

PCT Patent Application

User Voice Activity Detection Methods, Devices, Assemblies, and Components

INVENTORS

Jiajin An

MICHAEL JON WURTZ

DAVID WURTZ

MANPREET KHAIRA

Amit Kumar

Shawn O'Connor

Shankar Rathoud

James Scanlan

Eric Sorenson

Docket: 200.002PCT

User Voice Activity Detection Methods, Devices, Assemblies, and Components

Copyright Notice and Permission

A portion of this patent document contains material subject to copyright protection. The copyright owner has no objection to the facsimile reproduction by anyone of the patent document or the patent disclosure, as it appears in the patent and trademark office patent files or records, but otherwise reserves all copyrights whatsoever. The following notice applies to this document: Copyright © 2017, AVNERA CORPORATION.

Related Applications

The present application claims priority to U.S. Patent Application 15/711,793 filed September 21, 2017; and to U.S. Provisional Patent Applications 62/456,100 filed February 7, 2017; 62/459,055 filed February 14, 2017; and 62/532,964 filed July 14, 2017. All four applications are incorporated herein by reference in their entirety.

Technical Field

Various embodiments of the invention relate generally to automatic detection of user voice activity in headsets of various types, such as those with automatic noise reduction.

Background

Headsets typically include a microphone along with one or two ear devices or earplugs which are worn over, on, or in the ears of users to facilitate electronic communications. Many contemporary headsets also include automatic noise reduction (ANR) or automatic noise cancellation (ANC) circuitry for automatically sensing and canceling significant portions of ambient noise, thereby improving user listening experience. ANC circuitry is simple in principle, but complex in implementation, with many headsets including as many six microphones (mics): two for feedback (fb) ANC, two for feedforward (ff) ANC, and one or two for picking up user voice signals.

One problem the present inventors have recognized is that while ANC headsets improve the listening experiences of users wearing them, they do little or nothing to improve the quality of the signals being transmitted out of the headset to

other devices. For example, in many noisy environments, users wearing headsets with their mobile phones are typically required to manually switch their phones to a mute mode to prevent background noise from being transmitted by the phone to other phones and devices. Although this is effective in preventing the unwanted noise transmission, it often leads to gaps in communications as users who muted their phones forget to unmute them when they start speaking again. Moreover, in conference calls with many users manually muting and unmuting, the potential for communications gaps is even greater.

In attempting to solve the problem, some designers have provided circuitry that monitors the voice mic output and automatically mutes or unmutes based on comparison of the mic output level to one or more thresholds. However, the monitoring circuitry suffers from at least two problems that has inhibited its use in mobile phones. First, it consumes considerable power and thus shortens battery life. Second, the circuitry creates its own communications gaps not only because its slow reaction to speaking stops and starts, but also because it confuses external noises, for example voices of nearby people, with the user voice.

Accordingly, the present inventors have recognized a need for better ways of reducing transmission of ambient noise through ANC and other types of headsets.

Summary

To address one or more of these and/or other needs or problems, the present disclosure includes, among other things, one or more exemplary systems, kits, methods, devices, assemblies, and/or components for automatically detecting the presence or absence of user speech and operating associated muting or other voice-or speech-dependent functionality. Some embodiments include a voice mic, at least two control mics, and processing circuitry, with the processing circuitry configured to mute the voice mic based on a relationship of the control mic output signals.

In an aspect, there is provided a communication assembly comprising:

an incoming audio signal node configured for connection to a communication device that provides an incoming audio signal Rx;

signal processing circuitry configured to output a voice activity detection signal indicative of a determined absence or presence of user speech, the signal processing circuitry configured to determine absence or presence of user speech based on a first signal ϕ , and a second signal θ , , with the ϕ and θ signals derived from two or more transducers responsive to at least the ambient noise a , user speech s , and an incoming audio signal u , with the ϕ and θ signals being determined from a transform matrix

$$\begin{bmatrix} \theta \\ \phi \end{bmatrix} = T \begin{bmatrix} S \\ a \\ u \end{bmatrix}$$

where T denotes the matrix of transfer functions, and is defined as

$$T = \begin{bmatrix} T_{s\theta} & T_{a\theta} & T_{u\theta} \\ T_{s\phi} & T_{a\phi} & T_{u\phi} \end{bmatrix}$$

wherein $T_{s\theta}$ represents a transfer function from user speech s to signal θ ; $T_{u\theta}$ represents a transfer function from incoming audio u to signal θ ;

$T_{a\theta}$ represents a transfer function from ambient noise a to signal θ ;

$T_{s\phi}$ represents a transfer function from user speech s to signal ϕ ; and

$T_{a\phi}$ represents a transfer function from ambient noise a to signal ϕ ; and

wherein the third signal is based at least in part on an externally generated audio signal Rx or at least one audio notification signal, the signal processing circuitry providing an indication that the user is speaking responsive to a ratio of a prediction error for θ to a variance of ϕ exceeding a detection threshold.

In an embodiment, there is provided a method of operating a communication assembly having a voice mic output signal and a set of two or more transducers, the method comprising: determining whether a user of the communication assembly is speaking based on the output of the two or more transducers; and in response to a determination that the use is not speaking, changing the operating state of one or more speech-dependent modules associated with the communication assembly to a resource conservation state.

In an embodiment there is provided a communication assembly comprising:

first and second transducers configured to be worn on a user head, with the first transducer configured for substantially greater acoustic coupling to a user vocal region than the second transducer; and

signal processing circuitry configured to output a voice activity detection signal indicative of a determined absence or presence of user speech in a predetermined frequency range when the first and second transducers are mounted to a user head; and

wherein the signal processing circuitry is further configured to indicate absence of user speech when the first and second transducers are positioned on a standard head and torso simulator (HATS) that is outputting a speech signal in the predetermined frequency range.

In a first embodiment, the two control mics, for example a left earpiece mounted mic and a right earpiece mounted mic, are configured for approximately symmetric acoustic coupling to a user vocal region. The processing circuitry determines whether a user is speaking based on measured temporal symmetry -- for example phase relationship-- of output signals from the two mics, with high symmetry indicative of user speech (arriving at both mics approximately simultaneously) and low symmetry indicative of ambient noise or non-user speech. The two control mics in

1 some variants of the first embodiment are the left and right feedback mics in an ANC
2 headset. Some other variants measure or estimate the phase relationship using
3 complex coherence of the sampled mic outputs, activating the muting circuitry
4 whenever the real part of the complex coherence falls below a threshold value.

