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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
24 September 2009 (24.09.2009)

PCT

(10) International Publication Number
WO 2009/117084 A2

(51) International Patent Classification:
H04B 3/20 (2006.01)

(21) International Application Number:
PCT/US2009/001667

(22) International Filing Date:
16 March 2009 (16.03.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
12/077,436 18 March 2008 (18.03.2008) US

(71) Applicant *(for all designated States except US)*: **AUDIENCE, INC.** [US/US]; 440 Clyde Avenue, Mountain View, CA 94043 (US).

(72) Inventor; and

(75) Inventor/Applicant *(for US only)*: **KLEIN, David** [US/US]; C/o Audience, Inc., 440 Clyde Avenue, Mountain View, CA 94043 (US).

(74) Agents: **BACHMANN, Stephen** et al.; 2200 Geng Road, Palo Alto, CA 94303 (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, BR, BW, BY, BZ, CA, CH, 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, KM, 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, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, 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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (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, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

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

(54) **Title**: SYSTEM AND METHOD FOR ENVELOPE-BASED ACOUSTIC ECHO CANCELLATION

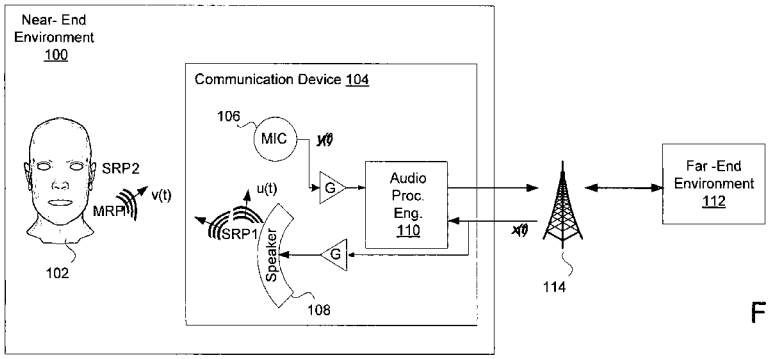


FIG. 1

(57) **Abstract**: Systems and methods for envelope-based acoustic echo cancellation in a communication device are provided. In exemplary embodiments, a primary acoustic signal is received via a microphone of the communication device, and a far- end signal is received via a receiver. Frequency analysis is performed on the primary acoustic signal and the far-end acoustic signal to obtain frequency sub- bands. An echo gain mask based on magnitude envelopes of the primary and far- end acoustic signals for each frequency sub-band is generated. A noise gain mask based on at least the primary acoustic signal for each frequency sub-band may also be generated. A combination of the echo gain mask and noise gain mask may then be applied to the primary acoustic signal to generate a masked signal. The masked signal is then output.

WO 2009/117084 A2

SYSTEM AND METHOD FOR ENVELOPE-BASED ACOUSTIC ECHO CANCELLATION

By: David Klein

BACKGROUND OF THE INVENTION

Field of Invention

[0001] The present invention relates generally to audio processing and, more particularly, to envelope-based acoustic echo cancellation in an audio system.

Description of Related Art

[0002] Conventionally, when audio from a far-end environment is presented through a speaker of a near-end communication device, sounds from a far-end audio signal may be picked up by microphones or other audio sensors of the near-end communication device. As such, the sounds from the far-end audio signal may be sent back to the far-end environment resulting in an echo to a far-end listener.

[0003] Conventionally, acoustic echo cancellation (AEC) systems may take the far-end audio signal and use it to predict the echo of the far-end audio signal (after being played through the speaker and picked up by the microphone). Typically, a transfer function that describes a path from the far-end audio signal, through the speaker, through an acoustic environment, and back to the microphone is linearly modeled to predict the echo. These AEC systems are performed in a waveform domain whereby the echo is predicted, inverted, delayed, and subtracted out from a near-end audio signal.

[0004] Disadvantageously, there are many problems with these conventional AEC systems. First, the transfer function (i.e., relationship between the far-end audio signal and the echo) is typically constantly changing, since the acoustic environment is rarely fixed. In the case of a handheld communication device (e.g., a cellular phone), there may also be reflections off a face of a user. The prior art AEC systems are adaptive and continually update the transfer function. However, errors

usually occur in the echo prediction due to the changing environment. If the echo prediction is even slightly incorrect, or an applied delay is incorrect, residual echo will remain.

[0005] A second disadvantage is that these prior art AEC systems typically use a linear model (i.e., linear filter) to predict the echo. However, the transfer function is often not linear (e.g., if there are non-linearities in the speaker which may cause distortion). As a result, poor echo prediction may occur.

[0006] Other AEC systems may attempt to overcome this disadvantage by introducing non-linearity to the echo prediction model. However, this results in more complexity. For example, non-linearity may become problematic in cellular phones or speakerphones with cheap components. The prediction may be difficult to obtain unless an exact model of the speaker (e.g., of the cellular phone or speakerphone) is known.

SUMMARY OF THE INVENTION

[0007] Embodiments of the present invention overcome or substantially alleviate prior problems associated with acoustic echo cancellation processing. Exemplary embodiments of the present invention utilize magnitude envelopes of acoustic signals to determine an echo envelope. An echo gain mask may then be generated based on the echo envelope.

[0008] In exemplary embodiments, a primary acoustic signal is received via a microphone of the communication device, and a far-end signal is received via a receiver. Because a speaker may provide audio (the far-end signal) that may be picked up by the microphone, the acoustic signal received by the microphone may include speaker leakage. As such, acoustic echo cancellation (AEC) is applied to the acoustic signal to obtain an AEC masked signal.

[0009] Frequency analysis is performed on the primary acoustic signal and the far-end acoustic signal to obtain frequency sub-bands. These frequency sub-bands may then be used to generate corresponding estimated energy spectra. The energy spectra are then used to determine echo-dominated and echo-free noise estimates.

[00010] An echo gain mask based on magnitude envelopes of the energy spectra of the primary and far-end acoustic signals for each frequency sub-band is generated. A noise gain mask based on at least the primary acoustic signal for each frequency sub-band may also be generated. The echo gain mask and the noise gain mask may then be combined. In one embodiment, the combination comprises selecting a minimum between the two gain masks. The combination of the echo gain mask and noise gain mask may then be applied to the primary acoustic signal to generate a masked signal, which may be output to a far-end environment.

BRIEF DESCRIPTION OF THE DRAWINGS

[00011] FIG. 1 is an environment in which embodiments of the present invention may be practiced.

[00012] FIG. 2 is a block diagram of an exemplary communication device implementing embodiments of the present invention.

[00013] FIG. 3 is a block diagram of an exemplary audio processing system.

[00014] FIG. 4 is a block diagram of an echo mask generator.

[00015] FIG. 5 is a block diagram of an alternative audio processing system.

[00016] FIG. 6 is a flowchart of an exemplary method for envelope-based acoustic echo cancellation.

[00017] FIG. 7 is a flowchart of an exemplary method for determining an echo gain mask.

DESCRIPTION OF EXEMPLARY EMBODIMENTS

[00018] The present invention provides exemplary systems and methods for providing envelope-based acoustic echo cancellation (EnvAEC). Exemplary embodiments perform the EnvAEC based on frequency sub-bands and prediction envelopes of echo waveforms, as opposed to details of actual echo waveforms. The envelopes, thus, show how energy in the waveforms may change over time.

[00019] Exemplary embodiments are configured to reduce and/or minimize effects of speaker signal leakage to one or more microphones in such a way that the far-end environment does not perceive an echo. While the following description will be discussed using a two microphone system, it should be noted that embodiments of the present invention may be applied to a single microphone envelope-based acoustic echo cancellation system.

[00020] Embodiments of the present invention may be practiced on any device that is configured to receive audio such as, but not limited to, cellular phones, phone handsets, headsets, and conferencing systems. While embodiments of the present invention will be described in reference to operation on a speakerphone, the present invention may be practiced on any audio device.

[00021] Referring to FIG. 1, environments in which embodiments of the present invention may be practiced is shown. A user in a near-end environment 100 acts as an acoustic source 102 to a communication device 104. In some embodiments, a mouth of the user comprises a point near a mouth of the acoustic source 102, while a speaker comprises a point near an ear of the acoustic source 102.

[00022] The exemplary communication device 104 comprises a microphone 106 (i.e., primary microphone), speaker 108, and an audio processing system 110 including an acoustic echo cancellation mechanism. The microphone 106 is configured to pick up audio from the acoustic source 102, but may also pick up noise from the near-end environment 100. The audio received from the acoustic source 102 will comprise a near-end microphone signal y(t), which will be sent back to a far-end environment 112.

[00023] In some embodiments, one or more additional microphones (not shown) may be present in the communication device 104. The one or more additional microphones may be located a distance away from the primary microphone 106. In some embodiments, the microphone(s) may comprise omni-directional microphones.

