



- (51) International Patent Classification:
G10L 25/78 (2013.01) *G10L 21/0216* (2013.01)
- (21) International Application Number:
PCT/US2013/058551
- (22) International Filing Date:
6 September 2013 (06.09.2013)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
13/631,716 28 September 2012 (28.09.2012) US
13/840,136 15 March 2013 (15.03.2013) US
- (71) Applicant: APPLE INC. [US/US]; 1 Infinite Loop, Cupertino, CA 95014 (US).
- (72) Inventors: DUSAN, Sorin, V.; 1 Infinite Loop, M/S 81-1TF, Cupertino, CA 95014 (US). ANDERSEN, Esge, B.; 1 Infinite Loop, M/S 26-AE, Cupertino, CA 95014 (US). LINDAHL, Aram; 1 Infinite Loop, M/S 305-3AE, Cupertino, CA 95014 (US). BRIGHT, Andrew, P.; 1 Infinite Loop, M/S 26-AE, Cupertino, CA 95014 (US).
- (74) Agents: SCHELLER, James, C. et al.; Blakely Sokoloff Taylor & Zafman LLP, 1279 Oakmead Parkway, Sunnyvale, CA 94085 (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, 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:

— with international search report (Art. 21(3))

(54) Title: SYSTEM AND METHOD OF DETECTING A USER'S VOICE ACTIVITY USING AN ACCELEROMETER

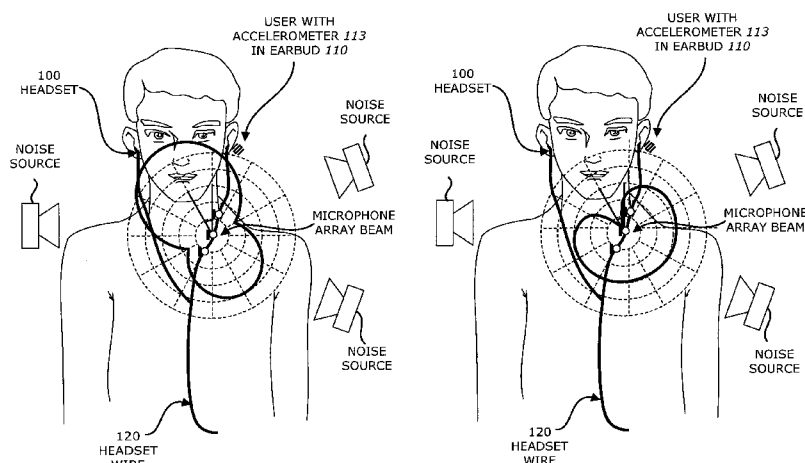


FIG. 1

(57) Abstract: A method of detecting a user's voice activity in a mobile device is described herein. The method starts with a voice activity detector (VAD) generating a VAD output based on (i) acoustic signals received from microphones included in the mobile device and (ii) data output by an inertial sensor that is included in an earphone portion of the mobile device. The inertial sensor may detect vibration of the user's vocal chords modulated by the user's vocal tract based on vibrations in bones and tissue of the user's head. A noise suppressor may then receive the acoustic signals from the microphones and the VAD output and suppress the noise included in the acoustic signals received from the microphones based on the VAD output. The method may also include steering one or more beamformers based on the VAD output. Other embodiments are also described.

SYSTEM AND METHOD OF DETECTING A USER'S VOICE ACTIVITY
USING AN ACCELEROMETER

CROSS REFERENCED APPLICATIONS

[001] This application is a continuation-in-part application of U.S. patent application No. 13/631,716, filed on September 28, 2012, currently pending, the entire contents of which are incorporated herein by reference.

FIELD

[002] An embodiment of the invention relate generally to an electronic device having a voice activity detector (VAD) that uses signals from an accelerometer included in the earbuds of a headset with a microphone array to detect the user's speech and to steer at least one beamformer. Another embodiment of the invention relates generally to an electronic device ("mobile device") having a VAD that uses signals from an accelerometer included in an earphone portion of the mobile device to detect the user's speech.

BACKGROUND

[003] Currently, a number of consumer electronic devices are adapted to receive speech via microphone ports or headsets. While the typical example is a portable telecommunications device (mobile telephone), with the advent of Voice over IP (VoIP), desktop computers, laptop computers and tablet computers may also be used to perform voice communications.

[004] When using these electronic devices, the user also has the option of using the speakerphone mode or a wired headset to receive his speech. However, a common complaint with these hands-free modes of operation is that the speech captured by the microphone port or the headset includes environmental noise such as secondary speakers in the background or other background noises. This environmental noise often renders the user's speech unintelligible and thus, degrades the quality of the voice communication.

[005] Similarly, when these electronic devices are used in a non-speaker phone mode which requires the user to hold the electronic device's earphone portion to the user's ear ("at ear position"), the speech that is captured by the microphone port may also be rendered unintelligible due to environmental noise.

SUMMARY

[006] Generally, the invention relates to using signals from an accelerometer included in an earbud of an enhanced headset for use with electronic devices to detect a user's voice activity. Being placed in the user's ear canal, the accelerometer may detect speech caused by the vibrations of the user's vocal chords. Using these signals from the accelerometer in combination with the acoustic signals received by microphones in the earbuds and a microphone array in the

headset wire, a coincidence defined as a “AND” function between a movement detected by the accelerometer and the voiced speech in the acoustic signals may indicate that the user’s voiced speech is detected. When a coincidence is obtained, a voice activity detector (VAD) output may indicate that the user’s voiced speech is detected. In addition to the user’s voiced speech, the user’s speech may also include unvoiced speech, which is speech that is generated without vocal chord vibrations (e.g., sounds such as /s/, /sh/, /f/). In order for the VAD output to indicate that unvoiced speech is detected, a signal from a microphone in the earbuds or a microphone in the microphone array or the output of a beamformer may be used. A high-pass filter is applied to the signal from the microphone or beamformer and if the resulting power is above a threshold, the VAD output may indicate the user’s unvoiced speech is detected. A noise suppressor may receive the acoustic signals as received from the microphone array beamformer and may suppress the noise from the acoustic signals or beamformer based on the VAD output. Further, based on this VAD output, one or more beamformers may also be steered such that the microphones in the earbuds and in the microphone array emphasize the user’s speech signals and deemphasize the environmental noise.

[007] In one embodiment of the invention, a method of detecting a user’s voice activity in a headset with a microphone array starts with a voice activity detector (VAD) generating a VAD output based on (i) acoustic signals received from microphones included in a pair of earbuds and the microphone array included on a headset wire and (ii) data output by a sensor detecting movement that is included in the pair of earbuds. The headset may include the pair of earbuds and the headset wire. The VAD output may be generated by detecting speech included in the acoustic signals, detecting a user’s speech vibrations from the data output by the accelerometer, coincidence of the detected speech in acoustic signals and the user’s speech vibrations, and setting the VAD output to indicate that the user’s voiced speech is detected if the coincidence is detected and setting the VAD output to indicate that the user’s voiced speech is not detected if the coincidence is not detected. A noise suppressor may then receive (i) the acoustic signals from the microphone array and (ii) the VAD output and suppress the noise included in the acoustic signals received from the microphone array based on the VAD output. The method may also include steering one or more beamformers based on the VAD output. The beamformers may be adaptively steered or the beamformers may be fixed and steered to a set location.

[008] In another embodiment of the invention, a system detecting a user’s voice activity comprises a headset, a voice activity detector (VAD) and a noise suppressor. The headset may include a pair of earbuds and a headset wire. Each of the earbuds may include earbud microphones and a sensor detecting movement such as an accelerometer. The headset wire may

include a microphone array. The VAD may be coupled to the headset and may generate a VAD output based on (i) acoustic signals received from the earbud microphones, the microphone array or beamformer and (ii) data output by the sensor detecting movement. The noise suppressor may be coupled to the headset and the VAD and may suppress noise from the acoustic signals from the microphone array based on the VAD output.

[009] In another embodiment of the invention, a method of detecting a user's voice activity in a mobile device starts with a voice activity detector (VAD) generating a VAD output based on (i) acoustic signals received from microphones included in the mobile device and (ii) data output by an inertial sensor that is included in an earphone portion of the mobile device, the inertial sensor to detect vibration of the user's vocal chords modulated by the user's vocal tract based on vibrations in bones and tissue of the user's head. In this embodiment, the inertial sensor being located in the earphone portion of the mobile device may detect the vibrations being detected at the user's ear or in the area proximate to the user's ear.

[0010] The above summary does not include an exhaustive list of all aspects of the present invention. It is contemplated that the invention includes all systems, apparatuses and methods that can be practiced from all suitable combinations of the various aspects summarized above, as well as those disclosed in the Detailed Description below and particularly pointed out in the claims filed with the application. Such combinations may have particular advantages not specifically recited in the above summary.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] The embodiments of the invention are illustrated by way of example and not by way of limitation in the figures of the accompanying drawings in which like references indicate similar elements. It should be noted that references to "an" or "one" embodiment of the invention in this disclosure are not necessarily to the same embodiment, and they mean at least one. In the drawings:

[0012] **Figure 1** illustrates an example of the headset in use according to one embodiment of the invention.

[0013] **Figure 2** illustrates an example of the right side of the headset used with a consumer electronic device in which an embodiment of the invention may be implemented.

[0014] **Figure 3** illustrates a block diagram of a system detecting a user's voice activity according to a first embodiment of the invention.

[0015] **Figure 4** illustrates a flow diagram of an example method of detecting a user's voice activity according to the first embodiment of the invention.

[0016] **Figure 5** illustrates a block diagram of a system detecting a user's voice activity according to a second embodiment of the invention.

[0017] **Figure 6** illustrates a flow diagram of an example method of detecting a user's voice activity according to the second embodiment of the invention.

[0018] **Figure 7** illustrates a block diagram of a system detecting a user's voice activity according to a third embodiment of the invention.

[0019] **Figure 8** illustrates a flow diagram of an example method of detecting a user's voice activity according to the third embodiment of the invention.

[0020] **Figure 9** illustrates a block diagram of a system detecting a user's voice activity according to a fourth embodiment of the invention.

[0021] **Figure 10** illustrates a flow diagram of an example method of detecting a user's voice activity according to the fourth embodiment of the invention.

[0022] **Figure 11** illustrates a block diagram of a system detecting a user's voice activity according to a fifth embodiment of the invention.

[0023] **Figure 12** illustrates a flow diagram of an example method of detecting a user's voice activity according to the fifth embodiment of the invention.

[0024] **Figure 13** illustrates an example of the headset in use according to the fifth embodiment of the invention.

[0025] **Figure 14** illustrates a block diagram of a system detecting a user's voice activity according to a sixth embodiment of the invention.

[0026] **Figure 15** illustrates a flow diagram of an example method of detecting a user's voice activity according to the sixth embodiment of the invention.

[0027] **Figure 16** illustrates an example of the headset in use according to the sixth embodiment of the invention.

[0028] **Figure 17** is a block diagram of exemplary components of an electronic device detecting a user's voice activity in accordance with aspects of the present disclosure.

[0029] **Figure 18** is a perspective view of an electronic device in the form of a computer, in accordance with aspects of the present disclosure.

[0030] **Figure 19** is a front-view of a portable handheld electronic device, in accordance with aspects of the present disclosure.

[0031] **Figure 20** is a perspective view of a tablet-style electronic device that may be used in conjunction with aspects of the present disclosure.

[0032] **Figure 21** shows a perspective view of a mobile device according to a seventh embodiment of the invention.

[0033] **Figure 22** is a block diagram of a system detecting a user's voice activity according to the seventh embodiment of the invention.

[0034] **Figure 23** illustrates a flow diagram of an example method of detecting a user's voice activity according to the seventh embodiment of the invention.

DETAILED DESCRIPTION

[0035] In the following description, numerous specific details are set forth. However, it is understood that embodiments of the invention may be practiced without these specific details. In other instances, well-known circuits, structures, and techniques have not been shown to avoid obscuring the understanding of this description.

[0036] Moreover, the following embodiments of the invention may be described as a process, which is usually depicted as a flowchart, a flow diagram, a structure diagram, or a block diagram. Although a flowchart may describe the operations as a sequential process, many of the operations can be performed in parallel or concurrently. In addition, the order of the operations may be re-arranged. A process is terminated when its operations are completed. A process may correspond to a method, a procedure, etc.

[0037] **Figure 1** illustrates an example of a headset in use that may be coupled with a consumer electronic device according to one embodiment of the invention. As shown in **Figures 1 and 2**, the headset 100 includes a pair of earbuds 110 and a headset wire 120. The user may place one or both the earbuds 110 into his ears and the microphones in the headset may receive his speech. The microphones may be air interface sound pickup devices that convert sound into an electrical signal. The headset 100 in **Figure 1** is double-earpiece headset. It is understood that single-earpiece or monaural headsets may also be used. As the user is using the headset to transmit his speech, environmental noise may also be present (e.g., noise sources in **Figure 1**). While the headset 100 in **Figure 2** is an in-ear type of headset that includes a pair of earbuds 110 which are placed inside the user's ears, respectively, it is understood that headsets that include a pair of earcups that are placed over the user's ears may also be used. Additionally, embodiments of the invention may also use other types of headsets.

[0038] **Figure 2** illustrates an example of the right side of the headset used with a consumer electronic device in which an embodiment of the invention may be implemented. It is understood that a similar configuration may be included in the left side of the headset 100.

[0039] As shown in **Figure 2**, the earbud 110 includes a speaker 112, a sensor detecting movement such as an accelerometer 113, a front microphone 111_F that faces the direction of the eardrum and a rear microphone 111_R that faces the opposite direction of the eardrum. The earbud 110 is coupled to the headset wire 120, which may include a plurality of microphones 121₁-121_M (M>1) distributed along the headset wire that can form one or more microphone

arrays. As shown in **Figure 1**, the microphone arrays in the headset wire 120 may be used to create microphone array beams (i.e., beamformers) which can be steered to a given direction by emphasizing and deemphasizing selected microphones 121₁-121_M. Similarly, the microphone arrays can also exhibit or provide nulls in other given directions. Accordingly, the beamforming process, also referred to as spatial filtering, may be a signal processing technique using the microphone array for directional sound reception. The headset 100 may also include one or more integrated circuits and a jack to connect the headset 100 to the electronic device (not shown) using digital signals, which may be sampled and quantized.

[0040] When the user speaks, his speech signals may include voiced speech and unvoiced speech. Voiced speech is speech that is generated with excitation or vibration of the user's vocal chords. In contrast, unvoiced speech is speech that is generated without excitation of the user's vocal chords. For example, unvoiced speech sounds include /s/, /sh/, /f/, etc. Accordingly, in some embodiments, both the types of speech (voiced and unvoiced) are detected in order to generate an augmented voice activity detector (VAD) output which more faithfully represents the user's speech.

[0041] First, in order to detect the user's voiced speech, in one embodiment of the invention, the output data signal from accelerometer 113 placed in each earbud 110 together with the signals from the front microphone 111_F, the rear microphone 111_R, the microphone array 121₁-121_M or the beamformer may be used. The accelerometer 113 may be a sensing device that measures proper acceleration in three directions, X, Y, and Z or in only one or two directions. When the user is generating voiced speech, the vibrations of the user's vocal chords are filtered by the vocal tract and cause vibrations in the bones of the user's head which is detected by the accelerometer 113 in the headset 110. In other embodiments, an inertial sensor, a force sensor or a position, orientation and movement sensor may be used in lieu of the accelerometer 113 in the headset 110.

[0042] In the embodiment with the accelerometer 113, the accelerometer 113 is used to detect the low frequencies since the low frequencies include the user's voiced speech signals. For example, the accelerometer 113 may be tuned such that it is sensitive to the frequency band range that is below 2000Hz. In one embodiment, the signals below 60Hz-70Hz may be filtered out using a high-pass filter and above 2000Hz-3000Hz may be filtered out using a low-pass filter. In one embodiment, the sampling rate of the accelerometer may be 2000Hz but in other embodiments, the sampling rate may be between 2000Hz and 6000Hz. In another embodiment, the accelerometer 113 may be tuned to a frequency band range under 1000Hz. It is understood that the dynamic range may be optimized to provide more resolution within a forced range that is expected to be produced by the bone conduction effect in the headset 100. Based on the outputs

of the accelerometer 113, an accelerometer-based VAD output (VADa) may be generated, which indicates whether or not the accelerometer 113 detected speech generated by the vibrations of the vocal chords. In one embodiment, the power or energy level of the outputs of the accelerometer 113 is assessed to determine whether the vibration of the vocal chords is detected. The power may be compared to a threshold level that indicates the vibrations are found in the outputs of the accelerometer 113. In another embodiment, the VADa signal indicating voiced speech is computed using the normalized cross-correlation between any pair of the accelerometer signals (e.g. X and Y, X and Z, or Y and Z). If the cross-correlation has values exceeding a threshold within a short delay interval the VADa indicates that the voiced speech is detected. In some embodiments, the VADa is a binary output that is generated as a voice activity detector (VAD), wherein 1 indicates that the vibrations of the vocal chords have been detected and 0 indicates that no vibrations of the vocal chords have been detected.

