

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. IPR2025-00565
U.S. Patent No. 11,069,337

DECLARATION OF JOSEPH MCALEXANDER

TABLE OF CONTENTS

I.	BACKGROUND	1
II.	PROFESSIONAL QUALIFICATIONS	1
III.	MATERIAL CONSIDERED	2
IV.	LEGAL STANDARDS	3
	A. Anticipation	3
	B. Obviousness.....	4
	C. Claim Construction.....	6
V.	LEVEL OF ORDINARY SKILL IN THE ART	8
VI.	TECHNOLOGY BACKGROUND.....	8
VII.	THE '337 PATENT.....	9
VIII.	CLAIM CONSTRUCTION	11
	A. Preliminary Construction of “User”	11
	B. Effect of Preliminary Construction on “Classify”	12
	C. Effect of Preliminary Construction on Two or More Voices	12
	D. Effect of Preliminary Construction on “Analyze a Voice ... to Classify the Voice ... based on the Distance Between the User and the ... Device”	20
IX.	OVERVIEW OF THE ALLEGED PRIOR ART.....	22
	A. Shin (EX-1003)	22
	B. Shimomura (EX-1004).....	22
	C. Kristjansson (EX-1013).....	23
X.	CLAIMS	24

XI.	THE CHALLENGED CLAIMS OF THE '337 PATENT ARE VALID	27
A.	Ground 1	27
1.	Shin Does Not Disclose Classification of a Voice as Either a First Voice or a Second Voice.....	27
2.	Shin Does Not Analyze the Acquired Voice to Classify the Acquired Voice Based on Calculated Distance Between the User and the Voice-Content Control Device.....	30
3.	Shin Does Not Disclose Generation of First and Second Output Sentence When the Acquired Voice has Been Classified as the First and Second Voice	32
4.	Shin Does Not Disclose the Claimed Components Arranged as in the Claims	33
B.	Ground 2: Shin Does Not Render the Challenged Claims Obvious ...	33
C.	Ground 3: Shimomura Alone Does Not Render the Challenged Claims Obvious.....	34
1.	Shimomura Does Not Disclose Classification of a Voice as Either a First Voice or a Second Voice.....	34
2.	Shimomura Does Not Analyze the Acquired Voice to Classify the Acquired Voice Based on Calculated Distance Between the User and the Voice-Content Control Device	37
3.	Shimomura Does Not Disclose Adjust the Sound Volume of Voice Data Such that the Sound Volume of Voice Data Obtained by Converting the First Output Sentence thereinto Differs from the Sound Volume of the Voice Data Obtained by Converting the Second Output Sentence thereinto	38
4.	Shimomura Does Not Disclose Generation of First and Second Output Sentence When the Acquired Voice has Been Classified as the First and Second Voice.....	44
D.	Ground 4: Shimomura in View of Shin Does Not Render the Challenged Claims Obvious.....	45

E. Ground 5: The Combination of Shin, Shimomura, and Kristjansson
Does Not Render the Challenged Claims Obvious 47

F. Grounds 2-5 47

I have personal knowledge of the facts set forth in this declaration and, if called to testify as a witness, would testify under oath as follows:

I. BACKGROUND

1. I have been retained as an expert on behalf of SoundClear Technologies LLC (“SoundClear”) in the above-captioned matter (IPR2025-00565).

2. I have been asked by SoundClear to provide my expert opinion regarding the validity of claims 1-5 of U.S. Patent No. 11,069,337 (“the ’337 patent”). For the reasons set forth below, it is my opinion that claims 1-5 of the ’337 patent are valid.

3. I have also been asked to assess a preliminary claim construction provided by the Board and opine on whether it is supported by the plain and ordinary meaning of the claim. For the reasons set forth below, it is my opinion that the preliminary construction would be problematic in a number of respects.

II. PROFESSIONAL QUALIFICATIONS

4. I am a Registered Professional Engineer (#79454) and recognized as an inventor on 31 U.S. and a number of foreign patents.

5. I am President of M^cAlexander Sound, Inc., and the Managing Director of M^cAlexander Sound Pte Ltd. I have focused my expertise to support a number of clients in product, process, and operations analysis and investigation.

6. I have extensive experience with voice-control audio devices and audio processing in general including speech detection and spatial audio processing,

beamforming, noise reduction, speech capture, near and far field detection, wireless audio transmission, audio device integration in automotive systems, spatial audio processing, audio system design and installation, and underwater acoustic signal detection.

7. Further details on various aspects of my professional experience and qualifications can be found in my curriculum vitae, which is provided as Exhibit 2014.

III. MATERIAL CONSIDERED

8. In preparing this declaration, I have reviewed the specification, claims and prosecution history of the '337 patent.

9. I considered the following documents in preparing this Declaration.

Exhibit No.	Description
Petitioner (Amazon) Exhibits	
1001	U.S. Patent No. 11,069,337
1002	Declaration of Amazone's expert, Dr. Richard Stern
1003	U.S. Patent App. Publ. No. 2017/0083281 ("Shin")
1004	English Translation of Shimomura from Ex. 1005
1007	Excerpts from File History of U.S. Patent No. 11,069,337
1009	U.S. Patent App. Publ. No. 2017/0154626 ("Kim")
1010	U.S. Patent App. Publ. No. 2017/0337921 ("Aoyama")
1011	U.S. Patent App. Publ. No. 2016/0284351 ("Ha")
1012	U.S. Patent No. 9,489,172 ("Iyer")

Exhibit No.	Description
1013	U.S. Patent No. 10, 147,439 (“Kristjansson”)
1015	Nicolae Duta, Natural Language Understanding and Prediction: From Formal Grammars to Large Scale Machine Learning, 131 Fundamenta Informaticae 425 (2014) (“Duta”)
Patent Owner (SoundClear) Exhibits	
2015	Transcript of Deposition of Petitioners’ Expert, Dr. Richard Stern
Other Documents	
Paper 1	Amazon’s Petition
Paper 12	Board’s Institution Decision

IV. LEGAL STANDARDS

10. I am not an attorney, and I have not independently researched the law. SoundClear’s counsel has explained certain legal principles to me that I have relied on in forming my opinions set forth in this declaration. I have applied these legal principles in arriving at my opinions expressed in this declaration. They are set forth below.

11. I have been informed and understand the following principles.

A. Anticipation

12. To anticipate a claim, a prior art reference must disclose each and every element of the claim. It must do so either explicitly or inherently.

13. To anticipate a claim, a prior art reference must also disclose those elements arranged as in the claim.

14. It is not proper to mix or combine incompatible embodiments in order to assert that a reference discloses each and every element of a claim.

B. Obviousness

15. An invention cannot be patented if the subject matter as a whole would have been obvious to a person of ordinary skill in the art at the time of the invention. Also, while the prior art is compared to each claim on an element-by-element basis, the claimed invention as a whole must have been obvious to one of ordinary skill in order for a claim to be invalid.

16. The fundamental question in analyzing obviousness is whether, at the time of the invention, the subject matter of the claimed invention as a whole would have been obvious to a person of ordinary skill in the art to which the subject matter pertains, taking into account: (a) the scope and content of the prior art; (b) the differences between the prior art and the claims at issue; (c) the level of ordinary skill in the art; and (d) any objective indicia of non-obviousness referred to as secondary considerations. The secondary considerations may include: (1) a long-felt but unmet need in the prior art that was satisfied by the claimed invention; (2) the failure of others; (3) skepticism; (4) commercial success of a product covered by the patent; (5) industry praise of the claimed invention; and (6) copying of the invention.