5 In a second embodiment, the two control mics are configured for assymetric
6 acoustic coupling to the user vocal region, that is, with one mic having superior
7 acoustic coupling to the user vocal region than the other. The processing circuitry is
8 configured to turn the muting on or off based on the relative magnitude or the energy
9 level of the two mic output signals. For example, in some ANC embodiments, the mic
10 with superior acoustic coupling is the feedback error mic in an ANC earpiece and the
11 other is the feedforward mic. The processing circuitry compares a ratio of feedback
12 error mic output to the feedforward mic output to a threshold in determining whether to
13 turn the muting function on or off. Some variants may use two mics placed on a
14 boom or cord, rather than the ANC control mics.

15 The processing circuitry compares a ratio of feedback error mic output to the
16 feedforward mic output to a threshold in determining whether to turn the muting
17 function on or off. Some variants may use two mics placed on a boom or cord, rather
18 than the ANC control mics. More generalized embodiments use two or more
19 microphones to produce two signals, with a first signal being more closely coupled to
20 user speech than the second, and the second signal being more closely coupled to
21 ambient noise or sound than the first.

22 A third embodiment estimates a transfer function that accounts for both the
23 phase and magnitude differences between the signals from two mics, such as the
24 feedback error mic and the feedforward ambient mic. Some variants of this
25 embodiment may lack ANC mic and use other pairs of mics, such as a voice mic and
26 one other mic. The transfer function can be estimated using a variety of methods,
27 such a Kalman filtering or other form of optimal estimator.

28 One or more embodiments also include other features. For example, xom
29 embodiments include a canceler to cancel music or other audio signals that are output
30 via a loudspeaker within an earpiece associated with one or more of the microphones.
31 Some embodiments include muting delay and hold functions to reduce risk of muting
32 between spoken words or during short pauses. Still other embodiments include a
33 buffer memory and look-ahead functionality to ensure timely unmuting of the voice mic
34 and thus avoid partial loss of user speech.

Brief Description of the Drawings

Various embodiments are described herein with reference to the following attached figures (Figs). These figures are annotated with reference numbers for various features and components, and these numbers are used in the following description as a teaching aid, with like numbers referring to the same or similar features and components.

Fig.1 is a block diagram of an exemplary headset system or assembly 100 corresponding to one or more embodiments of the present invention.

Fig.2 is a block diagram of an exemplary method of operating a user voice activity detector portion of assembly 100, corresponding to one or more embodiments of the present invention.

Fig. 3 is a conceptual system diagram of one half of a headset system 300 corresponding to one or more embodiments of the present invention.

Fig. 4 is a block diagram of an generic electroacoustic system having multiple audio transducers and audio input signals, corresponding to one or more embodiments of the invention.

Fig. 5 is a block diagram of an user voice activity detector corresponding to one or more embodiments of the invention.

Detailed Description of the Exemplary Embodiment(s)

This document, which incorporates drawings and claims, describes one or more specific embodiments of one or more inventions. These embodiments, which are offered not to limit but only to exemplify and teach the invention, are shown and described in sufficient detail to enable those skilled in the art to implement or practice the invention(s). Moreover, where appropriate to avoid obscuring the invention(s), the description may omit certain information known to those of skill in the art.

Fig. 1 shows an exemplary ANR headset assembly or system 100 mounted in ears 102A and 102B of a user's head 101 (top view). User head 101 includes respective left and right ears 102A and 102B, a mouth 103, and a user vocal region 104. User vocal region 104 include the user mouth and voice box (not shown). Assembly 100 includes left and right earpieces 110A and 110B, an optional headband 110C, an in-line or boom voice microphone 120, processing circuitry 130, and a host or external device 140.

1 Earpieces 110A and 110B, which may take form of in-the-ear, over-the-ear, or
2 on-the-ear configurations and define a perpendicular axis 110D, include respective
3 ANR drivers DA and DB, feedforward mics FFA and FFB, and respective feedback
4 mics FBA and FBB. ANR drivers DA and DB divide the internal volumes of respective
5 earpieces 110A and 110B into front cavity and back cavity portions (not separately
6 labeled). Feedforward ambient mics FFA and FFB are positioned within or adjacent to
7 the back cavities, further away from respective ears 102A and 102B than feedback
8 mics FBA and FBB, ensuring that they provide greater acoustic coupling to ambient
9 noise and lesser acoustic coupling to user head 101, to the user ear canals, and to
10 vocal region 104 than feedback error mics FBA and FBB. Feedback error mics FBA
11 and FBB are positioned similarly within or on their respective earpieces, ensuring
12 generally identical or symmetric acoustic coupling via head or bone (body) conduction
13 paths 104A and 104B to user vocal region 104, which is on axis 110D. Additionally,
14 the feedback mics have a generally symmetric acoustic coupling via air conduction
15 paths P1 and P2 to off-axis external noise sources, such as ambient noise source N.
16 (The feedback and feedforward mics in the same earpiece have a very assymmetric
17 response to user speech and ambient.) The microphones and drivers are coupled via
18 respective wired or wireless communication links 131 and 132 to processing circuitry
19 130.

20 Processing circuitry 130, which in some embodiments take the form of a digital
21 signal processor with associated memory and other integrated componentry fully
22 separated or partially or fully integrated within device 140, includes a set of
23 input/output nodes 133, an ANR processing module 134, a user voice detection
24 module 135, a muting module 136, speech dependent modules 137, and on-the-ear
25 detection module 137.

26 Input/output nodes 133 includes a voice mic output signal node Vm, a feedback
27 mic left signal node FBA (fbL), a feedback mic right signal node FBB (fbR), a
28 feedforward mic left signal node FFA (ffL), a feedforward mic right signal node FFB
29 (ffR), incoming device audio/music signal nodes RxA RxB. (Rx as used herein may
30 refer to the RxA or RxB singly or collectively and may include telephone audio signal.)

31 ANR processing module 134 provides circuitry and machine-executable
32 instructions for canceling ambient noise within the earpiece front volumes, including
33 within ear cavities of the users. In particular, module 134 receives output signals from
34 feedback mics FFA and FFB, representing a summation of the total acoustic energy

when the user is speaking and substantially different at least in phase when the user is not. In other words, the output signals from the feedback mics have approximately the same speech data components, and thus exhibit high coherence (in phase relationship) when the user is speaking. Moreover, the distance from ambient noise to the two mics are usually not equal or asymmetric, as indicated by paths P1 and P2 in Fig. 1, meaning that the coherence will be rotated in phase. (It still may be coherent, meaning the magnitude may still be very close to 1, but the angle will not be 0, which indicates an 'in-phase' situation.

