

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
Patent No. 11,069,337

DECLARATION OF RICHARD STERN, PH.D.

<i>Amazon v. SoundClear</i> US Patent 11,069,337 Amazon Ex. 1002
--

TABLE OF CONTENTS

I.	BACKGROUND	1
A.	Experience and Qualifications	1
B.	Materials Considered	3
II.	APPLICABLE LEGAL STANDARDS.....	4
A.	Claim Construction	4
B.	Anticipation.....	6
C.	Obviousness	7
III.	PERSON OF ORDINARY SKILL IN THE ART	10
IV.	TECHNOLOGY BACKGROUND.....	12
A.	Classifying Users Based on Proximity.....	13
B.	Tailoring Content Based on Distance	15
C.	Adjusting Volume Based on Distance	17
D.	Adjusting Both Volume and Content Based on Distance.....	18
V.	THE '337 PATENT.....	21
A.	Overview	21
B.	Priority Date	24
VI.	CLAIMS 1-5 OF THE '337 PATENT ARE ANTICIPATED AND WOULD HAVE BEEN OBVIOUS.....	24
A.	Claims 1-5 Are Anticipated by Shin.	24
1.	Claim 1	24
a.	1[pre]: Voice-Content Control Device	24

b.	1[a]: Proximity Sensor.....	25
c.	1[b]: Voice Classifying Unit	28
d.	1[c]: Process Executing Unit.....	33
e.	1[d]: Voice-Content Generating Unit.....	35
f.	1[e]: Output Controller	37
g.	1[f]: Generate a First Output Sentence.....	37
h.	1[g]: Generate a Second Output Sentence.....	39
i.	1[h]: Output Controller Adjusts Volume of Voice Data	42
2.	Claim 2.....	44
a.	2[a]: Process Executing Unit.....	44
b.	2[b]: Voice-Content Generating Unit.....	47
3.	Claim 3.....	48
4.	Claims 4 and 5	49
B.	Claims 1-5 Would Have Been Obvious in View of Shin.	52
1.	Claim 1.....	52
2.	Claim 2.....	55
3.	Claim 3.....	58
4.	Claims 4 and 5	60
C.	Claims 1-5 Would Have Been Obvious in View of Shimomura.	60
1.	Claim 1.....	60

a.	1[pre]: Voice-Content Control Device	60
b.	1[a]: Proximity Sensor.....	63
c.	1[b]: Voice Classifying Unit	64
d.	1[c]: Process Executing Unit.....	67
e.	1[d]: Voice-Content Generating Unit.....	69
f.	1[e]: Output Controller	70
g.	1[f]: Generate a First Output Sentence.....	72
h.	1[g]: Generate a Second Output Sentence.....	75
i.	1[h]: Output Controller Adjusts Volume of Voice Data	78
2.	Claim 2.....	85
a.	2[a]: Process Executing Unit.....	86
b.	2[b]: Voice-Content Generating Unit.....	89
3.	Claim 3.....	90
4.	Claims 4 and 5	92
D.	Claims 1-5 Would Have Been Obvious in View of Shimomura and Shin.	93
1.	Claim 1	93
2.	Claim 2.....	96
3.	Claim 3.....	96
4.	Claims 4 and 5	99
E.	Claims 1-5 Would Have Been Obvious in View of Shin and/or Shimomura and Further in View of Kristjansson.	99

VII. SECONDARY CONSIDERATIONS OF NONOBVIOUSNESS	102
VIII. CONCLUSION.....	103

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

1. I am making this declaration at the request of Petitioners Amazon.com, Inc., Amazon.com Services LLC, and Amazon Web Services, Inc. (collectively, “Amazon”). I have been retained by Amazon as a technical expert in this matter.

2. I am being compensated for my work on this case. My compensation does not depend on the content of this Declaration or the outcome of these proceedings. I do not own any stock in Amazon and, to my knowledge, I have no financial interest in Amazon.

I. BACKGROUND

A. Experience and Qualifications

3. I am a Professor at Carnegie Mellon University in the Department of Electrical and Computer Engineering, the Department of Computer Science, and the Language Technologies Institute. I have been on the faculty of Carnegie Mellon since 1977.

4. I received the S.B. degree from the Massachusetts Institute of Technology (MIT) in 1970, the M.S. from the University of California, Berkeley, in 1972, and the Ph.D. from MIT in 1977, all in electrical engineering.

5. While I was a graduate student, I worked as a Teaching and Research Assistant in the Department of Electrical Engineering at MIT, from 1973 to 1976.

My teaching experience was in the areas of signal processing and acoustics. My research at MIT had been in the area of auditory processing and perception.

6. I am a fellow of the Institute of Electrical and Electronics Engineers (IEEE), the Acoustical Society of America, and the International Speech Communication Association (ISCA). I was the ISCA 2008-2009 Distinguished Lecturer, a recipient of the Allen Newell Award for Research Excellence in 1992, and I served as the General Chair of Interspeech 2006. Interspeech is the world's largest technical conference focused on speech processing and application.

7. Much of my current research is in spoken language systems, where I am particularly concerned with the development of techniques with which automatic speech recognition can be made more robust with respect to changes in environment and acoustical ambience. Among my longstanding work on speech processing are several projects involving speaker identification. Furthermore, as a Professor, I have taught 15 courses, many of which are at the graduate level, in the general areas of signal processing, communication theory, and acoustics.

8. My relevant publications are available on Carnegie Mellon's web site at <http://www.cs.cmu.edu/afs/cs/user/robust/www/papers.html>.

9. I understand a copy of my current curriculum vitae, which lists my publications for the last ten years, is being submitted as Exhibit 1006.

B. Materials Considered

10. In preparing this Declaration, I have considered the following materials:

Exhibit No.	Description
1001	U.S. Patent No. 11,069,337
1003	U.S. Patent App. Publ. No. 2017/0083281 (“Shin”)
1004	English Translation of Shimomura from Ex. 1005
1005	Declaration of Gwen Snorteland for Translation of Jap. Unexamined Patent App. Publ. 2005/202076 (“Shimomura”)
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”)
1013	U.S. Patent No. 10, 147,439 (“Kristjansson”)
1015	Nicolae Duta, <i>Natural Language Understanding and Prediction: From Formal Grammars to Large Scale Machine Learning</i> , 131 Fundamenta Informaticae 425 (2014) (“Duta”)

11. I understand that Shimomura published as a Japanese Patent Application, and that it has been translated by a certified translator. (See Ex. 1005.) In performing my analysis, I relied on the certified English translation provided by the translator, which I understand will be submitted as Exhibit 1004.

12. I have also relied on my education, training, and experience, and my knowledge of pertinent literature in the field of the '337 patent.

II. APPLICABLE LEGAL STANDARDS

13. I have been asked to provide my opinion as to whether the claims of the '337 patent would have been obvious to a person of ordinary skill in the art at the time of the alleged invention, in view of the prior art.

14. I am an electrical engineer by training and profession. The opinions I am expressing in this report involve the application of my training and technical knowledge and experience to the evaluation of certain prior art with respect to the '337 patent.

15. Although I have been involved as a technical expert in patent matters before, I am not an expert in patent law. Therefore, the attorneys from Knobbe, Martens, Olson & Bear, LLP have provided me with guidance as to the applicable patent law in this matter. The paragraphs below express my understanding of how I must apply current principles related to patent validity to my analysis.

A. Claim Construction

16. It is my understanding that in determining whether a patent claim is anticipated by or obvious in view of the prior art, the Patent Office construes the claim by giving the claim terms their plain and ordinary meaning, as they would

have been understood by a person of ordinary skill in the art at the time of the invention in view of the intrinsic record (patent specification and file history). For the purposes of this review, and to the extent necessary, I have interpreted each claim term in accordance with its plain and ordinary meaning as it would have been understood by a person of ordinary skill in the art at the time of the invention, in view of the intrinsic record. I have been instructed that the time of the invention is March 6, 2018, which I understand to be the earliest claimed priority date of the '337 patent.

17. I understand that a patent and its prosecution history are considered “intrinsic evidence” and are the most important sources for interpreting claim language in a patent. I also understand that in reading the claim, I must not import limitations from the specification into the claim terms; in other words, I must not narrow the scope of the claim terms by implicitly adding disclosed limitations that have no express basis in the claims. The prosecution history of related patents and applications can also be relevant.

18. I understand that 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 that such extrinsic sources cannot be used to contradict the unambiguous meaning of the claim language that is evident from the intrinsic evidence.

19. Unless expressly stated herein, I have applied the plain and ordinary meaning of the claim terms, which I understand is the meaning that a person of ordinary skill in the art would have given to terms in March 2018 based on a review of the intrinsic evidence.

B. Anticipation

20. It is my understanding that documents and materials that qualify as prior art can render a patent claim unpatentable as anticipated. I understand that all prior art references are to be looked at from the viewpoint of a person of ordinary skill in the art.

21. It is my understanding that a challenged claim is unpatentable as “anticipated” under 35 U.S.C. § 102 if all the limitations of the claim are described in a single prior art reference. It is my understanding that, to anticipate a claim, a prior art reference must disclose, either expressly or inherently, each and every limitation of that claim and enable one of ordinary skill in the art to make and use the invention.

22. It is my understanding that something is “inherent in,” and therefore taught by, the prior art, if it necessarily flows from the explicit disclosure of the prior art. I understand that the fact that a certain result or characteristic *may be* present in the prior art is not sufficient to establish inherency. However, if the result or characteristic is *necessarily* present based upon the explicit disclosure in the prior art, it is inherent in the prior art and is therefore disclosed.

C. Obviousness

23. It is my understanding that a claim is “obvious” if the claimed subject matter as a whole would have been obvious to a person of ordinary skill in the art at the time of the alleged invention. I understand that an obviousness analysis involves a number of considerations. I understand that the following factors must be evaluated to determine whether a claim would have been obvious: (i) the scope and content of the prior art; (ii) the differences, if any, between each claim of the ’337 patent and the prior art; (iii) the level of ordinary skill in the art in March 2018; and (iv) additional considerations, if any, that indicate that the invention was obvious or not obvious. I understand that these “additional considerations” are often referred to as “secondary considerations” of non-obviousness or obviousness.

24. I also understand that the frame of reference when evaluating obviousness is what a hypothetical person of ordinary skill in the pertinent art would have known in March 2018. I understand that the hypothetical person of ordinary skill is presumed to have knowledge of all pertinent prior art references.

25. I understand that a prior art reference may be a pertinent prior art reference (or “analogous art”) if it is in the same field of endeavor as the patent or if it is pertinent to the problem that the inventors were trying to solve. A reference is reasonably pertinent if it logically would have commended itself to an inventor’s attention in considering the problem at hand. If a reference relates to the same problem

as the claimed invention, that supports use of the reference as prior art in an obviousness analysis. Here, all of the references relied on in my obviousness analysis below are in the same field of endeavor as the '337 patent, e.g., controlling a device's voice output. The references are also pertinent to a particular problem the inventor was focused on, e.g., adjusting voice output to improve user interaction.

26. It is my understanding that the law recognizes several rationales for combining references or modifying a reference to show obviousness of claimed subject matter. Some of these rationales include:

- combining prior art elements according to known methods to yield predictable results;
- simple substitution of one known element for another to obtain predictable results;
- a predictable use of prior art elements according to their established functions;
- using known techniques to improve similar devices (methods, or products) in the same way;
- applying a known technique to a known device (method, or product) ready for improvement to yield predictable results;

- choosing from a finite number of identified, predictable solutions, with a reasonable expectation of success (in which case a claim would have been obvious to try);
- known work in one field of endeavor may prompt variations of it for use in either the same field or a different one based on design incentives or other market forces if the variations would have been predictable to one of ordinary skill in the art; and
- some teaching, suggestion, or motivation in the prior art that would have led one of ordinary skill to modify the prior art reference or to combine prior art reference teachings to arrive at the claimed invention.

27. I understand that “secondary considerations” must be considered as part of the obviousness analysis when present. I further understand that 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 by experts; (4) commercial success of a product covered by the patent; (5) unexpected results achieved by the claimed invention; (6) industry praise of the claimed invention; (7) deliberate copying of the invention; and (8) teaching away by others. I also understand that evidence of the independent and nearly simultaneous “invention” of the claimed subject matter by others is a secondary consideration supporting an obviousness determination and may support a conclusion that a claimed invention was

within the knowledge of a person of ordinary skill as of March 2018. I am not aware of any evidence of secondary considerations that would suggest that the claims of the '337 patent would have been nonobvious in March 2018.

28. I understand that when assessing obviousness, using hindsight is impermissible; that is, what is known today or what was learned from the teachings of the patent should not be considered. The patent should not be used as a road map for selecting and combining items of prior art. Rather, obviousness must be considered from the perspective of a person of ordinary skill at the time the alleged invention was made – March 2018 in this case.

29. I also understand that an obviousness analysis must consider the invention as a whole, as opposed to just a part or element of the invention. I understand this “as a whole” assessment to require showing that one of ordinary skill in the art at the time of invention, confronted by the same problems as the inventor and with no knowledge of the claimed invention, would have selected the elements from the prior art and combined them in the claimed manner.

III. PERSON OF ORDINARY SKILL IN THE ART

30. It is my understanding that when interpreting the claims of the '337 patent and evaluating whether a claim is anticipated or would have been obvious, I must do so based on the perspective of a person of ordinary skill in the art at the

relevant priority date. I have been instructed to assume for the purposes of my opinions that the relevant priority date of the '337 patent is March 6, 2018.

31. I understand that in determining the level of ordinary skill in the art, several factors are considered. Those factors may include: (i) the type of problems encountered in the art; (ii) prior art solutions to those problems; (iii) the rapidity with which innovations are made; (iv) the sophistication of the technology; and (v) the educational level of active workers in the field. A person of ordinary skill in the art must have the capability of understanding the scientific and engineering principles applicable to the pertinent art.

32. Based on my review of the specification and claims of the '337 patent, it is my opinion that a person of ordinary skill in the art would have had a minimum of a bachelor's degree in computer engineering, computer science, electrical engineering, or a similar field, and approximately two years of industry or academic experience in a field related to controlling the audio output of electronic devices. Work experience could substitute for formal education and additional formal education could substitute for work experience.

33. My conclusions below that the claims of the '337 patent are anticipated and would have been obvious would remain the same even if the priority date, field of endeavor, or level of ordinary skill were slightly different.

34. I meet the above definition of a person of ordinary skill in the art, and

did so as of March 2018. Also, I have worked with persons of ordinary skill in the art through my professional and academic experiences, and I have an understanding of their skill level around March 2018.

IV. TECHNOLOGY BACKGROUND

35. The '337 patent describes an electronic device, such as a speaker or phone, that receives and processes a user's voice input and outputs a voice response. (Ex. 1001 ('337 patent), Fig. 1, 2:62-3:1.) The patent claims priority to March 2018, years after Apple launched Siri in its iPhones (2011), Amazon launched Alexa-enabled devices (2014), and Google launched its voice assistant (2016). The patent admits that devices that detect voice, perform processing according to the user's intent, and provide a voice output were known. (*Id.*, 1:21-28.) Public reporting confirms that both Siri- and Alexa-enabled devices performed these functions. (*E.g.*, *Amazon Just Surprised Everyone with a Crazy Speaker that Talks to You*, <https://www.theverge.com/2014/11/6/7167793/amazon-echo-speaker-announced> (Echo "will provide information, music, news, weather, and more whenever you ask for it," "can hear users from across the room," and can be used for "telling you the local weather" or asking "general questions and get[ting] answers from Wikipedia"); *iPhone 4S: a Siri-ously Slick, Speedy Smartphone*, <https://arstechnica.com/gadgets/2011/10/iphone-4s-a-siri-ously-slick-speedy-smartphone/> ("Siri

offers iPhone 4S users a way to interact with their devices, apps, and data with natural language voice commands. Users can send and receive texts or e-mails simply by talking to Siri, find out whether there will be rain today, create new Reminders and calendar items, ask for directions, move appointments, and more.”.)