[00024] An acoustic signal x(t) comprising speech from the far-end environment 112 may be received via a communication network 114 by the communication device 104. The received acoustic signal x(t) may then be provided to the near-end environment 100 via the speaker 108. The audio output from the speaker 108 may leak back into (i.e., be picked up by) the microphone 106. This leakage may result in an echo perceived at the far-end environment 112.

[00025] The exemplary audio processing system 110 is configured to remove u(t) (i.e., echoes of x(t)) from y(t), while preserving a near-end voice signal v(t). In exemplary embodiments, the removal of u(t) is performed without introducing distortion to a far-end listener. This may be achieved by calculating and applying time and frequency varying multiplicative gains or masks that render the acoustic echo inaudible. Ideally, the gains are less than 1 (i.e., less than 0dB) to result in signal attenuation. In various embodiments, the attenuation is strong when echo dominates over other components of the signal.

[00026] Referring now to FIG. 2, the exemplary communication device 104 is shown in more detail. In exemplary embodiments, the communication device 104 is an audio receiving device that comprises a receiver 200, a processor 202, the primary microphone 106, an optional secondary microphone 204, the audio processing system 110, and an output device 206. The communication device 104 may comprise more or other components necessary for communication device 104 operations. Similarly, the communication device 104 may comprise fewer components that perform similar or equivalent functions to those depicted in FIG. 2.

[00027] The exemplary receiver 200 is an acoustic sensor configured to receive the far-end signal x(t) from the network 114. In some embodiments, the receiver 200

may comprise an antenna device. The received far-end signal x(t) may then be forwarded to the audio processing system 110 and the output device 206.

[00028] The audio processing engine 110 is also configured to receive the acoustic signals from the acoustic source 102 via the primary and optional secondary microphones 106 and 204 (e.g., primary and secondary acoustic sensors) and process the acoustic signals. The primary and secondary microphones 106 and 204 may be spaced a distance apart in order to allow for an energy level differences between them. After reception by the microphones 106 and 204, the acoustic signals may be converted into electric signals (i.e., a primary electric signal and a secondary electric signal). The electric signals may themselves be converted by an analog-to-digital converter (not shown) into digital signals for processing in accordance with some embodiments. In order to differentiate the acoustic signals, the acoustic signal received by the primary microphone 106 is herein referred to as the primary acoustic signal, while the acoustic signal received by the secondary microphone 204 is herein referred to as the secondary acoustic signal. It should be noted that embodiments of the present invention may be practiced utilizing any number of microphones. In exemplary embodiments, the acoustic signals from the secondary microphone 204 are used for total noise estimation as will be discussed further below.

[00029] The output device 206 is any device which provides an audio output to a listener (e.g., the acoustic source 102). For example, the output device 206 may comprise the speaker 108, an earpiece of a headset, or handset on the communication device 104.

[00030] In various embodiments, where the primary and secondary microphones are omni-directional microphones that are closely-spaced (e.g., 1-2 cm apart), a beamforming technique may be used to simulate a forwards-facing and a backwards-facing directional microphone response. A level difference may be obtained using the simulated forwards-facing and the backwards-facing directional microphone. The level difference may be used to discriminate speech and noise in the time-frequency domain which can be used in noise estimation.

[00031] FIG. 3 is a detailed block diagram of the exemplary audio processing system 110, according to exemplary embodiments of the present invention. In exemplary embodiments, the audio processing engine 110 may be embodied within a memory device. The exemplary audio processing system 110 provides acoustic echo cancellation (AEC) and noise suppression. As a result, an acoustic signal sent from the communication device 104 to the far-end environment 112 comprises noise suppression as well as reduced or eliminated echo from speaker leakage. In accordance with one embodiment, the audio processing system 110 may comprise a frequency analysis module 302, an energy module 304, a noise estimate module 306, an echo mask generator 308, a total noise estimate module 310, a noise mask generator 312, an echo and noise mask integration module 314, and a modifier/reconstructor module 316.

[00032] In operation, the acoustic signals received from the primary and secondary microphones 106 and 204 and the far-end acoustic signal x(t) are converted to electric signals and processed through the frequency analysis module 302. The frequency analysis module 302 takes the acoustic signals and mimics the frequency analysis of the cochlea (i.e., cochlear domain) simulated by a filter bank. In one embodiment, the frequency analysis module 302 separates the acoustic signals into frequency bands or sub-bands. Alternatively, other filters such as short-time Fourier transform (STFT), Fast Fourier Transform, Fast Cochlea transform, sub-band filter banks, modulated complex lapped transforms, cochlear models, a gamma-tone filter bank, wavelets, or any generalized spectral analysis filter/method, can be used for the frequency analysis and synthesis.

[00033] Because most sounds (e.g., acoustic signals) are complex and comprise more than one frequency, a sub-band analysis on the acoustic signal may be performed to determine what individual frequencies are present in the acoustic signal during a frame (e.g., a predetermined period of time). According to one embodiment, the frame is 5 - 20 ms long (e.g., 40 to 160 samples for a system audio sampling rate of 8000 Hz). Alternative embodiments may utilize other frame

lengths. Data may be pushed through the audio processing system 110 in these frames (i.e., blocks of buffered samples).

[00034] The output of the frequency analysis module 302 comprises a plurality of waveforms. Thus, if the acoustic signal comprises high frequency bands, the resulting waveforms are more energetic. As will be discussed further below, the envelopes of these waveforms in these frequency bands are analyzed for echo suppression. Specifically, the envelopes of the far-end acoustic signal are used to predict the echo envelopes that will be present in the near-end acoustic signal (e.g., primary acoustic signal).

[00035] Once the frequencies are determined, the signals are forwarded to the energy module 304 which computes energy/power estimates for the primary, secondary, and far-end acoustic signals during an interval of time for each frequency sub-band (i.e., power estimates). As such an average power output for each frequency sub-band (i.e., power spectrum) may be calculated for each of the acoustic signals in frames. In exemplary embodiments, the frames comprise 5ms time periods. Thus, buffers are filled up with 5ms of frequency analysis module 302 output. An average power per frame may then be determined.

[00036] The exemplary energy module 304 is a component which, in some embodiments, can be represented mathematically by the following equation:

$$E(t,\omega) = \lambda_E \left|X(t,\omega)\right|^2 + (1-\lambda_E)E(t-1,\omega)$$

where λ_E is a number between zero and one that determines an averaging time constant, $X(t,\omega)$ is the acoustic signal being processed (e.g., the primary, secondary, or far-end acoustic signal) in the cochlea domain, ω represents the frequency, and t represents time. Given a desired time constant T (e.g., 4 ms) and sampling frequency f_s (e.g., 16 kHz), the value of λ_E can be approximated as

$$\lambda_E = 1 - e^{-\frac{1}{Tf_s}}$$

As provided, the energy level for the acoustic signal, $E(t,\omega)$, is dependent upon a previous energy level of the acoustic signal, $E(t\text{-}1,\omega)$.

[00037] The exemplary noise estimate module 306 is configured to determine a noise estimate based on the primary acoustic signal. In some embodiments, this noise estimate module 306 may produce a stationary noise estimate based on constant ambient noise in the near-end environment 100. This stationary noise estimate may be later augmented by non-stationary noise components as a function of both the primary microphone 106 and the optional secondary microphone 204.

[00038] In one embodiment, the noise estimate module 306 comprises a minimum statistics tracker (MST) which receives the energy of the primary acoustic signal from the signal path for processing. The determination of the noise estimate, according to one embodiment, is discussed in more detail in connection with U.S. Patent Application No. 12/072,931 filed February 29, 2008 and entitled "System and Method for Providing Single Microphone Noise Suppression Fallback," which is incorporated by reference. The noise estimate is then provided to the echo mask generator 308 and the total noise estimate module 310.

[00039] The exemplary echo mask generator 308 is configured to generate an echo gain mask that will render echo inaudible. The echo gain mask is generated based on predicted envelopes (of the echo waveforms) based on the far-end acoustic signal. By analyzing the near-end signal (e.g., primary acoustic signal) and the predicted envelopes, a determination may be made as to where and when echo may be audible in frequency sub-bands. Echo suppression may then be applied to these frequency sub-bands. The output of the echo mask generator 308 comprises a gain value per frequency sub-band and frame. The echo mask generator 308 will be discussed in more detail in connection with FIG. 4.

[00040] The result of the echo mask generator 308 along with the noise estimate and the optional secondary acoustic signal energy are forwarded to the total noise estimate module 310. In some embodiments, the total noise estimate module 310 may comprise an echo-free noise estimate module. The total noise estimate module 310 is configured to compute an estimate of the near-end noise power spectra (e.g., time and frequency dependent portion of the acoustic signal that is not from the

acoustic source 102). In exemplary embodiments, the total noise estimate module 310 refines the noise estimate received from the noise estimate module 306, which may be corrupted by echo power.

