UNITED STATES PATENT AND TRADEMARK OFFICE

BEFORE THE PATENT TRIAL AND APPEAL BOARD

SONOS, INC.,
Petitioner,

v.

GOOGLE LLC,
Patent Owner.

Case No. TBD
U.S. Patent No. 11,024,311

SONOS, INC.'S PETITION FOR *INTER PARTES* REVIEW OF U.S. PATENT NO. 11,024,311

TABLE OF CONTENTS

TABLE OF AUTHORITIES

Cases

Statutes

LISTING OF EXHIBITS

Exhibit	Description
1001	U.S. Patent No. 11,024,311
1002	Declaration of Dr. Michael T. Johnson
1003	US Patent Publication 2017/0357478 ("Piersol")
1004	US Patent Publication 2017/0083285 ("Meyers")
1005	Japanese Publication No. 2003/223188 ("Masahide")
1006	English translation of Masahide
1007	US Patent Publication 2013/0073293 ("Jang")
1008	*Certain Audio Players and Components Thereof*, Inv. No. 337-TA-1330, Order No. 14
1009	US Patent No. 9,318,107 ("'107 Patent)
1010	Relevant Excerpts of '107 Patent file history
1011	K. Paliwal, Filter-Bank Energies For Robust Speech Recognition (1999), *available at* https://maxwell.ict.griffith.edu.au/spl/publications/papers/euro99_kkp_fbe.pdf
1012	K. Paliwal, Filter-Bank Energies For Robust Speech Recognition (1999), *available at* https://www.semanticscholar.org/paper/FILTER-BANK-ENERGIES-FOR-ROBUST-SPEECH-RECOGNITION-Paliwal/a7777b5132f4b82430925190f8850d3d2090cbc6
1013	US Patent Publication 2014/0163978 ("Basye")
1014	Reetika & Saini, Impact of Router's Buffer Size on Packet Loss Due To Congestion in TCP, IJERT, Vol. 3 Issue 5 (May 2014), *available at* https://www.ijert.org/research/impact-of-routers-buffer-size-on-packet-loss-due-to-congestion-in-tcp-IJERTV3IS051297.pdf

1015	Krzyzanowski, Quality of Service (Sept. 29, 2012), *available at* https://people.cs.rutgers.edu/~pxk/417/notes/03-qos.html.
1016	Goetze and Appell, Voice activity detection driven acoustic event classification for monitoring in smart homes, IEEE Xplore (Dec. 2010), *available at* https://www.researchgate.net/publication/224215429_Voice_activity_detection_driven_acoustic_event_classification_for_monitoring_in_smart_homes
1017	Venter and Hanekom, Automatic detection of African elephant (Loxodonta africana) infrasonic vocalisations from recordings, ScienceDirect (2010).
1018	Miralles et al., SAMARUC a programmable system for passive acoustic monitoring of cetaceans (2013).
1019	Solera-Urena *et al.*, Real-Time Robust Automatic Speech Recognition Using Compact Support Vector Machines, IEEE Transactions on Audio Speech and Language Processing (May 2012), *available at* https://www.researchgate.net/publication/228109902_Real-Time_Robust_Automatic_Speech_Recognition_Using_Compact_Support_Vector_Machines.
1020	*Certain Audio Players and Components Thereof*, Inv. No. 337-TA-1330, Redacted excerpts of Google's Pre-Hearing Brief (May 31, 2023)
1021	*Certain Audio Players and Components Thereof*, Inv. No. 337-TA-1330, Redacted excerpts of Google's Post-Hearing Brief (July 10, 2023)
1022	*Certain Audio Players and Components Thereof*, Inv. No. 337-TA-1330, Google's Amended Complaint (Feb. 6, 2023)
1023	Infringement Claim Chart for U.S. Patent No. 11,024,311, *Google LLC v. Sonos, Inc.*, Case No. 22-CV-4553, Dkt. No. 1-8 (N.D. Cal. Aug. 8, 2022)

I. Introduction

Petitioner Sonos, Inc. ("Sonos") respectfully requests institution of *Inter Partes* Review ("IPR") of claims 1-3, 8-12, 14-20 ("Challenged Claims") of U.S. Patent 11,024,311 (the "'311 Patent"; Ex.1001). The '311 Patent issued on June 1, 2021 and is currently assigned to Google LLC ("Google" or "Patent Owner").

This Petition demonstrates at least a reasonable likelihood that Sonos will prevail with respect to at least one Challenged Claim. 35 U.S.C. §314(a). Sonos asserts that the Challenged Claims are anticipated and/or rendered obvious by the asserted prior art.

Pursuant to 37 CFR §42.22, Petitioner respectfully requests the Board to review the asserted prior art and below analysis, institute a trial for IPR of the Challenged Claims, and cancel those claims as unpatentable.

II. Mandatory Notices

A. Real Party-in-Interest (37 C.F.R. §42.8(b)(1))

Sonos is the real party-in-interest.

B. Related Matters (37 C.F.R. §42.8(b)(2))

Google asserted, *inter alia*, the '311 Patent and U.S. Patent 9,812,128 (the "'128 Patent")[1] against Sonos in the following ITC and district court proceedings:

[1] The '128 Patent is related to the '311 Patent, and the Board recently instituted IPR2022-01594 challenging claims of the '128 Patent.

- *Certain Audio Players and Components Thereof*, Inv. No. 337-TA-1330 ("1330 Investigation");
- *Google LLC v. Sonos, Inc.*, Case No. 3:22-CV-04553 (N.D. Cal.) ("NDCA-Action").

Sonos was served with the NDCA-Action complaint on August 9, 2022. Thus, this Petition is timely under 35 U.S.C. § 315(b).

Sonos previously filed a petition challenging claims 1-3, 8-12, 14-18, and 20 of the '311 Patent, and the Board instituted IPR of those claims. *See* IPR2022-01592 ("First Petition"). Unlike this Petition, the First Petition did not challenge claim 19, as Google had not asserted that claim against Sonos at that time. Pursuant to the Consolidated Trial Practice Guide, Sonos has provided a ranking of the petitions, an explanation of the material differences between the petitions, and why the Board should exercise discretion to institute this Petition. *See* Sonos's Ranking Paper (concurrently filed herewith).

Google also asserted U.S. Patents 10,134,398 and 10,593,330 against Sonos in *Certain Audio Players and Components Thereof*, Inv. No. 337-TA-1329 ("1329 Investigation") and *Google LLC v. Sonos, Inc.*, Case No. 3:22-CV-04552 (N.D. Cal.). These patents are related to the '311 Patent. Sonos also challenged various claims of these patents, and the Board instituted IPR of those claims. *See* IPR2023-00118; IPR2023-00119.

C. Counsel Information (37 C.F.R. §42.8(b)(3))

Sonos designates the following counsel:

Lead Counsel:

Michael P. Boyea, Reg. No. 70,248
LEE SULLIVAN SHEA & SMITH LLP
656 West Randolph Street, Suite 5W
Chicago, IL 60661
(312) 754-9602
boyea@ls3ip.com

Back-up Counsel:

Jae Y. Pak, Reg. No. 72,428
LEE SULLIVAN SHEA & SMITH LLP
656 West Randolph Street, Suite 5W
Chicago, IL 60661
(312) 754-9602
pak@ls3ip.com

Cole B. Richter, Reg. No. 65,398
LEE SULLIVAN SHEA & SMITH LLP
656 West Randolph Street, Suite 5W
Chicago, IL 60661
(312) 754-9602
richter@ls3ip.com

Pursuant to 37 C.F.R. § 42.10(b), a Power of Attorney accompanies this Petition. The above identified lead and backup counsel are registered practitioners associated with Customer No. 130085 listed in that Power of Attorney.

D. Service Information (37 C.F.R. §42.8(b)(4)) and Payment of Fees

Sonos consents to service by electronic mail at boyea@ls3ip.com,

pak@ls3ip.com, richter@ls3ip.com, and LS3_team@ls3ip.com.

The fee set forth in 37 C.F.R. § 42.15(a) for this Petition has been paid. Any additional fees due in connection with this Petition may be charged to Deposit Account 506632.

III. Requirements for IPR

A. Grounds for Standing

Sonos certifies that the '311 Patent is available for IPR and that Sonos is not barred or estopped from requesting IPR. Sonos has filed this Petition within one year of service of the NDCA-Action complaint on August 9, 2022. Sonos has not filed a civil action challenging the validity of any claim of the '311 Patent.

B. Identification of Challenge

Sonos requests IPR of Challenged Claims 1-3, 8-12, and 14-20. The application underlying the '311 Patent was filed February 10, 2020, but claims an earlier priority date as a continuation-in-part of an April 1, 2015 filing. Ex.1001. However, the Challenged Claims are not entitled to the April-2015 priority date because the specification of the earlier ancestor—the '107 Patent—fails to provide written description support for Google's narrow interpretation of "detecting a voice input" recited in the '311 claims. *Infra* §VIII. At best, the effective priority date for the Challenged Claims is October 3, 2016. *Id.*

Accordingly, review is requested in view of US Patent Publication Nos. 2017/0357478 ("Piersol";Ex.1003), 2017/0083285 ("Meyers";Ex.1004),

2013/0073293 ("Jang";Ex.1007), and Japanese Publication No. 2003/223188 ("Masahide";Ex.1005;Ex.1006[2]). These references qualify as prior art under 35 U.S.C. §§102(a)(1)-(2)(AIA).

Sonos requests institution based on the following grounds:

Ground 1A	Claims 1-3, 8-11, 14-20 anticipated by Piersol
Ground 1B	Claims 1-3, 8-12, 14-20 rendered obvious by Piersol and Meyers
Ground 2A	Claims 1-3[3], 8-11, 14, 16-17, 19-20 anticipated by Masahide
Ground 2B	Claims 1-3, 8-11, 14, 16-17, 19-20 rendered obvious by Masahide and Jang

These grounds are supported by a Declaration from Dr. Michael T. Johnson. Ex.1002.

IV. Discretionary Factors Favor Institution

A. Denial Under §314(a) Would Be Inappropriate[4]

Consistent with the goals of *Fintiv* and Director Vidal's June 21, 2022 Interim Procedure ("I.P."), Sonos requests that the Board grant this Petition and institute IPR. The I.P. states "the PTAB will not discretionarily deny petitions based on applying *Fintiv* to a parallel ITC proceeding." I.P., 7. Thus, the parallel 1330 Investigation does not implicate *Fintiv.*

---

[2] Certified English translation.
[3] Masahide alone also renders claim 3 obvious.
[4] Denial based on filing of multiple petitions would also be inappropriate for the reasons explained in Sonos's Ranking Paper.

Moreover, the *Fintiv* factors in view of the parallel NDCA-Action weigh strongly in favor of institution. Specifically, *Fintiv* factors 1-5 favor institution because the court stayed the NDCA-Action pending the outcome of the 1330 Investigation (Target Date is January 16, 2024) and any appeals. NDCA-Action, Dkt.22. There has been minimal investment in the NDCA-Action by the court and parties, and the Board is likely to address the '311 Patent via this Petition before the district court. Thus, there are no concerns involving conflicting decisions or duplicative work by the Board and district court. Additionally, the compelling strength of the grounds presented herein (*Fintiv* factor 6) further favors institution.

B. Denial Under §325(d) Would Be Inappropriate

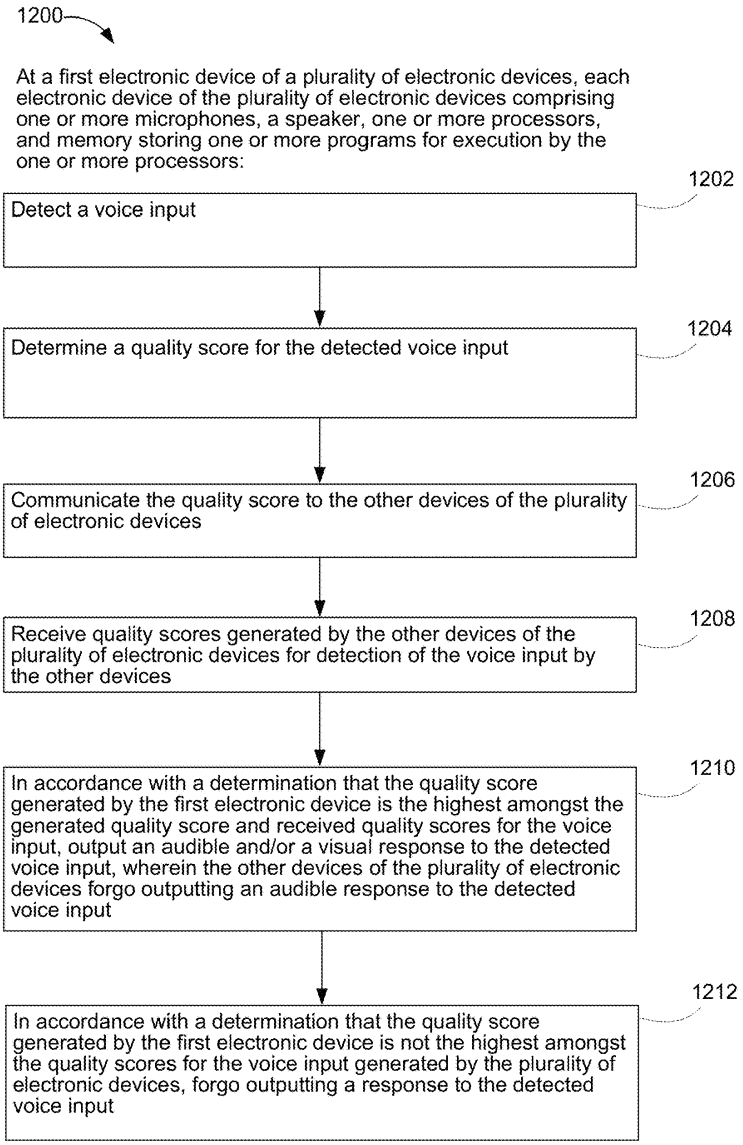
The Board should not exercise its discretion to deny institution under §325(d) for three main reasons: (1) the asserted prior art references were ***not*** considered during prosecution, (2) the Examiner erred in a manner material to patentability by overlooking at least Masahide and Jang, both of which were published before the ***earliest possible*** priority date (October 9, 2014), and (3) Sonos's Petition relies on evidence not previously available to the USPTO, including an expert declaration demonstrating why the Challenged Claims are not entitled to the earliest possible priority date and how the prior art would be understood by a POSITA. *See Liquidia Techs., Inc. v. United Therapeutics Corp.,* IPR2021-00406, 2021 WL 3553779, at *9 (PTAB Aug. 11, 2021) (finding *Becton* factors (e) and (f) weigh in Petitioner's

favor based on submission of expert declarations not previously available to USPTO).

V. ’311 Patent Overview

The ’311 Patent “relates generally to… methods and systems for coordination amongst multiple voice interface devices.” Ex.1001, 1:44-47. Much like the prior art, the ’311 Patent explains that “[a] location… may include multiple devices that include voice assistant systems” and that it is “desirable for there to be a leader amongst the voice assistant devices that is responsible for responding to the user’s voice inputs, in order to reduce user confusion.” *Id.*, 1:60-2:3; *see also, e.g.*, *id.*, 7:7-14.

To that end, the ’311 Patent discloses “coordination” or “leadership negotiation” techniques amongst “voice-activated electronic devices” (or simply, “electronic devices”), which are best summarized by FIG. 6:



Id., FIG. 6; *see also, e.g.*, *id.*, 20:55-22:20, 29:5-30:3. As shown, the disclosed "coordination" techniques seek to select a single "leader amongst themselves to respond to the user…." *Id.*, 7:4-14. This general process is more generally known in the field as "arbitration."

VI. Level of Ordinary Skill in the Art

A POSITA in the technology field of the '311 Patent as of the effective priority date (i.e., no earlier than October 3, 2016) would have had the equivalent of a four-year degree from an accredited institution (typically denoted as a B.S. degree) in electrical or computer engineering, or a related field, and approximately 2-4 years of experience in the field of speech processing, or an equivalent level of skill, knowledge, and experience. *See* Ex.1002, ¶42-45.

VII. Claim Construction

A. "Detecting a Voice Input"

Each '311 claim recites *detecting a voice input from a user* (cls. 1, 10) or *detect a voice input* (cl. 16).[5] Google contends that *detecting a voice input* (i) refers "to a process for identifying that an audio input includes voice" (IPR2022-01592, Patent Owner's Response (June 30, 2023) ("IPR2022-01592 POR"), p.12), (ii) "refers to a process separate from performing speech recognition (e.g., speech to text) on the voice input" (*id.*, p.13), and (iii) can be satisfied by "voice activity detection" or "hotword/wakeword detection" (*id.*, pp.3-9, 16). In other words, Google contends *detecting a voice input* requires identifying voice ***without***

[5] For clarity, claim language is italicized throughout this Petition.

performing "speech recognition," which Google narrowly interprets as being a ***mutually exclusive*** process from hotword/wakeword detection.[6]

While Sonos disagrees with this narrow interpretation, Sonos assumes that Google's interpretation of *detecting a voice input* is correct for purposes of this Petition.

B. "Second Values… From Other Electronic Devices"

Each '311 claim recites *receiving second values associated with the voice input from other electronic devices of the plurality of electronic devices* (cls. 1, 10) or *receive respective second values from other electronic devices of the plurality of electronic devices* (cl. 16).

In the 1330 Investigation, Google successfully argued that each of these terms—*second values* and *other electronic devices*—"can be satisfied by a single device, score, or value." Ex.1008, p.35. In other words, Google successfully argued that the '311 claims only require a claimed *electronic device* to receive a single *second value* from a single *other electronic device*. *Id.*, pp.39-40.

Sonos assumes that Google's interpretation of *second values… from other electronic devices of the plurality of electronic devices* is correct for purposes of this Petition.

---

[6] All emphasis set forth herein has been added unless noted otherwise.

C. Dependent Claim 11

Google contends claim 11 "requires a single device programmed to both arbitrate based on a 'first value' (claim 10) ***and*** to 'respond[]' to a voice command in accordance with a determination that the command is related to a device." IPR2022-01592 POR, p.26. In other words, Google contends claim 11 requires the device be "configured to perform arbitration based on both a 'first value' and the 'relevance' of a command to" the device. *Id.*, pp.26-27.

According to Google, claim 11 covers: (i) a device "***overriding*** the outcome of the arbitration based on scores based on a determination that a type of the command is related to the device" (*id.*, p.27), and (ii) the disclosure at 22:65-23:8 of the '311 Patent describing "relevance of the command" being used to "narrow leadership candidates" before arbitrating on "scores." *Id.*, p.26.

Sonos assumes that Google's interpretation of claim 11 is correct for purposes of this Petition.

D. "Confidence Level of Detection of the Voice Input"

Dependent claim 18 depends on claim 17. Dependent claim 17 recites *[t]he system of claim 16, wherein the first value comprises a first quality score of the voice input.* In turn, dependent claim 18 recites *[t]he system of claim 17, wherein the first quality score comprises a confidence level of detection of the voice input.*

The ’311 specification informs a POSITA that *a confidence level of detection of the voice input* refers to a likelihood or probability of the quality or closeness of the match between a voice model and the voice input perceived by the device. Ex.1002, ¶58-61 (*citing* Ex.1001, 24:35-49, 30:44-46). However, Google argued in the 1330 Investigation that this term means “a measure reflecting the confidence that the audio input is voice.” Ex.1020, p.6. In other words, Google broadly contends *a confidence level of detection* “is a measure of the likelihood that an audio signal includes speech, and indicates a level of confidence that a voice input was detected.” Ex.1021, p.6.