A. Classifying Users Based on Proximity

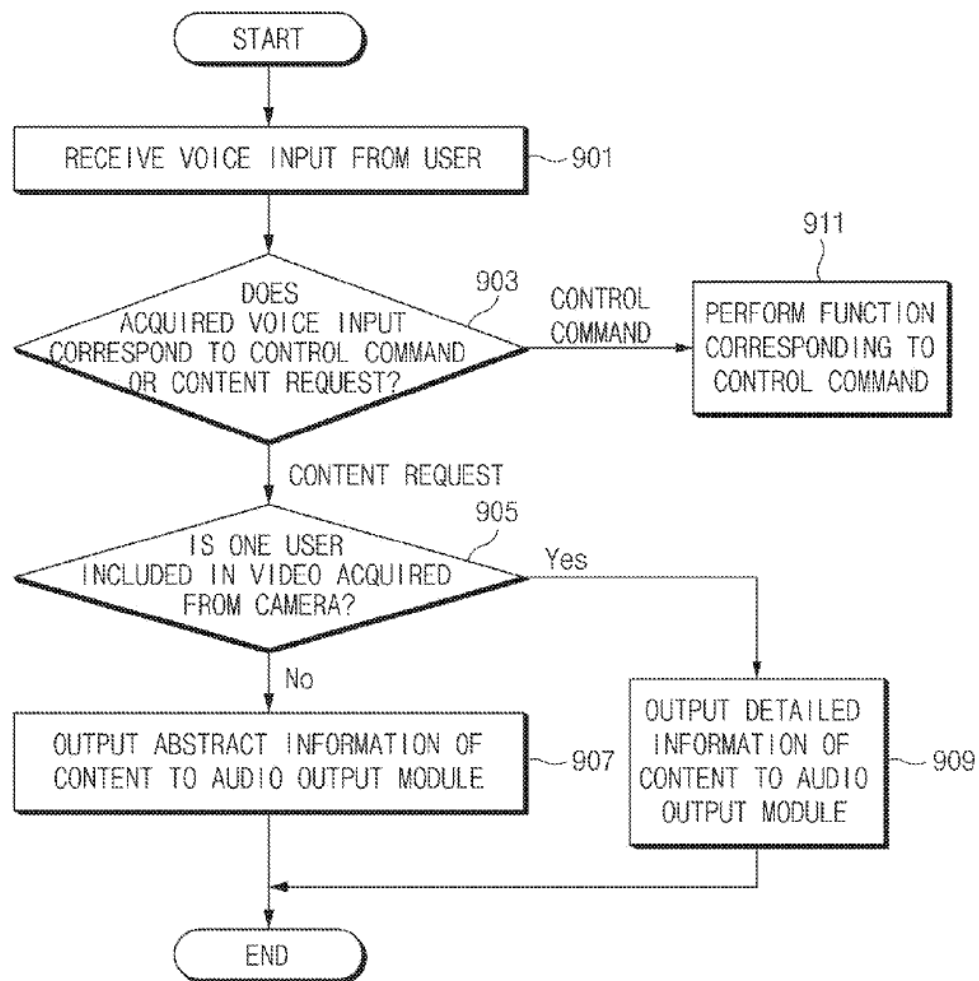
36. Many devices that classified users based on their proximity to the device were known in the art. For example, Kim, which was published in 2017, disclosed a device with this functionality. Kim disclosed that its device “collects context information associated with the user” after the device receives a user’s voice input. (Ex. 1009 (Kim) ¶¶[0045], [0163] (“If the voice is collected, ... the processor 120 may collect context information. If a user voice signal is received, the processor 120 ... may collect user related information, using a sensor[.]”).) The device can obtain user information using a variety of sensors, but notably, it can use a “proximity sensor” to detect “distance information” (i.e. the distance between a user and the device). (*Id.* ¶[0068] (“The sensor 180 may include ... a proximity sensor ... According to an example embodiment, the sensor 180 may collect ... distance information from [the user]”).) The device can determine whether a user is a “first specified user” based on the “context information.” (*Id.* ¶[0169] (“the processor 120 ... may determine whether the context information meets a first specified condition ... The first specified condition may include a condition in which a user who asks is a

first specified user.”); *see also id.* ¶[0244] (“the electronic device 1500 may use a sensor ... for identifying at least one user”).)

37. Aoyama, which also published in 2017, provides another example of an “information processing device” which classified users based on proximity. (Ex. 1010 (Aoyama), Abstract.) Aoyama discloses a device that determines a “user state” based on analysis of “image” and “sound input.” (*Id.* ¶[0114] (“the user state estimation unit 15 mainly estimates the user state on the basis of the analysis results of an image and the sound input has been described above”).) The device can determine various “user states,” such as whether a user is “stopped,” “walking,” or “running.” (*Id.* ¶[0106].) Based on this determination, the device may classify users as “calm” or “hurried.” (*Id.*) A person of ordinary skill in the art would have understood that, in order to make the determination that a user is either “stopped” or “running,” the device necessarily must calculate its proximity to the user at sequential time intervals. Further, Aoyama notes that “the type of information used to estimate the user state is not particularly limited as long as the information is information such as an image, a sound input, or detection results by various sensors.” (*Id.* ¶[0114]; *see also id.* ¶[0107] (“the estimation method is not particularly limited as long as the user state estimation unit 15 can estimate ... a motion of the user”).)

B. Tailoring Content Based on Distance

38. Devices that tailor content based on the distance between the device and the user were also known in the art. In 2016, Ha described an electronic device that “provides content in response to [] voice input.” (Ex. 1011 (Ha), Abstract.) The device can provide either “abstract” or “detailed” information to a user. (*Id.* ¶[0042] (“the processor 120 may determine to control the audio output module 150 to output abstract information of content”); *see also id.* ¶¶[0044], [0048], [0050], [0052].) Ha’s device may determine which type of content to provide to a user based on “a video of the vicinity of the electronic device 100 shot by the camera 170.” (*Id.* ¶[0050].) Ha discloses that the device may determine if “a user is present in the vicinity of the electronic device.” (*Id.* ¶[0114].) When the device makes the determination that a user is in its vicinity, it can provide “detailed information.” (*Id.*) Ha’s Figure 9 illustrates this process:



(*Id.*, Fig. 9.) If the device’s camera registers a user after receiving voice input, it would “output detailed information.” (*Id.*) Alternatively, if the camera did not register a user (e.g. because the user was too far from the camera), or alternatively, if it registered multiple users, it would output “abstract information.” (*Id.*; see also *id.* ¶¶[0109]-[0116].) Ha’s claims provide further support for this functionality. Specifically, claim 14 states that the device comprises “a camera configured to shoot a video of a vicinity of the electronic device” and a “processor [] further configured to

determine the output scheme of the content to be outputted through the audio output module or the display based on the video of the vicinity of the electronic device.”

(*Id.*, claim 14.)

C. Adjusting Volume Based on Distance

39. Devices that adjust volume based on the distance between the device and the user were known in the art. For example, Iyer, which issued in 2016, describes an “electronic device includ[ing] a voice control interface.” (Ex. 1012 (Iyer), Abstract). When the device receives a “speech command” from a “first distance,” it produces “an audible output at a first output level.” (*Id.*) If a user gives a speech command from a “second distance [that] is less than the first distance” (i.e. the distance between the user and the device has decreased), the device produces audible output at a “second output level less than the first level.” (*Id.*; *see also, e.g., id.*, 3:30-36 (in the “first mode,” the device produces, “through a loudspeaker and in response to the speech command, an audible output at a first output level”), 3:19-27 (the device is also capable of a secondary “‘discreet’ mode of operation”), 3:58-61 (in the discreet mode, the device generates audible “responses at a level that others will not overhear”), 7:47-8:3, 12:6-17.)

40. Kristjansson, which issued in 2017, provides a further example in its “speech-capturing device.” (Ex. 1013 (Kristjansson), Abstract.) The device first

detects a user's "spoken utterance." (*Id.*; *see also id.*, 3:3-4.) The device then "determines the distance of the user from the device." (*Id.*; *see also id.*, 3:8-10 (determining the user's "location relative to the device" through "range finding"), 15:57-16:5 (the system can use a "proximity/distance detector" to calculate a user's position), claims 1, 5, 12.) Based on the calculated distance, the device then "adjust[s] the gain for output audio (such as a spoken response to the utterance) to ensure that the output audio is at a certain desired sound pressure when it reaches the location of the user." (*Id.*, Abstract; *see also id.*, 3:16-19 (calculating "a gain for output audio data to ensure a desired volume of the output audio when it reaches the user's location"), 14:43-15:16 (describing the calculations used to determine the "appropriate volume for output audio" based on the distance to the user); *see generally id.*, Figs. 6-9.)

D. Adjusting Both Volume and Content Based on Distance

41. Using a proximity sensor to determine a user's distance from a device and then tailoring the information provided and the volume based on that distance has been known for decades. In 2005, Shimomura described an electronic device, such as a robot, that converses with a user. (Ex. 1004 (Shimomura), Abstract.) The device included cameras for determining the distance between the device and the user. (*Id.* ¶¶[0039], [0093].) The device classified the user's speech based on distance (e.g., whether the user was more or less than 350 cm from the device) and then

adjusted the output content (e.g., including or omitting words) and speech volume accordingly. (*Id.* ¶¶[0049]-[0051], [0057], [0086]-[0088].)

42. In 2017, Shin disclosed a device (e.g., smartphone, portable multimedia player, or home appliance) that acquires a user’s voice input via a microphone, analyzes it, and provides content as voice output via a speaker. (Ex. 1003 (Shin) ¶¶[0031], [0037].) Shin used a “distance detection module” or “proximity sensor” to “compute a distance between a user and the electronic device[.]” (*Id.* ¶¶[0066], [0153], [0053], [0075], [0077], Figs. 5A, 9.)

43. Shin’s device classified voice input based on distance. For users within 1 meter, it outputted non-abbreviated content (e.g., “detailed content” in which the content is “described most precisely”) at a lower volume (e.g., 40 dB). (*Id.* ¶¶[0051], [0088], Tables 1, 3.) For users more than 1 meter away, it outputted “abbreviated content” at a higher volume (e.g., 45 dB or more). (*Id.*) An example of this procedure in Shin’s system is shown in Figure 5A:

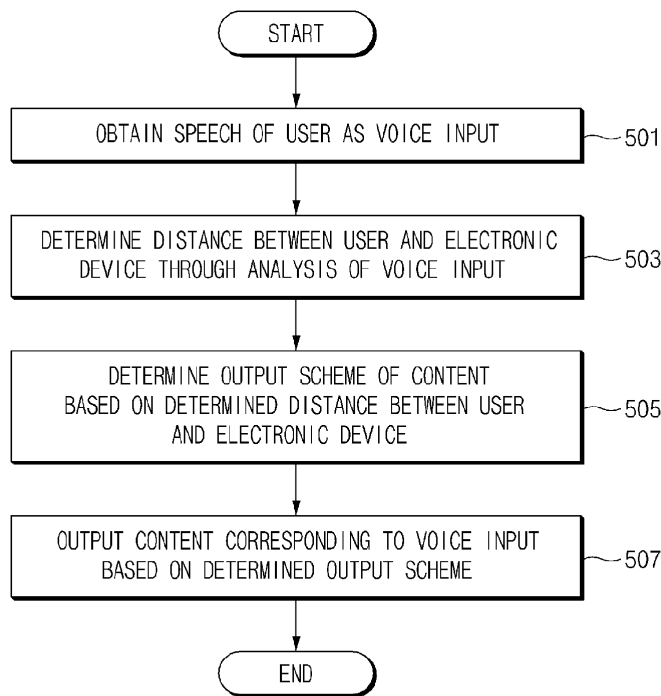


FIG. 5A

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

44. As another example, if a user said, “Let me know today’s weather,” Shin’s device obtained weather-related information and responded with, for example: “The weather in Jul. 1, 2015 is (①) rainy after cloudy (②). The highest temperature is 28° C., and the lowest temperature is 18° C. (③), the rainfall is 10 mm (④). Prepare your umbrella when you go out. (⑤).” (*Id.* ¶[0088].) If the voice was classified as a near voice (e.g., less than 1 meter away), all five categories of information were provided. (*Id.*) If the voice was classified as farther away (e.g., user between 1 and 2 meters away), category ⑤ was omitted. (*Id.*) And, if the voice

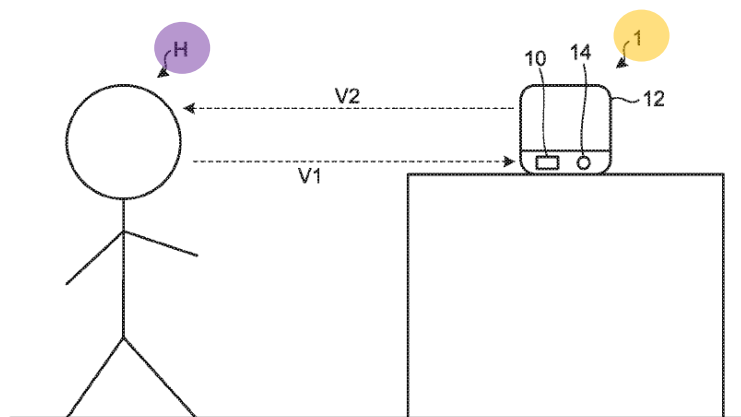
was classified as even farther away (e.g., greater than 4 meters), all but category ② was omitted. (*Id.*) Thus, Shin’s device used a proximity sensor to calculate distance between the user and the device, classify the user’s voice based on that distance, and tailor the response’s content and volume based on the classification. This allowed Shin’s device to provide a “suitable” amount of information at a “suitable” volume based on distance. (*Id.* ¶¶[0089], [0080].)

V. THE ’337 PATENT

A. Overview

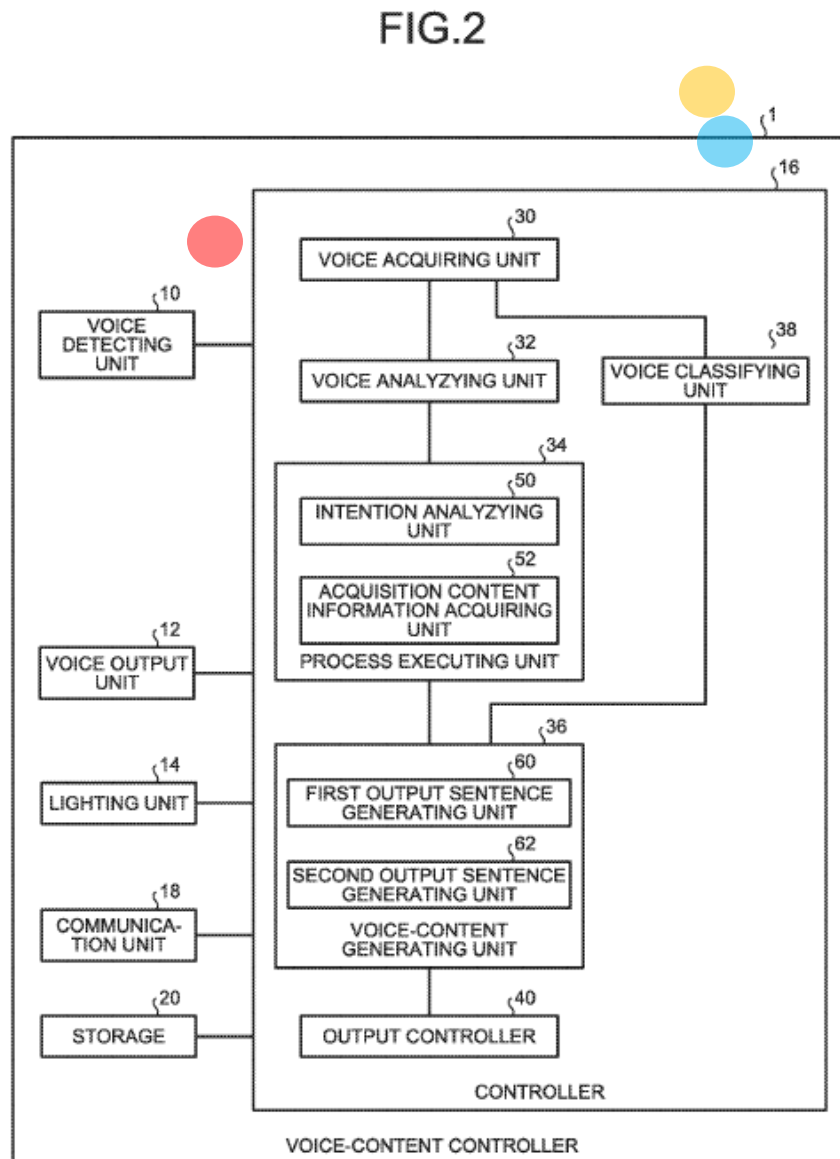
45. The ’337 patent describes a device, referred to as a “voice-content control device,” that analyzes a user’s voice and generates a response. (Ex. 1001 (’337 patent), Abstract.) Figure 1 illustrates such a device 1 (orange) that detects the voice V1 of a user H (purple), processes it, and outputs a responsive voice V2:

FIG.1



(*Id.*, Fig. 1¹, 2:62-3:1.)

46. The device 1 includes various components, such as a voice detecting unit 10 (red) and a controller 16 (blue):



¹ Some figures and tables in this Declaration may be colored or otherwise annotated for clarity.