17. A combination of prior art references must disclose, teach or suggest all of the recited limitations of the claim for the invention to be found obvious.

18. It is important in the obviousness inquiry to identify whether a reason existed at the time of the invention that would have motivated a person of ordinary skill in the art in the relevant field to combine or modify the prior art references in the manner proposed by the petitioner so as to arrive at the claimed invention. Put another way, a finding of obviousness should be supported by an apparent reason to combine or modify the prior art references as proposed by the petitioner.

19. It is important in the obviousness inquiry to understand how the combination of references is supposed to work. An explanation of the operation of the combined references is often a prerequisite to showing that a person of ordinary skill in the art would have had an apparent reason to make the proposed combination and would have had a reasonable expectation of success.

20. A finding of obviousness requires the additional showing that a person of ordinary skill at the time of the invention would have selected and combined those prior art elements in the normal course of research and development to yield the claimed invention. In addition, a person of ordinary skill in the art should have had a reasonable expectation of success in combining the references as proposed.

21. In assessing obviousness, one must consider both the ordinary creativity and common sense of the person of ordinary skill in the art. It is impermissible for common sense to be applied to fill gaps in prior art that fails to teach or suggest a limitation of the claim.

22. Multiple references can be combined, with one another and/or with the knowledge of a person of ordinary skill in the art, in rendering a claim obvious. However, obviousness cannot be established by simply demonstrating that each element was independently known in the prior art. Rather, it may be necessary to identify a reason, such as a teaching, suggestion, or motivation, which would have prompted a person of ordinary skill in the art to combine the elements in the way the claimed invention does.

23. Obviousness cannot be established through hindsight. That means the claimed invention cannot be used as a roadmap to combine elements from different pieces of prior art, or different embodiments of a single prior art reference, to create the claimed invention. The claimed invention as a whole must be compared to the prior art as a whole, and one must avoid aggregating pieces of prior art through hindsight that would not have been combined absent the patent inventor's insight.

C. Claim Construction

24. Claim construction is the process of determining the meaning of words used in a patent claim in order to understand the meaning of the claim. The proper construction of a term is how a person of ordinary skill in the art ("POSITA"), at the time of the invention, would have understood the term based on its use in the claims, specification, and prosecution history (the intrinsic record) of the patent.

25. The claims are not to be reviewed in this proceeding under the “broadest reasonable interpretation” standard.

26. Instead, the claims are to be reviewed under the *Phillips* standard in which the claims are to be interpreted based on their ordinary and customary meaning as understood by a person skilled in the art.

27. The claim construction inquiry begins and ends in all cases with the actual words of the claim. The language of the claim frames and ultimately resolves all issues of claim interpretation.

28. The inquiry into how a person of ordinary skill in the art understands a claim term provides an objective baseline from which to begin claim interpretation. That starting point is based on the well-settled understanding that inventors are typically persons skilled in the field of the invention and that patents are addressed to and intended to be read by others of skill in the pertinent art. Importantly, the person of ordinary skill in the art is deemed to read the claim term not only in the context of the particular claim in which the disputed term appears, but in the context of the entire patent, including the specification.

29. A patent—including the specification and claims—and its prosecution history are considered “intrinsic evidence.” They provide aid in interpreting claim language in a patent. In reading the claim, it is improper to import limitations from the specification into the claim terms. It is also improper to narrow the scope of the

claim terms by implicitly adding disclosed limitations that have no express basis in the claims.

30. Further, one must not provide a definition of claim terms divorced from the context of the patent and must not then turn to the specification only to look for grounds for limiting the claim scope.

31. The prosecution history of related patents and applications can also be relevant. Sources extrinsic to a patent and its prosecution history (such as dictionary definitions and technical publications) may also be used to help interpret the claim language, but such extrinsic sources cannot be used to contradict the unambiguous meaning of the claim language that is evident from the intrinsic evidence.

V. LEVEL OF ORDINARY SKILL IN THE ART

32. A POSITA at the time of the invention of the '337 patent would have: (a) a Bachelor of Science degree in electrical or computer engineering (or a related academic field), and at least two (2) additional years of work experience in the design and development of electronic devices or systems related to controlling audio output based on inputs.

VI. TECHNOLOGY BACKGROUND

33. I agree with Dr. Stern's recognition of the existence of voice-controlled devices in the 2010-era. Those included Alexa-enabled devices and Google's voice

assistant. Each is centered on determining the content of speech, which is provided by a voice. They are not mere measurement devices.

VII. THE '337 PATENT

34. The '337 patent was filed on March 4, 2019 as U.S. Patent Application No. 16/290,983. The '337 patent also claims priority to Japanese Patent Application No. JP2018-039754 that was filed March 6, 2018. Consequently, I understand that the priority date of the '337 patent is at least as early as March 6, 2018.

35. The '337 patent discloses a “voice-content control device” that includes “a voice classifying unit configured to analyze a voice spoken by a user.” EX-1001, Abstract. The focus is on user interaction with a device by speaking to it. More particularly, it relates to a device that analyzes a user’s voice. That voice includes multiple aspects, including the “content” contained in it. The '337 patent is not a mere measurement device. It is not simply a “volume measurement device” or “distance determination device.” It is focused on human voices and what makes voices different than any singular characteristic, such as volume or distance.

36. As mentioned above, the '337 patent describes “a voice-content control device.” EX-1001, 1:46. The voice-content control device “compris[es]: a voice classifying unit configured to analyze a voice spoken by a user and acquired by a voice acquiring unit to classify the voice as either one of a first voice or a second voice.” *Id.*, 1:48-52. The patent also discloses a “voice-content control method,

comprising: acquiring a voice spoken by a user; analyzing the acquired voice to classify the acquired voice as either one of a first voice and a second voice.” *Id.*, 1:65-2:1.

37. According to the ’337 patent, one embodiment of the voice classifying unit “analyzes the voice V1 acquired by the voice acquiring unit 30, and classifies the voice V1 to either one of a first voice V1A and a second voice V1B.” *Id.*, 7:51-54. For example, the voice classifying unit can determine if a voice is a whisper or not and classify the voice accordingly. *See id.*, 7:57-62 (classifying the whisper as V1A). As another example, the voice classifying unit can perform Fourier-transformation on the time waveform to determine the intensity of the voice. *See id.*, 7:64-8:1 (“the voice classifying unit 38 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.”). That embodiment of the voice classifying unit can then classify the voice as a first voice V1A or 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.” *Id.*, 8:3-4.

38. Other embodiments of the voice classifying unit use the peak frequency of the sound waves, the volume of the voice, the speaking speed of the voice, the volume ratio between the speech and wind noise, the distance between the speaker and the device, or the mel-frequency cepstrum coefficient. *See id.*, 8:3-34 (explaining that

the voice classifying unit sets certain feature values that distinguish between different voices). Further, “[t]he voice classifying unit 38 can use either one of the classification methods explained above, or can perform the classification by combining the classification methods explained above.” *Id.*, 8:56-59.

VIII. CLAIM CONSTRUCTION

39. I have been informed and understand that neither SoundClear nor Amazon proposed any claim construction at the institution stage of this proceeding. But, in the Board’s Institution Decision, the Board stated that claim construction of “user” was necessary. Board’s Institution Decision, p. 14.

40. I have reviewed the Board’s preliminary construction of the term “user” and the explanation stated in the institution decision. I have also reviewed the intrinsic record of the ’337 patent as it pertains to the term “user.”

