----------9

E. Whisper ---------- 11

II. PROPOSED SUBSTITUTE CLAIM 8 ---------- 11

A. Observations Regarding Proposed Substitute Claim 8 -------- 12

1. The Claim Recites Only Collecting Information, Analyzing Information, and Presenting Results -------- 12

2. The Claim Recites Results Without Reciting Solutions ---------- 14

3. The Claim Is Analogous to Human Activity ---------- 15

4. The Claim Relies on Conventional Components ------- 16

B. Proposed Substitute Claim 8 Is Indefinite ---------- 16

1. "a predetermined voice" ---------- 17

2. "whether a peak frequency is equal to or lower than a threshold" ---------- 17

C. Proposed Substitute Claim 8 Lacks Written Description and Enablement ---------- 19

D. Proposed Substitute Claim 8 Would Have Been Obvious ---- 20

1. Proposed Substitute Claim 8's New Limitation Was Known ---------- 21

a. Raitio ---------- 22

b. Chen ---------- 24

c. Additional References ---------- 24

2. The References Disclose or Suggest Determining Whisper Based on Distance and Peak Frequency ------ 26

3. Motivation to Combine ---------- 26

I, Richard Stern, Ph.D., do hereby declare:

1. In preparing this Declaration, I considered the Declaration of Joseph McAlexander (EX-2016), the exhibits cited therein, and all other references identified herein.

2. I understand that my prior declaration was submitted as Exhibit 1002, and I refer to certain portions of that declaration herein. I remain unaware of any evidence that would suggest that the claims of the ’675 patent would not have been obvious.

3. I understand that Mr. McAlexander has applied a slightly different definition than I have for a person of ordinary skill in the art. (EX-2016 ¶34.) While I disagree with Mr. McAlexander as to that definition, my opinions would not change under his definition.

4. In this Declaration, I provide opinions and enumerate ways in which I disagree with Mr. McAlexander and his opinions as stated in his Declaration (EX-2016). However, this Declaration does not enumerate all ways in which I disagree with Mr. McAlexander, and the fact that I do not address some of Mr. McAlexander’s statements should not be taken to mean that I agree with them.

I. CLAIMS 1-7 WOULD HAVE BEEN OBVIOUS.

A. Claim Construction

5. In Paragraph 39 of his declaration, Mr. McAlexander states that in one

embodiment, the voice classifying unit can "determine if a voice is a whisper or not and classify the voice accordingly." (EX-2016 ¶39.) In Paragraph 40, he notes that "[o]ther embodiments of the voice classifying unit use the … distance between the speaker and the device[.]" (*Id.* ¶40.) In Paragraph 45, he discusses that the specification describes using the "user-to-device distance" and a threshold "for the purpose of distinguishing one class of voice from another." (*Id.* ¶45.) I agree with these statements and they confirm my opinion, stated in my prior declaration, that the voice classifying unit may classify the user's voice based solely on the user-to-device distance. (EX-1002 ¶59.)

6. The '675 patent also confirms my opinion that the classification can be based solely on the measured distance. The plain language of the claims requires calculating the user-to-device distance "to classify the voice into either a first voice or a second voice based on the calculated distance." (EX-1001 ('675 patent), claim 1.) The claims do not require the classification to be based on anything other than distance. Nothing in the claims precludes classification based on distance alone. The specification also explicitly discloses classifying based on distance alone. (*Id.*, 10:32-38 ("the distance can be used as a feature value to perform the classification to the first voice V1A and the second voice V1B"), 10:42-46 ("the voice classifying unit 38 sets a threshold of the feature value, and classifies the voice V1 … based on whether the feature value exceeds the threshold").)

B. Obviousness Generally

1. The References Disclose Classifying a Voice Based on Calculated Distance.

7. In Paragraph 71, Mr. McAlexander discusses my testimony that the prior art discloses measuring variables such as distance and modifying content based on those variables. (EX-2016 ¶71 (citing EX-2017, 95:12-17).) My testimony is entirely consistent with the prior art's disclosure of voice classification.

8. I understand the Patent Owner has argued that the references in my prior declaration "only classify *distances* and not voices." I disagree. The prior art discloses devices that acquire a user's voice, calculate the user-to-device distance, and classify the voice based on the distance. That is exactly what the '675 patent does.