[00041] The results from the total noise estimate module 310 may then be used by the noise mask generator 312 to determine a noise suppression gain mask. Various embodiments of the exemplary total noise estimate module 310 and the exemplary noise mask generator 312 are further discussed in U.S. Patent Application No. 12/004,899 filed December 21, 2007 and entitled "System and Method for 2-Channel and 3-Channel Acoustic Echo Cancellation," and U.S Patent Application No. 12/004,896 filed December 21, 2007 and entitled "System and Method for Blind Subband Acoustic Echo Cancellation Postfiltering" which are both hereby incorporated by reference.

[00042] In some embodiments, the echo gain mask may be refined in each sub-band to reduce near-end speech distortion. In exemplary embodiments, the echo and noise mask integration module 314 takes into account the near-end noise level (from the echo-free noise estimate determined by the total noise estimate module 310) and the noise suppression gain mask (from the noise mask generator 312). In various embodiments, the echo gain mask may be limited such that a total output power is not more than a certain amount (e.g., 6dB) below the output noise power that will be produced by applying the noise suppression gain mask to the near-end noise. This process may reduce perception of output noise modulation correlated with the echo, while still ensuring the echo remains inaudible.

[00043] The echo gain mask is combined with the noise gain mask by the echo and noise mask integration module 314. In one embodiment, the echo and noise mask integration module 314 may select a minimum between the two gain masks (i.e., noise gain mask and echo gain mask) for each sub-band, with a lower limit on how far below the noise gain mask the combined mask can be. A possible mathematical description of this combination operation is:

$$g_{tot} = \min\left(g_N, \max\left(g_E, g_N \cdot \sqrt{Pn/Py} \cdot \gamma\right)\right),$$

where g_{tot} is a final gain mask, min(x,y) is a minimum of x and y, max(x,y) is a maximum of x and y, g_N is the noise gain mask, g_E is the echo gain mask, Py is the total power in the frame, Pn is the estimated noise power in the frame, and γ is maximum tolerated modulation on the output noise. For example, if the amount of tolerated modulation is -6dB (i.e., 6dB down), γ will equal $10^{-6/20}$ or around 0.5.

[00044] The noise gain mask is produced from the total noise estimate and may be defined such that the noise power is reduced at the system output but not rendered completely inaudible. As such, the echo and noise mask integration module 314 may be used to negotiate different objectives of the noise suppression and echo cancellation masks. Furthermore, it should be noted that the final gain mask may be defined such that no noise reduction, or at least multiplicative noise masking, is performed in the system. This is equivalent to taking the noise mask generator 312 out of the audio processing system 110. The result is a final gain mask that may be used for modification of the primary acoustic signal from the primary microphone 106. Accordingly in exemplary embodiments, gain masks may be applied to an associated frequency band of the primary acoustic signal in the modifier/reconstructor module 316.

[00045] Next, the post-AEC and noise suppressed frequency bands are converted back into time domain from the cochlea domain. The conversion may comprise taking the post-AEC and noise suppressed frequency bands and adding together phase shifted signals of the cochlea channels in the modifier/reconstructor module 316. In one embodiment, the reconstruction portion of the modifier/reconstructor module 316 comprises a frequency synthesis module. Once conversion is completed, the synthesized, masked acoustic signal may be output (e.g., forwarded to the communication network 114 and sent to the far-end environment 112).

[00046] In the embodiment of FIG. 3, the stationary and non-stationary noise estimates are calculated separately. This architecture is advantageous in situations where cues being used to estimate the non-stationary noise components are not able to discriminate echo from ambient noise. In these situations, the input into the echo mask generator 308 may cause an adaptation control module to incorrectly freeze adaptation, as will be described further in connection with FIG. 4. Thus, it may be desirable to utilize noise estimation components, such as stationary noise, that are not likely to be strongly corrupted by the echo in the echo mask generator 308. Subsequently, since the echo mask generator 308 output indicates frames and frequencies where echo dominates, the output may be taken into account for a "total noise estimate" operation. One example of this is to freeze update of the noise estimate when the echo mask indicates echo is dominating the signal.

[00047] It should be noted that the system architecture of the audio processing system 110 of FIG. 3 is exemplary. Alternative embodiments may comprise more components, fewer components, or equivalent components and still be within the scope of embodiments of the present invention.

[00048] Referring now to FIG. 4, a block diagram of the echo mask generator 308 is shown. The echo mask generator 308 is configured to generate an echo gain mask which will attenuate a far-end echo (ideally to render the far-end echo inaudible) while minimizing disturbance of near-end sources. In exemplary embodiments, the echo mask generator 308 receives as inputs per frame a power spectra of the near-end signal (e.g., primary acoustic signal), a power spectra of the far-end signal, and, optionally, a noise power spectrum estimate from the noise estimate module 306.

[00049] The inputs are then subjected to a square-root (sqrt) operation in the sqrt module 402. The sqrt operation transforms the power spectrum into magnitudes. For a given frequency band, a sequence of such magnitudes over time comprises an envelope, or one amplitude per frame.

[00050] In exemplary embodiments, a prediction of an echo in a near-end envelope is used to determine whether the near-end envelope is dominated by echo to a point that the echo is audible. The prediction may be based on the far-end signal (and thus the far-end envelope). For example, if there is no energy in the far-end envelope, there is likely no echo and thus no echo suppression. However, if there is energy both in the far-end and the near-end envelopes, an echo within the near-end envelopes is more likely. In these examples, echo suppression will be applied. The echo mask generator 308 comprises logic which analyzes the prediction versus an observation in order to determine where and when to apply echo suppression.

[00051] In some embodiments, the amount of echo suppression may be limited by perceptual effects. For example, if a loud sound is followed by a soft sound in quick succession, forward masking may occur whereby the soft sound is not really perceived. In this situation, suppression may not be necessary.

[00052] The determination of whether the near-end envelope is dominated by echo may occur in a mask generator 404. If the echo is audible, the mask generator 404 will have a low gain value (e.g., 0). Otherwise, the mask generator 404 may have an "all pass" gain value (e.g., 1). A mathematical representation of this determination is

$$m = H\left[\frac{y^2 - \hat{u}^2}{\hat{u}^2} - \tau_1\right], \qquad (1)$$

where m is the mask generated by the mask generator 404, y is the near-end envelope, $\hat{u}$ is the echo prediction, τ_1 is a dominance threshold for a ratio of non-echo power to echo power, and H is a Heavyside step function (which is 1 if its input is greater than 1 and 0 otherwise). A reasonable value for τ_1, for example, is 1 (0 dB). The lower τ_1 is, the higher the echo power has to be relative to the total power before it is attenuated.

[00053] An echo predictor 406 is configured to use the far-end envelope (per frequency band) to predict its echo in the near-end envelope. In one embodiment,

the echo predictor 406 may be simplified by assuming that the echo envelope is just a delayed and scaled version of the far-end envelope. Mathematically, this assumption may be represented by

$$\hat{u}(n) = g \cdot x(n-d), \quad (2)$$

where g is a scale factor, n is a frame index, and d is a number of frames delayed. In some embodiments, this representation may provide an accurate estimate due to low temporal resolution of the envelope (which is bolstered by smoothing). This allows the prediction to be robust to changes in the delay and echo tail length. Often, the number of frames delayed d can be measured and stored as a priori information. In other embodiments, d may be estimated during operation (e.g., by finding a peak of a cross-correlation between the near-end and far-end signals).

[00054] It may also be possible to extend the prediction to a multiple reflection model, such as

$$\hat{u}(n) = \sum_{d} g(d) \cdot x(n-d), \quad (3)$$

where the echo is predicted as a summation of far-end echoes at various delays and scale factors. However, in most embodiments, such an extreme model may not be necessary.

[00055] Smoothers 408 and 410 may smooth the magnitudes, and thus envelopes, of the far-end and near-end signals. The temporal smoothing removes "wiggles" in the envelope that may not be important for echo suppression. These "wiggles" if not removed, may result in glitches in the echo prediction. As a result, low temporal resolution can be further enforced by applying the temporal smoothing to the envelopes. An example of a simple form of smoothing is a leaky integrator, mathematically represented by,

$$x(n) = \alpha \cdot x(n-1) + (1-\alpha) \cdot X(n), \quad (4)$$

where α is a smoothing factor and x is an unsmoothed envelope. In this embodiment, more smoothing is performed at higher α. As a result, the system may be less susceptible to delay and echo tail changes.

[00056] Smoothing may also be beneficial for filtering out hard-to predict non-linear interactions (e.g., "cross-terms") between the echo and other near-end signal components inconsequential to determining echo dominance. However, if too much smoothing is performed, the system may not react quickly enough for echo dynamics. An example of a reasonable value for α is 0.85 for a frame rate of 200 Hz, which corresponds to a time constant of around $-200/\ln(0.85) = 30$ ms. For this value, the system is able to accurately predict the echo envelope, even if the echo tail is up to 600 ms long.

[00057] Exemplary embodiments are robust to changes in echo delay and tail length. As such, the accuracy of the echo prediction will be largely determined by the accuracy of the gain g. In most embodiments, this gain g may not be known *a priori* with enough precision, and so it must be estimated during operation. Further complicating matters, g can significantly change at any time due to echo path or speaker volume changes.