A. Preliminary Construction of “User”

41. I have been informed and understand that the Board preliminarily construed the claimed “a voice spoken by a user” to encompass “one or more voices spoken by a single user, or by plural users.” Board’s Institution Decision, p. 15. I have been informed and understand that SoundClear does not contest the Board’s conclusion that the claims recite “one or more users.” I understand that the Board explained that its “preliminary claim construction *only* requires petitioner to show a single user who may perform both a first and second voice *depending upon the single user’s*

changing distance (and associated changing voice volume) from the voice-content control device.” Board’s Institution Decision, p. 16 (emphasis added).

B. Effect of Preliminary Construction on “Classify”

42. In my opinion, the Board’s preliminary construction of “a voice spoken by a user” contradicts how a POSITA would have understood the term “classify” in the context of the claims. *See, e.g.*, EX-1001, 18:42-46.

43. A POSITA would have understood “classify” to require something discrete, not a mere measurement. A POSITA would not have understood “classify” to be used in the context of measuring a single, continuously changing variable (like volume or distance). A POSITA would have understood the claims of the ’337 patent to require a classification, or distinction between, at least two separately identifiable voices.

C. Effect of Preliminary Construction on Two or More Voices

44. In my opinion, the preliminary construction is also problematic in that it changes the plain and ordinary meaning of the word “voice.” The word “voice” does not equate to the mere incidental change in “volume” due to a measurement device being moved away from a speaker. It is especially problematic in the context of a user who speaks at a constant volume but simply moves away from a measuring device.

45. The word “voice” connotes the quality of information conveyed by a human. That information is not conveyed with simple changes in volume. It is conveyed via a multivariate notion of voice. The nature of a (singular) voice can change with volume. But volume alone does not create a new voice. Dr. Stern testified as follows:

Q Okay. But I do think this is within the scope of the patent. The patent talks about different voices and increasing and decreasing volume. So I’m just asking you, does the nature of the voice change when you increase the volume of your voice?

...

A The patent -- the patent really does not address that, the nature of the changes, that do happen when you increase and decrease your voice.

EX-2015, 59:13-50. When asked to clarify, Dr. Stern stated that something more than volume or intensity must change before voice changes. *Id.* 60:3-62:16. He acknowledged that something more is necessary before one can “discriminate between speech that’s shouted from speech that’s spoken quietly, regardless of what the volume control is” *Id.*

46. In my opinion, the preliminary construction conflates multiple terms that have plain and ordinary meanings in the art and that are used throughout the patent. A POSITA would have understood each of “voice,” “volume,” and “distance” to be

distinct concepts. In fact, the claims expressly provide for volume to change while voice remains constant. *See, e.g.*, Limitations 1[e]-1[h] below.

47. The full context of the claim is clear. The claims recite “**voice**” separate from both “**volume**” and “**distance**.” For instance, claim 1 reads as follows:

1. A **voice**-content control device, comprising:

- [a] a proximity sensor configured to calculate a **distance** between a user and the **voice**-content control device;
- [b] a **voice** classifying unit configured to analyze a **voice** spoken by a user and acquired by a **voice** acquiring unit to classify the **voice** as either one of a first **voice** or a second **voice** based on the **distance** between the user and the **voice**-content control device;
- [c] a process executing unit configured to analyze the **voice** acquired by the **voice** acquiring unit to execute processing required by the user;
- [d] a **voice**-content generating unit configured to generate, based on content of the processing executed by the process executing unit, output sentence that is text data for a **voice** to be output to the user; and
- [e] an output controller configured to adjust a sound **volume** of **voice** data obtained by converting the output sentence thereinto, wherein
- [f] the **voice**-content generating unit is further configured to generate a first output sentence as the output sentence when the acquired **voice** has been classified as the first **voice**, and

[g] generate a second output sentence in which information is omitted as compared to the first output sentence as the output sentence when the acquired **voice** has been classified as the second **voice**,

[h] wherein the output controller is further configured to adjust the sound **volume** of **voice** data such that the sound **volume** of **voice** data obtained by converting the first output sentence thereinto differs from the sound **volume** of the **voice** data obtained by converting the second output sentence thereinto.

'337 Patent (EX-1001), claim 1 (annotated). The distinction between voice and volume is unmistakable. The claim recites measuring a distance (limitation [a]) and then three limitations ([b]-[d]) regarding classifying voice, analyzing voice, and generating content for a voice. Each of these three limitations is conducted without reciting "volume." The claim recites that volume is a characteristic that can be "adjusted." Volume is not "classified." And that volume adjustment is not recited until a separate limitation is introduced (limitation [e]). There is no basis to conflate concepts that the claim recites separately.

48. A POSITA would have understood that the full context of the claim is helpful here. Voice is "classified." It is a discrete concept. The incoming voice can either be one voice or one or more other voices. A POSITA would have understood that there is no middle ground. It is not a single variable that can have any value (on a continuous scale). It is not something that is understood to be calculated, measured,

or “adjusted” like volume. Further, the claim requires that the classification include an assessment of distance. A POSITA would have understood that whether or not the classification is based on volume does not effect a POSITA’s understanding of the word “voice.” There is also no indication that the patentee recited distance as a mere proxy for volume, especially when volume could be measured directly. Dr. Stern admitted that, while distance may sometimes be sufficient to classify voice, it is not always sufficient. EX-20154, 48:19-49:7. That statement means that distance cannot be synonymous with “voice.” It is just one aspect of “voice.”

49. The ’337 patent is entitled “Voice-Content Control Device, Voice-Content Control Method, and Non-Transitory Storage Medium,” and “relates to a voice-content control device, a voice-content control method, and a non-transitory storage medium.” EX-1001, code (54), 1:15-17. A POSITA would understand that it is not, for example, merely a “volume-content control device.” The ’337 patent discloses “a voice processing device that outputs voice.” EX-1001, 1:29-32, 1:39-42. It does not disclose a “volume processing device.” Nor does it disclose a device that outputs “volume.”

50. The entire disclosure demonstrates what any POSITA would understand: that voices contain many characteristics. Those characteristics include, for example, volume, frequency, intensity, slopes of peak frequency, content, speaking speed, volume ratio with wind noise, or others. EX-1001, 7:64-8:34. The specification

expressly defines voice as incorporating each of those characteristics, not just volume. It is a discrete concept, whereas characteristics like volume, frequency, intensity, and slopes of frequencies are continuously varying attributes. Those attributes can, for example, *decrease* or can be *too much*.

In that case, the output voice itself may be *decreased* in sound volume, but if the output voice is *decreased* in sound volume *too much*, the output voice may be hard to be heard by the user himself and the user may be unable to understand the content of the voice.

EX-1001, 1:34-38 (emphasis added). The above quote from the '337 patent again confirms that a single voice (*the* output voice) can have multiple volumes (can be increased or decreased to any value).

51. The specification gives a specific example. It describes classifying “the voice V1 to the second voice V1B *when the voice V1 is determined to be a whisper as a result of analysis*,” and “to the first voice V1A *when the voice V1 is determined not to be a whisper*.” EX-1001, 7:51-63 (emphasis added). The specification explains that whether voice V1 is a whisper or not may be determined based on one or more “feature value,” including “a peak frequency that is equal to or higher than a predetermined intensity in the spectrum,” “a slope of the peak in the spectrum,” “a volume of the voice V1, a speaking speed of the user in the voice V1, [] a volume ratio between a speech of the user and a wind noise,” “a distance between the user H and the voice-content control device 1,” and “a Mel-frequency cepstrum

coefficient.” EX-1001, 7:64-8:34; *see also id.*, 15:22-54 (“When the user H determines that the voice V2 is not desired to influence any person around the voice-content control device1, the user H changes the voice V1 by, for example, uttering the voice V1 as *a whispered voice*” (emphasis added)).

