

ACOUSTIC ECHO CANCELLER

FIELD OF THE INVENTION

5 The present invention relates generally to the field of acoustic echo cancellation in telecommunications, and particularly to a pseudo spectrum-based acoustic echo canceller which adaptively cancels echoes arising in hands-free audio and video teleconferencing and related systems without requiring a state machine or training.

10

BACKGROUND OF THE INVENTION

 Acoustic echo cancellers and their applications in the field of telecommunication are well known to those skilled in the art. Many such
15 cancellers and related technologies have been described in various publications including the following patent documents:

 U.S. Patent No. 5,548,642
 U.S. Patent No. 5,530,724
 U.S. Patent No. 5,506,901
20 U.S. Patent No. 5,428,562
 U.S. Patent No. 5,406,583
 U.S. Patent No. 5,394,392
 U.S. Patent No. 5,384,806
 U.S. Patent No. 5,329,586
25 U.S. Patent No. 5,206,854
 U.S. Patent No. 5,163,044
 U.S. Patent No. 5,146,494
 U.S. Patent No. 5,016,271
 U.S. Patent No. 5,001,701

U.S. Patent No. 4,918,685

U.S. Patent No. 4,817,081

U.S. Patent No. 4,464,545

A typical acoustic echo canceller currently available uses what-is-
5 known-as an adaptive filter which employs a well-known algorithm such as
the algorithm known as the Least-Mean-Square algorithm, or LMS. This
algorithm continuously adapts to changes in the placement of both the
speaker and microphone and to changes in loudspeaker volume. For these
cancellers, a state machine is needed to automatically determine each of
10 the four states, i.e., receiving, transmitting, double-talk, and idle. In addition,
in order to cancel the echoes, these cancellers much be trained, that is, they
must "learn" the loudspeaker-to-microphone acoustic response function for
the room it is servicing. Also, the acoustic compensation length is
determined by the length of the filter that is determined by the host resource
15 availability.

OBJECT OF THE INVENTION

It is an object of the present invention to provide an acoustic echo
20 canceller which adaptively cancels echo arising in hands-free audio and
video conferencing systems and other related systems where echo
cancellation is required.

It is an another object of the present invention to provide an acoustic
echo canceller which provides high-quality and low cost full duplex speech
25 communication typical of dedicated video conferencing systems.

It is yet another object of the present invention to provide an acoustic
echo canceller which does not require a state machine.

It is still yet another object of the present invention to provide an acoustic echo canceller which does not require training.

It is still yet another object of the present invention to provide an acoustic echo canceller which continuously adapts to changes in
5 microphone and loudspeaker placement, loudspeaker volume setting, and the movement of people.

It is still yet another object of the present invention to provide an acoustic echo canceller which is independent of any standard.

It is still yet another object of the present invention to provide an
10 acoustic echo canceller which can be connected directly to a PC soundcard and an ordinary telephone set.

SUMMARY OF THE INVENTION

15 A microphone array is used together with a block adaptive algorithm to effectively suppress acoustic echo arising in hands free voice communication. At the same time, the system is also capable of suppressing environmental noise.

The present echo canceller utilizes the principle that the spectrum
20 pattern of human speech does not change much in the short run. The present echo canceller takes 256 overlap 128 samples in 16 ms intervals, or sample blocks. The power spectrum taken at time 0 and at any time within the 16 ms interval are essentially the same. This is true even though the waveform of the speech may change over time even in the short run. The
25 echoes are simply a delayed form of a speech signal. Therefore, in

following the principle described above, the spectrum of the speech signal and the spectrum of the echo taking are substantially the same.

The inputs to the present echo canceller are $x(t)$ and $y(t)$, $y(t)$ representing the incoming speech signal from a far-end speaker and $x(t)$ representing the combination of speech signal from a near-end speaker and the echo. The well-known normalized cross-correlation estimation between $x(t)$ and $y(t)$ is performed to determine the level of correlation between $x(t)$ and $y(t)$ which is quantitatively represented by the correlation coefficient C , a value of 1 for C being perfect correlation.

When the far-end speaker is speaking and the near-end speaker is not speaking, $x(t)$ comprises of only the echo portion which is essentially a delayed form of $y(t)$. In that case, there is almost a perfect correlation between $x(t)$ and $y(t)$ and the C value is near 1. When the near-end speaker is speaking and the far-end speaker is not speaking, the $x(t)$ comprises only of the signal and the C value is near 0. When both the near-end and the far-end speakers are speaking simultaneously, the C value may be between 0 and 1, but typically near to 0 since the two speech signals will not be highly correlated. And of course, silence would result in a near 0 also, since respective noises will not be highly correlated. Certain decisions are based on whether the C value exceeds certain thresholds.

Since the echo is essentially a delayed $y(t)$, the amount of delay is estimated by measuring the time shift required to produce the maximum C value. Once the delay is determined, the two channels of inputs are aligned by time-shifting $x(t)$ to match $y(t)$. The amplitude of the $x(t)$ and $y(t)$ is

then normalized by first determining a certain gain factor, and then multiplying $y(t)$ by the gain factor.