(Ex. 1001 ('337 patent), Fig. 2, 3:12-17.) The controller includes several “units”—a voice acquiring unit, a voice analyzing unit, a process executing unit, and a voice-content generating unit—for acquiring the user’s voice, analyzing it, processing its request, and generating a response. (*Id.*, 3:50-59.) For example, the process executing unit analyzes the user’s speech (e.g., “How’s the weather today?”) detected with the voice detecting unit (e.g., microphones) and obtains the requested information (e.g., weather, air temperature, and chance of rainfall). (*Id.*, 4:12-14, 4:56-61, 7:17-27.)

47. The device may include a “proximity sensor” for calculating the distance to the user. (Ex. 1001 ('337 patent), 8:20-26.) This generic sensor, which the patent describes in a single sentence, allows the device to classify a detected voice as either a first voice or a second voice. (*Id.*) When the voice is classified as the first voice (e.g., nearer than a threshold distance), the device may generate a first sentence and output it at a first volume. (*Id.*, 9:38-50, 13:62-65.) When the voice is classified as the second voice (e.g., farther than a threshold distance), the device may generate a second sentence that omits some information from the first sentence and output it at a different volume. (*Id.*)

48. These features purportedly allow the device to provide content that is understandable and that accounts for the user’s environment. (Ex. 1001 ('337 patent), 1:29-42.)

B. Priority Date

49. I understand that the '337 patent issued on July 20, 2021, from U.S. Patent Application 16/290,983 (“the '983 application”), filed March 4, 2019. I also understand that the '983 application claims priority to Japanese Application No. 2018-039754, filed on March 6, 2018. Thus, I understand that the '337 patent’s earliest possible priority date is March 6, 2018.

VI. CLAIMS 1-5 OF THE '337 PATENT ARE ANTICIPATED AND WOULD HAVE BEEN OBVIOUS.

A. Claims 1-5 Are Anticipated by Shin.

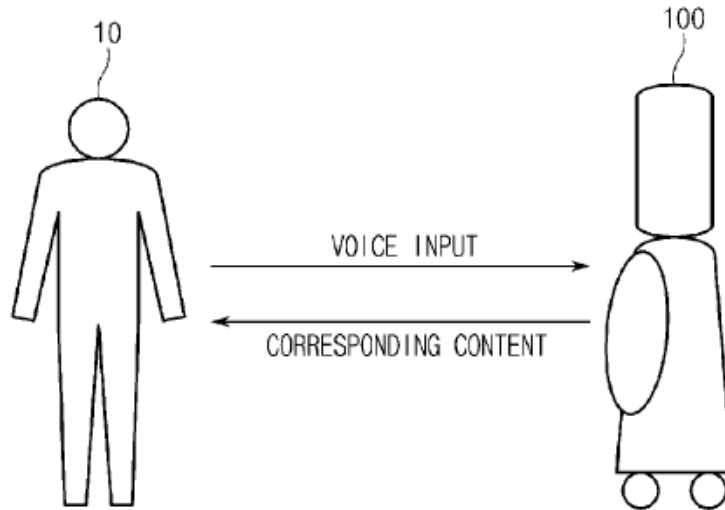
50. For at least the reasons I discuss below, Shin anticipates claims 1-5.

1. Claim 1

a. 1[pre]: Voice-Content Control Device

51. The preamble of claim 1 recites a “voice-content control device.” Shin discloses an electronic device that receives voice input from a user through a microphone, “generate[s] content corresponding to a result of analyzing the voice input,” and “provide[s] the generated content as sound through an embedded audio output module (e.g., a speaker).” (Ex. 1003 (Shin) ¶[0037]; *see also id.*, Abstract, ¶¶[0003], [0008]-[0011], [0038]-[0039], [0048]-[0060], [0069]-[0072], claims 1, 15, 20.) For example, if a user says, “Let me know what time it is now,” Shin’s device may respond with, “The current time is nine ten AM.” (*Id.* ¶[0037].) The device may be a

smartphone, PC, or home appliance. (*Id.* ¶[0031].) Figure 1A of Shin shows an example of the device 100:



(*Id.*, Fig. 1A.) Thus, Shin discloses a “voice-content control device.”

b. 1[a]: Proximity Sensor

52. Claim element 1[a] recites “a proximity sensor configured to calculate a distance between a user and the voice-content control device.” Shin discloses this claim element.

53. Shin discloses that the device includes a “distance detection module” that “compute[s] a distance between a user and the electronic device[.]” (Ex. 1003 (Shin) ¶[0066]; *see also id.* ¶¶[0053] (“the processor 120 may determine a distance between the user and the electronic device 101 based on ... the distance computed, calculated, or measured by the distance detection module 180”), [0075] (“the electronic device 101 may measure a distance between the user and the electronic device

101 through the distance detection module 180”), [0077] (processor 120 can “determine [the] distance between the user and the electronic device 101 based on ... the distance computed by the distance detection module 180”), Fig. 5A (after the device “obtain[s] [the] speech of [the] user as voice input,” it then “determine[s] distance between user and electronic device though analysis of [the] voice input”).)

54. The distance detection module 180 may include “a depth camera like a time-of-flight (TOF) camera, a stereo camera computing depth information using triangulation, a charge coupled device (CCD) camera computing a distance through an image processing, or the like” and/or “various sensors,” such as “an infra-red sensor, an RF sensor, an ultrasonic sensor, and the like.” (Ex. 1003 (Shin) ¶[0066]; *see also id.* ¶[0042] (electronic device may include “a sensor”).)

55. Shin’s Figure 3 shows the electronic device 101 comprising the distance detection module 180 (green):

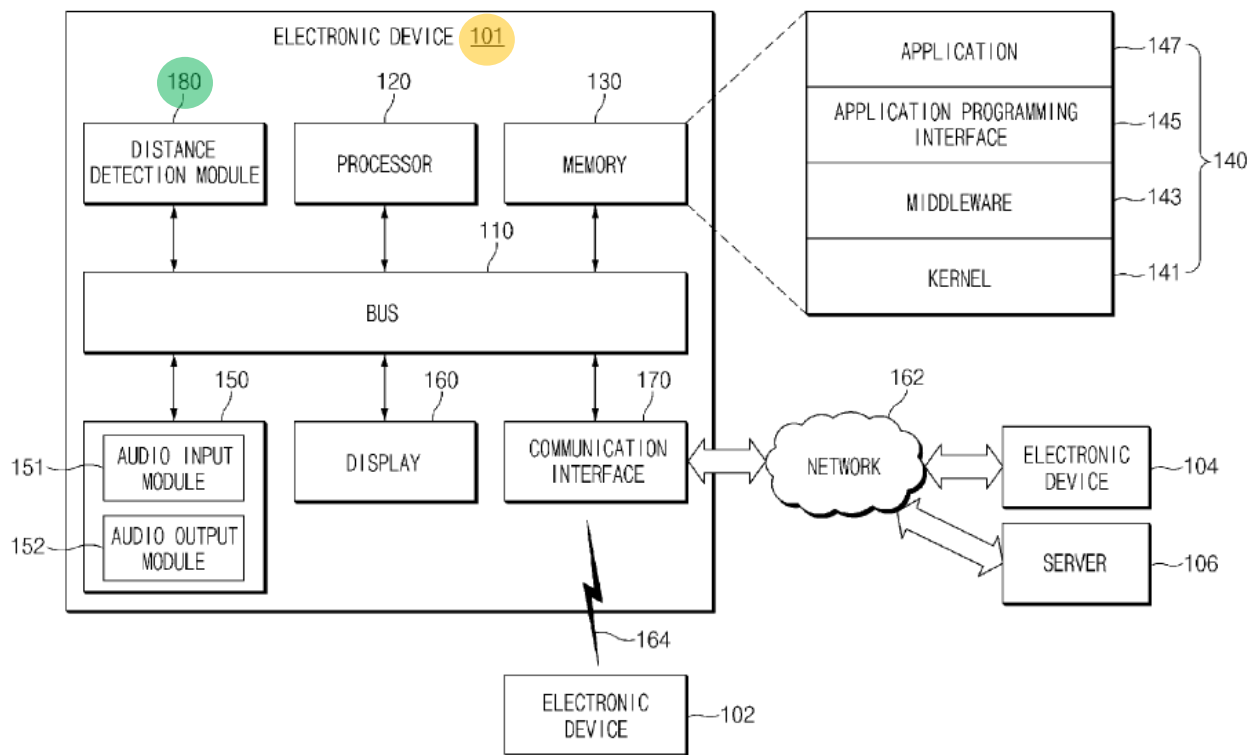


FIG. 3

(Ex. 1003 (Shin), Fig. 3, ¶¶[0045]; *id.* ¶¶[0053] (“the processor 120 may determine an output scheme of corresponding content based on a distance between a user and the electronic device”), [0054], [0077] (“the electronic device 101 may determine a distance between the user and the electronic device”), [0096].) I note that, while Figure 3 shows “electronic device 101,” Shin discloses that its electronic device 100 may be implemented with the modules of electronic device 101. (*Id.* ¶¶[0041], [0043].)

56. Shin’s claims further show the device’s proximity sensing capability. For example, claim 5 discloses that the “distance detection module” is “configured

to compute a distance between the user and the electronic device.” (Ex. 1003 (Shin), claim 5; *see also id.*, claim 9.)

57. Shin also expressly discloses that the electronic device 101 may use “a proximity sensor.” (Ex. 1003 (Shin) ¶[0153], Fig. 9 (showing “proximity sensor” 940G in “sensor module” 940); *see also id.* ¶[0143] (electronic device 901 may be included in electronic devices 100, 101).)

58. Thus, Shin discloses a proximity sensor configured to calculate a distance between a user and the voice-content control device.

c. 1[b]: Voice Classifying Unit

59. Claim element 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.” Shin discloses this claim element.

60. The ’337 patent discloses that a voice classifying unit may be part of a controller, which performs processes by reading software/program stored in a storage. (Ex. 1001 (’337 patent), 3:50-59, Fig. 2.) This controller can be a central processing unit (CPU). (*Id.*) Thus, according to the ’337 patent, the “voice classifying unit” refers 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.

61. The '337 patent explains that, when a proximity sensor is used for 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. (Ex. 1001 ('337 patent), 8:20-26.) Specifically, the unit “sets a threshold of the feature value, and classifies the voice” as the first or second voice “based on whether the feature value exceeds the threshold.” (*Id.*, 8:30-34.) Shin discloses the same thing.

62. First, Shin discloses a voice classifying unit that is configured to, and does, “analyze a voice spoken by a user and acquired by a voice acquiring unit,” as recited. Specifically, Shin discloses that the device 101 (orange) includes a processor 120 (blue) and an audio input module 151 (red):

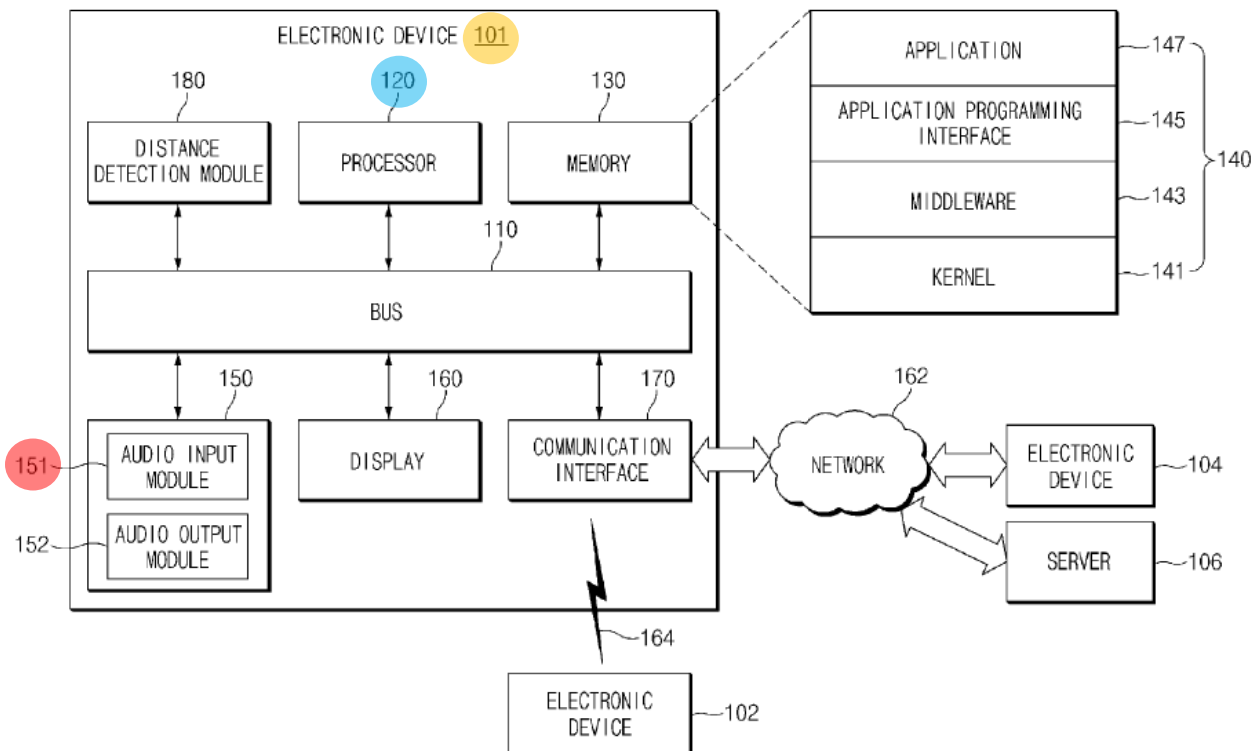


FIG. 3

(Ex. 1003 (Shin), Fig. 3, ¶[0045].) The processor 120, which may be a CPU (*id.* ¶[0029]), “analyze[s] a voice input received through [the] audio input module 151.” (*Id.* ¶[0048]; *see also id.*, Abstract, ¶¶[0003]-[0007], [0008] (the device “analyzes a voice input of a user's speech and determine[s] an output scheme of content based on the various speech features obtained from results of the analysis”), [0009]-[0011], [0037]-[0039], [0041], [0060] (“the application 147 may include a voice recognition application ... that is capable of analyzing a voice input”), [0061] (“The audio module 150 includes, for example, the audio input module 151”).) In fact, these functions are explicitly described in Shin’s claims. (*Id.*, claim 1 (“An electronic device comprising: an audio input module configured to receive a speech of a user as a voice input ... and a processor configured to determine an output scheme of the content based on ... an analysis of the voice input.”); *see also id.*, claims 15, 20.) The audio input module 151 can be “implemented with a microphone and the like” to “obtain a user’s speech as a voice input.” (*Id.* ¶[0062]; *see also id.* ¶¶[0037] (“The electronic device 100 may receive the voice input through an embedded audio input module (e.g., a microphone”), [0155], [0158] (“The audio module 980 may process, for example, sound information that is inputted ... through ... the microphone 988.”).) Thus, Shin discloses a voice classifying unit (processor) configured to analyze a voice spoken by a user (user’s speech) and acquired by a voice acquiring unit

(e.g., audio input module and/or portion of the processor that receives voice from the audio input module).

63. Second, Shin discloses that the voice classifying unit (processor) analyzes the voice to classify it as either a first or second voice based on the distance between the user and the device. Shin’s processor executes a voice recognition application to process the voice input and generates corresponding content for output. (Ex. 1003 (Shin) ¶¶[0047], [0060]; *see also id.* ¶¶[0037], [0039].) The 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[.]” (*Id.* ¶¶[0078], [0086]; *see also id.* ¶¶[0053], [0077]-[0080], claims 6, 17.) To do so, Shin’s processor classifies the voice based on the distance.

64. An example is shown in Shin’s Table 3, which discloses providing a different amount of information based on how far the user is from the electronic device:

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 (Shin), Table 3 (consolidated).) Thus, for example, 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:

TABLE 3		
	Distance (D) between user and electronic device	Output amount of information of content
First Voice	$D < 1 \text{ m}$	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
Second Voice	$D \geq 4 \text{ m}$	Content of fourth abbreviation level (abstract content)

(*Id.*, Table 3; *see also id.*, Table 1, ¶¶[0051] (content may be “classified dichotomously”), [0079]-[0081], [0086]-[0087].)

65. I note that any of the distance ranges up to 4 meters would satisfy the claimed “first voice.” For example, a voice between 3 and 4 meters away could be considered the recited “first voice.” Alternatively, all of the distances less than 4 meters could collectively be considered a first voice (e.g., a voice less than 4 meters away).