52. Accordingly, a POSITA would have understood that, when read in view of the specification, the claimed “first voice” and “second voice” are different classes of *voices* (e.g., a whisper voice, *i.e.*, “an unvoiced voice spoken without vibrating the vocal chords”). EX-1001, 7:59-63. And the specification’s description of “feature values” (including user-to-device distance, voice volume, and speaking speed) and related thresholds are determined and set for the purpose of distinguishing one class of voice from another. EX-1001, 8:20-26 (“[T]he distance can be used as a feature value to perform the classification to the first voice V1A and the second voice V1B.”), 8:35-41 (“As described above, ... the voice classifying unit 38 classifies the voice V1 to the second voice V1B when the voice V1 is a whisper, and to the first voice V1A when the voice V1 is not a whisper even when meaning of the voice V1 is the same.”).

53. The claimed classifying unit uses those characteristics to classify voices. Notably, each can be used alternatively or in different combinations. But no such changes alter the nature of what is being classified: voice. Certainly, a POSITA would not have understood use of a measurement to mean that any single

characteristic defines the to-be-classified voice. That would be akin to stating that deciding how sunny it is outside is the same as classifying celestial objects—simply because measuring luminosity is one method that one might use to determine whether something in the sky is a planet or a star.

54. Just because a POSITA might choose to use one variable to classify voice does not mean that the classification will be perfectly effective. Nor does it mean that any error associated with such single-variable classification would have transformed one speaking voice into a new voice. The claims of the '337 patent require that a classification be made and that the classification render a determination as to at least two discrete voices. A POSITA understands that simply detecting volume is not enough.

55. The Board's reasoning recognizes as much. The Board acknowledges that, at a minimum, a voice is something a user "perform[s]." Board's Institution Decision, p. 16. A user that speaks at the same volume while walking away from a device has not changed any aspect of their performance. The user has not even changed the volume of their own voice. In that instance, the device does not detect any aspect of user performance. It is the volume that the device receives that is different. In that way, the preliminary construction would defeat the entire purpose of the patent. The classification unit would simply be a volume measurement.

D. Effect of Preliminary Construction on “Analyze a Voice ... to Classify the Voice ... based on the Distance Between the User and the ... Device”

56. Limitation 1[b] recites: “a voice classifying unit configured to analyze a voice spoken by a user and acquired by a voice acquiring unit to classify the voice as either one of a first voice or a second voice based on the distance between the user and the voice-content control device.” EX-1001, Claim 1. Similarly, limitations 4[b] and 5[b] recite: “analyzing the acquired voice to classify the acquired voice as either one of a first voice and a second voice based on the distance between the user and the voice-content control device.” *Id.*, Claims 4-5. Thus, the plain language of limitations 1[b], 4[b], and 5[b] expressly require **(1)** a classification that must be based on **(2)** an analysis of the acquired voice and **(3)** a distance, which is calculated (*e.g.*, by a proximity sensor) in limitations 1[a], 4[a], and 5[a].

57. The specification consistently describes analyzing a user’s voice for purposes of determining whether the user’s voice is a whisper and classifying the user’s voice on that basis. *Id.*, 7:51-8:34. The specification describes calculating distance with a proximity sensor:

Furthermore, a proximity sensor can be provided in the voice-content control device 1, *a distance between the user H and the voice-content control device 1 is calculated from a detection result of the proximity sensor*, and the distance can be used as a feature

value to perform the classification to the first voice V1A and the second voice V1B.

Id., 8:20-26 (emphasis added). The specification further provides that “[t]he voice classifying unit 38 ... can perform the classification *by combining* the classification methods explained” therein, including by combining voice analyses (*e.g.*, to determine whether the user’s voice is a whisper) with distance calculation (*e.g.*, by a proximity sensor), as required by the plain language of the claims. *Id.*, 8:56-59, 7:64-8:34 (emphasis added).

58. Further, the specification discloses the “distance between the user H and the voice-content control device 1” as a potential “feature value” for determining whether the user’s voice is a whisper and classifying the user’s voice based on the same. *Id.* A POSITA would have recognized that such distance (*i.e.*, between the user and the device) alone is not a good indicator of whether the user’s voice is a whisper. In view of the specification’s disclosure, a POSITA further would have understood that distance between the user and the device can be used in conjunction with other “feature value[s]” (*e.g.*, peak frequency equal to or higher than a predetermined intensity, slope of the peak in the frequency spectrum, voice volume, speaking speed, volume ratio between speech and noise, mel-frequency cepstrum coefficient) to determine whether the user’s voice is a whisper for purposes of classifying the user’s voice. EX-1001, 7:64-8:34.

59. Thus, in view of (and consistent with) the specification, a POSITA would have understood the plain meaning of limitations 1[b], 4[b], and 5[b] to require the claimed (1) voice classification be based on (2) analysis of the acquired voice and (3) distance calculated (*e.g.*, by a proximity sensor) in limitations 1[a], 4[a], and 5[a].

IX. OVERVIEW OF THE ALLEGED PRIOR ART

A. Shin (EX-1003)

60. Shin (U.S. Patent App. Publ. No. 2017/0083281), entitled “Method and Electronic Device for Providing Content” was filed on September 19, 2016. *See* EX-1003. Shin claims priority to a Korean patent application filed on September 18, 2015. EX-1003 ¶ [0001].

61. Shin discloses an electronic device for providing content in response to user input. EX-1003, ¶ [0003]. Shin discloses adjusting content based on “least one of a speech rate of *the* speech, a volume of *the* speech, or a keyword included in *the* speech . . .” *Id.* ¶ [0010] (emphasis added). The Petition seems to agree: Shin operates “*based on* the user-to-device distance.” Petition, p. 19 (emphasis added).

B. Shimomura (EX-1004)

62. Shimomura (Japanese Unexamined Patent App. Publ. No. 2005/202076), entitled “Speech Control Device and Method and Robot Device” was filed on January 14, 2004. *See* EX-1004.

63. Shimomura is concerned only with distance and the volume output based on distance. Shimomura thus lacks all voice classification limitations, especially those based on an analysis of the voice and distance:

[Problem]

Depending on the *distance* from the device or robot device to the conversation partner, the conversation partner may have difficulty *hearing* the speech of the device or robot device.

[Means for Solving]

By changing the speech style of the device or robot device when interacting with the user as necessary depending on the *distance* between the device or robot device and the user

EX-1004, at p. 1 (emphasis added); *id.* ¶ [0012] (“the speech form of the device during a dialogue with the user is changed *as necessary according to the distance* between the device and the user.”) (emphasis added)).

64. Shimomura discloses a “change[] in intonation.” EX-1004, ¶ [0092] (“changes in intonation (emphasis at the end of each phrase) ...”).

C. Kristjansson (EX-1013)

65. Kristjansson et al. (U.S. Patent No. 10,147,439), entitled “Volume Adjustment for Listening Environment” was filed on March 30, 2017. *See* EX-1013.

66. Kristjansson discloses a device that uses distance to determine sound pressure where the user is located. EX-1013, Abstract.

