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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
31 December 2014 (31.12.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/210392 A2

(51) **International Patent Classification**:
*G10L 15/04* (2013.01)

(21) **International Application Number**:
PCT/US2014/044455

(22) **International Filing Date**:
26 June 2014 (26.06.2014)

(25) **Filing Language**: English

(26) **Publication Language**: English

(30) **Priority Data**:
13/929,540 27 June 2013 (27.06.2013) US

(71) **Applicant**: **RAWLES LLC** [US/US]; 103 Foulk Road, Suite 100, Wilmington, DE 19803 (US).

(72) **Inventors**: **POGUE, Michael, Alan**; c/o Rawles LLC, 103 Foulk Road, Suite 100, Wilmington, DE 19803 (US). **HILMES, Philip, Ryan**; c/o Rawles LLC, 103 Foulk Road, Suite 100, Wilmington, DE 19803 (US).

(74) **Agents**: **HAYES, Daniel, L.** et al.; Lee & Hayes, PLLC, 601 W. Riverside Ave, Suite 1400, Spokane, WA 99201 (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, 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, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, 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, 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).

**Published**:

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

(54) **Title**: DETECTING SELF-GENERATED WAKE EXPRESSIONS

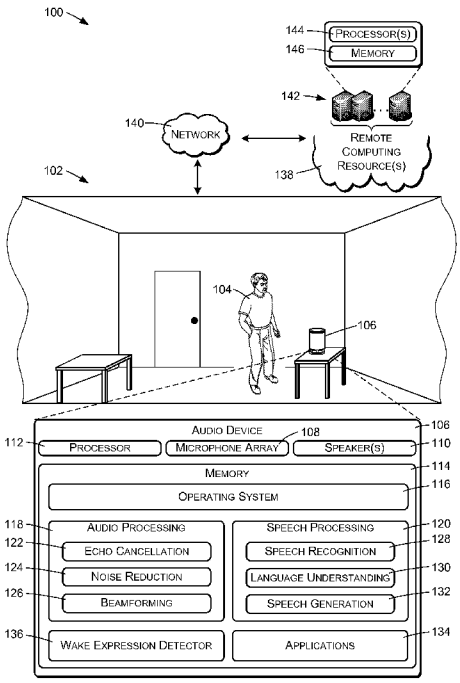


FIG. 1

(57) **Abstract**: A speech-based audio device may be configured to detect a user-uttered wake expression and to respond by interpreting subsequent words or phrases as commands. In order to distinguish between utterance of the wake expression by the user and generation of the wake expression by the device itself, directional audio signals may by analyzed to detect whether the wake expression has been received from multiple directions. If the wake expression has been received from many directions, it is declared as being generated by the audio device and ignored. Otherwise, if the wake expression is received from a single direction or a limited number of directions, the wake expression is declared as being uttered by the user and subsequent words or phrase are interpreted and acted upon by the audio device.

WO 2014/210392 A2

# DETECTING SELF-GENERATED WAKE EXPRESSIONS

## RELATED APPLICATIONS

**[0001]** The present application claims priority to US Patent Application No. 13/929,540 filed on June 27, 2013, entitled "Detecting Self-Generated Wake Expressions", which is incorporated by reference herein in its entirety.

## BACKGROUND

**[0002]** Homes, offices, automobiles, and public spaces are becoming more wired and connected with the proliferation of computing devices such as notebook computers, tablets, entertainment systems, and portable communication devices. As computing devices evolve, the way in which users interact with these devices continues to evolve. For example, people can interact with computing devices through mechanical devices (e.g., keyboards, mice, etc.), electrical devices (e.g., touch screens, touch pads, etc.), and optical devices (e.g., motion detectors, camera, etc.). Another way to interact with computing devices is through audio devices that capture and respond to human speech.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0003]** The detailed description is described with reference to the accompanying figures. In the figures, the left-most digit(s) of a reference number identifies the figure in which the reference number first appears. The use of the same reference numbers in different figures indicates similar or identical components or features.

**[0004]** FIG. 1 is a block diagram of an illustrative voice interaction computing architecture that includes a voice-controlled audio device.

**[0005]** FIG. 2 is a view of a voice-controlled audio device such as might be used in the architecture of FIG. 1.

**[0006]** FIGS. 3 and 4 are block diagrams illustrating functionality that may be implemented to discriminate between user-uttered wake expressions and device-produced wake expressions.

**[0007]** FIG. 5 is a flow diagram illustrating an example process for learning reference parameters, which may be used to detecting device-produced wake expressions.

**[0008]** FIG. 6 is a flow diagram illustrating an example process for discriminating between user-uttered wake expressions and device-produced wake expressions.

## DETAILED DESCRIPTION

**[0009]** This disclosure pertains generally to a speech interface device or other audio device that provides speech-based interaction with a user. The audio device has a speaker that produces audio within the environment of a user and a microphone that captures user speech. The audio device may be configured to respond to user speech by performing functions and providing services. User commands may be prefaced by a wake expression, also referred to as a trigger expression, such as a predefined word, phrase, or other sound. In response to detecting the wake expression, the audio device interprets any immediately following words or phrases as actionable input or commands.

**[0010]** In providing services to the user, the audio device may itself generate the wake expression at its speaker, which may cause the audio device to react as if the user has spoken the wake expression. To avoid this, the audio device may be configured to evaluate the direction or directions from which the wake expression has

been received. Generally, a wake expression generated by the audio device will be received omnidirectionally. A wake expression generated by a user, on the other hand, will be received from one direction or a limited number of directions. Accordingly, the audio device may be configured to ignore wake expressions that are received omnidirectionally, or from more than one or two directions. Note that a user-uttered wake expression may at times seem to originate from more than a single direction due to acoustic reflections within a particular environment.

**[0011]** More particularly, an audio device may be configured to perform wake expression detection with respect to multiple directional audio signals. The audio device may be further configured to compare the number or pattern of the directional audio signals containing the wake expression to a reference. The reference may indicate a threshold number of directional input signals or a pattern or set of the directional signals. When the reference comprises a threshold, the wake expression is considered to have been generated by the audio device if the number of directional input audio signals containing the wake expression exceeds the threshold. When the reference comprises a pattern or set, the wake expression is evaluated based on whether the particular directional input audio signals containing the wake expression match those of the pattern or set.

**[0012]** In some implementations, the audio device may be configured to learn or to train itself regarding patterns of audio characteristics are characteristic of device-generated wake expressions. For example, the audio device may be configured to generate the wake expression or another sound upon initialization, and to identify a combination of the directional audio signals in which the expression or sound is detected. Subsequently, the audio device may be configured to ignore the wake expression when it is detected in the learned combination of directional audio signals.

**[0013]** Other conditions or parameters may also be analyzed or considered when determining whether a detected wake expression has been generated by the audio device rather than by the user. As examples, such conditions or parameters may include the following: presence and/or loudness of speaker output; whether the speaker output is known to contain speech; echo characteristics input signals and/or effectiveness of echo reduction; loudness of received audio signals including directional audio signals.

**[0014]** Machine learning techniques may be utilized to analyze various parameters in order to determine patterns of parameters that are typically exhibited when a wake expression has been self-generated.

**[0015]** FIG. 1 shows an illustrative voice interaction computing architecture 100 set in an environment 102, such as a home environment, that includes a user 104. The architecture 100 includes an electronic, voice-controlled audio device 106 with which the user 104 may interact. In the illustrated implementation, the audio device 106 is positioned on a table within a room of the environment 102. In other implementations, the audio device 106 may be placed in any number of locations (e.g., ceiling, wall, in a lamp, beneath a table, under a chair, etc.). Furthermore, more than one audio device 106 may be positioned in a single room, or one audio device 106 may be used to accommodate user interactions from more than one room.

**[0016]** Generally, the audio device 106 may have a microphone array 108 and one or more audio speakers or transducers 110 to facilitate audio interactions with the user 104 and/or other users. The microphone array 108 produces input audio signals representing audio from the environment 102, such as sounds uttered by the user 104 and ambient noise within the environment 102. The input audio signals may also contain output audio components that have been produced by the speaker 110. As

will be described in more detail below, the input audio signals produced by the microphone array 108 may comprise directional audio signals or may be used to produce directional audio signals, where each of the directional audio signals emphasizes audio from a different direction relative to the microphone array 108.