9. In Paragraph 75, Mr. McAlexander states that the prior art's "measurements or processing can be completely independent of one another." (EX-2016 ¶75.) But the references explicitly disclose that the portions he claims are independent are actually used together. For example, Shin expressly discloses that the embodiments described in Tables 1 and 3 may be used together. (*E.g.*, EX-1003 (Shin) ¶¶[0049] (disclosing determining "the output scheme" which "may include an output volume level … and an output amount of information"), [0053], [0140], claim 17 (disclosing "adjusting at least one of an output volume level of the content … or an output amount of information of the content based on the distance between the user and the

electronic device"); *see also id.* ¶¶[0183]-[0184].) And Shimomura's device receives a "character string" of potential output, and based on the distance, classifies the input, determines whether to alter the output string, and then passes the resulting (potentially altered) string to be output. (EX-1004 (Shimomura) ¶¶[0059]-[0066], Figs. 7-8.) Then, it again uses the distance-based classification to determine whether to alter the output volume. (*Id.* ¶¶[0066]-[0070], Figs. 9, 13.) Shimomura explicitly discloses that the first process is the input to the second; thus, they are used in conjunction.

10. In Paragraph 165, Mr. McAlexander states that "the authors of Shin and Shimomura … purposely chose simpler, more discrete approaches" than that of the '675 patent, leading a person of ordinary skill to disfavor modifying them. (EX-2016 ¶165.) I disagree. The references disclose classifying voice input by distance in the same way as the '675 patent.

C. Shin and Aoyama

1. Shin Discloses Classifying Voice Based on Distance.

11. In Paragraphs 82-93, Mr. McAlexander opines that Shin does not disclose voice classification. (EX-2016 ¶¶82-93.) I disagree. Shin expressly discloses "analyz[ing] a voice input[.]" (EX-1003 (Shin) ¶[0048].) Indeed, Shin discloses "receiv[ing] a speech of a user *as a voice input* … to determine an output scheme …

from *an analysis of the voice input*[.]" (*Id.*, Abstract (emphases added).) Shin discloses analyzing a voice input—for example, evaluating speech rate, volume, presence of a keyword, and/or user-to-device distance—to alter the output, just as in the '675 patent. (*See, e.g.*, *id.* ¶¶[0049], [0053].) Shin's Figures also plainly disclose analyzing a voice spoken by a user:

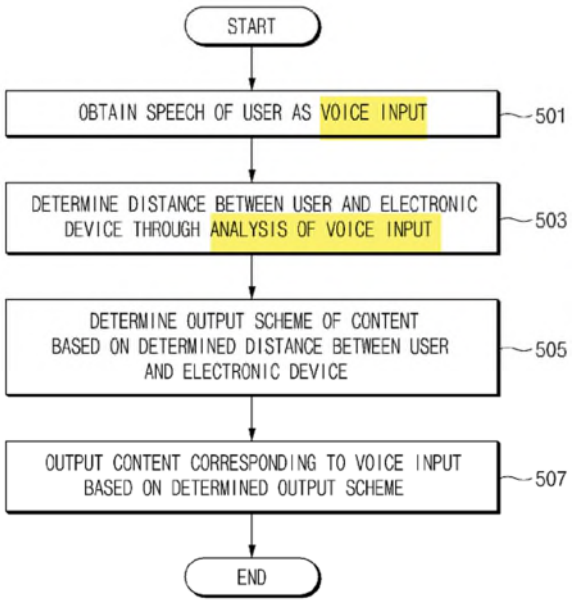


FIG.5A

(*Id.*, Fig. 5A (annotated).)

12. In Paragraph 84, Mr. McAlexander opines that Shin's Table 3 "discloses only a calculation of distance and nothing else[.]" (EX-2016 ¶84.) I disagree. Table 3 does not disclose *calculation* of distance. Shin discloses that the distance is calculated by the proximity sensor and, optionally, the processor. (*See, e.g.*,

EX-1003 (Shin) ¶¶[0075]-[0077], claim 9.) Rather than calculations, Table 3 expressly discloses *classifications based on distance*.

TABLE 3

Distance (D) between user and electronic device	Output amount of information of content
D < 1 m	Amount of information of original content
1 m ≤ D < 2 m	Content of first abbreviation level
2 m ≤ D < 3 m	Content of second abbreviation level
3 m ≤ D < 4 m	Content of third abbreviation level
D ≥ 4 m	Content of fourth abbreviation level (abstract content)

(*Id.*, Table 3 (consolidated).) The left column of Table 3 includes several categories (or classifications) into which the distance D may fall. (*Id.*) And the right column of Table 3 describes the result of each such classification on the device's output. (*Id.*)

13. This is exactly how the '675 patent describes voice classification. The '675 patent describes two classes: (1) $D \leq x$; and (2) $D > x$, where x is a threshold value. (EX-1001 ('675 patent), 10:42-46.) In the '675 patent, the device receives a voice, calculates the user-to-device distance D, determines which distance-based category (class) the received voice belongs into, and the appropriate output is provided. (*Id.*, 11:5-11.) Shin discloses the same thing. The only difference is that the '675 patent gives the two classes generic names: "V1A" (or "first voice") and "V1B" (or "second voice"), whereas Shin does not expressly refer to the classes' *names*. But Shin still discloses the classifications.