[0043] Using at least one of the microphones in the headset 110 (e.g., one of the microphones in the microphone array 121_I-121_M, front earbud microphone 111_F, or back earbud microphone 111_R) or the output of a beamformer, a microphone-based VAD output (VADm) may be generated by the VAD to indicate whether or not speech is detected. This determination may be based on an analysis of the power or energy present in the acoustic signal received by the microphone. The power in the acoustic signal may be compared to a threshold that indicates that speech is present. In another embodiment, the VADm signal indicating speech is computed using the normalized cross-correlation between any pair of the microphone signals (e.g. 121_I and 121_M). If the cross-correlation has values exceeding a threshold within a short delay interval the VADm indicates that the speech is detected. In some embodiments, the VADm is a binary output that is generated as a voice activity detector (VAD), wherein 1 indicates that the speech has been detected in the acoustic signals and 0 indicates that no speech has been detected in the acoustic signals.

[0044] Both the VADa and the VADm may be subject to erroneous detections of voiced speech. For instance, the VADa may falsely identify the movement of the user or the headset 100 as being vibrations of the vocal chords while the VADm may falsely identify noises in the environment as being speech in the acoustic signals. Accordingly, in one embodiment, the VAD output (VADv) is set to indicate that the user's voiced speech is detected (e.g., VADv output is set to 1) if the coincidence between the detected speech in acoustic signals (e.g., VADm) and the user's speech vibrations from the accelerometer output data signals is detected (e.g., VADa). Conversely, the VAD output is set to indicate that the user's voiced speech is not detected (e.g., VADv output is set to 0) if this coincidence is not detected. In other words, the VADv output is obtained by applying an AND function to the VADa and VADm outputs.

[0045] Second, the signal from at least one of the microphones in the headset 100 or the output from the beamformer may be used to generate a VAD output for unvoiced speech (VADu), which indicates whether or not unvoiced speech is detected. It is understood that the VADu output may be affected by environmental noise since it is computed only based on an analysis of the acoustic signals received from a microphone in the headset 100 or from the beamformer. In one embodiment, the signal from the microphone closest in proximity to the user's mouth or the output of the beamformer is used to generate the VADu output. In this embodiment, the VAD may apply a high-pass filter to this signal to compute high frequency energies from the microphone or beamformer signal. When the energy envelope in the high frequency band (e.g. between 2000Hz and 8000Hz) is above certain threshold the VADu signal is set to 1 to indicate that unvoiced speech is present. Otherwise, the VADu signal may be set to 0 to indicate that unvoiced speech is not detected. Voiced speech can also set VADu to 1 if significant energy is detected at high frequencies. This has no negative consequences since the VADv and VADu are further combined in an "OR" manner as described below.

[0046] Accordingly, in order to take into account both the voiced and unvoiced speech and to further be more robust to errors, the method may generate a VAD output by combining the VADv and VADu outputs using an OR function. In other words, the VAD output may be augmented to indicate that the user's speech is detected when VADv indicates that voiced speech is detected or VADu indicates that unvoiced speech is detected. Further, when this augmented VAD output is 0, this indicates that the user is not speaking and thus a noise suppressor may apply a supplementary attenuation to the acoustic signals received from the microphones or from beamformer in order to achieve additional suppression of the environmental noise.

[0047] The VAD output may be used in a number of ways. For instance, in one embodiment, a noise suppressor may estimate the user's speech when the VAD output is set to 1 and may estimate the environmental noise when the VAD output is set to 0. In another embodiment, when the VAD output is set to 1, one microphone array may detect the direction of the user's mouth and steer a beamformer in the direction of the user's mouth to capture the user's speech while another microphone array may steer a cardioid or other beamforming patterns in the opposite direction of the user's mouth to capture the environmental noise with as little contamination of the user's speech as possible. In this embodiment, when the VAD output is set to 0, one or more microphone arrays may detect the direction and steer a second beamformer in the direction of the main noise source or in the direction of the individual noise sources from the environment.

[0048] The latter embodiment is illustrated in **Figure 1**, the user in the left part of **Figure 1** is speaking while the user in the right part of **Figure 1** is not speaking. When the VAD output

is set to 1, at least one of the microphone arrays is enabled to detect the direction of the user's mouth. The same or another microphone array creates a beamforming pattern in the direction of the user's mouth, which is used to capture the user's speech. Accordingly, the beamformer outputs an enhanced speech signal. When the VAD output is 0, the same or another microphone array may create a cardioid beamforming pattern in the direction opposite to the user's mouth, which is used to capture the environmental noise. When the VAD output is 0, other microphone arrays may create beamforming patterns (not shown in **Figure 1**) in the directions of individual environmental noise sources. When the VAD output is 0, the microphone arrays is not enabled to detect the direction of the user's mouth, but rather the beamformer is maintained at its previous setting. In this manner, the VAD output is used to detect and track both the user's speech and the environmental noise.

[0049] The microphone arrays are generating beams in the direction of the mouth of the user in the left part of **Figure 1** to capture the user's speech and in the direction opposite to the direction of the user's mouth in the right part of **Figure 1** to capture the environmental noise.

[0050] **Figure 3** illustrates a block diagram of a system detecting a user's voice activity according to a first embodiment of the invention. The system 300 in **Figure 3** includes the headset having the pair of earbuds 110 and the headset wire and an electronic device that includes a VAD 130 and a noise suppressor 140. As shown in **Figure 3**, the VAD 130 receives the accelerometer's 113 output signals that provide information on sensed vibrations in the x, y, and z directions and the acoustic signals received from the microphones 111_F, 111_R and microphone array 121_I-121_M. It is understood that a plurality of microphone arrays (beamformers) on the headset wire 120 may also provide acoustic signals to the VAD 130 and the noise suppressor 140.

[0051] The accelerometer signals may be first pre-conditioned. First, the accelerometer signals are pre-conditioned by removing the DC component and the low frequency components by applying a high pass filter with a cut-off frequency of 60Hz-70 Hz, for example. Second, the stationary noise is removed from the accelerometer signals by applying a spectral subtraction method for noise suppression. Third, the cross-talk or echo introduced in the accelerometer signals by the speakers in the earbuds may also be removed. This cross-talk or echo suppression can employ any known methods for echo cancellation. Once the accelerometer signals are pre-conditioned, the VAD 130 may use these signals to generate the VAD output. In one embodiment, the VAD output is generated by using one of the X, Y, Z accelerometer signals which shows the highest sensitivity to the user's speech or by adding the three accelerometer signals and computing the power envelope for the resulting signal. When the power envelope is above a given threshold, the VAD output is set to 1, otherwise is set to 0. In another

embodiment, the VAD signal indicating voiced speech is computed using the normalized cross-correlation between any pair of the accelerometer signals (e.g. X and Y, X and Z, or Y and Z). If the cross-correlation has values exceeding a threshold within a short delay interval the VAD indicates that the voiced speech is detected. In another embodiment, the VAD output is generated by computing the coincidence as a “AND” function between the VAD_m from one of the microphone signals or beamformer output and the VAD_a from one or more of the accelerometer signals (VAD_a). This coincidence between the VAD_m from the microphones and the VAD_a from the accelerometer signals ensures that the VAD is set to 1 only when both signals display significant correlated energy, such as the case when the user is speaking. In another embodiment, when at least one of the accelerometer signal (e.g., x, y, z) indicates that user's speech is detected and is greater than a required threshold and the acoustic signals received from the microphones also indicates that user's speech is detected and is also greater than the required threshold, the VAD output is set to 1, otherwise is set to 0.

[0052] The noise suppressor 140 receives and uses the VAD output to estimate the noise from the vicinity of the user and remove the noise from the signals captured by at least one of the microphones 121₁-121_M in the microphone array. By using the data signals outputted from the accelerometers 113 further increases the accuracy of the VAD output and hence, the noise suppression. Since the acoustic signals received from the microphones 121₁-121_M and 111_F, 111_R may wrongly indicate that speech is detected when, in fact, environmental noises including voices (i.e., distractors or second talkers) in the background are detected, the VAD 130 may more accurately detect the user's voiced speech by looking for coincidence of vibrations of the user's vocal chords in the data signals from the accelerometers 113 when the acoustic signals indicate a positive detection of speech.

[0053] **Figure 4** illustrates a flow diagram of an example method of detecting a user's voice activity according to the first embodiment of the invention. Method 400 starts with a VAD detector 130 generating a VAD output based on (i) acoustic signals received from microphones 111_F, 111_R included in a pair of earbuds 110 and the microphone array 121₁-121_M included on a headset wire 120 and (ii) data output by a sensor detecting movement 113 that is included in the pair of earbuds 120 (Block 401). At Block 402, a noise suppressor 140 receives the acoustic signals from the microphone array 121₁-121_M and (ii) the VAD output from the VAD detector 130. At Block 403, the noise suppressor may suppress the noise included in the acoustic signals received from the microphone array 121₁-121_M based on the VAD output.

[0054] **Figure 5** illustrates a block diagram of a system detecting a user's voice activity according to a second embodiment of the invention. The system 500 is similar to the system 300 in **Figure 3** but further includes a fixed beamformer 150 to receive the acoustic signals received

from the microphone array 121_1 - 121_M and its output is provided to the noise suppressor 140 and to the VAD Block 130. The fixed beamformer is steered in a direction of the user's mouth during a normal wearing position of the headset. This direction may be pre-defined setting in the headset 100. By steering the fixed beamformer in the direction of the user's mouth during a normal wearing position, the fixed beamformer may provide the user's speech signal with significant attenuation of the noises in the environment. Accordingly, the fixed beamformer outputs a main speech signal to the noise suppressor 140. In other embodiments, the microphone array based on the microphones 111_F , 111_R in the earbuds 110 and the plurality of microphones 121_1 - 121_M are generating and steering the fixed beamformer 150 in the direction of the mouth of the user as corresponding to normal wearing conditions.

[0055] **Figure 6** illustrates a flow diagram of an example method of detecting a user's voice activity according to the second embodiment of the invention. In this embodiment, after the VAD output is generated at Block 401 in **Figure 4**, the fixed beamformer 150 receives the acoustic signals from the microphone array at Block 601. The fixed beamformer 150 is then steered in the direction of the user's mouth during normal wearing position of the headset at Block 602 and the noise suppressor 140 receives the acoustic signals as outputted by the fixed beamformer 150 (i.e., the main speech signal). In this embodiment, the noise suppressor 140 may suppress the noise included in the acoustic signals as outputted by the fixed beamformer 150 as using the additional information in the VAD output received from the VAD 130.

[0056] **Figure 7** illustrates a block diagram of a system detecting a user's voice activity according to a third embodiment of the invention. Due to the user's movements and changing positions the headset 100 and the microphone arrays 121_1 - 121_M included therein may also change orientation with regards to the user's mouth. Thus, system 700 is similar to the system 300 in **Figure 3** but further includes a source direction detector 151 and a first beamformer 152 to implement voice-tracking principles. As shown in **Figure 7**, the source direction detector 151 also receives the VAD output from the VAD 130 as well as the acoustic signals from the microphone array 121_1 - 121_M . The source direction detector 151 may detect the user's speech source based on the VAD output and provide the direction of the user's speech source to the first beamformer 152. For instance, when the VAD output is set to indicate that the user's speech is detected (e.g., VAD output is set to 1), the source direction detector 151 estimates the direction of the user's mouth relative to the microphone array 121_1 - 121_M . Using this directional information from the source direction detector 151, when the VAD output is set to 1, the first beamformer 152 is adaptively steered in the direction of the user's speech source. The output of the first beamformer 152 may be the acoustic signals from the microphone array 121_1 - 121_M as captured by the first beamformer 152. As shown in **Figure 7**, the output of the first beamformer

152 may be the main speech signal that is then provided to the noise suppressor 140.

Accordingly, when the VAD output is set to 1, the source direction detector 151 computes the direction of user's mouth. Thus, the microphone array's beam direction can be adaptively adjusted when the VAD output is set to 1 to track the user's mouth direction. When the VAD output indicates that the user's speech is not detected (e.g., VAD output set to 0), the direction of the first beamformer 152 may be maintained at the direction corresponding to its position the last time the VAD output was set to 1.

[0057] In one embodiment, the source direction detector 151 may perform acoustic source localization based on time-delay estimates in which pairs of microphones included in the plurality of microphones 121_1 - 121_M and 111_F , 111_R in the headset 100 are used to estimate the delay for the sound signal between the two of the microphones. The delays from the pairs of microphones may also be combined and used to estimate the source location using methods such as the generalized cross-correlation (GCC) or adaptive eigenvalue decomposition (AED). In another embodiment, the source direction detector 151 and the first beamformer 152 may work in conjunction to perform the source localization based on steered beamforming (SBF). In this embodiment, the first beamformer 152 is steered over a range of directions and for each direction the power of the beamforming output is calculated. The power of the first beamformer 152 for each direction in the range of directions is calculated and the user's speech source is detected as the direction that has the highest power.

[0058] As shown in **Figure 7**, the noise suppressor 140 receives the output from the first beamformer 152 which is a main speech signal (i.e., the acoustic signals from the microphone array 121_1 - 121_M as captured by the first beamformer 152). In this embodiment, the noise suppressor 140 may suppress the noise included in the main speech signal based on the VAD output.

[0059] **Figure 8** illustrates a flow diagram of an example method of detecting a user's voice activity according to the third embodiment of the invention. In this embodiment, after the VAD output is generated at Block 401 in **Figure 4**, the source direction detector 151 receives the acoustic signals from the microphone array 121_1 - 121_M at Block 801 and detects the user's speech source based on the VAD output at Block 802. When the VAD output is set to indicate that the user's speech is detected, the first beamformer is adaptively steered in the direction of the detected user's speech source at Block 803. In this embodiment, the noise suppressor 140 may suppress the noise included in the acoustic signals as outputted by the first beamformer 152 (i.e., the main speech signal) based on the VAD output received from the VAD 130.

[0060] **Figure 9** illustrates a block diagram of a system detecting a user's voice activity according to a fourth embodiment of the invention. System 900 is similar to the system 700 in

Figure 7 but further includes a second beamformer 153 to provide a noise estimation of the environment noise that is present in the acoustic signals from the microphone array 121_I-121_M. As shown in **Figure 9**, the second beamformer 153 may have a cardioid pattern and may be adaptively steered with a null towards the mouth direction. In other words, the second beamformer 153 may be adaptively steered in a direction opposite to the mouth's direction to provide a signal representing an estimate of the environmental noise.

[0061] As shown in **Figure 9**, the noise suppressor 140 in this embodiment receives the outputs from the first beamformer 152 and the second beamformer 153 as well as the VAD output. Thus, the noise estimate from the second beamformer is provided to the noise suppressor 140 together with the user's speech signal included in the acoustic signals as outputted by the first beamformer. In this embodiment, the noise suppressor 140 may further suppress the noise included in the main speech signal outputted from the first beamformer 152 based on the outputs of the second beamformer 153 (i.e., the signal representing the environmental noise) and the VAD output.

[0062] Referring back to **Figure 1**, the adaptively steered first beamformer is illustrated on the left side of **Figure 1** while the adaptively steered second beamformer is illustrated on the right side of **Figure 1**. In this example, when the VAD output is set to 1, the first beamformer may be adaptively steered towards the user's mouth (e.g., left side of **Figure 1**) and the second different beamformer may be adaptively steered to form a cardioid pattern in the direction opposite to the user's mouth (e.g., right side of **Figure 1**). When the VAD output is set to 0, both the first and second beamformers 152, 153 may be maintained at the directions corresponding to their respective positions the last time the VAD output was set to 1.

[0063] **Figure 10** illustrates a flow diagram of an example method of detecting a user's voice activity according to the fourth embodiment of the invention. In this embodiment, after the first beamformer is adaptively steered in the direction of the detected user's speech source at Block 803 in **Figure 8**, the second beamformer 153 is adaptively steered with a null towards the detected user's speech source. In this embodiment, the second beamformer has a cardioid pattern and outputs a signal representing environmental noise when the VAD output is set to indicate that the user's speech is not detected. In this embodiment, the noise suppressor 140 may suppress the noise included in the main speech signal as outputted by the first beamformer 152 based on the noise estimate as outputted from the second beamformer 153 and the VAD output received from the VAD 130.

[0064] **Figure 11** illustrates a block diagram of a system detecting a user's voice activity according to a fifth embodiment of the invention. System 1100 is similar to the system 900 in **Figure 9** but in lieu of the second beamformer 153, system 1100 includes a third beamformer

154 to provide a noise estimation of the environment noise that is present in the acoustic signals from the microphone array 121₁-121_M. The third beamformer 154 differs from the second beamformer 153 in that the third beamformer 154 is used to detect the strongest environmental noise. The third beamformer 154 may then be adaptively steered in the direction of the strongest environmental noise location when the VAD output is set to indicate that the user's speech is not detected. Accordingly, the third beamformer 154 provides an estimate of the main environmental noise that is present in the acoustic signals from the microphone array 121₁-121_M. It is understood that the third beamformer 154 may also be adaptively steered to in a direction of a plurality of strongest environmental noise locations. In this embodiment, the noise suppressor 140 may suppress the noise included in the main speech signal as outputted by the first beamformer 152 based on the noise estimate of the main environmental noise as outputted from the third beamformer 154 and the VAD output received from the VAD 130.

[0065] **Figure 12** illustrates a flow diagram of an example method of detecting a user's voice activity according to the fifth embodiment of the invention. In this embodiment, after the first beamformer is adaptively steered in the direction of the detected user's speech source at Block 803 in **Figure 8**, the third beamformer 154 is adaptively steered in a direction of the strongest environmental noise location when the VAD output indicates that the user's speech is not detected. In this embodiment, the noise suppressor 140 receives a noise estimate of the main environmental noise from the third beamformer 154 and suppresses the noise included in the main speech signal as outputted from the first beamformer 152 based on the output from the third beamformer 154 and the VAD output.