66. Thus, Shin discloses that the voice classifying unit (processor) is configured to analyze the voice to classify the voice as either one of a first voice (e.g.,

voice closer than 1 meter) or a second voice (e.g., voice farther than 4 meters) based on the distance between the user and the voice-content control device.

d. 1[c]: Process Executing Unit

67. Claim element 1[c] recites “a process executing unit configured to analyze the voice acquired by the voice acquiring unit to execute processing required by the user.” The ’337 patent explains that, like the voice classifying unit, a process executing unit may be part of a controller that performs processes by reading software/program stored in a storage. (Ex. 1001 (’337 patent), 3:50-59, Fig. 2.) In the ’337 patent, the “process executing unit” analyzes the user’s speech (e.g., “How’s the weather today?”) and obtains the requested information (e.g., weather, air temperature, and chance of rainfall). (*Id.*, 4:12-14, 4:56-61, 7:17-27.) The patent admits that devices that performed these functions were known. (*Id.*, 1:21-28; *see* ¶35, above.) Shin discloses this claim element.

68. Shin’s processor analyzes the voice acquired by the voice acquiring unit (e.g., the “audio input module” and/or the portion of the processor that receives voice from the audio input module). (Ex. 1003 (Shin) ¶¶[0048], [0062].) As I discuss above, Shin’s processor executes a voice recognition application to process the voice input and generate corresponding content for output. (*See* ¶¶59-66, above.) Additionally, Shin discloses that the voice input requests content in which the user is interested, indicating what processing is required or intended. (Ex. 1003 (Shin)

¶¶[0084] (“a user who requests content with a voice input to the electronic device 101 while being far away from the electronic device 101 may want a quick response”), [0089], [0059] (“determin[ing] that a user has an interest in the specific content”); *see also id.* ¶¶[0058], [0067], [0103], [0107], [0120].) The processor analyzes these requests to generate and/or provide the content. (*Id.* ¶¶[0048] (“the processor 120 may analyze a voice input ... and may output content corresponding to the voice input”), [0058], [0059] (“when it is determined that a user has an interest in the specific content, the processor 120 may provide more detailed information by increasing an output amount of information with respect to the content in which the user has an interest.”), [0060]; *see also id.* ¶¶[0037], [0071], [0088], [0126]-[0127], [0141].)

69. Indeed, Shin discloses analyzing the acquired voice “to execute processing required by the user” in the same way as the ’337 patent. For example, Shin describes analyzing a user’s speech (“Let me know today’s weather”) to execute processing required by the user to obtain the requested information (weather, air temperature, and rainfall). (Ex. 1003 (Shin) ¶[0088].)

70. Thus, Shin discloses a process executing unit (processor) configured to analyze the voice (speech) acquired by the voice acquiring unit to execute processing required by the user.

e. 1[d]: Voice-Content Generating Unit

71. Claim element 1[d] recites “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.” In the ’337 patent, the “voice-content generating unit” is the portion of the processor (CPU) that performs the recited function. (Ex. 1001 (’337 patent), 3:50-59, Fig. 2.) In the ’337 patent’s weather example, the “voice-content generating unit” performs this function when it generates the following output sentence: “On the day Z of the month Y of the year X, in Tokyo, the weather is partly cloudy, the chance of rainfall is 20%, the highest air temperature is 25 degrees, and the lowest air temperature is 15 degrees.” (*Id.*, 9:16-21.) Again, the patent admits that devices that performed these functions were known. (*Id.*, 1:21-28; *see* ¶¶35, above.) And, Shin discloses this claim element.

72. Shin discloses a voice-content generating unit that generates an output sentence as text data for a voice to be output to the user. As I discussed above for claim elements 1[b]-1[c], Shin’s processor generates the content for output to the user. (*See* ¶¶59-70, above.) For example, Shin discloses that the content generated can be a sentence including text data, such as: “The current time is nine ten AM.” (Ex. 1003 (Shin) ¶[0037].) In another example, Shin’s processor performs the claimed function when it generates the output sentence: “The weather in Jul. 1, 2015

is (①) rainy after cloudy (②). The highest temperature is 28° C., and the lowest temperature is 18° C. (③), the rainfall is 10 mm (④). Prepare your umbrella when you go out. (⑤).” (*Id.* ¶[0088].) For example, the output sentence may be “The weather in Jul. 1, 2015 is rainy after cloudy.” This content may be converted into a voice using text-to-speech technology and outputted audibly. (*Id.* ¶[0048] (“the content may be provided to a user as sound based on a text to speech (TTS) technology”); *see also id.* ¶¶[0007], [0037], [0039], [0062], [0072], [0109], [0131], [0139], [0158].) Thus, Shin discloses a voice-content generating unit (processor) configured to generate an output sentence that is text data for a voice to be output to the user.

73. Shin also discloses that the voice-content generating unit generates the output sentence based on content of the processing executed by the process executing unit. As I discussed above for claim element 1[c], Shin’s processor analyzes the voice input to generate the content which aligns with the user’s interests. (*See* ¶¶67-70, above.) Thus, Shin discloses a voice-content generating unit (processor) configured to generate, based on content (e.g., weather or time information) of the processing executed by the process executing unit (processor), output sentence that is text data for a voice to be output to the user.

f. 1[e]: Output Controller

74. Claim element 1[e] recites “an output controller configured to adjust a sound volume of voice data obtained by converting the output sentence thereinto.” Shin discloses this claim element.

75. As I discussed above regarding claim element 1[d], Shin’s processor converts the content to voice. (See ¶¶71-73, above.) Shin’s processor can “adjust” the “output volume level[.]” (Ex. 1003 (Shin) ¶[0053]; *see also id.* ¶¶[0076]-[0081], [0098]-[0101], [0105], claims 6, 17.) Thus, Shin discloses an output controller (processor) configured to adjust a sound volume of voice data obtained by converting the output sentence thereinto. I discuss claim element 1[h], which is similar, in detail below, which further shows that Shin discloses this claim element. (See ¶¶86-89, below.)

g. 1[f]: Generate a First Output Sentence

76. Claim element 1[f] recites that “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[.]” Shin discloses this claim element.

77. Shin’s processor adjusts the amount of information provided to the user based on the user-to-device distance. (Ex. 1003 (Shin) ¶[0053]; *see also id.* ¶¶[0051], [0086]-[0090], claims 8, 17.) Specifically, when a voice is classified as a

first voice (e.g., closer than 1 meter), the device generates a first output comprising all available information. (*Id.* ¶[0087] (when distance “is less than 1 m, an output amount of information of the content may be set highest” and provides “the most detailed original content corresponding to a voice input”).) In this scenario, the content output may include “rich and extended information” and may be “described most precisely.” (*Id.* ¶[0051].) An example of the amount of information provided when the voice is classified as the first voice is shown in Table 3:

TABLE 3

	Distance (D) between user and electronic device	Output amount of information of content
First Voice	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.) Thus, when classified as a first voice (e.g., closer than 1 meter), the processor outputs the content without abbreviation. (*Id.* ¶[0087].)

78. Shin discloses an example in which the “original content,” and thus the output sentence when the voice is classified as a first voice (e.g., closer than 1 meter), is: “The weather in Jul. 1, 2015 is (①) rainy after cloudy (②). The highest temperature is 28° C., and the lowest temperature is 18° C. (③), the rainfall is 10 mm (④). Prepare your umbrella when you go out. (⑤).” (Ex. 1003 (Shin) ¶[0088].) In this

example, the first output sentence may be “The weather in Jul. 1, 2015 is rainy after cloudy.” (*Id.*)

79. Thus, Shin discloses that the voice-content generating unit (processor) is further configured to generate a first output sentence (e.g., containing information of original content) as the output sentence when the acquired voice has been classified as the first voice (e.g., voice closer than 1 meter).

h. 1[g]: Generate a Second Output Sentence

80. Claim element 1[g] recites that the voice-content generating unit is further configured to “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.” Shin discloses this claim element.

81. Shin discloses that, whenever the user-to-device “distance increases by 1 m, the electronic device 101 may decrease the output amount of information by making the abbreviation level of the original content high.” (Ex. 1003 (Shin) ¶[0087]; *id.* ¶[0086] (“output amount of information of the content decreases as a distance between the user and the electronic device 101 increases”).) The varying abbreviation levels based on distance ranges are shown in Shin’s Table 3:

TABLE 3

	Distance (D) between user and electronic device	Output amount of information of content
First Voice	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
Second Voice	D ≥ 4 m	Content of fourth abbreviation level (abstract content)

(*Id.*, Table 3.) As shown, when the user is more than 4 meters from the device, the voice falls into a second classification and the processor outputs “Content of fourth abbreviation level (abstract content).” (*Id.* ¶[0087].) In other words, when the voice is classified as a second voice (e.g., greater than 4 meters), the output content is abbreviated, i.e., it omits some information provided when the voice is classified as a first voice. (*Id.* ¶[0051] (content may be classified into a second category that “includes only a gist” and is provided in a format that includes the “least” amount of information and the content “is described most simply”); *id.* (the processor may “extract[]” a portion of the content).)

82. In the weather example I discuss above, Shin’s device provides information ① – ⑤ when the voice is classified as a first voice. (Ex. 1003 (Shin) ¶[0088].) That includes generating the first output sentence “The weather in Jul. 1, 2015 is rainy after cloudy.” (*Id.*) In contrast, a voice from more than 4 meters away is classified as a second voice and the processor generates the second output sentence

“rainy after cloudy.” (*Id.* (“the content of a fourth abbreviation level (abstract content) may only include information of ②”).) Thus, Shin discloses omitting some information (e.g., ①, “The weather in Jul. 1, 2015 is”) when the voice is classified as the second voice. (*Id.*)

83. Shin also discloses that a voice between 1 and 2 meters away may be classified as a second voice and the processor generates an output sentence that includes information ① – ④ “while excluding ⑤.” (Ex. 1003 (Shin) ¶[0088] (discussing “content of a first abbreviation level”).) Thus, Shin also discloses omitting some information (“Prepare your umbrella when you go out”) when a voice is classified as between 1 and 2 meters away. (*Id.*)

84. Likewise, in the time example I discuss above, when Shin’s device is prompted with “voice input” such as “[l]et me know what time it is now,” the device returns the output that “the current time is nine ten AM.” (Ex. 1003 (Shin) ¶[0037].) Alternatively, if “the user [] moves away from the electronic device,” it may return an abbreviated response such as simply “nine ten.” (*Id.* ¶[0039].)

85. Thus, Shin discloses that the voice-content generating unit (processor) is further configured to generate a second output sentence (e.g., content of a fourth abbreviation level) in which information is omitted (e.g., information of ①) as compared to the first output sentence (e.g., information of original content) as the output

sentence when the acquired voice has been classified as the second voice (e.g., voice greater than 4 meters away).

i. 1[h]: Output Controller Adjusts Volume of Voice Data

86. Claim element 1[h] recites that “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.” Shin discloses this claim element.

87. As I discussed above for claim element 1[e], Shin’s processor adjusts the content’s output volume level based on the user-to-device distance. (See ¶¶74-75, above; Ex. 1003 (Shin) ¶¶[0053], [0077]-[0081].) Shin specifically discloses that the “electronic device 101 may set an output volume level of the content to 40 dB when the distance between the user and the electronic device 101 is less than 1 m and may increase the output volume level by 5 dB whenever the distance increases by 1m.” (*Id.* ¶[0080].) Shin’s Table 1 also discloses this:

TABLE 1

Distance (D) between user and electronic device	Output volume level of content
$D < 1 \text{ m}$	40 dB
$1 \text{ m} \leq D < 2 \text{ m}$	45 dB
...	...
$D \geq 5 \text{ m}$	65 dB

(*Id.*, Table 1.) Thus, Shin discloses that the output volume when the distance is between 4 and 5 meters may be 60 dB. (*Id.* ¶[0080], Table 1.)

88. Considering Shin’s disclosure regarding the output information and output volume together, Shin discloses that when a voice is classified as a first voice (e.g., closer than 1 meter), a first output sentence containing the “original content” is provided (Table 3) at a volume level of 40 dB (Table 1). When a voice is classified as a second voice (e.g., greater than 4 meters), a second output sentence containing “[c]ontent of fourth abbreviation level” is provided (Table 3) at a volume level of 60 or 65 dB (Table 1 and ¶[0080]). Shin repeatedly discloses that both content and volume may vary based on distance. (Ex. 1003 (Shin) ¶¶[0039] (when “the user 10 moves away from the electronic device 100, the electronic device 100 may provide the content as sound with a relatively louder output volume level”), [0049], [0053], [0140].)

89. Thus, Shin discloses that the output controller (processor) is further configured to adjust the sound volume of voice data such that the sound volume (e.g., 40 dB) of voice data obtained by converting the first output sentence (e.g., information of original content) thereinto differs from the sound volume (e.g., 60 or 65 dB) of the voice data obtained by converting the second output sentence (e.g., content at fourth abbreviation level) thereinto.

90. Accordingly, for at least the reasons I discuss above, Shin anticipates claim 1.

2. Claim 2

91. Claim 2 depends from claim 1 and recites two additional claim elements, which Shin discloses. Moreover, these claim elements reflect the operation of known voice processing systems like Alexa and Siri, and were therefore well known to people of ordinary skill in the art. (*See* ¶35, above.)

a. 2[a]: Process Executing Unit

92. Claim element 2[a] recites that the “process executing unit” comprises: [i] “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 [ii] an acquisition content information acquiring unit configured to acquire acquisition content information which is notified to the user based on the extracted intention information.” The ’337 patent discloses that the “intention analyzing unit” extracts intention information indicating the user’s intent when, for example, it analyzes the phrase “How’s the weather today?” and recognizes that the user wants weather-related information. (Ex. 1001 (’337 patent), 4:56-66.) The “information acquiring unit” in the ’337 patent acquires content information that is notified to the user based

on the extracted intent information when, for example, it acquires the weather-related information. (*See, e.g., id.*, 7:14-27, 7:28-36, 9:9-16 (intention information is “weather” and acquisition information is “partly cloudy”).)

93. As I discussed above for claim element 1[c], Shin’s processor analyzes the requests from the voice input to identify the content the user is interested in, and then generates the requested content. (*See* ¶¶67-70, above.) Shin expressly discloses that the device can “receive the voice input ... and may generate content corresponding to a result of analyzing the voice input.” (Ex. 1003 (Shin) ¶[0037].) As I discussed above, in response to a user’s statement, “Let me know what time it is now,” the device “may generate content, such as ‘The current time is nine ten AM[.]’” (*Id.*; *see also id.* ¶[0039].) Shin further discloses that, in response to the voice input, “Let me know today’s weather,” the device generates a response containing weather-related information, such as “cloudy.” (*Id.* ¶[0088].) To obtain such content, the device may store an application in memory that “analyz[es] a voice input[] and generat[es] and/or provid[es] the content corresponding to the voice input.” (*Id.* ¶[0060].) The device can also delegate some processing to another device (e.g., another electronic device or a server), which then delivers the results. (*Id.* ¶[0067] (“all or a part of operations that the electronic device 101 will perform may be executed by another or plural electronic devices”); *see also id.* ¶¶[0064]-[0065], [0174]-

[0176].) Accordingly, Shin discloses acquiring the information from memory or from another device.

94. Even if Shin had not expressly disclosed this claim element, Shin inherently discloses it. Shin's processor necessarily includes an "intention analyzing unit" and an "acquisition content information acquiring unit," as recited in claim element 2[a], because such units would have been required for Shin's device to function as described. For example, to provide content such as weather or time information in response to voice input, Shin's processor necessarily analyzes the voice input to determine the user's intent (e.g., a request for weather or time) and necessarily acquires the corresponding content information (e.g., "cloudy" or "nine ten AM"). These functionalities are required for the processor to obtain and deliver the requested content to the user in the manner Shin describes.

95. Thus, Shin expressly or inherently discloses that the process executing unit (processor) comprises: an intention analyzing unit configured to extract intention information (e.g., today's weather or current time) indicating an intention of the user based on the voice acquired by the voice acquiring unit. Shin also expressly or inherently discloses that the process executing unit (processor) comprises an acquisition content information acquiring unit configured to acquire acquisition content information (e.g., weather-related information or time information) which is notified to the user based on the extracted intention information (e.g., weather or time).

96. Moreover, the '337 patent admits this claim element was known. (Ex. 1001 ('337 patent), 1:21-28 (known devices analyze user voice, perform processing “according to the user’s intention,” and output, via voice, content of a user’s inquiry).)

b. 2[b]: Voice-Content Generating Unit

97. Claim element 2[b] recites that “the voice-content generating unit is further configured to generate the text data including the acquisition content information as the output sentence.” Shin discloses this claim element.

