(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
16 November 2017 (16.11.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/197312 A2

(51) International Patent Classification:
G10L 15/22 (2006.01)

(21) International Application Number:
PCT/US2017/032488

(22) International Filing Date:
12 May 2017 (12.05.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/335,981 13 May 2016 (13.05.2016) US
62/375,543 16 August 2016 (16.08.2016) US

(71) Applicant: **BOSE CORPORATION** [US/US]; The Mountain, MS 3B1, Framingham, Massachusetts 01701-9168 (US).

(72) Inventors: **DALEY, Michael J.**; c/o BOSE CORPORATION, The Mountain, MS 3B1, Framingham, Massachusetts 01701-9168 (US). **CRIST, David Rolland**; c/o BOSE CORPORATION, The Mountain, MS 3B1, Framingham, Massachusetts 01701-9168 (US). **BERARDI, William**; c/o BOSE CORPORATION, The Mountain, MS 3B1, Framingham, Massachusetts 01701-9168 (US).

(74) Agent: **HILL, Misha K.**; BOSE CORPORATION, The Mountain, MS 3B1, Framingham, Massachusetts 01701-9168 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) Title: PROCESSING SPEECH FROM DISTRIBUTED MICROPHONES

(57) Abstract: A plurality of microphones are positioned at different locations. A dispatch system in communication with the microphones derives a plurality of audio signals from the plurality of microphones, computes a confidence score for each derived audio signal, compares the computed confidence scores. Based on the comparison, the dispatch system selects at least one of the derived audio signals for further handling, receives a response to the further processing, and outputs the response using an output device. The output device does not correspond to the microphone that captured the selected audio signals.

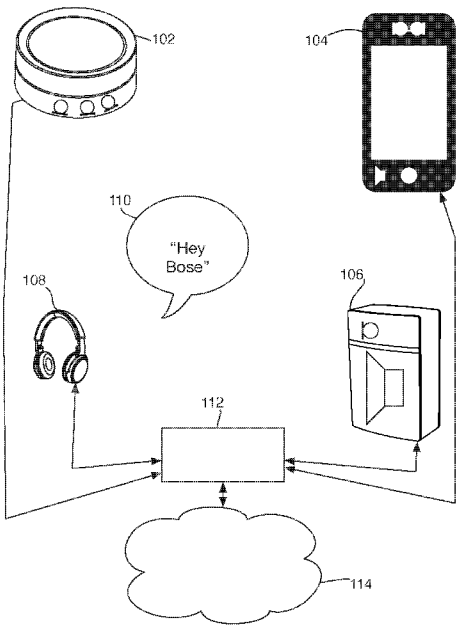

Fig. 1

WO 2017/197312 A2

[Continued on next page]

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

PROCESSING SPEECH FROM DISTRIBUTED MICROPHONES

CLAIM TO PRIORITY AND CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to provisional U.S. patent applications 62/335,981, filed May 13, 2016, and 62/375,543, filed August 16, 2016, the entire contents of which are incorporated here by reference. This application is related to U.S. patent application 15/373,541, filed December 9, 2016, the entire contents of which are incorporated here by reference.

BACKGROUND

[0002] This disclosure relates to processing speech from distributed microphones.

[0003] Current speech recognition systems assume one microphone or microphone array is listening to a user speak and taking action based on the speech. The action may include local speech recognition and response, cloud-based recognition and response, or a combination of these. In some cases, a "wake-up word" is identified locally, and further processing is provided remotely based on the wake-up word.

[0004] Distributed speaker systems may coordinate the playback of audio at multiple speakers, located around a home, so that the sound playback is synchronized between locations.

SUMMARY

[0005] In general, in one aspect, a system includes a plurality of microphones positioned at different locations, and a dispatch system in communication with the microphones. The dispatch system derives a plurality of audio signals from the plurality of microphones, computes a confidence score for each derived audio signal, and compares the computed confidence scores. Based on the comparison, the dispatch system selects at least one of the derived audio signals for further handling.

[0006] Implementations may include one or more of the following, in any combination. The dispatch system may include a plurality of local processors each connected to at least one of the microphones. The dispatch system may include at least a first local processor and at least a second processor available to the first processor over a network. Computing the confidence score for each derived audio signal may include computing a confidence in one or more of whether the signal may include speech, whether a wakeup word may be included in the signal, what wakeup word may be included in the signal, a quality of speech contained in the signal, an identity of a user whose voice may be recorded in the signal, and a location of the user relative to the microphone locations. Computing the confidence score for each derived audio signal may also include determining that the audio signal appears to contain an utterance and whether the utterance includes a wakeup word. Computing the confidence score for each derived audio signal may also include identifying which wakeup word from a plurality of wakeup words is included in the speech. Computing the confidence score for each derived audio signal further may include determining a degree of confidence that the speech includes the wakeup word.