[0066] **Figure 13** illustrates an example of the headset in use according to the fifth embodiment of the invention. In **Figure 13**, the voice tracking using the first beamformer 152 (e.g., left side of **Figure 13**) and noise tracking using the third beamformer 154 (e.g., right side of **Figure 13**) are illustrated. When the VAD output is set to 1, the first beamformer 152 is adaptively steered in the direction of the user's mouth (e.g., left side of **Figure 13**). When the VAD output is set to 0, the third beamformer 154 will detect the direction of the most significant noise source and be adaptively steered in this direction. Accordingly, this noise estimate may be passed together with the user's speech signal included in the output of the first beamformer 152 to the noise suppressor 140, which removes the noise based on the noise estimate and the VAD output. The noise suppressor 140 removes residual noise from main speech signal received from the first beamformer 152.

[0067] **Figure 14** illustrates a block diagram of a system detecting a user's voice activity according to a sixth embodiment of the invention. System 1400 is similar to the system 1100 in **Figure 11**, in that the third beamformer 154 is used to detect the direction of the strongest

environmental noise location when the VAD output indicates that the user's speech is not detected (e.g., VAD output is set to 0). However, in system 1400, the direction of the strongest environmental noise location detected by the third beamformer 154 is provided to the first beamformer 152 and the nulls of the first beamformer 152 may be adaptively steered towards the direction of the strongest environmental noise location while keeping the main beam of the first beamformer 152 in the direction of the user's mouth as detected when the VAD output is set to 1. The adaptive steering of the nulls of the first beamformer 152 may be performed when the VAD output is 1 or 0. Further, it is understood that the strongest environmental noise location may include one or more directions. In this embodiment, the noise suppressor 140 receives the main speech signal being outputted from the first beamformer 152. This main speech signal may include the acoustic signals from the microphones 121_I - 121_M as captured by the first beamformer 152 having a main beam directed to the user's mouth and nulls directed to the location(s) of the main environmental noise(s). In this embodiment, the noise suppressor 140 suppresses the noise included in the main speech signal outputted from the first beamformer 152 based on the VAD output.

[0068] **Figure 15** illustrates a flow diagram of an example method of detecting a user's voice activity according to the sixth embodiment of the invention. In this embodiment, after the first beamformer is adaptively steered in the direction of the detected user's speech source at Block 803 in **Figure 8**, the third beamformer 154 detects a direction of the strongest environmental noise location when the VAD output indicates that the user's speech is not detected at Block 1501. At Block 1502, the null of first beamformer 152 is adaptively steered in a direction of the strongest environmental noise location. In some embodiments, the nulls of the first beamformer 152 may be adaptively steered in the directions of a plurality of detected strongest environmental noise locations, respectively. The adaptive steering of the null(s) of the first beamformer 152 in Block 1502 may be performed when the VAD output indicates that the user's speech is detected or when the VAD output indicates that the user's speech is not detected. In this embodiment, the noise suppressor 140 suppresses the noise included in the main speech signal as outputted from the first beamformer 152 based on the VAD output.

[0069] **Figure 16** illustrates an example of the headset in use according to the sixth embodiment of the invention. As shown in **Figure 16**, when the VAD output is set to 1, the first beamformer 152 is adaptively steered such that the main beam is directed towards the user's mouth and maintained in that direction when the VAD output is set to 0. The third beamformer 154 detects the directions of the main environment noise locations when the VAD output is set to 0. Using the directions detected by the third beamformer 154, the nulls of the first beamformer 152 are adaptively steered in these directions of the main environment noise locations.

Accordingly, the first beamformer 152 emphasizes the user's speech using the main beam and deemphasizes the noise locations using the nulls.

[0070] A general description of suitable electronic devices for performing these functions is provided below with respect to **Figures 17-20**. Specifically, **Figure 17** is a block diagram depicting various components that may be present in electronic devices suitable for use with the present techniques. **Figure 18** depicts an example of a suitable electronic device in the form of a computer. **Figure 19** depicts another example of a suitable electronic device in the form of a handheld portable electronic device. Additionally, **Figure 20** depicts yet another example of a suitable electronic device in the form of a computing device having a tablet-style form factor. These types of electronic devices, as well as other electronic devices providing comparable voice communications capabilities (e.g., VoIP, telephone communications, etc.), may be used in conjunction with the present techniques.

[0071] Keeping the above points in mind, **Figure 17** is a block diagram illustrating components that may be present in one such electronic device 10, and which may allow the device 10 to function in accordance with the techniques discussed herein. The various functional blocks shown in **Figure 17** may include hardware elements (including circuitry), software elements (including computer code stored on a computer-readable medium, such as a hard drive or system memory), or a combination of both hardware and software elements. It should be noted that **Figure 17** is merely one example of a particular implementation and is merely intended to illustrate the types of components that may be present in the electronic device 10. For example, in the illustrated embodiment, these components may include a display 12, input/output (I/O) ports 14, input structures 16, one or more processors 18, memory device(s) 20, non-volatile storage 22, expansion card(s) 24, RF circuitry 26, and power source 28.

[0072] **Figure 18** illustrates an embodiment of the electronic device 10 in the form of a computer 30. The computer 30 may include computers that are generally portable (such as laptop, notebook, tablet, and handheld computers), as well as computers that are generally used in one place (such as conventional desktop computers, workstations, and servers). In certain embodiments, the electronic device 10 in the form of a computer may be a model of a MacBook[™], MacBook[™] Pro, MacBook Air[™], iMac[™], Mac[™] Mini, or Mac Pro[™], available from Apple Inc. of Cupertino, Calif. The depicted computer 30 includes a housing or enclosure 33, the display 12 (e.g., as an LCD 34 or some other suitable display), I/O ports 14, and input structures 16.

[0073] The electronic device 10 may also take the form of other types of devices, such as mobile telephones, media players, personal data organizers, handheld game platforms, cameras, and/or combinations of such devices. For instance, as generally depicted in **Figure 19**, the

device 10 may be provided in the form of a handheld electronic device 32 that includes various functionalities (such as the ability to take pictures, make telephone calls, access the Internet, communicate via email, record audio and/or video, listen to music, play games, connect to wireless networks, and so forth). By way of example, the handheld device 32 may be a model of an iPodTM, iPodTM Touch, or iPhoneTM available from Apple Inc.

[0074] In another embodiment, the electronic device 10 may also be provided in the form of a portable multi-function tablet computing device 50, as depicted in **Figure 20**. In certain embodiments, the tablet computing device 50 may provide the functionality of media player, a web browser, a cellular phone, a gaming platform, a personal data organizer, and so forth. By way of example, the tablet computing device 50 may be a model of an iPadTM tablet computer, available from Apple Inc.

[0075] **Figure 21** shows a perspective view of a mobile device 10 according to a seventh embodiment of the invention. In this embodiment, the mobile device 10 may be used in an at-ear position. The at-ear position is one in which the device 10 is being held to the user's ear. Referring to **Figure 21**, the mobile device 10 may include input-output components such as ports and jacks. For example, opening 61 may form the microphone port and opening 62 may form a speaker port. The sound during a telephone call is emitted through opening 63 which may form a speaker port for a telephone receiver that is placed adjacent to the user's ear during a call when the mobile device 10 is in the at-ear position. The portion of the mobile device 10 that is placed adjacent to the user's ear during a call when the mobile device 10 is in the at-ear position may be referred to as the earphone portion. Accordingly, in the at-ear position, the earpiece speaker port 63 may be used as a close-to-the-ear receiver port such that the sound during a telephone call is emitted through an earphone portion of the mobile device 10. When the mobile device 10 is in the at-ear position, the earphone speaker port 63 is "sealed" by the contact of the ear to the device housing the region surrounding the earphone speaker's opening 63. It should be noted that the closure of the ear around the speaker port 63 may not be perfectly "sealed," but such term is simply used to generally characterize the closed environment around the speaker port 63 formed by the ear and the device 10.

[0076] In one embodiment, the microphone port 61, the speaker ports 62 and 63 may be coupled to the communications circuitry to enable the user to participate in wireless telephone. In one embodiment, the microphone port 61 is coupled to microphones included in the mobile device 10. The microphones may be a microphone array similar to the microphone array 121_I-121_M in the headset 100 as described above. As further illustrated in **Figure 22**, the mobile device 10 may include an inertial sensor that is included in an earphone portion of the mobile device 10. The inertial sensor may be an accelerometer 114 that detects vibration of the user's

vocal chords modulated by the user's vocal tract based on vibrations in bones and tissue of the user's head. In one embodiment, the accelerometer 114 has a sampling rate greater than 2000Hz. In another embodiment, the sampling rate of the accelerometer 114 may be between 2000 Hz and 6000 Hz. By being included in the earphone portion of the mobile device 10, the accelerometer 114 may detect the vibrations of the user's vocal chords modulated by the user's vocal tract based on vibrations from portions of the user's ear and head that are in contact with the earphone portion of the mobile device 10 when the mobile device 10 is being used in an at-ear position.

[0077] **Figure 22** is a block diagram of a system 2200 detecting a user's voice activity according to a seventh embodiment of the invention. The system 2200 in **Figure 22** includes the mobile device 10 having a microphone array 122₁-122_M and an accelerometer included in the earphone portion of the mobile device 10. The system 2200 also includes a VAD 130 and a noise suppressor 140. In one embodiment, the VAD 130 and the noise suppressor 140 may be included the mobile device 10. In this embodiment, the components of system 2200 as illustrated in **Figure 22** are all included in the mobile device 10. As shown in **Figure 22**, the VAD 130 receives the accelerometer's 114 output signals that provide information on sensed vibrations in the x, y, and z directions and the acoustic signals received from the microphone array 122₁-122_M. It is understood that a plurality of microphone arrays (beamformers) in the mobile device 10 may also provide acoustic signals to the VAD 130 and the noise suppressor 140.

[0078] Similar to the embodiment in **Figure 3** as described above, the embodiment as illustrated in **Figure 22** may also pre-condition the accelerometer signals from accelerometer 114. Once the accelerometer 114's signals are pre-conditioned, the VAD 130 may use these signals to generate the VAD output as described in each embodiment described above. For instance, in one embodiment, the VAD output is generated by using one of the X, Y, Z accelerometer signals which shows the highest sensitivity to the user's speech or by adding the three accelerometer signals and computing the power envelope for the resulting signal. When the power envelope is above a given threshold, the VAD output is set to 1, otherwise is set to 0. In another embodiment, the VAD signal indicating voiced speech is computed using the normalized cross-correlation between any pair of the accelerometer signals (e.g. X and Y, X and Z, or Y and Z). If the cross-correlation has values exceeding a threshold within a short delay interval the VAD indicates that the voiced speech is detected. In another embodiment, the VAD output is generated by computing the coincidence as a "AND" function between the VAD_m from one of the microphone signals or beamformer output and the VAD_a from one or more of the accelerometer signals (VAD_a). This coincidence between the VAD_m from the microphones and the VAD_a from the accelerometer signals ensures that the VAD is set to 1 only when both

signals display significant correlated energy, such as the case when the user is speaking. In another embodiment, when at least one of the accelerometer signal (e.g., x, y, z) indicates that user's speech is detected and is greater than a required threshold and the acoustic signals received from the microphones also indicates that user's speech is detected and is also greater than the required threshold, the VAD output is set to 1, otherwise is set to 0.

[0079] As illustrated in **Figure 22**, the noise suppressor 140 receives and uses the VAD output to estimate the noise from the vicinity of the user and removes the noise from the signals captured by at least one of the microphones 122₁-122_M in the microphone array. By using the data signals outputted from the accelerometer 114 further increases the accuracy of the VAD output and hence, the noise suppression.

[0080] **Figure 23** illustrates a flow diagram of an example method of detecting a user's voice activity according to the seventh embodiment of the invention. Method 2300 starts with a VAD detector 130 generating a VAD output based on (i) acoustic signals received from microphones included in the mobile device 10 and (ii) data output by an inertial sensor 114 that is included in an earphone portion of the mobile device 10 (Block 2301). The microphones included in the mobile device 10 may be a microphone array. The inertial sensor 114 may detect vibration of the user's vocal chords modulated by the user's vocal tract based on vibrations in bones and tissue of the user's head. At Block 2302, a noise suppressor 140 receives the acoustic signals from the microphones included in the mobile device 10 and (ii) the VAD output from the VAD detector 130. At Block 2303, the noise suppressor may suppress the noise included in the acoustic signals received from the microphones (e.g., microphone array 122₁-122_M) included in the mobile device 10 based on the VAD output.

[0081] It is contemplated that when the headset 100 is not being used by the user during a telephone call but rather the user is holding the mobile device 10 to his ear (i.e., at-ear position), the signals from the accelerometer 114 and the microphone array 122₁-122_M as illustrated in **Figure 22** may be used in lieu of signals from the accelerometer 113, and signals from the microphones 111_R, 111_F and microphone array 121₁-121_M. Further, it is contemplated that the second to sixth embodiments, as illustrated in **Figures 5 to 16**, may also be modified such that the signals from the accelerometer 114 and the microphone array 122₁-122_M as illustrated in **Figure 22** may be used in lieu of signals from the accelerometer 113, and signals from the microphones 111_R, 111_F and microphone array 121₁-121_M to generate a VAD output, generate and steer beamformers, and suppress noise, when the mobile device 10 is being used at an at-ear position.

[0082] While the invention has been described in terms of several embodiments, those of ordinary skill in the art will recognize that the invention is not limited to the embodiments

described, but can be practiced with modification and alteration within the spirit and scope of the appended claims. The description is thus to be regarded as illustrative instead of limiting. There are numerous other variations to different aspects of the invention described above, which in the interest of conciseness have not been provided in detail. Accordingly, other embodiments are within the scope of the claims.

CLAIMS

1. A method of detecting a user's voice activity in a mobile device comprising:
generating by a voice activity detector (VAD) a VAD output based on (i) acoustic signals received from microphones included in the mobile device and (ii) data output by an inertial sensor that is included in an earphone portion of the mobile device, the inertial sensor to detect vibration of the user's vocal chords modulated by the user's vocal tract based on vibrations in bones and tissue of the user's head.
2. The method of claim 1, wherein inertial sensor is an accelerometer.
3. The method of claim 2, wherein the accelerometer has a sampling rate greater than 2000Hz.
4. The method of claim 2, wherein the accelerometer has a sampling rate between 2000 Hz and 6000 Hz.
5. The method of claim 2, wherein the microphones included in the mobile device are a microphone array.
6. The method of claim 5, wherein the vibrations in the bones and tissue of the user's head further comprises the vibrations detected from portions of the user's ear and head that are in contact with the earphone portion of the mobile device.
7. The method of claim 6, wherein the mobile device is being used in an at-ear position.
8. The method of claim 6, wherein generating the VAD output comprises:
computing a power envelope of at least one of x, y, z signals generated by the accelerometer; and
setting the VAD output to 1 to indicate that the user's voiced speech is detected if the power envelope is greater than a threshold and setting the VAD output to 0 to indicate that the user's voiced speech is not detected if the power envelope is less than the threshold.
9. The method of claim 6, wherein generating the VAD output comprises:
computing the normalized cross-correlation between any pair of x, y, z direction signals generated by the accelerometer;

setting the VAD output to 1 to indicate that the user's voiced speech is detected if normalized cross-correlation is greater than a threshold within a short delay range, and setting the VAD output to 0 to indicate that the user's voiced speech is not detected if the normalized cross-correlation is less than the threshold.

10. The method of claim 6, wherein generating the VAD output comprises:
 - detecting voiced speech included in the acoustic signals;
 - detecting the vibration of the user's vocal chords from the data output by the accelerometer;
 - computing the coincidence of the detected speech in acoustic signals and the vibration of the user's vocal chords; and
 - setting the VAD output to indicate that the user's voiced speech is detected if the coincidence is detected and setting the VAD output to indicate that the user's voiced speech is not detected if the coincidence is not detected.
11. The method of claim 10, wherein generating the VAD output comprises:
 - detecting unvoiced speech in the acoustic signals by:
 - analyzing at least one of the acoustic signals;
 - if an energy envelope in a high frequency band of the at least one of the acoustic signals is greater than a threshold, a VAD output for unvoiced speech (VADu) is set to indicate that unvoiced speech is detected; and
 - setting the global VAD output to indicate that the user's speech is detected if the voiced speech is detected or if the VADu is set to indicate that unvoiced speech is detected.
12. The method of claim 11, further comprising:
 - receiving the acoustic signals from the microphone array by a fixed beamformer; and
 - steering the fixed beamformer in a direction of the user's mouth when the mobile device is in an at-ear position.
13. The method of claim 12, further comprising:
 - receiving by a noise suppressor (i) a main speech signal from the fixed beamformer and (ii) the VAD output; and
 - suppressing by the noise suppressor noise included in the main speech signal based on the VAD output.