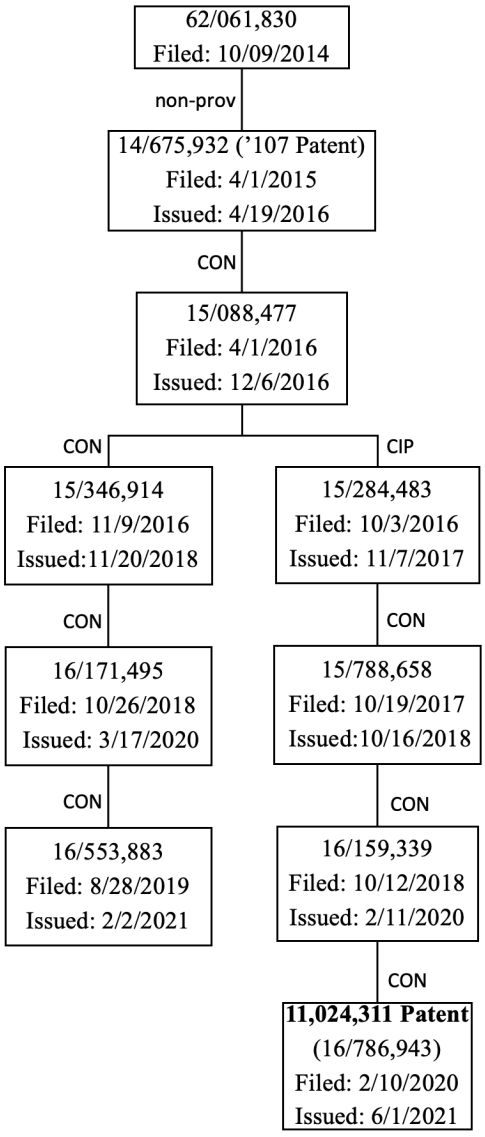
While Sonos disagrees with this broad interpretation, Sonos assumes it is correct for purposes of this Petition.

VIII. Effective Priority Date

A. ’311 Patent Family

The ’311 Patent issued on June 1, 2021 from an application filed on February 10, 2020. As shown below, the ’311 Patent is part of a chain of filings ultimately claiming priority back to provisional application 62/061,830 (the “’830 Provisional”) filed on October 9, 2014 and non-provisional application 14/675,932 filed on April 1, 2015, which issued as US Patent 9,318,107 (the “’107 Patent”;Ex.1009). The ’311 Patent is part of a continuation-in-part branch of the family tree that developed on October 3, 2016 when application 15/284,483 (the

"'483 CIP Application") was filed.



The '311 Patent's specification is purportedly the same as the specification of the '483 CIP Application but materially differs from the '107 Patent's specification.

B. '311 Patent's Effective Priority Date Is No Earlier Than October 3, 2016

A petitioner may challenge an asserted priority date of a patent by showing the priority application does not provide written description support for a challenged

claim. *See Intel Corp. v. Tela Innovations, Inc.,* IPR2019-01636, Paper 46, p.36 (PTAB Mar. 10, 2021). The patent owner then has the burden to "come forward with evidence to prove entitlement to claim priority to an earlier filing date." *PowerOasis, Inc. v. T-Mobile USA, Inc.*, 522 F.3d 1299, 1305-306 (Fed. Cir. 2008).

For a claim to receive the benefit of priority to an earlier-filed application, the earlier-filed application must support every claim element "in the manner provided by section 112(a)." 35 U.S.C. §119(e)(1); 35 U.S.C. §120. In other words, a claim is entitled to an earlier date ***only if*** an earlier-filed application discloses every claim element sufficiently to meet the written description requirement, such that a POSITA can "clearly conclude that the inventor invented the claimed invention as of the filing date sought.'" *Anascape, Ltd. v. Nintendo of Am. Inc.*, 601 F.3d 1333, 1335 (Fed. Cir. 2010).

Here, the Challenged Claims are ***not*** entitled to the '107 Patent's (or '830 Provisional's) filing date because there is no written description support for Google's interpretation of *detecting a voice input* discussed above. *Supra* §VII.A.; Ex.1002, ¶64-83. In fact, Google's narrow interpretation conflicts with the '107 Patent's express teachings for two reasons.

First, the '107 Patent discloses that a user's "utterance" (which is *a voice input*) is "detected" simply by a device's microphone perceiving or sensing the utterance, and any "process[ing]" occurs after the voice has been detected:

- "[S]ystem 100 illustrates a user 102 speaking an utterance 104 that is ***detected by microphones*** of computing devices 106, 108, and 110." Ex.1009, 3:56-58.
- "When the user 102 speaks the utterance 104, 'OK computer,' each computing device that has a microphone in the vicinity of the user 102 ***detects and processes*** the utterance 104. Each computing device ***detects*** the utterance 104 ***through*** an audio input device such as ***a microphone***." *Id*., 5:3-7.

Thus, the '107 Patent discloses *detecting a voice input* simply refers to a device perceiving or sensing a voice input via the device's microphone and does not require any sort of "process[ing]," much less "process[ing] for identifying that an audio input includes voice," as Google argued. IPR2022-01592 POR, p.12; Ex.1002, ¶65-66.

Second, the '107 Patent clearly explains that hotword/wakeword detection ***is*** a form of "speech recognition."

- "In an example environment, the ***hotword*** used to invoke the system's attention are the words 'OK computer.' Consequently, ***each time*** the words 'OK computer' are spoken, it is picked up by a microphone, conveyed to the system, which ***performs speech recognition*** techniques ***to determine whether the hotword was spoken*** and, if so, awaits an ensuing command or query." Ex.1009, 1:58-64.

Thus, hotword/wakeword detection and speech recognition are ***not*** mutually exclusive processes at least in the context of the '311 Patent. Ex.1002, ¶67.

Indeed, during prosecution, the Examiner confirmed the '107 Patent's specification discloses that hotword/wakeword detection is a form of "speech recognition." Specifically, in its August 20, 2015 Office Action Response, Google amended its pending claims to distinguish hotword detection from "automated speech recognition," as shown in relevant part below:

> 1. (Currently amended) A computer-implemented method comprising:
> receiving, by a first computing device, audio data that corresponds to an utterance;
> determining, based on an analysis of the audio data and without performing automated speech recognition processing on the audio data, a first value ~~corresponding to~~ that reflects a first likelihood that the utterance includes a particular hotword;

Ex.1010, p.2. In the October 1, 2015 Office Action, however, the Examiner correctly pointed out that there was no written description support in the '107 Patent's specification for distinguishing hotword detection from "speech recognition":

> **NOTE**: During current analysis of the present invention it is realized that ***contradictory to the claim language***, ***the specification in fact states that speech recognition is performed***, which from a technical standpoint is ***almost inherent*** given the nature of command and control. ***In order to spot a keyword some sort of speech recognition must be performed***. There is no indication of waveform analysis that would suggest otherwise in the specification.

Id., p.15. The Examiner also cited specific passages in the '107 Patent's specification (including 1:58-64 cited above) that directly contradicted Google's

attempt to distinguish hotword detection from "automated speech recognition." *See id.*, pp.16-17.

In recognition of this glaring problem, Google abandoned its attempt to distinguish hotword detection from "automated speech recognition" in its After-Final Response dated December 7, 2015 and amended the pending independent claims, including claim 1, as follows:

> 1. (Currently amended) A computer-implemented method comprising:
>
> receiving, by a first computing device, audio data that corresponds to an utterance;
>
> before beginning automated speech recognition processing on the audio data, processing the audio data using a classifier that classifies audio data as including a particular hotword or as not including the particular hotword;
>
> determining, based on ~~an analysis of the audio data and without performing automated speech recognition processing on the audio data~~ the processing of the audio data using the classifier that classifies audio data as including a particular hotword or as not including the particular hotword, a first value that reflects a first likelihood that the utterance includes ~~a particular~~ the particular hotword;
>
> receiving a second value that reflects a second likelihood that the utterance includes the particular hotword, as determined by a second computing device;
>
> ~~transmitting, to the second device, the first value that reflects the first likelihood that the utterance includes the particular hotword;~~
>
> comparing the first value that reflects the first likelihood that the utterance includes the particular hotword and the second value that reflects the second likelihood that the utterance includes the particular hotword; and
>
> based on comparing the first value that reflects the first likelihood that the utterance includes the particular hotword to the second value that reflects the second likelihood that the utterance includes the particular hotword, determining whether to ~~perform~~ begin performing automated speech recognition processing on the audio data.

See id., pp.33, 36-37, 39-40.

A POSITA would appreciate that these amendments ***cannot*** be interpreted to cover detecting a hotword ***without*** performing speech recognition for several reasons. Ex.1002, ¶73-82.[7]

First, as discussed, the Examiner objected to Google's attempt to amend the claims to recite hotword detection "without performing automated speech recognition processing on the audio data" because it contradicted the '107 Patent's disclosure, and Google did not traverse and instead, abandoned this attempt by further amending the claim language.[8]

Second, a POSITA would appreciate that the claimed "classifier" utilizes speech recognition, and in the context of the '107 Patent, the amendment requiring "[a] classifier that classifies audio data as including a particular hotword or as not including the particular hotword" merely refers to limited-vocabulary speech recognition, which is speech recognition limited to identifying a particular hotword, as opposed to large-vocabulary speech recognition. Ex.1002, ¶75.

---

[7] *See, e.g.*, *Ajinomoto Co. v. Int'l Trade Comm'n*, 932 F.3d 1342, 1351 (Fed. Cir. 2019) ("[W]hen a word is changed during prosecution, the change tends to suggest that the new word differs in meaning in some way from the original word.").

[8] *See, e.g.*, *Rheox, Inc. v. Entact, Inc.*, 276 F.3d 1319, 1326 (Fed. Cir. 2002) ("[P]atentee [may] relinquish[] a particular claim construction based on the totality of the prosecution history, which includes amendments to claims….").

Indeed, Google added the following dependent claims in the same After-Final Response, each identifying specific speech recognition techniques for the claimed "classifier" of independent claim 1:

> 23. (New) The method of claim 1, wherein processing the audio data using a classifier that classifies audio data as including a particular hotword or as not including the particular hotword comprises:
>
> extracting filterbank energies or mel-frequency cepstral coefficients from the audio data.
>
> 24. (New) The method of claim 1, wherein processing the audio data using a classifier that classifies audio data as including a particular hotword or as not including the particular hotword comprises:
>
> processing the audio data using a support vector machine or a neural network.

Ex.1010, p. 41. These amendments appear to correspond to disclosure from the '107 Patent. *See* Ex.1009, 5:12-28.

As of the '107 Patent's earliest possible priority date—October 9, 2014—a POSITA would have understood that extracting "filterbank energies or mel-frequency cepstral coefficient" and "processing [] audio data using a support vector machine or a neural network" to classify speech data as containing a hotword or not refer to well-known ***speech recognition*** techniques. Ex.1002, ¶78 (*citing* Ex.1011 (1999 publication stating "Mel frequency cepstral coefficients (MFCCs) are perhaps the ***most widely used*** features ***for speech recognition***," stating "filter-bank energies (FBEs)" "have been used in 50's, 60's and 70's ***for speech recognition***," and showing "FBEs perform at least as good as (and sometimes even better than) the

MFCCs ***for*** robust ***speech recognition***."); Ex.1019, pp. 1, 4 (2012 publication comparing use of support vector machines and neural networks in speech recognition applications); Ex.1004, ¶26-27).

Third, the other elements added by Google (i.e., "before beginning automated speech recognition processing on the audio data" and "determining whether to begin performing automated speech recognition processing on the audio data") do ***not*** suggest that "processing the audio data using a classifier that classifies audio data as including a particular hotword or as not including the particular hotword" is done without speech recognition. Ex.1002, ¶80-82. Instead, these elements are consistent with the specification's ***repeated*** disclosure of a voice-enabled device first performing speech recognition to detect a hotword in the user's utterance and then initiating speech recognition on the speech in the utterance ***following*** the hotword. *See* Ex.1009, 1:58-2:3, 3:67-4:3, 6:24-30, 6:50-55, 6:56-63, 7:2-4, 7:27-29, 7:46-49, 7:56-62, 8:3-15, 9:28-33, 10:66-11:2, 11:51-54.

Considering the Examiner's rejection, a POSITA's knowledge, and the '107 Patent expressly stating "***each time*** the [hotword is] spoken, it is picked up by a microphone, conveyed to the system, which performs speech recognition techniques to determine whether the hotword was spoken and, if so, awaits an ensuing command

or query" (Ex.1009, 1:58-64),[9] a POSITA would conclude that the claimed "classifier" utilizes speech recognition. Ex.1002, ¶82.

In summary, the '107 Patent does not provide written description support for *detecting a voice input* as interpreted by Google. Thus, under Google's interpretation, the Challenged Claims are ***not*** entitled to the '107 Patent's priority date, and instead, have a priority date no earlier than ***October 3, 2016***—the filing date of the '483 CIP Application. *See Smith & Nephew, Inc. v. Arthrex, Inc.,* IPR2017-00275, 2017 WL 1969743, at *5 (PTAB May 10, 2017) ("For a claim in a later-filed application to be entitled to the filing date of an earlier application, the earlier application must provide written description support for the claimed subject matter.") (*citing Anascape*, 601 F.3d at 1337); *Dr. Reddy's Labs. S.A. v. Indivisor UK Ltd.*, IPR2019-00329, Paper 49, 11, 81-82 (PTAB June 2, 2020) (finding priority date was when challenged claims were filed because challenged claims lacked written description support in earlier-filed application).

---

[9] *See, e.g., Honeywell Int'l, Inc. v. ITT Indus., Inc.*, 452 F.3d 1312, 1319 (Fed. Cir. 2006) ("Where, as here, the written description clearly identifies what his invention is, an expression by a patentee during prosecution that he intends his claims to cover more than what his specification discloses is entitled to little weight."); *Biogen, Inc. v. Berlex Lab'ys, Inc.*, 318 F.3d 1132, 1140 (Fed. Cir. 2003) ("Representations during prosecution cannot enlarge the content of the specification....").

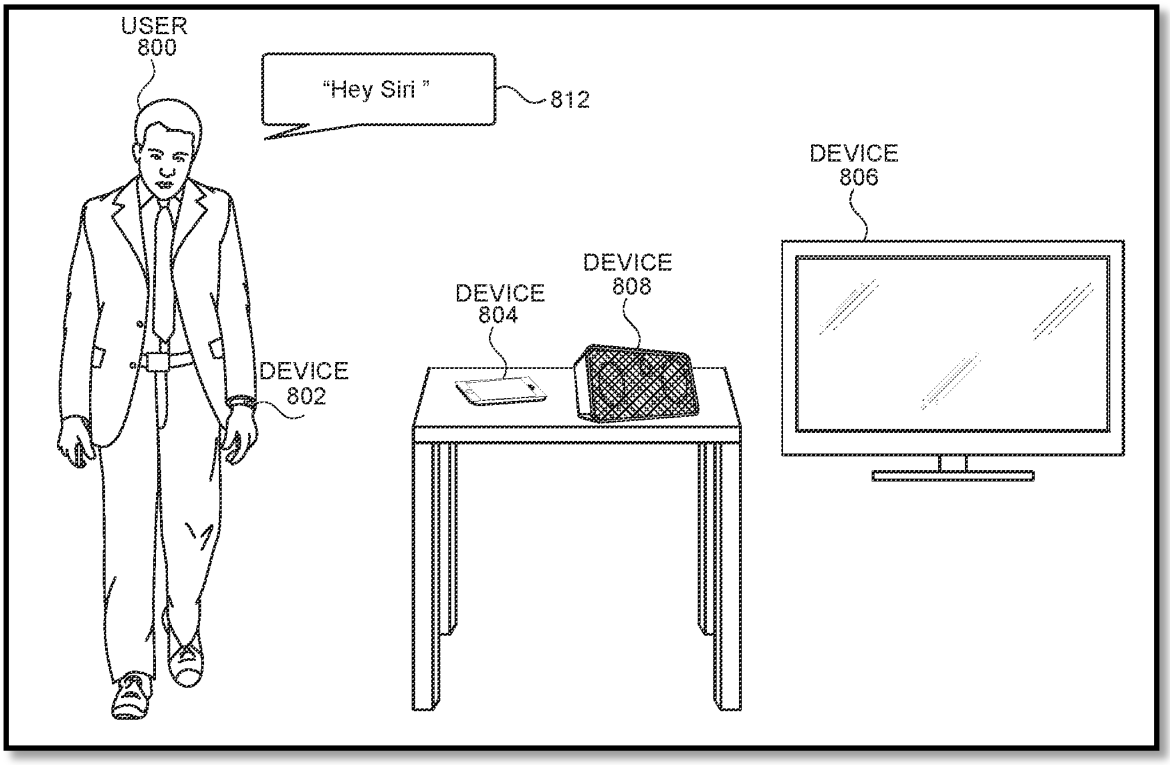
IX. Prior Art Overview

A. Piersol

Piersol was filed on September 16, 2016 (Ex.1003) and thus, qualifies as prior art since this date is before the '311 Patent's effective priority date (October 3, 2016).

Piersol generally addresses the problem where multiple "electronic devices having virtual assistant services" in a system detect the same voice input from a user containing a request. Ex.1003, ¶3. Piersol discloses that the voice input may cause each device to handle the user's request, which may "result in a confusing experience for the user" and/or "duplicative or conflicting operations" being performed when each device handle the user's request. *Id*; *see also id.*, ¶33, 295.

Piersol depicts an example system comprising multiple "electronic devices" that detect the same voice input from a user:



Id., FIG. 8B. Piersol summarizes its disclosed solution as follows:

> [A] first electronic device samples an audio input using a microphone. The first electronic device broadcasts a first set of one or more values based on the sampled audio input. Furthermore, the first electronic device receives a second set of one or more values, which are based on the audio input, from a second electronic device. Based on the first set of one or more values and the second set of one or more values, the first electronic device determines whether to respond to the audio input or forego responding to the audio input.

Id., Abstract. In this way, Piersol discloses techniques whereby, "[i]n response to detecting [a] spoken instruction [], electronic devices 802-808 initiate an arbitration

process to identify… an electronic device for responding to the spoken instruction…." *Id.*, ¶247.

B. Masahide

Masahide published on August 8, 2003 (Ex.1006) and thus, qualifies as prior art since this date is before the '311 Patent's earliest possible priority date (October 9, 2014).

Masahide generally relates to "voice input systems, voice input methods, and voice input programs where the speech of a user may be input into ***a plurality*** of voice inputs." Ex.1006, ¶1. Masahide recognized that, when multiple voice input devices are located in a room, it may be necessary for the user to specify a particular target device for the voice input, and additionally, voice input may need to be suppressed for devices other than the intended voice input device. *Id.*, ¶2. Accordingly, Masahide seeks to "solve[] the question of what to do with each voice input device when the voice of the user is input to ***a plurality*** of networked voice input devices." *Id.*, ¶11.

Masahide summarizes its disclosed solution as follows:

> [T]he voice input method of the present invention is characterized by detecting voice information input into a plurality of voice input devices connected to a network, respectively, and sharing information regarding said voice detected by each voice input device connected to the network with other voice input devices via the network as determinative information, with each voice input

> device connected to the network using the determinative information in the applicable voice input device itself, and other voice input devices to determine processing and execution of the voice information.

Id., ¶5.

Masahide explains that the "determinative information" used to perform arbitration and select one voice input device to interact with the user can take various forms, including signal-to-noise ratio information. *See, e.g.*, *id.*, ¶49-51, FIG. 6, ¶52, 54-56, FIG. 7.

C. Meyers

Meyers was filed on September 21, 2015 (Ex.1004) and thus, qualifies as prior art since this date is before the '311 Patent's effective priority date (October 3, 2016).

Meyers generally addresses the problem where multiple "speech interface devices" in a system detect the same "user utterance" and "independently attempt to process and respond to the user utterance as if it were two separate utterances" by disclosing "techniques for ***avoiding such duplicate efforts and responses***, among other things." Ex.1004, ¶14; *see also id.*, ¶31, 36, 66.