[00058] Thus, embodiments of the present invention adaptively estimate g using a gain updater 412. In some embodiments, g is updated quickly when echo is dominating the near-end signal, and updated slowly or not at all otherwise. In exemplary embodiments, an adaptation control module 414 provides the information necessary to determine how quickly to update g. In some embodiments, the information may come in a form of a frame- and frequency-varying parameter μ. The larger μ is, the faster g is updated.

[00059] Given μ, an efficient way of updating g is given by a normalized least-mean-square (NLMS) algorithm, such as,

$$g_d(n+1) = g_d(n) + \frac{\mu \cdot x_d \cdot (y - \hat{u})}{x_d^{\,2} + \Delta}, \qquad (5)$$

where $x_d = x(n-d)$. d subscripts denote that the expression is valid for an a priori known delay d.

[00060] If a vector of delays is being considered in conjunction with the prediction model (3), there may be well known and simple extensions to Eq. (5) that

apply. Additionally, Δ is a regularization parameter that slows the update of g when the far-end signal is small. Eq. (5) may be re-written, using (2), as

$$g_d(n+1) = (1-\mu\beta)\cdot g_d(n) + \mu\beta\cdot\frac{y}{x}, \qquad (6)$$

where $\beta = \frac{x^2}{x^2+\Delta}$ is between 0 and 1. Eq. (6) gives an intuitive explanation that the gain estimate is an adaptively smoothed version of a ratio of near-end to far-end envelopes. The output of the gain updater 412 may be fed back to the echo predictor 406 in the next frame. It should be noted that if any upper or lower limits on g are known a priori, these limits may be applied before passing the result to the echo predictor 406.

[00061] The exemplary adaptation control module 414 is configured to notify the gain updater 412 when and how fast to safely adapt the gain g via the parameter μ (which is between 0 and 1 with 1 being the fastest). In exemplary embodiments, a determination of when the echo sufficiently dominates the near-end signal (e.g., when the ratio y/x in Eq. (6) is a meaningful observation of g) is performed. The output of the adaptation control module 414 encapsulates an adaptation speed.

[00062] In accordance with some embodiments, multiple cues may be utilized to help strike a balance between updating too quickly (e.g., the system reacts erroneously to near-end sources) and updating too slowly (e.g., the system doesn't react quickly enough to echo path changes). The cues may comprise, for example, adaptive far-end noise floor, near-end noise floor, and maximum gain. While these examples, as will be discussed below, provide outputs of either 0 or 1, alternative cues can produce values between 0 and 1.

[00063] The adaptive far-end floor cue is an adaptive version of the Δ in Eq. (5) (i.e., an adaptive threshold on the far-end energy). As a result, only the most powerful times and frequencies of x enable fast adaptation, since the echo is most likely to dominate where x is its most powerful. One method of insuring this result is by performing a long-term average/smoothing of past values of x, and then

determining whether or not the current value of x is far enough above the long term average x_{LT}.

[00064] Besides smoothing (e.g., Eq. (4)), an alternative method to produce x_{LT} is to convert x to decibel units and apply a linear slew rate. This may be mathematically represented by,

$$x_{LT}(n+1) = \begin{cases} x_{LT}(n) + \gamma_{up}; & dB[x(n)] > x_{LT}(n) \\ x_{LT}(n) - \gamma_{down}; & dB[x(n)] \le x_{LT}(n) \end{cases},$$

where dB[x] denotes a conversion of x to dB units. This method may result in smoother tracking behavior. Furthermore, independent control of upward and downward slew rates via γ_{up} and γ_{down} allows a user to control at what percentile in a range of x the long-term average settles (e.g., it is higher for higher ratios of γ_{up} to γ_{down}).

[00065] The output of this adaptive far-end floor cue can be represented as

$$Q1 = H[dB[x] > x_{LT} + \tau_2], \quad (8)$$

where higher values of τ_2 are more selective to higher values of dB[x].

[00066] The near-end noise floor cue utilizes a noise power spectrum input, N, such that adaptation is prevented if the near-end envelope is not high enough above the noise envelope. In these embodiments, echo cannot dominate the signal if the near-end envelope is not high enough above the noise envelope. Therefore, mathematically,

$$Q2 = H[y > N \cdot \tau_3], \quad (9)$$

where higher values of τ_3 require higher minimum dominance over the ambient noise.

[00067] In exemplary embodiments, the maximum gain cue may be useful if there is an upper limit $g_{\max}$ expected on the ratio y/x (the gain observation) if y is dominated by echo, while that maximum is often exceeded when y is dominated by near-end sources. In such an embodiment, a condition may be mathematically set such that

$$Q3 = H[y < g_{\max} \cdot x]. \quad (10)$$

[00068] Alternative cues may be devised to increase robustness of the system. The final output of the adaptation control module 414 may be obtained by combining all of the cues and applying an absolute maximum on the update speed, according to exemplary embodiments. This may be mathematically represented by,

$$\mu = Q1 \cdot Q2 \cdot Q3 \cdot \mu_{\max}, \quad (11)$$

[00069] In exemplary embodiments, the combination of all the cues being true (e.g., value = 1) will result in adaptation. In contrast, if one or more the cues is false (e.g., value = 0), adaptation will not occur.

[00070] An alternative embodiment of the audio processing system is shown in FIG. 5. In this embodiment, the total noise estimate module 310 is removed from the embodiment of FIG. 3. This embodiment may be utilized when the total noise estimate used in the system is not likely to be corrupted by echo. In one embodiment, the noise estimate may not be corrupted by echo when the system is designed to perform stationary noise reduction only. In this embodiment, the noise estimate from the noise estimate module 306 is the stationary noise estimate. In another embodiment, the noise estimate is not corrupted by echo when the non-stationary noise estimate is not corrupted by echo in a non-stationary noise reduction system. In these embodiments, there is no need for the echo mask generator 308 to influence the noise estimation process or for the second stage of the noise estimation (i.e., use of the total noise estimate module 310). Thus, the noise estimate module 306 may, in some embodiments, utilize the energy estimates for the primary and/or secondary acoustic signals to determine the noise estimate that will be used by the echo mask generator 308. The noise estimate is then provided to the echo mask generator 308 along with the energy from the far end signal. It should be noted that the embodiment of FIG. 5 may be applicable in situations where only the stationary noise component from the noise estimate module 306 is utilized or in

situations where a noise estimate including both stationary and non-stationary noise components are utilized.

[00071] Referring now to FIG. 6, a flowchart 600 of an exemplary method for envelope-based acoustic echo cancellation is shown. In step 602, the acoustic signals are received by the communication device 104. In exemplary embodiments, a near-end acoustic signal is received by the primary microphones 106. A far-end acoustic signal may also be received via a receiver 200. If the far-end acoustic signal is being output through the speaker 108, then audio from the speaker 108 may leak back to the primary microphones 106. This may result in an echo being provided back to a listener at the far-end environment 112. In some embodiments, an optional secondary microphone 204 may also receive a secondary acoustic signal.

[00072] The acoustic signals are then converted to electric signals and processed through the frequency analysis module 302 to obtain frequency sub-bands in step 604. In one embodiment, the frequency analysis module 302 takes the acoustic signals and mimics the frequency analysis of a cochlea (i.e., cochlear domain) simulated by a filter bank. The result comprises frequency sub-bands.

[00073] In step 606, energy estimates for the acoustic signals are computed. In one embodiment, the energy estimates are determined by the energy module 304. The exemplary energy module 304 utilizes a present acoustic signal and a previously calculated energy estimate to determine the present energy estimate for each acoustic signal at each frequency sub-band.

[00074] Subsequently, the noise estimate is determined in step 608. According to embodiments of the present invention, the noise estimate for each frequency sub-band is based on the acoustic signal received at the primary microphone 106. The noise estimate may then be provided to the echo mask generator 308 and the total noise estimate module 310.

[00075] The echo mask is then generated in step 610. Step 610 will be discussed in more detail in connection with FIG. 7 below.

[00076] A noise suppression gain mask is generated in step 612. In exemplary embodiments a total noise estimate that may be echo free is determined by the total noise estimate module 310. In exemplary embodiments, the total noise estimate is determined for each frequency sub-band. In exemplary embodiments, the total noise estimate module 310 is configured to compute an estimate of the near-end noise power spectrum (e.g., time and frequency dependent portion of the acoustic signal that is not from the acoustic source 102). In some embodiments, the total noise estimate module 310 may refine the noise estimate received from the noise estimate module 306, which may be corrupted by echo power. The noise suppression gain mask may then be generated using the total noise estimate by the noise mask generator 312.

[00077] In step 614, a combined echo/noise suppression mask is generated. In exemplary embodiments, the combined echo and noise mask integration module 314 generates the combined echo/noise suppression mask. In one embodiment, the echo and noise mask integration module 314 may select a minimum between the two gain masks (i.e., noise gain mask and echo gain mask) for each sub-band, with a lower limit on how far below the noise gain mask the combined mask can be.