[0028] Thus, the exemplary embodiment uses the complex coherence of the left and right feedback error mic signals fbL and fbR within a threshold test to determine if user voice activity is likely to be present or not. If the threshold criteria is satisfied, then a mute command signal is provided to mute module 136, and if not, an unmute command signal is provided. More specifically, the threshold test takes the form of

[0029] If $\text{mean}(\text{real}(C_{xy}(\text{freq_range}))) > CT$ then Unmute; else Mute, where $\text{mean}()$ denotes an average function (over frequency); $\text{real}()$ denotes the real part of a complex argument; $C_{xy}(\text{freq_range})$ denotes the complex coherence of signals x and y over a frequency range denoted freq_range ; subscript x denotes the fbL error mic signal (left fb mic signal, FBA) and subscript y denotes the fbR error mic signal (right fb mic signal, FBB); CT denotes the coherence threshold. The real part of the complex coherence C_{xy} is used as this is a measure of how “in phase” the two signals are. Note: $\text{abs}(C_{xy})$ is 1 if only one average is taken, but this can still be useful as the important information is still in the phase. The freq_range over which the mean is computed can vary. However, because the body-conducted acoustics of the human voice box are primarily low frequency, some embodiments use the range of 70-700Hz or 100-500Hz. In response to an indication from the C_{xy} detector circuitry or logic that the user is not talking (that is user speech is absent), the exemplary embodiment mutes or attenuates the output of one or more microphones, such as the voice mic, and/or the feedforward microphones, using mute module 136.

[0030] Also, in some embodiments, the coherence threshold CT is in the range of 0.7 to 0.9, inclusive, with slight variations in performance. Setting it too close to 1 makes the detector more prone to false negatives (speech not detected when it exists) if there is significant background noise, and setting it too low yields more false positives (that is, unmuting when there is no speech). In some embodiments, the coherence threshold may be

1 mute command signal is provided to mute module 136, and if not, an unmute
 2 command signal is provided. More specifically, the threshold test takes the form of
 3 If $\text{mean}(\text{real}(C_{xy}(\text{freq_range}))) > CT$ then Unmute; else Mute,
 4 where mean denotes an average function (over frequency); $\text{real}()$ denotes the real
 5 part of a complex argument; $C_{xy}(\text{freq_range})$ denotes the complex coherence of
 6 signals x and y over a frequency range denoted freq_range ; subscript x denotes the
 7 fbL error mic signal (left fb mic signal, FBA) and subscript y denotes the fbR error mic
 8 signal (right fb mic signal, FBB); CT denotes the coherence threshold. The real part
 9 of the complex coherence C_{xy} is used as this is a measure of how "in phase" the two
 10 signals are. .Note: $\text{abs}(C_{xy})$ is 1 if only one average is taken, but this can still be
 11 useful as the important information is still in the phase. The freq_range over which the
 12 mean is computed can vary. However, because the body-conducted acoustics of the
 13 human voice box are primarily low frequency, some embodiments use the range of
 14 70-700Hz or 100-500Hz. In response to an indication from the C_{xy} detector circuitry
 15 or logic that the user is not talking (that is user speech is absent), the exemplary
 16 embodiment mutes or attenuates the output of one or more microphones, such as the
 17 voice mic, and/or the feedforward microphones, using mute module 136.

18 Also, in some embodiments, the coherence threshold CT is in the range of 0.7
 19 to 0.9, inclusive, with slight variations in performance. Setting it too close to 1 makes
 20 the detector more prone to false negatives (speech not detected when it exists) if
 21 there is significant background noise, and setting it too low yields more false positives
 22 (that is, unmuting when there is no speech). In some embodiments, the coherence
 23 threshold may be adjusted dynamically based on system or environmental
 24 parameters, such as signal (speech) to noise ratio. That is, if the speech is much
 25 stronger than noise, then some embodiments shift the threshold closer to 1, for
 26 example 0.9, and if it is very noisy, these embodiments decrease the threshold level to
 27 avoid muting user speech. An exemplary threshold between 0.7 and 0.8, such as
 28 0.75, is a good compromise.

29 This left-right symmetry based approach is effective, but may be vulnerable to
 30 on-axis ambient noise which can occur, for example, if a second speaker is on axis
 31 with the user (or generally equidistant from the left and right earpieces). It also
 32 suffers from poor signal-to-noise ratio (SNR.) On-axis noise will exhibit high
 33 coherence at the feedback error microphones and thus be falsely recognized as
 34 speech. Interestingly, one way to confirm this vulnerability is to place a headset using

1 this UVAD approach (for example with left and right feedforward or left and right
2 feedback mics, or more generally two mics configured to be approximately equidistant
3 from the vocal region) on a conventional Head and Torso Simulator (HATS), such as
4 B&K Type 4128-C simulator manufactured by Brüel & Kjær Sound & Vibration
5 Measurement A/S of Nærum, Denmark. The headsets can then be seen to
6 automatically mute in response to noise sources approximately equidistance from the
7 mics, for example, directly in front, behind, and above the HATS.

8 To reject on-axis ambient noise sources, such as non-user speech, and enable
9 better performance in noisy environments, some embodiments use a second method
10 of user voice activity detection, which has the added advantage of also being useful
11 for a single-earpiece systems, or more generally to any situation where two mics
12 having asymmetric coupling to a user and ambient sound. This method leverages
13 the insight that the magnitude of energy occurring at the feedback mics (more
14 generally, the mic with better acoustic coupling to the user vocal region) is
15 substantially greater than that occurring at the feedforward mics when the user is
16 speaking than when not speaking, due to asymmetry in the acoustic paths of voice
17 signals to the feedback and feedforward mics. The feedback mics are positioned in
18 the front cavity of the earpieces in or near the user ear canal. With this positioning,
19 the feedback mics receive user voice energy via bone conduction paths 104A and
20 104B with very low attenuation and receive ambient noise via air conduction with
21 strong attenuation due to the seal that is typically present in ANR systems. Therefore,
22 the feedback mic to feedforward mic output ratios (fb/ff energy ratio) when user voice
23 is present is much higher than that of the ambient noise without speech, regardless of
24 the noise direction. To this end, some embodiments determine user voice activity
25 using the following logic:

26
27 If $fb_left/ff_left < DT$ then Mute; else Unmute,

28 where the fb_left/ff_left denotes the ratio of the energy in output signal of the left
29 feedback mic (FBA) to the energy in the output signal of the left feedforward mic
30 (FFA), and DT denotes a selected detection threshold for user speech. In this
31 embodiment, DT is platform dependent; however, generally, when the user is
32 speaking, the ratio will increase appreciably over no-speech. (The ratio is a transfer
33 ratio, and in this case is a real number.)

Notably, some embodiments may use a complex transfer ratio, which includes both magnitude and phase, and thus provides improved performance. In dealing with a threshold in this complex case, the threshold is no longer a scaler, but a line (perhaps curved or pieced together, that divides the complex plane. For example, a line for the imaginary part >0 could be one threshold. Or the threshold might be that positive real and imaginary parts indicates speech, defining a quadrant of the complex plane.) Note that some embodiments may use the right feedback and feedforward microphones, rather than the left. Still other embodiments may automatically determine whether to use the right or left earpiece. For example, if the left earpiece is removed, on-ear detection circuitry determines that only the right earpiece is in place, and operates the voice detection module based on right feedback and feedforward microphones.