Meyers depicts an example system comprising multiple "speech interface devices" (102(a), 102(b)) that detect the same voice input from a user:

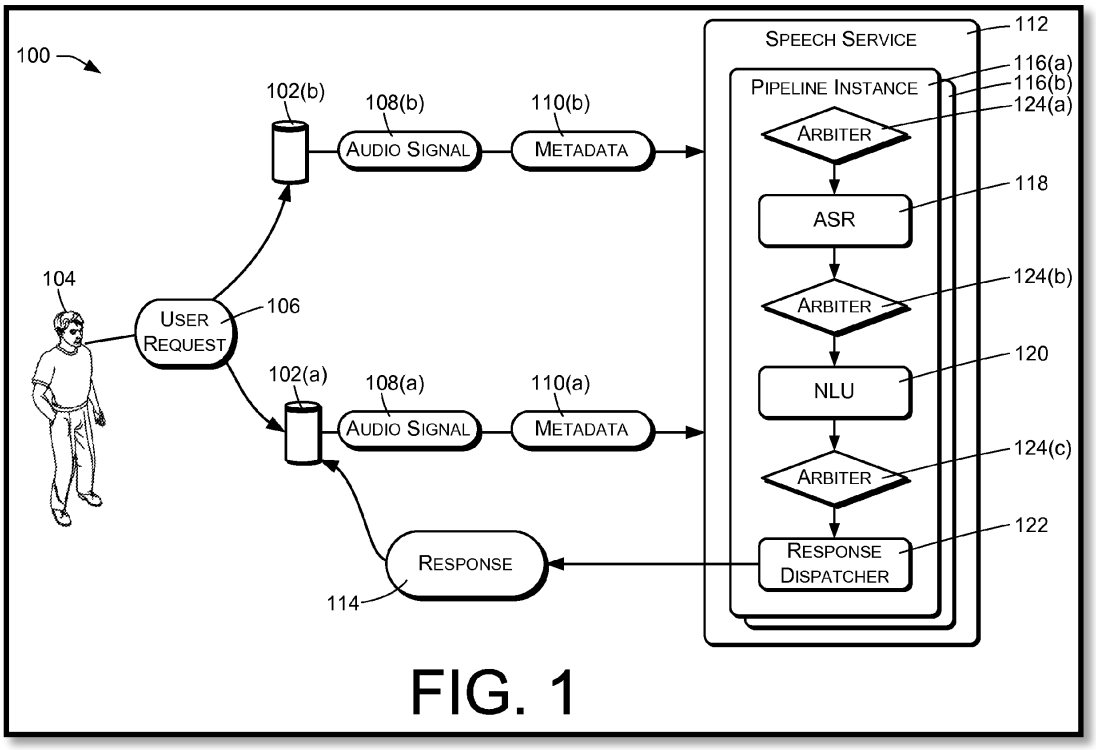


FIG. 1

Id., FIG. 1.

To address the foregoing problem, Meyers discloses arbitration techniques "for determining which of the devices 102 the user 104 most likely wants to respond to the user request 106." *Id.*, ¶31, Abstract.

For instance, Meyers discloses that a "speech service 112" performs the disclosed arbitration techniques, and the "speech service 112" could take the form of "a network-accessible service implemented by multiple server computers that support devices 102 in the homes or other premises of many different users" or "one or more of the devices 102 may include or provide the speech service 112." *Id.*, ¶30.

Meyers explains that each speech interface device is configured to "detect" voice input containing "a user request 106." *See, e.g.*, *id.*, ¶15, 21. Meyers further

explains that "the user request 106 may be prefaced by a wakeword or other trigger expression that is spoken by the user 104 to indicate that subsequent user speech is intended to be received and acted upon by one of the devices 102." *Id.*, ¶23. Meyers explains "[t]he device 102 may ***detect*** the wakeword and interpret subsequent user speech as being directed to the device 102" where "[a] wakeword in certain embodiments may be a reserved keyword that is ***detected*** locally by the speech interface device 102." *Id.*

In particular, Meyers' speech interface device includes a "voice activity detector 922" that performs "voice activity detection" (VAD) and "wake word detector 920" that performs "wakeword detection":

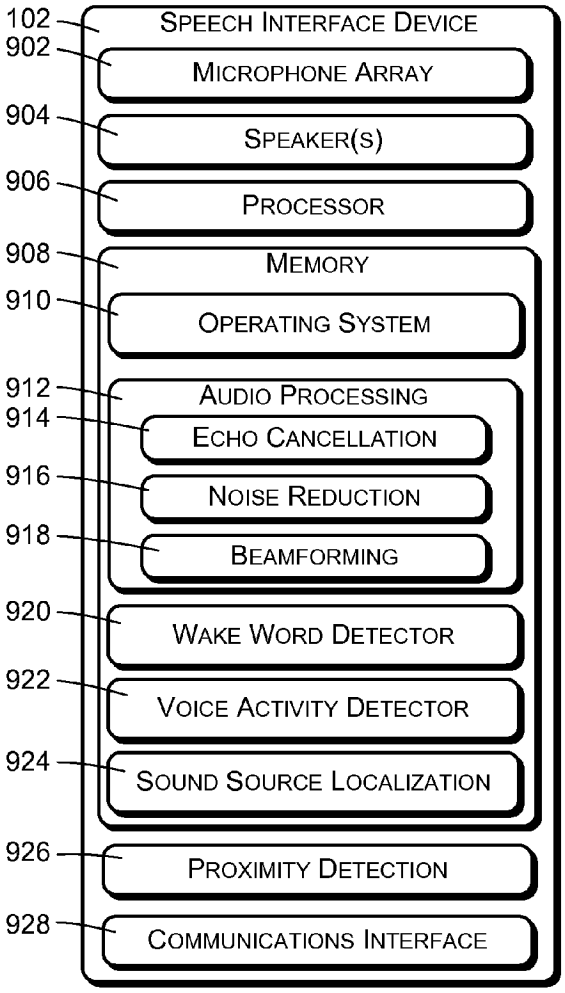


FIG. 9

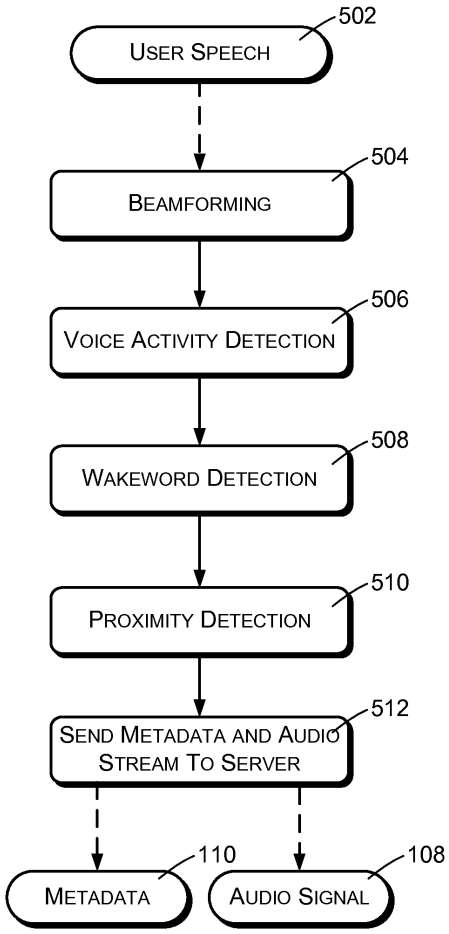


FIG. 5

Meyers explains that "VAD ***determines*** the level of ***voice presence*** in an audio signal by analyzing a portion of the audio signal to evaluate features of the audio signal such as signal energy and frequency distribution." *Id.*, ¶102; *see also id.*, ¶120. And, Meyers explains that the wake word detector (or "expression detector") and wakeword detection involves "analyz[ing] an audio signal produced by a microphone of the device 102 to ***detect the wakeword***" and "may be implemented using keyword spotting technology" to "***generate***[] ***a true/false*** output to indicate

whether or not the predefined word… was represented in the audio signal" "***[r]ather than producing a transcription*** of the words of the speech." *Id.*, ¶24; *see also id.*, ¶104, 119. Meyers also explains that wakeword detection may be performed on an audio signal "within which voice activity has been detected…." *Id.*, ¶103.

Upon detecting a voice signal, Meyers discloses that each speech interface device is configured to determine "metadata" associated with the detected voice signal that includes "one or more signal attributes." *See, e.g.*, *id.*, ¶16, 40. Meyers discloses numerous signal attributes, including (i) "the amplitude of the audio signal," (ii) "the signal-to-noise ratio of the audio signal," (iii) "the level of voice presence detected in the audio signal," and (iv) "the confidence level with which a wakeword was detected in the audio signal," among other examples. *Id.*, ¶40; *see also id.*, ¶70, 108. Specifically, Meyers describes "the level of voice presence detected in the audio signal" and "the confidence level with which a wakeword was detected in the audio signal" attributes as a likelihood of detection of speech. *See id.*, ¶102, 108; *see also id.*, ¶27, 106, 119.

Meyers further provides examples where a device having the highest "amplitude," "level of detected voice presence," "signal-to-noise ratio," or "level of confidence" is selected as the arbitration winner. *Id.*, ¶66, 70.

D. Jang

Jang published on March 21, 2013 (Ex.1007) and thus, qualifies as prior art

since this date is before the '311 Patent's earliest possible priority date (October 9, 2014).

Jang discloses a system of "electronic devices" (*see, e.g.*, Ex.1007, ¶30, 53, FIGs. 1-2), where each "electronic device" includes a "voice recognition unit" that "can carry out voice recognition upon voice signals input through the microphone 122" of the electronic device. *Id.*, ¶110. Jang explains that the "voice recognition unit" enables each electronic device to recognize a user's voice command, and there can be scenarios where these devices recognize the same voice command and "generate[] the same output in response to the user's voice command." *Id.*, ¶120.

To address this duplicate response problem, Jang provides arbitration techniques such that, "[i]n an environment involving a plurality of devices, it can be determined by communication between the devices or by a third device managing the plurality of devices which of the plurality of devices is to carry out a user's voice command." *Id.*, ¶122. In this way, Jang's embodiments generally either involve each electronic device performing arbitration independently (*see, e.g.*, *id.*, ¶148-52, FIG. 9) or a "managing" device performing arbitration on behalf of the electronic devices in the system.

As one example, Jang discloses that the "voice recognition unit" enables each electronic device to "recognize the input voice signals by ***detecting voice activity from the input voice signals***, carrying out sound analysis thereof, and recognizing

the analysis result as a recognition unit." *Id.*, ¶110; *see also id.*, ¶127. Each electronic device may then compute a respective "voice recognition result," share those results with one another, and use them to determine which "electronic device" should respond to the voice input. *Id.*, ¶130-32, 135-40, 149-51.

Jang discloses that a "voice recognition result" (or "voice command result") may take various forms, including "a magnitude (gain value) of the recognized voice signal, voice recognition ratio of each device, type of content or application in execution by each device upon voice recognition, and remaining power." *Id.*, ¶135.

X. Ground 1A—Piersol (§102)

A. Piersol Anticipates Independent Claims 1, 10, 16

This Petition addresses independent claim 16 first as representative of the independent claims. Ex.1002, ¶119, 159-62.

1. Independent Claim 16

a. [16.0] *A system comprising a plurality of electronic devices, each of the electronic devices including one or more microphones, one or more processors, and memory storing one or more programs for execution by the one or more processors, each of the electronic devices programmed to:*

To the extent that the preamble is limiting, Piersol discloses it. Ex.1002, ¶120-28.

For instance, FIG. 1 below depicts "a system and environment for implementing a digital assistant according to various examples." Ex.1003, ¶14, 37.

In this "system 100," a "digital assistant can include client-side portion 102… executed on user device 104," which "can be any suitable electronic device," and "[s]econd user device 122 can be similar or identical to user device 104" and communicatively coupled with one another "via a wired or wireless network, such as a local Wi-Fi network." *Id.*, ¶39, 41, 44.

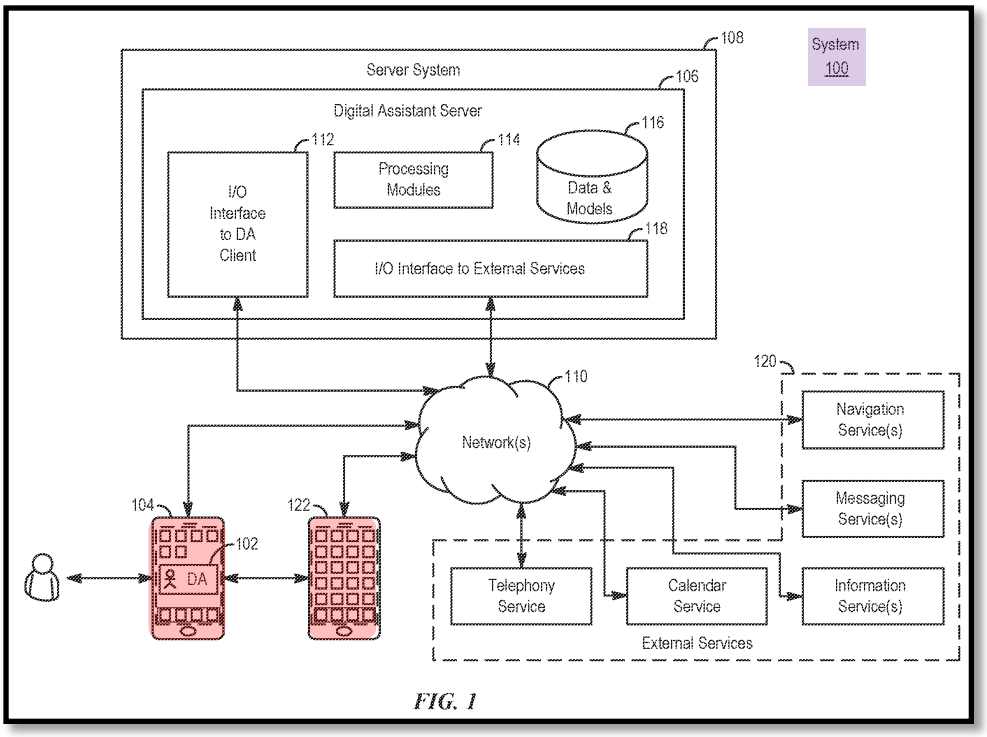
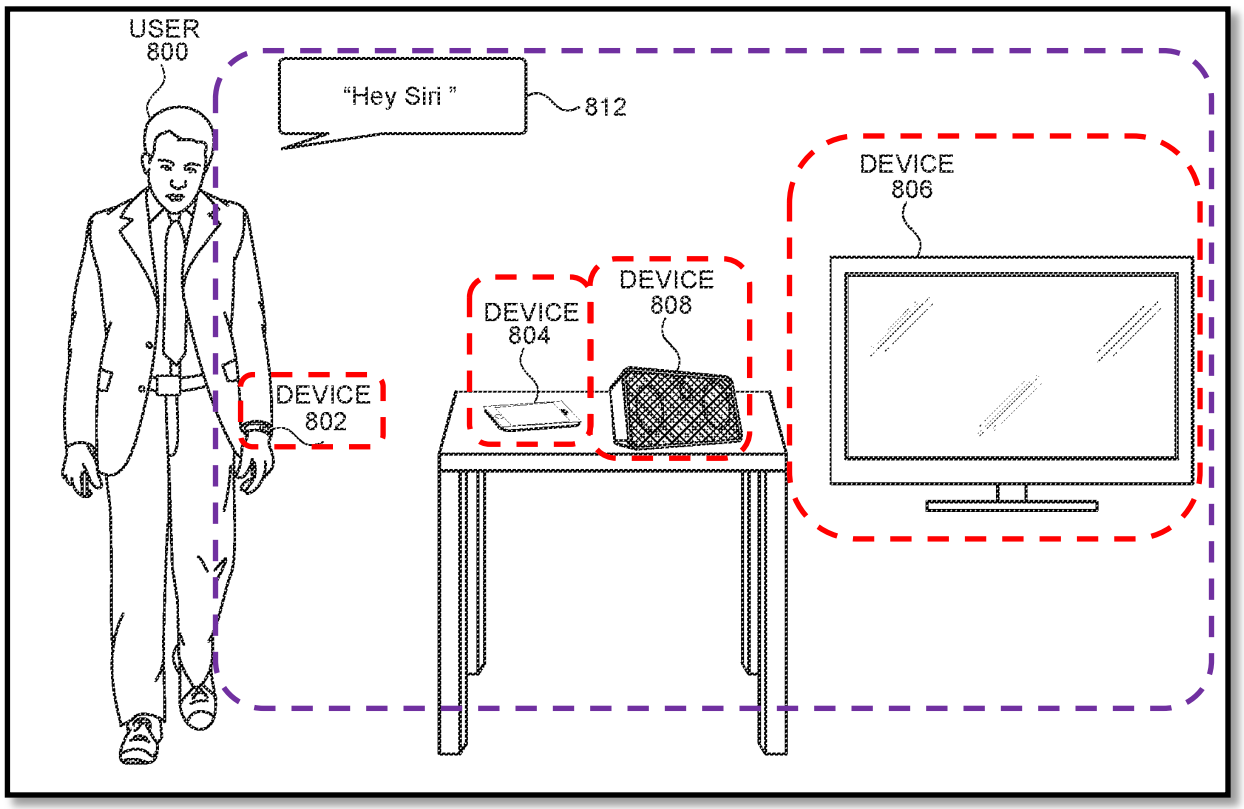


FIG. 1

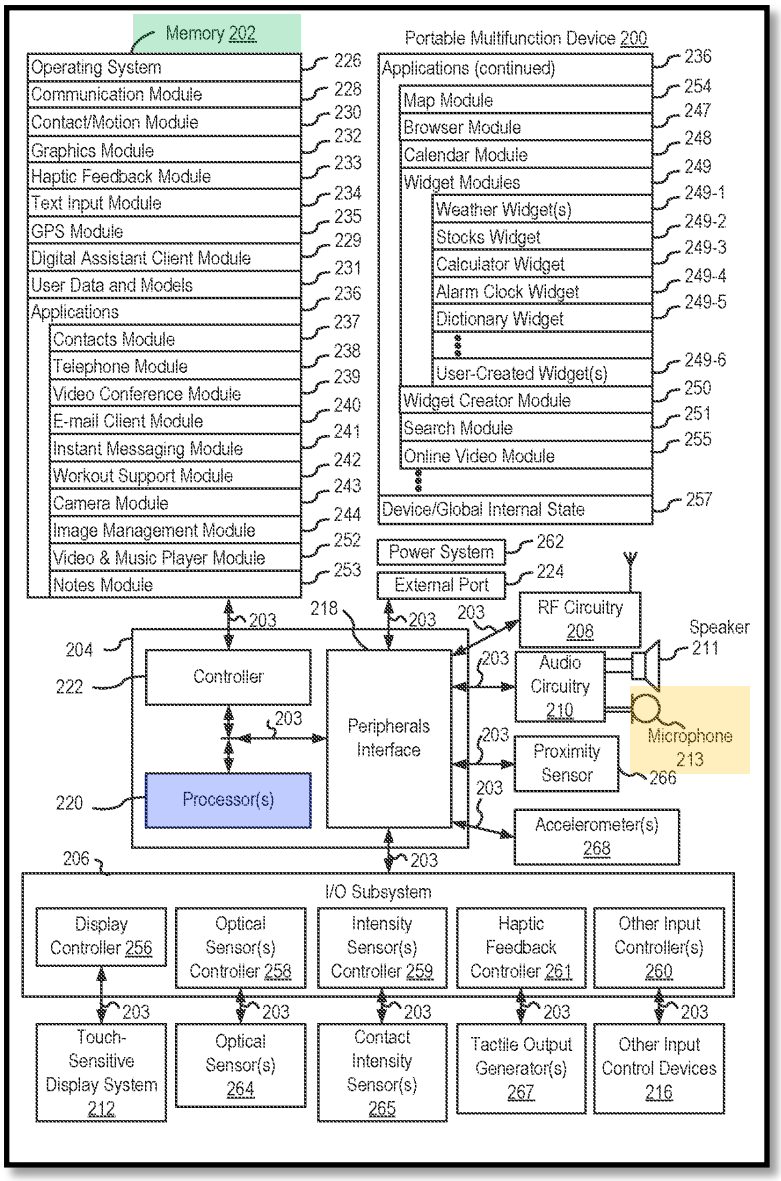
Id., FIG. 1 (annotations added); ¶41 (listing examples of "user device 104").

Piersol also depicts a system and environment for implementing a digital assistant according to various examples in FIGs. 8A-8C that show "electronic devices 802, 804, 806, and 808 of user 800" where "[o]ne or more of the devices 802-808 may be any of devices 104, 122, 200, 400, 600, and 1200 (FIGS. 1, 2A, 3, 4, 5A, 6A-6B, and 12) in some embodiments." *Id.*, ¶245.