14. The method of claim 11, further comprising:
receiving the acoustic signals from the microphone array by a source direction detector;
detecting by the source direction detector the user's speech source based on the VAD output;
adaptively steering a first beamformer in a direction of the detected user's speech source when the VAD output is set to indicate that the user's speech is detected, the first beamformer outputting a main speech signal.
15. The method of claim 14, wherein detecting by the source direction detector the user's speech source based on the VAD output comprises:
determining a delay for a sound signal between microphones in the microphone array;
and
detecting the main acoustic source location using generalized cross correlation (GCC) or adaptive eigenvalue decomposition (AED).
16. The method of claim 14, detecting by the source direction detector the user's speech source based on the VAD output comprises:
steering the first beamformer over a range of directions; and
calculating a power of the first beamformer for each direction in the range of directions, wherein the user's speech source is detected as a direction in the range of directions having the highest power.
17. The method of claim 14, further comprising:
adaptively steering a second beamformer with a null towards the user's speech source, wherein the second beamformer has a cardioid pattern, wherein the second beamformer outputs a signal representing environmental noise when the VAD output is set to indicate that the user's speech is not detected;
receiving by a noise suppressor (i) a main speech signal from the first beamformer, (ii) the signal representing the environmental noise from the second beamformer, and (iii) the VAD output; and
suppressing by the noise suppressor noise included in the main speech signal based on the signal representing the environmental noise and the VAD output.
18. The method of claim 14, further comprising:

adaptively steering a second beamformer in a direction of strongest environmental noise location when the VAD output is set to indicate that the user's speech is not detected, wherein the second beamformer outputs a signal representing the strongest environmental noise;

receiving by a noise suppressor (i) a main speech signal from the first beamformer, (ii) the signal representing the strongest environmental noise outputted from the second beamformer, and (iii) the VAD output; and

suppressing by the noise suppressor noise included in the main speech signal based on the signal representing the strongest environmental noise and the VAD output.

19. The method of claim 14, further comprising:

detecting by a second beamformer a direction of strongest environmental noise location when the VAD output is set to indicate that the user's speech is not detected;

adaptively steering the nulls of the first beamformer in the direction of the strongest environmental noise location to output a main speech signal from the first beamformer;

receiving by a noise suppressor (i) the main speech signal being output from the first beamformer, and (ii) the VAD output; and

suppressing by the noise suppressor noise included in the main speech signal based on the VAD output.

20. A mobile device detecting a user's voice activity comprising:

an accelerometer to detect vibration of the user's vocal chords modulated by the user's vocal tract based on vibrations in bones and tissue of the user's head, wherein the accelerometer is included in an earphone portion of the mobile device;

a voice activity detector (VAD) coupled to the accelerometer, the VAD to generate a VAD output based on (i) acoustic signals received from microphones included in the mobile device and (ii) data output by the accelerometer; and

a noise suppressor coupled to the microphones and the VAD, the noise suppressor to suppress noise from the acoustic signals from the microphones based on the VAD output.

21. The mobile device of claim 20, wherein accelerometer has a sampling rate greater than 2000Hz.

22. The mobile device of claim 20, wherein the accelerometer has a sampling rate between 2000 Hz and 6000 Hz.

23. The mobile device of claim 20, wherein the microphones included in the mobile device are a microphone array.

24. The mobile device of claim 23, wherein the vibrations in the bones and tissue of the user's head further comprises the vibrations detected from portions of the user's ear and head that are in contact with the earphone portion of the mobile device.

25. The mobile device of claim 24, wherein the mobile device is being used in an at-ear position.

26. The mobile device of claim 24, wherein the VAD generates the VAD output by:
computing a power envelope of at least one of x, y, z signals generated by the accelerometer; and

setting the VAD output to indicate that the user's voiced speech is detected if the power envelope is greater than a threshold and setting the VAD output to indicate that the user's voiced speech is not detected if the power envelope is less than the threshold.

27. The mobile device of claim 24, wherein the VAD generates the VAD output by:
computing the normalized cross-correlation between any pair of x, y, z direction signals generated by the accelerometer; and

setting the VAD output to indicate that the user's voiced speech is detected if normalized cross-correlation is greater than a threshold within a short delay range, and setting the VAD output to indicate that the user's voiced speech is not detected if the normalized cross-correlation is less than the threshold.

28. The mobile device of claim 24, wherein the VAD generates the VAD output by:
detecting speech included in the acoustic signals;

detecting the vibrations of the user's vocal chords from the data output by the accelerometer;

computing the coincidence of the detected speech in acoustic signals and the vibrations of the user's vocal chords; and

setting the VAD output to indicate that the user's voiced speech is detected if the coincidence is detected and setting the VAD output to indicate that the user's voiced speech is not detected if the coincidence is not detected.

29. The mobile device of claim 28, wherein generating the VAD output comprises:
detecting unvoiced speech in the acoustic signals by:
analyzing at least one of the acoustic signals;
if an energy envelope in a high frequency band of the at least one of the acoustic signals is greater than a threshold, a VAD output for unvoiced speech (VADu) is set to indicate that unvoiced speech is detected; and
setting the VAD output to indicate that the user's speech is detected if the voiced speech is detected or if the VADu is set to indicate that unvoiced speech is detected.
30. The mobile device of claim 27, further comprising:
a fixed beamformer receiving the acoustic signals from the microphone array, wherein the fixed beamformer is steered in a direction of the user's mouth when the mobile device is in an at-ear position to output a main speech signal.
31. The mobile device of claim 30, wherein the noise suppressor suppresses the noise included in the main speech signal outputted by the fixed beamformer based on the VAD output.
32. The mobile device of claim 27, further comprising:
a source direction detector receiving the acoustic signals from the microphone array and detecting the user's speech source based on the VAD output; and
a first beamformer being adaptively steered in a direction of the detected user's speech source when the VAD output is set to indicate that the user's voiced speech is detected, wherein the first beamformer outputs a main speech signal.
33. The mobile device of claim 32, wherein the source direction detector detects the user's speech source based on the VAD output by:
determining a delay for a sound signal between microphones in the microphone array;
and
detecting the main acoustic source location using generalized cross correlation (GCC) or adaptive eigenvalue decomposition (AED).
34. The mobile device of claim 32, wherein the source direction detector detects the user's speech source based on the VAD output by:
steering the first beamformer over a range of directions; and

calculating a power of the first beamformer for each direction in the range of directions, wherein the user's speech source is detected as a direction in the range of directions having the highest power.

35. The mobile device of claim 32, further comprising:

a second beamformer being adaptively steered to direct a null of the second beamformer towards the user's speech source, wherein the second beamformer has a cardioid pattern, wherein the second beamformer outputs a signal representing environmental noise when the VAD output is set to indicate that the user's voiced speech is not detected,

wherein the noise suppressor suppresses the noise included in the main speech signal based the signal representing environmental noise outputted from the second beamformer and the VAD output.

36. The mobile device of claim 32, further comprising:

a second beamformer being adaptively steered in a direction of strongest environmental noise location when the VAD output is set to indicate that the user's speech is not detected, wherein the second beamformer outputs a signal representing the strongest environmental noise,

wherein the noise suppressor suppresses the noise included in the main speech signal based on the signal representing the strongest environmental noise outputted from the second beamformer and the VAD output.

37. The mobile device of claim 32, further comprising:

a second beamformer detecting a direction of strongest environmental noise location when the VAD output is set to indicate that the user's speech is not detected, wherein the nulls of the first beamformer are adaptively steered in the direction of the strongest environmental noise location.

1/23

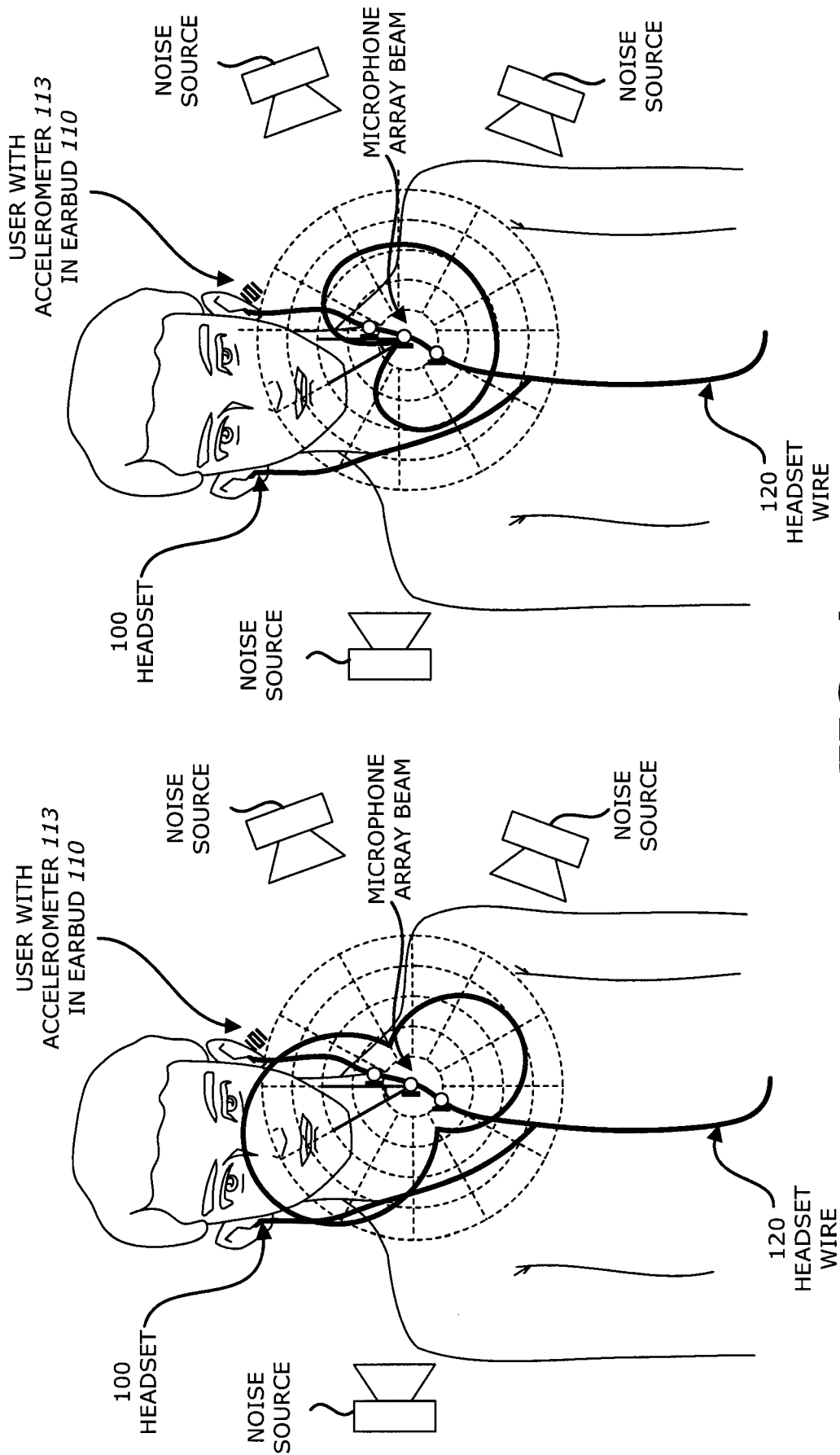


FIG. 1

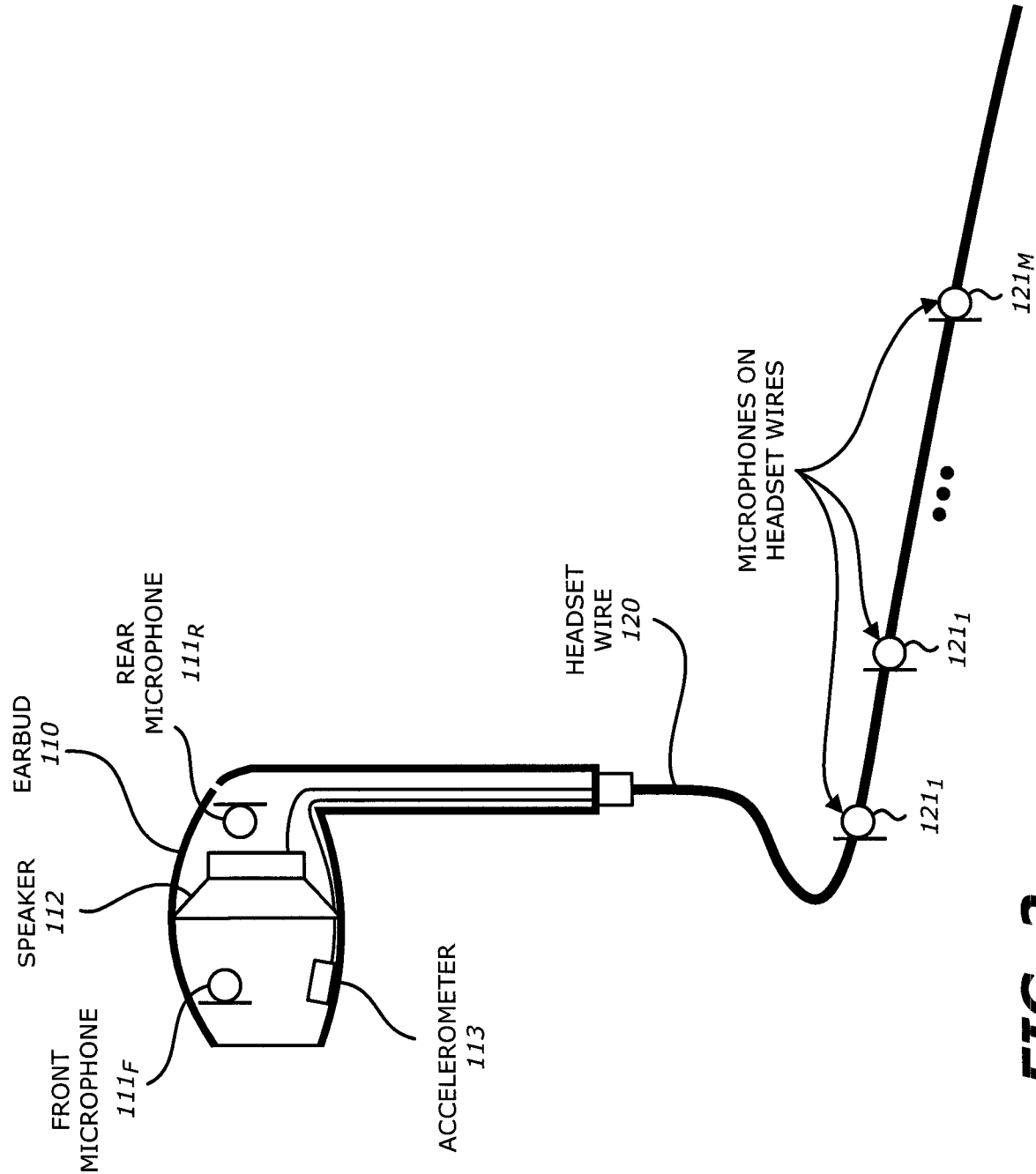
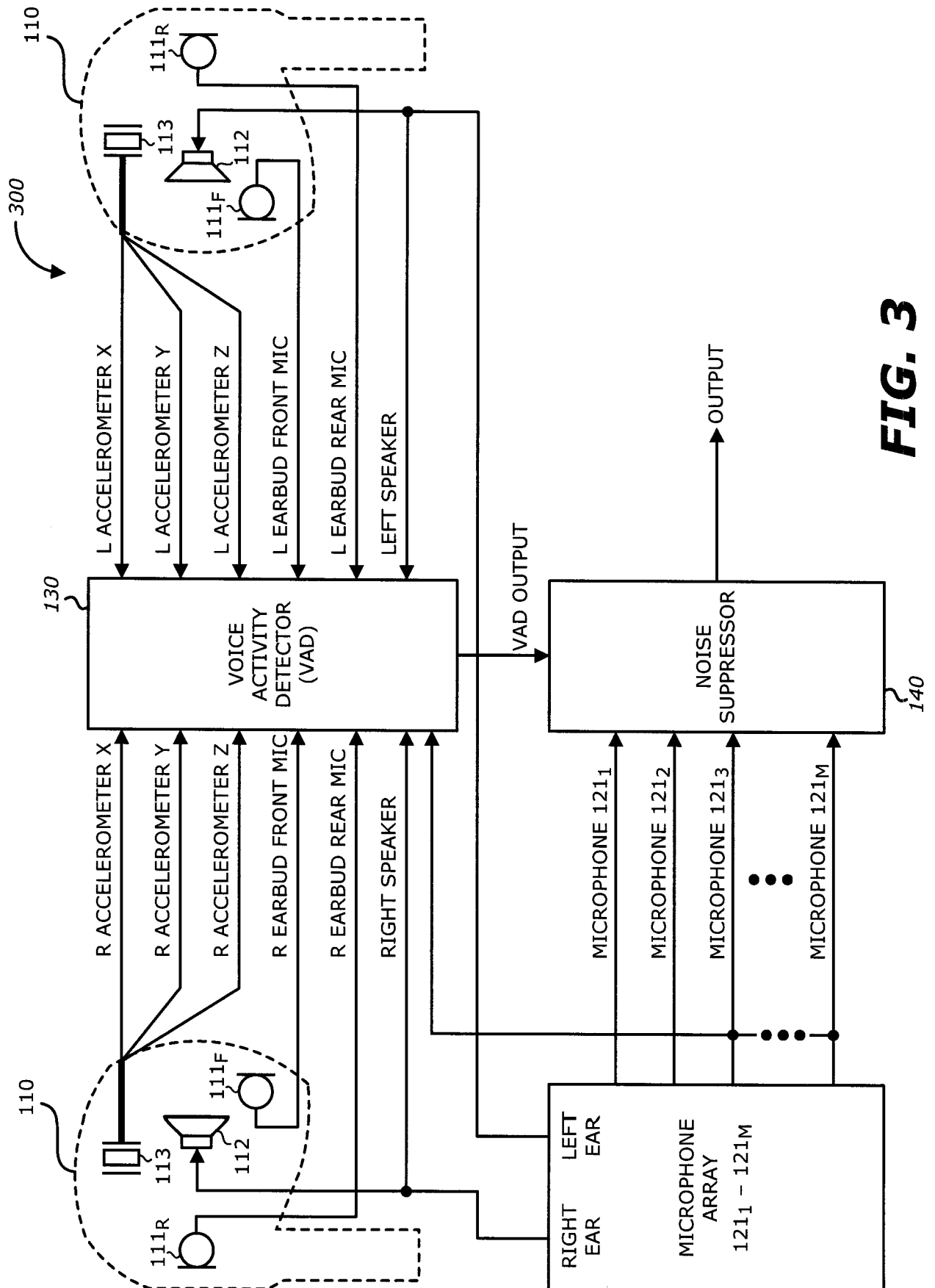
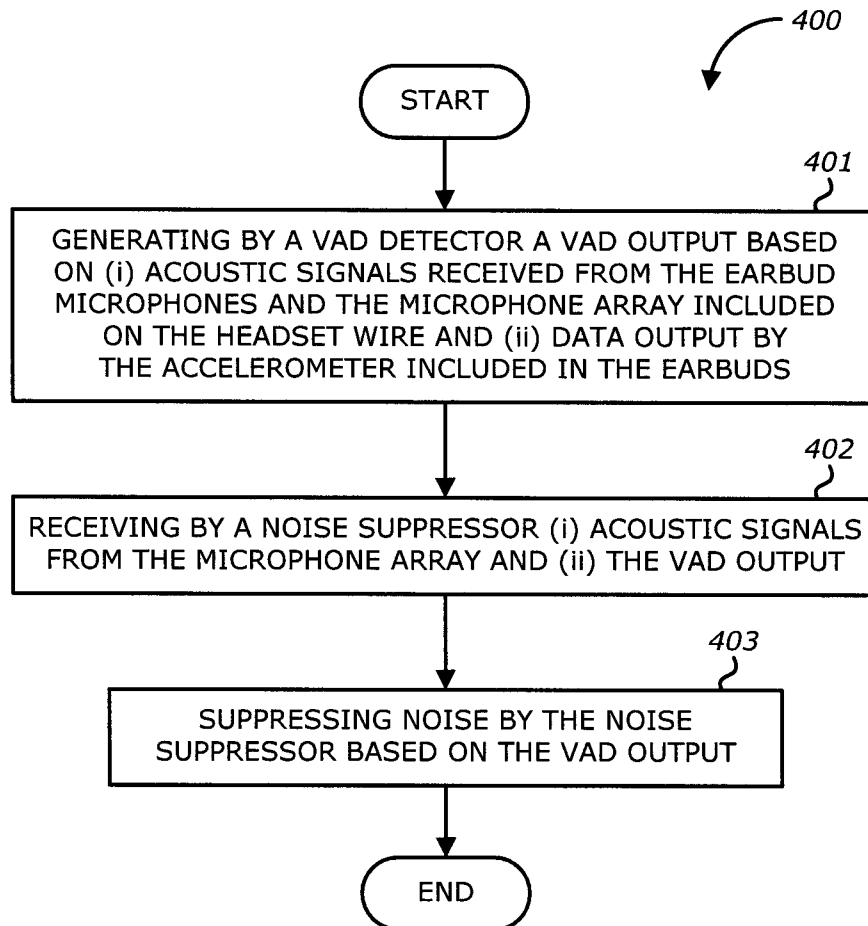