In determining the energy levels in the output of the feedback and feedforward microphones, one can compute a complex transfer function (TF) T_{xy} as

$$T_{xy} = S_{xy} / S_{xx},$$

where S_{xy} is the cross spectral density between x and y , and S_{xx} is the power spectral density of x . So S_{xy} can be estimated via FFTs in the following way:

$$S_{xy_hat} = \text{sum}(\text{fft}(y) * \text{conjugate}(\text{fft}(x))) / \text{sum}(\text{fft}(x) * \text{conj}(\text{fft}(x))).$$

or if only one fft is used (and no averaging) T_{xy} is defined as

$$T_{xy} = \text{fft}(y) / \text{fft}(x)$$

For the ratio of fb (feedback mic) energy to ff (feed forward mic) energy, x is the left feedforward mic signal ffL (FFA) and y is the left feedback mic signal fbL (FBA). The ratio of fb energy to ff energy is actually the $|T_{ff2fb}|^2$.

So in embodiments that use fft's as the core basis, one can threshold the ratio of absolute value squared of the fft result. Or if one were implementing without fft's, one could simply compute a moving average of the sampled time signals, after passing it through a bandpass filter and accomplish the similar results. Using the complex transfer function estimate (T_{xy} ..) one can take advantage of the significant phase information.

1 The detection threshold DT is set generally based on the physical
2 implementation of the earpieces and the quality of seal they provide for the front cavity
3 volume. For a good ANC headphone, one might expect 20dB of attenuation when the
4 user is not speaking. This rises to roughly 0dB with the user talking because of the
5 bone conduction mechanism.

6 However, the degree of change is more pronounced in in-the-ear, less
7 pronounced in on-the-ear, and least pronounced in around-the-ear types of
8 headphones, meaning use of different thresholds for each type of headset. For
9 example, in ITE headphones, it is expected that at certain frequencies, such as 100
10 Hz, there would be approximately 20~30dB more sound pressure in the blocked ear
11 (fb mic) than outside it (ff mic) due to user speech. This effect can also be leveraged
12 in circumaurals, but the difference in in-ear sound pressure for speech at 100 Hz
13 relative to external sound pressure at feedforward mic is perhaps only a few decibels.

14 Some embodiments may also vary the threshold dynamically based on
15 measured parameters such as ambient noise or average signal-to-noise ratios, or
16 alternatively provide a user adjustment mechanism via an app hosted by device 140,
17 for example. Some embodiments employ a dynamic detection threshold because the
18 fb/ff ratio is a function of total attenuation within the ANC headset and thus may vary
19 over a fairly wide range. In one embodiment, the detection threshold is estimated
20 during the two-ear smart mute period, specifically as a function of the moving average
21 of energy for mute and the moving average of energy for unmute. Some embodiments
22 measure the attenuation of the system, active plus passive, with the threshold
23 ultimately being a function of the active attenuation.

24 For the TF estimates, some embodiments update the average conditionally.
25 Also some embodiments recognize that the feedback mic signal includes multiple
26 components, namely audio input signal Rx (from device 140), ambient noise, user
27 speech, and measurement noise. This ensures good signal levels that are
28 uncorrelated with the ambient noise. Alternatively, for estimating the Noise transfer
29 function, some embodiments update the average when the energy ratio of fb/ff is low,
30 or highly correlated, ensuring that the TF estimates converge faster than they
31 otherwise would.

32 This second approach to user voice activity detection (UVAD) based on the
33 complex transfer characteristic of two control mics within certain frequency range is
34 believed to be particularly robust when used with any two mics that have a transfer

1 characteristic satisfying the following constraints. The first constraint is that the
2 transfer characteristic changes in a significant way for speech compared to
3 interference. The second constraint is that the transfer characteristic remain relatively
4 clear (i.e., relatively unambiguous) with changes in the relative direction of the
5 interference. These conditions are met with the feedforward and feedback mics in an
6 in-the-ear headset (and in other situations where one mic is more strongly acoustically
7 coupled to the user voice region than the other.)

8 For ambient noise, the output of the ambient pressure at the ambient mic
9 responds first, has leading phase relative to the ear-coupled feedback mic, for two
10 reasons. One is that it is directly coupled to the ambient and usually closer in acoustic
11 path length to the noise source. Two is that a typical headset earpiece will have some
12 amount of passive attenuation, and this is more or less a sort of low pass filter. i.e.
13 having no difference at very low frequencies, but attenuating the ear coupled mic (fb
14 mic) more significantly as frequencies rise. All causal low pass filters induce a phase
15 lag and all physical filters are causal.

16 For user speech, besides the acoustic path from the user's mouth to the ear,
17 and then the ear coupled mic, there is another path through the body from the vocal
18 cords. The speed of sound through the body is significantly faster, approximately 4~5
19 times faster or sub 1ms versus 3~4.5 milliseconds. (The speed of sound in air is
20 approximately 340 meters/second; in the flesh it is roughly 1500 meters/second; and
21 in skull and bone over 4000 meters per second.) As a result, the sound from the user
22 speech conducted through the users body arrives much sooner than does the air
23 transmission acoustic signal via the mouth.

24 When looking at the transfer characteristics between an ambient-coupled and
25 ear-coupled mics in close proximity, the ambient microphone will lead ear-coupled
26 mics independent of direction arrival. For user speech, the ear-coupled mic will lead
27 the ambient microphone. Thus it can be clearly seen that asymmetry in complex
28 transfer function (or any other basis) and the lack of large ambiguity caused by the
29 direction of the interference are far superior for the ambient and ear coupled
30 microphones from the same ear.

31 To take full advantage of the bone conduction path of the ear coupled mic,
32 referred to here as the feedback mic, fb, some embodiments account for incoming
33 audio "Rx" from an external device, such as device 140. (In some embodiments, Rx
34 may be mixed with one or more internally generated audio notification signals, such as

beeps or tones indicative of system events, prompts, etc.) This incoming audio will typically have characteristics strongly resembling bone conducted speech, meaning that it has much stronger fb_mic amplitude than the amplitude of the ambient mic (feed-forward mic, ff), and therefore can cause false user-speech detections.

One approach to mitigate this issue is to use an Rx canceler, which mathematically cancels or removes the effect of the Rx signal from the UVAD computations. The exemplary embodiment uses an Rx canceler based on decomposing the fb signal as follows:

$$fb = fb_{Rx} + fb_{ambient} + fb_{speech_BC},$$

where:

fb_{Rx} Rx is the fb mic signal that is due to the Rx signal, played out the ear coupled speaker;

$fb_{ambient}$ is the fb mic signal due to the ambient noise; and

fb_{speech_BC} is the fb mic signal due to bone conduction.