2. Shin and Aoyama Disclose Replacing Words.

a. Shin

14. In Paragraph 95, Mr. McAlexander states that "Shin provides no indication that the word 'abbreviation' means or includes 'replacing words'" in output content. (EX-2016 ¶95.) But, as I previously testified, a person of ordinary skill in the art would have readily understood exactly that. (EX-1002 ¶92; *see* EX-1207 (Merriam-Webster's Collegiate Dictionary), 2 (defining "abbreviation" as "a shortened form of a written word or phrase used in place of the whole").) For example, a person of ordinary skill would have understood Shin to disclose that the original output content may include the full name of a place (e.g., "Patent and Trademark Office") and the abbreviated content may include an abbreviated reference to that place (e.g., "PTO"), which involves word replacement. Thus, to a person of ordinary skill, Shin discloses replacing words.

15. In Paragraph 96, Mr. McAlexander states that, in Shin, "abbreviating content involves excluding certain portions of output content." (EX-2016 ¶96.) That is true in the example he cites. But the mere inclusion of that example in Shin does not limit its disclosure to *only* abbreviating output by excluding words rather than using abbreviations or otherwise replacing words.

b. Motivation to Combine Shin and Aoyama

16. In Paragraph 99, Mr. McAlexander states that "Aoyama teaches that

words are *only* to be chosen in response to 'speaking style,'" and thus "[a] POSITA would have understood Aoyama as teaching away from replacing words based on distance." (EX-2016 ¶99.) Mr. McAlexander is incorrect for several reasons.

17. First, Aoyama does not teach that word replacement occurs *only* in response to speaking style; rather, Aoyama teaches speaking style as one option for determining word replacement. (*See, e.g.*, EX-1010 (Aoyama) ¶¶[0128] ("in the example illustrated in Fig. 3"), [0162] ("The example of the functional configuration of the information processing device 1 according to the embodiment has been described with reference to Figs. 2 to 7").)

18. Second, Mr. McAlexander states that word replacement based on distance "would have led to a response with which 'the user may feel uncomfortable,'" citing an example in Aoyama of a device responding in military time to a user who prefers normal time format. (EX-2016 ¶99.) But, as I discussed in my prior declaration, Aoyama discloses many types of word replacement, including replacing full names with nicknames and replacing complicated date formats with simpler ones. (EX-1002 ¶95.) Such replacements would have made the device's output easier to understand at greater distances. (*Id.* ¶97.)

3. Shin Discloses Classifying Voice When the User Whispers.

19. In Paragraph 106, Mr. McAlexander states that claim 5 "requires a determination that 'the voice spoken by the user is a whisper.'" (EX-2016 ¶106.) I disagree. Claim 5 requires that the device be configured to determine that "the voice" is "the predetermined voice" in circumstances where the user whispers to the device. (EX-1001 ('675 patent), claim 5.) Nothing in the claim requires the device to *determine* whether the received voice is a whisper.

20. The portion of the specification Mr. McAlexander cites does not address the same situation. (EX-2016 ¶106; *compare* EX-1001 ('675 patent), claim 5 ("when the voice spoken by the user is a whisper") *with id.*, 9:64-10:1 ("when the voice V1 is determined to be a whisper as a result of analysis").) Thus, Mr. McAlexander's testimony is premised on an improperly narrow reading of the claim. I also note that, as I discuss below and in my prior declaration, the claim is still invalid under Mr. McAlexander's reading. (EX-1002 ¶¶225-231.)

D. Shimomura and Aoyama

1. Shimomura Discloses Voice Classification.

21. In Paragraph 126, Mr. McAlexander argues that Shimomura's "Figure 7 discloses one and only one thing: a calculation of distance." (EX-2016 ¶126.) I

disagree. Shimomura's Figure 7 discloses classification of voice input based on distance. I note that in Paragraph 125, Mr. McAlexander acknowledges that Figure 7 discloses "classes of distances." (*Id.* ¶125.) Indeed, Figure 7 describes classes or categories of distances *from which the voice is received.* Thus, Shimomura discloses classifying a received voice based on whether the distance to the user exceeds a threshold (e.g., 350cm), exactly as the '675 patent describes classifying a voice based on distance. (EX-1001 ('675 patent), 10:42-46 ("the voice classifying unit 38 … classifies the voice V1 … based on whether the feature value exceeds the threshold").) The top row of Figure 7 includes two discrete classifications into which the voice may fall. (EX-1004 (Shimomura), Fig. 7.) And the bottom row describes the result of each such classification on the device's output. (*Id.*)

22. In Paragraph 129, Mr. McAlexander concludes that Shimomura "directly correlates the output content and volume based on distance without performing the step of classifying the voice." (EX-2016 ¶129.) I disagree. Shimomura uses the distance to classify the received voice as coming from less than a threshold distance or more than a threshold distance. The result is used to determine the content and volume of the output. Again, this is precisely what the '675 patent describes and claims. (EX-1001 ('675 patent), 10:42-46.)