**[0017]** The audio device 106 includes operational logic, which in many cases may comprise a processor 112 and memory 114. The processor 112 may include multiple processors and/or a processor having multiple cores. The memory 114 may contain applications and programs in the form of instructions that are executed by the processor 112 to perform acts or actions that implement desired functionality of the audio device 106, including the functionality specifically described below. The memory 114 may be a type of computer storage media and may include volatile and nonvolatile memory. Thus, the memory 114 may include, but is not limited to, RAM, ROM, EEPROM, flash memory, or other memory technology.

**[0018]** The audio device 106 may have an operating system 116 that is configured to manage hardware and services within and coupled to the audio device 106. In addition, the audio device 106 may include audio processing components 118 and speech processing components 120.

**[0019]** The audio processing components 118 may include functionality for processing input audio signals generated by the microphone array 108 and/or output audio signals provided to the speaker 110. As an example, the audio processing components 118 may include an acoustic echo cancellation or suppression component 122 for reducing acoustic echo generated by acoustic coupling between the microphone array 108 and the speaker 110. The audio processing components 118 may also include a noise reduction component 124 for reducing noise in received audio signals, such as elements of audio signals other than user speech.

**[0020]** The audio processing components 118 may include one or more audio beamformers or beamforming components 126 that are configured to generate an audio signal that is focused in a direction from which user speech has been detected. More specifically, the beamforming components 126 may be responsive to spatially separated microphone elements of the microphone array 108 to produce directional audio signals that emphasize sounds originating from different directions relative to the audio device 106, and to select and output one of the audio signals that is most likely to contain user speech.

**[0021]** The speech processing components 120 receive an audio signal that has been processed by the audio processing components 118 and perform various types of processing in order to understand the intent expressed by human speech. The speech processing components 120 may include an automatic speech recognition component 128 that recognizes human speech in the audio represented by the received audio signal. The speech processing components 120 may also include a natural language understanding component 130 that is configured to determine user intent based on recognized speech of the user 104.

**[0022]** The speech processing components 120 may also include a text-to-speech or speech generation component 132 that converts text to audio for generation at the speaker 110.

**[0023]** The audio device 106 may include a plurality of applications 134 that are configured to work in conjunction with other elements of the audio device 106 to provide services and functionality. The applications 134 may include media playback services such as music players. Other services or operations performed or provided by the applications 134 may include, as examples, requesting and consuming entertainment (e.g., gaming, finding and playing music, movies or other content, etc.),

personal management (e.g., calendaring, note taking, etc.), online shopping, financial transactions, database inquiries, and so forth. In some embodiments, the applications may be pre-installed on the audio device 106, and may implement core functionality of the audio device 106. In other embodiments, one or more of the applications 134 may be installed by the user 104, or otherwise installed after the audio device 106 has been initialized by the user 104, and may implement additional or customized functionality as desired by the user 104.

**[0024]** In certain embodiments, the primary mode of user interaction with the audio device 106 is through speech. For example, the audio device 106 may receive spoken commands from the user 104 and provide services in response to the commands. The user may speak a predefined wake or trigger expression (e.g., "Awake"), which may be followed by instructions or directives (e.g., "I'd like to go to a movie. Please tell me what's playing at the local cinema."). Provided services may include performing actions or activities, rendering media, obtaining and/or providing information, providing information via generated or synthesized speech via the audio device 106, initiating Internet-based services on behalf of the user 104, and so forth.

**[0025]** The audio device 106 may include wake expression detection components 136, which monitor received input audio and provide event notifications to the speech processing components 120 and/or applications 134 in response to user utterances of a wake or trigger expression. The speech processing components 120 and/or applications 134 may respond by interpreting and acting upon user speech that follows the wake expression. The wake expression may comprise a word, a phrase, or other sound.

**[0026]** In some instances, the audio device 106 may operate in conjunction with or may otherwise utilize computing resources 138 that are remote from the

environment 102. For instance, the audio device 106 may couple to the remote computing resources 138 over a network 140. As illustrated, the remote computing resources 138 may be implemented as one or more servers or server devices 142. The remote computing resources 138 may in some instances be part of a network-accessible computing platform that is maintained and accessible via a network 140 such as the Internet. Common expressions associated with these remote computing resources 138 may include "on-demand computing", "software as a service (SaaS)", "platform computing", "network-accessible platform", "cloud services", "data centers", and so forth.

**[0027]** Each of the servers 142 may include processor(s) 144 and memory 146. The servers 142 may perform various functions in support of the audio device 106, and may also provide additional services in conjunction with the audio device 106. Furthermore, one or more of the functions described herein as being performed by the audio device 106 may be performed instead by the servers 142, either in whole or in part. As an example, the servers 142 may in some cases provide the functionality attributed above to the speech processing components 120. Similarly, one or more of the applications 134 may reside in the memory 146 of the servers 142 and may be executed by the servers 142.

**[0028]** The audio device 106 may communicatively couple to the network 140 via wired technologies (e.g., wires, universal serial bus (USB), fiber optic cable, etc.), wireless technologies (e.g., radio frequencies (RF), cellular, mobile telephone networks, satellite, Bluetooth, etc.), or other connection technologies. The network 140 is representative of any type of communication network, including data and/or voice network, and may be implemented using wired infrastructure (e.g., coaxial

cable, fiber optic cable, etc.), a wireless infrastructure (e.g., RF, cellular, microwave, satellite, Bluetooth®, etc.), and/or other connection technologies.

**[0029]** Although the audio device 106 is described herein as a voice-controlled or speech-based interface device, the techniques described herein may be implemented in conjunction with various different types of devices, such as telecommunications devices and components, hands-free devices, entertainment devices, media playback devices, and so forth.

**[0030]** FIG. 2 illustrates details of microphone and speaker positioning in an example embodiment of the audio device 106. In this embodiment, the audio device 106 is housed by a cylindrical body 202. The microphone array 108 comprises six microphones 204 that are laterally spaced from each other so that they can be used by audio beamforming components to produce directional audio signals. In the illustrated embodiment, the microphones 204 are positioned in a circle or hexagon on a top surface 206 of the cylindrical body 202. Each of the microphones 204 is omnidirectional in the described embodiment, and beamforming technology is used to produce directional audio signals based on signals form the microphones 204. In other embodiments, the microphones may have directional audio reception, which may remove the need for subsequent beamforming.

**[0031]** In various embodiments, the microphone array 108 may include greater or less than the number of microphones shown. For example, an additional microphone may be located in the center of the top surface 206 and used in conjunction with peripheral microphones for producing directionally focused audio signals.

**[0032]** The speaker 110 may be located at the bottom of the cylindrical body 202, and may be configured to emit sound omnidirectionally, in a 360 degree pattern around the audio device 106. For example, the speaker 110 may comprise a round

speaker element directed downwardly in the lower part of the body 202, to radiate sound radially through an omnidirectional opening or gap 208 in the lower part of the body 202.

**[0033]** FIG. 3 illustrates an example 300 of detecting wake expressions, such as might be performed in conjunction with the architecture described above. The speaker 110 is configured to produce audio in the user environment 102. The microphone array 108 is configured as described above to receive input audio from the user environment 102, which may include speech utterances by the user 104 as well as components of audio produced by the speaker 110. The microphone array 108 produces a plurality of input audio signals 302, corresponding respectively to each of the microphones of the microphone array 108.

**[0034]** The audio beamformer 126 receives the input audio signals 302 and processes the signals 302 to produce a plurality of directional or directionally-focused audio signals 304. The directional audio signals 304 represent or contain input audio from the environment 102, corresponding respectively to different areas or portions of the environment 102. In the described embodiment, the directional audio signals 304 correspond respectively to different radial directions relative to the audio device 106.

**[0035]** Audio beamforming, also referred to as audio array processing, uses a microphone array having multiple microphones that are spaced from each other at known distances. Sound originating from a source is received by each of the microphones. However, because each microphone is potentially at a different distance from the sound source, a propagating sound wave arrives at each of the microphones at slightly different times. This difference in arrival time results in phase differences between audio signals produced by the microphones. The phase

differences can be exploited to enhance sounds originating from chosen directions relative to the microphone array.

**[0036]** Beamforming uses signal processing techniques to combine signals from the different microphones so that sound signals originating from a particular direction are emphasized while sound signals from other directions are deemphasized. More specifically, signals from the different microphones are combined in such a way that signals from a particular direction experience constructive interference, while signals from other directions experience destructive interference. The parameters used in beamforming may be varied to dynamically select different directions, even when using a fixed-configuration microphone array.