X. CLAIMS

67. The following claims appear in the '337 patent.

Claims
1. A voice-content control device, comprising: [a] a proximity sensor configured to calculate a distance between a user and the voice-content control device; [b] a voice classifying unit configured to analyze a voice spoken by a user and acquired by a voice acquiring unit to classify the voice as either one of a first voice or a second voice based on the distance between the user and the voice-content control device; [c] a process executing unit configured to analyze the voice acquired by the voice acquiring unit to execute processing required by the user; [d] a voice-content generating unit configured to generate, based on content of the processing executed by the process executing unit, output sentence that is text data for a voice to be output to the user; and [e] an output controller configured to adjust a sound volume of voice data obtained by converting the output sentence thereinto, wherein [f] the voice-content generating unit is further configured to generate a first output sentence as the output sentence when the acquired voice has been classified as the first voice, and [g] generate a second output sentence in which information is omitted as compared to the first output sentence as the output sentence when the acquired voice has been classified as the second voice, [h] wherein the output controller is further configured to adjust the sound volume of voice data such that the sound volume of voice data obtained by converting the first output sentence thereinto differs from the sound volume of the voice data obtained by converting the second output sentence thereinto.
2. The voice-content control device according to claim 1, wherein the

Claims
process executing unit comprises: [a] an intention analyzing unit configured to extract intention information indicating an intention of the user based on the voice acquired by the voice acquiring unit; and [b] an acquisition content information acquiring unit configured to acquire acquisition content information which is notified to the user based on the extracted intention information, and [c] the voice-content generating unit is further configured to generate the text data including the acquisition content information as the output sentence.
3. The voice-content control device according to claim 1, wherein, on generating the second sentence, the voice-content generating unit is further configured to omit a part of information included in the voice spoken by the user.
4. A voice-content control method, comprising: [a] calculating a distance between a user and a voice-content control device; [b] acquiring a voice spoken by a user; [c] analyzing the acquired voice to classify the acquired voice as either one of a first voice and a second voice based on the distance between the user and the voice-content control device; [d] analyzing the acquired voice to execute processing intended by the user; [e] generating, based on content of the executed processing, output sentence that is text data for a voice to be output to the user; and [f] adjusting a sound volume of voice data obtained by converting the output sentence thereinto, [g] wherein at the generating, a first output sentence is generated as the output sentence when the acquired voice has been classified as the first voice, and a second output sentence is generated as the output sentence in which a part of information included in the first output sentence is

Claims
omitted when the acquired voice has been classified as the second voice, [h] wherein at adjusting the sound volume of voice data, further adjusting the sound volume of voice data such that the sound volume of voice data obtained by converting the first output sentence thereinto differs from the sound volume of voice data obtained by converting the second output sentence thereinto.
5. A non-transitory storage medium that stores a voice-content control program that causes a computer to execute: [a] calculating a distance between a user and a voice-content control device; [b] acquiring a voice spoken by a user; [c] analyzing the acquired voice to classify the acquired voice as either one of a first voice and a second voice based on the distance between the user and the voice-content control device; [d] analyzing the acquired voice to execute processing intended by the user; [e] generating, based on content of the executed processing, output sentence that is text data for a voice to be output to the user; and [f] adjusting a sound volume of voice data obtained by converting the output sentence thereinto, [g] wherein at the generating, a first output sentence is generated as the output sentence when the acquired voice has been classified as the first voice, and a second output sentence is generated as the output sentence in which a part of information included in the first output sentence is omitted when the acquired voice has been classified as the second voice, [h] wherein at adjusting the sound volume of voice data, further adjusting the sound volume of voice data such that the sound volume of voice data obtained by converting the first output sentence thereinto differs from the sound volume of voice data obtained by converting the second

Claims
output sentence thereinto.

XI. THE CHALLENGED CLAIMS OF THE '337 PATENT ARE VALID

A. Ground 1

1. Shin Does Not Disclose Classification of a Voice as Either a First Voice or a Second Voice

68. The Petition asserts that “Shin discloses that the voice classifying unit (processor) analyzes the voice to classify it as either a first or second voice based on the user-to-device distance.” Petition, p. 13 (1st full ¶) (citing only expert declaration). But the only express disclosure of “voice” the Petition identifies in Shin that relates to processing voice is in connection with “generate[ing] . . . content.” *Id.* Petition, p. 13 (1st full ¶, 2d sentence). A POSITA would have understood that to refer to standard natural language processing to determine user intent. The Petition references Shin ¶ [0051]. But that paragraph only mentions “classification” of “content.” EX-1003, ¶ [0051]. And Shin mentions that classification of content solely in the context of output, and not voice input. EX-1003, Table 1, ¶ [0051].

69. The Petition references Table 3 of Shin at the bottom of page 13. The Petition includes an annotated version of the table, where Dr. Stern added the “First Voice”

and “Second Voice” labels. EX-1002, ¶ 64. But annotations aside, Table 3 discloses one thing: a calculation of distance—nothing else:

TABLE 3	
Distance (D) between user and electronic device	Output amount of information of content
$D < 1\text{ m}$	Amount of information of original content
$1\text{ m} \leq D < 2\text{ m}$	Content of first abbreviation level
$2\text{ m} \leq D < 3\text{ m}$	Content of second abbreviation level
$3\text{ m} \leq D < 4\text{ m}$	Content of third abbreviation level
$D \geq 4\text{ m}$	Content of fourth abbreviation level (abstract content)

EX-1003, Table 3.

70. That determination of distance cannot expressly disclose a classification because it is a measurement of a continuous variable, not any discrete classification. Nor can it expressly disclose an analysis of a “voice,” which is not at all synonymous with “distance.”

71. One example is sufficient to demonstrate the failed reasoning here. Assume two users, with two different voices, were standing the same distance from the device disclosed in Shin. Table 3 shows that Shin would not distinguish between the voices of those users. The only information determined by Shin would be distance. And that information would be the same for each voice.

72. Shin itself states that it only contemplates a single voice in connection with Table 3: “For example, the content that has the greatest output amount of information

may correspond to the most detailed original content corresponding to a voice input.” EX-1003, ¶ [0087] (emphasis added). And Shin confirms the distinction between “voice” and “distance.” Shin states that “processor 120 may determine a distance between the user and the electronic device 101 based on at least one of the *volume* of the user’s speech obtained through an analysis of a *voice* input or the *distance* computed, calculated, or measured by the distance detection module 180.” EX-1003, ¶ [0053] (emphasis added). Thus, a POSITA would not have understood Shin’s disclosure of “distance” to be synonymous with “voice.” Nor would a POSITA have understood Shin’s determination of a “distance” to be synonymous with a classification.

73. Moreover, Shin directly correlates content output to distance between the user and the device. Shin does not classify the voice input in any way, let alone classifying the voice input as a first voice or a second voice.

74. Even if a new voice could theoretically be created based solely on a volume change based on distance (under the Board’s proposed preliminary construction), that would not mean that a new voice is always created simply because distance is changed. In other words, the concepts of “volume” and “distance” would still be different than the concept of a “voice.” The Petition provides no basis upon which the Board could conclude that Shin discloses the concept of determining a voice as opposed to simply making a measurement of “distance.” And a POSITA would never

have understood Shin’s disclosure of calculating a change in distance of a user, who otherwise does not change their voice at all, as any sort of voice classification—even if volume received at the device changed as well.