[0007] Computing the confidence score for each derived audio signal may include comparing one or more of a timing between when the microphones detected sounds corresponding to each of the audio signals, signal strength of the derived audio signals, signal-to-noise ratio of the derived audio signals, spectral content of the derived audio signals, and reverberation within the derived audio signals. Computing the confidence score for each derived audio signal may include, for each audio signal, computing a distance between an apparent source of the audio signal and at least one of the microphones. Computing the confidence score for each derived audio signal may include computing a location of the source of each audio signal relative to the locations of the microphones. Computing the location of the source of each audio signal may include triangulating the location based on computed distances distance between each source and at least two of the microphones.

[0008] The dispatch system may transmit at least a portion of the selected signal or signals to a speech processing system to provide the further handling. Transmitting the selected audio signal or signals may include selecting at least one speech processing system from a plurality of speech processing systems. At least one speech processing system of the plurality of speech processing systems may include a speech recognition service provided over a wide-area network. At least one speech processing system of the plurality of speech processing systems may include a speech recognition process executing on the same processor on which the dispatch system is executing. The selection of the speech processing system may be based on one or more of preferences associated with a user, the computed confidence scores, orcontext in which the audio signals are derived. The context may include one or more of an identification of a user that may be speaking, which microphones of the plurality of microphones produced the selected derived audio signals, a location of the user relative to the microphone locations, operating state of other devices in the system, and time of day. The selection of the speech processing system may be based on resources available to the speech processing systems.

[0009] Comparing the computed confidence scores may include determining that at least two selected audio signals appear to contain utterances from at least two different users. The determining that the selected audio signals appear to contain utterances from at least two different users may be based on one or more of voice identification, location of the users relative to the locations of the microphones, which of the microphones produced each of the selected audio signals, use of different wakeup words in the two selected audio signals and visual identification of the users. The dispatch system may also send the selected audio signals corresponding to the two different users to two different selected speech processing systems. The selected audio signals may be assigned to the selected speech processing systems based on one or more of preferences of the users, load balancing of the speech processing systems, context of the selected audio signals, and use of different wakeup words in the two selected audio signals. The dispatch system may also send the selected audio signals

corresponding to the two different users to the same speech processing system as two separate processing requests.

[0010] Comparing the computed confidence scores may include determining that at least two received audio signals appear to represent the same utterance. The determining that the selected audio signals represent the same utterance may be based on one or more of voice identification, location of the source of the audio signals relative to the locations of the microphones, which of the microphones produced each of the selected audio signals, time of arrival of the audio signals, correlations between the audio signals or between outputs of microphone array elements, pattern matching, and visual identification of the person speaking. The dispatch system may also send only one of the audio signals appearing to represent the same utterance to the speech processing system. The dispatch system may also send both of the audio signals appearing to represent the same utterance to the speech processing system. The dispatch system may also transmit at least one selected audio signal to each of at least two speech processing systems, receive responses from each of the speech processing systems, and determine an order in which to output the responses.

[0011] The dispatch system may also transmit at least two selected audio signals to at least one speech processing system, receive responses from the speech processing system corresponding to each of the transmitted signals, and determine an order in which to output the responses. The dispatch system may be further configured to receive a response to the further processing, and output the response using an output device. The output device may not correspond to the microphone that captured the audio. The output device may not be located at any of the locations where the microphones are located. The output device may include one or more of a loudspeaker, headphones, a wearable audio device, a display, a video screen, or an appliance. Upon receiving multiple responses to the further processing, the dispatch system may determine an order in which to output the responses by combining the responses into a single output. Upon receiving multiple responses to the further processing, the dispatch system may determine an order in which to output the responses by selecting fewer than all

of the responses to output, or sending different responses to different output devices.. The number of derived audio signals may be not equal to the number of microphones. At least one of the microphones may include a microphone array. The system may also include non-audio input devices. The non-audio input devices may include one or more of accelerometers, presence detectors, cameras, wearable sensors, or user interface devices.

[0012] In general, in one aspect, a system includes a plurality of devices positioned at different locations, and a dispatch system in communication with the devices receives a response from a speech processing system in response to a previously-communicated request, determines a relevance of the response to each of the devices, and forwards the response to at least one of the devices based on the determination.

[0013] Implementations may include one or more of the following, in any combination. The at least one of the devices may include an audio output device, and forwarding the response may cause that device to output audio signals corresponding to the response. The audio output device may include one or more of a loudspeaker, headphones, or a wearable audio device. The at least one of the devices may include a display, a video screen, or an appliance. The previously-communicated request may have been communicated from a third location not associated with any of the plurality of locations of the devices. The response may be a first response, and the dispatch system may also receive a response from a second speech processing system. The dispatch system may also forward the first response to a first one of the devices, and forward the second response to a second one of the devices. The dispatch system may also forward both the first response and the second response to a first one of the devices. The dispatch system may also forward only one of the first response and the second response to any of the devices.