One can also further define fb_Rx and $fb_ambient$ as follows:

$$fb_{Rx} = Rx * T_{rx2fb};$$

$$fb_{ambient} = ff * T_{ff2fb};$$

where T_{rx2fb} is the transfer function from Rx to the fb mic with all other inputs zero and T_{ff2fb} is the transfer function from the feedforward mic to feedback mic with no speech or Rx, and only noise excitation. T_{rx2fb} and T_{ff2fb} can be estimated using several methods. For example, some embodiments use a Kalman filter, or traditional estimate based on the auto and cross spectra with care taken to, for example, not update the averages for T_{ff2fb} when Rx is present. Care also needs to be taken not to update estimates when user speech is present, but this is a much relaxed problem as the UVAD for this step need not catch all speech, but have high confidence that speech is not present.

Once estimates are available for these primarily static transfer functions one can then use them to estimate the fb_{speech_BC} signal in near real time. The estimates of T_{rx2fb} and T_{ff2fb} would be averaged over time. The exemplary embodiment uses fast fourier transforms (FFTs) to compute the estimates; however, some embodiments use any basis that sufficiently spans the subspace containing the bone conduction information.

1 Knowing $Trx2fb$ and $Tff2fb$, one can express fb_{speech_BC} as

$$2 \quad fb_{speech_BC} = fb - fb_{Rx} - fb_{ambient}$$

3 or as

$$4 \quad fb_{speech_BC}^{\wedge} = fb - Rx * T_{rx2fb}^{\wedge} - ff * T_{ff2fb}^{\wedge}$$

5 where $fb_{speech_BC}^{\wedge}$ is the estimate of fb_{speech_BC} .

6 Thus, the user speech detection is based on an estimated signal that is primarily free
7 of interference from ambient noise and incoming audio.

8 Notably this version of the asymmetrical approach (using the same side
9 feedback and feedforward mics) relies on the bone conduction path between the user
10 vocal region and the feedback mic. Thus, placement of a headset employing this
11 asymmetrical approach on a conventional HATS (such as the B&K 4128-C simulator
12 mentioned above) would generally prevent proper operation of the UVAD, because
13 conventional HATS lack the bone conduction path. In other words, the headset
14 mounted to the HATS would fail to mute and unmute appropriately (or otherwise fail to
15 correctly detect user voice activity) in response to user voice signal of the appropriate
16 frequency range input to the HATS (A path from the vocal region to the feedback mic
17 would exist due to vibration, but this would be very weak coupling in comparison to
18 actual bone conduction.)

19 Some embodiments combine the symmetry and asymmetry based threshold
20 tests together, as follows:

21

22 if (coherence_of_left_and_right < CT), then Mute;

23 else { if ((fb_left/ff_left) < ET, Mute; else Unmute }

24 Notably, implementing this detection logic entails use of three control mics, the left
25 and right feedback error mics and one of the feedforward mics. Also, this logic only
26 allows the asymmetrical threshold test (the ratio of feedback mic energy to feedforward
27 mic energy) to control the unmuting. Other embodiments could allow both to trigger
28 unmuting.

29 Additionally, some embodiments provide buffers within processing circuitry 130
30 to perform voice detection on a delayed version of the relevant signal data. More
31 specifically, some embodiments employ an X-millisecond look ahead buffer to allow
32 the detector time enough to detect presence of user speech in the headset and
33 unmute the system without loss of any speech content. For example, one

1 embodiment stores 20 millisecond set of sampled data from the system mics, allowing
2 the detector to detect user speech at sample n and then to unmute all prior samples
3 taken in the prior 20 millisecond period, and thus avoid skipping or muting the first part
4 of a user phrase. In some embodiments, the length the look ahead period may be
5 adjusted or calibrated by the user, and in others it may be dynamically determined
6 based on a detected speech cadence of the user, for example a rolling average of
7 distance between peaks in the speech signal.

8 Mute module 136 provides the muting function in response to the command
9 signals from user voice detection module 135. In some embodiment, this entails
10 turning off the signal paths associated with the voice mic and/or one or more other
11 mics within the assembly. However, to improve user experience, some embodiments
12 delay activation or onset of the mute for 3, 5, 10, 15, or 20 milliseconds, adding a
13 margin of safety to avoid clipping off the ends of phrases or muting between words. In
14 some embodiments, the period of this delay may be set by the user or dynamically
15 determined based on measured speech cadence of the user. Also, in some
16 embodiments, a visual, audible, or haptic indication is provided in response to
17 activation and deactivation of the mute function to alert the user of the change in mute
18 state. In some embodiments, one or more of these indicators are provided on the
19 headset itself and/or on device 140. In some instances, the visual indicator takes the
20 form of an illuminated and/or flashing LED (light emitting diode) on the headset and/or
21 an illuminated or flashing or changed coloration or shape of a microphone icon on the
22 device display. In some embodiments, the user may override the mute function via a
23 manual control element on the headset and/or device, with the override having effect
24 for predetermined time period, for example, 1, 2, or 3 minutes, or until termination of
25 current communication session, such as phone call. At the end of the override period,
26 automatic muting and unmuting would resume.

27 More generally, some embodiments ramp or deliberately slow down or decay
28 the state changes between mute on and mute off (and vice versa) to avoid very fast
29 microphone gain changes that might otherwise create audible pops or clicks that are
30 annoying and indicative of poor quality audio components. This is generally dealt with
31 by making the gain change gradually rather than instantly. For example, in one
32 embodiment, the “attack” off to on occurs over approximately 100 msec, which is slow
33 enough to avoid pop and long enough to minimize look ahead memory requirements.
34 One embodiment uses a decay function with the following form:

```
1      If start mute {  
2          static float decayCoef = 1;  
3          if (decayCoef > 0.00316) decayCoef *= 0.995      // -25dB  
4          Scale the output with decayCoef  }
```

5 Some embodiments facilitate more effective auto-muting by detecting when a
6 two earpiece system has only one ear-piece properly mounted. Some of these
7 embodiments rely on an On Ear Detector (OED) to optimize performance. Details of
8 the OED are further described in co-owned U.S. Patent Application 14/850,859 filed
9 September 10, 2015, the disclosure of which is incorporated herein by reference in its
10 entirety.

11 Speech-dependent module 136 represent one or more other functions (of
12 processor circuitry 130 and/or device 140) that receive a binary speech-present or
13 speech-not-present signal from voice activity detection module 135. Some of these
14 modules use the signal to activate or deactivate the function of the module,
15 conserving processing power, memory, and/or battery life. For example, in some
16 embodiments, speech-dependent module 137 includes a speech or keyword
17 recognition module that is configured to listen for specific keyword commands or to
18 perform more generalized speech recognition functions.