98. As I discussed above for claim element 2[a], Shin’s processor acquires information, such as time-related or weather-related information, for generating a response to the user. (See ¶¶92-96, above; Ex. 1003 (Shin) ¶¶[0037], [0039], [0088].) This information is incorporated into the content provided via voice to the user, as demonstrated by the examples I discussed. (*Id.*) The content is provided to the user “as sound based on a text to speech (TTS) technology.” (*Id.* ¶[0048].) Thus, Shin discloses that the voice-content generating unit (processor) is configured to generate the text data including the acquisition content information (e.g., weather-related information or time information) as the output sentence.

99. Accordingly, for at least the reasons I discuss above, Shin anticipates claim 2.

3. Claim 3

100. Claim 3 depends from claim 1 and further recites “wherein, on generating the second [output] sentence, the voice-content generating unit is further configured to omit a part of information included in the voice spoken by the user.” Shin discloses such a voice-content generating unit.

101. As I discussed above for claim element 1[g], Shin discloses classifying a voice received from a distance of 4 meters or more as a second voice. (*See* ¶¶80-85, above; Ex. 1003 (Shin), Table 3.) For that second voice, in response to the request “Let me know today’s weather,” the second output sentence is simply: “rainy after cloudy.” (Ex. 1003 (Shin) ¶[0088].) This output omits all user-spoken words, as well as information such as “Jul. 1, 2015” (corresponding to “today’s”) included in the user’s voice. (*Id.*; *see also id.* ¶[0039] (the input “What time is it now?” results in output “nine ten,” omitting all words and information included in the voice spoken by the user).) Thus, Shin discloses, on generating the second output sentence, the voice-content generating unit (processor) is further configured to omit a part of information included in the voice spoken by the user.

102. Alternatively, Shin’s processor may generate the second output sentence to include content at a first level of abbreviation when the acquired voice is received from a distance between 1 and 2 meters away. (*See* ¶83, above; Ex. 1003 (Shin), Table 3.) This first level of abbreviation also omits part of the information

included in the voice spoken by the user. For example, in response to the request, “Let me know today’s weather,” the processor outputs information that omits the information “[l]et me know.” (Ex. 1003 (Shin) ¶[0088].)

103. Accordingly, for at least the reasons I discuss above, Shin anticipates claim 3.

4. Claims 4 and 5

104. Claim 4 recites “[a] voice-content control method” and is even broader than claim 1. The steps of claim 4 generally correspond to elements of claim 1 and are therefore disclosed by Shin for the reasons I discuss above. Indeed, Shin discloses both electronic devices and methods of operating the electronic devices. (Ex. 1003 (Shin) ¶[0181]; *see also id.* ¶¶[0003], [0008]-[0011], [0016]-[0021], [0043], [0047], [0092]-[0093], Title, Abstract, claims 15-19.)

105. Claim 5 recites “[a] non-transitory storage medium that stores a voice-content control program that causes a computer to execute” a series of steps, which are the same as the steps recited in claim 4. Thus, the steps recited in claim 5 are disclosed by Shin for the reasons I discuss above. And, Shin discloses a non-transitory storage medium. Specifically, Shin discloses electronic devices implemented by instructions stored in a computer-readable storage media. (Ex. 1003 (Shin) ¶[0181]; *see also id.* ¶¶[0011], [0182], [0179]-[0180], claim 20.)

106. I have compared the elements in claims 4 and 5 to the elements in claim 1. As shown, other than their preambles, claims 4 and 5 are identical and merely recite the functions performed by the various units recited in claim 1. The table below demonstrates how the elements correspond. As I discuss above for claim 1, Shin discloses all the steps recited in claims 4 and 5 and thus anticipates claims 4 and 5. (See ¶¶51-90, above.)

	Claim 4	Claim 5		Claim 1
4[pre]/5[pre]	A voice-content control method, comprising:	A non-transitory storage medium that stores a voice-content control program that causes a computer to execute:	1[pre]	A voice-content control device, comprising:
4[a]/5[a]	calculating a distance between a user and a voice-content control device;		1[a]	a proximity sensor configured to <i>calculate a distance between a user and the voice-content control device</i> ;
4[b]/5[b]	acquiring a voice spoken by a user;		1[b]	a voice classifying unit configured to <i>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</i> ;
4[c]/5[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;		1[b]	

4[d]/5[d]	analyzing the acquired voice to execute processing intended by the user;	1[c]	a process executing unit configured to <i>analyze the voice acquired by the voice acquiring unit to execute processing required by the user;</i>
4[e]/5[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	1[d]	a voice-content generating unit configured to <i>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</i>
4[f]/5[f]	adjusting a sound volume of voice data obtained by converting the output sentence thereinto,	1[e]	an output controller configured to <i>adjust a sound volume of voice data obtained by converting the output sentence thereinto,</i>
4[g]/5[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,	1[f]- 1[g]	wherein the voice-content generating unit is further configured to <i>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,</i>

4[h]/5[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.	1[h]	wherein the output controller is further configured to <i>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.</i>
-----------	--	------	--

B. Claims 1-5 Would Have Been Obvious in View of Shin.

107. As I have discussed above, Shin discloses each limitation of claims 1-5 and thus anticipates those claims. However, even if Shin failed to disclose any particular limitation of claims 1-5, such a limitation would have been obvious to a person of ordinary skill in the art over Shin.

1. Claim 1

108. Even if claim 1 were to require the user's voice to be classified into no more than two classifications based on distance, Shin would render this obvious. A person of ordinary skill in the art 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.

109. First, Shin suggests doing so. Shin discloses that content can be divided into two classifications. (Ex. 1003 (Shin) ¶[0051] (“detailed content” and “abstract content”).) In fact, Shin explicitly states that the detailed and the abstract content “may be classified dichotomously.” (*Id.*) Based on that disclosure, a person of ordinary skill in the art would have been motivated to also use two volume levels, one for each classification.

110. Second, Shin discloses that its “embodiments ... are not limited to the examples of Table 1, Table 2, and Table 3.” (Ex. 1003 (Shin) ¶[0090]; *see also id.* ¶[0081] (“Table 1 is merely provided as an example, but embodiments of the present disclosure are not limited to the example of Table 1.”).) This flexibility in Shin’s disclosure would have motivated a person of ordinary skill in the art to implement a simpler, two-classification system—such as “near” and “far”—sufficient for applications requiring only basic distinctions to adjust content output based on user distance.

111. Third, simplifying complex systems into fewer categories is a routine and well-established approach in user interfaces and voice-controlled devices to reduce computational load and improve efficiency. Binary classification (classifying into two categories) is a standard approach in many fields, including voice-controlled devices (*see e.g.*, Ex. 1004 (Shimomura), Figs. 7, 12), to simplify user experience. Many consumer devices adopt binary distinctions, such as “low” and “high”

or “near” and “far.” A person of ordinary skill in the art would have recognized binary classification as an efficient, user-friendly solution for Shin’s described needs. For example, many systems developed in my own work rely on binary classification such as detecting the presence or absence of speech (*e.g.*, Vuong et al., 2021, 2023) or determining whether the source of detected speech is actually a person in the room or a recording (*e.g.*, Lindsey et al., 2023).

112. Fourth, such a modification reflects merely applying a known technique (binary classification) to achieve predictable results (a streamlined user interface requiring fewer processing resources, having faster response times, and providing straightforward user experience with content output adjusted based on either “near” or “far” proximity classifications).

113. Fifth, such a modification reflects merely applying a known technique (binary classification) to a known device (Shin’s device and method) that is ready for improvement and yields predictable results (consistent and reliable adjustments to content detail and volume based on user proximity, effectively aligning with Shin’s goal of enhancing user interaction).

114. Sixth, such a modification would have been obvious to try. Shin’s system accommodates adaptable classifications, and a person of ordinary skill in the art would have reasonably expected success in using binary classification to meet Shin’s

functional goals while simplifying implementation and minimizing computational demand.

115. A person of ordinary skill in the art would have reasonably expected success in modifying Shin in this way. Doing so would have been trivial, merely involving configuring Shin’s processor with revised tables for two proximity-based categories and optimizing performance for specific uses.

116. Accordingly, for at least the reasons I discuss above, Shin renders obvious claim 1.

2. Claim 2

117. Even if Shin had not explicitly disclosed the additional limitations of claim 2, the claim would have been obvious over Shin. As I discuss above, Shin’s processor analyzes the voice input to identify the content the user is interested in, and then acquires and provides that content. (See ¶¶67-70, 92-96, above.) A person of ordinary skill in the art would have recognized that implementing an “intention analyzing unit” to extract the user’s intent from user voice data, alongside an “acquisition content information acquiring unit” to retrieve content based on the extracted intention, would be a predictable and logical extension of Shin’s functionalities.

118. As I discuss above, Shin describes voice processing capabilities enabling the device to interpret user intent, such as identifying requests for time or

weather, and generate appropriate responses. (See ¶¶92-96, above.) A person of ordinary skill in the art would have been motivated to structure the processor with an intention analyzing unit and a content acquisition unit for greater efficiency. This modular design would facilitate consistent intent extraction, reliable content retrieval, and improved clarity and organization of the device’s processes. It would also merely reflect the way voice assistants such as Apple’s Siri and Amazon’s Alexa worked at the time of the invention. (See ¶35, above.) For example, Duta discloses that conventional voice analysis systems in 2014 included an “[i]ntent classifier” module that is used to “find[] the query intent (what the user asks for).” (Ex. 1015 (Duta), 431-432, Fig. 1.) Such systems also included a “dialog manager” module that is used to respond to the user, for example by asking for additional information or confirmation. (*Id.*, 432-433, Fig. 1.) Duta states that its description of conventional systems “generalizes all systems built in the past” and is illustrated in Figure 1:

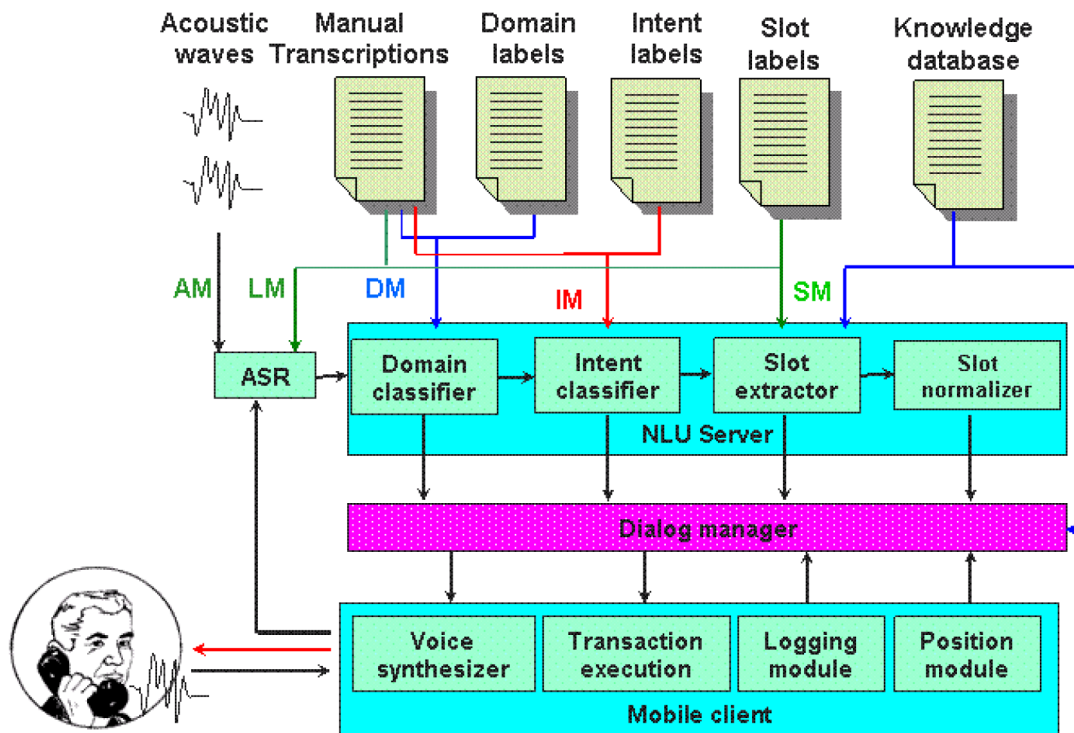


Figure 1. The architecture of a client-server voice-based transactional system including dialog.

(*Id.*, 432, Fig. 1.) Thus, this approach merely reflects known voice analysis techniques. (*Id.*, 431-433, Fig. 1.) This is further demonstrated by my discussion of Kristjansson below. (See ¶¶216-223, below.)

119. This approach would have been straightforward for a person of ordinary skill in the art to implement, leading to a reasonable expectation of success, because Shin's device already integrates voice analysis and content generation. Implementing these functionalities as processor units would have been an obvious design choice, enhancing modularity, maintainability, and troubleshooting. Using these units, Shin's device could improve response accuracy and handle diverse user requests with greater flexibility.

120. Accordingly, for at least the reasons I discuss above, Shin renders obvious claim 2.

3. Claim 3

121. Even if Shin had not explicitly disclosed the additional limitations of claim 3, the claim would have been obvious over Shin. A person of ordinary skill in the art would have been motivated for several reasons to configure Shin's processor to selectively omit parts of a user's spoken information when generating an output sentence.

122. First, this would have been a predictable and desirable implementation of Shin's processor, which processes and responds to user inputs. A person of ordinary skill in the art would have understood that repeating all user-spoken information (e.g., the user speaks, "What time is it?" and the response begins with "The time is ...") would be unnecessary, as the user already knows that information. Avoiding repetition improves efficiency and prevents delays in response delivery.

123. Second, market demands and user expectations for efficient, streamlined interactions with automated systems would have motivated this implementation. In my experience, users expect minimal delays and concise responses. For example, my own research group has found that users indeed prefer faster and more concise responses, and that delays of longer than 500 ms will motivate users to avoid

interacting with the system. (Teal and Rudnicky, 1992.) Omitting user-spoken information would align Shin's processor with these expectations, improving user satisfaction. A person of ordinary skill in the art would have recognized this as an obvious solution to meet these demands.

124. Third, such an implementation would have been a matter of routine experimentation and common sense. Given Shin's aims to enhance user interactions (Ex. 1003 (Shin) ¶¶[0007]-[0009], [0037], [0140]-[0141]), reducing unnecessary repetition to streamline responses would have been a logical step and would predictably improve performance without requiring significant innovation.

125. Fourth, such an implementation would have been obvious to try. Shin's device presents only two design options for responding: include (1) all information from a user's request or (2) less than all. Omitting some information to enhance efficiency and conciseness would have been an obvious choice.

126. A person of ordinary skill in the art would have reasonably expected success in implementing Shin's processor to omit some user-spoken information when responding. Based on my experience, this trivial implementation would have involved straightforward changes to configure the processor for more efficient and streamlined interactions and was well within the skill of a person of ordinary skill in the art.

127. Accordingly, for at least the reasons I discuss above, Shin renders obvious claim 3.

4. Claims 4 and 5

128. Claims 4 and 5 would have been obvious in view of Shin for the reasons I discuss above regarding claim 1. (*See* ¶¶51-90, 104-106, above.)

C. Claims 1-5 Would Have Been Obvious in View of Shimomura.

129. For at least the reasons I discuss below, Shimomura renders obvious claims 1-5.

1. Claim 1

a. 1[pre]: Voice-Content Control Device

130. The preamble of claim 1 recites a “voice-content control device.” To the extent the preamble is limiting, Shimomura discloses a “speech control device that controls the speech of a device[.]” (Ex. 1004 (Shimomura), claim 1.) Indeed, Shimomura’s title is, “Speech control device and method and robot device.” (*Id.*, Title.) Shimomura discloses an electronic device, such as a robot, with input and output units corresponding to human sensory systems. (*Id.* ¶¶[0020]-[0022].) Inputs include “a pair of CCD (Charge Coupled Device) cameras 20” (eyes) “a pair of microphones 21” (ears). (*Id.* ¶[0021].) Outputs include “a speaker 23” (mouth). (*Id.* ¶[0022].) Figure 1 shows an example, robot 1:

[FIG. 1]

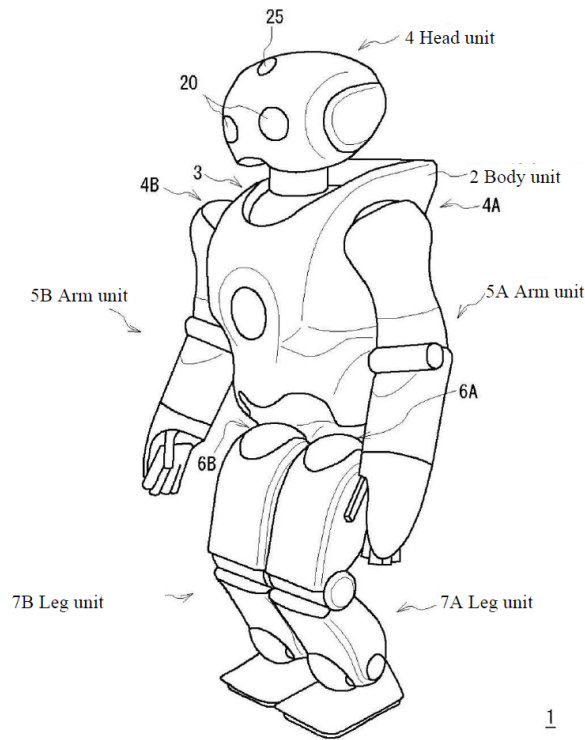


FIG. 1 Hardware constitution of the robot according to this embodiment (1)

(*Id.*, Fig. 1.)