[0014] Determining the relevance of the response may include determining which of the devices were associated with the previously-communicated request. Determining the relevance of the response may include determining which of the devices may be closest to a user associated with the previously-communicated

request. Determining the relevance of the response may be based on preferences associated with a user of the claimed system. Determining the relevance of the response may include determining a context of the previously-communicated request. The context may include one or more of an identification of a user that may have been associated with the request, which microphone of a plurality of microphones may have been associated with the request, a location of the user relative to the device locations, operating state of other devices in the system, and time of day. Determining the relevance of the response may include determining capabilities or resource availability of the devices.

[0015] A plurality of output devices may be positioned at different output device locations, and the dispatch system may also receive a response from the speech processing system in response to the transmitted request, determine a relevance of the response to each of the output devices, and forward the response to at least one of the output devices based on the determination. The at least one the output devices may include an audio output device, and forwarding the response causes that device to output audio signals corresponding to the response. The audio output device may include one or more of a loudspeaker, headphones, or a wearable audio device. The at least one of the output devices may include a display, a video screen, or an appliance. Determining the relevance of the response may include determining a relationship between the output devices and the microphones associated with the selected audio signals. Determining the relevance of the response may include determining which of the output devices may be closest to a source of the selected audio signal. Determining the relevance of the response may include determining a context in which the audio signals were derived. The context may include one or more of an identification of a user that may have been speaking, which microphone of the plurality of microphones produced the selected derived audio signals, a location of the user relative to the microphone locations and the device locations, operating state of other devices in the system, and time of day. Determining the relevance of the response may include determining capabilities or resource availability of the output devices.

[0016] In general, in one aspect, a system includes a plurality of microphones positioned at different microphone locations, a plurality of loudspeakers positioned at different loudspeaker locations, and a dispatch system in communication with the microphones and loudspeakers. The dispatch system derives a plurality of voice signals from the plurality of microphones, computes a confidence score about the inclusion of a wakeup word for each derived voice signal, compares the computed confidence scores, and based on the comparison, selects at least one of the derived voice signals and transmits at least a portion of the selected signal or signals to a speech processing system. The dispatch system receives a response from a speech processing system in response to the transmission, determines a relevance of the response to each of the loudspeakers, and forwards the response to at least one of the loudspeakers for output based on the determination.

[0017] Advantages include detecting a spoken command at multiple locations and providing a single response to the command. Advantages also include providing a response to a spoken command at a location more relevant to the user than the location where the command was detected.

[0018] All examples and features mentioned above can be combined in any technically possible way. Other features and advantages will be apparent from the description and the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

[0019] Figure 1 shows a system layout of microphones and devices that may respond to voice commands received by the microphones.

DESCRIPTION

[0020] As more and more devices implement voice-controlled user interfaces (VUIs), a problem arises that multiple devices may detect the same spoken command and attempt to handle it, resulting in problems ranging from redundant responses to contradictory actions being taken at different points of action. Similarly, if a spoken command can result in output or action by multiple devices, which device should take action may be ambiguous. In some VUIs, a

special phrase, referred to as a "wakeup word," "wake word," or "keyword" is used to activate the speech recognition features of the VUI – the device implementing the VUI is always listening for the wakeup word, and when it hears it, it parses whatever spoken commands came after it. This is done to conserve processing resources, by not parsing every sound that is detected, and can help disambiguate which system was the target of the command, but if multiple systems are listening for the same wakeup word, such as because the wakeup word is associated with a service provider and not individual pieces of hardware, the problem remains of determining which device should handle the command.

[0021] Figure 1 shows a potential environment, in which a stand-alone microphone array 102, a smart phone 104, a loudspeaker 106, and a set of headphones 108 each have microphones that detect a user's speech (to avoid confusion, we refer to the person speaking as the "user" and the device 106 as a "loudspeaker;" discrete things spoken by the user are "utterances"). Each of the devices that detects the utterance 110 transmits what it heard as an audio signal to a dispatch system 112. In the case of the devices having multiple microphones, those devices may combine the signals rendered by the individual microphones to render single combined audio signal, or they may transmit a signal rendered by each microphone.

[0022] This disclosure refers to various different types of audio and related signals. For clarity, the following conventions are used. "Acoustic signal" refers to physical signals, that is, physical sound pressure waves that are interpreted as sound by humans, such as the utterances mentioned above. "Audio signal" refers to electrical signals that represent sound. Audio signals may be generated from a microphone responding to acoustic audio, or they may be received from other electronic sources, such as recordings, computer-generated signals, or streamed data. "Audio output" refers to acoustic signals generated by a loudspeaker based on an audio signal input to the speaker.

[0023] The dispatch system 112 maybe a cloud-based service to which each of the devices is individually connected, a local service running on one of the same

devices or an associated device, a distributed service running cooperatively on some or all of the devices themselves, or any combination of these or similar architectures. Due to their different microphone designs and their differing proximity to the user, each of the devices may hear the utterance 110 differently, if at all. For example, the stand-alone microphone array 102 may have a high-quality beam-forming capability that allows it to clearly hear the utterance regardless of where the user is, while the headphones 108 and the smart phone 104 have highly directional near-field microphones that only clearly pick up the user's voice if the user is wearing the headphones and holding the phone up to their face, respectively. Meanwhile, the loudspeaker 106 may have a simple omnidirectional microphone that detects the speech well if the user is close to and facing the loudspeaker, but produces a low-quality signal otherwise.