FIG. 2

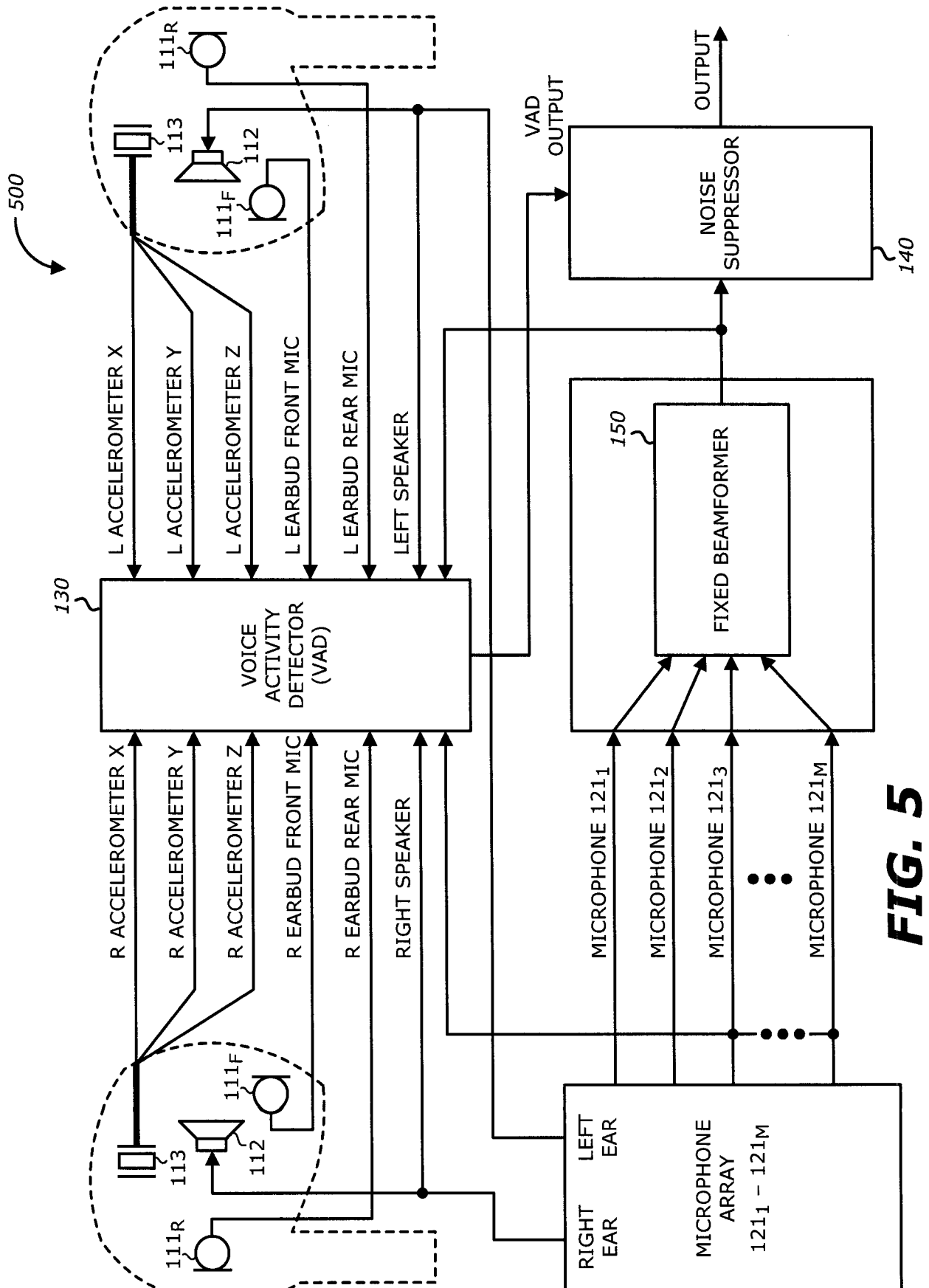
3/23



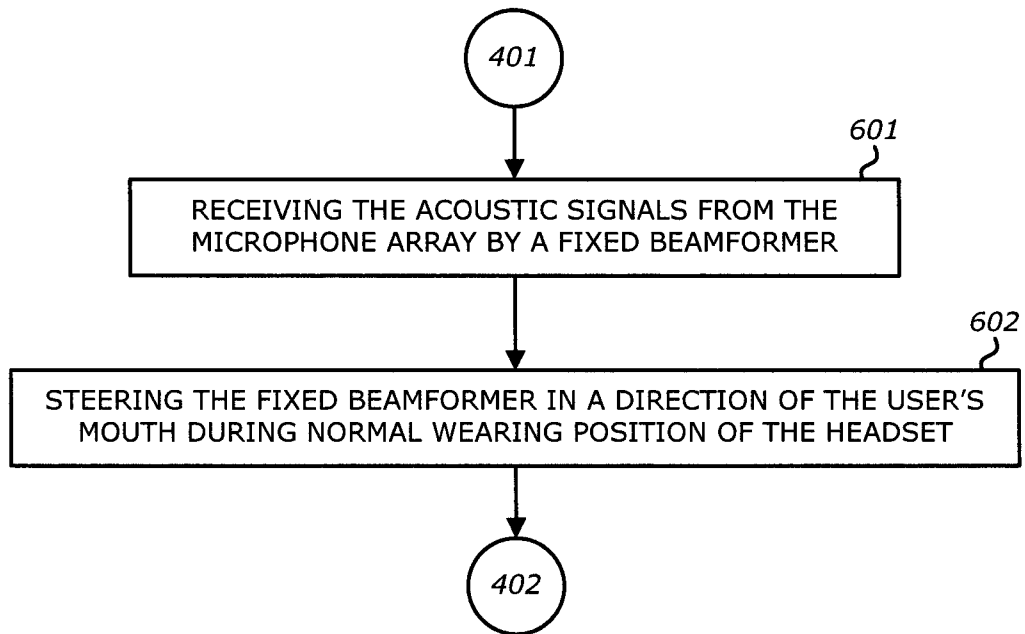
4/23

**FIG. 4**

5/23



6/23

**FIG. 6**

7/23

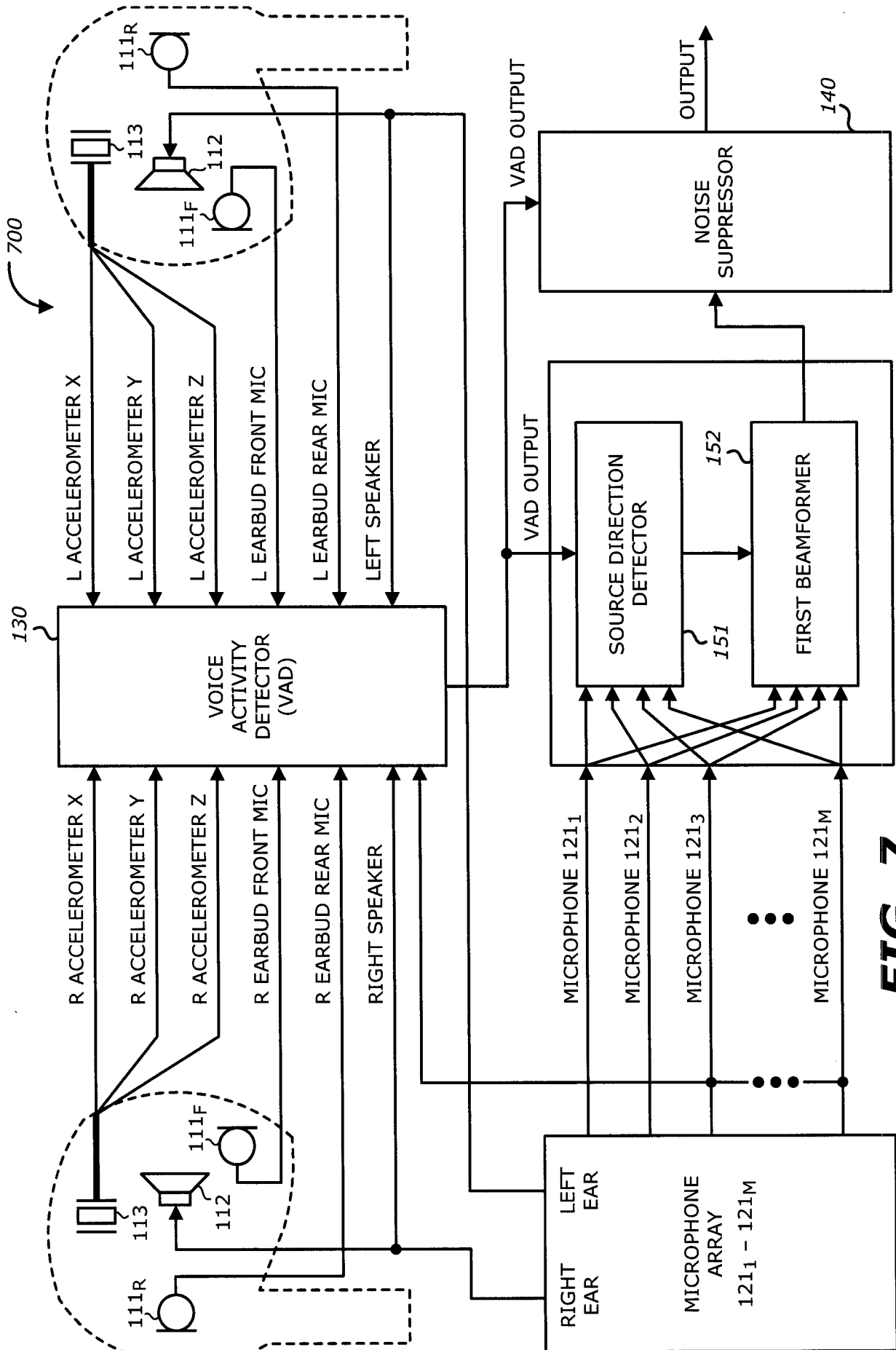
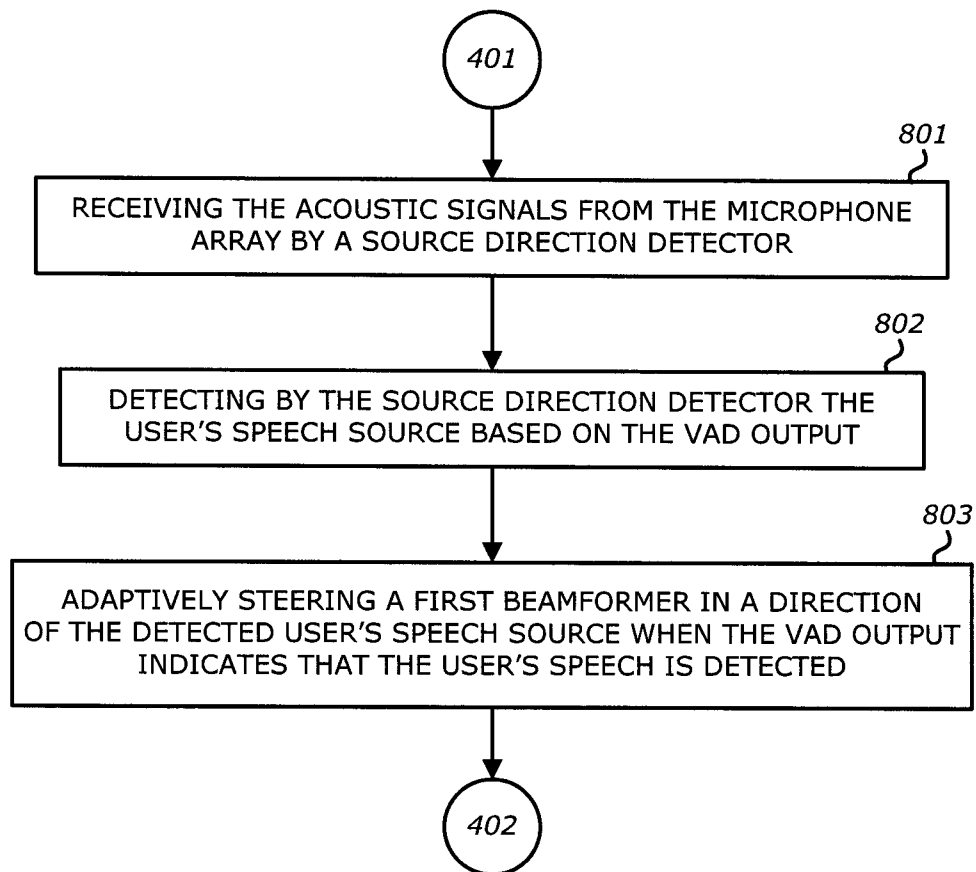