Id., FIG. 8B (annotations added); *see also* FIGs. 8A, 8C.

With respect to FIG. 2A below, Piersol discloses that "[d]evice 200 includes memory 202…, one or more processing units (CPUs) 220,… microphone 213," among other components. *Id.*



Id., FIG. 2A (annotations added); *see also, e.g.*, *id.*, ¶17-21, 30, 321-23, FIGs. 3-4, 5A, 6A-6B, 12.

Piersol explains "[t]he one or more processors 220 run or execute various software programs and/or sets of instructions stored in memory 202 to perform various functions for device 200 and to process data." *Id.*, ¶53; *see also id.*, ¶51-52, 294-95.

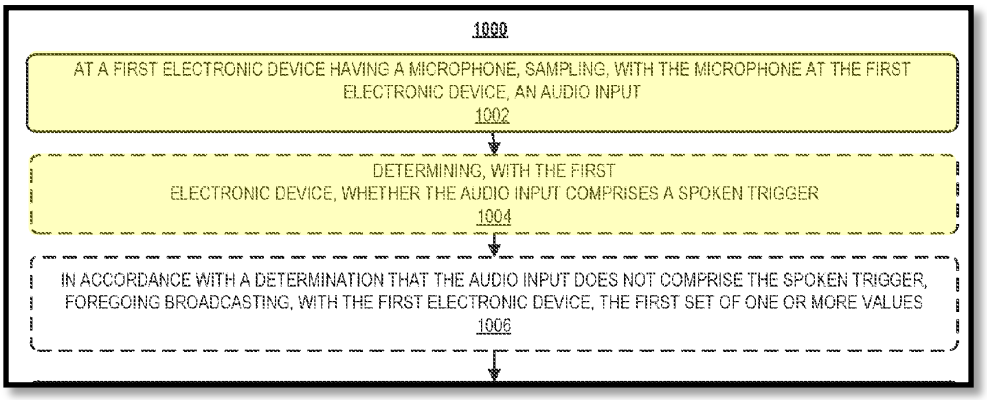
b. [16.1] *detect a voice input*

Piersol discloses this element under Google's narrow interpretation (*supra* §VII.A.) of *detect[ing] a voice input*. Ex.1002, ¶129-35.

For instance, Piersol explains that each electronic device[10] is configured to detect a "spoken trigger" (e.g., "Hey Siri"). Ex.1003, ¶254; *see also id.*, ¶246. A POSITA would appreciate that Piersol's "spoken trigger" is synonymous with a "hotword" or "wakeword." *See* Ex.1002, ¶131; Ex.1001, 6:2-5 ("[A] 'hotword' is a user defined or predefined word or phrase used to 'wake-up' or ***trigger*** a voice-activated device to attend/respond to a spoken command that is issued subsequent to the hotword").

Piersol explains that the electronic device detects a spoken trigger by its "microphone" first "sampl[ing] an audio input" and then "determin[ing] whether the audio input comprises a spoken trigger." Ex.1003, ¶297-98; *see also id.*, ¶330. This is illustrated in FIG. 10A below:

[10] This Petition refers to Piersol's "device" as an "electronic device" for consistency with claim language. *See* Ex.1002, ¶127.



Id., FIG. 10A (cropped; annotated). In this way, Piersol discloses *detect[ing] a voice input* by virtue of hotword detection. Ex.1002, ¶133.

Notably, consistent with Google's interpretation, Piersol discloses detecting a spoken trigger ***independent of*** "speech-to-text" or "ASR systems." Ex.1002, ¶134. Indeed, Piersol describes "digital assistant system 700"—which may be implemented on a cloud server (Ex.1003, ¶202) or on an electronic device (*id.*, ¶205)—as including "speech-to-text (STT) processing module 730" (*id.*, ¶212) that "can include one or more ASR systems" (*id.*, ¶215). Piersol explains "each ASR system can include one or more speech recognition models… and can implement one or more speech recognition engines" that "can be used to process the extracted representative features of the front-end speech pre-processor to produce intermediate recognitions results…, and ultimately, text recognition results…." *Id.* Piersol further explains, "[o]nce STT processing module 730 produces recognition results containing a text string…, the recognition result can be passed to natural language processing module 732 for intent deduction."

c. [16.2] *determine a first value associated with the detected voice input*

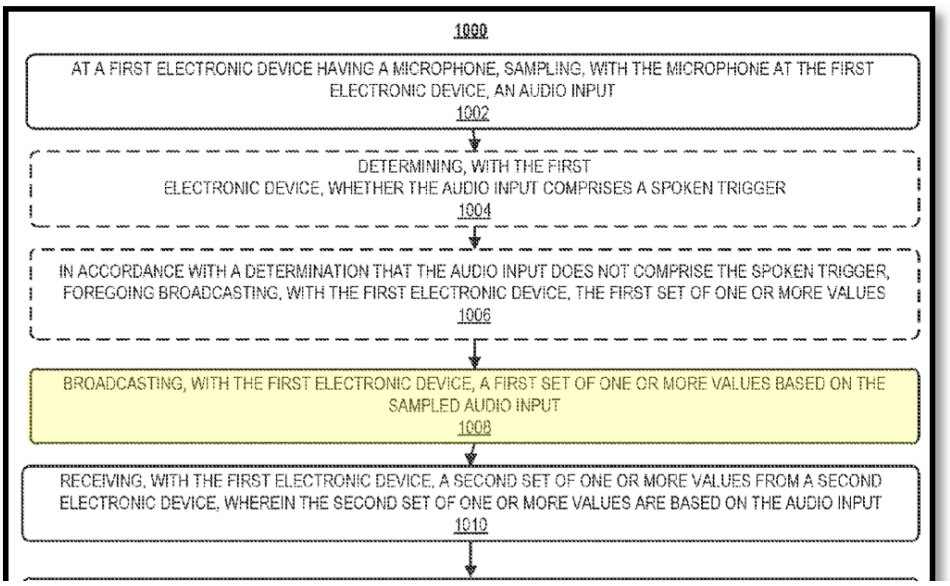
Piersol discloses this element. Ex.1002, ¶136-44. For instance, Piersol explains that each electronic device is configured to determine (or generate) a "score" (also referred to as a set of one or more "values") "in response to detecting a spoken trigger." Ex.1003, ¶254 ("[T]he electronic device ***determines respective values*** and/or broadcasts the values ***in response to detecting a spoken trigger***."), 261 ("In some examples, each of electronic devices 802-808 broadcasts a single ***score***…."), 267.

Specifically, in response to detecting a spoken instruction comprising the "spoken trigger," Piersol explains "electronic devices 802-808 ***initiate an arbitration process*** to identify… an electronic device for responding to the spoken instruction 834 from user 800." *Id.*, ¶247. Each electronic device determines and broadcasts a set of values/score that "may include one or more values" "based on the audio input as sampled on the respective device." *Id.*

Piersol discloses that these one or more values/score can take various forms (e.g., a single value/score derived from multiple values with different weights) (*id.*, ¶253, 261, 267) and comprise various information "relevant to implementing intelligent device arbitration," such as a value/score that could indicate (i) an "audio quality" (e.g., "an energy level of the audio input") comprising "signal-to-noise ratio, sound pressure, or a combination thereof" (*id.*, ¶248, 261), (ii) "an acoustic

fingerprint of the audio input" or "a confidence value that expresses the likelihood that the audio input originates from [a] particular user" (*id.*, ¶249), and/or (iii) "a type of the electronic device" in instances "when the spoken instruction specifies a task to be performed" (*id.*, ¶250).

Piersol describes an electronic device determining and then broadcasting a first set of values in connection with FIG. 10A below.



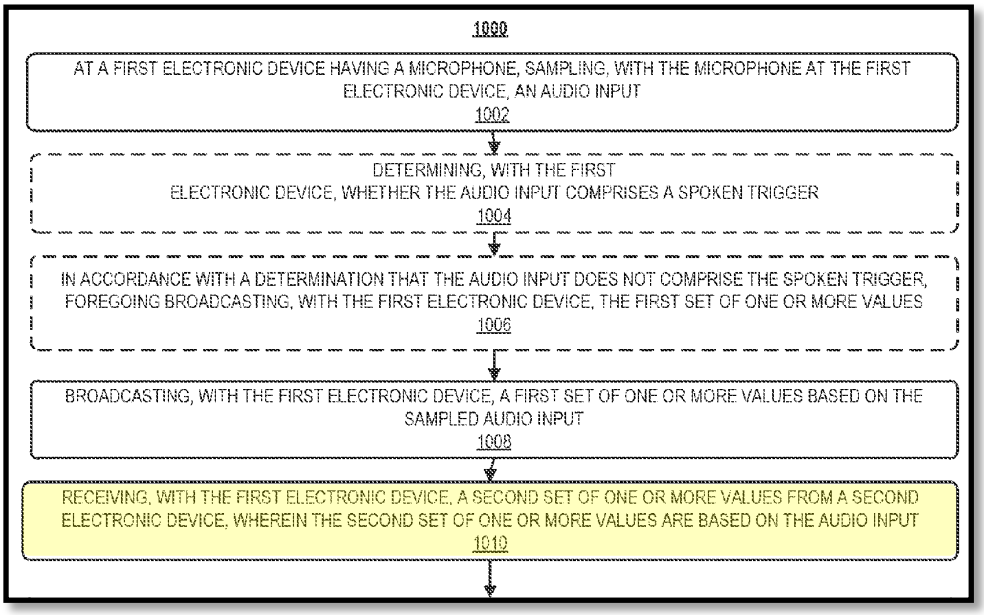
Id., ¶297, FIG. 10A (cropped; annotated).

d. [16.3] *receive respective second values from other electronic devices of the plurality of electronic devices*

Piersol discloses this element. Ex.1002, ¶145-49. For instance, Piersol discloses that "[e]ach of the electronic devices 802-808 may ***receive sets of values from other devices***." Ex.1003, ¶255. Piersol explains that this is accomplished by

virtue of another electronic device "broadcast[ing]" its one or more values/score. *Id.*, ¶261, 267.

Piersol describes an electronic device receiving one or more values from another electronic device in connection with FIG. 10A below.



Id., ¶299, FIG. 10A (cropped; annotated).

e. [16.4] *compare the first value and the respective second values*

Piersol discloses this element. Ex.1002, ¶150-53. For instance, Piersol discloses embodiments where each electronic device is configured to compare its own determined one or more values/score (e.g., "'energy level' value," "'state' values," "'acoustic fingerprint' values," or "'device type' values") with the one or more values/score received from another electronic device. *See* Ex.1003, ¶256

(Each of the electronic devices 802-808 ***compares*** a respective "energy level" value with the "energy level" values broadcasted by other electronic devices.), ¶257-59.

f. **[16.5]** *in accordance with a determination that the first value is higher than the respective second values, respond to the voice input*

Piersol discloses this element. Ex.1002, ¶154-58. For instance, Piersol discloses that each electronic device is configured to (i) "determine[] whether the [] electronic device is to respond to the audio input by ***determining*** whether a value of the first set of one or more values [e.g., "energy level" value (*id.*, ¶256, 267)] is ***higher than*** a corresponding value of the second set of one or more values" from another electronic device and (ii) "respond[] to the audio input" "***in accordance with a determination*** that the [] electronic device is to ***respond to the audio input***." Ex.1003, ¶300-301. This is illustrated in FIG. 10A below:

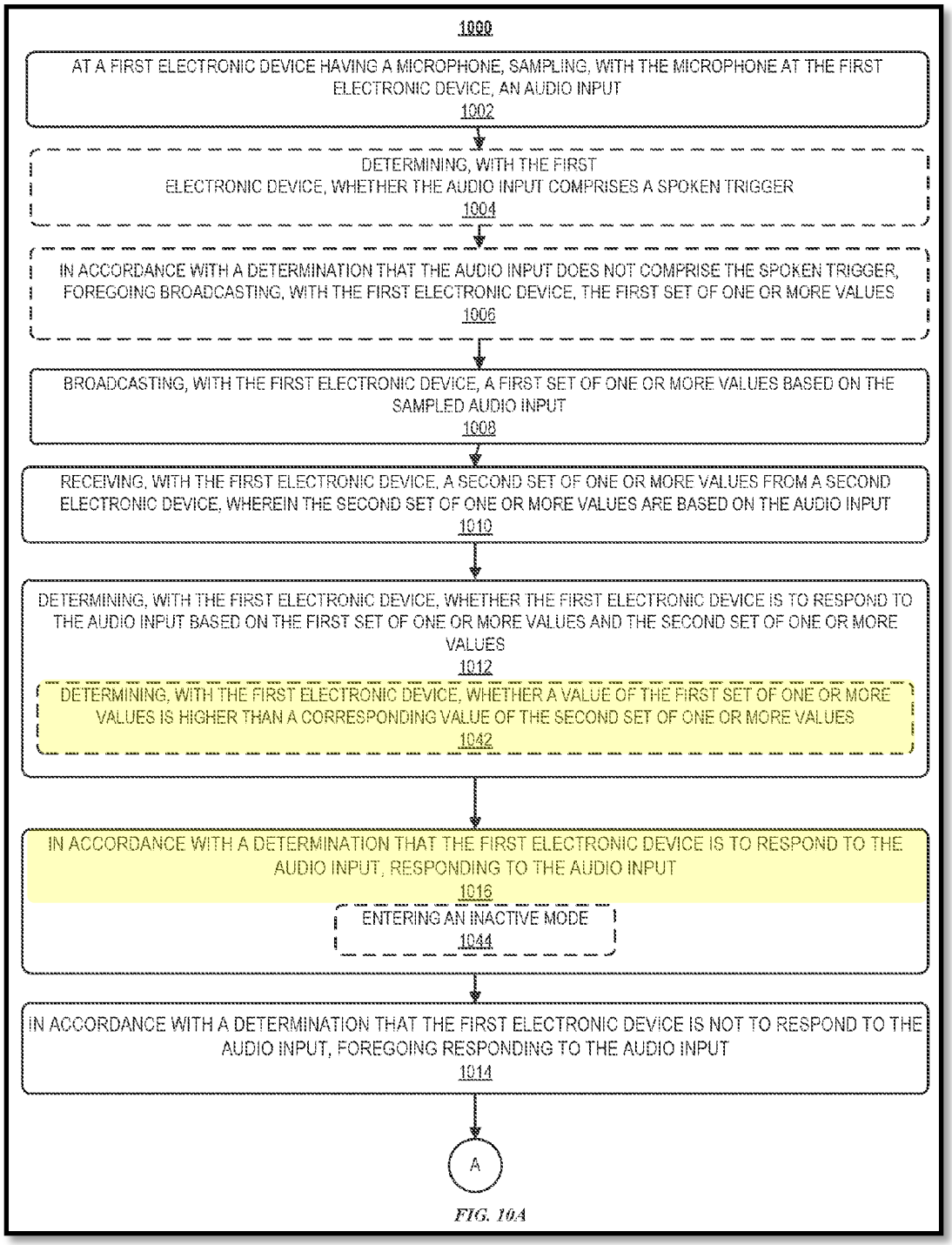


FIG. 10A

Id., FIG. 10A (annotated); *see also, e.g., id.*, ¶339.

Piersol discloses that the electronic device with the highest value/score can respond to the user's voice input in various manners, including by outputting an audible and/or visual response. *See, e.g.*, *id.*, ¶38, 240-41, 262; *infra* §X.B.3.

2. Independent Claim 1

Independent claim 1 recites a *method performed at a first electronic device* that is broader[11] than the recited functions of independent claim 16. In particular, independent claim 1 recites:

> **[1.0]** *A method performed at a first electronic device of a plurality of electronic devices, each of the plurality of electronic devices comprising one or more microphones, one or more processors, and memory storing one or more programs for execution by the one or more processors, the method comprising:*
> **[1.1]** *detecting a voice input from a user;*
> **[1.2]** *determining a first value associated with the voice input;*
> **[1.3]** *receiving second values associated with the voice input from other electronic devices of the plurality of electronic devices; and*
> **[1.4]** *in accordance with a determination that the first value is higher than the second values, responding to the detected input.*

As shown, elements 1.0, 1.1, 1.2, 1.3, and 1.4 of claim 1 map to elements 16.0, 16.1, 16.2, 16.3, and 16.5 of claim 16, respectively. Thus, Piersol discloses claim 1 for the same reasons it discloses claim 16. Ex.1002, ¶159-63.

---

[11] Claim 1 is broader than claim 16 because it does not expressly require the *comparing* function of element 16.4.

3. Independent Claim 10

Independent claim 10 recites a *first electronic device* provisioned with *one or more programs comprising instructions for* performing functions that are broader[12] than the recited functions of independent claim 16. In particular, independent claim 10 recites:

> **[10.0]** *A first electronic device of a plurality of electronic devices, the first electronic device comprising: one or more microphones; one or more processors; and memory storing one or more programs to be executed by the one or more processors, the one or more programs comprising instructions for:*
> **[10.1]** *detecting a voice input from a user;*
> **[10.2]** *determining a first value associated with the voice input;*
> **[10.3]** *receiving second values associated with the voice input from other electronic devices of the plurality of electronic devices; and*
> **[10.4]** *in accordance with a determination that the first value is higher than the second values, responding to the detected input.*

As shown, elements 10.0, 10.1, 10.2, 10.3, and 10.4 of claim 10 map to elements 16.0, 16.1, 16.2, 16.3, and 16.5 of claim 16, respectively. Thus, Piersol discloses claim 10 for the same reasons it discloses claim 16. Ex.1002, ¶159-63.

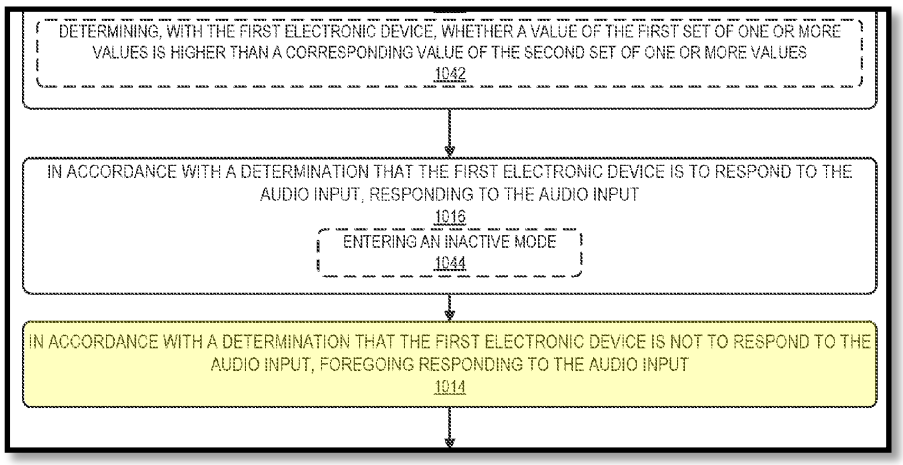
[12] Claim 10 is broader than claim 16 because it does not expressly require the *comparing* function of element 16.4.

B. Piersol Anticipates Dependent Claims 2-3, 8-9, 11, 14-15, 17-20

1. Dependent Claim 2

Claim 2 depends on claim 1 and further recites **[2.1]** "*in accordance with a determination that the first value is not higher than the second scores, forgoing responding to the detected input.*"[13] Piersol discloses this additional element. Ex.1002, ¶166-70.

For instance, Piersol discloses that each electronic device is configured such that, upon determining that its values/score is not higher than values/score received from another electronic device (i.e., "[i]f the electronic device determines not to respond to the audio input"), it "***forgo[es] responding*** to the audio input by entering an inactive mode (e.g. sleep mode)." Ex.1003, ¶262; *see also id.*, ¶266-67, FIG. 8B. This is illustrated in FIG. 10A below:



[13] This Petition assumes "*the second scores*" (which lacks antecedent basis) refers to "*the second values*" recited in claim 1.

