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 signs is provided. More specifically, the threshold test takes the form of

[0029] If mean (real(Cxy(freq_range))) > CT then Unmute; else Mute,

where mean () denotes an average function (over frequency); real () denotes the real part of a complex argument; Cxy (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 Cxy is used as this is a measure of how "in phase" the two signals are. Note: abs(Cxy) 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 Cxy 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

adjusted dynamically based on system or environmental parameters, such as signal (speech) to noise ratio. That is, if the speech is much stronger than noise, then some embodiments shift the threshold closer to 1, for example 0.9, and if it is very noisy, these embodiments decrease the threshold level to avoid muting user speech. An exemplary threshold between 0.7 and 0.8, such as 0.75, is a good compromise.

[0031] This left-right symmetry based approach is effective, but may be vulnerable to on-axis ambient noise which can occur, for example, if a second speaker is on axis with the user (or generally equidistant from the left and right earpieces). It also suffers from poor signal-to-noise ratio (SNR.) On-axis noise will exhibit high coherence at the feedback error microphones and thus be falsely recognized as speech. Interestingly, one way to confirm this vulnerability is to place a headset using this UVAD approach (for example with left and right feedforward or left and right feedback mics, or more generally two mics configured to be approximately equidistant from the vocal region) on a conventional Head and Torso Simulator (HATS), such as B&K Type 4128-C simulator manufactured by Brüel & Kjær Sound & Vibration Measurement A/S of Nærum, Denmark. The headsets can then be seen to automatically mute in response to noise sources approximately equidistance from the mics, for example, directly in front, behind, and above the HATS.

[0032] To reject on-axis ambient noise sources, such as non-user speech, and enable better performance in noisy environments, some embodiments use a second method of user voice activity detection, which has the added advantage of also being useful for a single-earpiece systems, or more generally to any situation where two mics having asymmetric coupling to a user and ambient sound. This method leverages the insight that the magnitude of energy occurring at the feedback mics (more generally, the mic with better acoustic coupling to the user vocal region) is substantially greater than that occurring at the feedforward mics when the user is speaking than when not speaking, due to asymmetry in the acoustic paths of voice signals to the feedback and feedforward mics. The feedback mics are positioned in the front cavity of the earpieces in or near the user ear canal. With this positioning, the feedback mics receive user voice energy via bone conduction paths 104A and 104B with very low attenuation and receive ambient noise via air conduction with strong attenuation due to the seal that is typically present in ANR systems. Therefore, the feedback mic to feedforward mic output ratios (fb/ff energy ratio) when user voice is present is much

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

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

wherein $T_{s\theta}$ represents a transfer function from user speech s to signal θ; $T_{a\theta}$ 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. With this model, reliable detection of user voice activity requires adherence to the following asymmetrical constraint:

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

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

[0072] Mapping the terms of this more generalized model to the specific electroacoustical implementation of Fig. 3, shows the corresponding T matrix as

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

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

[0073] Some embodiments may use an alternative asymmetrical constraint of the form

$$\left|\frac{Ts\theta}{Ta\theta}\right|^2 > \left|\frac{Ts\phi}{Ta\phi}\right|^2$$

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

[0074] Fig. 4 shows a block diagram of a general electroacoustical system 400 illustrating generation of signal θ from a linear combination of mic inputs and Rx/music (incoming audio) inputs. System 400 includes a set of microphone or transducer measurements or inputs $M_1...M_N$ and Rx/music signal references U1 and U2 which are provided via respective gain modules or blocks K_1-K_N to filters H having respective transfer

a keyword recognition module configured to recognize one or more spoken words or phrases; and

a side tone module configured to generate a side tone signal based on the voice mic output signal.

16. The method of claim 14, wherein determining whether a user of the communication assembly is speaking based on the output of the two or more transducers, comprises

deriving a first signal Φ, a second signal θ, and a third signal, with the ϕ and θ signals derived from the two or more transducers transducing an ambient noise signal a, user speech signal s, and a third signal, with the ϕ and θ signals satisfying a constraint

$$\left|\frac{T_{S\theta}}{T_{a\theta}} - \frac{T_{S\varphi}}{T_{a\varphi}}\right|^2 > 0,$$

wherein $T_{s\theta}$ represents a transfer function from user speech s 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 a function of at least an incoming audio input signal from an external device or at least one audio notification signal.

17. 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.

18. The communication assembly of claim 17,

wherein the assembly includes an earpiece having a loud speaker, a feedforward ambient mic, a feedback error mic, a voice mic, and a an incoming audio signal node configured for connection to a communication device that provides an incoming audio signal Rx;

wherein the signal processing circuitry is configured to determine absence or presence of user speech based on a first signal Φ, a second signal θ, and a third signal, with the ϕ and θ signals derived at least from output signals of at least two of the feedforward ambient mic, the feedback error mic, and the voice mic in response to ambient noise a, user speech s, and a third signal, with the ϕ and θ signals satisfying a constraint

$$\left|\frac{T_{S\theta}}{T_{a\theta}} - \frac{T_{S\varphi}}{T_{a\varphi}}\right|^2 > 0,$$

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

$T_{s\theta}$ 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 incoming audio signal Rx or at least one audio notification signal.

19. The communication assembly of claim 18:

wherein the processing circuitry is further 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 assembly further comprises a set of one or more speech-dependent modules, with each speech dependent module responsive to the voice activity detection signal indicating absence of user speech to enter a power or memory conservation state and responsive to the voice activity detection signal