FIG. 7

8/23

**FIG. 8**

9/23

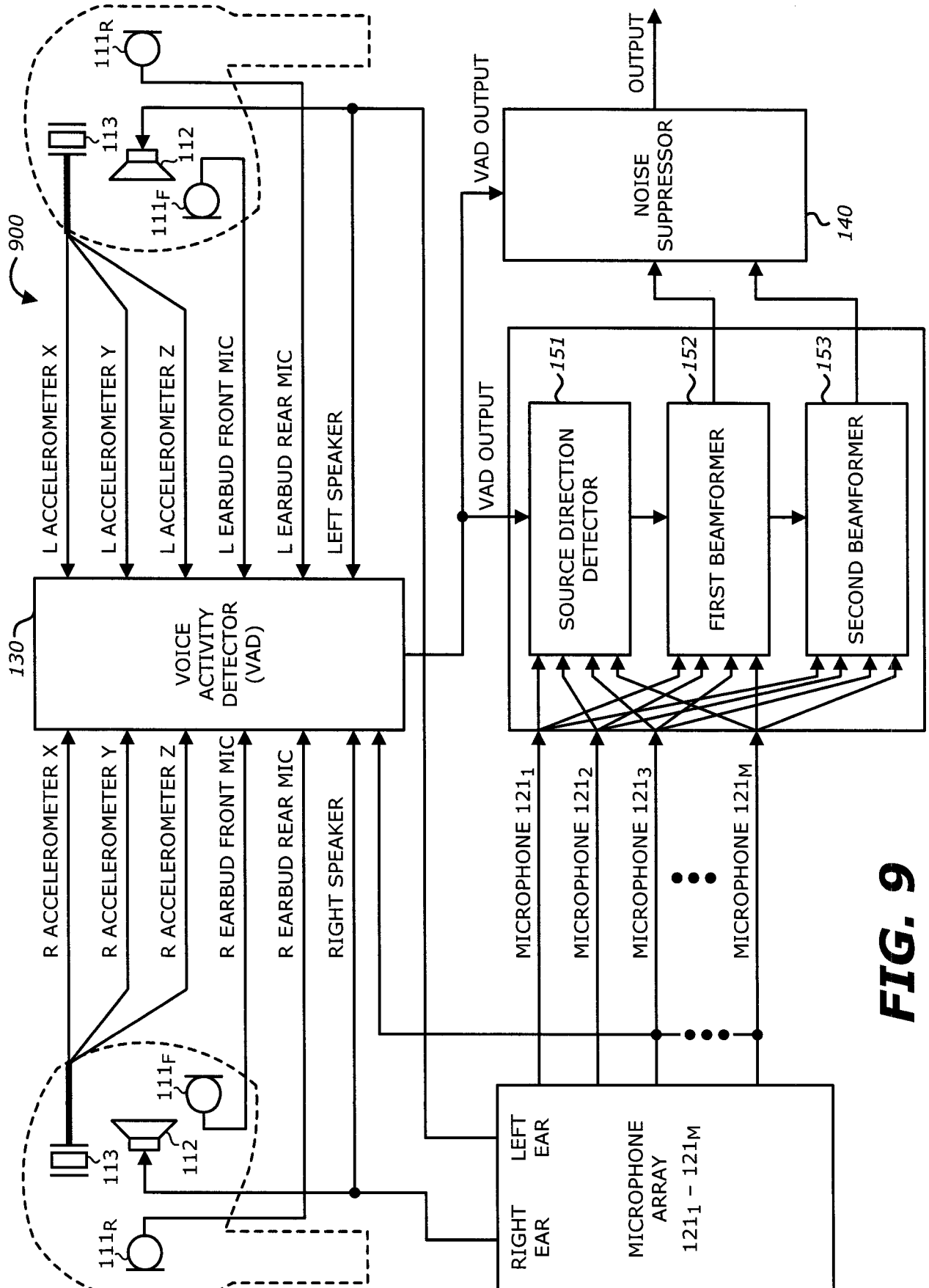
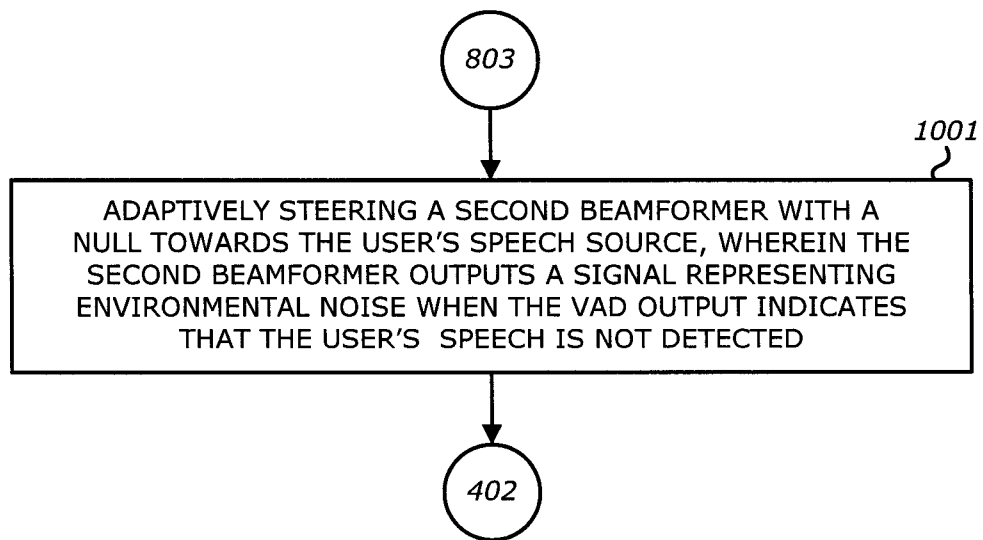


FIG. 9

10/23

**FIG. 10**

11/23

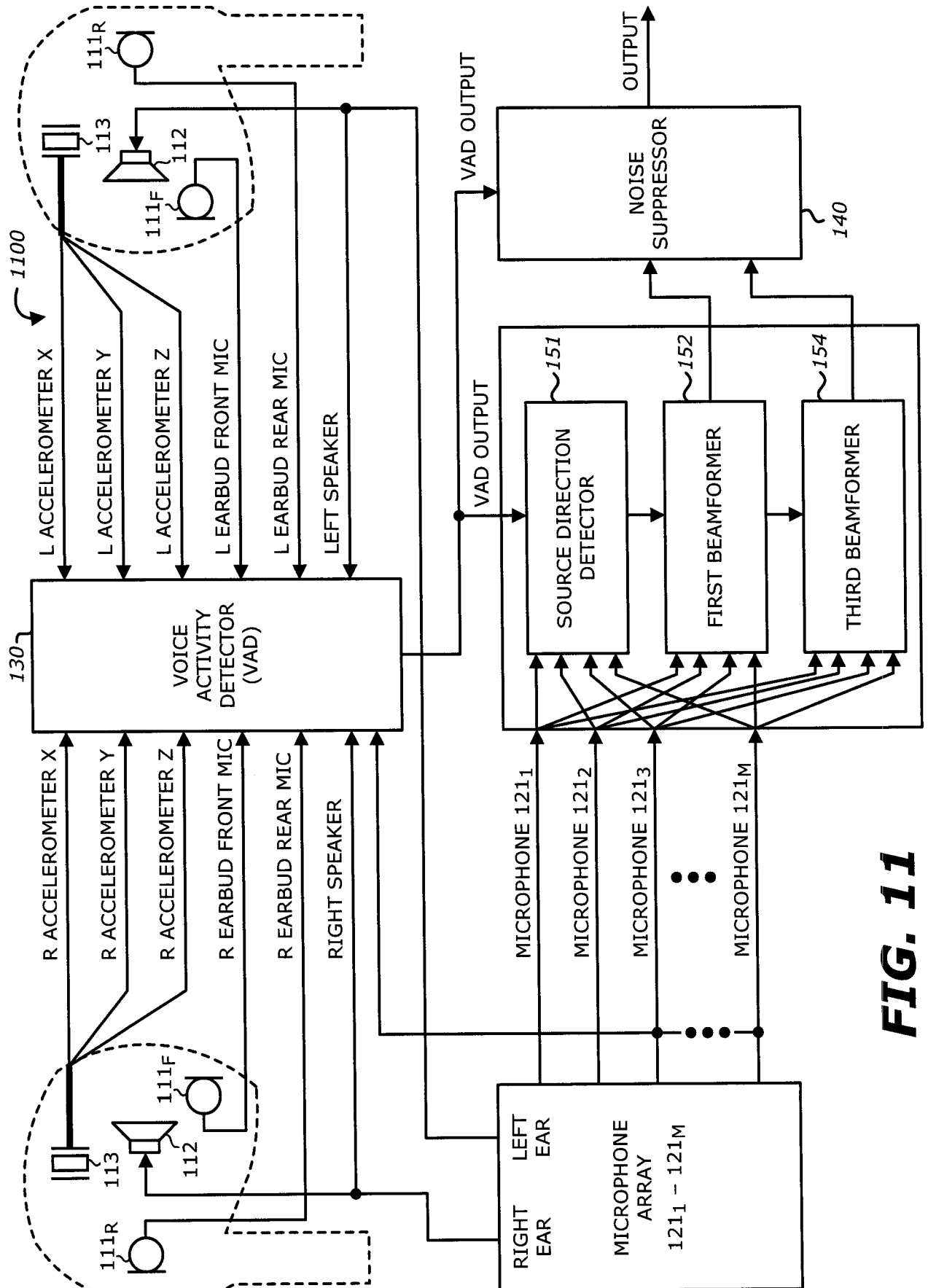
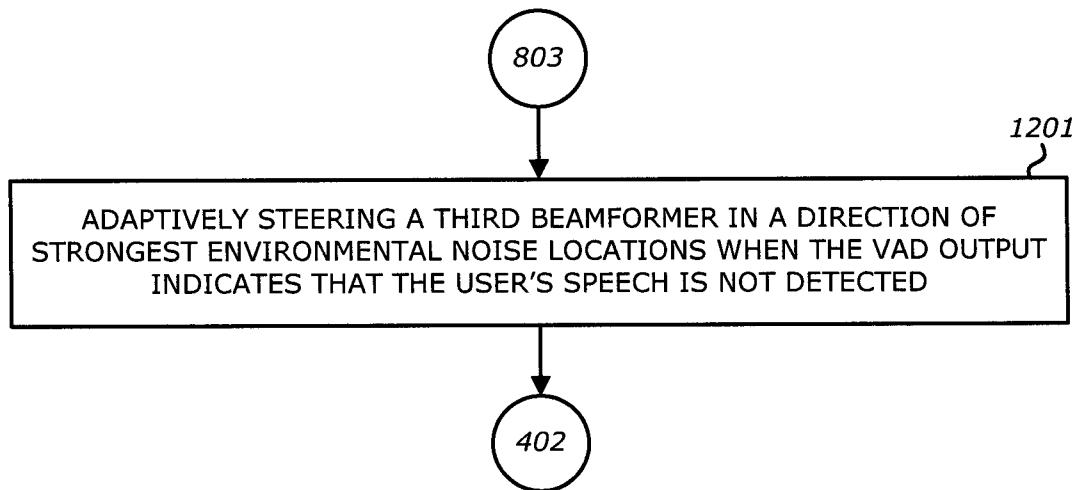


FIG. 11

12/23

**FIG. 12**

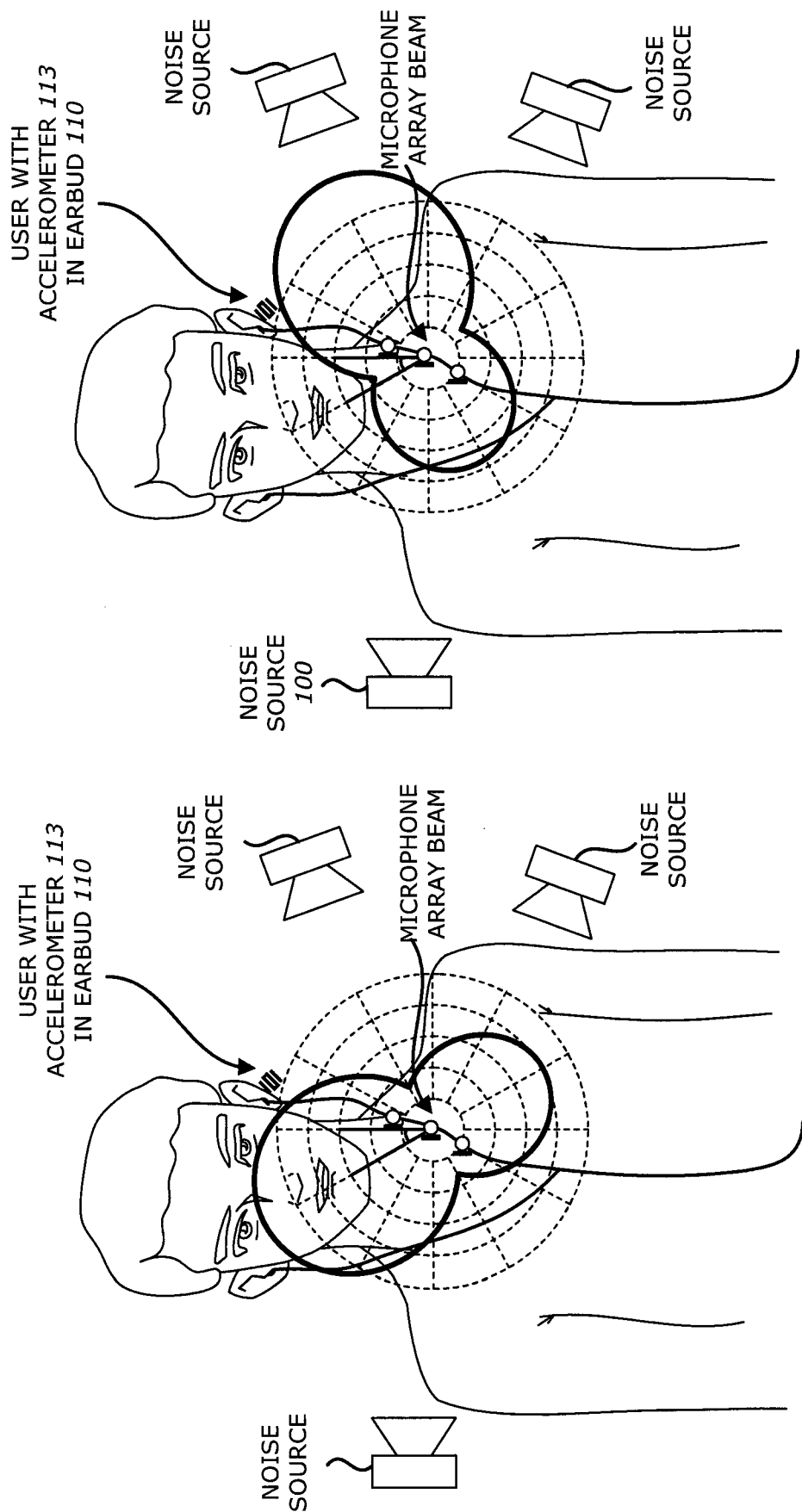
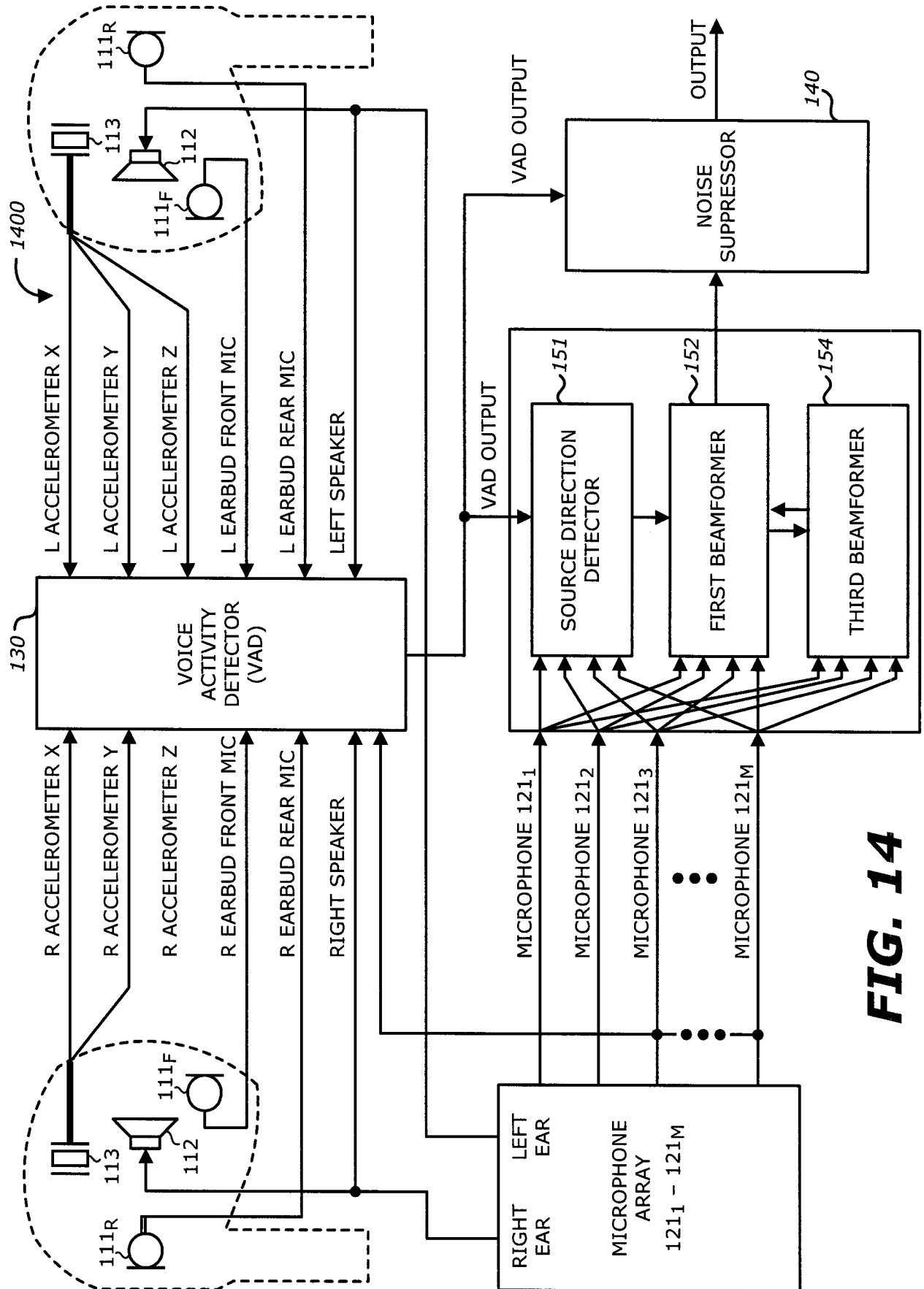
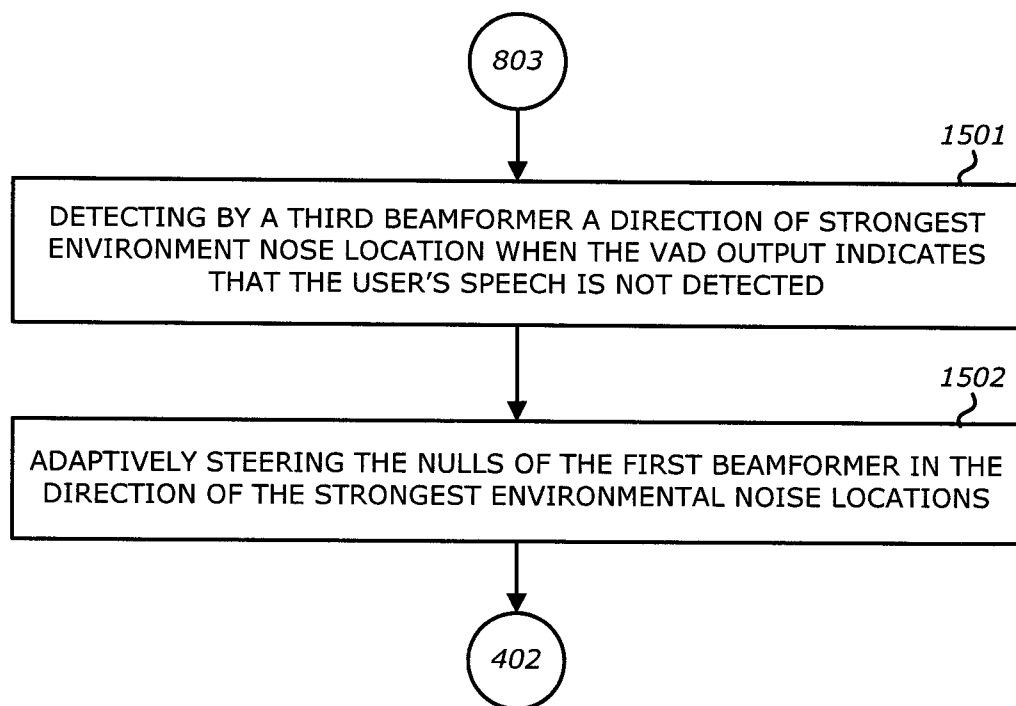