E. Whisper

23. The ’675 patent contemplates using microphone signal analysis such as volume or peak frequency to classify voice alongside distance determined by a proximity sensor. (EX-1001 (’675 patent), 9:58-10:46, 11:1-4 (classification can be performed “by combining the classification methods explained above”).) The ’675 patent describes no problems that had to be overcome in doing so and provides no guidance as to how the variables should be analyzed in combination. (*See id.*) However, I note that a person of ordinary skill in the art at the time would have understood how to implement a system that used microphone signal analysis to detect whisper and would have reasonably expected success in doing so. I discuss several references below that disclose exactly that. (*See* ¶¶52-59, below.)

II. PROPOSED SUBSTITUTE CLAIM 8

24. I understand the Patent Owner has proposed a substitute claim in the event claim 5 is found unpatentable. The substitute claim I understand Patent Owner has proposed is listed below, with new elements underlined:

8. (substitute for claim 5 if found unpatentable)
An output-content control device comprising:
- a voice acquiring unit configured to acquire a voice spoken by a user;
- a voice classifying unit configured to calculate a distance between the user and the output-content control device by a proximity sensor to classify the voice into either a first voice or a second voice based on the calculated distance;
- an intention analyzing unit configured to analyze the voice acquired by the voice acquiring unit to detect intention information indicating what kind of information is wished to be acquired by the user;

a notification-information acquiring unit configured to acquire notification information which includes content information as a content information to be notified to the user based on the intention information; ~~and~~

an output-content generating unit configured to generate an output sentence as sentence data to be output to the user based on the notification information, wherein the output-content generating unit is further configured to generate, when the voice is determined to be the first voice, a first output sentence in which at least one word selected among words included in the content information of the notification information is replaced with another word, and to generate, when the voice is determined to be the second voice, a second output sentence which includes all of the intention information and the content information; and

the voice classifying unit is further configured to determine the voice to be ~~the~~ a predetermined voice when the voice spoken by the user is determined to be a whisper based upon whether a peak frequency is equal to or lower than a threshold.

A. Observations Regarding Proposed Substitute Claim 8

1. The Claim Recites Only Collecting Information, Analyzing Information, and Presenting Results.

25. The limitations of proposed substitute claim 8 reflect nothing more than collecting information, analyzing the information, and presenting results.

26. First, the claim recites a unit configured to "acquire a voice spoken by a user." This unit merely collects information of a particular type (i.e., sound).

27. Second, the claim recites a "voice classifying unit" configured to calculate the distance between the user and the device. Because "calculating" requires numerical inputs, which are a type of information, this step requires analyzing information to determine a distance. Although the claim recites a "proximity sensor" for

collecting the distance information, the patent does not identify the proximity sensor as a focus of the claimed advance over the prior art. Indeed, the specification provides no explanation of how the sensor operates. (EX-1001 ('675 patent), 10:32-39.) The sensor is undisputedly conventional and generic.

28. That unit is also configured to "classify the voice … based on the calculated distance." The specification describes only one way of doing so: by comparing the calculated distance to a threshold value and classifying the voice into one category if the distance is above the threshold, and into another category if the distance is below the threshold. (EX-1001 ('675 patent), 10:17-38.) This process also requires merely analyzing information to compare the distance to the threshold value.

29. The voice classifying unit is further configured to "determine the voice to be a predetermined voice" when the voice is "determined to be a whisper based upon whether a peak frequency is equal to or lower than a threshold." This step merely analyzes information (frequency characteristics and threshold).

30. Third, the claim recites a unit configured to "analyze the voice" to "detect intention information[.]" The claim does not provide any detail on how this analysis occurs. Thus, this unit merely analyzes the collected information.

31. Fourth, the claim recites a unit configured to "acquire notification information" based on the analysis of the acquired voice. This step simply refers to collecting additional information.

32. Fifth, the claim requires a unit configured to "generate" output information (an "output sentence") to provide to the user. The output is generated based on the analysis of the voice and how the voice is classified—certain information is provided if the voice is classified as a "first voice" and additional information is provided if the voice is classified as a "second voice." This unit is merely analyzing information and presenting the results to the user.

33. Thus, the claim as a whole focuses on collecting information, analyzing the information, and presenting the results.

2. The Claim Recites Results Without Reciting Solutions.

34. Proposed substitute claim 8 fails to specify *how* to achieve the results the claim recites. For example, the claim recites calculating a distance "to classify the voice," "analyz[ing] the acquired voice" to detect intent information, "acquir[ing] notification information" based on the intent information, "generat[ing]" a first output sentence wherein "at least one word selected among words included in the content information of the notification information is replaced with another

word," and "generat[ing]" a second output sentence which "includes all of the intention information and the content information," but never explains how any of these functions are performed. Rather, the claim recites only generic "units" configured to produce the desired results.