[0024] Based on these and similar factors, the dispatch system 112 computes a confidence score for each audio signal (this may include the devices themselves scoring their own detection before sending what they heard, and sending that score along with their respective audio signals). Based on a comparison of the confidence scores, to each other, to a baseline, or both, the dispatch system 112 selects one or more of the audio signals for further processing. This may include locally performing speech recognition and taking direct action, or transmitting the audio signal over a network 114, such as the Internet or any private network, to another service provider. For example, if one of the devices produces an audio signal with a high confidence that the signal contains the wakeup word "OK Google," that audio signal may be sent to Google's cloud-based speech recognition system for handling. In the case that the audio signal is transmitted to a remote service, the wakeup word may be included along with whatever utterance followed it, or the utterance alone may be sent.

[0025] The confidence scoring may be based on a large number of factors, and may indicate confidence in more than one parameter as well. For example, the score may indicate a degree of confidence about which wakeup word was used (including whether one was used at all), or where the user was located relative to the microphone. The score may also indicate a degree of confidence in

whether the audio signal is of high quality. In one example, the dispatch system may score the audio signals from two devices as both having a high confidence score that a particular wakeup word was used, but score one of them with a low confidence in the quality of the audio signal, while the other is scored with a high confidence in the audio signal quality. The audio signal with the high confidence score for signal quality would be selected for further processing.

[0026] When more than one device transmits an audio signal, one of the critical things to determine confidence in is whether the audio signals represent the same utterance or two (or more) different utterances. The scoring itself may be based on such factors as signal level, signal-to-noise ratio (SNR), amount of reverberation in the signal, spectral content of the signal, user identification, knowledge about the user's location relative to the microphones, or relative timing of the audio signals at two or more of the devices. Location-related scoring and user identity-related scoring may be based on both the audio signals themselves and on external data such as visual systems, wearable trackers worn by users, and identity of the devices providing the signals. For example, if a smart phone is the source of the audio signal, a confidence score that the owner of that smart phone is the user whose voice was heard would be high. User location may be determined based on the strength and timing of acoustic signals received at multiple locations, or at multiple microphones in an array at a single location.

[0027] In addition to determining which wakeup word was used and which signal is best, the scoring may provide additional context that informs how the audio signal should be handled. For example, if the confidence scores indicate that the user was facing the loudspeaker, than it may be that a VUI associated with the loudspeaker should be used, over one associated with the smart phone. Context may include such things as which user was speaking, where the user was located and facing relative to the devices, what activity was the user engaged in (e.g., exercising, cooking, watching TV), what time of day it is, or what other devices are in use (including devices other than those providing the audio signals).

[0028] In some cases, the scoring indicates that more than one command was heard. For example, two devices may each have high confidence that they heard different wakeup words, or that they heard different users speaking. In that case, the dispatch system may send two requests – one request to each system for which a wakeup word was used, or two different requests to a single system that both users invoked. In other cases, more than one of the audio signals may be sent – for example, to get more than one response, to let the remote system decide which one to use, or to improve the voice recognition by combining the signals. In addition to selecting an audio signal for further handling, the scoring may also lead to other user feedback. For example, a light may be flashed on whichever device was selected, so that the user knows the command was received.

[0029] Similar considerations come into play when a response is received from whatever service or system the dispatch system sent the audio signal to for handling. In many cases, the context around the utterance will also inform the handling of the response. For example, the response may be sent to the device from which the selected audio signal was received. In other cases, the response may be sent to a different device. For example, if the audio signal from the stand-alone microphone array 102 was selected, but the response back from the VUI is to start playing an audio file, the response should be handled by the headphones 108 or the loudspeaker 106. If the response is to display information, the smart phone 104 or some other device with a screen would be used to deliver the response. If the microphone array audio signal was selected because the scoring indicated that it had the best signal quality, additional scoring may have indicated that the user was not using the headphones 108 but was in the same room as the loudspeaker 106, so the loudspeaker is the likely target for the response. Other capabilities of the devices would also be considered – for example, while only audio devices are shown, voice commands could address other systems, such as lighting or home automation systems. Hence, if the response to the utterance is to turn down lights, the dispatch system may conclude that it is referring to the lights in the room where the strongest audio signal was detected. Other potential output devices include displays, screens

(e.g., the screen on the smart phone, or a television monitor), appliances, door locks, and the like. In some examples, the context is provided to the remote system, and the remote system specifically targets a particular output device based on a combination of the utterance and the context.