131. The device receives user speech via microphones, performs speech recognition, and determines the “content to be spoken by the robot.” (Ex. 1004 (Shimomura) ¶¶[0046]; *see also id.*, Abstract, Title, ¶¶[0001], [0030].) For example, if a user asks about the weather, the device may respond, “Today’s weather in Tokyo is sunny.” (*Id.* ¶¶[0065].) Shimomura’s Fig. 4 provides an overview of the system and its components:

[FIG. 4]

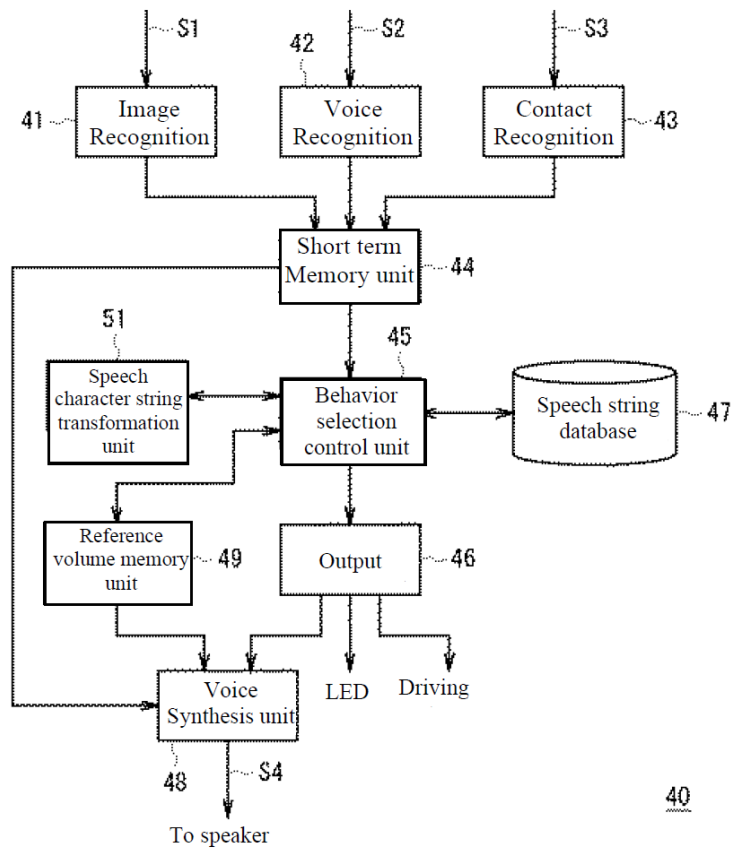


FIG. 4 Constitution of a behavior control system in a robot in accordance with the present embodiment

(*Id.*, Fig. 4.)

132. The device includes “a speech control function that controls speech forms such as speech volume, speech speed, intonation, and intervals between phrases according to the physical distance to the conversation partner.” (Ex. 1004 (Shimomura) ¶¶[0049]-[0051]; *see also id.* ¶¶[0009]-[0010], [0016], [0057], [0085], [0088], [0090], [0092], claims 1-24.) Shimomura’s device may be an “entertainment

robot[]” or “other device[] ... for other applications having interactive functions.”

(*Id.* ¶[0096].) Thus, Shimomura discloses a “voice-content control device.”

b. 1[a]: Proximity Sensor

133. Claim element 1[a] recites “a proximity sensor configured to calculate a distance between a user and the voice-content control device.” Shimomura discloses this claim element.

134. Shimomura discloses that its device includes a “distance detection means for detecting the distance to the conversation partner user[.]” (Ex. 1004 (Shimomura) ¶[0093].) The distance detection means may include “a pair of CCD cameras 20 ... that detects the distance by a stereo vision method based on the output of the cameras” or a “distance sensor[.]” (*Id.*; *id.* ¶¶[0039] (the “image recognition unit” detects a distance to the user “by means of a so-called stereo visual method based on the image signal S1 from the CCD camera [sic] 20”), [0059] (“distance information indicating the distance to the conversation partner detected by the image recognition unit 41” is sent to speech character string transformation unit), [0070] (describing calculated distances), claims 1 (“a distance detection means for detecting the distance between the device and the user”), 9, 17.)

135. Thus, Shimomura discloses a proximity sensor (image recognition unit, distance detection means, or distance sensor) configured to calculate a distance between a user and the voice-content control device.

c. 1[b]: Voice Classifying Unit

136. Claim element 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.” Shimomura discloses this claim element.

137. First, Shimomura discloses a voice classifying unit that is configured to “analyze a voice spoken by a user and acquired by a voice acquiring unit” as recited. Shimomura discloses that its device, robot 1 (orange), includes a control unit 10 (blue) and a microphone 21 (red):

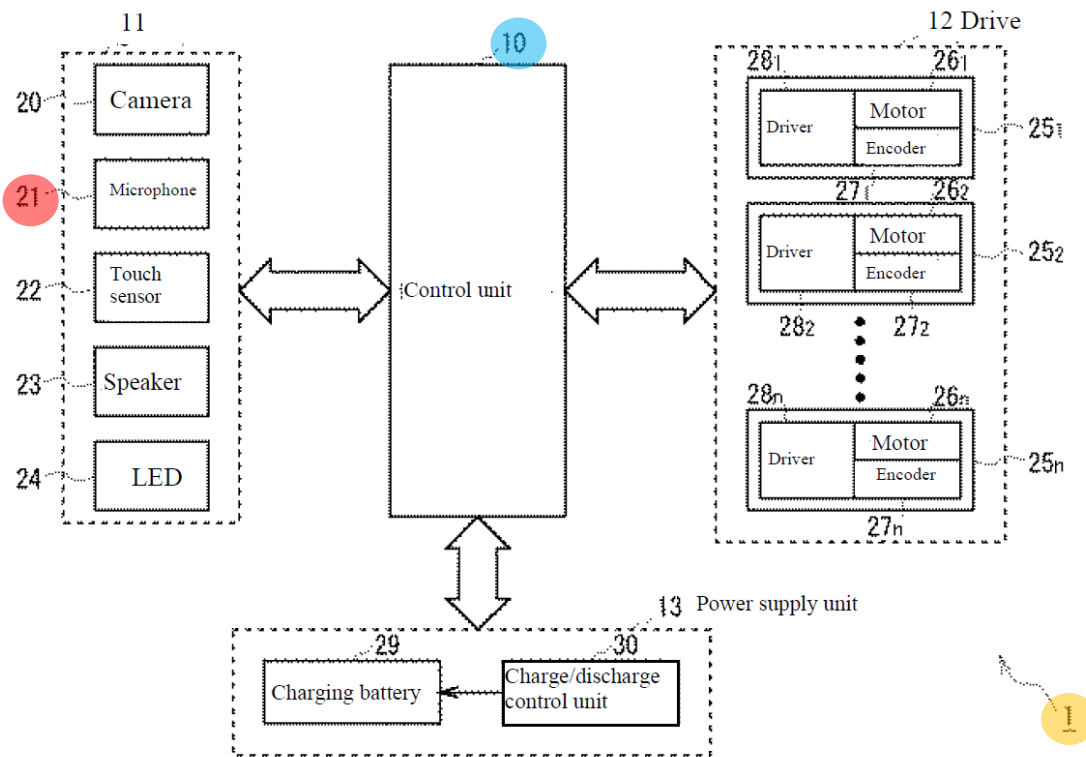


FIG. 2 Hardware constitution of the robot according to this embodiment (2)

(Ex. 1004 (Shimomura), Fig. 2.) The control unit 10, which includes a CPU, controls overall operation and other data processing. (*Id.* ¶¶[0020], [0027], Fig. 3; *see also id.* ¶¶[0028], [0032], [0034].) The control unit implements a “behavior control system” that includes a “speech recognition unit.” (*Id.* ¶¶[0036]-[0038], [0042], Fig. 4.) “The speech recognition unit 42 performs various sound-related recognition processes such as speech recognition, speaker recognition, and sound source direction recognition based on the speech signal S2 provided by the microphone 21.” (*Id.* ¶[0040].) Thus, Shimomura discloses a voice classifying unit (control unit) that is configured to analyze a voice spoken by a user (user’s speech) and acquired by a

voice acquiring unit (speech recognition unit and/or portion of the processor that receives voice from the microphone).

138. Second, Shimomura discloses that the voice classifying unit (control unit) analyzes the voice to classify it as either a first or second voice based on the distance between the user and the device. Shimomura's control unit implements the behavior control system, which also includes a "speech synthesis unit" and a "speech character string transformation unit." (Ex. 1004 (Shimomura) ¶¶[0036]-[0037], [0046], [0058], [0067], Fig. 4.) The speech synthesis unit and the speech character string transformation unit may "transform the character string" of the content spoken by the robot or change its "volume," "speed," "pause length," or "intonation" "according to the distance to the conversation partner." (*Id.* ¶¶[0058], [0067]-[0070]; *see also id.* ¶¶[0060]-[0063], [0073]-[0084], Abstract, claims 1-24.) To do so, Shimomura's control unit first classifies the voice.

139. For example, Shimomura discloses classifying voice into two categories based on the user-to-device distance: (1) a first voice when the distance is greater than 350 cm, and (2) a second voice when the distance is less than 350 cm. These classifications are shown in the character string transformation determination table of Figure 7:

[FIG. 7]

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

FIG. 7 Character string transformation determination table

(Ex. 1004 (Shimomura), Fig. 7, ¶¶[0060], [0061] (determining “whether or not to transform the character string based on ... the distance to the conversation partner recognized based on the distance information from the action selection control unit 45 and the character string transformation determination table”), [0062]-[0064].)

140. Thus, Shimomura discloses the voice classifying unit (processor) is configured to analyze the voice to classify the voice as either one of a first voice (e.g., voice farther than 350 cm) or a second voice (e.g., voice closer than 350 cm) based on the distance between the user and the voice-content control device.

d. 1[c]: Process Executing Unit

141. Claim element 1[c] recites “a process executing unit configured to analyze the voice acquired by the voice acquiring unit to execute processing required by the user.” Shimomura discloses this claim element.

142. Shimomura’s control unit is configured to analyze the voice acquired by the voice acquiring unit. (Ex. 1004 (Shimomura) ¶[0040].) As I discuss above,

It recognizes speech from the microphone's signal and, based on that speech, determines the content to be spoken to the user. (*See* ¶¶136-140, above.) Shimomura describes the user as a conversation partner of the device, confirming the device extracts user intent during conversation and responds by performing the required or intended processing. (Ex. 1004 (Shimomura) ¶¶[0050], [0052], [0057]-[0059], [0067], [0086]-[0087], [0090], claims 1 (“[a] speech control device that controls the speech of a device having a dialogue function with a user during dialogue between the device and the user”), 9, 17; *see also id.* ¶¶[0003]-[0004].)

143. Indeed, Shimomura discloses analyzing voice “to execute processing required by the user” just like the '337 patent. Shimomura describes analyzing a user's request (e.g., for weather) and obtaining the requested information (“Today's weather in Tokyo is sunny”). (Ex. 1004 (Shimomura) ¶[0065].) In another example, Shimomura describes analyzing a user's voice to execute processing (volume change). (*Id.* ¶¶[0053]-[0055].)

144. Thus, Shimomura discloses a process executing unit (control unit) configured to analyze the voice acquired by the voice acquiring unit to execute processing required by the user. To the extent this claim element is not expressly disclosed by Shimomura, it would have been inherent because, in order to converse with a user, as repeatedly disclosed by Shimomura, the control unit must necessarily analyze the acquired voice input to formulate a relevant response.

e. 1[d]: Voice-Content Generating Unit

145. Claim element 1[d] recites “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.” Shimomura discloses this claim element.

146. Shimomura discloses a voice-content generating unit that generates an output sentence as text data for a voice for output to the user. As I discussed above for claim elements 1[b]-1[c], Shimomura’s control unit generates the content for output to the user. (See ¶¶136-144, above.) The content can be a sentence including text data like: “Today’s weather in Tokyo is sunny.” (Ex. 1004 (Shimomura) ¶[0065]; *see also id.* ¶¶[0076] (speech synthesis unit determines how long to pause between sentences), [0078] (speech synthesis unit determines how to intonate within sentences), [0079], [0092].) The control unit implements the behavior control system with an “output management unit” and a “behavior selection control unit.” (*Id.* ¶¶[0036]-[0037], [0046]-[0047], Fig. 4.) The output management unit “sends the character string provided by the behavior selection control unit 45 to the speech synthesis unit 48, while the speech synthesis unit 48 generates a speech signal S4 of synthetic speech based on the supplied character string and sends it to the speaker 23” for audible presentation. (*Id.* ¶[0047]; *see also id.* ¶¶[0046], [0049]-[0050],

[0063], [0083], Fig. 13.) Shimomura’s claims provide further clarity on this sequence. For example, claim 6 states that the device comprises “a character string output means for outputting a character string corresponding to a content to be spoken by the device; a speech means for generating a speech signal of synthetic voice corresponding to the character string; and a speaker for outputting audio.” (*Id.*, claim 6; *see also id.*, claim 14.) Thus, Shimomura discloses a voice-content generating unit (control unit) configured to generate an output sentence that is text data for a voice to be output to the user.

147. Shimomura also discloses the voice-content generating unit generates the output sentence based on content of the processing executed by the process executing unit. As I discussed above for claim element 1[c], Shimomura’s control unit determines the content spoken to the user based on the recognized speech so that the device converses with the user. (*See* ¶¶141-144, above.) Thus, Shimomura discloses a voice-content generating unit (control unit) configured to generate, based on content of the processing executed by the process executing unit (control unit), output sentence that is text data for a voice to be output to the user.

f. 1[e]: Output Controller

148. Claim element 1[e] recites “an output controller configured to adjust a sound volume of voice data obtained by converting the output sentence thereinto.” Shimomura discloses this claim element.

149. As I discuss above for claim element 1[d], Shimomura's control unit generates the speech signal and sends it to a speaker. (See ¶¶145-147, above.) Shimomura's control unit also implements the speech synthesis unit to "store[] in the ROM 33 ... a volume change amount specification table 52 that specifies the amount of change in the speech volume of the robot 1 with respect to the distance to the conversation partner." (Ex. 1004 (Shimomura) ¶[0067]; *see also id.* ¶¶[0050] ("The robot 1 is equipped with a speech control function that controls speech forms such as speech volume"), [0068] ("the voice synthesis unit 48 changes the voice volume"), [0069] ("the speech synthesis unit ...refers to the volume change amount specification table 52 [] and changes the parameter value of the volume parameter defining the speech volume"), [0070]-[0073], [0075], [0083], [0084] ("controlling the speech type such as speech volume ... depending on the physical distance to the conversation partner"), [0090], [0092].) Furthermore, Shimomura's claim 2 states that the device "changes the speech volume" according to the "distance between the device and the user." (*Id.*, claim 2; *see also id.*, claims 10, 18.) Likewise, in Shimomura's Figure 9, a "Volume Change Table" illustrates a set of volume change parameters based on distance between the device and the user:

[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

(*Id.*, Fig. 9; *see also id.* ¶[0067], Fig. 13.)

150. As shown in Figure 9, Shimomura’s speech synthesis unit “determines the reference volume read from the reference volume memory unit 49 as the current speech volume, and when the distance to the conversation partner is in the range of 50 to 80 cm, determines a volume that is a predetermined specified value (for example, ‘1’) higher than the reference volume as the current speech volume, and similarly thereafter, when the distance to the user is 80 to 150 cm, 150 to 250 cm, or 250 cm or more, determines a volume that is ‘2’, ‘3’, or ‘5’ higher than the reference volume as the current speech volume.” (Ex. 1004 (Shimomura) ¶[0070].) Thus, Shimomura discloses an output controller configured to adjust a sound volume of voice data obtained by converting the output sentence thereinto.

g. 1[f]: Generate a First Output Sentence

151. Claim element 1[f] recites that “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[.]” Shimomura discloses this claim element.