19 In some other embodiments, module 137 further includes a noise-reduction
20 module that provides further processing to reduce noise in the voice mic signal. This
21 noise reduction module, in some embodiments, can be tuned to the specific
22 environment of the user. And in still other embodiments, speech-dependent module
23 136 includes a side-tone module or circuitry that receives voice mic output and
24 produces a 3-10% user side-tone signal for one or both earpieces. Generating the
25 side-tone consumes power, so switching this function off if the user is not speaking
26 conserves battery life. See U.S. Provisional Patent Application 62/530,049 filed July
27 7, 2017, which is incorporated herein by reference.

28

29 Fig. 2 shows a flow chart 200 of an enhanced automatic muting system for a
30 ANR headset with two earpieces, Flow chart 200 includes process blocks 210-280.

31 Block 210 entails performing OED (on-ear detection) to determine the status of
32 the earpieces. (See co-pending U.S. Patent Application 14/850,859 filed on

September 10, 2015), which is incorporated herein by reference.) Execution then continues at block 220 which entails determining whether the ambient noise level is low or not. If the ambient noise is low, the smart mute capability of module 134 is disabled at block 230 and execution returns to block 220; otherwise execution continues at block 240.

Block 240 entails determining whether both earpieces are mounted to the user. In the exemplary embodiment, this entails another call to OED module. If both earbuds are installed, execution branches to block 250 which determines whether to mute or not, using the symmetry based or combined symmetry-assymetry muting functions described above, both of which require signals from both earpieces. Execution from here cycles back to block 240 to determine if both earpieces are still mounted. If both earbuds are determined not to be mounted, execution proceeds to block 260, which determines whether one earpiece is mounted. (It also determines which one is mounted.) If one is mounted, execution branches to block 270, to perform smart muting function based on the assymetrical thresholding test as described above. Execution then cycles back to block 260 to determine if one earpiece is still mounted. (In some embodiments, the earpiece may still be mounted, but have insufficient battery power.) If there is not one earpiece mounted, no smart mute is performed and execution branches back to block 220.

Fig. 3 shows a monaural system model 300 of system 100, including the following blocks: T_p is the passive attenuation transfer function; T_{dm} is the driver-to-feedback-mic transfer function; H_{ff} is the feedforward filter; H_{fb} is the feedback filter; V is the user-speech-to-feedforward-mic acoustic path (transfer function); W is the user-speech-to-feedback-mic bone-conduction path (transfer function). The model also includes following signals: s is the user speech signal; a is the ambient noise signal; n is the feedforward mic measurement (or more generally, the mic farthest from or less acoustically coupled to the vocal region); m is the feedback mic measurement (or more generally mic farthest from or more acoustically coupled to the vocal region); u is the RX signal and/or any system audio notification signals; and d is the DAC (driver) output.

More particularly, the Fig. 3 system has both feedforward and feedback filters H_{ff} and H_{fb} present. However, some embodiments omit these filters, meaning that ($H_{ff} = H_{fb} = 0$) and the headset is passive. The exemplary embodiment uses the following linear statistical model:

$$M = T_{dm}D + T_pN + W$$

$$D = H_{fb}M + H_{ff}N + U$$

$$N = A + VS$$

Substituting D into M gives

$$M = T_{dm}(H_{fb}M + H_{ff}N + U) + T_pN + WS$$

Collecting terms yields

$$(1 - T_{dm}H_{fb})M = (T_{dm}H_{ff} + T_p)N + T_{dm}U + WS$$

Substituting N gives

$$(1 - T_{dm}H_{fb})M = (T_{dm}H_{ff} + T_p)(A + VS) + T_{dm}U + WS$$

Collecting terms yields

$$(1 - T_{dm}H_{fb})M = (T_{dm}H_{ff} + T_p)A + T_{dm}U + ((T_{dm}H_{ff} + T_p)V + W)S$$

$$M = F_1A + F_2U + (F_1V + F_s)S$$

where

$$F_1 = \frac{T_{dm}H_{ff} + T_p}{1 - T_{dm}H_{fb}}$$

$$F_2 = \frac{T_{dm}}{1 - T_{dm}H_{fb}}$$

$$F_s = \frac{W}{1 - T_{dm}H_{fb}}$$

The objective in the linear modeling is to decompose the feedback microphone measurement M into a linear combination of independent components (ambient noise, Rx, and user speech). This model is applicable for narrowband signals, i.e. for a particular frequency bin. To model a broadband system one would add frequency indices to each term.

Variances can be expressed as follows:

$$E[AA^*] = \sigma_A^2$$

$$E[SS^*] = \sigma_S^2$$

$$E[NN^*] = E[(A + VS)(A + VS)^*]$$

$$E[NN^*] = \sigma_A^2 + VV^*\sigma_S^2$$

use of a reference to U , simplifies the estimation problem, though it is not required. Some embodiments formulate the estimation problem in a state-space framework using the Measurement Model $M(t)$ defined as

$$M(t) = G(t)N(t) + H(t)U(t) + \sqrt{\sigma_r}r(t)$$

where $r(t)$ is a Gaussian random variable with zero mean and unit variance, σ_r is a tunable parameter that explains unmodeled behavior (i.e. slight nonlinearities in the measurements). $H(t)$ is a regression coefficient that explains the contribution of the rx/music signal to the feedback mic measurement.

[0069] Some embodiments use the following process model:

$$\begin{bmatrix} G(t+1) \\ H(t+1) \end{bmatrix} = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} G(t) \\ H(t) \end{bmatrix} + \begin{bmatrix} (1 - \alpha_1^2)^{1/2} & 0 \\ 0 & (1 - \alpha_2^2)^{1/2} \end{bmatrix} \begin{bmatrix} q_1(t) \\ q_2(t) \end{bmatrix}$$

Where $q_1(t)$ and $q_2(t)$ are independent Gaussian random variables with zero means and unit variances. α_1 and α_2 are tunable parameters that govern how quickly $G(t)$ and $H(t)$ can change over time.

[0070] The state-space framework is useful because there are efficient algorithms for state estimation Recursive Least Square (RLS), Least Mean Square (LMS), and Kalman filter, for example. Some embodiments estimate the states, $G(t)$ and $H(t)$, in several frequency bins by using a Kalman filter in each bin.

[0071] Some embodiments incorporate a more generalized approach to user voice activity detection which avoids the need for explicitly characterizing the electroacoustical parameters of the system. The generalized approach uses two signals θ (theta) and ϕ (phi) which are each a function of at least two of user speech s , ambient noise a , and/or incoming audio u , according to the matrix of transfer functions T . (Incoming audio u can be a mix of externally generated audio signals such as Rx/music from a communication device and internally generated audio signals, such as system event prompts, notifications, or alarms) The matrix T of transfer functions is determined by how speech, ambient and incoming audio u appear on the two or more transducers, and how the transducers are combined with a reference to the Rx/music to generate phi and theta. Mathematically, this can be expressed as

$$\begin{bmatrix} \theta \\ \phi \end{bmatrix} = T \begin{bmatrix} s \\ a \\ u \end{bmatrix}$$

where T denotes the matrix of transfer functions, and is defined as

measurements). $H(t)$ is a regression coefficient that explains the contribution of the rx/music signal to the feedback mic measurement.