[0030] As mentioned, the dispatch system may be a single computer or a distributed system. The speech processing provided may similarly be provided by a single computer or a distributed system, coextensive with or separate from the dispatch system. They each may be located entirely locally to the devices, entirely in the cloud, or split between both. They may be integrated into one or all of the devices. The various tasks described – scoring signals, detecting wakeup words, sending a signal to another system for handling, parsing the signal for a command, handling the command, generating a response, determining which device should handle the response, etc., may be combined together or broken down into more sub-tasks. Each of the tasks and sub-tasks may be performed by a different device or combination of devices, locally or in a cloud-based or other remote system.

[0031] When we refer to microphones, we include microphone arrays without any intended restriction on particular microphone technology, topology, or signal processing. Similarly, references to loudspeakers and headphones should be understood to include any audio output devices – televisions, home theater systems, doorbells, wearable speakers, etc.

[0032] Embodiments of the systems and methods described above comprise computer components and computer-implemented steps that will be apparent to those skilled in the art. For example, it should be understood by one of skill in the art that instructions for executing the computer-implemented steps may be stored as computer-executable instructions on a computer-readable medium such as, for example, floppy disks, hard disks, optical disks, Flash ROMS, nonvolatile ROM, and RAM. Furthermore, it should be understood by one of skill in the art that the computer-executable instructions may be executed on a variety of processors such as, for example, microprocessors, digital signal processors, gate arrays, etc. For ease of exposition, not every step or element of

the systems and methods described above is described herein as part of a computer system, but those skilled in the art will recognize that each step or element may have a corresponding computer system or software component. Such computer system and/or software components are therefore enabled by describing their corresponding steps or elements (that is, their functionality), and are within the scope of the disclosure.

[0033] A number of implementations have been described. Nevertheless, it will be understood that additional modifications may be made without departing from the scope of the inventive concepts described herein, and, accordingly, other embodiments are within the scope of the following claims.

WHAT IS CLAIMED IS:

1. A system comprising:
a plurality of microphones positioned at different locations; and
a dispatch system in communication with the microphones and configured to:
derive a plurality of audio signals from the plurality of microphones,
compute a confidence score for each derived audio signal, and
compare the computed confidence scores, and based on the comparison, select at least one of the derived audio signals for further handling.

2. The system of claim 1, wherein the dispatch system comprises a plurality of local processors each connected to at least one of the microphones.

3. The system of claim 1, wherein the dispatch system comprises at least a first local processor and at least a second processor available to the first processor over a network.

4. The system of claim 1, wherein computing the confidence score for each derived audio signal comprises computing a confidence in one or more of whether the signal comprises speech, whether a wakeup word is included in the signal, what wakeup word is included in the signal, a quality of speech contained in the signal, an identity of a user whose voice is recorded in the signal, or a location of the user relative to the microphone locations.

5. The system of claim 1, wherein computing the confidence score for each derived audio signal comprises determining that the audio signal appears to contain an utterance and whether the utterance includes a wakeup word.

6. The system of claim 5, wherein computing the confidence score for each derived audio signal further comprises identifying which wakeup word from a plurality of wakeup words is included in the speech.

7. The system of claim 5, wherein computing the confidence score for each derived audio signal further comprises determining a degree of confidence that the utterance includes the wakeup word.

8. The system of claim 1, wherein computing the confidence score for each derived audio signal comprises comparing one or more of a timing between when the microphones detected sounds corresponding to each of the audio signals, signal strength of the derived audio signals, signal-to-noise ratio of the derived audio signals, spectral content of the derived audio signals, and reverberation within the derived audio signals.

9. The system of claim 1, wherein computing the confidence score for each derived audio signal comprises, for each audio signal, computing a distance between an apparent source of the audio signal and at least one of the microphones.

10. The system of claim 1, wherein computing the confidence score for each derived audio signal comprises computing a location of the source of each audio signal relative to the locations of the microphones.

11. The system of claim 10, wherein computing the location of the source of each audio signal comprises triangulating the location based on computed distances distance between each source and at least two of the microphones.

12. The system of claim 1, wherein the dispatch system is further configured to transmit at least a portion of the selected signal or signals to a speech processing system to provide the further handling.

13. The system of claim 12, wherein transmitting the selected audio signal or signals comprises selecting at least one speech processing system from a plurality of speech processing systems.

14. The system of claim 13, wherein at least one speech processing system of the plurality of speech processing systems comprises a speech recognition service provided over a wide-area network.

15. The system of claim 13, wherein at least one speech processing system of the plurality of speech processing systems comprises a speech recognition process executing on the same processor on which the dispatch system is executing.

16. The system of claim 13, wherein the selection of the speech processing system is based on one or more of preferences associated with a user of the claimed system, the computed confidence scores, or context in which the audio signals are derived.

17. The system of claim 16 wherein the context includes one or more of an identification of a user that is speaking, which microphones of the plurality of microphones produced the selected derived audio signals, a location of the user relative to the microphone locations, operating state of other devices in the system, and time of day.