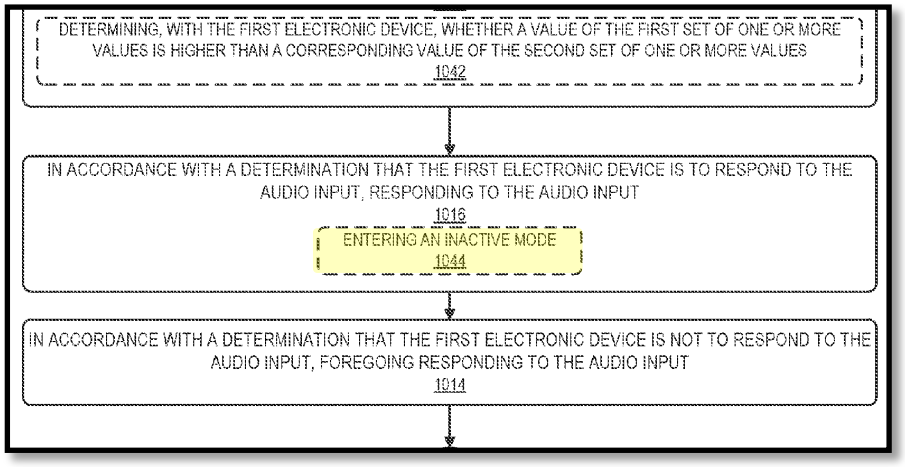
Id., FIG. 10A (cropped; annotated); *see also id.*, ¶303.

Piersol further provides an example described in connection with FIG. 8B, where devices that ***do not*** determine that they have the "highest 'energy level' value" "forego responding to the audio input." *Id.*, ¶267.

2. Dependent Claim 3

Claim 3 depends on claim 2 and further recites **[3.1]** "*forgoing responding to the detected input further comprises changing a state of operation of the first electronic device.*" Piersol discloses this additional element. Ex.1002, ¶171-75.

For instance, as just explained, Piersol discloses that an "electronic device can forego responding to the audio input ***by entering an inactive mode (e.g. sleep mode)***." Ex.1003, ¶262; *see also id.*, ¶266, 305, 336. This is illustrated in FIG. 10A below:



Id., FIG. 10A (cropped; annotated). A POSITA would understand that Piersol's electronic device "***entering***" such a mode/state would necessarily involve changing

its current mode/state from an active or non-sleep mode/state to the "inactive mode (e.g. sleep mode)." Ex.1002, ¶174. Otherwise, Piersol's electronic device would simply ***remain*** in the "inactive mode." *Id.*

3. Dependent Claim 8

Claim 8 depends on claim 1 and further recites **[8.1]** "*responding to the detected input comprises outputting an audible and/or a visual response to the detected voice input.*" Piersol discloses this additional element. Ex.1002, ¶176-78. For instance, Piersol discloses that the electronic device with the highest values/score can respond to the detected voice input in various manners, such as by "provid[ing] ***a visual output, an auditory output***, a haptic output, or ***a combination thereof***." Ex.1003, ¶307; *supra* §X.A.1.f.

4. Dependent Claim 9

Claim 9 depends on claim 1 and further recites **[9.1]** "*the first electronic device further comprises a speaker, and* **[9.2]** *responding to the detected input further comprises outputting an audible response.*" Piersol discloses these additional elements. Ex.1002, ¶179-82.

For instance, as just discussed, Piersol discloses that each electronic device can respond to the user's voice input by providing "an auditory output." *Supra* §X.B.3.; Ex.1003, ¶307. Piersol discloses that each electronic device comprises a speaker (e.g., "speaker 211") to enable outputting such responses. *Id.*, ¶55, 82.

5. Dependent Claim 11

Claim 11 depends on independent claim 10 and further recites **[11.1]** "*recognizing a command in the voice input; and* **[11.2]** *in accordance with a determination that a type of the command is related to the first electronic device, responding to the detected voice input.*" Piersol discloses these additional elements under Google's interpretation (*supra* §VII.C.). Ex.1002, ¶183-95.

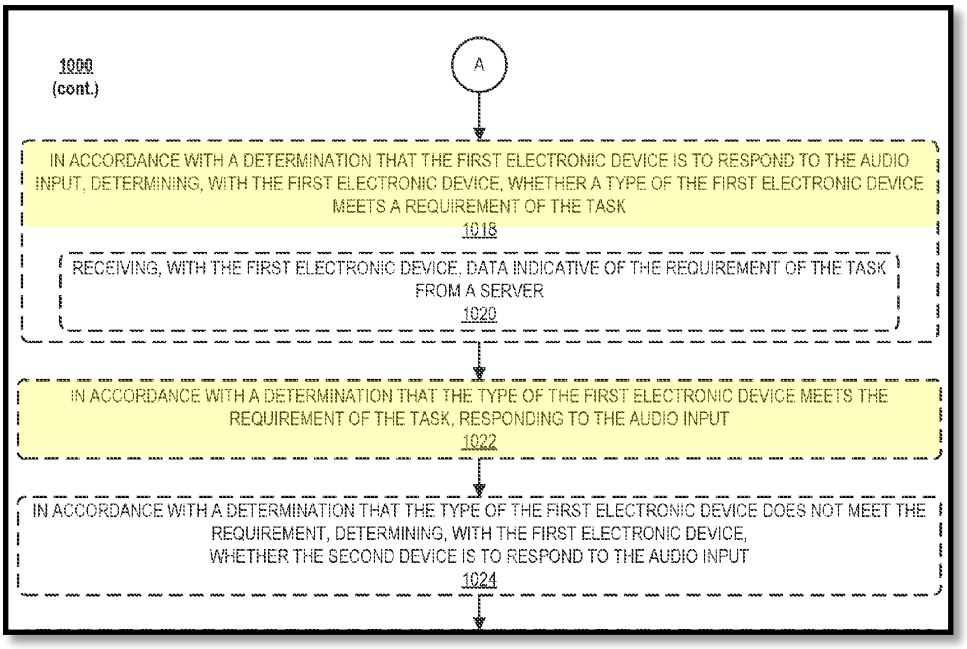
First, Piersol discloses **[11.1]**. For instance, Piersol discloses embodiments where "the functions of a digital assistant can be implemented… on [an electronic] device." Ex.1003, ¶46. Piersol discloses that such functions involve performing speech recognition on a voice input and deriving the user's intent using natural language understanding. *Id.*, ¶83 ("[U]ser data and models 231 can include various models (e.g., ***speech recognition models***,… ***natural language processing models***,… etc.) for processing user input and ***determining user intent***."); *see also id.*, ¶215, 217, 219-20.

Notably, Piersol discloses that "[d]evice type is particularly relevant to intelligent device arbitration when the ***user's input specifies a task to be performed***, such as 'Hey Siri, find me a TV episode' or 'Hey Siri, drive me there.'" *Id.*, ¶259. Piersol explains that an electronic device can identify a user's specified task "by locally analyzing the audio input…." *Id.*, ¶260. In other words, "spoken instruction 834 can be processed and resolved into one or more tasks locally…." *Id.*, ¶264.

Second, Piersol discloses **[11.2]**. For instance, Piersol discloses that each electronic device is configured to "obtain[] a list of acceptable ***device types*** that can ***handle the specified task***…, and ***determine***[] ***whether the electronic device is to respond to*** the audio input based on the broadcasted 'device type' values and the list of acceptable device types." *Id.*, ¶260.

As one example, Piersol discloses that each electronic device "***compares a respective 'device type' value(s)*** with the values broadcasted by other electronic devices" during arbitration, and a given "electronic device is to ***respond to the audio input if*** … [it] is ***the only device of a type that can handle the task specified in the audio input***." *Id.*, ¶259. Piersol explains that its arbitration process of determining whether to respond can look at "device type" separate from another score/value of the electronic device or use "device type" as part of determining a single, aggregated score/value. *Id.*, ¶261.

Piersol also discloses **[11.2]** in connection with FIG. 10B below:



Id., FIG. 10B (cropped; annotated); *see also id.*, ¶301-304, 331.

Finally, Piersol provides a specific example that demonstrates how it discloses the additional elements of claim 11 where "electronic device 806" responds to a voice input despite not having the highest "'energy level' value" because it determines that it is of a device type that can handle the task the user requested be performed. *See id.*, ¶263-64, 266.

6. Dependent Claim 14

Claim 14 depends on claim 10 and further recites **[14.1]** "*the plurality of electronic devices is communicatively coupled through a local network; and* **[14.2]** *the receiving is performed through the local network.*" Piersol discloses these additional elements. Ex.1002, ¶196-201.

For instance, Piersol discloses that the electronic devices are communicatively coupled through a local network, such as a "local Wi-Fi network" or "Bluetooth" network. Ex.1003, ¶44; *see also id.*, ¶42, 247. Piersol further discloses that each electronic device is configured to receive "one or more values" (or score) "***broadcasted***" by another electronic device through the local network. *Id.*, ¶247.

7. Dependent Claim 15

Claim 15 depends on claim 10 and further recites **[15.1]** "*the first electronic device is configured to be awakened by a plurality of affordances including a voice-based affordance; and* **[15.2]** *detecting the voice input is in response to the awakening of the first electronic device*." Piersol discloses these additional elements. Ex.1002, ¶202-207.

First, Piersol discloses that each electronic device is configured to be awakened by a voice-based affordance. *See* Ex.1003, ¶246 ("[O]ne or more of electronic devices 802-808 begin to sample audio input ***in response to a particular input by user 800, such as a spoken trigger***."). Piersol also discloses that each electronic device is configured to be awakened by a plurality of other affordances, including, a touch-based affordance (*id.*, ¶156), a user-proximity-based affordance (*id.*, ¶246), and a device-position based affordance (*id.*, ¶245, 269).

Second, Piersol discloses that each electronic device is configured such that detecting a voice input is responsive to the awakening of the electronic device. *See,*

e.g., id., ¶246 ("[O]ne or more of electronic devices 802-808 ***begin to sample audio input*** in response to a particular input by user 800, such as a spoken trigger."), ¶269.

8. Dependent Claim 17

Claim 17 depends on independent claim 16 and further recites **[17.1]** "*the first value comprises a first quality score of the voice input.*" Piersol discloses this additional element. Ex.1002, ¶208-11. For instance, Piersol discloses that each electronic device is configured to determine a "score" (or one or more "values") that can be "calculated based on a predetermined weighting of received ***audio quality*** (e.g., ***signal-to-noise ratio***[14])…." Ex.1003, ¶261; *see also, id.*, ¶248, 297.

9. Dependent Claim 18

Claim 18 depends on claim 17 and further recites **[18.1]** "*the first quality score comprises a confidence level of detection of the voice input.*" Piersol discloses this additional element under Google's broad interpretation. Ex.1002, ¶212-17.

For instance, Piersol discloses that each electronic device is configured to determine a "score" (or one or more "values") that can comprise "a ***confidence value*** indicative of a likelihood that the audio input was provided by a particular user." Ex.1003, ¶297; *see also id.*, ¶249. A POSITA would understand that Piersol's

[14] Claim 19 recites that the *first quality score comprises a signal-to-noise rating of detection of the voice input*, and the '311 Patent explains that "signal-to-noise ratio" is one example of *a signal-to-noise rating*. Ex.1001, 30:47-50.

"confidence value" satisfies Google's interpretation of *a confidence level of detection* (*supra* §VII.D.) because it is a "measure ***reflecting*** the confidence that the audio input is voice," as opposed to some non-voice sound. Ex.1002, ¶215.

Indeed, if Piersol's electronic device detected ***non-voice***, Piersol's "confidence value" would be extremely low (or would not even be calculated); whereas, if Piersol's electronic device detected ***voice***, Piersol's "confidence value" would be relatively higher when that voice matches a voice model corresponding to a particular user's voice. *Id.* In this way, Piersol's "confidence value" is a ***specific*** type of "measure ***reflecting*** the confidence that the audio input is voice." *Id.*[15]

10. Dependent Claim 19

Claim 19 depends on claim 17 and further recites **[19.1]** "*the first quality score comprises a signal-to-noise rating of detection of the voice input.*" Piersol discloses this additional element for the same reasons it discloses **[17.1]**. Ex.1002, ¶218-19; *supra* §X.B.8.

11. Dependent Claim 20

Claim 20 depends on claim 16 and further recites **[20.1]** "*each of the respective second values is indicative of a quality of the voice input detected by a*

[15] Piersol's "confidence value" is also no different than the ***only*** example of a "confidence level" disclosed in the '311 Patent. *See* Ex.1002, ¶216 (*citing* Ex.1001, 24:35-48).

respective one of the other electronic devices." Piersol discloses this additional element for the same reasons it discloses **[17.1]**. Ex.1002, ¶220-23; *supra* §X.B.8.

XI. Ground 1B—Piersol + Meyers (§103)

Ground 1B is presented to the extent Google argues that Piersol does not adequately disclose **[1.1]/[10.1]/[16.1]** *detect[ing] a voice input*. Other than Meyers making up for this hypothetical deficiency, Piersol and Meyers invalidates independent claims 1, 10, and 16 and dependent claims 2-3, 8-9, 11, 14-15, and 17-20 for the same reasons as Ground 1A.

Additionally, Ground 1B establishes that (i) Piersol combined with Meyers renders obvious dependent claim 12, and (ii) while Piersol anticipates dependent claim 18, Piersol combined with Meyers renders it obvious as well.

A. *Detecting a Voice Input.*

1. Meyers Discloses [1.1]/[10.1]/[16.1]

Meyers discloses an electronic device **[1.1]/[10.1]/[16.1]** *detect[ing] a voice input*. Ex.1002, ¶¶229-37.

For instance, Meyers discloses a system of "speech interface devices" (Ex.1004, FIG. 1), each configured to "detect" a voice input containing "a user request 106." *Id.*, ¶15 ("[E]ach speech interface device ***detects*** that a user is speaking a command…."), ¶21.

Meyers explains that "the user request 106 may be prefaced by a wakeword… that is spoken by the user 104 to indicate that subsequent user speech is intended to

be received and acted upon by one of the devices 102." *Id.*, ¶23. According to Meyers, "[t]he device 102 may ***detect*** the wakeword and interpret subsequent user speech as being directed to the device 102" where "[a] wakeword in certain embodiments may be a reserved keyword that is ***detected*** locally by the speech interface device 102." *Id.*

Specifically, Meyers' speech interface device includes a "voice activity detector 922" that performs "voice activity detection" (VAD) and "wake word detector 920" that performs "wakeword:

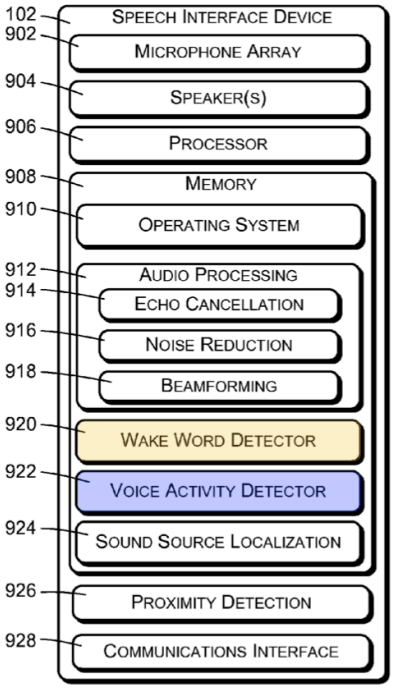


FIG. 9

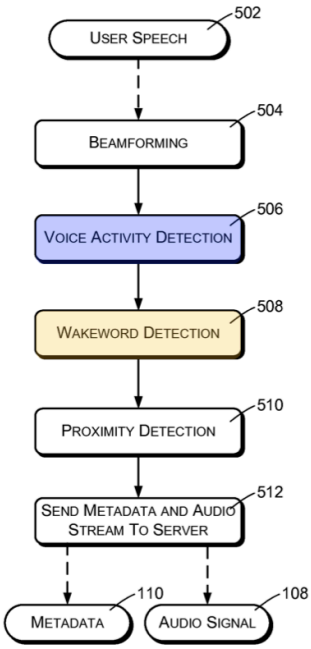


FIG. 5

Id., FIGs. 5, 9 (annotated).

Meyers explains that "VAD ***determines*** the level of ***voice presence*** in an audio signal by analyzing a portion of the audio signal to evaluate features of the audio

signal such as signal energy and frequency distribution" (*id.*, ¶102; *see also id.*, ¶120), and the wake word detector (or "expression detector") "analyzes an audio signal produced by a microphone of the device 102 to ***detect the wakeword***" and "may be implemented using keyword spotting technology" to "***generate***[] ***a true/false*** output to indicate whether or not the predefined word… was represented in the audio signal" "***[r]ather than producing a transcription*** of the words of the speech." *Id.*, ¶24; *see also id.*, ¶104, 119. Meyers also explains that wakeword detection may be performed on an audio signal "within which voice activity has been detected…." *Id.*, ¶103.

Upon detecting a voice signal, Meyers discloses that each speech interface device is configured to determine "metadata" associated with the detected voice signal that includes "one or more signal attributes," such as (i) "the amplitude of the audio signal," (ii) "the signal-to-noise ratio of the audio signal," (iii) "the level of voice presence detected in the audio signal," and (iv) "the confidence level with which a wakeword was detected in the audio signal," among other examples. *Id.*, ¶¶16, 40, 70, 108.

2. Motivation to Combine

For the reasons below, a POSITA would have been motivated to combine Piersol with Meyers before the effective priority date (October 3, 2016), such that Piersol's electronic device would *detect a voice input* utilizing Meyers's "voice

activity detector" and/or local "wake word detector" discussed above, which in turn would "initiate [Piersol's] arbitration process to identify (e.g., determine) an electronic device for responding to the spoken instruction [] from user 800." Ex.1002, ¶238-48 (*citing* Ex.1003, ¶247).

First, Piersol and Meyers are in the same field of endeavor, relate to voice-enabled devices, and seek to address the same problem—namely, the duplicate response problem, where multiple devices detect the same user utterance. *See, e.g.*, Ex.1003, ¶3 (disclosing duplicate response problem "may result in a confusing experience for the user" and/or "multiple electronic devices may perform duplicative or conflicting operations"); Ex.1004, ¶14 (disclosing "techniques for avoiding such duplicate efforts and responses….").

Second, Piersol and Meyers have interrelated teachings, both disclosing very similar "arbitration" processes to solve the duplicate response problem. *See, e.g.*, Ex.1003, Title ("Intelligent device ***arbitration*** and control"), ¶274; Ex.1004, Abstract ("***[A]rbitration*** is employed to select one of the multiple speech interface devices to respond to the user utterance."), ¶66. Indeed, both references disclose arbitration techniques where electronic devices determine, receive, and compare signal-to-noise ratio (SNR) information associated with a detected voice input and select the electronic device with the highest SNR to respond to the user. *See, e.g.*, Ex.1003, ¶248, 256; Ex.1004, ¶44, 70. A POSITA evaluating Piersol would have

been motivated to seek out references that disclose similar arbitration techniques, like Meyers, to obtain additional details regarding such techniques. Ex.1002, ¶241.

Third, Piersol already discloses *detect[ing] a voice input* but does not provide specific implementation details. Thus, a POSITA evaluating Piersol would have been motivated to seek out references, like Meyers, that disclose implementation details for this, especially when implementing a system based on Piersol's teachings. *Id*., ¶242.

Fourth, Piersol and Meyers both disclose techniques where arbitration is initiated based on detecting a voice input, and Meyers's methods of performing this function via VAD and/or hotword/wakeword detection were well known before the effective priority date. *Id.*, ¶243-44; Ex.1013, ¶21 (disclosing a "speech detection module 108" that utilizes "voice activity detection (VAD) techniques" and a "speech processing module 110" configured to "detect a keyword in the speech"), ¶54 (disclosing "an automatic speech recognition engine" separate and distinct from "speech processing module 110" and responsible for recognizing speech).

Fifth, a POSITA would appreciate that applying Meyer's teachings to Piersol would have simply involved applying known methods to similar electronic devices to achieve predictable results. Ex.1002, ¶245-47. Indeed, the '311 specification provides no suggestion to a POSITA that *detect[ing] a voice input* under Google's narrow interpretation was anything more than a design choice among other choices

already in the prior art. *Id.*, ¶246. Thus, a POSITA would have had a reasonable expectation of success, especially because Meyers already discloses using techniques for *detecting a voice input* in performing arbitration similar to Piersol. *Id.*, ¶247.

B. Dependent Claim 12

1. Meyers Discloses Element [12.1]

Claim 12 depends on claim 10 and further recites **[12.1]** "*communicating to the other devices of the plurality of electronic devices to forgo responding to the detected output.*"[16] Meyers discloses this additional element. Ex.1002, ¶251-56.