**[0037]** The wake expression detector 136 receives the directional audio signals 304 and detects occurrences of the wake expression in the audio represented by the individual directional audio signals 304. In the described embodiment, this is performed by multiple expression recognizers or detectors 306, corresponding respectively to each of the directional audio signals 304. The expression recognizers are configured to identify which of the directional audio signals 304 are likely to contain or represent the wake expression. In some embodiments, the expression recognizers 406 may be configured collectively to identify a set of the directional audio signals 304 in which the wake expression is detected or in which the wake expression is likely to have occurred.

**[0038]** Each of the expression recognizers 306 implements automated speech recognition to detect the wake expression in the corresponding directional audio signal 304. In some cases, implementation of the automated speech recognition by the expression recognizers 306 may be somewhat simplified in comparison to a full recognition system because of the fact that only a single word or phrase needs to be

detected. In some implementations, however, elements or functionality provided by the speech recognition component 128 may be used to perform the functions of the expression recognizers 306.

**[0039]** The expression recognizers 306 produce a set of recognition indications or parameters 308 that provide indications of whether the audio of the corresponding directional audio signals 304 contain the wake expression. In some implementations, each parameter or indication 308 may comprise a binary, true/false value or parameter regarding whether the wake expression has been detected in the audio of the corresponding directional audio signal 304. In other implementations, the parameters or indications 308 may comprise confidence levels or probabilities, indicating relative likelihoods that the wake expression has been detected in the corresponding directional audio signals. For example, a confidence level may be indicated as a percentage ranging from 0% to 100%.

**[0040]** The wake expression detector 136 may include a classifier 310 that distinguishes between generation of the wake expression by the speaker 110 and utterance of the wake expression by the user 104, based at least in part on the parameters 308 produced by the expression recognizers 306 regarding which of the directional audio signals are likely to contain the wake expression.

**[0041]** In certain embodiments, each of the recognizers 306 may be configured to produce a binary value indicating whether or not the wake expression has been detected or recognized in the corresponding directional audio signal 304. Based on this binary indication, the classifier 310 identifies a set of the directional audio signals 304 that contain the wake expression. The classifier 310 then determines whether a wake expression has been generated by the speaker 110 or uttered by the user 104,

based on which of the directional audio signals are in the identified set of directional audio signals.

**[0042]** As an example, it may be assumed in certain situations that a user-uttered wake expression will be received from a single direction or directional cone with respect to the audio device 106, and that a wake expression produced by the speaker 110 will be received from all directions or multiple directional cones. Based on this assumption, the classifier 310 may evaluate a wake expression as being generated by the speaker 110 if the wake expression is detected in all or a majority (i.e., more than half) of the directional audio signals 304. If the wake expression is detected in only one of the directional audio signals, or in a relatively small set of the directional audio signals corresponding to a single direction, the classifier 310 may evaluate the wake expression as being uttered by the user 104. For example, it may be concluded that the wake expression has been uttered by the user if the wake expression occurs in multiple directions or directional signals that are within a single cone shape extending from an apex at the audio device.

**[0043]** In some cases, a user-uttered wake expression may be received from more than a single direction or directional cone due to acoustic reflections within the environment 102. Accordingly, the classifier 310 may be configured to determine that a wake expression has been uttered by the user 104 if the wake expression is detected in directional audio signals corresponding to two different directions, which may be represented by two cone shapes extending from one or more apexes at the audio device. In some cases, the wake expression may be deemed to have been uttered by the user if the wake expression is found in less than all of the directional audio signals 304, or if the wake expression is found in a number of the directional audio signals 304 that is less than a threshold number. Similarly, the classifier 310

may conclude that a wake expression has been generated by the speaker 110 if all or a majority of the directional audio signals 304 are identified by the expression recognizers 306 as being likely to contain the wake expression.

**[0044]** In some implementations, the expression recognizers 306 may produce non-binary indications regarding whether the wake expression is likely to be present in the corresponding directional audio signals 304. For example, each expression recognizer 306 may provide a confidence level indicating the likelihood or probability that the wake expression is present in the corresponding directional audio signal 304. The classifier may compare the received confidence levels to predetermined thresholds or may use other means to evaluate whether the wake expression is present in each of the directional audio signals.

**[0045]** In some situations, the classifier 310 may be configured to recognize a pattern or set of the directional audio signals 304 that typically contain the wake expression when the wake expression has been generated by the speaker 110. A reference pattern or signal set may in some cases be identified in an initialization procedure by generating the wake expression at the speaker 110 and concurrently recording which of the directional audio signals 304 are then identified as containing the wake expression. The identified signals are then considered members of the reference set. During normal operation, the classifier 310 may conclude that a detected wake expression has been generated by the speaker 110 when the observed pattern or signal set has the same members as the reference pattern or signal set.

**[0046]** If the classifier 310 determines that a detected wake expression has been uttered by the user 104, and not generated by the speaker 110, the classifier 310 generates or provides a wake event or wake notification 312. The wake event 312

may be provided to the speech processing components 120, to the operating system 116, and/or to various of the applications 134.

**[0047]** FIG. 4 illustrates further techniques that may be used in some environments for evaluating whether a wake expression has been uttered by a user or has been self-generated. In this case, a classifier 402 receives various parameters 404 relating to received audio, generated audio, and other operational aspects of the audio device 106, and distinguishes between user-uttered wake expressions and self-generated wake expressions based on the parameters 404.

**[0048]** The parameters 404 utilized by the classifier 402 may include recognition parameters 404(a) such as might be generated by the expression recognizers 306 of FIG. 3. The recognition parameters 404(a) may comprise confidence levels corresponding respectively to each of the directional audio signals. Each of the recognition parameters 404(a) may indicate the likelihood of the corresponding directional audio signal 304 containing the wake expression. Confidence values or likelihoods may be indicated as values on a continuous scale, such as percentages that range from 0% to 100%.

**[0049]** The parameters 404 may also include echo or echo-related parameters 404(b) that indicate the amount of echo present in each of the directional audio signals or the amount of echo reduction that has been applied to each of the directional audio signals. These parameters may be provided by the echo cancellation component 122 (FIG. 1) with respect to each of the directional audio signals 304 or to the directional audio signals collectively. The echo-related parameters 404(b) may be indicated as values on a continuous scale, such as by percentages ranging from 0% to 100%.

**[0050]** The parameters 404 may also include loudness parameters 404(c), indicating the current loudness or volume level at which audio is being generated by

the speaker 110 and/or the loudness of each of the received directional audio signals. As with the previously described parameters, the loudness parameters 404(c) may be indicated as values on a continuous scale, such as a percentage that ranges from 0% to 100%. Loudness may be evaluated on the basis of amplitudes of the signals, such as the amplitude of the output audio signal or the amplitudes of the input audio signals.

**[0051]** The parameters 404 may include informational parameters 404(d), indicating other aspects of the audio device 102. For example, the informational parameters 404(d) may indicate whether speech or other audio (which may or may not contain the wake expression) is currently being produced by the speaker 110. Similarly, the informational parameters 404(d) may indicate whether the wake expression is currently being generated by the text-to-speech component 132 of the audio device 106 or is otherwise known to be present in the output of the speaker 110.

**[0052]** The parameters 404 may be evaluated collectively to distinguish between wake expressions that have been uttered by a user and wake expressions that have been produced by a device speaker. As examples, the following factors may indicate the probability of a speaker-generated wake expression:

the speaker is known to be producing speech, music, or other audio;

high speaker volume;

low degree of echo cancellation;

high wake expression recognition confidence in many directions; and

high input audio volume levels from many directions.

**[0053]** Similarly, the following factors may indicate the probability of a user-generated wake expression:

the speaker is not producing speech, music, or other audio;

low speaker volume;

high degree of echo cancellation;

high wake expression recognition confidence in one or two of the directional audio signals; and

high input audio volume levels from one or two directions.

**[0054]** The classifier 402 may be configured to compare the parameters 404 to a set of reference parameters 406 to determine whether a detected wake expression has been uttered by the user 104 or whether the wake expression has been generated by the speaker 110. The classifier 310 may generate the wake event 312 if the received parameters 404 match or are within specified tolerances of the reference parameters.

**[0055]** The reference parameters 406 may be provided by a system designer based on known characteristics of the audio device 106 and/or its environment. Alternatively, the reference parameters may be learned in a training or machine learning procedure, an example of which is described below with reference to FIG. 5. The reference parameters 406 may be specified as specific values, as values and allowed deviations, and/or as ranges of allowable values.