18. The system of claim 13, wherein the selection of the speech processing system is based on resources available to the speech processing systems.

19. The system of claim 1, wherein the number of derived audio signals is not equal to the number of microphones.

20. The system of claim 1, wherein at least one of the microphones comprises a microphone array.

21. The system of claim 1, further comprising non-audio input devices.

22. The system of claim 21, wherein the non-audio input devices comprise one or more of accelerometers, presence detectors, cameras, wearable sensors, or user interface devices.

23. A method of processing audio signals, comprising:
receiving audio signals form a plurality of microphones positioned at different locations; and
in a dispatch system in communication with the microphones:
deriving a plurality of audio signals from the plurality of microphones,
computing a confidence score for each derived audio signal,
comparing the computed confidence scores, and based on the comparison,
selecting at least one of the derived audio signals for further handling.

24. The method of claim 23, wherein computing the confidence score for each derived audio signal comprises computing a confidence in one or more of whether the signal comprises speech, whether a wakeup word is included in the signal, what wakeup word is included in the signal, a quality of speech contained in the signal, an identity of a user whose voice is recorded in the signal, or a location of the user relative to the microphone locations.

25. The system of claim 23, wherein computing the confidence score for each derived audio signal comprises determining that the audio signal appears to contain an utterance and whether the utterance includes a wakeup word.

26. A system comprising:
a plurality of microphones positioned at different locations; and
a dispatch system in communication with the microphones and configured to:
derive a plurality of audio signals from the plurality of microphones,

compute a confidence score for each derived audio signal, and compare the computed confidence scores, and based on the comparison, select at least two of the derived audio signals for further handling;

wherein comparing the computed confidence scores comprises determining that at least the two selected audio signals appear to contain utterances from at least two different users.

27. The system of claim 26, wherein the determining that the selected audio signals appear to contain utterances from at least two different users is based on one or more of voice identification, location of the users relative to the locations of the microphones, which of the microphones produced each of the selected audio signals, use of different wakeup words in the two selected audio signals and visual identification of the users.

28. The system of claim 26, wherein the dispatch system is further configured to send the selected audio signals corresponding to the two different users to two different selected speech processing systems.

29. The system of claim 28, wherein the selected audio signals are assigned to the selected speech processing systems based on one or more of preferences of the users, load balancing of the speech processing systems, context of the selected audio signals, and use of different wakeup words in the two selected audio signals.

30. The system of claim 26, wherein the dispatch system is further configured to send the selected audio signals corresponding to the two different users to the same speech processing system as two separate processing requests.

31. A system comprising:

a plurality of microphones positioned at different locations; and

a dispatch system in communication with the microphones and configured to:

derive a plurality of audio signals from the plurality of microphones,

compute a confidence score for each derived audio signal,

compare the computed confidence scores, and based on the comparison, select at least two of the derived audio signals for further handling;

wherein comparing the computed confidence scores comprises determining that at least the two selected audio signals appear to represent the same utterance.

32. The system of claim 31, wherein the determining that the selected audio signals represent the same utterance is based on one or more of voice identification, location of the source of the audio signals relative to the locations of the microphones, which of the microphones produced each of the selected audio signals, time of arrival of the audio signals, correlations between the audio signals or between outputs of microphone array elements, pattern matching, and visual identification of the person speaking.

33. The system of claim 31, wherein the dispatch system is further configured to send only one of the audio signals appearing to represent the same utterance to the speech processing system.

34. The system of claim 31, wherein the dispatch system is further configured to send both of the audio signals appearing to represent the same utterance to the speech processing system.

35. The system of claim 31, wherein the dispatch system is further configured to:

transmit at least one selected audio signal to each of at least two speech processing systems,

receive responses from each of the speech processing systems, and

determine an order in which to output the responses.

36. The system of claim 31, wherein the dispatch system is further configured to:
transmit at least two selected audio signals to at least one speech processing system,
receive responses from the speech processing system corresponding to each of the transmitted signals, and
determine an order in which to output the responses.

37. A method of processing audio signals, comprising:
receiving audio signals from a plurality of microphones positioned at different locations; and
in a dispatch system in communication with the microphones:
deriving a plurality of audio signals from the plurality of microphones,
computing a confidence score for each derived audio signal,
comparing the computed confidence scores, and, based on the comparison,
selecting at least two of the derived audio signals for further handling;
wherein comparing the computed confidence scores comprises determining that at least the two selected audio signals appear to contain utterances from at least two different users.

38. The method of claim 37, wherein determining that the selected audio signals appear to contain utterances from at least two different users is based on one or more of voice identification, location of the users relative to the locations of the microphones, which of the microphones produced each of the selected audio signals, use of different wakeup words in the two selected audio signals and visual identification of the users.

39. The method of claim 37, further comprising sending the selected audio signals corresponding to the two different users to two different selected speech processing systems.