35. The proposed amendment merely adds a step of determining whether the voice is a whisper based on a threshold. Even then, it does not recite what the threshold should be or *how* the threshold is determined to facilitate the whisper/non-whisper determination.

3. The Claim Is Analogous to Human Activity.

36. Proposed substitute claim 8 is analogous to a listener choosing how to respond to a friend. The friend asks where the listener would like to meet. If the friend is close, the listener responds that they are meeting at "the Patent & Trademark Office" (first output sentence). If the friend is far away, the listener responds that they are meeting at "the agency" (second output sentence). This allows the listener to shout without informing other people who could hear where the two are meeting. As in claim 8, the response is chosen by classifying a voice based on distance. Furthermore, as in claim 8, the listener can also determine whether the friend is whispering and, if so, may respond with the first output sentence.

37. This analogy demonstrates that claim 8 focuses on a fundamental practice that humans have long used to communicate: tailoring the information provided based on the other person's proximity.

4. The Claim Relies on Conventional Components.

38. Claim 8 recites only generic processing "units." It is technology agnostic.

39. The only other hardware recited in claim 8 is a generic, known proximity sensor. I understand that the Patent Owner has asserted that "[t]he use of distance calculation to classify voices in connection with such voice-control devices was not conventional." I disagree. Proximity sensors were available and conventional as a way to calculate distance between a user and a device.

40. Even if claim 8 had recited other hardware, the specification confirms that only conventional components, such as a microphone, speaker, and CPU are used to perform the claimed methods. (EX-1001 ('675 patent), 3:27-28, 3:36-39, 3:55-56.)

B. Proposed Substitute Claim 8 Is Indefinite.

41. I understand that a patent is invalid for indefiniteness if its claims, read in light of the specification delineating the patent, and the prosecution history, fail to inform, with reasonable certainty, those skilled in the art about the scope of the invention.

1. "a predetermined voice"

42. Proposed substitute claim 8 recites that the voice classifying unit is configured to "determine the voice to be a *predetermined voice* when the voice spoken by the user is determined to be a whisper[.]"

43. A person of ordinary skill in the art would not know what "a predetermined voice" means. Limitation 8[b] states that the voice is classified as "a first voice" or "a second voice" based on distance. It is unclear whether the "predetermined voice" refers to the first voice, second voice, or some other category of voice. Nor does the claim provide any indication of how such a voice is "predetermined."

44. The specification provides no clarification. It discloses an embodiment in which "[t]he first voice V1A corresponds to a predetermined voice" and "classifies the voice V1 to the first voice V1A when the voice V1 is determined to be a whisper as a result of analysis (determines as the predetermined voice)." (EX-1001 ('675 patent), 9:58-10:1.) However, the specification does not explain what "a predetermined voice" is or whether or when it corresponds to the first and second voices, including when the classification as a first or second voice is based on distance.

2. "whether a peak frequency is equal to or lower than a threshold"

45. A POSITA would not know the boundaries of this claim limitation. The only relevant disclosure in the specification is as follows:

> [T]he voice classifying unit 42 acquires a time waveform of intensity of the voice V1, and performs Fourier-transformation on the time waveform, thereby acquiring a spectrum that indicates intensity of the voice V1 per frequency. The voice classifying unit 42 classifies the voice V1 either to the first voice V1A or the second voice V1B by using a peak frequency that is equal to or higher than a predetermined intensity in the spectrum as a feature value. For example, the voice classifying unit 42 determines the voice as a whisper to classify to the first voice V1A when the peak frequency is equal to or lower than the threshold, and determines the voice as not a whisper to classify to the second voice V1B when the peak frequency is larger than the threshold.

(EX-1001 ('675 patent), 10:13-22; EX-1007 (File History), 143.) Thus, the specification describes classifying a voice by using a "peak frequency that is equal to or higher than a predetermined intensity in the spectrum as a feature value." I note that a person of ordinary skill in the art would have understood "peak frequency" to mean the frequency at which the energy, power, intensity, or amplitude (which are equivalent measures in this context) is maximum. A person of ordinary skill would thus have understood the specification to mean that the voice is classified based on whether the intensity at the peak frequency is equal to or lower than a threshold. Otherwise, it would require comparing a frequency (measured in Hertz) to an intensity threshold (measured in decibels), which is not possible.