**[0056]** The wake event 312 may comprise a simple notification that the wake expression has occurred. Alternatively, the wake event 312 may comprise or be accompanied by information allowing the audio device 106 or applications 134 to evaluate whether the wake expression has occurred. For example, the wake event 312 may indicate or be accompanied by a confidence level, indicating the evaluated probability that the wake expression has occurred. A confidence level may indicate probability on a continuous scale, such as from 0% to 100%. The applications 134 may respond to the wake event in different ways, depending on the confidence level. For example, an application may respond to a low confidence level by lowering the volume of output audio so that a repeated utterance of the wake expression is more

likely to be detected. As another example, an application may respond to a wake event having a low confidence level by verbally prompting the user for confirmation. As another example, an application may alter its behavior over time in light of receiving wake events with low confidence levels.

**[0057]** The wake event 312 may indicate other information. For example, the wake event 312 may indicate the identity of the user who has uttered the wake expression. As another example, the wake event 312 may indicate which of multiple available wake expressions has been detected. As a further example, the wake event 312 may include recognition parameters 404 or other parameters based on or related to the recognition parameters 404.

**[0058]** FIG. 5 illustrates an example method 500 that may be used to learn or generate the reference parameters 406. In some cases, the example method 500 may be implemented as machine learning to dynamically learn which of the directional audio signals are likely to contain the wake expression when the wake expression is known to occur in the output audio. In other cases, the example method may be implemented as machine learning to dynamically learn various parameters and/or ranges of parameters that may be used to detect generation of the wake expression by the speaker 110 of the audio device 106.

**[0059]** An action 502 comprises producing or generating the wake expression at the speaker 110. The action 502 may be performed upon startup or initialization of the audio device 106 and/or at other times during operation of the audio device 106. In some implementations, the action 502 may comprise generating the wake expression as part of responding to user commands. For example, the wake expression may be contained in speech generated by the speech generation component 132, and may be generated as part of providing services or responses to the user 104.

The audio device 106 may be configured to learn the reference parameters or to refine the reference parameters in response to any such known generation of the wake expression by the speaker 110.

**[0060]** An action 504 comprises receiving input audio at the microphone array 108. Because of acoustic computing between the speaker 110 and the microphone array 108, the input audio contains the wake expression generated in the action 502.

**[0061]** An action 506 comprises producing and/or receiving directional audio signals based on the received input audio. The directional audio signals may in some embodiments be produced by beamforming techniques. In other embodiments, the directional audio signals may be produced by other techniques, such as by directional microphones or microphones placed in different areas of a room.

**[0062]** An action 508 comprises performing wake expression detection with respect to each of the produced or received directional audio signals. The action 508 may comprise evaluating the produced or received directional audio signals to generate respectively corresponding indications of whether the directional audio signals contain the wake expression. Detection of the wake expression in an individual directional audio signal may be indicated by recognition parameters as described above, which may comprise binary values or non-binary probabilities.

**[0063]** An action 510 comprises receiving recognition parameters, such as the recognition parameters 404(a) described above with reference to FIG. 4, which may include the results of the wake expression detection 508. In some implementations, the recognition parameters may indicate a set of the directional audio signals in which the wake expression has been detected. In other implementations, the recognition parameters may comprise probabilities with respect to each of the directional audio

signals, where each probability indicates the likelihood that the corresponding directional audio signal contains the wake expression.

**[0064]** The action 510 may also comprise receiving other parameters or indications, such as the echo parameters 404(b), the loudness parameters 404(c), and the information parameters 404(d), described above with reference to FIG. 4.

**[0065]** An action 512 may comprise generating and saving a set of reference parameters, based on the parameters received in the action 510. The reference parameters may include the values of the parameters 404 at the time the wake expression is detected. The method 500 may be performed repeatedly or continuously, during operation of the audio device, to tune and retune the learned reference parameters.

**[0066]** FIG. 6 shows a process 600 of detecting a wake expression and determining whether it has been uttered by the user 104 or generated by the audio device 106.

**[0067]** An action 602 comprises producing output audio at the speaker 110 in the user environment 102. The output audio may comprise generated speech, music, or other content, which may be generated by the audio device 106 or received from other content sources. The output audio may from time to time include the wake expression.

**[0068]** An action 604 comprises receiving input audio, which may include components of the output audio due to acoustic coupling between the speaker 110 and the microphone array 108. The input audio may also include speech uttered by the user 104, which may include the wake expression.

**[0069]** An action 606 comprises producing and/or receiving a plurality of directional audio signals corresponding to input audio from different areas of the user

environment 102. The directional audio signals contain audio components from different areas or portions of the user environment 102, such as from different radial directions relative to the audio device 106. The directional audio signals may be produced using beamforming techniques based on an array of non-directional microphones, or may be received respectively from a plurality of directional microphones.

**[0070]** An action 608 comprises generating and/or receiving device parameters or indications relating to operation of the audio device 106. In some embodiments, the action 608 may comprise evaluating the directional audio signals to generate respectively corresponding recognition parameters or other indications of whether the directional audio signals contain the wake expression. The parameters or indications may also include parameters relating to speech generation, output audio generation, echo cancelation, etc.

**[0071]** An action 610 comprises evaluating the device parameters or indications to determine whether the wake expression has occurred in the input audio, based at least in part on expression recognition parameters. This may comprise determining whether the wake expression has occurred in any one or more of the directional audio signals, and may be performed by the individual expression recognizers 306 of FIG. 3.

**[0072]** If the wake expression has not occurred, no further action is taken. If the wake expression has occurred in at least one of the directional audio signals, an action 612 is performed. The action 612 comprises determining when a detected occurrence of the wake expression in the input audio is a result of the wake expression occurring in the output audio and/or of being produced by the speaker 110 of the audio device

106. The action 612 is based at least in part on the recognition parameters generated by the action 608.

**[0073]** In some embodiments, the determination 612 may be made in light of the number or pattern of the directional audio signals in which the wake expression is found. For example, detecting the wake expression in all or a majority of the directional audio signals may be considered an indication that the wake expression has been generated by the speaker 110, while detection of the wake expression in less than a majority of the directional audio signals may be considered an indication that the wake expression has been generated by a user who is located in a particular direction relative to the audio device 106. As another example, the action 612 may comprise identifying a number of the directional audio signals that are likely to contain the wake expression, and comparing the number to a threshold. More specifically, the wake expression may be considered to have been uttered by the user if the number of directional signals identified as being likely to contain the threshold is less than or equal to a threshold of one or two.

**[0074]** As another example, the action 612 may comprise identifying a set of the directional audio signals that are likely to contain the wake expression and comparing the identified set to a predetermined set of the directional audio signals, wherein the predetermined set includes directional audio signals that are known to contain the wake expression when the wake expression occurs in the output audio. The predetermined set may be learned in an initialization process or at other times when the audio device 106 is known to be producing the wake expression. More particularly, a learning procedure may be used to determine a particular set of the directional audio signals which can be expected to contain the wake expression when the wake expression has been produced from the speaker 110. Similarly, a learning

procedure may be used to determine a pattern or group of the directional audio signals which can be expected to contain the wake expression when the wake expression has been uttered by the user.

**[0075]** As another example, the pattern of directional audio signals in which the wake expression is detected may be analyzed to determine whether the wake expression was received as an omnidirectional input or whether it was received from a single direction corresponding to the position of a user. In some cases, a user-uttered wake expression may also be received as an audio reflection from a reflective surface. Accordingly, a wake expression originating from two distinct directions may in some cases be evaluated as being uttered by the user.

**[0076]** Certain embodiments may utilize more complex analyses in the action 612, with reference to a set of reference parameters 614. The reference parameters 614 may be specified by a system designer, or may comprise parameters that have been learned as described above with reference to FIG. 5. The reference parameters may include expression recognition parameters indicating which of the directional audio signals contain or are likely to contain the wake expression. The reference parameters may also include parameters relating to speech generation, output audio generation, echo cancelation, and so forth. Machine learning techniques, including neural networks, fuzzy logic, and Bayesian classification, may be used to formulate the reference parameters and/or to perform comparisons of current parameters with the reference parameters.