For instance, Meyers discloses that "one or more of the devices 102 may include or provide the speech service 112" (Ex.1004, ¶30), and when a "first device" wins arbitration, the "speech service 112" of the "first device" may perform "action 216" (*see id.*, ¶57) that involves sending a "message" to an arbitration loser "informing the device [] not to expect a further response from the speech service 112," thereby "***canceling any response*** that might otherwise have been provided…." *Id.*, ¶67. Meyers explains that such a "message" may further cause the arbitration loser to enter a "listening mode" and/or "take some… action indicating that the device is not going to respond to the user request." *Id.*, ¶53.

---

[16] For purposes of this Petition, Sonos assumes *the detected* ***output*** refers to *the detected* ***input*** in claim 10.

2. Motivation to Combine

For the reasons below, a POSITA would have been motivated to combine Piersol with Meyers before the effective priority date, such that Piersol's electronic device (alone or as modified by Meyer's teachings regarding **[10.1]**) that won arbitration would be configured to communicate to other electronic devices that lost arbitration to forgo responding to detected voice input, as taught by Meyers. Ex.1002, ¶257-65.

First, a POSITA would have been motivated to combine Piersol with Meyers for the same reasons discussed above. *See supra* §XI.A.2.

Second, a POSITA would have been motivated to incorporate Meyers's cancellation message when implementing a system in accordance with Piersol's teachings on Wi-Fi networks (*see, e.g.*, Ex.1003, ¶247) because such networks were known for message/packet loss, whereby transmitted messages/packets did not reach their intended recipient. Ex.1002, ¶260 (*citing* Ex.1014, p.1; Ex.1015, p.1).

In this regard, Piersol's arbitration winner would communicate to electronic devices that lost arbitration to forgo responding to detected voice input (e.g., by using Meyers's cancellation message) to account for such known problems, which may prevent the arbitration winner's score/values from reaching the other electronic devices in the first instance. *Id.*, ¶261. In this way, even if the winner's score/values was not received by the other electronic devices (which would mean an arbitration

loser would mistakenly believe it won arbitration), the cancellation message would ensure Piersol's stated objectives of eliminating user confusion and excessive power consumption resulting from "redundant responses" (Ex.1003, ¶3, 33) are still achieved. Ex.1002, ¶261.

Third, a POSITA would have appreciated that applying Meyers' teachings of **[12.1]** to Piersol would have simply involved applying known methods to similar electronic devices to achieve predictable results. *Id*., ¶262-64. Indeed, the '311 specification provides no suggestion to a POSITA that *communicating to the other devices of the plurality of electronic devices to forgo responding to the detected [input]* was anything more than a design choice already in the prior art. *Id*., ¶263. Thus, a POSITA would have had a reasonable expectation of success, especially because Meyers already discloses using such techniques in performing arbitration similar to Piersol. *Id*., ¶264.

C. Dependent Claim 18

1. Meyers Discloses Element [18.1]

Like Piersol, Meyers discloses a quality score comprising *a confidence level of detection of the voice input*. Ex.1002, ¶268-71.

For instance, Meyers discloses that each "speech interface device" is configured such that, upon detecting a voice signal, it determines "metadata" associated with the detected voice signal that includes "one or more signal

attributes," including (i) "the level of voice presence detected in the audio signal" and (ii) "the confidence level with which a wakeword was detected in the audio signal," among other examples. Ex.1004, ¶40, 70, 108.

Specifically, Meyers describes "the level of voice presence detected in the audio signal" attribute as a likelihood of detection of speech. *Id.*, ¶102 ("The score is used as ***an indication of the detected or likely level of speech presence in the audio signal***."), ¶108 ("[T]he VAD 506 may produce a voice presence level, ***indicating the likelihood a person is speaking*** in the vicinity of the device 102.").

With regard to "the confidence level with which a wakeword was detected in the audio signal"—or "the level of confidence with which a wakeword or other trigger expression was detected" (*id.*, ¶70)—attribute, Meyers explains "[t]he wakeword [detector] may produce ***a wakeword confidence level***, corresponding to ***the likelihood that the user 104 has uttered the wakeword***." *Id.*, ¶108; *see also id.*, ¶27, 106, 119. In this way, Meyers also describes this attribute as a likelihood of detection of speech. Ex.1002, ¶270.

Thus, each of the foregoing likelihood attributes satisfies *a confidence level of detection of [a] voice input* under Google's broad interpretation. *Id.*, ¶271.

2. Motivation to Combine

For the following reasons, a POSITA would have been motivated to combine Piersol with Meyers before the effective priority date to achieve claim 18, such that

Piersol's electronic device (alone or as modified by Meyer's teachings regarding **[16.1]**) would determine and include one or both of Meyers' likelihood attributes as part of Piersol's arbitration score/values used for arbitration. Ex.1002, ¶272.

First, a POSITA would have been motivated to combine Piersol with Meyers for the same reasons discussed above. *See supra* §XI.A.2.

Second, Piersol itself explains that its arbitration score/values can take various forms and include a variety of information. *See* Ex.1003, ¶253, 261. Piersol already discloses that its arbitration score/values may include a "confidence value." *Supra* §X.B.9. In this way, Piersol itself provides a motivating reason to modify its arbitration score/values, and Piersol and Meyers make clear that use of a "confidence level" for a detected voice input was well known before the effective filing date. Ex.1002, ¶276. Thus, a POSITA would have appreciated that applying Meyers' teachings of **[18.1]** to Piersol would have simply involved applying known measures to similar electronic devices to achieve predictable results. *Id.*

Fourth, a POSITA would have had at least a reasonable expectation of success of modifying Piersol with Meyers' teachings regarding confidence levels of detection of a voice input. Ex.1002, ¶277. Indeed, Meyers itself discloses using such attributes in its own arbitration process. This demonstrates that the proposed modification to Piersol was well within a POSITA's skill. *Id.*

Moreover, the '311 Patent merely informs a POSITA that *a confidence level of detection of the voice input* is but one design choice picked from the prior art to be used in a predictable way. *Id.*, ¶278; Ex.1001, 30:44-50.

XII. Ground 2A—Masahide (§102)

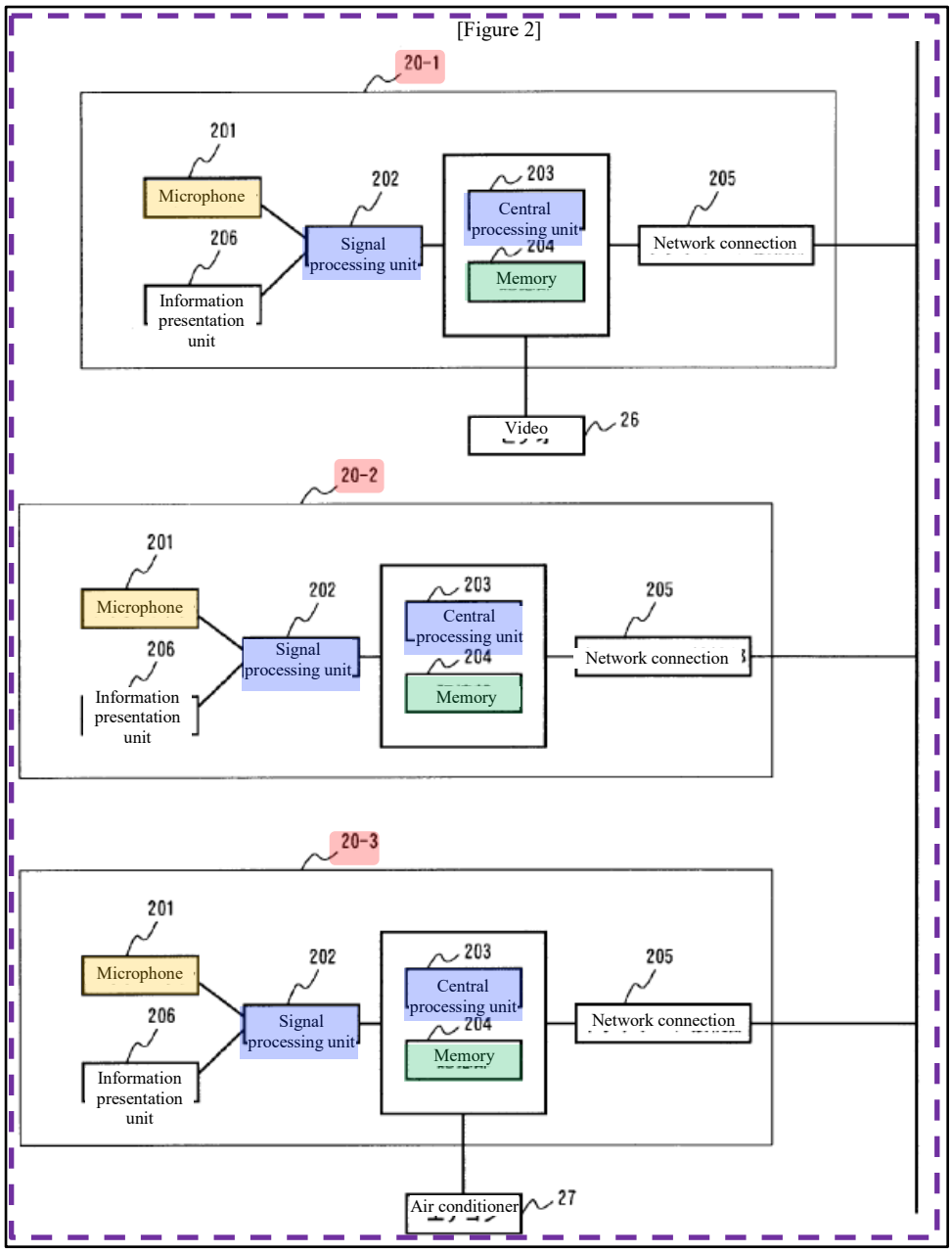
A. Masahide Anticipates Independent Claims 1, 10, 16

As explained, independent claim 16 is representative of the independent claims. *Supra* §§X.A.2.-X.A.3. Thus, Masahide invalidates claims 1 and 10 for the reasons Masahide invalidates representative claim 16, discussed below.

1. [16.0] *A system comprising a plurality of electronic devices, each of the electronic devices including one or more microphones, one or more processors, and memory storing one or more programs for execution by the one or more processors, each of the electronic devices programmed to:*

To the extent that the preamble is limiting, Masahide discloses it. Ex.1002, ¶283-87.

For instance, in connection with FIG. 2 below, Masahide discloses a voice input system that includes "voice input devices" that "are connected to network 21," where "[e]ach voice input device… consists of a microphone 201, a signal processing unit 202, a central processing unit 203, a storage unit 204…." Ex.1006, ¶13-14.



Id., FIG. 2 (annotated).

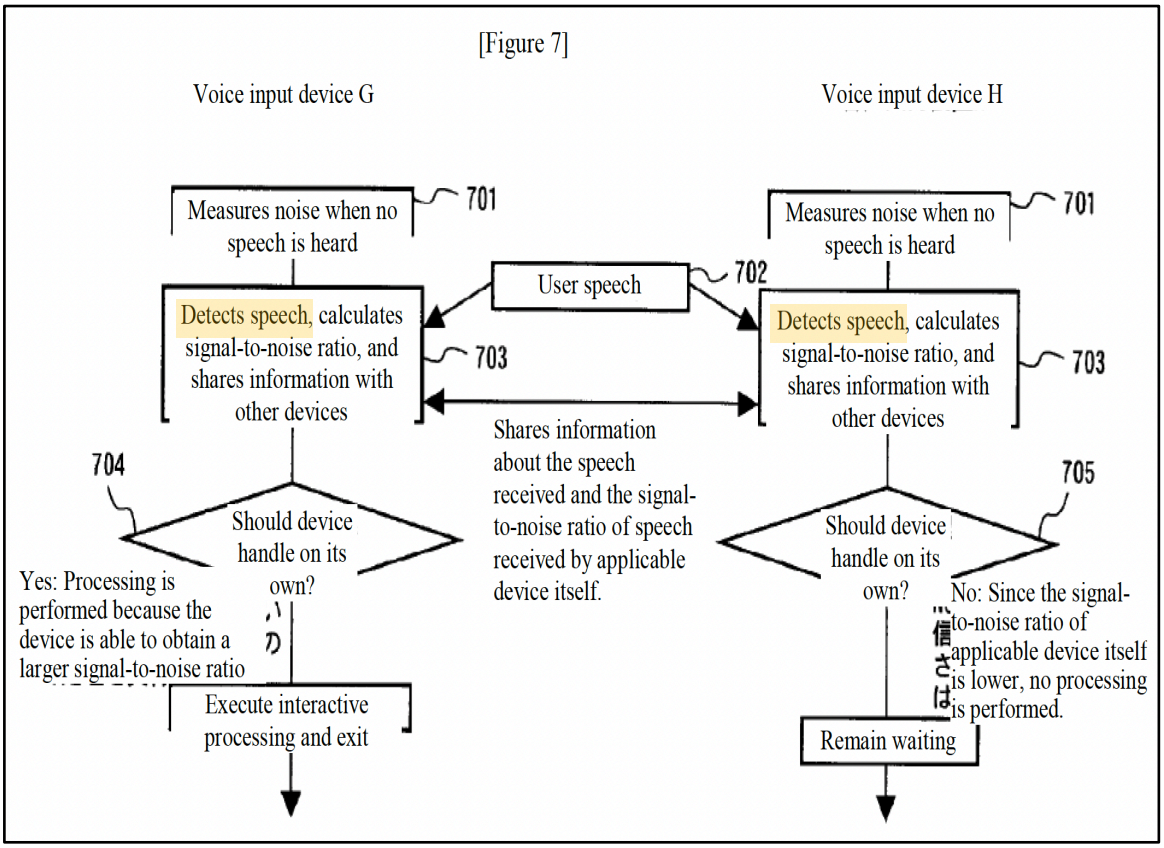
Masahide explains that "[c]entral processing unit 203 controls the state of the voice input device and sends ***instructions*** to each processing mechanism as necessary," and "[t]he content to be controlled can be determined ***based on information from*** signal processing unit 202, network connection unit 205 and

memory unit 204." *Id.*, ¶21; *see also id.*, ¶28. A POSITA would understand that Masahide's memory/storage unit 204 includes one or more programs for execution by one or more of Masahide's processing units to perform the functions disclosed in Masahide. Ex.1002, ¶286.

2. [16.1] *detect a voice input*

Masahide discloses this element Ex.1002, ¶288-96.[17] For instance, Masahide discloses that each voice input device is configured to "***detect[]*** voice from the user captured by microphone 201 at signal processing unit 202…." Ex.1006, ¶55; *see also, e.g.*, *id.*, ¶30 ("[W]hen a user speaks to voice input device A and voice input device B (step 301), ***voice is detected*** from a user captured by microphone 201 and signal is processed in signal processing unit 202 …."), ¶66, 73, 88, 114. This is illustrated in several figures, including FIG. 7 below:

[17] Because Google's interpretation of this element (*supra* §VII.A.) is ***narrower*** than its plain and ordinary meaning (*see* Ex.1002, ¶65-66), Masahide (and Jang) discloses this element even if Google's interpretation is rejected.

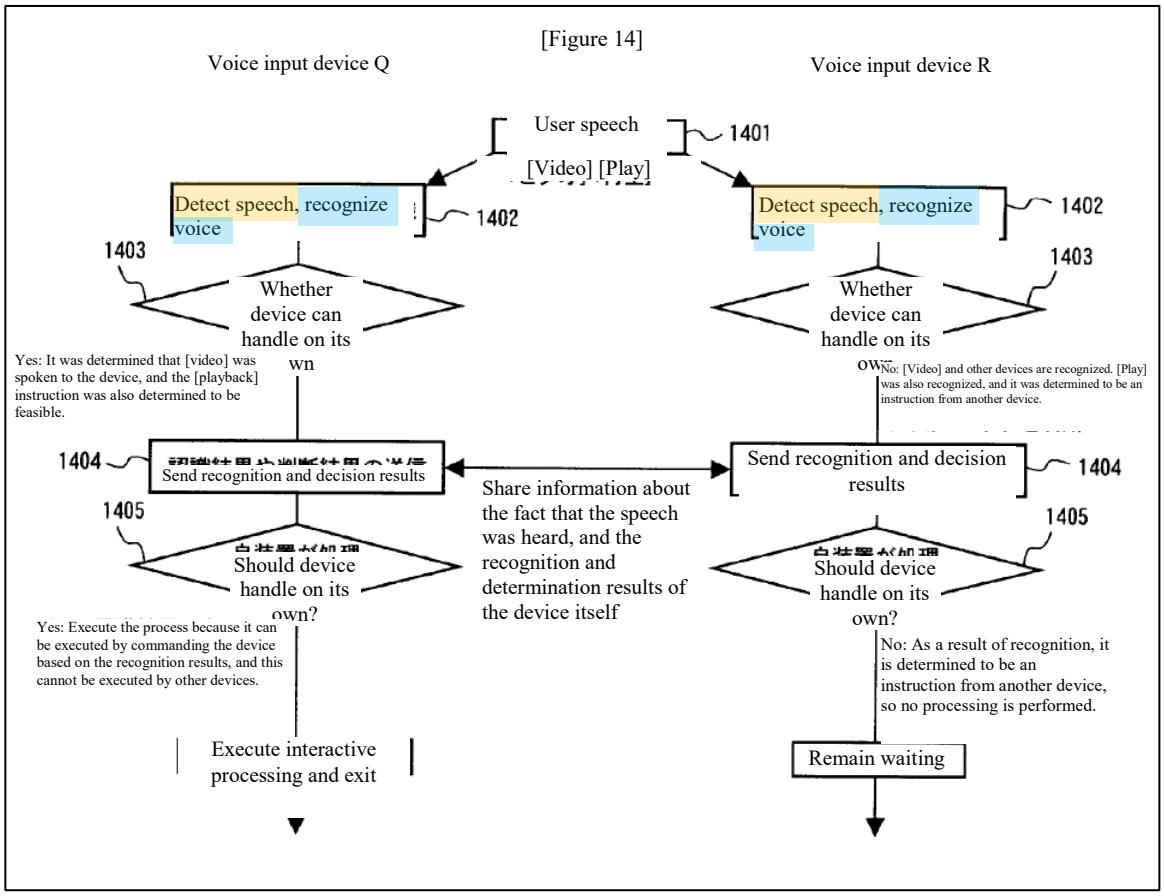


Id., FIG. 7 (annotated); *see also id.*, FIGs. 3-6, 8-10, 12, 14.

Masahide repeatedly states that each voice input device's ability to "detect" voice is separate and distinct from its ability to "recognize"[18] voice. *See, e.g.*, *id.*, ¶66 ("In the first instance, when the user says [play] as associated with voice input device K (step 901[])], each voice input device ***detects and recognizes*** the voice (step 902).") (brackets around "play" original), ¶88 ("When the user says 'video'

[18] Masahide explains that each voice input device recognizes voice by performing "speech recognition" (or "voice recognition") using various methods, "such as HMM and DP matching using mixed normal distribution as models," and the voice input devices in the system might have a "common" "vocabulary" or different vocabularies. *Id.*, ¶64, 87-91.