40. The method of claim 39, further comprising assigning the selected audio signals to the selected speech processing systems based on one or more of preferences of the users, load balancing of the speech processing systems, context of the selected audio signals, and use of different wakeup words in the two selected audio signals.

41. The method of claim 37, further comprising sending the selected audio signals corresponding to the two different users to the same speech processing system as two separate processing requests.

42. A method of processing audio signals comprising:
receiving audio signals form a plurality of microphones positioned at different locations; and
in a dispatch system in communication with the microphones:
deriving a plurality of audio signals from the plurality of microphones,
computing a confidence score for each derived audio signal,
comparing the computed confidence scores, and based on the comparison,
selecting at least two of the derived audio signals for further handling;
wherein comparing the computed confidence scores comprises determining that at least the two selected audio signals appear to represent the same utterance.

43. The method of claim 42, wherein determining that the selected audio signals represent the same utterance is based on one or more of voice identification, location of the source of the audio signals relative to the locations of the microphones, which of the microphones produced each of the selected audio signals, time of arrival of the audio signals, correlations between the audio signals or between outputs of microphone array elements, pattern matching, and visual identification of the person speaking.

44. The method of claim 42, further comprising sending only one of the audio signals appearing to represent the same utterance to the speech processing system.

45. The method of claim 42, further comprising sending both of the audio signals appearing to represent the same utterance to the speech processing system.

46. The method of claim 42, is further comprising:
transmitting at least one selected audio signal to each of at least two speech processing systems,
receiving responses from each of the speech processing systems, and
determining an order in which to output the responses.

47. The method of claim 42, further comprising:
transmitting at least two selected audio signals to at least one speech processing system,
receiving responses from the speech processing system corresponding to each of the transmitted signals, and
determining an order in which to output the responses.

48. A system comprising:
a plurality of microphones positioned at different locations;
an output device; and
a dispatch system in communication with the microphones and configured to:
derive a plurality of audio signals from the plurality of microphones,
compute a confidence score for each derived audio signal,
compare the computed confidence scores,
based on the comparison, select at least one of the derived audio signals for further handling,
receive a response to the further processing, and

output the response using the output device;

wherein the output device does not correspond to the microphone that captured the selected audio signals.

49. The system of claim 48, wherein the output device comprises one or more of a loudspeaker, headphones, a wearable audio device, a display, a video screen, or an appliance.

50. The system of claim 48, wherein upon receiving multiple responses to the further processing, the dispatch system determines an order in which to output the responses by combining the responses into a single output.

51. The system of claim 48, wherein upon receiving multiple responses to the further processing, the dispatch system determines an order in which to output the responses by selecting fewer than all of the responses to output.

52. The system of claim 48, wherein upon receiving multiple responses to the further processing, the dispatch system sends different responses to different output devices.

53. A method of processing audio signals, comprising:

receiving audio signals from a plurality of microphones positioned at different locations;

in a dispatch system in communication with the microphones:

deriving a plurality of audio signals from the plurality of microphones,

computing a confidence score for each derived audio signal,

comparing the computed confidence scores,

based on the comparison, selecting at least one of the derived audio signals for further handling,

receiving a response to the further processing, and

output the response using an output device;

wherein the output device does not correspond to the microphone that captured the selected audio signals.

54. The method of claim 53, wherein the output device is not located at any of the locations where the microphones are located.

55. A system comprising:
a plurality of devices positioned at different locations; and
a dispatch system in communication with the devices and configured to:
receive a response from a speech processing system in response to a previously-communicated request,
determine a relevance of the response to each of the devices, and
forward the response to at least one of the devices based on the determination.

56. The system of claim 55, wherein the at least one of the devices comprises an audio output device, and forwarding the response causes that device to output audio signals corresponding to the response.

57. The system of claim 55, wherein the at least one of the devices comprises a display, a video screen, or an appliance.

58. The system of claim 55, wherein the response is a first response, and the dispatch system is further configured to receive a response from a second speech processing system.

59. The system of claim 58, wherein the dispatch system is further configured to forward the first response to a first one of the devices, and forward the second response to a second one of the devices.

60. The system of claim 58, wherein the dispatch system is further configured to forward both the first response and the second response to a first one of the devices.

61. The system of claim 58, wherein the dispatch system is further configured to forward only one of the first response and the second response to any of the devices.

62. The system of claim 55, wherein determining the relevance of the response comprises determining which of the devices were associated with the previously-communicated request.

63. The system of claim 55, wherein determining the relevance of the response comprises determining which of the devices is closest to a user associated with the previously-communicated request.

64. The system of claim 55, wherein determining the relevance of the response is based on preferences associated with a user of the claimed system.

65. The system of claim 55, wherein determining the relevance of the response comprises determining a context of the previously-communicated request.

66. The system of claim 65, wherein the context includes one or more of an identification of a user that was associated with the request, which microphone of a plurality of microphones was associated with the request, a location of the user relative to the device locations, operating state of other devices in the system, and time of day.