[00078] The result of the echo and noise mask integration module 314 is a final gain mask that may be used for modification of the primary acoustic signal from the primary microphone 106 in step 616. The modified signal may then be reconstructed and output.

[00079] Referring now to FIG. 7, an exemplary flowchart of a method for generating the echo gain mask (step 610) is shown. In step 702, envelopes of the echo-free noise, far-end signal, and near-end signal (i.e., primary acoustic signal) are determined. In exemplary embodiments, the echo mask generator 308 receives the noise estimate, far-end signal energy spectra, and near-end signal energy spectra. These estimates and spectra are subjected to a square-root (sqrt) operation in the sqrt module 402. The sqrt operation transforms the power spectrum into magnitudes.

For a given frequency band, a sequence of such magnitudes over time comprises an envelope (e.g., one amplitude per frame).

[00080] The far-end and near-end envelopes are then smoothed in step 704. The temporal smoothing removes "wiggles" in the envelope that may not be important for echo suppression. In one embodiment, the smoothing may be performed using a leaky integrator.

[00081] In step 706, the echo envelope is predicted. The echo predictor 406 is configured to use the far-end envelope (per frequency band) to predict its echo in the near-end envelope. In one embodiment, the echo predictor 406 may be simplified by assuming that the echo envelope is just a delayed and scaled version of the far-end envelope. In some embodiments, this representation may provide an accurate estimate due to low temporal resolution of the envelope (which is bolstered by smoothing).

[00082] The gain mask may be generated or updated in step 708. The gain mask may be generated by looking at a ratio between the echo prediction and the echo prediction power from the total near-end power.

[00083] Embodiments of the present invention adaptively estimate gain using the gain updater 412. In some embodiments, gain is updated quickly when echo is dominating the near-end signal, and updated slowly or not at all otherwise. In exemplary embodiments, the adaptation control module 414 provides the information necessary to determine how quickly to update gain. In some embodiments, the information may come in a form of a frame- and frequency-varying parameter μ. The larger μ is, the faster g is updated.

[00084] The above-described modules can be comprised of instructions that are stored on storage media. The instructions can be retrieved and executed by the processor 202. Some examples of instructions include software, program code, and firmware. Some examples of storage media comprise memory devices and integrated circuits. The instructions are operational when executed by the processor 202 to direct the processor 202 to operate in accordance with embodiments of the

present invention. Those skilled in the art are familiar with instructions, processor(s), and storage media.

[00085] The present invention is described above with reference to exemplary embodiments. It will be apparent to those skilled in the art that various modifications may be made and other embodiments can be used without departing from the broader scope of the present invention. For example, embodiments of the present invention may be applied to any system (e.g., non speech enhancement system) utilizing AEC. Therefore, these and other variations upon the exemplary embodiments are intended to be covered by the present invention.

CLAIMS

1. A method for envelope-based acoustic echo cancellation, comprising:

receiving a primary acoustic signal via a primary microphone and a far-end signal via a receiver;

performing frequency analysis on the primary acoustic signal and the far-end acoustic signal to obtain frequency sub-bands for the primary and far-end acoustic signals;

applying the echo gain mask to the primary acoustic signal to generate a masked signal; and

outputting the masked signal.

2. The method of claim 1 further comprising generating a noise gain mask based on at least the primary acoustic signal for each frequency sub-band.

3. The method of claim 2 wherein applying the echo gain mask to the primary acoustic signal to generate the masked signal comprises applying a combination of the echo gain mask and the noise gain mask to the primary acoustic signal to generate the mask signal.

4. The method of claim 3 wherein applying the combination of the echo gain mask and noise gain mask comprises selecting a minimum between the echo gain mask and noise gain mask for each sub-band.

5. The method of claim 3 wherein generating the noise gain mask comprises determining a substantially echo-free noise estimate.

6. The method of claim 1 further comprising receiving a secondary acoustic signal via a secondary microphone.

7. The method of claim 6 wherein generating the noise gain mask comprises factoring in the secondary acoustic signal.

8. The method of claim 1 wherein generating the echo gain mask comprises taking a square root of the far-end acoustic signal and the primary acoustic signal to form the magnitude envelopes of the far-end and primary acoustic signals.

9. The method of claim 8 further comprising smoothing the magnitude envelopes of the far-end and primary acoustic signals.

10. The method of claim 1 wherein generating the echo gain mask comprises determining an echo dominated noise estimate and using a square root of the echo dominated noise estimate to generate a noise magnitude envelope.

11. The method of claim 1 wherein generating the echo gain mask comprises predicting an echo envelope based on the magnitude envelopes of the far-end and primary acoustic signals.

12. The method of claim 1 wherein generating the echo gain mask comprises updating the gain based on one or more adaptation control cues.

13. The method of claim 12 wherein the adaptation control cue comprises an adaptive far-end noise floor.

14. The method of claim 12 wherein the adaptation control cue comprises a noise floor.

15. The method of claim 12 wherein the adaptation control cue comprises a maximum gain.

16. The method of claim 1 further comprising synthesizing the masked signal into time domain for outputting

17. A system for envelope-based acoustic echo cancellation in a communication device, comprising:

acoustic sensors configured to receive a primary acoustic signal and a far-end acoustic signal;

a frequency analysis module configured to performing frequency analysis on the primary acoustic signal and the far-end acoustic signal to obtain frequency sub-bands for the primary and far-end acoustic signals;

an echo mask generator configured to generate an echo gain mask based on magnitude envelopes of the primary and far-end acoustic signals for each frequency sub-band;

a modifier/reconstructor module configured to apply the echo gain mask to the primary acoustic signal to generate a masked signal; and

an output device configured to output the masked signal.

18. The system of claim 17 further comprising a noise mask generator configured to generate a noise gain mask based on at least the primary acoustic signal for each frequency sub-band.

19. The system of claim 18 wherein the modifier/reconstructor module configured to apply the echo gain mask to the primary signal to generate the masked signal comprises the modifier/reconstructor module configured to apply a combination of the echo gain mask and the noise gain mask to the primary acoustic signal to generate the masked signal.

20. The system of claim 19 further comprising an echo and noise mask integration module configured to generate the combination of the echo gain mask and noise gain mask

21. The system of claim 17 wherein the modifier/reconstructor module is further configured to convert the masked signal into time domain for outputting.

22. The system of claim 17 wherein the echo mask generator comprises an echo predictor configured to predict echo envelopes.

23. A machine readable medium having embodied thereon a program, the program providing instructions for a method for envelope-based acoustic echo cancellation in a communication device, comprising:

receiving a primary acoustic signal via a primary microphone and a far-end signal via a receiver;

performing frequency analysis on the primary acoustic signal and the far-end acoustic signal to obtain frequency sub-bands for the primary and far-end acoustic signals;

generating an echo gain mask based on magnitude envelopes of the primary and far-end acoustic signals for each frequency sub-band;

generating a noise gain mask based on at least the primary acoustic signal for each frequency sub-band;

applying a combination of the echo gain mask and noise gain mask to the primary acoustic signal to generate a masked signal; and

outputting the masked signal.