46. Therefore, a person of ordinary skill would have understood that the claim limitation *includes* determining a whisper based upon whether the peak frequency's *intensity* is equal to or lower than an *intensity* threshold. However, a person of ordinary skill in the art would not have understood whether the claim element also

covers determining a whisper based on other characteristics. I understand that Mr. McAlexander testified that the claimed threshold could include a threshold frequency, threshold intensity, threshold mel value, or some other threshold. A person of ordinary skill in the art would not have known which of those options are covered by the claim.

C. Proposed Substitute Claim 8 Lacks Written Description and Enablement.

47. I understand that a claimed invention lacks adequate written description if the patent's specification does not convey with reasonable clarity to a person of ordinary skill in the art that the inventor had possession of the claimed invention at the time the patent application was filed. I further understand that a claim is enabled only if a person of ordinary skill, reading the specification, could practice the claimed invention without undue experimentation. I understand that the specification itself, and not the knowledge of a person of ordinary skill, must supply the novel aspects of an invention to adequately enable it.

48. Proposed substitute claim 8 recites determining whether a voice is a whisper based on peak frequency. To the extent the claim is construed to encompass determining a whisper based on (i) whether the intensity at a peak frequency is less than a threshold and/or (ii) whether the peak frequency is less than a *frequency* threshold, as I understand that Mr. McAlexander testified, the patent's specification

neither describes nor enables such determinations. The specification describes using a predetermined *intensity* (measured in dB) as a feature value, but never discloses what an appropriate intensity threshold would be to distinguish between whispers and non-whispers. It does not describe using a frequency threshold (measured in Hz) at all, let alone describe an appropriate frequency threshold for detecting whisper. Moreover, to the extent the claim can be construed to encompass threshold frequencies, intensities, mel values, and other thresholds, as I understand Mr. McAlexander testified, the specification provides *no* details about many of those options.

49. I understand that Mr. McAlexander testified that proposed substitute claim 8 should be construed to require a classification of an acquired voice based on both user-to-device distance and peak frequency. The specification neither enables nor shows possession of such an alleged invention. The specification only details voice classification based on distance *or* peak frequency. (EX-1001 ('675 patent), 10:9-46.) Although the specification teaches that the device may "combin[e] the classification methods," it provides no details as to how to do so. (*Id.*, 11:1-4.)

D. Proposed Substitute Claim 8 Would Have Been Obvious.

50. Because my prior declaration, and the paragraphs above, show how each of the limitations in the original claims are obvious over the prior art, I will address only the new and/or different limitations of proposed substitute claim 8, e.g.,

that "the voice classifying unit is further configured to determine the voice to be a predetermined voice when the voice spoken by the user is determined to be a whisper based upon whether a peak frequency is equal to or lower than a threshold."

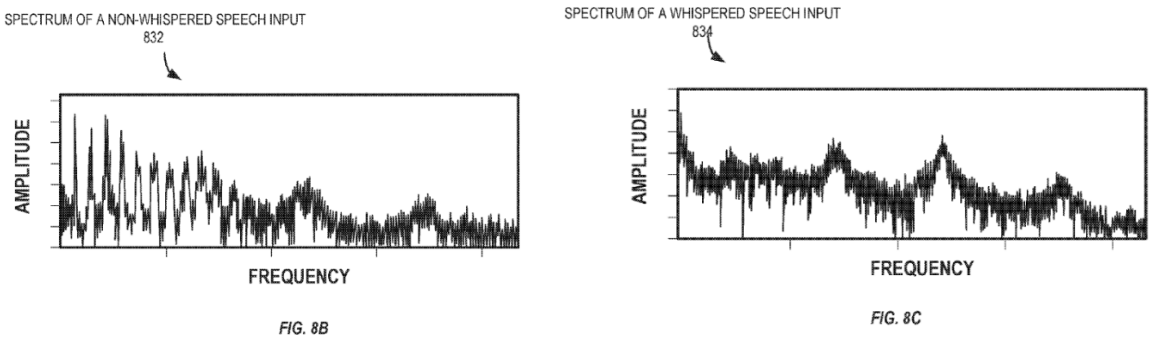
51. Each reference I discuss below is from the same field as the '675 patent (e.g., controlling a device's voice output) and pertinent to a problem the inventors focused on (e.g., adjusting voice output to improve user interaction).

1. Proposed Substitute Claim 8's New Limitation Was Known.

52. Schuster discloses determining whether a received voice is a whisper. (EX-1016 (Schuster), 3:1-11.) Schuster discloses determining the voice to be a whisper based on cepstral analysis, but further discloses that whisper can be detected using "any suitable method." (*Id.*) In my view, detecting whisper based on whether the intensity at a peak frequency is less than a threshold intensity was known in the art and would have been considered a "suitable method" for detecting whisper. Thus, it is my opinion that substitute claim 8 would have been obvious to a person of ordinary skill in view of the references in my previous declaration, including Schuster. I understand the Patent Owner may argue that determining whisper using a peak-frequency threshold was not known as a "suitable method." If so, claim 8 still would have been obvious in view of the references I discuss below.

a. Raitio

53. Raitio discloses a device that includes a "[w]hispered speech determination module 820" that "can determine whether the speech input includes a whispered speech input based on one or more differences between" whispered and non-whispered speech. (EX-1209 (Raitio) ¶[0243].) Raitio discloses that "the non-whispered speech input and the whispered speech input can have differing spectrums" (*id.* ¶[0244]) and includes figures showing exemplary spectrums:



(*Id.*, Figs. 8B-8C.)