Some embodiments use the following process model:

$$\begin{bmatrix} G(t+1) \\ H(t+1) \end{bmatrix} = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} G(t) \\ H(t) \end{bmatrix} + \begin{bmatrix} (1-\alpha_1^2)^{1/2} & 0 \\ 0 & (1-\alpha_2^2)^{1/2} \end{bmatrix} \begin{bmatrix} q_1(t) \\ q_2(t) \end{bmatrix}$$

where $q_1(t)$ and $q_2(t)$ are independent Gaussian random variables with zero means and unit variances α_1 and α_2 are tunable parameters that govern how quickly $G(t)$ and $H(t)$ can change over time.

The state-space framework is useful because there are efficient algorithms for state estimation Recursive Least Square (RLS), Least Mean Square (LMS), and Kalman filter, for example). Some embodiments estimate the states, $G(t)$ and $H(t)$, in several frequency bins by using a Kalman filter in each bin.

Some embodiments incorporate a more generalized approach to user voice activity detection which avoids the need for explicitly characterizing the electroacoustical parameters of the system. The generalized approach uses two signals θ (theta) and ϕ (phi) which are each a function of at least two of user speech s , ambient noise a , and/or incoming audio u , according to the matrix of transfer functions T . (Incoming audio u can be a mix of externally generated audio signals such as Rx/music from a communication device and internally generated audio signals, such as system event prompts, notifications, or alarms) The matrix T of transfer functions is determined by how speech, ambient and incoming audio u appear on the two or more transducers, and how the transducers are combined with a reference to the Rx/music to generate phi and theta. Mathematically, this can be expressed as

$$\begin{bmatrix} \theta \\ \phi \end{bmatrix} = T \begin{bmatrix} s \\ a \\ u \end{bmatrix}$$

where T denotes the matrix of transfer functions, and is defined as

$$T = \begin{bmatrix} T_{s\theta} & T_{a\theta} & T_{u\theta} \\ T_{s\phi} & T_{a\phi} & T_{u\phi} \end{bmatrix}$$

wherein $T_{s\theta}$ represents a transfer function from user speech s to signal θ ; $T_{u\theta}$ represents a transfer function from incoming audio u to signal θ ; $T_{a\theta}$ represents a transfer function from ambient noise a to signal θ ; $T_{s\phi}$ represents a transfer function from user speech s to signal ϕ ; and $T_{a\phi}$ represents a transfer function

1 from ambient noise a to signal Φ . With this model, reliable detection of user voice
 2 activity requires adherence to the following asymmetrical constraint:

$$3 \quad \left| \frac{T_{s\theta}}{T_{a\theta}} - \frac{T_{s\phi}}{T_{a\phi}} \right|^2 > z$$

4 where z denotes the asymmetry threshold z . Which indicates that the speech-to-
 5 ambient sensitivity ratios must be different in magnitude and/or phase. In some
 6 embodiments z equals zero and in other embodiments z equals 0.01, 0.1, 0.5, 0.75,
 7 1.0, 2.0. In still other embodiments, z is greater than 2.

8 Mapping the terms of this more generalized model to the specific electroacoustical
 9 implementation of Fig. 3, shows the corresponding T matrix as

$$10 \quad T = \begin{bmatrix} \frac{W}{1 - T_{dm}H_{fb}} & \frac{T_{dm}H_{ff} + T_p}{1 - T_{dm}H_{fb}} & \frac{T_{dm}}{1 - T_{dm}H_{fb}} \\ V & 1 & 0 \end{bmatrix}$$

11 where $T_{s\theta} = \frac{W}{1 - T_{dm}H_{fb}}$, $T_{s\phi} = V$, $T_{a\theta} = \frac{T_{dm}H_{ff} + T_p}{1 - T_{dm}H_{fb}}$, $T_{a\phi} = 1$, $T_{u\theta} = \frac{T_{dm}}{1 - T_{dm}H_{fb}}$, $T_{u\phi} = 0$.

12 Some embodiments may use an alternative asymmetrical constraint of the form

$$13 \quad \left| \frac{T_{s\theta}}{T_{a\theta}} \right|^2 > \left| \frac{T_{s\phi}}{T_{a\phi}} \right|^2$$

14 which requires that the ratio of speech to ambient signal power in signal θ is greater
 15 than the ratio of speech to ambient signal power in signal Φ .

16

17 Fig. 4 shows a block diagram of a general electroacoustical system 400
 18 illustrating generation of signal θ from a linear combination of mic inputs and Rx/music
 19 (incoming audio) inputs. System 400 includes a set of microphone or transducer
 20 measurements or inputs $M_1 \dots M_N$ and Rx/music signal references U_1 and U_2 which
 21 are provided via respective gain modules or blocks $K_1 \dots K_N$ to filters H having respective
 22 transfer functions $H_1 \dots H_N$. The filter outputs feed into a summer, which produces θ .
 23 Thus, signal θ is a filtered combination of the transducer and reference inputs.

24 Fig. 5 shows a block diagram of generalized UVAD module 500 which can be
 25 used as part of UVAD module 135 and within the Fig. 2 process. In addition to input
 26 signals user speech s , ambient noise a , and incoming device audio u , module 500
 27 includes an electroacoustic system model block 510, estimator block 520, summer

1 530, variance ratio estimator block 540, and decision block 550. Electroacoustic
2 system model block 510 is generally representative of T , the matrix of transfer
3 functions T and any electroacoustic system, such as system 100 or 500.

4 Estimator block 520 iteratively predicts or estimates (θ) from ϕ and u , with
5 prediction error signal e from summer block 530 being fed back to update each new
6 prediction. In the exemplary embodiment, estimator block 520 takes the form of a
7 Kalman filter (as described above). However, other embodiments use forms of linear
8 estimators, such as RLS and LMS estimators. $\theta \phi$.

9 Variance ratio estimator block 540 estimates the variance for the prediction
10 error, S_e and the variance of signal ϕ , S_ϕ , and computes the ratio S_e/S_ϕ . The ratio is
11 provided to decision block 550, which compares the ratio to detection threshold DT . If
12 the ratio exceeds the threshold the user voice activity detection signal is set to 1
13 indicating the presence of user speech. Otherwise, the detection signal is set to 0.

14 15 Conclusion

16 In the foregoing specification, specific exemplary embodiments have been
17 described. However, one of ordinary skill in the art appreciates that various
18 modifications and changes can be made without departing from the scope of the
19 invention as set forth in the claims below. Accordingly, the specification and figures
20 are to be regarded in an illustrative rather than a restrictive sense, and all such
21 modifications are intended to be included within the scope of present teachings.