152. Shimomura's control unit implements the speech character string transformation unit to insert a word like "ne" after each phrase in the content to "chang[e] the interval between phrases so that the division of meaning can be clearly understood according to the physical distance to the conversation partner." (Ex. 1004 (Shimomura) ¶¶[0057]-[0058]; *see also id.* ¶¶[0059], [0060] ("When the speech character string transformation unit transformation unit 51 is given character string and distance information from the behavior selection control unit 45, it judges whether or not to transform the character string based on ... based on the distance information and the character string transformation determination table 50"), [0061]-[0066], [0085], [0092] ("It goes without saying that words other than 'ne' can be used as words to be inserted after each phrase end to create a gap between phrases."), Figs. 7-8.) When the voice is classified as the first voice (e.g., voice farther than 350 cm), the device generates a first output comprising the original content after being transformed to appropriately insert "ne" or another word after phrases. (*Id.* ¶¶[0062], [0065], [0092].) Figure 7 shows when "implementation" of the transformation (e.g., adding "ne") is performed:

[FIG. 7]

Less than 350 cm	350 cm or more
None	Implementation

First Voice

FIG. 7 Character string transformation determination table

(*Id.*, Fig. 7.) Figure 8 illustrates this process:

[FIG. 8]

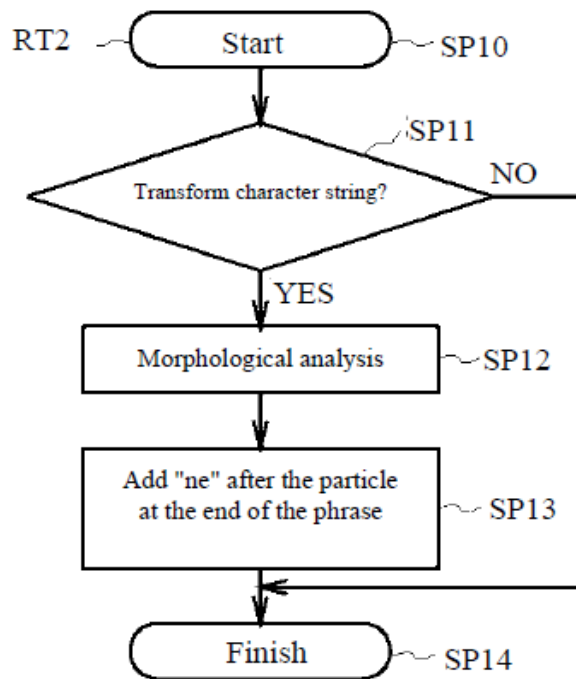


FIG. 8 Character string transformation processing procedure

(*Id.*, Fig. 8.) As shown, the control unit implements the transformation and outputs transformed content when the user is greater than 350 cm from the device. (*Id.* ¶[0062].)

153. Shimomura discloses an example of generating transformed content, and thus the output sentence when the voice is classified as the first voice (e.g., voice farther than 350 cm): “generate a character string such as ‘Today’s, Tokyo’s, the weather is, sunny’ by adding the character ‘ne’ to the end of each clause-final auxiliary verb, ‘Today’s,’ ‘Tokyo’s,’ and ‘The weather is.’” (Ex. 1004 (Shimomura) ¶[0065].)

154. Thus, Shimomura discloses the voice-content generating unit is further configured to generate a first output sentence (transformed content) as the output sentence when the acquired voice has been classified as the first voice (e.g., voice farther than 350 cm).

h. 1[g]: Generate a Second Output Sentence

155. Claim element 1[g] recites that the voice-content generating unit is further configured to “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.” Shimomura discloses this claim element.

156. The '337 patent explains that omitted information may include “a polite Japanese expression such as ‘densu’,” “a Japanese postpositional particle ‘ha’,” or in English, “an article or the like.” (Ex. 1001 ('337 patent), 9:31-35.) This is the same type of information omitted by Shimomura’s device.

157. Shimomura discloses that, whereas a voice classified as a first voice (e.g., farther than 350 cm) results in a transformed response containing “ne” or another word added to “the end of each clause-final auxiliary verb,” the speech character string transformation unit “judges that the character string should not be transformed if the distance to the conversation partner recognized by the image recognition unit 41 is less than 350 [cm].” (Ex. 1004 (Shimomura) ¶¶[0062], [0065], [0092]; *see also id.* ¶¶[0057]-[0058], [0085] (“depending on the distance to the user, it ... inserts the word ‘ne’ after each phrase.”).) Thus, when the voice is classified as a second voice (e.g., closer than 350 cm), the device generates a second output sentence comprising the original content without the transformation to add “ne” or another “word[] to be inserted after each phrase end to create a gap between phrases.” (*Id.* ¶¶[0062], [0092].) This is also shown in Shimomura’s Figure 7:

[FIG. 7] **Second Voice** First Voice

Less than 350 cm	350 cm or more
None	Implementation

FIG. 7 Character string transformation determination table

(*Id.*, Fig. 7.) As shown, the control unit outputs the original content *without* transformation when the user is closer than 350 cm. (*Id.* ¶[0062].) In other words, when the acquired voice is classified as a second voice (e.g., closer than 350 cm), the output content lacks the added words, i.e., it omits some information provided when the voice is classified as a first voice.

158. Shimomura discloses an example in which the original content is not transformed, and thus the output sentence when the voice is classified as the second voice (e.g., closer than 350 cm) is: “Today’s weather in Tokyo is sunny.” (Ex. 1004 (Shimomura) ¶[0065].) Thus, Shimomura discloses omitting some information (i.e., “ne” after phrases) when the voice is classified as the second voice. (*Id.*)

159. Thus, Shimomura discloses the voice-content generating unit is further configured to generate a second output sentence (original content) in which information is omitted (e.g., “ne”) as compared to the first output sentence (transformed

content) as the output sentence when the acquired voice has been classified as the second voice (e.g., voice closer than 350 cm).

i. 1[h]: Output Controller Adjusts Volume of Voice Data

160. Claim element 1[h] recites that “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.” Shimomura renders this claim element obvious.

161. As I discuss above for claim element 1[e], Shimomura’s control unit adjusts the output volume of the content based on the user-to-device distance. (See ¶¶148-150 above; Ex. 1004 (Shimomura) ¶[0067].) Specifically, it implements the speech synthesis unit to “determine[] the reference volume read from the reference volume memory unit 49 as the current speech volume, and when the distance to the conversation partner is in the range of 50 to 80 cm, determines a volume that is a predetermined specified value (for example, ‘1’) higher than the reference volume as the current speech volume, and similarly thereafter, when the distance to the user is 80 to 150 cm, 150 to 250 cm, or 250 cm or more, determines a volume that is ‘2’, ‘3’, or ‘5’ higher than the reference volume as the current speech volume.” (Ex. 1004 (Shimomura) ¶[0070].) Figure 9 shows this in the volume change table:

[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

(*Id.*, Fig. 9.)

162. As I discuss above, Shimomura discloses generating a first output sentence when the distance is more than 350 cm and a second output sentence when the distance is less than 350 cm. (See ¶¶151-159, above.) Although Shimomura discloses altering the volume based on distance, it does not specify that the first and second output sentences have different volumes. (*Id.*) However, this modification would have been obvious to a person of ordinary skill in the art and trivial to implement.

163. A person of ordinary skill in the art would have been motivated to modify Shimomura's volume change table to include a range ending at 350 cm. This would, for example, have been implemented by extending the 150-250 cm range to end at 350 cm, as shown in the highlighted column:

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

Alternatively, this would have been implemented by adding a new range from 250 cm to 350 cm, as shown in the highlighted columns:

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

One of ordinary skill in the art would have been motivated to make such a modification for several reasons.

164. First, a modification such as the one shown in the first modified table would have been the simple substitution of a known element (350 cm threshold as in Shimomura's Figures 7 and 10-11) for another known element (250 cm threshold as in Shimomura's Figure 9) to obtain predictable results (a device that adjusts output volume based on whether user is closer or farther than 350 cm).

165. Second, such a modification reflects nothing more than a combination of known elements (different distance thresholds, including one at 350 cm as in Shimomura's Figures 7 and 10-11, for controlling output volume based on distance from user) to achieve an expected result (adjusting output volume based on user's distance up to 350 cm to ensure appropriate audibility at distances around 350 cm).

166. Third, such a modification reflects nothing more than the application of a known technique (adjusting output volume based on whether user is closer or farther than 350 cm as in Shimomura's Figures 7 and 10-11) to achieve predictable

results (adjusting output volume so that a user closer or farther than 350 cm from Shimomura's device can clearly hear content).

167. Fourth, such a modification reflects nothing more than applying a known technique (adjusting output volume based on whether user is closer or farther than 350 cm as in Shimomura's Figures 7 and 10-11) to a known device and method (Shimomura's control unit that varies output volume with user distance) that is ready for improvement and yields predictable results (a device that varies output volume based on whether user is closer or farther than 350 cm to provide appropriate volume for users closer or farther than 350 cm).

168. Fifth, such a modification would be a matter of routine experimentation and common sense for a person of ordinary skill in the art. Given Shimomura's goal of enhancing user interactions by adjusting volume with distance (Ex. 1004 ¶¶[0005], [0010]-[0011], [0016]-[0017], Abstract), a person of ordinary skill in the art would have logically modified the threshold (as in the first modified table) or added a threshold (as in the second modified table) to ensure appropriate audibility around the 350 cm threshold indicated in Shimomura's Figures 7 and 10-11. This simple modification would yield predictable improvements without requiring significant innovation.

169. Sixth, such a modification would have been obvious to try. In environments where Shimomura's device operates at varied distances, users may interact

from beyond 250 cm. Extending the “More than 250 cm” range to 350 cm, as in the first modified table, addresses this without adding discrete categories, ensuring predictable responses up to 350 cm—a logical limit aligned with Shimomura’s Figures 7 and 10–11. This approach simplifies design by minimizing programming complexity. Alternatively, adding a range from 250 cm to 350 cm, as in the second modified table, is an incremental adjustment continuing the trend of volume increases with distance. Shimomura’s existing incremental volume changes make this a predictable solution for ensuring audibility at greater distances.

170. A person of ordinary skill in the art would have reasonably expected success in modifying Shimomura’s volume change table in this way. This trivial modification would have involved configuring Shimomura’s control unit with the revised table to provide volume levels based on whether the user is closer or farther than 350 cm.

171. Considering the first modified volume change table and Figure 7’s character string transformation determination table together, it would have been obvious to a person of ordinary skill in the art from Shimomura that, when a voice is classified as a first voice (e.g., voice farther than 350 cm), a first output sentence containing the original content after transformation is provided (character string transformation determination table) at a volume level of +5 higher than the reference volume (first modified volume change table) and, when a voice is classified as a

second voice (e.g., voice closer than 350 cm, such as 300 cm), a second output sentence containing original content without transformation may be provided (character string transformation determination table) at a volume level of +3 higher than the reference volume (first 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

[FIG. 7]

Less than 350 cm	350 cm or more
None	Implementation

FIG. 7 Character string transformation determination table

(Ex. 1004 (Shimomura), Fig. 7.) Similarly, considering the second modified volume change table and Figure 7's character string transformation determination table together, it would have been obvious to a person of ordinary skill in the art from Shimomura that, when a voice is classified as a first voice (e.g., voice farther than 350 cm), a first output sentence containing the original content after transformation is provided (character string transformation determination table) at a volume level

of +7 higher than the reference volume (second modified volume change table) and, when a voice is classified as a second voice (e.g., voice closer than 350 cm, such as 300 cm), a second output sentence containing original content without transformation may be provided (character string transformation determination table) at a volume level of +5 higher than the reference volume (second modified volume change table).

				Second Voice	First Voice
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

[FIG. 7]

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

FIG. 7 Character string transformation determination table

(*Id.*) Shimomura expressly discloses that both the content and volume may vary.

(*Id.* ¶¶[0085]) (“[T]he robot 1 speaks louder, more slowly, and with larger gaps between phrases the farther away the user/conversation partner is, while depending on the distance to the user, it speaks with an intonation that emphasizes the end of a

phrase and inserts the word ‘ne’ after each phrase.”), [0083] (“After the speech synthesis unit 48 has finished setting various parameters such as the volume parameter ..., it proceeds to step SP30, generates a speech waveform according to the character string based on the given character string information, and transforms this speech waveform according to the volume parameter ... and phrase-end emphasis parameter set.”); *see also id.* ¶¶[0084], [0092], Fig. 4, claims 6, 14, 21.)

172. Thus, it would have been obvious to a person of ordinary skill in the art to modify Shimomura’s output controller (control unit) to be configured to adjust the sound volume of voice data such that the sound volume of voice data obtained by converting the first output sentence (transformed content) thereinto differs from the sound volume of the voice data obtained by converting the second output sentence (original content) thereinto.

173. Accordingly, for at least the reasons I discuss above, Shimomura renders obvious claim 1.

2. Claim 2

174. Claim 2 depends from claim 1 and recites two additional claim elements, which Shimomura discloses. Moreover, these claim elements reflect the operation of known voice processing systems like Alexa and Siri. (*See* ¶35, above.)

a. 2[a]: Process Executing Unit

175. Claim element 2[a] recites that the “process executing unit” comprises: [i] “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 [ii] an acquisition content information acquiring unit configured to acquire acquisition content information which is notified to the user based on the extracted intention information.” Shimomura discloses and/or renders obvious this claim element.

176. As I discuss above for claim element 1[c], Shimomura’s control unit determines the content spoken to the user from the recognized speech by extracting and responding to a user’s intention. (See ¶¶141-144, above.) Shimomura expressly discloses that the device controls its speech “during a dialogue between the device and a user,” enabling the device to “converse with the user smoothly.” (Ex. 1004 (Shimomura) ¶¶[0010], [0086].) Shimomura’s control unit acquires from memory the content such as “Today’s weather in Tokyo is sunny” for presenting to the user. (*Id.* ¶¶[0044], [0065].)

177. Even if Shimomura did not expressly disclose this element, Shimomura’s control unit inherently includes an “intention analyzing unit” and an “acquisition content information acquiring unit,” as recited in claim element 2[a], because these would have been necessary for Shimomura’s device to function as described. These functionalities are essential for interpreting user input, determining

intent, and generating appropriate responses. For example, providing weather updates requires interpreting user intent and retrieving relevant content. A person of ordinary skill in the art would have understood that extracting semantic and contextual speech information corresponds to the “intention analyzing unit,” while retrieving relevant content discloses the “acquisition content information acquiring unit.” These units are inherently required for Shimomura’s control unit to perform as disclosed.

178. Thus, Shimomura expressly or inherently discloses that the process executing unit (control unit) comprises: 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. Shimomura also expressly or inherently discloses that the process executing unit (control unit) comprises an acquisition content information acquiring unit configured to acquire acquisition content information which is notified to the user based on the extracted intention information.

179. Even if Shimomura had not disclosed this claim element (expressly or inherently), the claim element would have been obvious to a person of ordinary skill in the art from Shimomura. Shimomura’s device engages in dialogue with the user, and a person of ordinary skill in the art would have been motivated to use an “intention analyzing unit” to determine user intent and an “acquisition content information acquiring unit” to retrieve content based on that intent.

180. Shimomura describes voice processing capabilities that allow the device to interpret user intent and generate appropriate responses. (See ¶¶141-144, above.) A person of ordinary skill in the art would have been motivated to structure Shimomura's control unit with units, including an intention analyzing unit and a content acquisition unit, to perform these functions more efficiently. This modular design would facilitate consistent intent extraction, reliable content retrieval, and improved clarity and organization of internal processes. Furthermore, as I discuss elsewhere, this approach would also merely reflect the way known voice analysis systems (such as Apple's Siri and Amazon's Alexa) worked at the time of the invention. (See ¶¶35, 118, above; ¶¶216-223, below.)

181. This approach would have been straightforward for a person of ordinary skill in the art, as Shimomura's device already integrates voice analysis and content generation. Implementing these functionalities as units within Shimomura's control unit would have been an obvious design choice, enhancing modularity, maintainability, and troubleshooting. This design would predictably improve response accuracy, while offering greater flexibility to handle various user requests.

182. Thus, this claim element would have been obvious over Shimomura. Moreover, as I discuss above, even if Shimomura had not disclosed or suggested this claim element, the '337 patent admits this claim element was known in the art. (See ¶35, above.)

b. 2[b]: Voice-Content Generating Unit

183. Claim element 2[b] recites that “the voice-content generating unit is further configured to generate the text data including the acquisition content information as the output sentence.” Shimomura discloses this claim element.