54. Raitio discloses determining whether a voice is a whisper based on "spectrum characteristics such as the amplitude, the energy, the volume, and the slope." (EX-1209 (Raitio) ¶[0245].) Raitio further discloses detecting whisper when "the amplitude of the spectrum of the speech input is smaller than a threshold amplitude below a pre-determined frequency (e.g., 800 Hz)[.]" (*Id.*; *id.* ¶[0247], claim 6.) Thus, Raitio discloses the new limitation in claim 8, i.e., determining the voice to be a predetermined voice (e.g., whisper or "first voice") when the voice

spoken by the user is determined to be a whisper based upon whether the peak frequency intensity (amplitude) is equal to or lower than a threshold.

55. I understand that the Patent Owner may argue that the claim limitation is limited to detecting whisper when the peak frequency is equal to or less than a *frequency* threshold. Even if this were correct, Raitio renders it obvious. Raitio renders it obvious because it discloses detecting whispered voice using a threshold frequency (e.g., 800 Hz). (EX-1209 (Raitio) ¶[0245].) Raitio also renders it obvious because a person of ordinary skill in the art would have recognized that, in the whispered speech spectrum of Fig. 8B, the peak frequency is at the extreme left side, whereas in the non-whispered speech spectrum of Fig. 8C, the peak frequency is rightward of the extreme left side. To a person of ordinary skill, this discloses, or at least suggests, that there exists some threshold peak frequency that distinguishes whispered and non-whispered speech. Moreover, Raitio discloses detecting whisper "based on one or more spectrum characteristics" and "using any desired classification methods and/or spectral estimation methods." (*Id.* ¶[0250].) A person of ordinary skill in the art knew that peak frequency is a "spectrum characteristic" and, therefore, Raitio suggests classifying whispered and non-whispered speech based on comparing peak frequency to a threshold.

b. Chen

56. Chen discloses a device that is configured to detect whispered speech. (EX-1211 (Chen) ¶¶[0046]-[0074], Figs. 1-2.) In particular, Chen discloses that whispered speech "can be detected by finding one or more energy levels about one or more predefined thresholds in one or more predefined frequencies." (*Id.* ¶¶[0065], [0068], [0070] (detecting normal speech rather than whispered speech "when the power exceeds a predefined level" in certain frequency ranges), [0002] (whispered speech "exhibits a lower energy level").) Thus, Chen discloses exactly what the '675 patent's specification describes, i.e., determining whether a voice is a whisper based on a threshold intensity (energy level). Notably, like the '675 patent, Chen also discloses using a "proximity sensor" and user-to-device distance to classify the received voice. (*Id.* ¶¶[0040], [0043], [0092].)

c. Additional References

57. Detecting whisper was widely known before March 2018. In addition to the above references, many other references disclosed methods for detecting whispers using speech characteristics and thresholds. For example, Basye disclosed "a system that may determine that an input utterance was whispered" "based on audio (and possibly non-audio) paralinguistic feature data[.]" (EX-1208 (Basye) ¶[0051]; *id.* ¶[0018].) Basye discloses a system that is "trained to recognize whispered speech based on resonance, volume, and/or other features of input audio." (*Id.* ¶[0051].)

Basye discloses detecting whispered speech when the speech "has resonance below a threshold[.]" (*Id.* ¶[0052]; *id.* ¶¶[0058], [0060] (output tailored based on whether the voice was whispered).) As another example, Jain disclosed detecting whisper based on "the absence of pitch[.]" (EX-1210 (Jain) ¶[0045].) Thus, even if Raitio and Chen did not expressly disclose it, detecting whisper based on a threshold intensity or threshold peak frequency would have been obvious to a person of ordinary skill in the art in view of additional references such as Basye and Jain.

58. Lastly, determining whisper based on a threshold intensity (or frequency) would have been obvious to try. The references above disclose detecting whisper by examining various speech properties, including resonance (EX-1208 (Basye) ¶[0051]), volume (EX-1208 (Basye) ¶[0051]; EX-1209 (Raitio) ¶[0244]), amplitude or energy at a particular frequency (EX-1209 (Raitio) ¶[0244]; EX-1211 (Chen) ¶[0065]), spectrum slope or tilt (EX-1209 (Raitio) ¶[0244]), and/or lack of pitch (EX-1210 (Jain) ¶[0045]). The references also disclose that combinations of these and other properties can be examined to determine whisper. (EX-1208 (Basye) ¶[0051]; EX-1209 (Raitio) ¶¶[0244]-[0245].) To the extent those references do not expressly disclose using the claimed peak-frequency threshold, a person of ordinary skill would have understood that peak frequency is a similar speech characteristic and would have found it obvious to try detecting whisper by comparing peak frequency to a threshold.