FIG. 13

14/23



15/23

**FIG. 15**

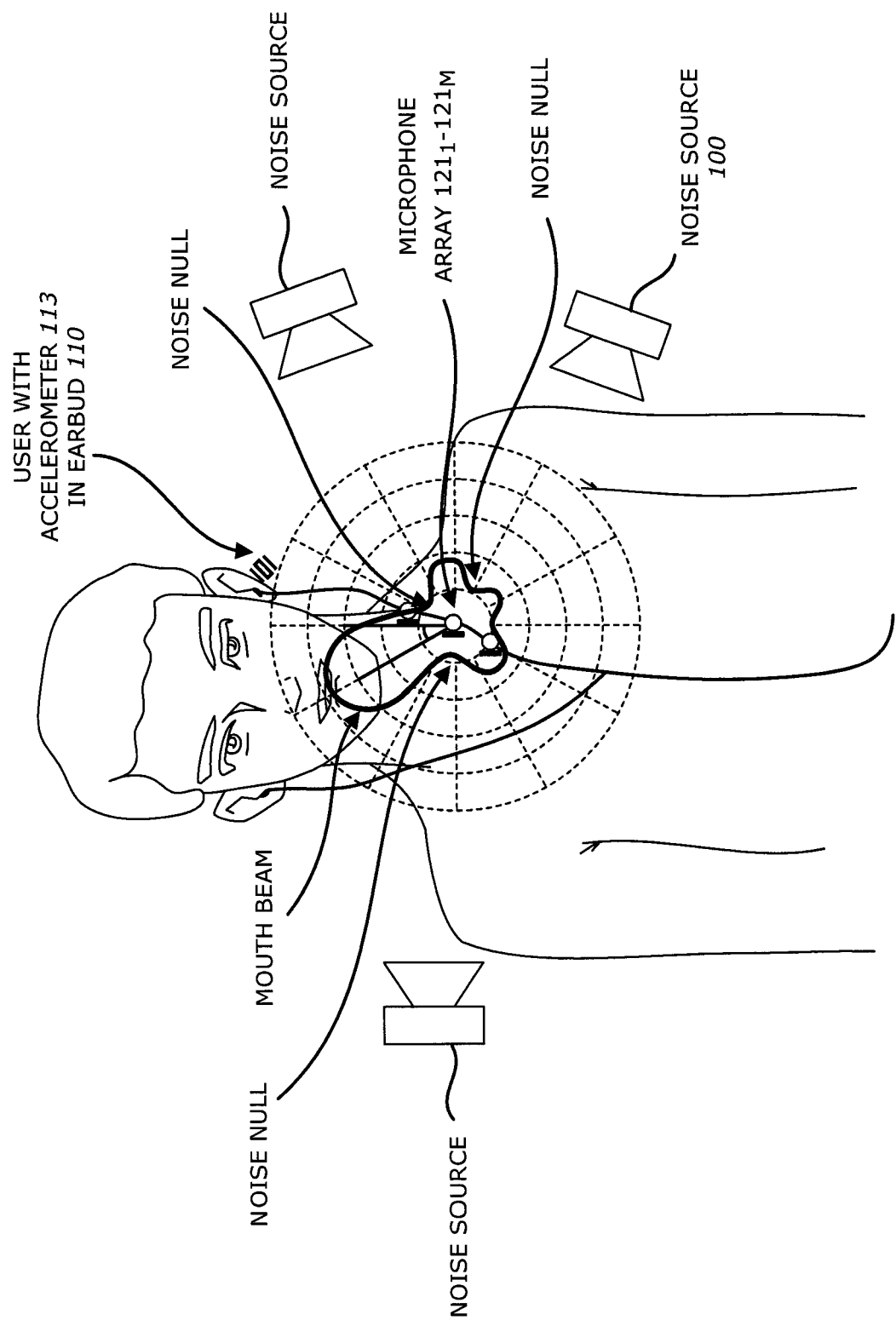


FIG. 16

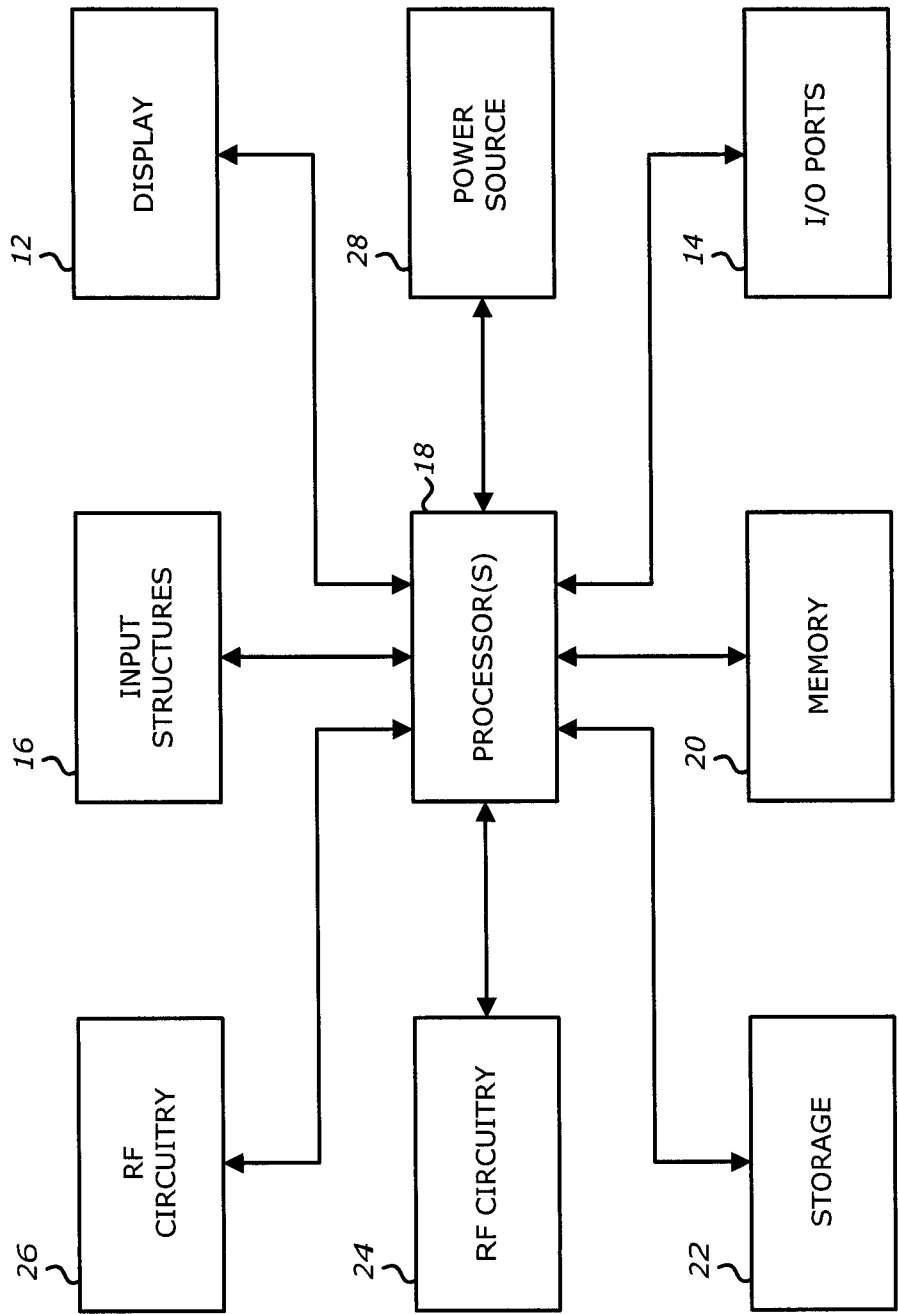


FIG. 17

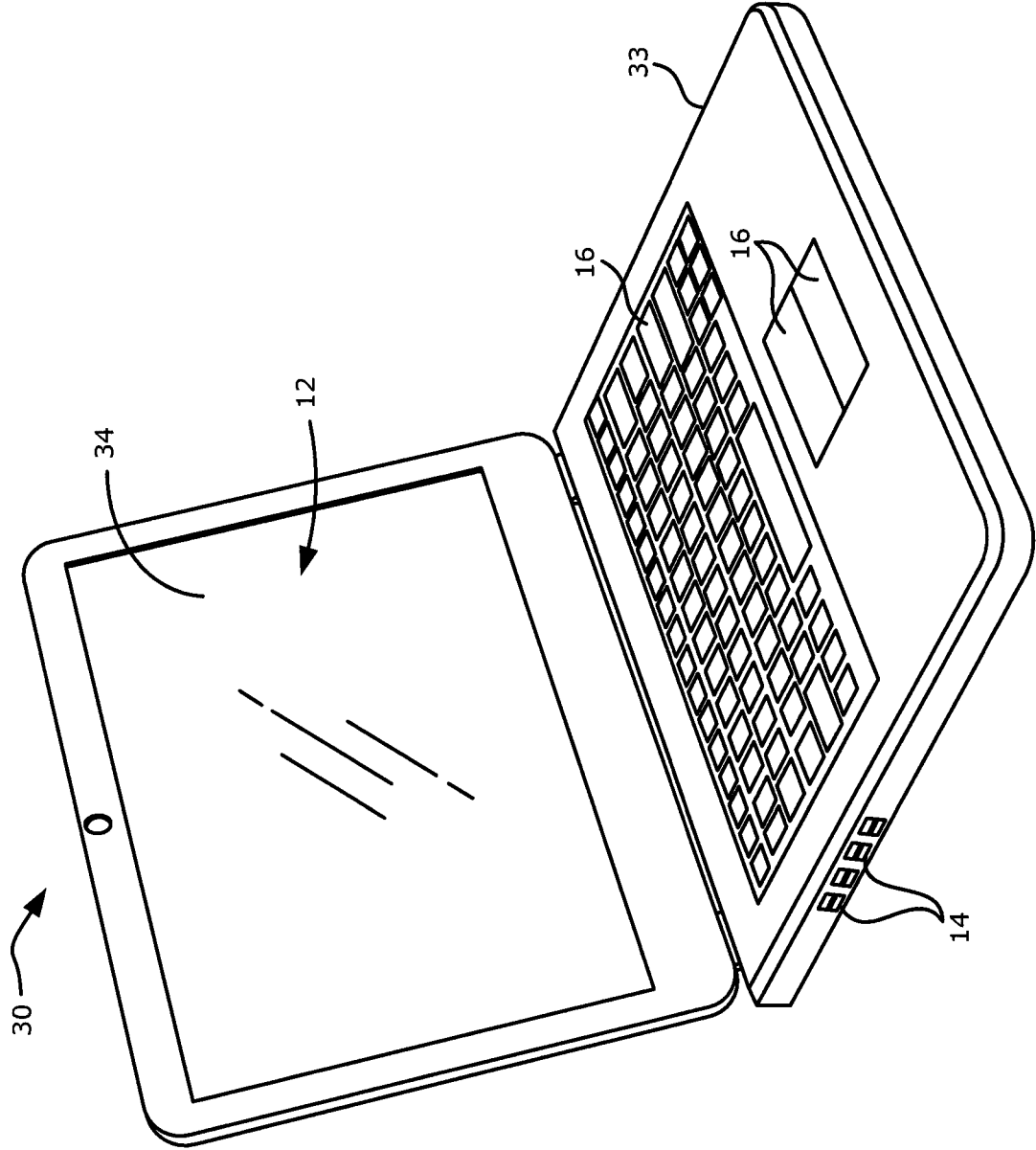
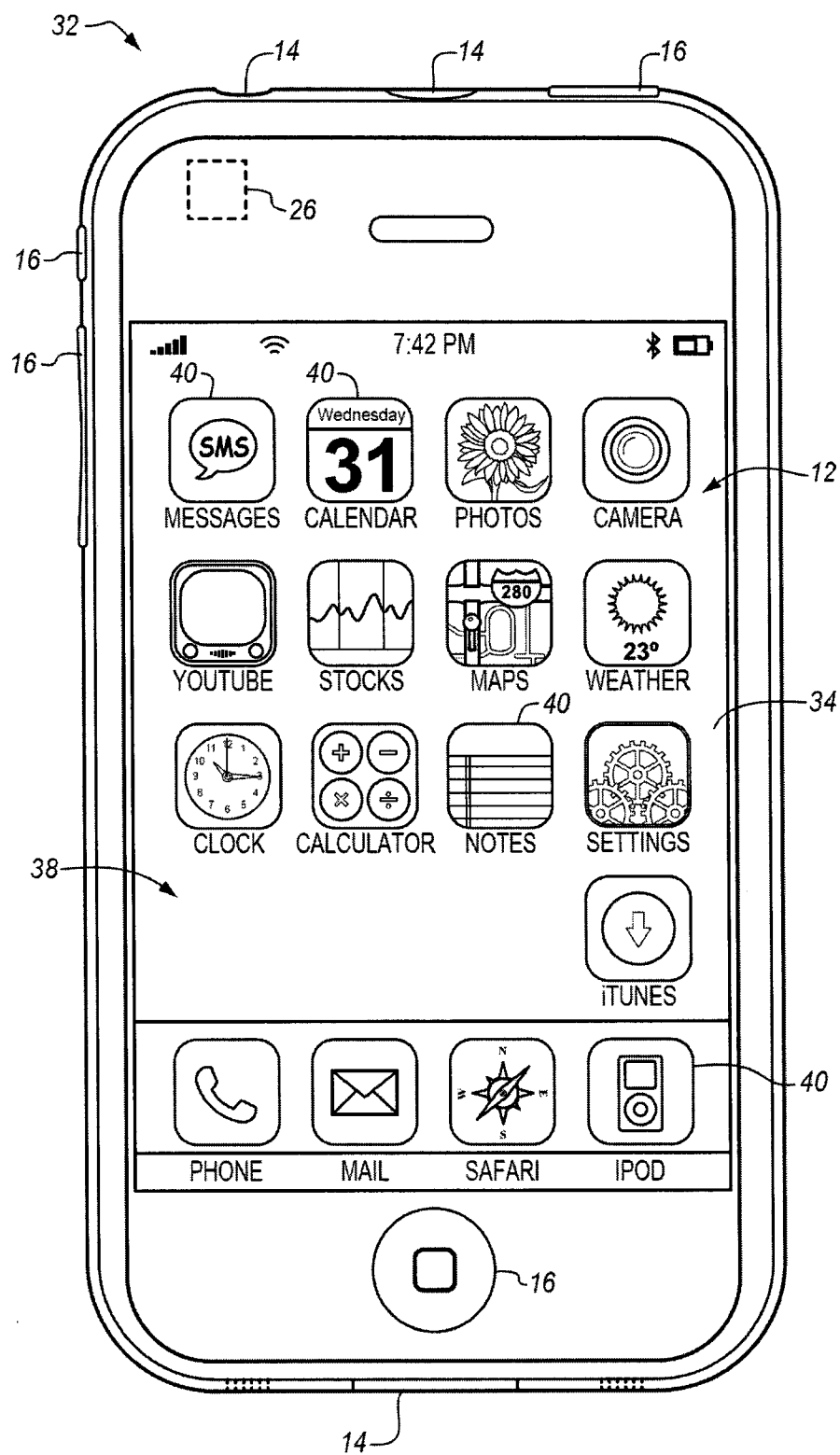