2. Shin Does Not Analyze the Acquired Voice to Classify the Acquired Voice Based on Calculated Distance Between the User and the Voice-Content Control Device

75. Limitation 1[b] recites: “a voice classifying unit configured to analyze a voice spoken by a user and acquired by a voice acquiring unit to classify the voice as either one of a first voice or a second voice based on the distance between the user and the voice-content control device.” EX-1001, Claim 1. Similarly, limitations 4[b] and 5[b] recite: “analyzing the acquired voice to classify the acquired voice as either one of a first voice and a second voice based on the distance between the user and the voice-content control device.” *Id.*, Claims 4-5.

76. These limitations require (1) a classification that must be based on (2) an analysis of the acquired voice and (3) a distance, which is calculated (*e.g.*, by a proximity sensor) in limitations 1[a], 4[a], and 5[a]. Shin does not disclose the claimed “analyz[e/ing] the [acquired] voice to classify the [acquired] voice.”

77. The Petition asserts that Shin discloses “analyz[ing] a voice spoken by a user and acquired by a voice acquiring unit” (as recited in claim 1 and similarly recited in claims 4-5). Petition, pp. 12-13. The Petition relies on Shin’s disclosure of a processor 120 that “analyze[s] a voice input received through [the] audio input

module 151.” Petition, pp. 12-13 (quoting EX-1003, ¶ [0048]), 29-30 (alleging that Shin discloses claims 4-5 for the same reasons discussed in analysis for claim 1). Paragraph [0048] of Shin, however, relates to *content* of speech—what a user is saying. It does not relate at all to the distance detection disclosure, let alone any classification of voices. The Petition then moves right to the output side of Shin’s disclosure. The Petition asserts that Shin’s “processor determines the ‘output scheme,’ which includes the content and its volume level, ‘based on the distance between the user and the electronic device 101[.]’” Petition, p. 13 (quoting EX-1003 ¶¶ [0078], [0086]). The Petition presents no analysis as to how either the input or the output functions that Shin discloses meet the claims. It does not explain where Shin discloses analysis of the user’s voice in order “*to classify the voice* as either one of a first voice or a second voice,” as claimed.

78. As discussed above, the Petition relies entirely on Table 3 as disclosing the claimed classification. The Petition states that Shin discloses the claimed voice classification because Table 3 shows “providing a different amount of information based on the user-to-device distance.” Petition, pp. 13-14. The Petition further states that “Shin discloses classifying the voice as a ‘first voice’ when the distance is less than 1 meter, and a ‘second voice’ when the distance is greater than 4 meters.” *Id.*

79. Under Amazon’s mapping, Shin makes clear that voice analysis is *not* used for that purpose. Shin “compute[s] a distance between a user and the electronic

device” using camera and/or sensors in its “distance detection module.” Petition, pp. 9-11. Thus, there is no express disclosure or any other reason to believe Shin discloses “analyze [the] voice ... *to classify the voice*,” as claimed.

3. Shin Does Not Disclose Generation of First and Second Output Sentence When the Acquired Voice has Been Classified as the First and Second Voice

80. Claim 1 recites: “generate a first output sentence as the output sentence when the acquired voice has been classified as the first voice, and generate a second output sentence in which information is omitted as compared to the first output sentence as the output sentence when the acquired voice has been classified as the second voice.” EX-1001, Claim 1. Similarly, Claims 4-5 recite: “generating, a first output sentence is generated as the output sentence when the acquired voice has been classified as the first voice, and a second output sentence is generated as the output sentence in which a part of information included in the first output sentence is omitted when the acquired voice has been classified as the second voice.” The Board’s preliminary construction does not affect this limitation.

81. None of the prior art references cited in the Petition disclose or suggest classifying the acquired voice as the first and second voice as claimed. Nor does the prior art disclose or suggest generating a first and second output sentence when the acquired voice has been classified as the first and second voice.

4. Shin Does Not Disclose the Claimed Components Arranged as in the Claims

82. The Petition does not identify where the components of claim 1 are actually disclosed in Shin. The Petition focuses on functions without identifying what *component* allegedly accomplishes those functions. For example, the Petition describes the '337 patent's "voice classifying unit" as referring to "a CPU configured to perform the claimed function, namely, to analyze the user's voice and to classify it as a first or second voice." Petition, p. 11. As another example, the Petition describes the '337 patent's proximity sensor as "calculating the user-to-device distance, the distance can be used as a 'feature value' to perform the classification as the first or second voice." *Id.*

83. The Petition introduces Kristjansson in Ground 5 to assert disclosure of certain recited components that perform certain functions. But even then, the Petition only attempts to identify a few of the recited components in dependent claim 2. Petition, pp. 65-68 (*e.g.*, "intention analyzing unit" and "acquisition content information acquiring unit").

B. Ground 2: Shin Does Not Render the Challenged Claims Obvious

84. In Ground 2, modifying Table 3, as Amazon suggests, would limit the amount of distance-based *outputs*. Forcing a limitation on the outputs would not necessarily mean that Shin would have had to change anything about how inputs are processed.

C. Ground 3: Shimomura Alone Does Not Render the Challenged Claims Obvious

1. Shimomura Does Not Disclose Classification of a Voice as Either a First Voice or a Second Voice

85. Limitations 1[b], 4[b], and 5[b] require classes of *voices* (*i.e.*, “a first *voice* and a second *voice*”). EX-1001. None of the asserted grounds and prior art therein discloses classifying a voice spoken by a user “as either a first voice [and/or] a second voice,” as claimed and described in the specification.

86. Shimomura does not disclose any manner of classifying voices. The only basis the Petition provides for “classification” of a “voice” is in connection with Figure 7. The Petition does not identify any express disclosure of either aspect. Indeed, the Petition does not cite to a single disclosure of a “voice.” All mentions of “voice” (including those with quotes around them) are taken from Dr. Stern or the ’337 patent. Neither the Petition nor Dr. Stern provides any analysis supporting the conclusion that Shin classifies voices. Thus, regardless of the construction (*e.g.*, the Board’s preliminary construction or the plain meaning), the Petition fails to demonstrate that Shimomura discloses a voice.

87. To the extent Shimomura discloses any required, claimed classification, and it does not, Shimomura discloses classes of distances, not voices. Amazon argues that Shimomura discloses, in connection with Figure 7, “classifying voice into *two*

categories based on the user-to-device distance: (1) a first voice at greater than 350 cm, and (2) a second voice at less than 350 cm.”

[FIG. 7]

Second Voice	First Voice
Less than 350 cm	350 cm or more
None	Implementation

Petition, p. 40 (emphasis added); EX-1004, Fig. 7 (annotated by Amazon).

88. Dr. Stern added the “First Voice” and “Second Voice” labels to Figure 7. EX-1002, ¶ 64. Annotations aside, Figure 7 discloses one and only one thing: a calculation of distance. That determination of distance cannot expressly disclose a classification because it is a measurement of a continuous variable, not discrete classification. Nor can it expressly disclose an analysis of a “voice,” which is not at all synonymous with “distance.” The multiple-users example described above applies here as well.

89. Shimomura confirms the distinction between “voice,” “distance,” and “volume.” Shimomura states the following:

However, conventional robots equipped with such *voice* recognition and dialogue control functions are constructed to always converse with the user at a constant speech *volume* set in advance, regardless of the physical *distance* between the robot and the user.

EX-1004, ¶ [0004] (emphasis added).

90. A POSITA would not have understood Shin's disclosure of distance to be synonymous with "voice." Nor would a POSITA have understood Shin's determination of a distance to be synonymous with a classification.