2. The References Disclose or Suggest Determining Whisper Based on Distance *and* Peak Frequency.

59. I understand that the Patent Owner may argue that claim 8 requires that a single voice classification be made based on both user-to-device distance and peak frequency. Even if that is the case, the prior art discloses this as well. As I discuss above, Chen discloses detecting whisper based on energy level at a frequency and "a proximity sensor for sensing a distance between a user and the mobile computerized device[.]" (EX-1211 (Chen) ¶¶[0040]-[0041].) Basye discloses detecting whisper based on audio data (e.g., resonance, volume) and non-audio data including the "proximity of the user to a device." (EX-1208 (Basye) ¶¶[0055], [0049], [0064].) Thus, the prior art discloses using, and it would have been obvious to use, both peak frequency and distance to classify acquired speech.

3. Motivation to Combine

60. A person of ordinary skill in the art would have been motivated for many reasons to combine the teachings of one or more of the prior art references discussed above (Raitio, Chen, Basye, and/or Jain) with the prior art discussed in my previous declaration to create a device capable of classifying voice as a whisper based on a peak-frequency threshold.

61. First, the prior art discloses the benefits of detecting whispered speech. For example, Raitio discloses that "the user may whisper speech to the device and/or

desire that responses be whispered" and thus it is "important that the device recognizes the user's whispered speech[.]" (EX-1209 (Raitio) ¶[0003]; *see also* EX-1208 (Basye) ¶¶[0014]-[0016]; EX-1211 (Chen) ¶[0004] ("[t]here is a growing need to detect whispered speech").) As another example, Schuster teaches that configuring a device to respond in a "private communication mode" when the user whispers to the device "allows the user to interact with the device in a private manner, [] avoid disturbing other people, [] keep their interaction confidential, or simply [] avoid social awkwardness[.]" (EX-1016 (Schuster), 2:59-67; *see also* EX-1012 (Iyer), Abstract, 3:8-27, 4:30-36.) Thus, a person of ordinary skill would have been motivated to configure Shin and Shimomura's devices to detect whisper, and do so by comparing peak-frequency intensity to a threshold as taught by Raitio or Chen, to achieve the benefits of detecting whispered speech.

62. Second, as I discussed in my prior declaration, a person of ordinary skill would have been motivated to combine Schuster's disclosure with the other references in my prior declaration to detect whispered speech. (EX-1002 ¶¶228-229.) However, Schuster does not disclose detailed methods for detecting whispered speech. A person of ordinary skill therefore would have looked to references like Raitio and Chen for details on implementing whisper detection.

63. Third, such a combination merely represents the addition of one known element (whispered speech detection using peak frequency) to another known element (Shin's or Shimomura's output-content control device) to obtain predictable results (a device capable of classifying speech as whispered and adjusting output accordingly). Fourth, such a combination applies a known technique (detecting whispered speech using peak frequency) to improve a similar device and method (e.g., Shin's or Shimomura's) in the same way. Fifth, such a combination applies a known technique (detecting whispered speech using peak frequency) to a known device and method (e.g., Shin's or Shimomura's) that is ready for improvement and yields predictable results (a device capable of classifying speech as whispered and adjusting output accordingly).

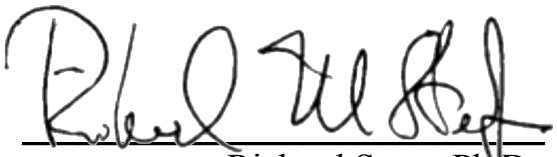
64. A person of ordinary skill would have reasonably expected success in combining the teachings. First, detecting whispered speech was well known, as demonstrated by the references above. Second, the references discussed above disclose very similar devices to Shin and Shimomura; for example, each reference discloses a device with a microphone for acquiring a voice signal and a processor for analyzing the voice signal and providing appropriate output. (*E.g.*, EX-1209 (Raitio) ¶[0036] (Raitio's device can be "any suitable electronic device" including "portable multifunctional device" or "personal electronic device"); EX-1211 (Chen) ¶[0039]

(Chen's device may be "a smartphone …, a media player, a streamer, or any other computerized system").)

65. Thus, the prior art renders proposed substitute claim 8 obvious.

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

Executed on March 18, 2026 at Buenos Aires, Argentina.



Richard Stern, Ph.D.