**[0077]** Learned reference parameters may be used in situations in which the audio produced by or received from a device speaker is not omnidirectional. Situations such as this may result from acoustic reflections or other anomalies, and/or in embodiments where the speaker of a device is directional rather than omnidirectional. In some

embodiments, a beamforming speaker, sometimes referred to as a sound bar, may be used to customize speaker output for optimum performance in the context of the unique acoustic properties of a particular environment. For example, the directionality of the speaker may be configured to minimize reflections and to optimize the ability to detect user uttered audio.

**[0078]** If the action 612 determines that a detected wake expression has been produced by the speaker 110, an action 516 is performed, which comprises ignoring the wake expression. Otherwise, if the action 612 determines that the detected wake expression has been uttered by the user 104, an action 618 is performed. The action 618 comprises declaring a wake event. The audio device 106 may respond to a declared wake event by interpreting and acting upon subsequently detected user speech.

**[0079]** The embodiments described above may be implemented programmatically, such as with computers, processors, as digital signal processors, analog processors, and so forth. In other embodiments, however, one or more of the components, functions, or elements may be implemented using specialized or dedicated circuits, including analog circuits and/or digital logic circuits. The term "component", as used herein, is intended to include any hardware, software, logic, or combinations of the foregoing that are used to implement the functionality attributed to the component.

**[0080]** Although the subject matter has been described in language specific to structural features, it is to be understood that the subject matter defined in the appended claims is not necessarily limited to the specific features described. Rather, the specific features are disclosed as illustrative forms of implementing the claims.

## CLAUSES

1. An audio device configured to respond to a trigger expression uttered by a user, comprising:

a speaker configured to generate output audio;

a microphone array configured to produce a plurality of input audio signals;

an audio beamformer configured to produce a plurality of directional audio signals based at least in part on the input audio signals, wherein the directional audio signals represent audio from respectively corresponding directions relative to the audio device;

one or more speech recognition components configured to detect whether the predefined expression occurs in the audio represented by each of the respective directional audio signals; and

an expression detector configured to (a) determine that the trigger expression has been uttered by the user if the trigger expression occurs in the audio represented by less than a threshold number of the directional audio signals; and (b) determine that the predefined expression has been not been generated by the speaker if the predefined expression occurs in the audio represented by all of the directional audio signals.

2. The audio device of clause1, wherein the expression detector is further configured to determine that the trigger expression has been uttered by the user if the trigger expression occurs in the audio from multiple directions that are within a single cone shape extending from an apex at the audio device.

3. The audio device of clause 1, wherein the expression detector is further configured to determine that the predefined expression has been uttered by the user if the predefined expression occurs in the audio from multiple directions that are within two cone shapes extending from apexes at the audio device.

4. The audio device of clause 1, wherein the expression detector is further configured to determine that the predefined expression has been generated by the speaker if the predefined expression occurs in the audio represented by more than half of the directional audio signals.

5. A method comprising:

producing output audio in a user environment;

receiving a plurality of audio signals representing input audio from respectively corresponding portions of the user environment;

generating one or more recognition parameters indicating which one or more of the directional audio signals contain a predefined expression; and

determining that an occurrence of the predefined expression in the input audio is a result of the predefined expression occurring in the output audio based at least in part on the one or more recognition parameters.

6. The method of clause 5, wherein the determining comprises:

determining whether the one or more recognition parameters indicate that all of the audio signals represent input audio containing the predefined expression; and

determining that the occurrence of the predefined expression in the input audio is a result of the predefined expression occurring in the output audio if the one or

more recognition parameters indicate that all of the input audio signals represent input audio containing the predefined expression.

7. The method of clause 5, wherein the determining comprises:

identifying a number of the audio signals that represent input audio containing the predefined expression based at least in part on the recognition parameters; and

determining that the occurrence of the predefined expression in the input audio is a result of the predefined expression occurring in the output audio if the number exceeds a threshold.

8. The method of clause 5, wherein:

the recognition parameters comprise individual parameters corresponding respectively to the audio signals;

each individual parameter indicates whether the corresponding audio signal represents input audio containing the predefined expression;

the determining further comprises identifying, based at least in part on the individual parameters, a number of the audio signals that represent input audio containing the predefined expression; and

the determining further comprises determining that the occurrence of the predefined expression in the input audio is a result of the predefined expression occurring in the output audio if the number exceeds a threshold.

9. The method of clause 5, wherein the determining comprises:

identifying an observed signal set, wherein the observed signal set has one or more members comprising one or more of the audio signals that are indicated by the

one or more recognition parameters to represent input audio containing the predefined expression;

determining that the occurrence of the predefined expression in the input audio is a result of the predefined expression occurring in the output audio if the observed signal set and a reference signal set have the same one or more members; and

wherein the reference signal set has one or more members comprising one or more of the audio signals that contain the predefined expression during an occurrence of the predefined expression in the output audio.

10. The method of clause 9, further comprising identifying the one or more members of the reference signal set during a known occurrence of the predefined expression in the output audio, wherein the one or more members of the reference signal set comprise one or more of the audio signals that are indicated by the one or more recognition parameters to represent input audio containing the predefined expression during the known occurrence of the predefined expression in the output audio.

11. The method of clause 5, wherein the one or more recognition parameters indicate one or more of the following:

loudness of the output audio;

whether the output audio is known to contain speech;

loudness of the input audio; or

echo characteristics of the audio signals.

12. The method of clause 11, further comprising using machine learning to perform the determining.

13. The method of clause 5, wherein the one or more recognition parameters correspond respectively to the directional audio signals, and wherein each of the one or more recognition parameters indicates whether the predefined expression is present in the corresponding audio signal.

14. The method of clause 5, wherein the one or more recognition parameters correspond respectively to the audio signals, and wherein each of the one or more recognition parameters indicates a probability of the predefined expression being present in the corresponding audio signal.

15. An audio device comprising:

one or more processors;

memory storing computer-executable instructions that, when executed by one or more processors, cause the one or more processors to perform acts comprising:

receiving a plurality of audio signals representing input audio from respectively corresponding portions of a user environment;

evaluating the audio signals to generate indications corresponding respectively to the audio signals, wherein each indication indicates whether the input audio represented by the corresponding audio signal contains a predefined expression; and

evaluating the indications to distinguish between utterance of the predefined expression by a user and production of the predefined expression by an audio speaker based at least in part on which one or more of the audio signals represent input audio containing the predefined expression.

16. The audio device of clause 15, wherein each of the indications comprises a binary parameter indicating whether the predefined expression occurs in the input audio represented by the corresponding audio signal .

17. The audio device of clause 15, wherein each of the indications comprises a probability that the predefined expression occurs in the input audio represented by the corresponding audio signal.

18. The audio device of clause 15, the acts further comprising:

identifying an observed signal set, wherein the observed signal set has one or more members comprising one or more of the audio signals that represent input audio containing the predefined expression;

wherein the evaluating comprises determining that an occurrence of the predefined expression in the input audio is a result of the predefined expression being produced by the audio speaker if the observed signal set and a reference signal set have the same one or more members; and

wherein the one or more members of the reference signal set comprise one or more of the audio signals that contain the predefined expression during production of the predefined expression by the audio speaker.

19. The audio device of clause 18, the acts further comprising identifying the one or more members of the reference signal set during known production of the predefined expression by the audio speaker, wherein the one or members of the reference signal set comprise one or more of the audio signals that are indicated by the indications to contain the predefined expression during the known production of the predefined expression by the output speaker.

20. The one audio device of clause 15, the acts further comprising generating a wake event indicating a probability of whether the predefined expression has been uttered by the user.

# CLAIMS

1. A method comprising:

producing output audio in a user environment;

receiving a plurality of audio signals representing input audio from respectively corresponding portions of the user environment;

generating one or more recognition parameters indicating which one or more of the directional audio signals contain a predefined expression; and

determining that an occurrence of the predefined expression in the input audio is a result of the predefined expression occurring in the output audio based at least in part on the one or more recognition parameters.

2. The method of claim 1, wherein the determining comprises:

determining whether the one or more recognition parameters indicate that all of the audio signals represent input audio containing the predefined expression; and

determining that the occurrence of the predefined expression in the input audio is a result of the predefined expression occurring in the output audio if the one or more recognition parameters indicate that all of the input audio signals represent input audio containing the predefined expression.

3. The method of claim 1, wherein the determining comprises:

identifying a number of the audio signals that represent input audio containing the predefined expression based at least in part on the recognition parameters; and

determining that the occurrence of the predefined expression in the input audio is a result of the predefined expression occurring in the output audio if the number exceeds a threshold.