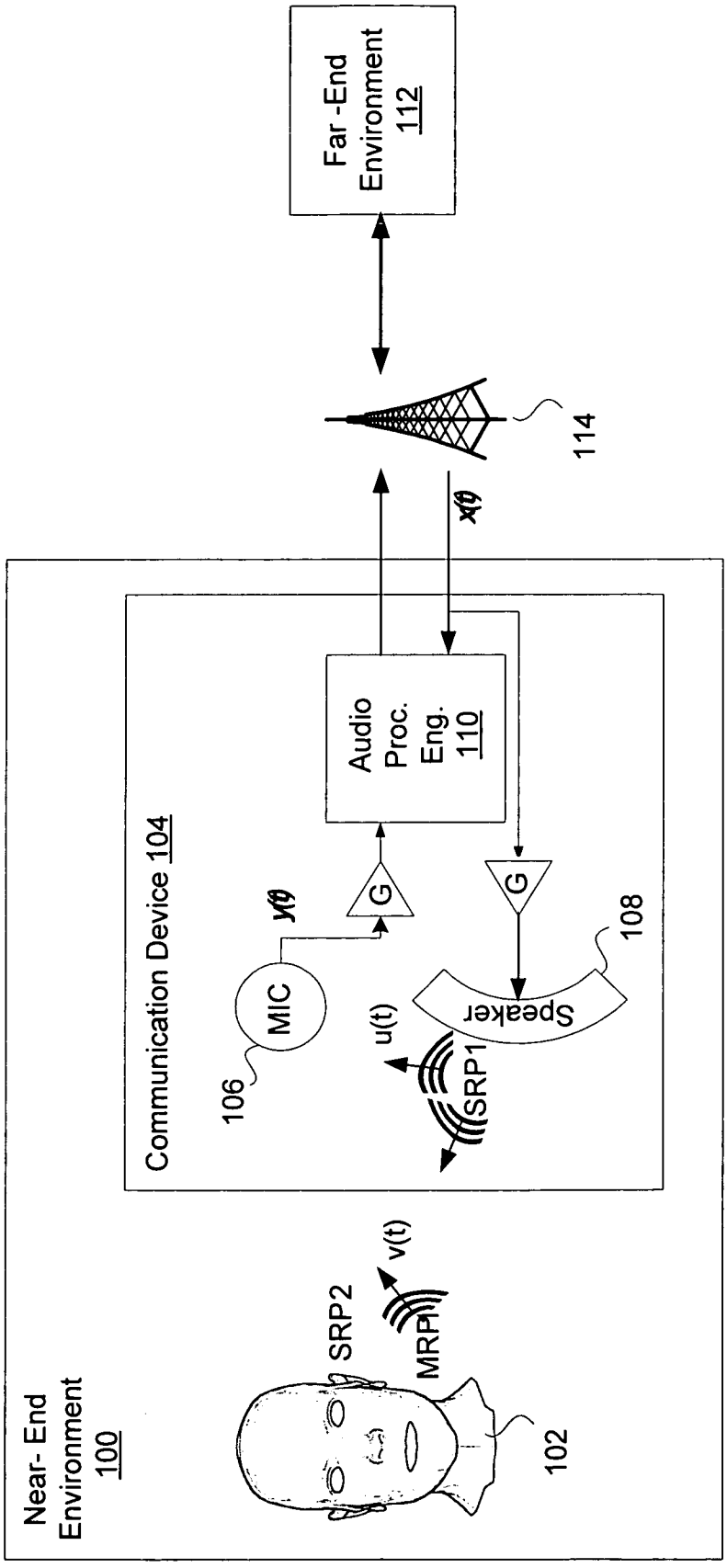

Near- End Environment 100
Communication Device 104
Audio Proc. Eng. 110
MIC
106
G
G
Speaker
108
u(t)
SRP1
SRP2
MRP
v(t)
102
114
Far -End Environment 112

FIG. 1

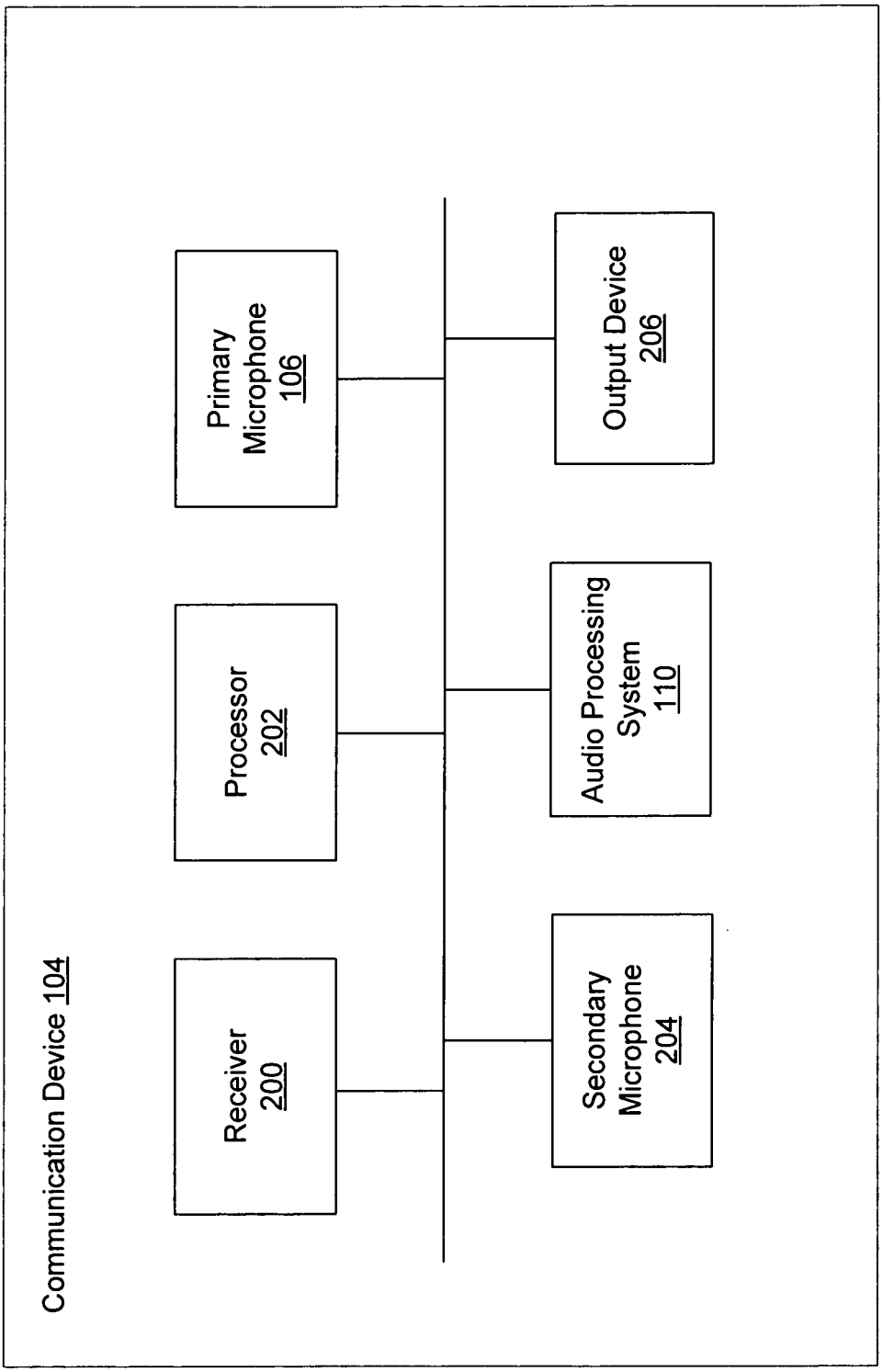

Communication Device 104
Receiver 200
Processor 202
Primary Microphone 106
Secondary Microphone 204
Audio Processing System 110
Output Device 206

FIG. 2

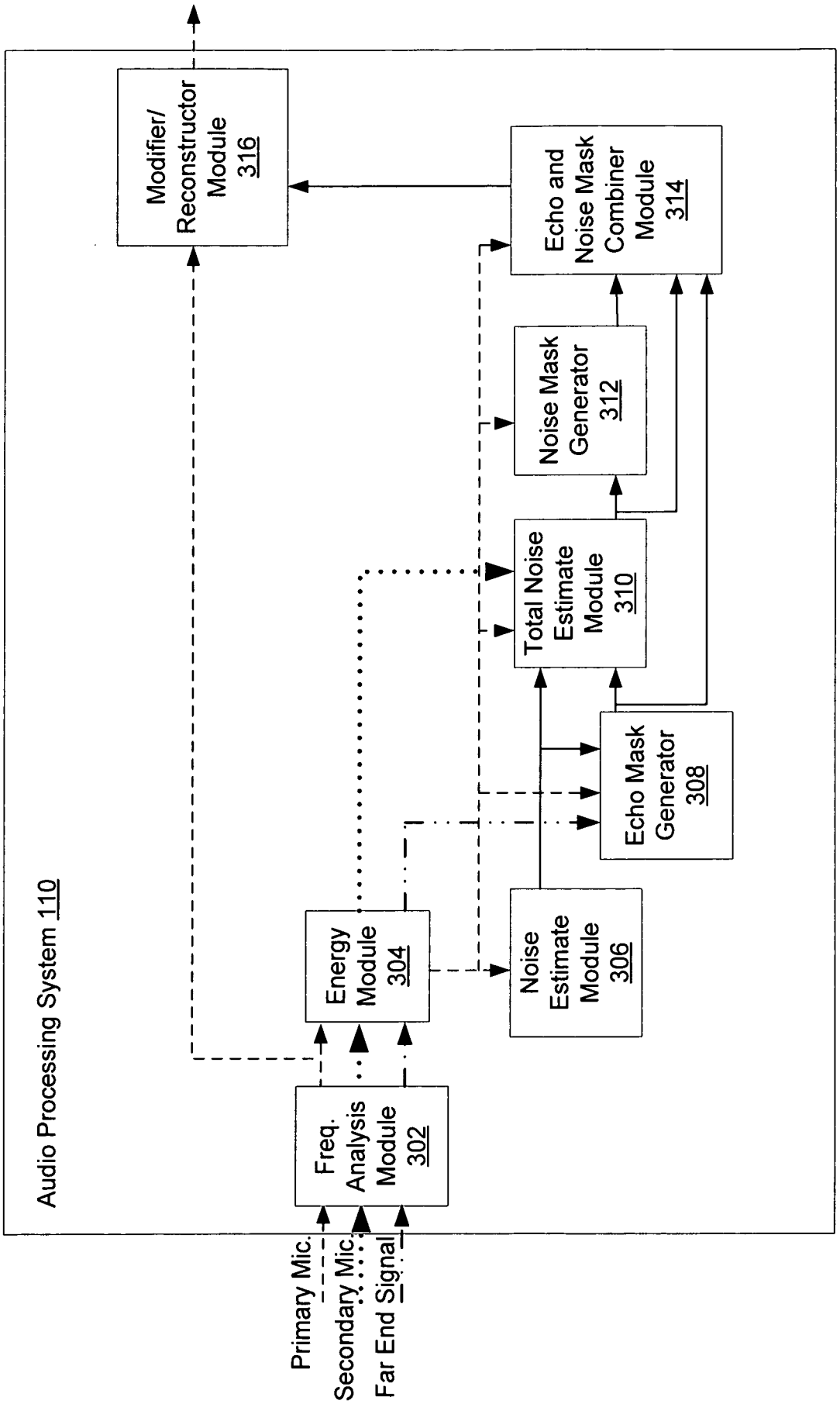

Audio Processing System 110
Modifier/ Reconstructor Module 316
Echo and Noise Mask Combiner Module 314
Noise Mask Generator 312
Total Noise Estimate Module 310
Echo Mask Generator 308
Noise Estimate Module 306
Energy Module 304
Freq. Analysis Module 302
Primary Mic.
Secondary Mic.
Far End Signal