'playback' as shown in the flow diagram in Figure 14 (step 1401), voice input device Q and voice input device R ***detect and recognize*** voice (step 1402)."). This is also illustrated in several figures, including FIG. 14 below:



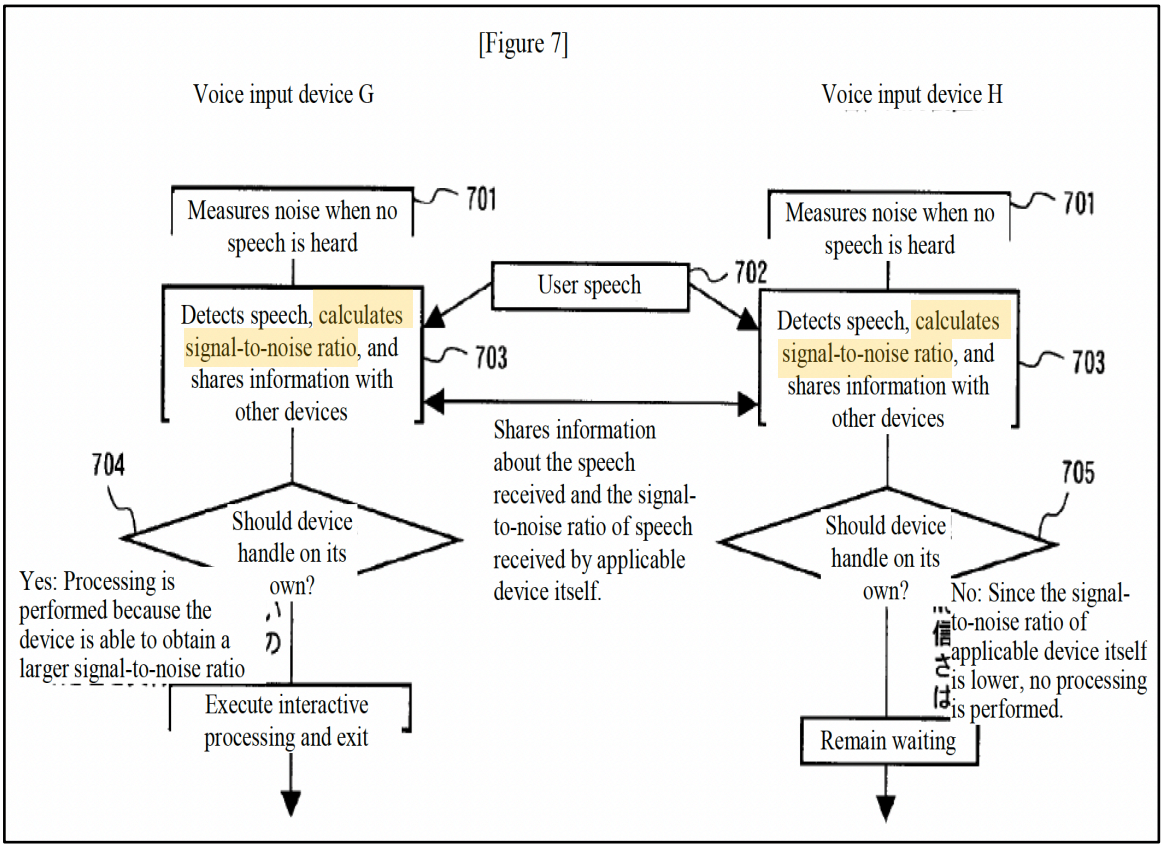
Id., FIG. 14 (annotated); *see also id.*, FIGs. 9, 12.

While Masahide discloses "an example of the present invention mainly describes the human voice of a user," Masahide explains the "sound of machinery and animals, may, depending on the purpose, be acceptable." *Id.*, ¶12; *see also id.*, ¶70, 76. Nevertheless, Google's interpretation of **[16.1]** expressly encompasses voice activity detection (VAD), and a POSITA would understand that VAD often

involved detecting sounds other than human voice. Ex.1002, ¶295 (*citing* Exs. 1016-1018).

3. [16.2] *determine a first value associated with the detected voice input*

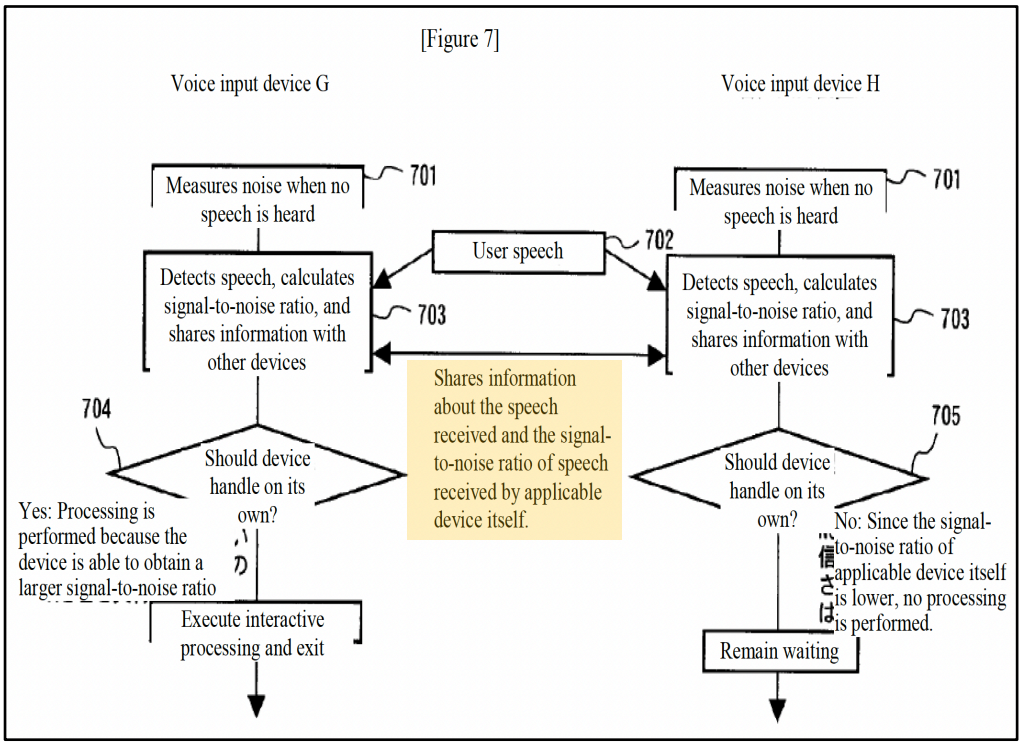
Masahide discloses this element. Ex.1002, ¶297-300. For instance, Masahide discloses that each voice input device is configured to determine (or calculate) "determinative information," such as "signal-to-noise ratio" (SNR) information, that is associated with the detected voice input. *See, e.g.*, Ex.1006, ¶4-6, 51-57, 116, FIG. 7, cls. 1, 3-6. This is illustrated in FIG. 7 below:



Id., FIG. 7 (annotated).

4. [16.3] *receive respective second values from other electronic devices of the plurality of electronic devices*

Masahide discloses this element. Ex.1002, ¶301-304. For instance, Masahide discloses that each voice input device is configured to receive "determinative information," such as SNR information, from one or more other voice input devices. *See, e.g.*, Ex.1006, ¶25 ("Network connection unit 205 is a mechanism for sending and ***receiving information between voice input devices*** through network 21…."), ¶55 ("[E]ach voice input device… calculates ***the signal-to-noise ratio*** based on the noise information when the user's voice is captured from the microphone, and ***communicates this to other voice input devices*** on the network (step 703)."). This is illustrated in FIG. 7 below, where "voice input device G" receives SNR information from "voice input device H":

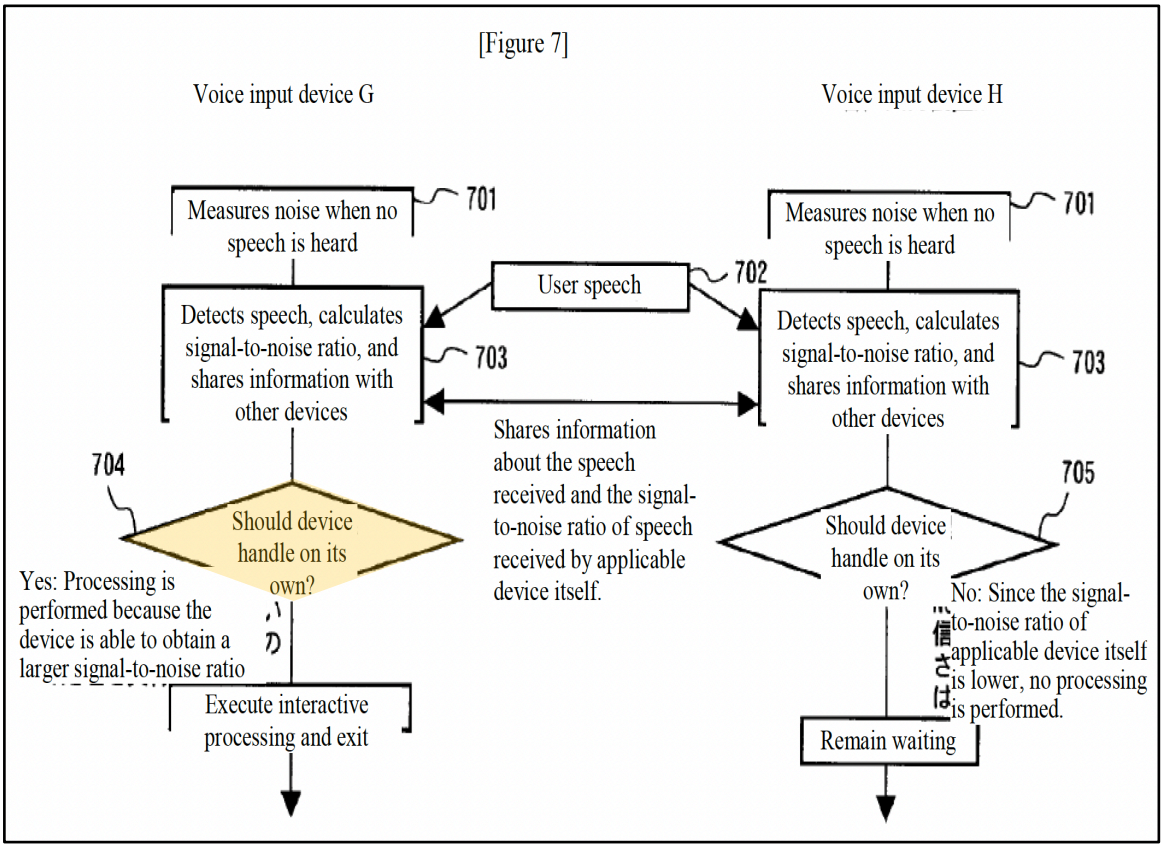


Id., FIG. 7 (annotated).

5. [16.4] *compare the first value and the respective second values*

Masahide discloses this element. Ex.1002, ¶305-308. For instance, Masahide discloses embodiments where each voice input device is configured to compare its own "determinative information" (e.g., SNR) with "determinative information" received from another voice input device. *See, e.g.*, Ex.1006, ¶56 ("Next, the signal-to-noise ratio information from other voice input devices that detected sound ***is compared*** with the signal-to-noise ratio information of the voice input device….").

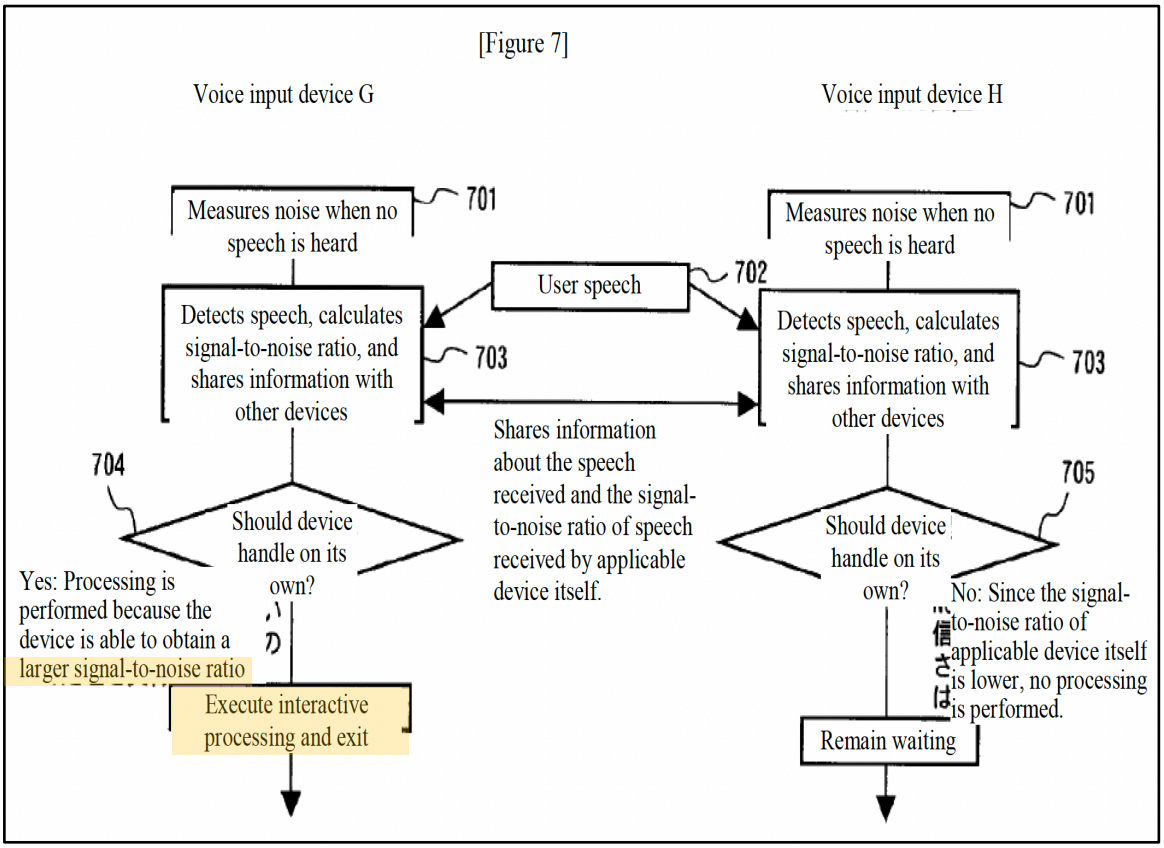
This is illustrated in FIG. 7 below:



Id., FIG. 7 (annotated).

6. [16.5] *in accordance with a determination that the first value is higher than the respective second values, respond to the voice input*

Masahide discloses this element. Ex.1002, ¶309-16. For instance, Masahide discloses that each voice input device is configured to respond to voice input in accordance with a determination that its "determinative information" (e.g., SNR) is higher than the "determinative information" from another voice input device. *See, e.g.*, Ex.1006, ¶56 ("[T]he signal-to-noise ratio information from other voice input devices that detected sound is compared with the signal-to-noise ratio information of the voice input device, and if the voice input device is ***the largest***, the ***voice is processed*** (step 704))."). This is illustrated in FIG. 7 below, where "voice input device G" responds to the voice input by "execut[ing] interactive processing" because it had a "larger" SNR than "voice input device H":



Id., FIG. 7 (annotated).

Masahide discloses to a POSITA that the reason for evaluating "determinative information" (e.g., SNR) is to determine "processing and execution ***of [the] voice information***" (*id.*, ¶4). Ex.1002, ¶312. In this regard, Masahide discloses that, once a voice input device determines to handle the detected voice input on its own, the device "execute[s] interactive processing" and exits the routine. Ex.1006, FIG. 3-10.

Specifically, Masahide introduces the "execute interactive processing and exit" language for the first time in connection with FIG. 3, where Masahide describes that "central processing unit 202 of voice input device B ***processes the voice***

captured according to the function of the voice input device, ***operates the device according to the content of the voice***, and waits again after the interaction ends (step 304)." *Id.*, ¶31. Subsequently, FIGs. 4-10 each include a step with the same "execute interactive processing and exit" language. A POSITA would understand that Masahide intended the same words from step 304 to have the same meaning in FIGs. 4-10. Ex.1002, ¶314.

Moreover, Masahide provides examples of a voice input device responding to a voice input, including starting playback of a video and/or outputting a synthetic audible voice "playback started" or "starting playback." Ex.1006, ¶89-90, 114-19. FIG. 13 further identifies examples of how a voice input device may respond to a voice input:

[Figure 13]

Recognition vocabulary	Target device	Processing details
[Video]	Video	Specify video
[Air conditioner]	Air conditioner	Specify air conditioner
[Power ON]	Shared Instructions	Turn on the target device
[Power OFF]	Shared Instructions	Turn off the target device
[Play]	Video	Play video
[Stop]	Video	Stop playing video
[Turn up temperature]	Air conditioner	Operates by increasing the temperature setting of the air conditioner
[Turn down temperature]	Air conditioner	Operation by lowering the air conditioner temperature setting

Id., FIG. 13 (annotated); *infra* §XII.B.3.

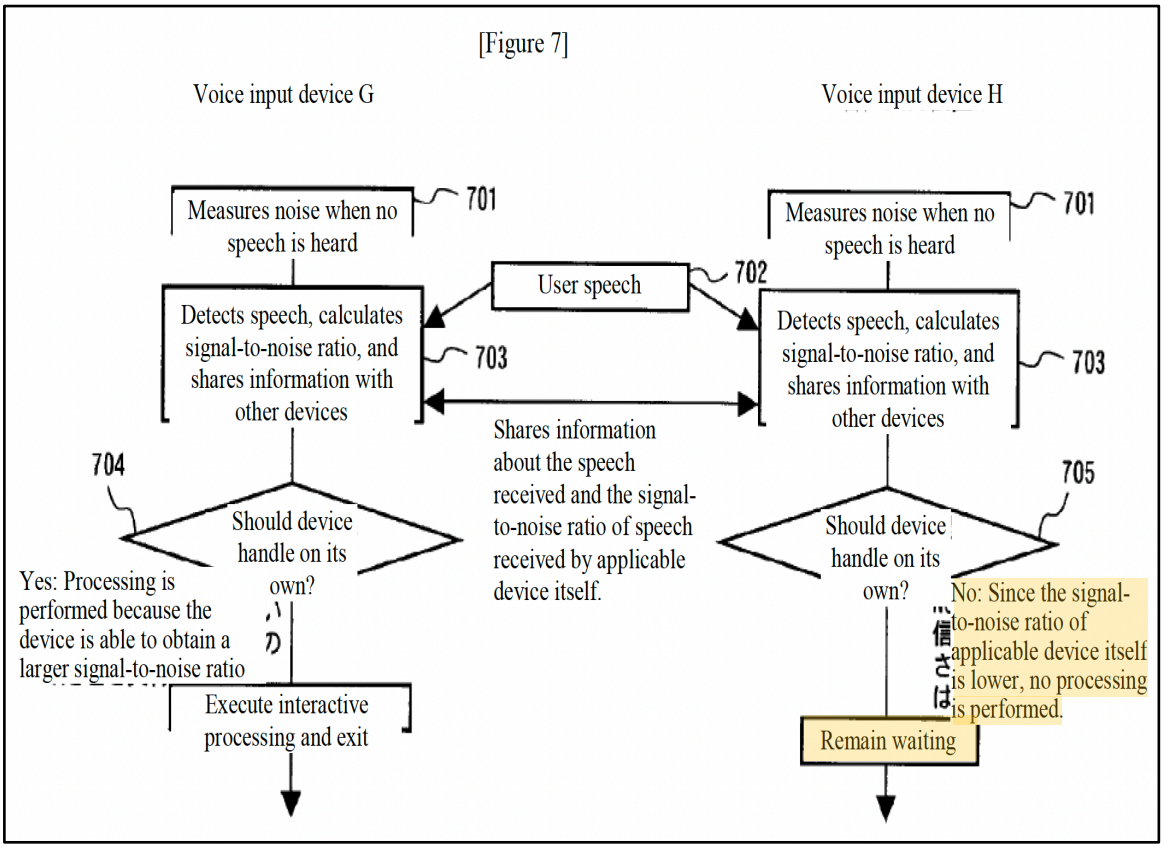
B. Masahide Invalidates Dependent Claims 2-3, 8-9, 11, 14, 17, 19-20

1. Dependent Claim 2

Claim 2 depends on claim 1 and further recites **[2.1]** "*in accordance with a determination that the first value is not higher than the second scores, forgoing responding to the detected input.*"[19] Masahide discloses this additional element. Ex.1002, ¶319-22.

For instance, Masahide discloses that each voice input device is configured such that, upon determining that its "determinative information" (e.g., SNR) is ***not*** higher than "determinative information" received from another electronic device, it forgoes responding to the detected voice input. *See, e.g.*, Ex.1006, ¶56 ("[T]he signal-to-noise ratio information from other voice input devices that detected sound is compared with the signal-to-noise ratio information of the voice input device, and if the voice input device is the largest, the voice is processed (step 704). ***Otherwise***, a decision is made that ***the voice is not to be processed*** (step 705)."); *see also, e.g.*, *id.*, ¶47, 51. This is illustrated in FIG. 7 below:

[19] This Petition assumes "*the second scores*" (which lacks antecedent basis) refers to "*the second values*" recited in claim 1.