4. The method of claim 1, wherein:

the recognition parameters comprise individual parameters corresponding respectively to the audio signals;

each individual parameter indicates whether the corresponding audio signal represents input audio containing the predefined expression;

the determining further comprises identifying, based at least in part on the individual parameters, a number of the audio signals that represent input audio containing the predefined expression; and

the determining further comprises determining that the occurrence of the predefined expression in the input audio is a result of the predefined expression occurring in the output audio if the number exceeds a threshold.

5. The method of claim 1, wherein the determining comprises:

identifying an observed signal set, wherein the observed signal set has one or more members comprising one or more of the audio signals that are indicated by the one or more recognition parameters to represent input audio containing the predefined expression;

determining that the occurrence of the predefined expression in the input audio is a result of the predefined expression occurring in the output audio if the observed signal set and a reference signal set have the same one or more members; and

wherein the reference signal set has one or more members comprising one or more of the audio signals that contain the predefined expression during an occurrence of the predefined expression in the output audio.

6. The method of claim 5, further comprising identifying the one or more members of the reference signal set during a known occurrence of the predefined

expression in the output audio, wherein the one or more members of the reference signal set comprise one or more of the audio signals that are indicated by the one or more recognition parameters to represent input audio containing the predefined expression during the known occurrence of the predefined expression in the output audio.

7. The method of claim 1, wherein the one or more recognition parameters indicate one or more of the following:

loudness of the output audio;

whether the output audio is known to contain speech;

loudness of the input audio; or

echo characteristics of the audio signals.

8. The method of claim 1, wherein the one or more recognition parameters correspond respectively to the directional audio signals, and wherein each of the one or more recognition parameters indicates whether the predefined expression is present in the corresponding audio signal.

9. The method of claim 1, wherein the one or more recognition parameters correspond respectively to the audio signals, and wherein each of the one or more recognition parameters indicates a probability of the predefined expression being present in the corresponding audio signal.

10. An audio device comprising:

one or more processors;

memory storing computer-executable instructions that, when executed by one or more processors, cause the one or more processors to perform acts comprising:

receiving a plurality of audio signals representing input audio from respectively corresponding portions of a user environment;

evaluating the audio signals to generate indications corresponding respectively to the audio signals, wherein each indication indicates whether the input audio represented by the corresponding audio signal contains a predefined expression; and

evaluating the indications to distinguish between utterance of the predefined expression by a user and production of the predefined expression by an audio speaker based at least in part on which one or more of the audio signals represent input audio containing the predefined expression.

11. The audio device of claim 10, wherein each of the indications comprises a binary parameter indicating whether the predefined expression occurs in the input audio represented by the corresponding audio signal .

12. The audio device of claim 10, wherein each of the indications comprises a probability that the predefined expression occurs in the input audio represented by the corresponding audio signal.

13. The audio device of claim 10, the acts further comprising:

identifying an observed signal set, wherein the observed signal set has one or more members comprising one or more of the audio signals that represent input audio containing the predefined expression;

wherein the evaluating comprises determining that an occurrence of the predefined expression in the input audio is a result of the predefined expression being produced by the audio speaker if the observed signal set and a reference signal set have the same one or more members; and

wherein the one or more members of the reference signal set comprise one or more of the audio signals that contain the predefined expression during production of the predefined expression by the audio speaker.

14. The audio device of claim 13, the acts further comprising identifying the one or more members of the reference signal set during known production of the predefined expression by the audio speaker, wherein the one or members of the reference signal set comprise one or more of the audio signals that are indicated by the indications to contain the predefined expression during the known production of the predefined expression by the output speaker.

15. The one audio device of claim 10, the acts further comprising generating a wake event indicating a probability of whether the predefined expression has been uttered by the user.

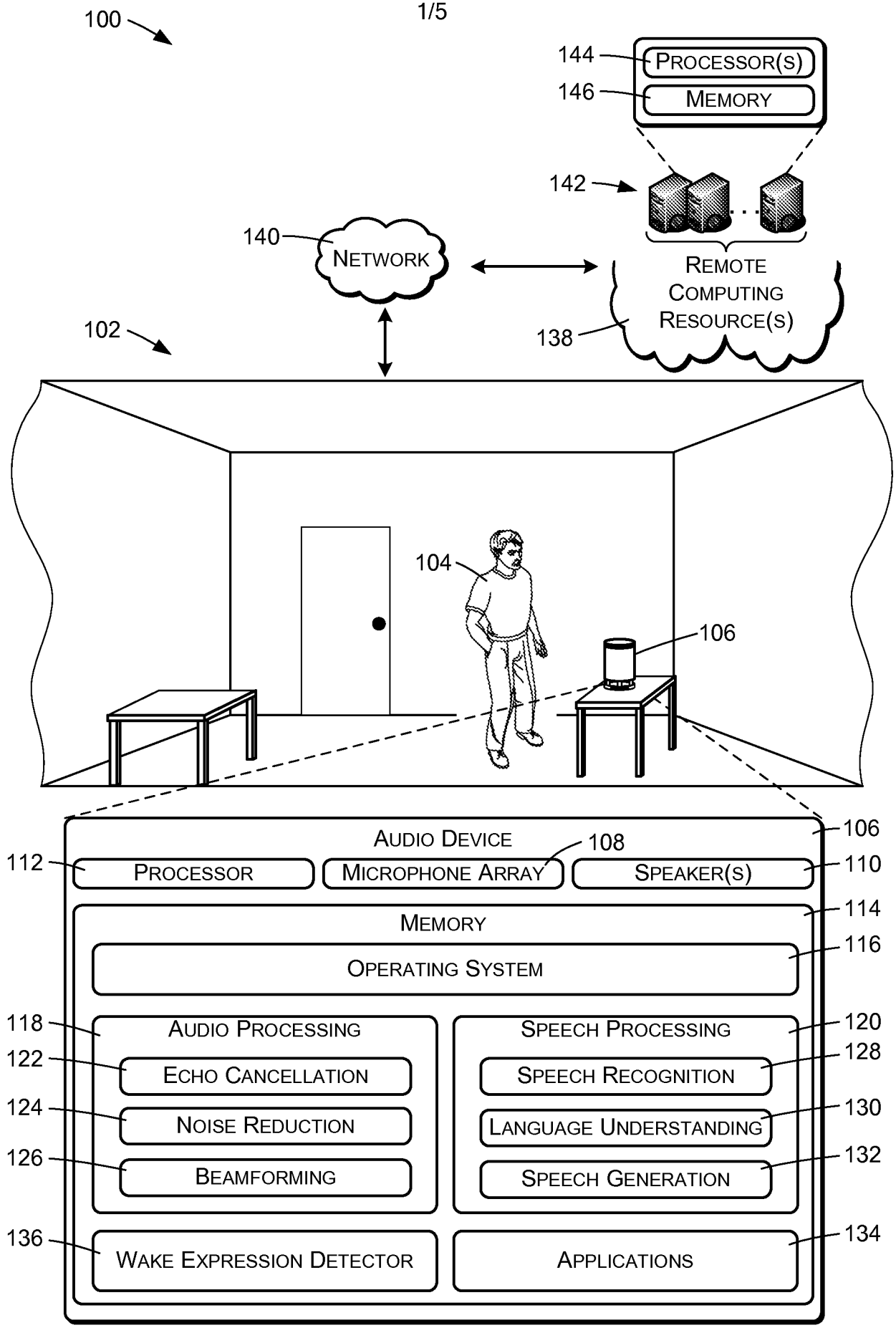
100
144
PROCESSOR(S)
146
MEMORY
142
140
NETWORK
REMOTE COMPUTING RESOURCE(S)
138
102
104
106
AUDIO DEVICE
108
112
PROCESSOR
MICROPHONE ARRAY
SPEAKER(S)
110
114
MEMORY
116
OPERATING SYSTEM
118
AUDIO PROCESSING
120
SPEECH PROCESSING
122
ECHO CANCELLATION
128
SPEECH RECOGNITION
124
NOISE REDUCTION
130
LANGUAGE UNDERSTANDING
126
BEAMFORMING
132
SPEECH GENERATION
136
WAKE EXPRESSION DETECTOR
APPLICATIONS
134