67. The system of claim 55, wherein determining the relevance of the response comprises determining capabilities or resource availability of the devices.

68. The system of claim 55, wherein determining the relevance of the response comprises determining a relationship between the output devices and the microphones associated with the selected audio signals.

69. The system of claim 55, wherein determining the relevance of the response comprises determining which of the output devices is closest to a source of the selected audio signal.

70. A system comprising:

a plurality of microphones positioned at different microphone locations;

a plurality of loudspeakers positioned at different loudspeaker locations; and

a dispatch system in communication with the microphones and loudspeakers and configured to:

derive a plurality of voice signals from the plurality of microphones;

compute a confidence score about the inclusion of a wakeup word for each derived voice signal;

compare the computed confidence scores,

based on the comparison, select at least one of the derived voice signals and transmit at least a portion of the selected signal or signals to a speech processing system,

receive a response from a speech processing system in response to the transmission,

determine a relevance of the response to each of the loudspeakers, and

forward the response to at least one of the loudspeakers for output based on the determination.

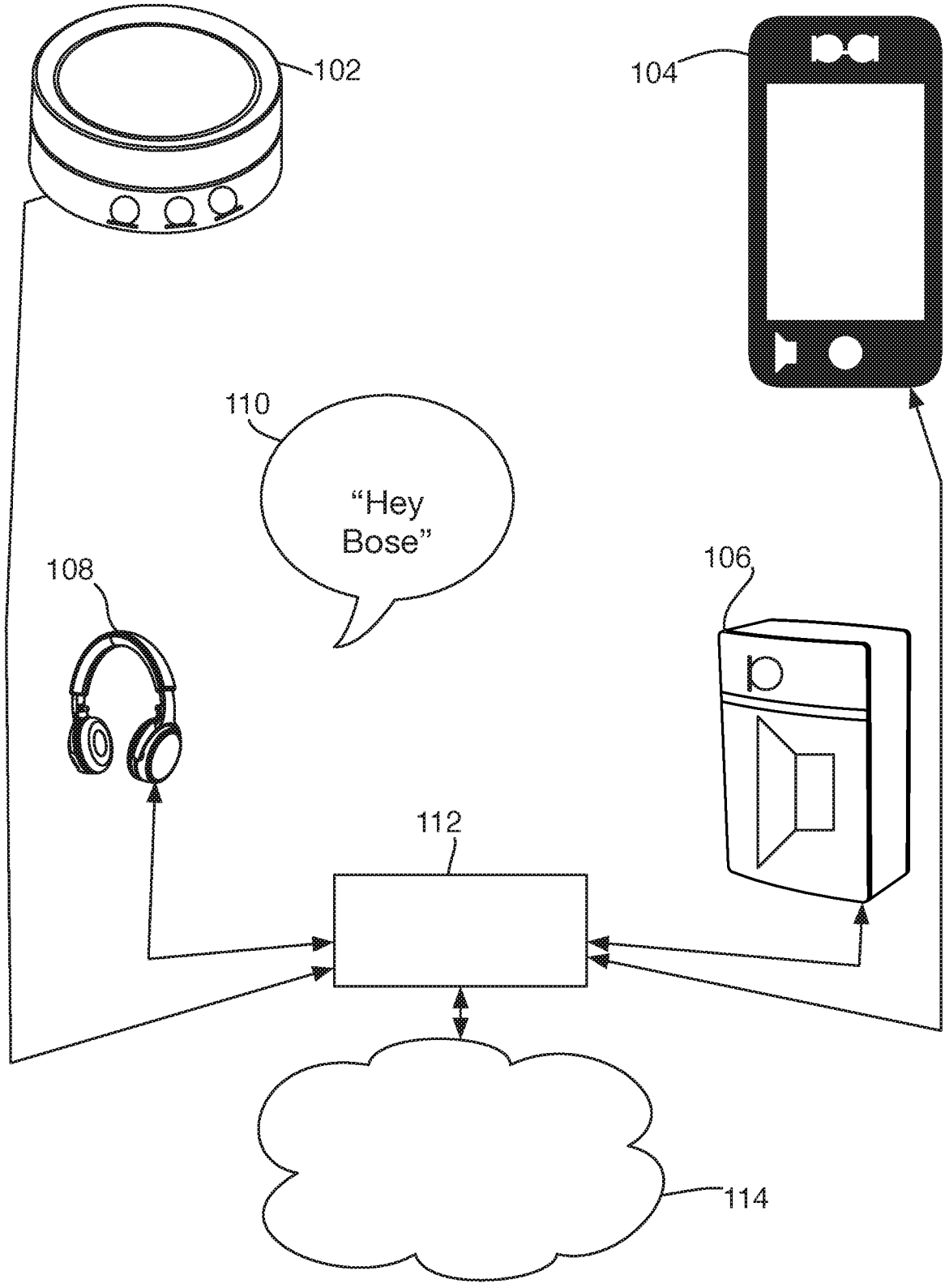

Fig. 1