184. As I discuss above for claim element 2[a], Shimomura’s control unit acquires information, such as weather information, to determine the content spoken to the user. (See ¶¶175-182, above; Ex. 1004 (Shimomura) ¶¶[0010], [0065], [0086].) This information is incorporated into the voice content provided to the user, as demonstrated by the example I discuss for claim element 2[a]. (See ¶176, above.) The control unit’s output management unit “sends the character string provided by the behavior selection control unit 45 to the speech synthesis unit 48, while the speech synthesis unit 48 generates a speech signal S4 of synthetic speech based on the supplied character string and sends it to the speaker 23.” (Ex. 1004 (Shimomura) ¶[0047].) Thus, Shimomura discloses that the voice-content generating unit (control unit) is configured to generate the text data including the acquisition content information as the output sentence.

185. Accordingly, for at least the reasons I discuss above, Shimomura renders obvious claim 2.

3. Claim 3

186. Claim 3 depends from claim 1 and further recites “wherein, on generating the second [output] sentence, the voice-content generating unit is further configured to omit a part of information included in the voice spoken by the user.” Shimomura discloses and/or renders obvious such a voice-content generating unit.

187. As I discuss above for claim elements 1[c] and 1[g], Shimomura’s control unit generates content, such as “Today’s weather in Tokyo is sunny,” and can do so while conversing with the user. (See ¶¶141-144, 155-159, above.) A person of ordinary skill in the art would have understood that this content could be prompted by specific user speech, such as by asking, “what is the weather today?” The interrogative word (“what”) prompts processing, and the device omits it in the response to provide a direct answer rather than echoing the question.

188. A person of ordinary skill in the art also would have been motivated for several reasons to configure Shimomura’s control unit to omit some user-spoken information when responding.

189. First, this would have been a predictable and desirable implementation of Shimomura’s control unit, which processes and responds to user inputs. A person of ordinary skill in the art would have understood that repeating all user-spoken information is unnecessary, as the user already knows it, and excessive repetition would be inefficient and delay responses.

190. Second, market demands and user expectations for efficient, streamlined interactions with automated systems would have motivated this implementation. As I discuss above, users expect minimal delays and concise responses. (*See* ¶123, above.) Omitting user-spoken information would align Shimomura's control unit with these expectations, improving user satisfaction. A person of ordinary skill in the art would have understood this as an obvious solution to meet these demands.

191. Third, such an implementation would have been a matter of routine experimentation and common sense for a person of ordinary skill. Given Shimomura's aims to enhance user interactions (Ex. 1004 (Shimomura) ¶¶[0010]-[0011], [0016]-[0017], Abstract), reducing unnecessary repetition to streamline responses would have been a logical step. A person of ordinary skill in the art would have recognized this simple, routine modification (if even a modification) would yield predictable improvements without requiring significant innovation.

192. Fourth, such an implementation would have been obvious to try. Shimomura's device presents only two design options for responding to a user request: include (1) all information from a user's request or (2) less than all. Omitting some information to enhance efficiency and conciseness would have been an obvious approach.

193. A person of ordinary skill in the art would have reasonably expected success in implementing Shimomura's control unit to omit some user-spoken information when responding. This implementation would have been trivial, involving straightforward changes to configure the control unit for more efficient and streamlined interactions.

194. Thus, Shimomura renders obvious that, on generating the second output sentence, the voice-content generating unit (control unit) is further configured to omit a part of information included in the voice spoken by the user.

195. Accordingly, for at least the reasons I discuss above, Shimomura renders obvious claim 3.

4. Claims 4 and 5

196. Shimomura discloses both a speech control device and method. (Ex. 1004 (Shimomura) ¶¶[0012]-[0013]; *see also id.* ¶¶[0001], [0009], [0016], [0057], Title, Abstract, Figs. 6, 8, 13, claims 9-16.) Shimomura also discloses electronic devices implemented by instructions stored in a computer-readable storage media. (Ex. 1004 (Shimomura) ¶¶[0036] (“FIG. 4 schematically illustrates a functional constitution of a behavior control system 40 of a robot 1 configured by a group of control programs stored in a ROM.”), [0027], [0030]; *see also id.* ¶¶[0028]-[0029], [0037]-[0047], [0051]-[0083], Figs. 3-4.) Thus, Shimomura discloses the preamble of claims 4 and 5.

197. As I discuss above, the steps of claims 4 and 5 are identical and merely recite the functions performed by the various units recited in claim 1. (*See* ¶¶104-107, above.) Accordingly, for the reasons I discuss for claim 1, Shimomura renders obvious claims 4 and 5. (*See* ¶¶130-173, above.)

D. Claims 1-5 Would Have Been Obvious in View of Shimomura and Shin.

198. As I have discussed above, Shimomura discloses and/or renders obvious each limitation of claims 1-5 and thus renders the claims obvious. However, even if Shimomura failed to render obvious any particular limitation of claims 1-5, such a limitation would have been obvious to a person of ordinary skill in the art further in view of Shin.

1. Claim 1

199. Even if Shimomura alone did not teach or motivate a person of ordinary skill in the art to alter the output's content and volume at the same threshold distance (e.g., 350 cm), doing so would have been obvious in view of Shin's teachings.

200. Both Shimomura and Shin teach proximity-based voice output adjustments for improving user interaction. (*See* ¶¶51-90, 130-173, above.) Shimomura adjusts content and volume based on proximity, while Shin complements this with content classification and dynamic output. (*Id.*) A person of ordinary skill in the art would have naturally combined these teachings to enhance adaptability and functionality in voice-controlled devices with predictable results.

201. Shin’s volume change table (Table 1) and amount of information provided table (Table 3) include matching distance ranges, correlating specific distances with volume levels and information amounts. (Ex. 1003 (Shin), Tables 1, 3.) For example, Shin could use Tables 1 and 3 to classify the user’s voice into a first category (e.g., first voice) based on a user distance of less than one meter, and a second category (e.g., second voice) based on a user distance between one to two meters:

TABLE 1		
	Distance (D) between user and electronic device	Output volume level of content
First Voice	$D < 1 \text{ m}$	40 dB
Second Voice	$1 \text{ m} \leq D < 2 \text{ m}$	45 dB
	...	...
	$D \geq 5 \text{ m}$	65 dB

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

(*Id.*, Tables 1, 3.) This would have motivated a person of ordinary skill in the art and provided a reasonable expectation of success in modifying Shimomura’s volume change table—including in the changes I discuss above (*see* ¶¶160-172, above)—to implement common distance thresholds for adjusting volume level and content. This modification would enable Shimomura’s device to classify a user’s 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.

202. Shin discloses that the distances and volume levels in its volume change table (Table 1), similar to Shimomura's, are "merely provided as an example" and that embodiments "are not limited to the example of Table 1." (Ex. 1003 (Shin) ¶¶0081]; *see also id.* ¶¶0090] ("[E]mbodiments ... are not limited to the examples of Table 1, Table 2, and Table 3.")). This would have motivated a person of ordinary skill in the art and provided a reasonable expectation of success in implementing alternative distances or volume levels in Shimomura's volume change table, such as those I discuss above, to achieve desired volume levels for specific user distances. (*See* ¶¶160-172, above.)

203. Moreover, Shin discloses that information may be classified "dichotomously" (i.e., two categories) or divided "into several levels," ranging from the most to the least detailed format. (Ex. 1003 (Shin) ¶¶0051].) A person of ordinary skill in the art would have been motivated to use two different volume levels for the dichotomous embodiment, with one corresponding to each category of content, aligning with Shin's disclosure, simplifying design, and being trivial to implement.

204. Accordingly, for at least the reasons I discuss above, Shimomura and Shin render obvious claim 1.

2. Claim 2

205. Claim 2 would have been obvious over Shimomura in view of Shin. Specifically, as I discuss above, Shin discloses that its processor analyzes the requests from the voice input to identify the content the user is interested in, and then acquires and provides that content. (See ¶¶67-70, 92-96, above.) Shimomura similarly describes this functionality. (See ¶¶141-144, 175-182, above.) A person of ordinary skill in the art would have been motivated by Shin’s disclosure to implement an “intention analyzing unit” to extract intention information from user voice data and an “acquisition content information acquiring unit” to retrieve content based on the extracted intention, in Shimomura’s device. Shin’s voice-processing capabilities, such as interpreting user intent for requests (e.g., time or weather), would have motivated structuring Shimomura’s processor with these units for efficient intent extraction and content retrieval. (See ¶¶118, 180, above.) This modular design would enhance process clarity, reliability, and efficiency. Accordingly, for at least the reasons I discuss above, Shimomura and Shin render obvious claim 2.

3. Claim 3

206. Even if Shimomura alone had not suggested the additional elements of claim 3, Shin discloses those elements, making claim 3 as a whole obvious over Shimomura in view of Shin. A person of ordinary skill in the art would have been motivated, for several reasons, to implement Shin’s feature of omitting some user-

spoken information into Shimomura’s control unit to achieve Shimomura’s stated goal of allowing “smooth interaction with the user.” (Ex. 1004 (Shimomura), Abstract.)

207. First, such an implementation would represent merely the simple substitution of a known element (a response that omits user-spoken information, as in Shin) for another known element (Shimomura’s response) to obtain predictable results (more efficient and concise user interaction with Shimomura’s device).

208. Second, such an implementation reflects merely a combination of known elements (Shimomura’s control unit processing and responding to user inputs and Shin’s output that omits some user-spoken information) to achieve an expected result (improving efficiency and streamlining user interactions with Shimomura’s device).

209. Third, such an implementation reflects nothing more than the application of a known technique (Shin’s omission of some user-spoken information in its output) to achieve predictable results (reducing unnecessary information to improve efficiency of communication with Shimomura’s device).

210. Fourth, such an implementation reflects nothing more than applying a known technique (Shin’s output that omits some user-spoken information) to a known device and method (Shimomura’s control unit and its method of responding

to user requests) that is ready for improvement and yields predictable results (more efficient and streamlined interaction between a user and Shimomura's device).

211. Fifth, such an implementation would be a straightforward optimization of Shimomura's control unit. A person of ordinary skill in the art would have recognized Shin's goal of enhancing user interaction efficiency by eliminating redundant information. Applying this to Shimomura's control unit would involve minimal modification (if any), yielding predictable benefits like faster communication and more concise responses.

212. Sixth, such an implementation would have been obvious to try. As I discuss above, Shimomura's device presents only two design options for responding to user requests: include (1) all information from a user's request or (2) less than all. Given Shin's disclosure of omitting information from user requests to enhance efficiency (*see* ¶124, above), it would have been obvious to apply a similar approach with Shimomura's device.

213. A person of ordinary skill in the art would have reasonably expected success in implementing Shin's feature to Shimomura's control unit. This implementation would have been trivial, involving straightforward changes to enhance the control unit's efficiency and streamline user interactions.

214. Accordingly, for at least the reasons I discuss above, Shimomura and Shin render obvious claim 3.

4. Claims 4 and 5

215. Claims 4 and 5 would have been obvious in view of Shimomura and Shin for the reasons I discuss above regarding claim 1. (*See* ¶¶51-90, 104-107, 130-173, 199-204, above.)

E. Claims 1-5 Would Have Been Obvious in View of Shin and/or Shimomura and Further in View of Kristjansson.

216. As I have discussed above, Shin and Shimomura disclose and/or render obvious each limitation of claims 1-5 and thus render the claims obvious. However, even if Shin or Shimomura failed to render obvious the claim limitations relating to the “process executing unit” (e.g., claim limitations 1[c], 2[a], 4[d], and 5[d]), those limitations would have been obvious to a person of ordinary skill in the art further in view of Kristjansson.

217. For example, claim element 2[a] recites that the “process executing unit” comprises: [i] “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 [ii] an acquisition content information acquiring unit configured to acquire acquisition content information which is notified to the user based on the extracted intention information.” This limitation, along with claim limitations 1[c], 4[d], and 5[d], refer to determining the user’s intent using natural language understanding (“NLU”) processing and acquiring the desired information. (Ex. 1001

(’337 patent), 4:31-33.) Kristjansson discloses and/or renders obvious these claim elements.

218. Kristjansson discloses a “speech processing system” that, similar to those disclosed in Shin and Shimomura, comprises “one or more speech-controlled devices 110 local to user 5.” (Ex. 1013 (Kristjansson), 2:45-67.) Kristjansson’s system is “capable of performing traditional speech processing” and “may execute certain commands, such as answering spoken utterances of the users 5[.]” (*Id.*) Kristjansson discloses that to do so, the system includes “an NLU component” that “takes textual input (such as processed from ASR [automatic speech recognition] 450 based on the utterance input audio 11 [from the user]) and ... derive[s] an intent or a desired action from the user[.]” (*Id.*, 7:26-67.) Upon determining the user’s intent, the system sends that information to a “command processor 490,” which retrieves content based on that intention. (*Id.*, 10:38-11:12.) For example, the command processor may retrieve a song requested by the user (*id.*, 10:44-48), results of a user’s search (*id.*, 10:49-56), or content from another source requested by the user (*id.*, 10:56-11:12). In each case, the retrieved content is then delivered to the user. (*Id.*, 10:38-11:12.) Thus, Kristjansson discloses claim limitation 2[a]. For the same reasons, Kristjansson discloses the broader corresponding limitations in claims 1, 4, and 5 (e.g., 1[c], 4[d], and 5[d]).

219. As I discuss above, Shin and Shimomura, like Kristjansson, disclose devices that receive voice input from a user and return voice output based on what the user requests. (See ¶¶92-96, 175-182, above.) A person of ordinary skill in the art would have been motivated to implement Kristjansson's NLU teachings in Shin and/or Shimomura's devices to process the user's input and determine and retrieve the proper output for several reasons.

220. First, Shin and Shimomura suggest doing so. Shin and Shimomura both disclose systems that analyze voice input to identify the content the user is interested in, and then generate the requested content. (See ¶¶92-96, 175-182, above.) Kristjansson's disclosure merely provides additional detail about how a person of ordinary skill would structure such a system, with NLU processing and content retrieval. Thus, a person of ordinary skill would have been motivated to look to Kristjansson when implementing Shin and/or Shimomura's systems.

221. Second, Kristjansson's disclosure represents a routine and well-established approach to providing a user with requested content. (See Ex. 1015, 431-433, Fig. 1.) Third, such a modification reflects merely applying a known technique (Kristjansson's NLU processing) to achieve predictable results (providing a user with requested content). Fourth, such a modification reflects merely applying a known technique (Kristjansson's NLU processing) to a known device (Shin's and/or

Shimomura's) that is ready for improvement and yields predictable results (providing a user with requested content).

222. A person of ordinary skill in the art would have reasonably expected success in integrating Kristjansson's teachings with Shin's and/or Shimomura's at least because Kristjansson's approach was well-known in the art (*e.g.*, Ex. 1015, 431-433, Fig. 1) and because Shin and Shimomura contemplate modification or addition to their systems by persons of ordinary skill in the art (*e.g.*, Ex. 1003 (Shin) ¶[0184]; Ex. 1004 (Shimomura) ¶[0089].) Moreover, Kristjansson discloses a system, like Shin and Shimomura, that uses a proximity sensor to determine the user-to-device distance and adjusts the volume of the output based on that distance. (Ex. 1013 (Kristjansson), claim 1, 3:1-27.)

223. Accordingly, for at least the reasons I discuss above, Shin and/or Shimomura, in further view of Kristjansson, render obvious claims 1-5.

VII. SECONDARY CONSIDERATIONS OF NONOBVIOUSNESS

224. I am not aware of any secondary considerations of nonobviousness. If Patent Owner identifies any alleged evidence of secondary considerations in the future, I reserve the right to respond to that information.

VIII. CONCLUSION

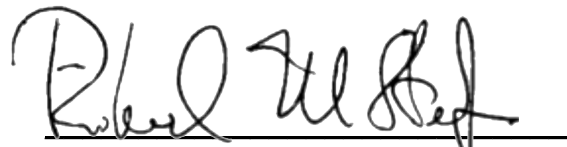
225. For the foregoing reasons, it is my opinion that claims 1-5 of the '337 patent are anticipated by, and would have been obvious to a person of ordinary skill in the art at the time of the alleged invention in view of, the prior art discussed above.

226. I reserve the right to supplement my opinions in the future to address or respond to any issues that the Patent Owner may raise, as well as new information including, but not limited to, any claim constructions advanced by the Patent Owner or adopted by the Board in the Institution Decision, and respond to any alleged secondary considerations as they become available to me.

Amazon.com, Inc. v. SoundClear Technologies LLC
Declaration of Richard Stern, Ph.D. – Patent 11,069,337

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 January 30, 2025 at Pittsburgh, Pennsylvania.



Richard Stern, Ph.D.