FIG. 1

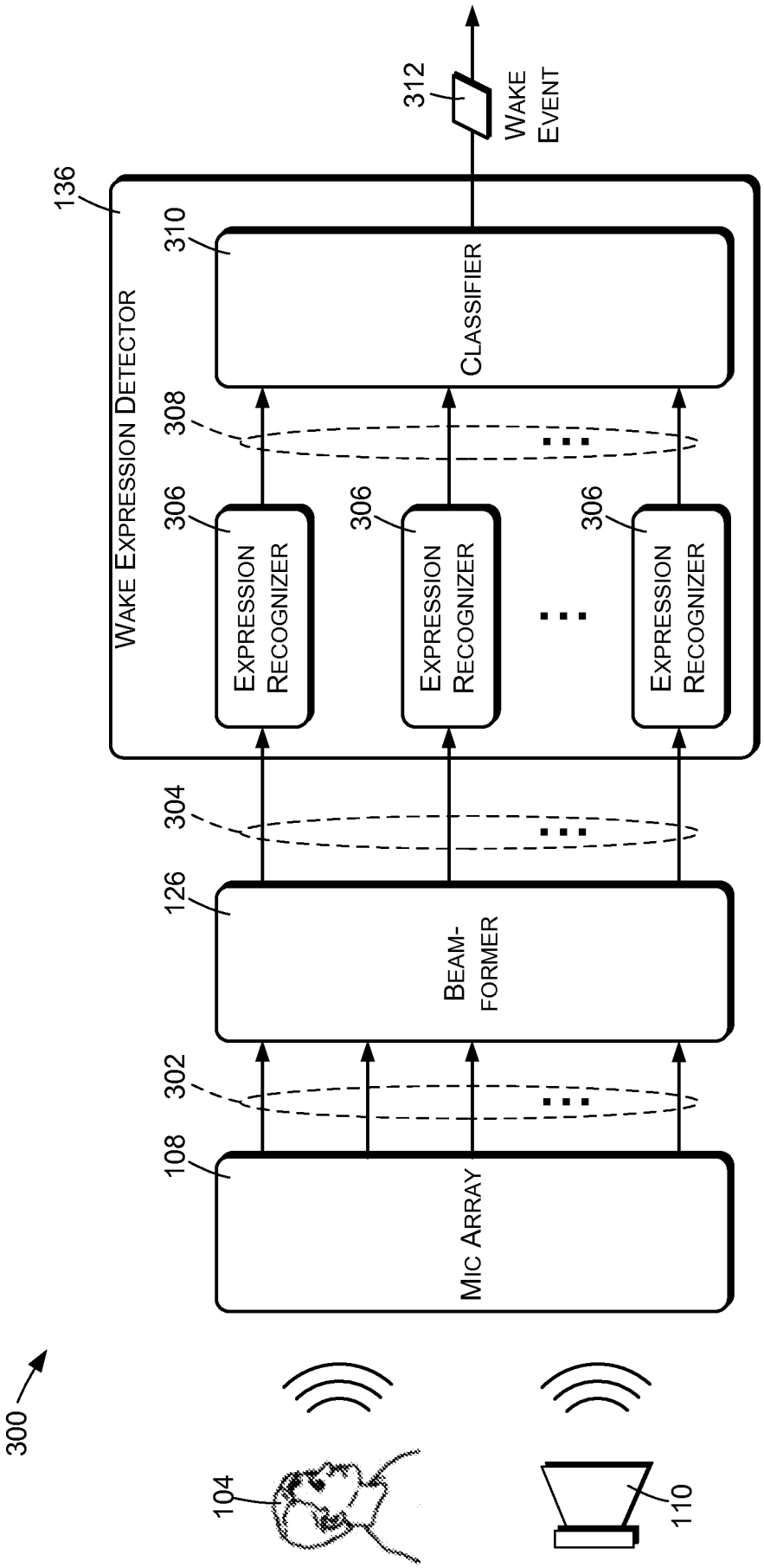
300
104
110
MIC ARRAY
108
302
BEAM-FORMER
126
304
WAKE EXPRESSION DETECTOR
136
EXPRESSION RECOGNIZER
306
EXPRESSION RECOGNIZER
306
EXPRESSION RECOGNIZER
306
308
CLASSIFIER
310
WAKE EVENT
312

FIG. 3

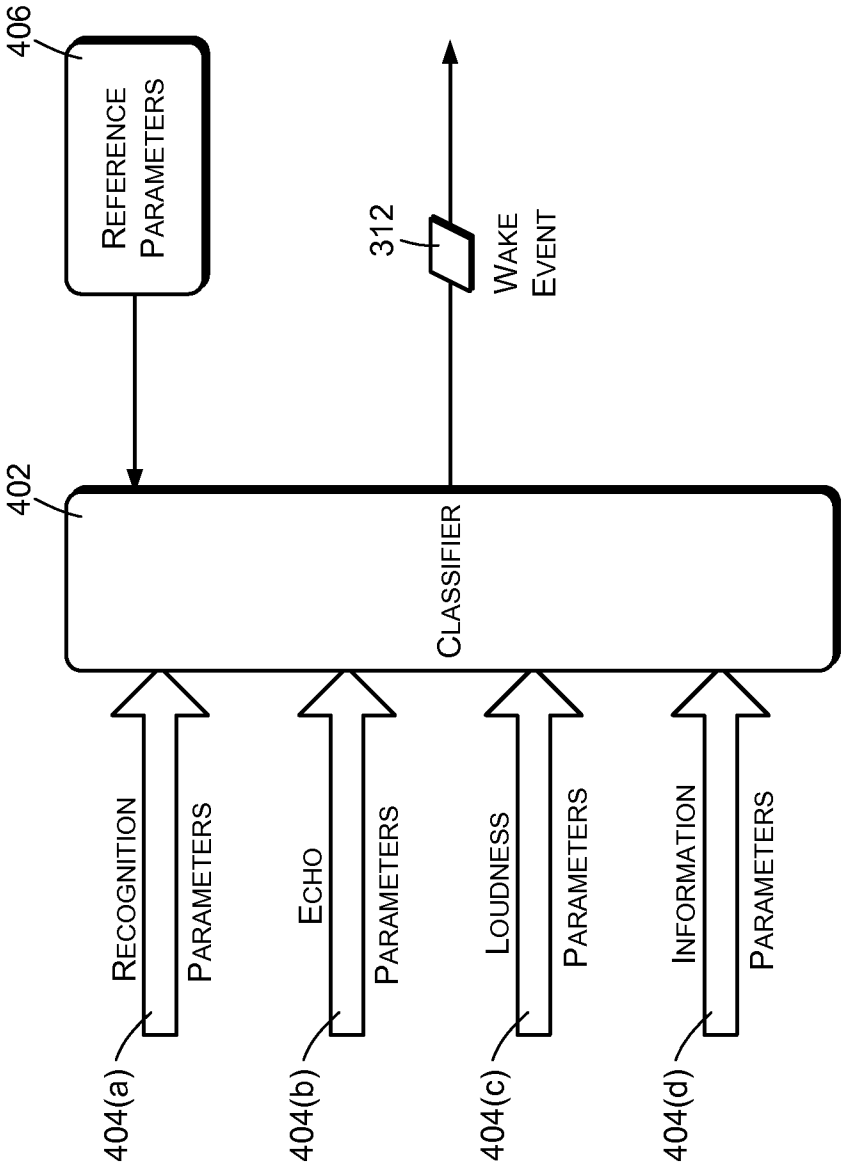
406
REFERENCE PARAMETERS
312
WAKE EVENT
402
CLASSIFIER
RECOGNITION PARAMETERS
ECHO PARAMETERS
LOUDNESS PARAMETERS
INFORMATION PARAMETERS
404(a)
404(b)
404(c)
404(d)