Id., FIG. 7 (annotated).

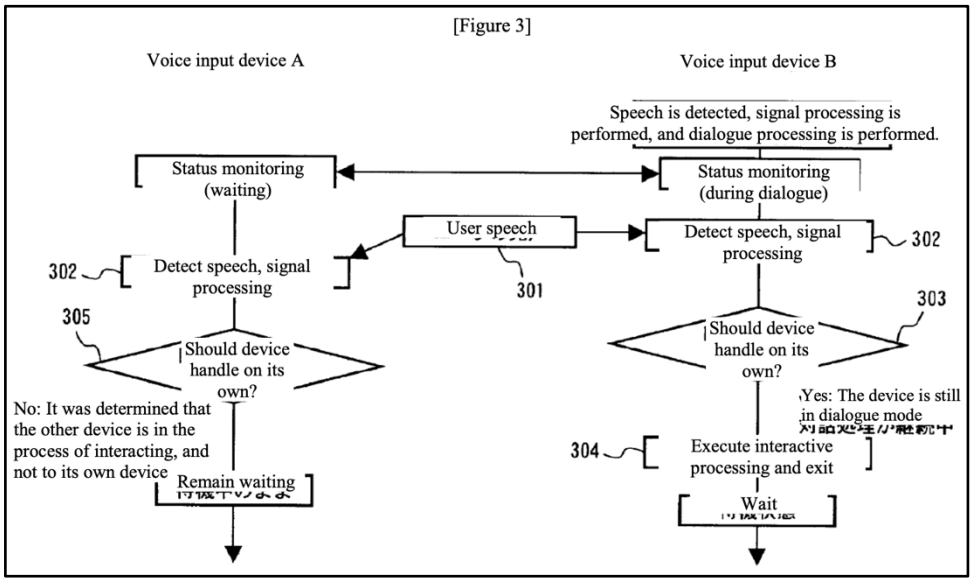
2. Dependent Claim 3

Claim 3 depends on claim 2 and further recites **[3.1]** "*forgoing responding to the detected input further comprises changing a state of operation of the first electronic device.*" Masahide discloses this additional element (and/or alone renders this additional element obvious). Ex.1002, ¶323-30.

To start, Masahide discloses multiple arbitration approaches, each involving different types of "determinative information" for making an arbitration decision. *See, e.g.*, Ex.1006, ¶29-34, FIG. 3 (disclosing arbitration based on processing state), ¶44-48, FIG. 5 (disclosing arbitration based on voice detection time), ¶49-51, FIG.

6 (disclosing arbitration based on volume of detected voice input), ¶52-57, FIG. 7 (disclosing arbitration based on SNR of detected voice input). Masahide also discloses that multiple, different types of information can be used for making the arbitration decision (*id.*, ¶104-107), such as "processing status, processing performance, recognizable vocabulary, and processing details of other voice processing systems" "in addition to information such as detection and recognition results when voice is detected" (e.g., voice detection time, volume, SNR, etc.). *Id.*, ¶104.

In particular, in connection with FIG. 3 showing voice input devices exchanging respective status information, Masahide explains, "if voice input device B is in the process of interacting with the user ***and other systems are not in an interactive state***, voice input device B makes a choice to process the content spoken by the user." *Id.*, ¶31.



From the foregoing disclosures, a POSITA would understand that Masahide discloses (or at least renders obvious) a scenario where Masahide's voice input devices consider ***both*** "processing status" (where multiple voice input devices are in "an interactive state" or "dialogue mode") and a detection result like SNR, such that a voice input device that loses arbitration is configured to forgo responding to the detected voice input by "***enter[ing]*** a waiting state" (*see id.*, ¶32), thereby changing its mode of operation from the "interactive state" (or "dialogue mode") to the "waiting state." Ex.1002, ¶328-29.

Indeed, a POSITA would appreciate that, in this scenario, "processing status" alone would not control the arbitration decision, and other "determinative information" (e.g., SNR) would also need to be evaluated to make the decision. *Id*., ¶328. Otherwise, multiple voice input devices with the same "processing status" would respond to the detected voice input, which would undesirably cause user confusion and/or redundant interactive processing. *Id*. (*citing* Ex.1006, ¶1-3, 11, 51, 57, 116).

3. Dependent Claim 8

Claim 8 depends on claim 1 and further recites **[8.1]** "*responding to the detected input comprises outputting an audible and/or a visual response to the detected voice input.*" Masahide discloses this additional element. Ex.1002, ¶331-33.

For instance, Masahide discloses that voice input devices are configured such that responding to a detected voice input can involve providing a "visual response[]" and/or audible output. *See, e.g.*, Ex.1006, ¶35 ("[T]he dialogue here is not limited to ***one-on-one voice interactions between humans and systems***, but rather may include the return of unilateral speech from humans to systems, ***visual responses*** from the system side, or ***responses from the system to any person***, and the same applies to the dialogue used in the following explanations."), ¶113 ("The information presentation unit of ***each voice input device*** in this example is ***equipped with a speaker***, and it is possible to ***return the voice of any text synthesized*** by the central processing unit and the signal-processing unit to the user."), ¶119 ("[A] standalone voice input device can convey to the user a message of ***'started playing' with a synthetic voice***."); *see also id.*, ¶17; *supra* §XII.A.6.

4. Dependent Claim 9

Claim 9 depends on claim 1 and further recites **[9.1]** "*the first electronic device further comprises a speaker, and* **[9.2]** *responding to the detected input further comprises outputting an audible response.*" Masahide discloses these additional elements. Ex.1002, ¶334-37.

For instance, as just discussed, Masahide discloses that voice input devices are configured such that responding to a detected voice input can involve providing

an audible output. *Supra* §XII.B.3. Masahide discloses that voice input devices can comprise a "speaker" to enable outputting such responses. *See* Ex.1006, ¶113.

5. Dependent Claim 11

Claim 11 depends on independent claim 10 and further recites **[11.1]** "*recognizing a command in the voice input; and* **[11.2]** *in accordance with a determination that a type of the command is related to the first electronic device, responding to the detected voice input.*" Mashide discloses these additional elements. Ex.1002, ¶338-46.

First, Masahide discloses **[11.1]**. For instance, Masahide discloses that each voice input device is configured to perform "speech recognition" (or "voice recognition") to recognize voice commands/instructions. Ex.1006, ¶63-64; *id.*, ¶82 ("[I]f the user says [video] [power ON] (step 1201), voice input device O and voice input device P ***recognize the device name and device instructions***… and ***determine whether they are instructions to their system***…."), ¶8, 43, 65-69, 79-91, 114-18, FIGs. 9, 12, 13, 14.

Second, Masahide discloses **[11.2]**. For instance, Masahide discloses an example where multiple voice input devices "***recognize the [same] device name and device instructions***" (e.g., "[video] [power ON]"), and "***determine whether they are instructions to their system***…." *Id.*, ¶82. Based on this determination, Masahide discloses that only one device responds. *See id.*, ¶83 ("[T]hese results and results

from other voice input devices can be used to ***determine if the voice input device should process these***…."), FIG. 12.

Masahide discloses another example in connection with "Figure 13 [that] represents the concept of linking [a] recognition word to the target device and processing content." *Id.*, ¶87. Specifically, Masahide discloses an example where multiple voice input devices recognize the same command(s) in a user's utterance (e.g., "[video] and [playback]") and only the "Video" "target device" responds when it determines the command is "an instruction ***for playback of the video.***" *Id.*, ¶88-90, FIGs. 13-14.

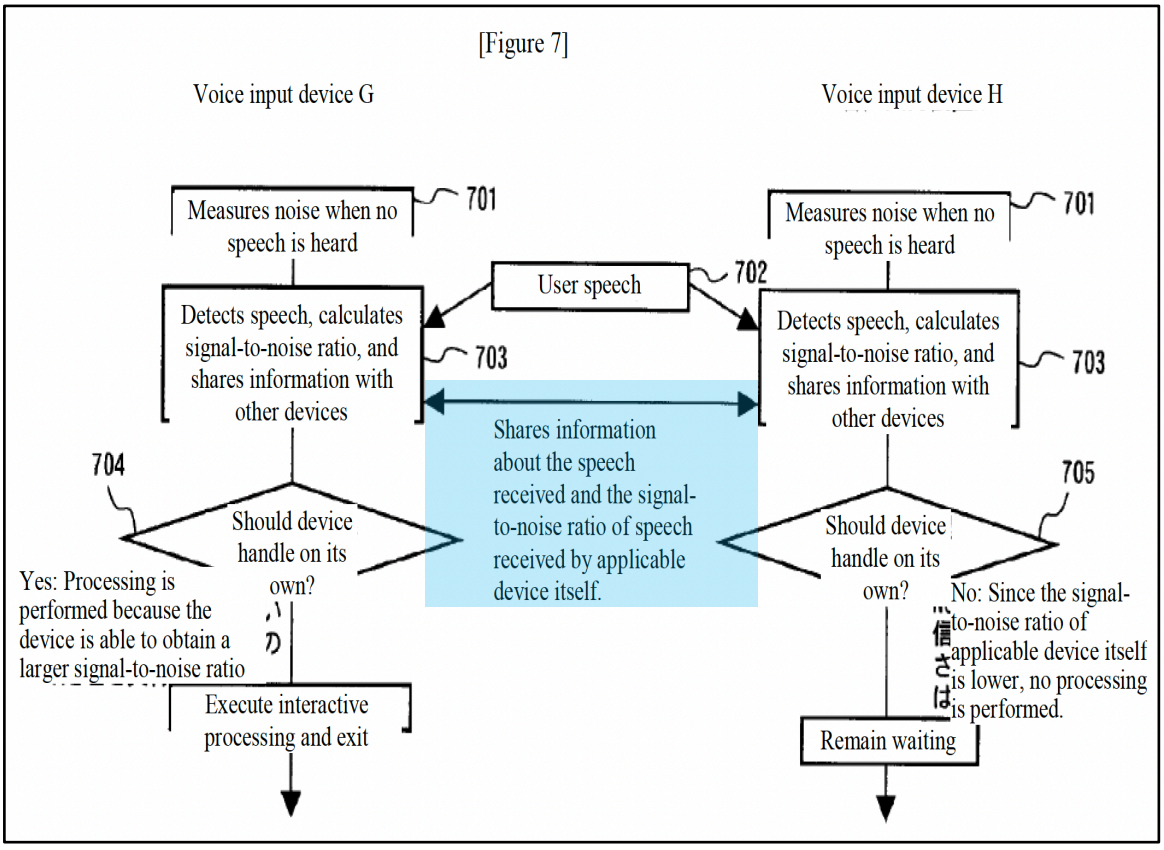
Masahide further discloses that its arbitration process of determining whether to respond can look at recognized voice commands and/or other "determinative information" (e.g., SNR). *See, e.g.*, *id.*, cls. 1, 8, ¶104-107.

6. Dependent Claim 14

Claim 14 depends on claim 10 and further recites **[14.1]** "*the plurality of electronic devices is communicatively coupled through a local network; and* **[14.2]** *the receiving is performed through the local network.*" Masahide discloses these additional elements. Ex.1002, ¶347-53.

For instance, Masahide's voice input devices are communicatively coupled through a "***LAN*** and wireless technology such as ***Bluetooth***." Ex.1006, ¶25; *see also id.*, ¶8, 14. Masahide discloses that each voice input device is configured to

receive "determinative information" (e.g., SNR) from another voice input device through such local network. *See, e.g.*, *id.*, ¶25 ("Network connection unit 205 is a mechanism for ***sending and receiving information between voice input devices*** through ***network*** 21, and can be achieved by… network connection over LAN and wireless technology such as Bluetooth."), ¶55 ("[E]ach voice input device … calculates the ***signal-to-noise ratio*** based on the noise information … and ***communicates this to other voice input devices on the network*** (step 703)."). This is illustrated in FIG. 7 below:



Id., FIG. 7 (annotated); *see also, e.g.*, *id.*, FIGs. 1, 11, 15-19.

7. Dependent Claim 17

Claim 17 depends on independent claim 16 and further recites that **[17.1]** "*the first value comprises a first quality score of the voice input.*" Masahide discloses this additional element. Ex.1002, ¶354-57. For instance, as explained, Masahide discloses that its "determinative information" associated with a detected voice input can include SNR information, which constitutes a "*quality score*." *Supra* §XII.A.3.; *supra* n.15.

8. Dependent Claim 19

Claim 19 depends on claim 17 and further recites **[19.1]** "*the first quality score comprises a signal-to-noise rating of detection of the voice input.*" Masahide discloses this additional element for the same reasons it discloses **[17.1]**. Ex.1002, ¶358-59; *supra* §XII.B.7.

9. Dependent Claim 20

Claim 20 depends on claim 16 and further recites **[20.1]** "*each of the respective second values is indicative of a quality of the voice input detected by a respective one of the other electronic devices.*" Masahide discloses this additional element for the same reasons it discloses **[17.1]**. Ex.1002, ¶360-63; *supra* §XII.B.7.

XIII. Ground 2B: Masahide + Jang (§103)

Ground 2B is presented to the extent Google argues that Masahide does not adequately disclose **[1.1]/[10.1]/[16.1]** *detect[ing] a voice input*. Ground 2B is otherwise the same as Ground 2A.

A. Jang Discloses [1.1]/[10.1]/[16.1]

Jang discloses this element. Ex.1002, ¶366-67. For instance, Jang discloses a system of "electronic devices" (*see, e.g.*, Ex.1007, ¶30, 53, FIGs. 1-2), each of which is configured to "detect" voice input:

> [T]he voice recognition unit 182 can recognize the input voice signals by ***detecting voice activity from the input voice signals***, carrying out sound analysis thereof, and recognizing the analysis result as a recognition unit.

Id., ¶110.

Accordingly, a POSITA would understand Jang discloses that each electronic device's "voice recognition unit 182" performs voice activity detection ***before*** performing speech/voice recognition. Ex.1002, ¶331; *see also* Ex.1007, ¶98-99 (describing Jang's "recognition dictionary"), ¶109-10.

B. Motivation to Combine

For the following reasons, a POSITA would have been motivated to combine Masahide with Jang before the ***earliest possible*** priority date of the '311 Patent (October 9, 2014) to achieve **[1.1]/[10.1]/[16.1]** (assuming Masahide does not already disclose this element), such that Masahide's voice input device would detect a voice input utilizing Jang's "voice recognition unit 182… detecting voice activity from the input voice signals" (Ex.1007, ¶110), which in turn would initiate Masahide's arbitration process to "solve[] the question of what to do with each voice input device when the voice of the user is input to a plurality of networked voice input devices" (Ex.1006, ¶11). *See* Ex.1002, ¶333-42.

First, Masahide and Jang are in the same field of endeavor, relate to voice-enabled devices, and seek to address the same duplicate response problem, where multiple devices detect the same user utterance. *See, e.g.*, Ex.1006, ¶11 ("The present invention solves the question of what to do with each voice input device when the voice of the user is input to a ***plurality*** of networked voice input devices."), Summary; Ex.1007, ¶119-22 ("[T]hat the plurality of devices simultaneously process the use's voice command may ***cause a multi process to be unnecessarily performed***…. In an environment involving a plurality of devices, it can be determined by communication between the devices or by a third device managing the plurality of devices ***which of the plurality of devices*** is to carry out a user's voice command.").

Second, Masahide and Jang have interrelated teachings, both disclosing similar arbitration processes to solve the duplicate response problem. *See, e.g.*, Ex.1006, FIGs. 6-7; Ex.1007, FIGs. 11-12.

Indeed, Masahide and Jang both disclose arbitration techniques where electronic devices are configured to determine, receive, and compare values associated with a detected voice input (e.g., magnitude (gain value) or volume of voice signal) and select the electronic device with the highest value to respond to the user. *See, e.g.*, Ex.1006, ¶49-51; Ex.1007, ¶122, 127-136, FIG. 11. A POSITA evaluating Masahide would have been motivated to seek out other references that

disclose similar arbitration techniques, like Jang, to obtain additional details regarding such techniques. Ex.1002, ¶372.

Third, Masahide already discloses *detect[ing] a voice input* (*supra* §XII.A.2.), but does not provide specific implementation details. Thus, a POSITA evaluating Masahide would have been motivated to seek out other references, like Jang, that disclose implementation details for this, especially when implementing a system based on Masahide's teachings. Ex.1002, ¶373.

Fourth, Masahide and Jang both disclose techniques where arbitration is initiated based on detecting a voice input, and Jang's methods of performing this function via VAD were well known before the ***earliest possible*** priority date (October 9, 2014). Ex.1007, ¶110; Ex.1013, ¶21 (disclosing a "speech detection module 108" that utilizes "voice activity detection (VAD) techniques" and a "speech processing module 110" configured to "detect a keyword in the speech"), ¶54 (disclosing "an automatic speech recognition engine" separate and distinct from "speech processing module 110" and responsible for recognizing speech).

Fifth, a POSITA would have appreciated that applying Jang's teachings to Masahide would have simply involved applying known methods to similar electronic devices to achieve predictable results. Ex.1002, ¶339-40. Indeed, the '311 specification provides no suggestion to a POSITA that *detect[ing] a voice input* under Google's interpretation was anything more than a design choice among other

choices already in the prior art. *Id.*, ¶376. Thus, a POSITA would have had a reasonable expectation of success, especially because Jang already discloses using techniques for *detecting a voice input* in performing arbitration in a manner similar to Masahide. *Id.*, ¶377.

XIV. Lack of Secondary Considerations

Sonos is unaware of any secondary considerations of non-obviousness. Even if they existed, they would not overcome the compelling *prima facie* showing of unpatentability. *See* Ex.1002, ¶379-80.

XV. Conclusion

For the foregoing reasons, Petitioner respectfully requests institution of IPR and cancellation of Challenged Claims 1-3, 8-12, 14-20.

Respectfully submitted,

LEE, SULLIVAN, SHEA & SMITH LLP

Dated: August 7, 2023

/Michael P. Boyea/

Michael P. Boyea, Reg. No. 70,248
LEE SULLIVAN SHEA & SMITH LLP
656 West Randolph Street, Suite 5W
Chicago, IL 60661
(312) 754-9602
boyea@ls3ip.com

Counsel for Petitioner Sonos, Inc.

CERTIFICATE OF WORD COUNT

The undersigned certifies that the foregoing PETITION FOR *INTER PARTES* REVIEW complies with the type volume limitation in 37 C.F.R. § 42.24(c)(1). According to the utilized word-processing system's word count, the petition—excluding the caption, table of contents, table of exhibits, mandatory notices, certificate of word count, and certificate of service—contains 13,823 words.

Dated: August 7, 2023

By: /Michael P. Boyea/

Michael P. Boyea, Reg. No. 70,248
LEE SULLIVAN SHEA & SMITH LLP
656 West Randolph Street, Suite 5W
Chicago, IL 60661
(312) 754-9602
boyea@ls3ip.com

Counsel for Petitioner Sonos, Inc.

CERTIFICATE OF SERVICE

The undersigned hereby confirms that the foregoing Petition for *Inter Partes* Review and associated exhibits were caused to be served on August 7, 2023 via overnight courier upon the following counsel of record for Patent Owner:

Marshall, Gerstein & Borun LLP (Google)
233 South Wacker Drive
6300 Willis Tower
Chicago, IL 60606-6357

Courtesy copies were also served via electronic mail on Patent Owner's counsel of record in a related district court litigation:

David A Nelson (davenelson@quinnemanuel.com)
Patrick D Curran (patrickcurran@quinnemanuel.com)
S. Alex Lasher (alexlasher@quinnemanuel.com)
Jeffrey S Gerchick (jeffgerchick@quinnemanuel.com)
Nina S. Tallon (ninatallon@quinnemanuel.com)

QUINN EMANUEL URQUHART & SULLIVAN, LLP

/Michael P. Boyea/

Michael P. Boyea, Reg. No. 70,248
LEE SULLIVAN SHEA & SMITH LLP
656 West Randolph Street, Suite 5W
Chicago, IL 60661
(312) 754-9602
boyea@ls3ip.com