91. In summary, like Shin, Shimomura directly correlates the output content and volume based on distance without performing the step of classifying the voice. EX-1004, ¶¶ [0058], [0067]-[0070]. To the extent Shimomura discloses classification, Shimomura discloses classes of distance—not voices similar to as discussed for Shin.

92. The above reasoning applies even under the Board's preliminary construction. Even if a new voice could be created based solely on a volume change based on distance, that does not mean that a new voice is always created simply because distance is changed. In other words, the concepts of "volume" and "distance" would still be broader than the concept of a "voice." And the Petition provides no basis upon which the Board could conclude that Shimomura discloses the concept of determining a voice as opposed to simply making a measurement of "distance." And a POSITA would never have understood Shin's disclosure of calculating a change in distance of a user, who otherwise does not change their voice at all, as any sort of voice classification—even if volume changed as well.

2. Shimomura Does Not Analyze the Acquired Voice to Classify the Acquired Voice Based on Calculated Distance Between the User and the Voice-Content Control Device

93. Limitation 1[b] recites: “a voice classifying unit configured to analyze a voice spoken by a user and acquired by a voice acquiring unit to classify the voice as either one of a first voice or a second voice based on the distance between the user and the voice-content control device.” EX-1001, Claim 1. Similarly, limitations 4[b] and 5[b] recite: “analyzing the acquired voice to classify the acquired voice as either one of a first voice and a second voice based on the distance between the user and the voice-content control device.” *Id.*, Claims 4-5.

94. These limitations require the claimed voice classification to be based on: **(1)** analysis of the acquired voice; and **(2)** distance separately calculated (*e.g.*, by a proximity sensor) in limitations 1[a], 4[a], and 5[a]. *See* Section §IV. None of the asserted grounds and prior art references therein discloses or suggests the claimed “analyz[e/ing] the [acquired] voice to classify the [acquired] voice.”

95. Amazon does not explain if or how Shimomura’s determination of user-to-device distance (*i.e.*, the basis of the alleged voice classification) involves any analysis of the user’s voice. Petition, pp. 37-40. In fact, Shimomura does not disclose determining user-to-device distance based on any voice analysis. Rather, Shimomura relies exclusively on camera 20 in image recognition unit 41 to detect distance. EX-1004, ¶¶ [0039] (“The image recognition unit 41 also detects the

distance to the imaging object by means of a so-called stereo visual method based on the image signal S1 from the CCD [] camera 20”), [0059] (“... distance information indicating the distance to the conversation partner detected by the image recognition unit 41....”).

96. Accordingly, Shimomura does not disclose “analyz[ing] a user’s voice ... to classify the voice” as recited in claim 1 (and similarly recited in claims 4-5) under both the plain meaning of the claims, the preliminary construction, or PO’s construction because no voice analysis whatsoever is involved in Shimomura’s determination of user-to-device distance based by image recognition unit 41.

3. Shimomura Does Not Disclose Adjust the Sound Volume of Voice Data Such that the Sound Volume of Voice Data Obtained by Converting the First Output Sentence thereinto Differs from the Sound Volume of the Voice Data Obtained by Converting the Second Output Sentence thereinto

97. Claim 1 requires the following: “the output controller is further configured to adjust the sound volume of voice data such that the sound volume of voice data obtained by converting the first output sentence thereinto differs from the sound volume of the voice data obtained by converting the second output sentence thereinto.” EX-1001, Claim 1. Similarly, Claims 4-5 require the following: “at adjusting the sound volume of voice data, further adjusting the sound volume of voice data such that the sound volume of voice data obtained by converting the first output sentence thereinto differs from the sound volume of voice data obtained by

converting the second output sentence thereinto.” The Board’s preliminary construction does not affect this limitation.

98. The Petition states that Shimomura “does not specify that the output sentences have different volumes.” Petition, p. 47. This is because Shimomura does not expressly disclose a connection between the processes associated with Figures 7 and 9. And Shimomura discloses, in connection with Figures 7 and 9, transforming the output content and changing volume at different distances:

[FIG. 7]

Second Voice	First Voice
Less than 350 cm	350 cm or more
None	Implementation

[FIG. 9]

Less than 50 cm	50~80cm	80~150cm	150~250cm	More than 250
0	+1	+2	+3	+5

52

FIG. 9 Volume Change Table

99. As shown above, I put EX-1004, Fig. 7 (transforming output content when user-to-device distance is 350cm or more) first followed by Fig. 9 (changing volume when user-to-device distance is less than 250cm).

100. The Petition states that “first output sentence” (*i.e.*, transformed character string at distance of 350 cm or more) and “second output sentence” (*i.e.*, original character string at distance of less than 350 cm) would have the same volume at

distance more than 250 cm (as provided in Figure 9 of Shimomura). Petition, pp. 43-46; EX-1004, Figs. 7, 9. The Petition further states that “[a] POSITA would have been motivated to modify Shimomura’s volume change table to include a range ending at 350 cm,” either by “extending the 150-250 cm range to end at 350 cm,” as shown in Amazon’s “First Modified Volume Change Table,” or by “adding a new range from 250 cm to 350 cm,” as shown in Amazon’s “Second Modified Volume Change Table”:

Less than 50 cm	50~80 cm	80~150 cm	150~350 cm	More than 350 cm
0	+1	+2	+3	+5

First Modified Volume Change Table

Less than 50 cm	50~80 cm	80~150 cm	150~250 cm	250~350 cm	More than 350 cm
0	+1	+2	+3	+5	+7

Second Modified Volume Change Table

Petition, pp. 47-48 (Both modified and annotated in yellow by Amazon)

101. A POSITA would not have had any reason to alter Shimomura’s expressed teachings in the way the Petition suggests, as discussed below.

102. Shimomura expressly discloses transforming the output content and changing volume at different distances. EX-1004, Figs. 7, 9. This is because Shimomura discloses different options that work at different distances. Shimomura recognizes

that “there is a certain limit to the speech volume of the robot 1,” such that “it may not be possible even if the speech volume should be determined to be a larger volume in accordance with the distance with the user when determining the volume in step SP22.” *Id.*, ¶¶ [0072]-[0073]. When the user is at a distance from Shimomura’s robot where further increase in volume is not possible or not effective in making the robot’s speech easier for the user to hear, Shimomura discloses changing the robot’s speech form (other than speech volume), including speech speed, pause length, phrase-end emphasis, and most drastically, transforming the speech character string by adding “ne” to the end of each phrase. *Id.*, ¶¶ [0060]-[0066], [0074]-[0079], [0084]-[0088].

103. As shown in Figures 7 and 10-12, Shimomura discloses changing the robot’s speech form at user-to-device distances that do not align with those disclosed in Figure 9 for changing speech volume:

[FIG. 9]

Less than 50 cm	50~80cm	80~150cm	150~250cm	More than 250
0	+1	+2	+3	+5

52

FIG. 9 Volume Change Table

[FIG. 7]

Less than 350 cm	350 cm or more
None	Implementation

FIG. 7 Character string transformation determination table

[FIG. 10]

Less than 200 cm	200~350cm	350 cm or more
0%	+20%	+50%

53

FIG. 10 Speed change table

[FIG. 11]

Less than 100 cm	100~350cm	350 cm or more
0%	+30%	+60%

54

FIG. 11 Pause length change table

[FIG. 12]

Less than 200 cm	Less than 200 cm
None	Raise

55

FIG. 12 Phrase-end emphasis determination table

Id., Figs. 7, 9-12.

104. Accordingly, a POSITA would have understood Shimomura to teach that increasing speech volume alone may not be possible or sufficient for a robot's speech to be easily understandable by a user, especially at a greater user-to-device distance.