FIG. 4

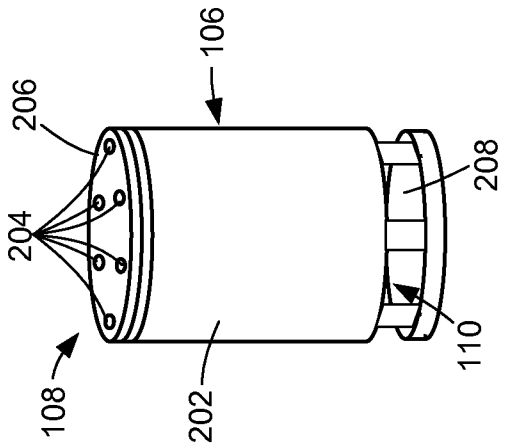
106
206
204
208
108
202
110

FIG. 2

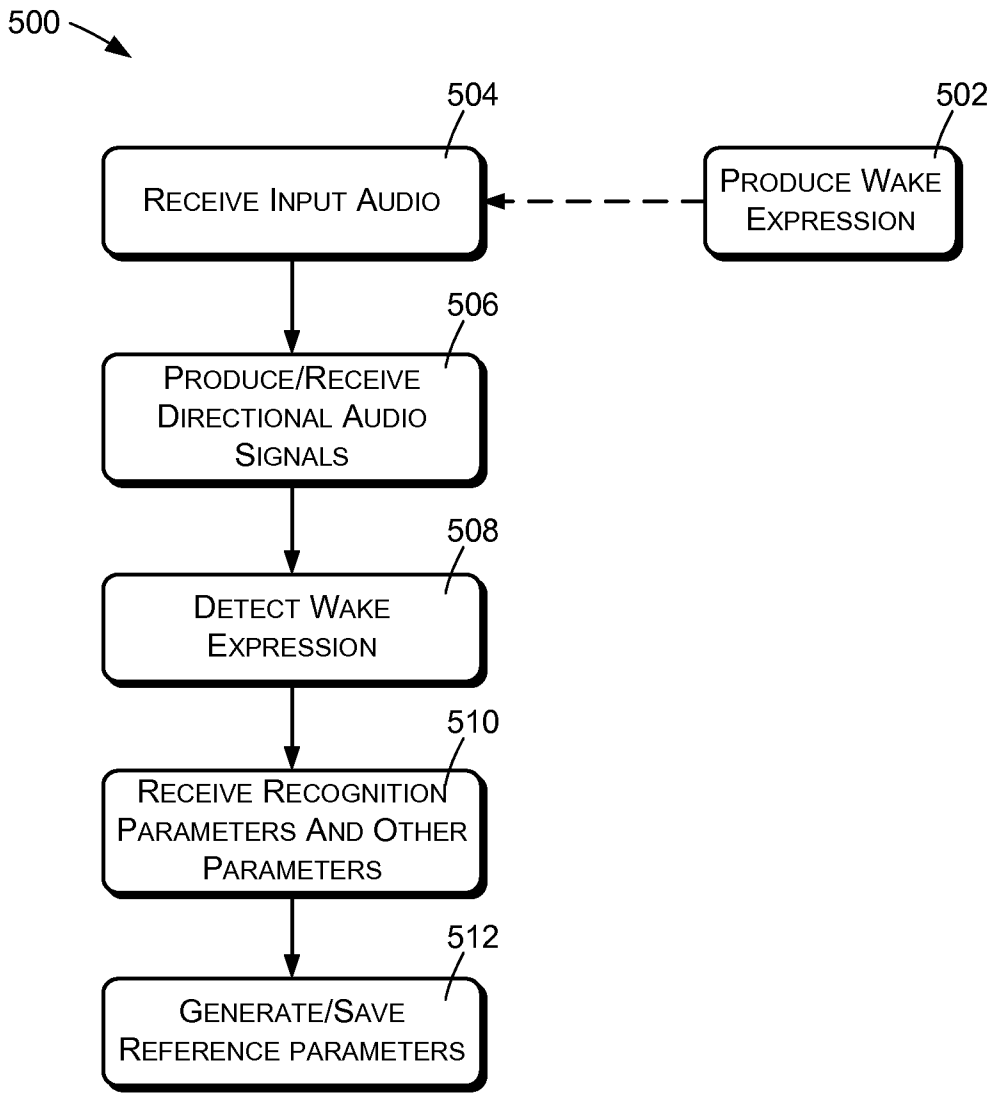
500
504
502
RECEIVE INPUT AUDIO
PRODUCE WAKE EXPRESSION
506
PRODUCE/RECEIVE DIRECTIONAL AUDIO SIGNALS
508
DETECT WAKE EXPRESSION
510
RECEIVE RECOGNITION PARAMETERS AND OTHER PARAMETERS
512
GENERATE/SAVE REFERENCE PARAMETERS

FIG. 5

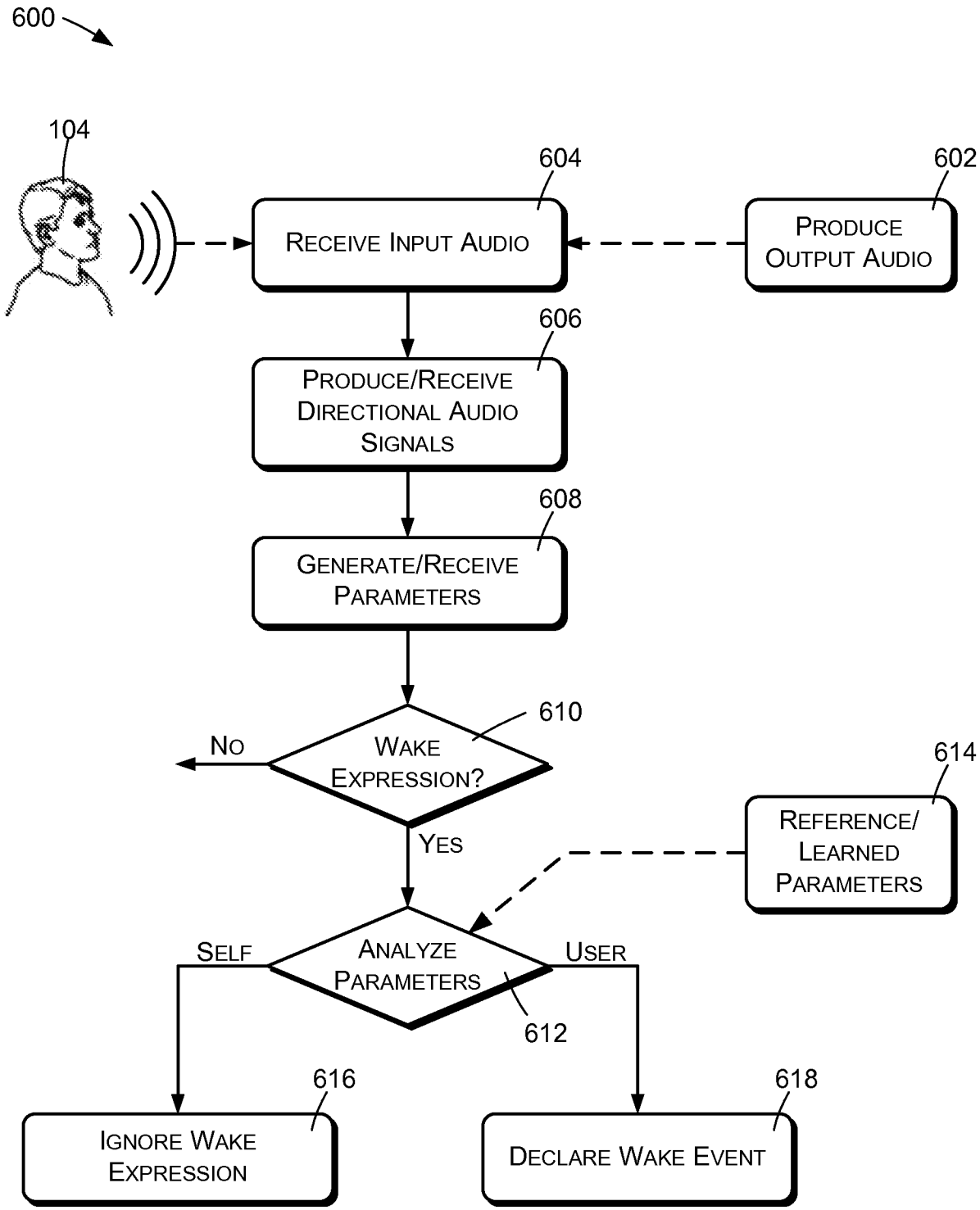
600
104
604
602
RECEIVE INPUT AUDIO
PRODUCE OUTPUT AUDIO
606
PRODUCE/RECEIVE DIRECTIONAL AUDIO SIGNALS
608
GENERATE/RECEIVE PARAMETERS
610
NO
WAKE EXPRESSION?
YES
614
REFERENCE/ LEARNED PARAMETERS
SELF
ANALYZE PARAMETERS
USER
612
616
618
IGNORE WAKE EXPRESSION
DECLARE WAKE EVENT

FIG. 6