FIG. 18

19/23

**FIG. 19**

20/23

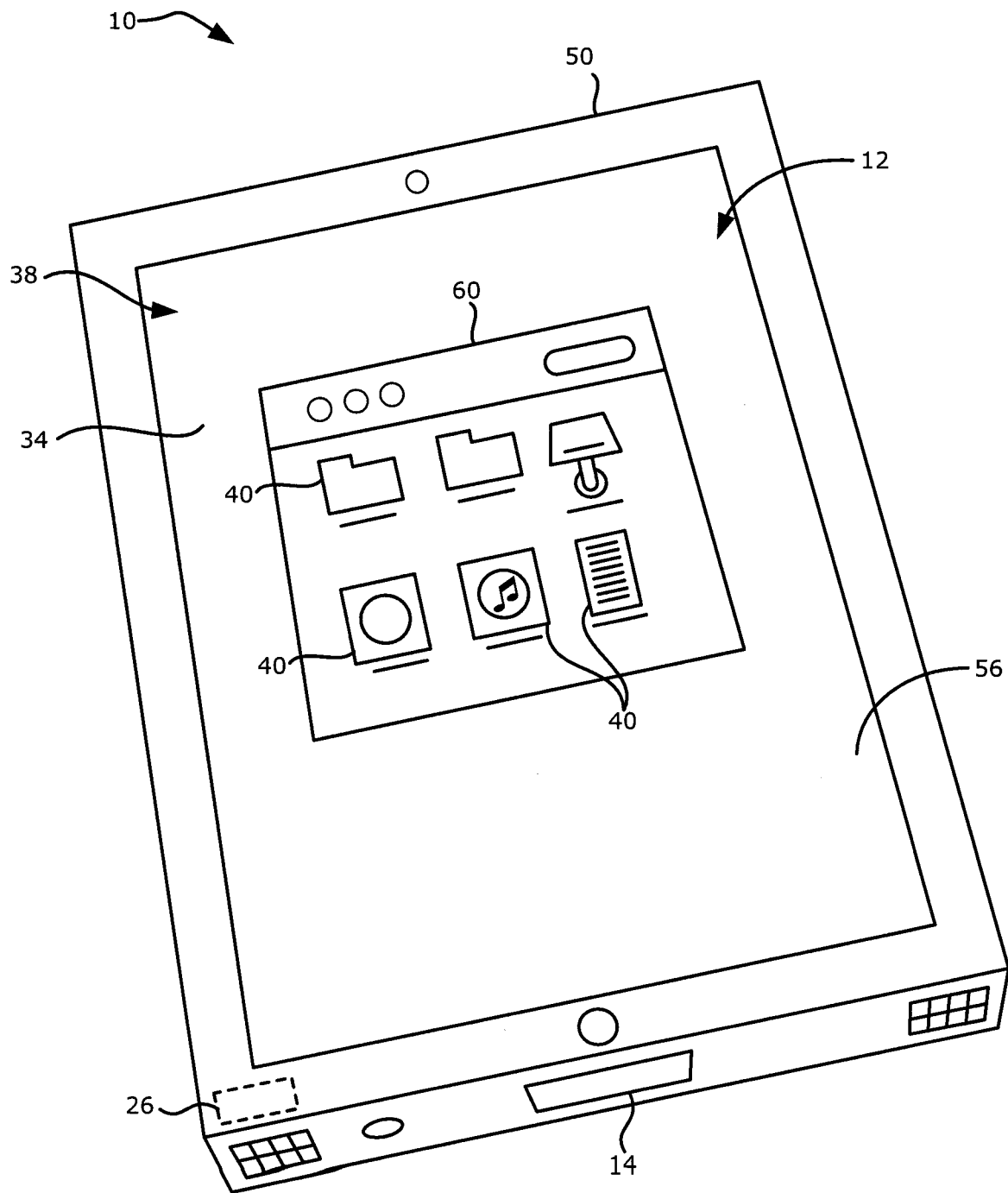
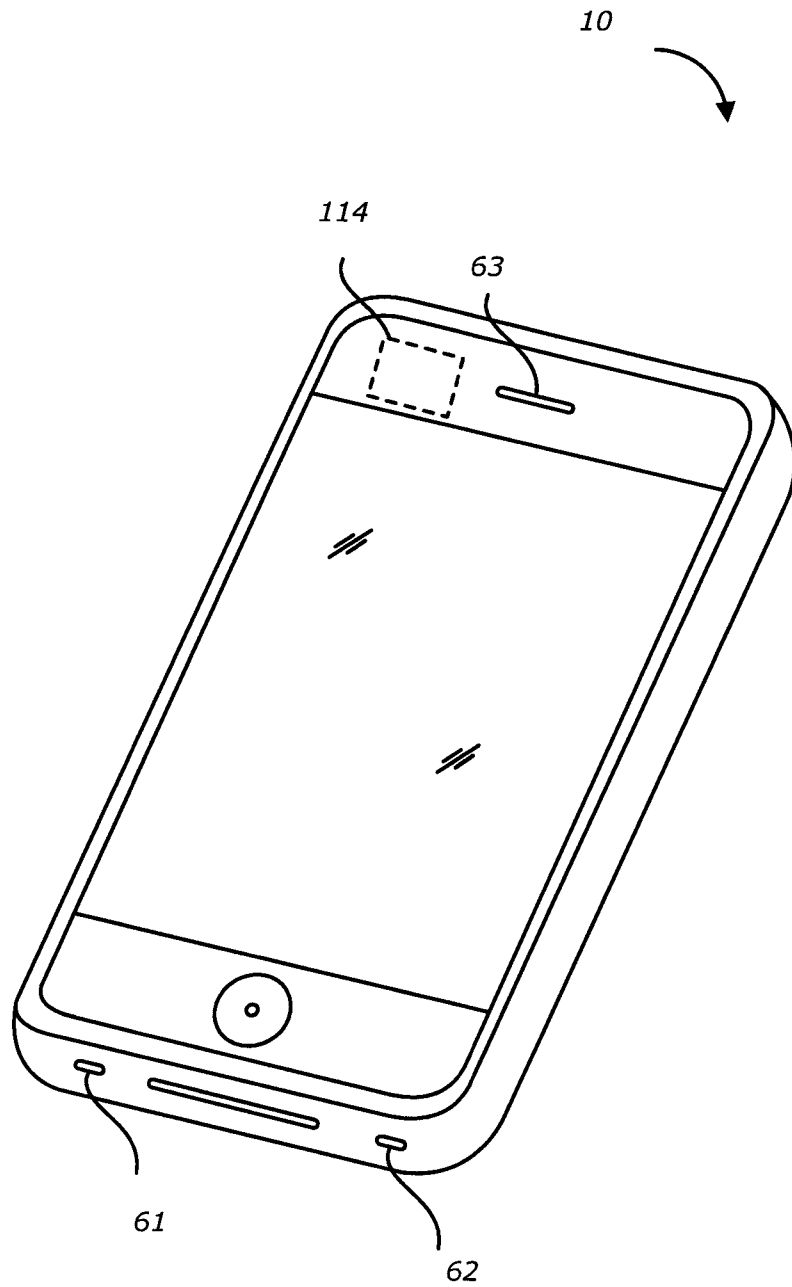
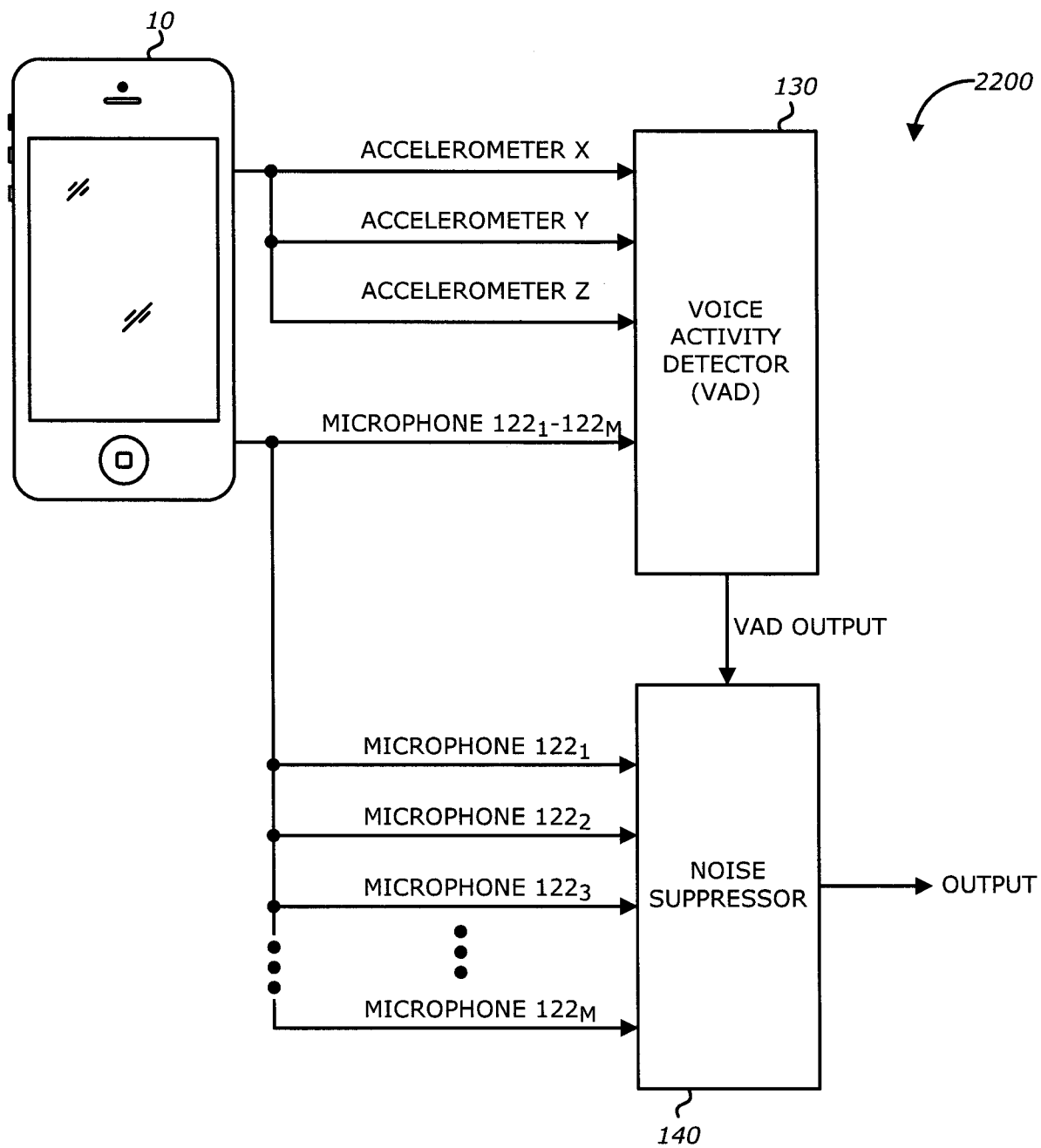


FIG. 20

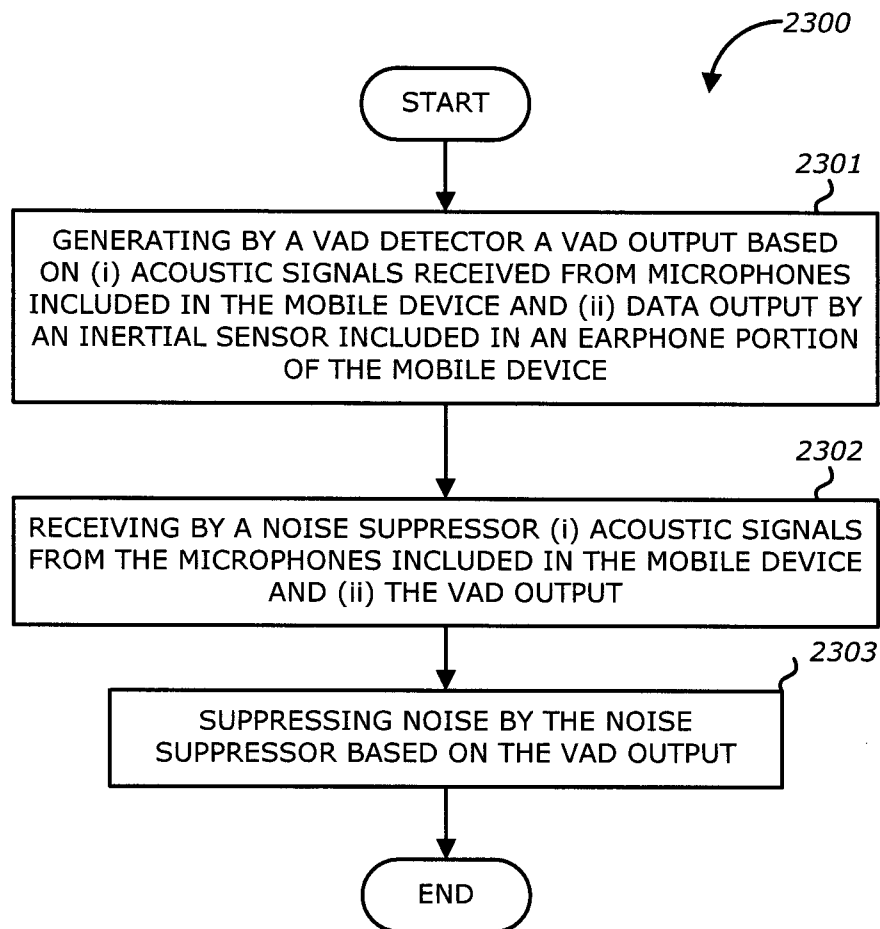
21/23

**FIG. 21**

22/23

**FIG. 22**

23/23

**FIG. 23**

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2013/058551

A. CLASSIFICATION OF SUBJECT MATTER

INV. G10L25/78 G10L21/0216
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G10L

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X A	US 2011/208520 A1 (LEE TE-WON [US]) 25 August 2011 (2011-08-25) abstract page 2, paragraphs [0026], [0030] - [0032] page 4, paragraph [0051] page 5, paragraph [0068] page 6, paragraph [0077] figure 1	1-13, 20-31 14-19, 32-37
A	US 2009/238377 A1 (RAMAKRISHNAN DINESH [US] ET AL) 24 September 2009 (2009-09-24) abstract page 3, paragraph [0041] - [0042] page 4, paragraphs [0056], [0065] page 5, paragraph [0075] - [0077] figures 3,4,8 ----- -/-	1,14-20, 32-37



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

15 November 2013

Date of mailing of the international search report

25/11/2013

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Greiser, Norbert

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2013/058551

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	EP 1 489 596 A1 (SONY ERICSSON MOBILE COMM AB [SE]) 22 December 2004 (2004-12-22) abstract page 2, paragraph [0004]-[0006] page 3, paragraph [0037] - page 4, paragraph [0045] figure 2 -----	14-19, 32-37
A	SHAHIDUR RAHMAN M ET AL: "Low-frequency band noise suppression using bone conducted speech", COMMUNICATIONS, COMPUTERS AND SIGNAL PROCESSING (PACRIM), 2011 IEEE PACIFIC RIM CONFERENCE ON, IEEE, 23 August 2011 (2011-08-23), pages 520-525, XP031971208, DOI: 10.1109/PACRIM.2011.6032948 ISBN: 978-1-4577-0252-5 abstract page 522, left-hand column, paragraph 2 - right-hand column, paragraph 1 figures 5,6 -----	1,20

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2013/058551

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
US 2011208520	A1	25-08-2011	CN	102770909 A	07-11-2012
			EP	2539887 A1	02-01-2013
			JP	2013520707 A	06-06-2013
			KR	20120125986 A	19-11-2012
			US	2011208520 A1	25-08-2011
			WO	2011106065 A1	01-09-2011

US 2009238377	A1	24-09-2009	CA	2705789 A1	24-09-2009
			CN	101911724 A	08-12-2010
			EP	2277323 A1	26-01-2011
			JP	5313268 B2	09-10-2013
			JP	2011515897 A	19-05-2011
			KR	20100116693 A	01-11-2010
			RU	2010142270 A	27-04-2012
			TW	200951942 A	16-12-2009
			US	2009238377 A1	24-09-2009
			WO	2009117471 A1	24-09-2009

EP 1489596	A1	22-12-2004	AT	339757 T	15-10-2006
			CN	1813284 A	02-08-2006
			DE	60308342 T2	06-09-2007
			EP	1489596 A1	22-12-2004
			US	2008091421 A1	17-04-2008
			WO	2004111995 A1	23-12-2004