Id., ¶¶ [0072]-[0073]. A POSITA also would have understood that, in addition to changing speech volume, other speech forms, such as speech speed, pause length, phrase-end emphasis, and speech character string could also have been changed depending on user-to-device distance to make Shimomura's robot's speech easier to understand. EX-1004, ¶¶ [0060]-[0066], [0074]-[0079], [0084]-[0088]. A POSITA would have understood from Shimomura that the user-to-device distance at which the various speech forms discussed above should be changed would have differed from one speech form to another. EX-1004, Figs. 7, 9-12. For instance, in view of

Figures 7 and 9-12 of Shimomura, a POSITA would have understood that, at a user-to-device distance of less than 350 cm, a combination of changes to speech volume, speech speed, pause length, and phrase-end emphasis would have been sufficient for Shimomura's robot's speech to be understandable without transforming the speech character string. EX-1004, Figs. 7, 9-12. This is for a good reason—at such closer distances, transforming the speech character string (*e.g.*, by adding “ne” to the end of each phrase, as taught by Shimomura) would have been unnecessary, unnatural, and potentially annoying to a user.

105. Thus, a POSITA would have recognized that Shimomura teaches transforming a speech character string (*e.g.*, by adding “ne” to the end of each phrase, as taught by Shimomura) only when changes to other speech forms (*e.g.*, speech volume, speech speed, pause length, phrase-end emphasis) are insufficient for the robot to be easily understandable by the user. To the extent a POSITA would have been motivated to modify Shimomura's user-to-device distance at which to change the various speech forms, the POSITA would not have been motivated to set the distance at which to change speech volume to be the same as the distance at which to transform the speech character string. This is because a POSITA would have recognized (in view of Shimomura's teachings in connection with Figures 7 and 9-12) that, even when Shimomura's robot output speech is at maximum volume, a combination of speech speed, pause length, and phrase-end emphasis could still have

been changed. Doing so would have improved the ease of comprehension of the robot's speech before it was necessary to transform the speech character string. A POSITA would have understood Shimomura's choice to set different distances as discouraging a POSITA from changing them to be the same.

4. Shimomura Does Not Disclose Generation of First and Second Output Sentence When the Acquired Voice has Been Classified as the First and Second Voice

106. Claim 1 recites: "generate a first output sentence as the output sentence when the acquired voice has been classified as the first voice, and generate a second output sentence in which information is omitted as compared to the first output sentence as the output sentence when the acquired voice has been classified as the second voice."

EX-1001, Claim 1. Similarly, Claims 4-5 recite: "generating, a first output sentence is generated as the output sentence when the acquired voice has been classified as the first voice, and a second output sentence is generated as the output sentence in which a part of information included in the first output sentence is omitted when the acquired voice has been classified as the second voice." The Board's preliminary construction does not affect this limitation.

107. The Petition relies on Shimomura to assert that the claimed omitted information is disclosed in the form of "ne." Petition, p. 46. But Shimomura states that such disclosure is a "change[] in intonation." EX-1004, ¶ [0092] ("changes in

intonation (emphasis at the end of each phrase) ...”). It is not an omission of information.

108. None of the prior art references cited in the Petition discloses or suggests classifying the acquired voice as the first and second voice as claimed. Nor does the prior art disclose or suggest generating a first and second output sentence when the acquired voice has been classified as the first and second voice.

D. Ground 4: Shimomura in View of Shin Does Not Render the Challenged Claims Obvious

109. I addressed Shin and Shimomura above. That discussion applies to the combination of Shimomura and Shin. Here, I further address the assertions of obviousness in connection with the proposed combination of those references.

110. Shimomura alone would not have rendered obvious “the output controller is further configured to adjust the sound volume of voice data such that the sound volume of voice data obtained by converting the first output sentence thereinto differs from the sound volume of the voice data obtained by converting the second output sentence thereinto” as required by claim 1, and “at adjusting the sound volume of voice data, further adjusting the sound volume of voice data such that the sound volume of voice data obtained by converting the first output sentence thereinto differs from the sound volume of voice data obtained by converting the second output sentence thereinto,” as required by claims 4-5.

111. For Ground 4, Amazon asserts that these limitations would have been obvious in view of Shin. Petition, pp. 59-62. But nothing in Shin would have motivated a POSITA to deviate from Shimomura’s expressed teachings. At best, Shin suggests to a POSITA that the user-to-device distance at which to change the output volume and output content *could* be the same. In the context of Shimomura, however, a POSITA would have recognized that doing so would have been undesirable as it would have burdened Shimomura’s robot and user with superfluous phrases (*e.g.*, “ne”) when they are unnecessary for the robot to be understood.

112. Amazon’s proposed motivation to combine would not have been known prior to the ’337 patent. The Petition alleges that the proposed modification of Shimomura in view of Shin “would enable Shimomura’s device to classify a voice into one of two distance-based categories and adjust output volume level and content based on that category, rather than referencing tables with distinct ranges.” Petition, p. 61. But Shin, like Shimomura, discloses tables with more than two distance ranges. EX-1003, Tables 1, 3. Neither Shin nor Shimomura discloses classifying a voice into one of two categories. Petition, pp. 30-35 (“A POSITA would have been motivated for several reasons to implement Shin’s volume change table (Table 1) and amount of information table (Table 3) with only two distance ranges.”).

113. In summary, a POSITA would not have been motivated by Shin to modify Shimomura in a manner that renders obvious “the output controller is further

configured to adjust the sound volume of voice data such that the sound volume of voice data obtained by converting the first output sentence thereinto differs from the sound volume of the voice data obtained by converting the second output sentence thereinto,” as required by claim 1, or “at adjusting the sound volume of voice data, further adjusting the sound volume of voice data such that the sound volume of voice data obtained by converting the first output sentence thereinto differs from the sound volume of voice data obtained by converting the second output sentence thereinto,” as required by claims 4-5.

E. Ground 5: The Combination of Shin, Shimomura, and Kristjansson Does Not Render the Challenged Claims Obvious

114. I addressed Shin, Shimomura, and the proposed combination of Shin and Shimomura above. That discussion applies to the combination of either or both of Shin and Shimomura with Kristjansson. Here, I further address the assertions of obviousness in connection with the proposed combination of either reference with Kristjansson.

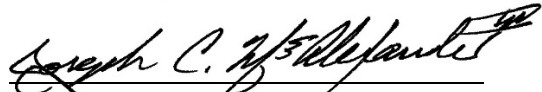
F. Grounds 2-5

115. The Petition relies on disclosures of discrete distance measurements (between a device and a user) in Shin and Shimomura as allegedly disclosing the claimed analysis and classification. Each would operate exactly the same regardless of whether any voice changed, so long as the distance remained constant. The

simplicity of that metric demonstrates that the authors of Shin and Shimomura specifically declined to conduct an analysis of voices to classify voices, let alone based on distance. They purposely chose simpler, more discrete approaches. Thus, adding the claimed analysis and voice-classification would have achieved the opposite. It would have mandated a more complex approach than implementing a single-variable measurement in simple, two-dimensional tables.

I, Joseph C. McAlexander, P.E., do hereby declare and state, that all statements are made herein of my own knowledge are true and that all statements made on information and belief are believed to be true. I am over the age of 21 and am competent to make this declaration. These statements were made with the knowledge that willful false statements are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code.

Dated: December 9, 2025



Joseph C. McAlexander, P.E.