FIG. 3

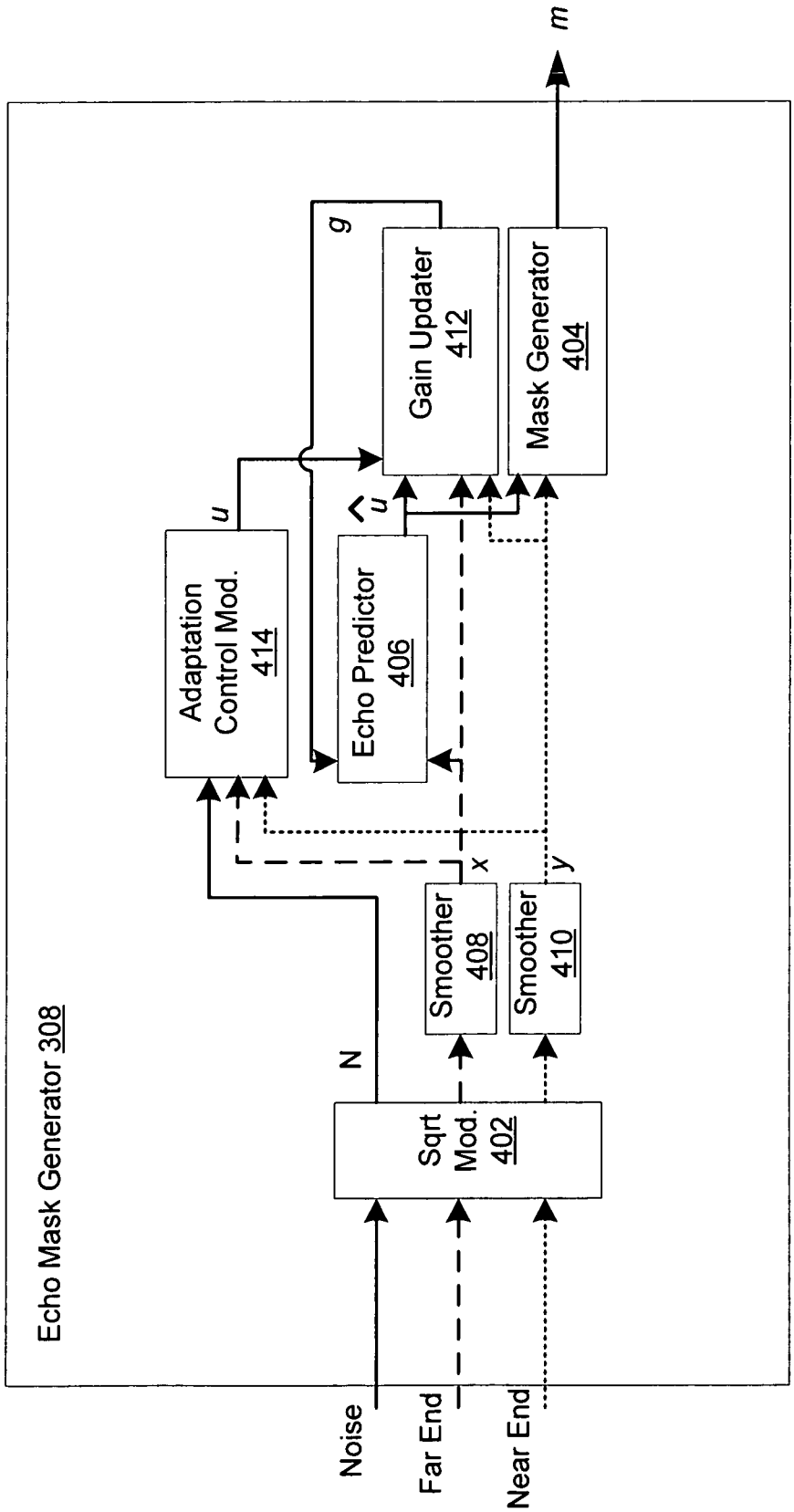

Echo Mask Generator 308
Noise
Far End
Near End
Sqrt Mod. 402
N
Smoother 408
Smoother 410
x
y
Adaptation Control Mod. 414
Echo Predictor 406
u
û
g
Gain Updater 412
Mask Generator 404
m

FIG. 4

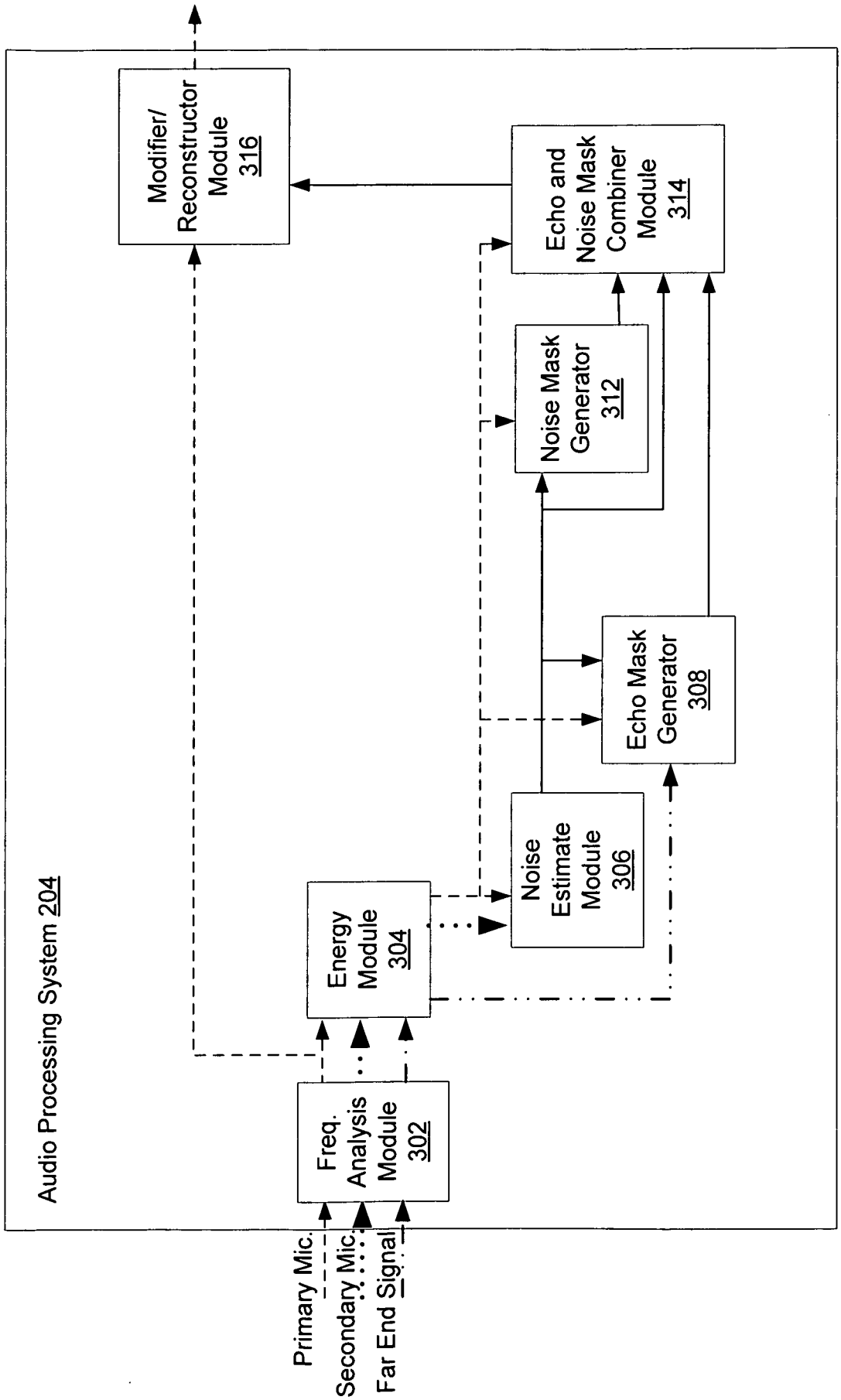

Audio Processing System 204
Modifier/ Reconstructor Module 316
Echo and Noise Mask Combiner Module 314
Noise Mask Generator 312
Echo Mask Generator 308
Noise Estimate Module 306
Energy Module 304
Freq. Analysis Module 302
Primary Mic.
Secondary Mic.
Far End Signal