22 The benefits, advantages, solutions to problems, and any element(s) that may
23 cause any benefit, advantage, or solution to occur or become more pronounced are
24 not to be construed as a critical, required, or essential features or elements of any or
25 all the claims. The invention is defined solely by the appended claims including any
26 amendments made during the pendency of this application and all equivalents of
27 those claims as issued.

28 Moreover in this document, relational terms, such as second, top and bottom,
29 and the like may be used solely to distinguish one entity or action from another entity
30 or action without necessarily requiring or implying any actual such relationship or
31 order between such entities or actions. The terms "comprises," "comprising," "has",
32 "having," "includes", "including," "contains", "containing" or any other variation thereof,
33 are intended to cover a non-exclusive inclusion, such that a process, method, article,
34 or apparatus that comprises, has, includes, contains a list of elements does not

1 include only those elements but may include other elements not expressly listed or
2 inherent to such process, method, article, or apparatus. An element proceeded by
3 “comprises a”, “has ...a”, “includes ...a”, “contains ...a” does not, without more
4 constraints, preclude the existence of additional identical elements in the process,
5 method, article, or apparatus that comprises, has, includes, contains the element.
6 The terms “a” and “an” are defined as one or more unless explicitly stated otherwise
7 herein. The terms “substantially”, “essentially”, “approximately”, “about” or any other
8 version thereof, are defined as being close to as understood by one of ordinary skill in
9 the art, and in one non-limiting embodiment the term is defined to be within 10%, in
10 another embodiment within 5%, in another embodiment within 1% and in another
11 embodiment within 0.5%. The term “coupled” as used herein is defined as connected,
12 although not necessarily directly and not necessarily mechanically. A device or
13 structure that is “configured” in a certain way is configured in at least that way, but
14 may also be configured in ways that are not listed. Also, the term “exemplary” is used
15 as an adjective herein to modify one or more nouns, such as embodiment, system,
16 method, device, and is meant to indicate specifically that the noun is provided as a
17 non-limiting example.

WHAT IS CLAIMED IS:

1. A communication assembly comprising:
an incoming audio signal node configured for connection to a communication device that provides an incoming audio signal Rx;
signal processing circuitry configured to output a voice activity detection signal indicative of a determined absence or presence of user speech, the signal processing circuitry configured to determine absence or presence of user speech based on a first signal ϕ , and a second signal θ , with the ϕ and θ signals derived from two or more transducers responsive to at least the ambient noise a , user speech s , and an incoming audio signal u , with the ϕ and θ signals being determined from a transform matrix

$$\begin{bmatrix} \theta \\ \phi \end{bmatrix} = \mathbf{T} \begin{bmatrix} s \\ a \\ u \end{bmatrix}$$

where $\mathbf{T}$ denotes the matrix of transfer functions, and is defined as

$$\mathbf{T} = \begin{bmatrix} T_{s\theta} & T_{a\theta} & T_{u\theta} \\ T_{s\phi} & T_{a\phi} & T_{u\phi} \end{bmatrix}$$

wherein $T_{s\theta}$ represents a transfer function from user speech s to signal θ ; $T_{u\theta}$ represents a transfer function from incoming audio u to signal θ ; $T_{a\theta}$ represents a transfer function from ambient noise a to signal θ ; $T_{s\phi}$ represents a transfer function from user speech s to signal ϕ ; and $T_{a\phi}$ represents a transfer function from ambient noise a to signal ϕ ; and $T_{u\phi}$ represents a transfer function from incoming audio u to signal ϕ ; and wherein the incoming audio signal u is based at least in part on an externally generated audio signal Rx or at least one audio notification signal, the signal processing circuitry providing an indication that the user is speaking responsive to a ratio of a prediction error for θ to a variance of ϕ exceeding a detection threshold.

2. The communication assembly of claim 1, wherein the two or more transducers include first and second mics, with the first mic positioned with substantially greater acoustic coupling to a user vocal region than the second mic.

3. The communication assembly of claim 2, wherein the first mic is positioned to receive more acoustic energy from the user vocal region via a bone conduction path than the second mic.

4. The communication assembly of claim 1, further comprising:
an automatic muting module configured to attenuate a voice mic output signal in response to the voice activity detection signal indicating a determined absence of user speech and to de-attenuate the voice mic output signal in response to the voice activity detection signal indicating presence of user speech.
5. The communication assembly of claim 1, wherein the automatic muting module is configured to mute or attenuate the voice mic output signal a predetermined amount of time after receipt of the voice activity detection signal.
6. The communication assembly of claim 1, further comprising:
a keyword recognition module configured to recognize one or more keywords within user speech signals derived from a voice mic output signal and responsive to the voice activity detection signal indicating a determined absence of user speech to enter a shutdown or sleep state to conserve power or memory.
7. The communication assembly of claim 1, further comprising: a side tone module responsive to user speech to provide a side tone signal to a loud speaker within a headset earpiece and responsive to the voice activity detection signal indicating a determined absence of user speech to enter a shutdown or sleep state to conserve power or memory.
8. The communication assembly of claim 1, further comprising: a side tone module responsive to user speech to provide a side tone signal to a loud speaker within a headset earpiece and responsive to the voice activity detection signal indicating a determined absence of user speech to enter a shutdown or sleep state to conserve power or memory.
9. The communication assembly of claim 1, further comprising: a noise reduction module responsive to the voice activity detection signal indicating a determined presence of user speech to reduce noise in a first way and responsive to the voice activity detection signal indicating a determined absence of user speech to reduce noise in a second way.

10. The communication assembly of claim 1, further comprising: at least one speech-dependent module configured for operation in a manner dependent on speech content of a voice mic output signal and responsive to the voice activity detection signal indicating a determined absence of user speech to enter a power and/or memory conservation mode.

11. The communication assembly of claim 1,
wherein the assembly includes a first node configured for connection to a loud speaker associated with a headset earpiece; a second node configured for connection to a feedforward ambient mic associated with the headset earpiece, and a third node configured for connection to a feedback error mic associated with headset earpiece;
wherein signal processing circuitry is coupled to the first, second and third nodes, and configured to provide an anti-noise signal to the loud speaker in response to signals originating from the feedforward ambient mic and the feedback error mic; and
wherein the ϕ and θ signals are derived from signals originating from the feedforward ambient mic and the feedback error mic.

12. The communication assembly of claim 11, wherein the signal processing circuitry is further configured to indicate absence of the user speech when the first and second transducers are placed on a standard head and torso simulator (HATS) that is outputting a speech signal in the range.

13. The communication assembly of claim 11, wherein the headset earpiece is an in-the-ear earpiece having an ear-canal portion containing the feedback error mic.