FIG. 5

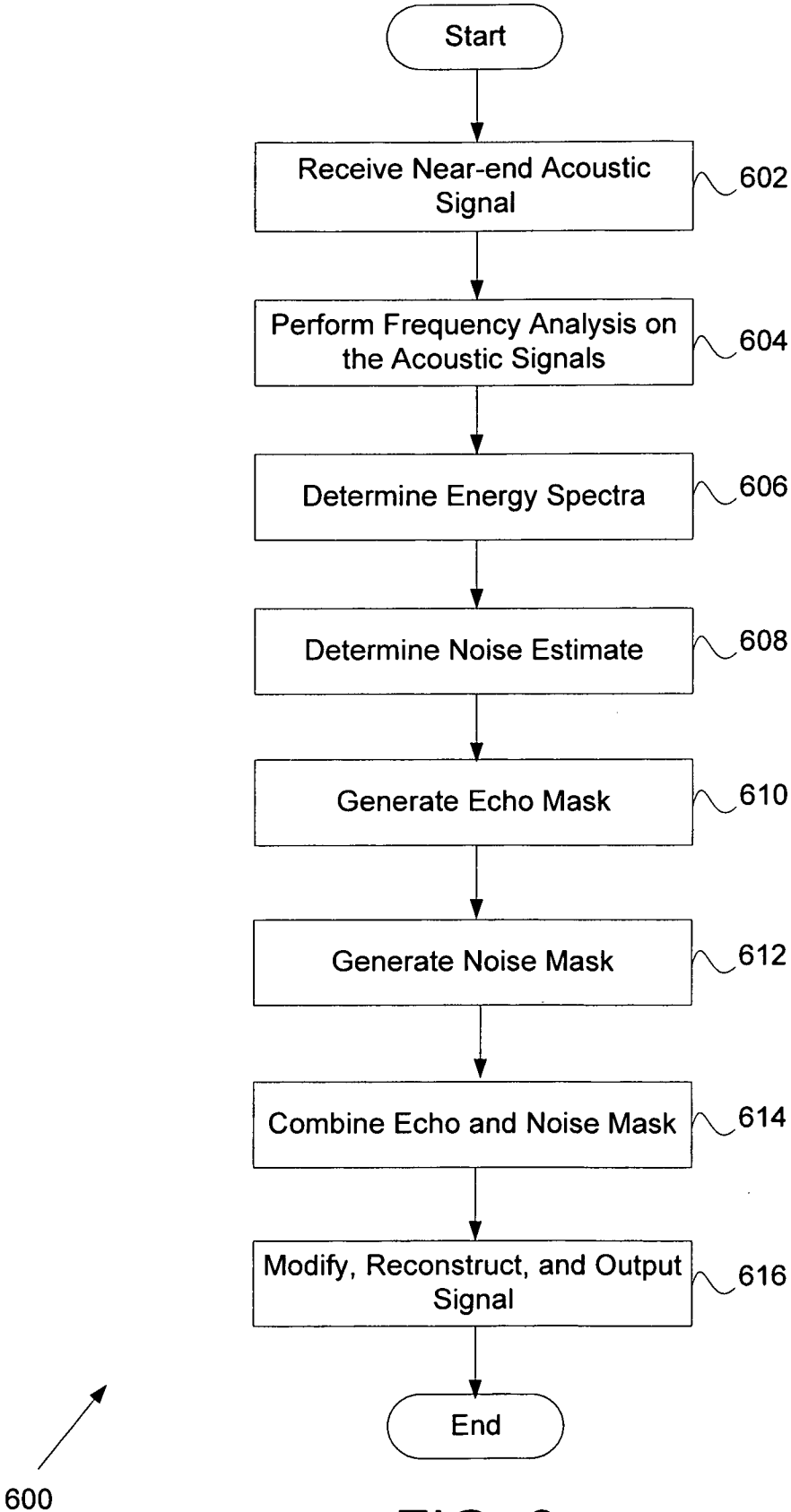

Start
Receive Near-end Acoustic Signal
602
Perform Frequency Analysis on the Acoustic Signals
604
Determine Energy Spectra
606
Determine Noise Estimate
608
Generate Echo Mask
610
Generate Noise Mask
612
Combine Echo and Noise Mask
614
Modify, Reconstruct, and Output Signal
616
End
600

FIG. 6

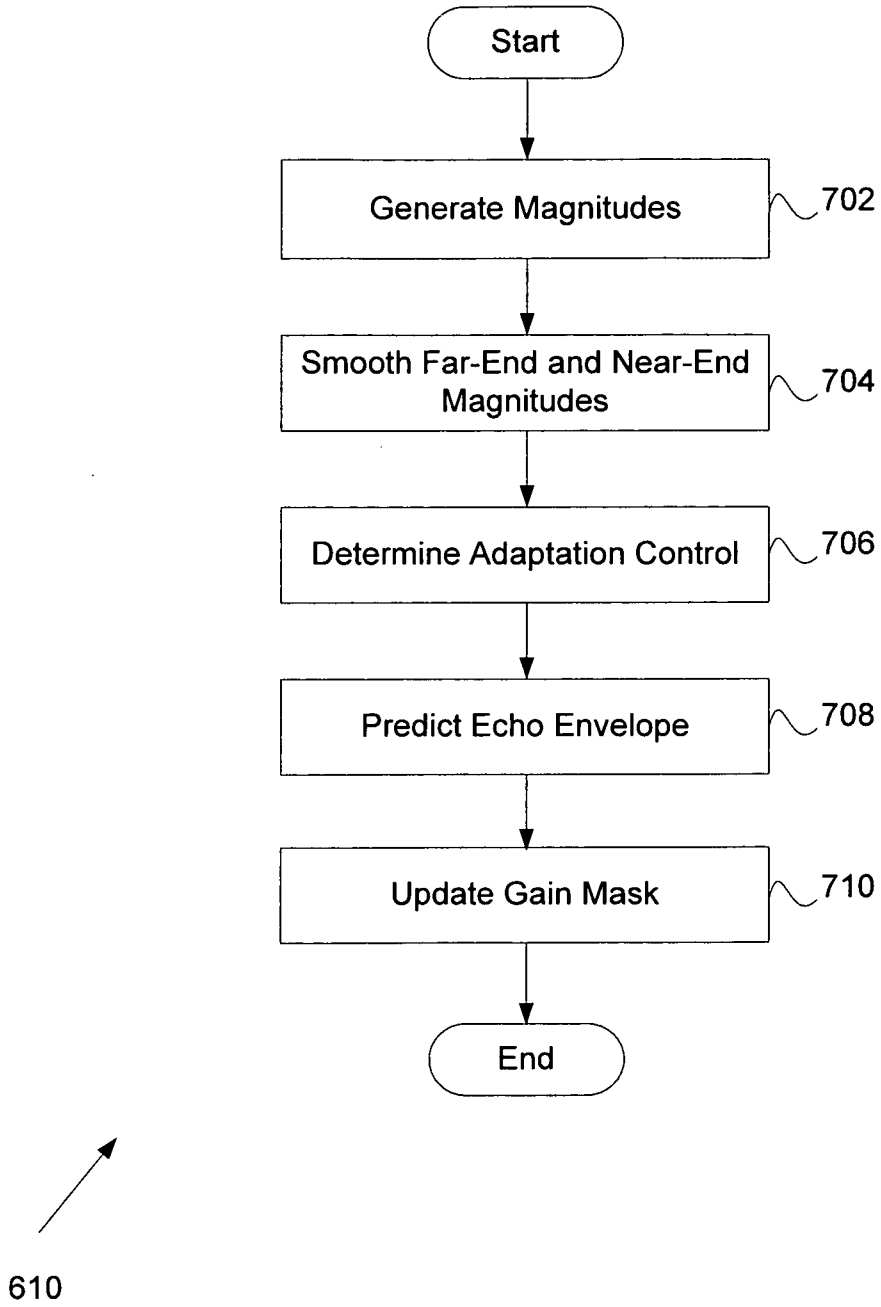

Start
Generate Magnitudes
702
Smooth Far-End and Near-End Magnitudes
704
Determine Adaptation Control
706
Predict Echo Envelope
708
Update Gain Mask
710
End
610

FIG. 7