The processed forms of the input $x(t)$ and $y(t)$ are next processed by applying the well-known Hanning window. They are then transformed
 5 into their respective frequency domain using the well-known fast Fourier transform (FFT) and then to Bark Scales, $P_x(b)$ and $P_y(b)$, using the Bark Frequency Warping technique. The transfer function $H(b)$ is then estimated using the Bark Scales. The transfer function is used to normalize $P_y(b)$, which, in turn together with $P_x(b)$, is used to estimate the gain
 10 $G(b)$ which will be used to suppress the echo. Subsequently, the Bark Scales are unwarped and the gain function is then used to suppress the echo from the input $x(t)$. The well-known inverse FFT (IFFT) and overlap add are performed to yield an echo-free signal.

15 BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a functional diagram illustrating the present echo canceller deployed in a teleconference room setting.

FIG. 2 is functional block diagram illustrating the circuitry of the
 20 present echo canceller.

FIGS. 3a through 3c is a continuous flow diagram illustrating the echo canceling process employed by the present echo canceller.

FIG. 4 is a lookup Table 1 listing values for G_s .

FIG. 5 is a lookup Table 2 listing values for $\bar{L}_b(b)$.

FIG. 6 is a lookup Table 3 listing values for W_i .

DETAILED DESCRIPTION OF THE INVENTION

5 FIG. 1 illustrates schematically the present echo canceller 1 placed in a telephone conference system operating in a room 10. The echo canceller 1 is serially connected to a telecommunications network through incoming line 15 and outgoing line 16. Room reverberative surfaces 18 define multiple echo paths which depend on room geometry. Two such echo
10 paths, 20 and 21, are illustrated. Speech originating from a far-end speaker (speaker not shown) emanating from the room loudspeaker 1 travels along the echo paths 20 and 21, among others paths, and enters microphone 25 with various time delays. Speech 31 from the near-end speaker 30 also enters the microphone 25. Both the speech signal and the echo, denoted
15 as $x(t)$, travel along the line 35 and into the echo canceller 1. The speech signal from the far-end speaker, denoted as $y(t)$, which is essentially the same as the echo without the delay, is also an input to the echo canceller 1 via line 15.

To optimize the performance of the present echo canceller, a
20 microphone array consisting of 3 microphones is used instead of a single microphone, and a well-known beam-forming technique is employed. This arrangement enhances the strength of the near-end speech signal while reducing the strength of the echo signal from the loudspeaker. This occurs because the array forms an acoustic beam at the signal direction and a null
25 in the speaker direction. It has been found that this microphone array significantly enhances the performance of the present invention.

The present echo canceller utilizes the principle that the spectrum pattern of human speech does not change much in the short run. The present echo canceller takes 256 overlap 128 samples in 16 ms intervals, or sample blocks. The power spectrum of a speech signal taken at time 0, for instance, and at any time within the 16 ms interval are essentially the same. This is true even though the waveform of the speech may change over time even in the short run. In referring to FIG. 1, the echo taking the paths 21 and 20, for instance, are simply a delayed form of $y(t)$. Therefore, in following the principle described above, the spectrum of the speech signal from the far-end speaker's speech signal $y(t)$ and the spectrum of the echo taking the paths 21 and 20 are substantially the same. The following description will make it clearer to the those skilled in the art, how this principle is utilized in the present echo canceller to cancel the echo in a manner which is more effective than the currently-available systems.

FIG. 2 illustrates a functional block diagram representing the circuitry for the echo canceller 1 referred to in FIG. 1. Typically, the circuitry would be implemented in a DSP chip or a microprocessor, though it can be implemented in other ways which are known to one skilled in the art. A brief description of the blocks will be given for FIG. 2. A more detailed flow diagram and description for the echo cancellation process employed by the circuit of FIG. 2 shall follow thereafter.

Referring to FIG. 2 in conjunction with FIG. 1, the inputs to the circuit are $x(t)$ and $y(t)$, $y(t)$ representing the incoming speech signal from the far-end speaker and $x(t)$ representing the combination of speech signal from the near-end speaker and the echo. The well-known normalized cross-correlation estimation between $x(t)$ and $y(t)$ is performed in block

100 to determine the level of correlation between $x(t)$ and $y(t)$ which is quantitatively represented by the correlation coefficient C , a value of 1 for C being perfect correlation.

When the far-end speaker is speaking and the near-end speaker 30
5 (see FIG. 1) is not speaking, $x(t)$ comprises of only the echo portion which is essentially a delayed form of $y(t)$. In that case, there is almost a perfect correlation between $x(t)$ and $y(t)$ and the C value is near 1. When the near-end speaker 30 is speaking and the far-end speaker is not speaking, the $x(t)$ comprises only of the signal and the C value is near 0. When both
10 the near-end 30 and the far-end speakers are speaking simultaneously, the C value may be between 0 and 1, but typically near to 0 since the two speech signals will not be highly correlated. And of course, silence would result in a near 0 also, since respective noises will not be highly correlated. Certain decisions are based on whether the C value exceeds certain
15 thresholds.

Since the echo is essentially a delayed $y(t)$, the amount of delay is estimated in block 120 by measuring the time shift required to produce the maximum C value. Once the delay is determined, the two channels of inputs are aligned in block 130 by time-shifting $y(t)$ to match $x(t)$. The
20 amplitude of the $x(t)$ and $y(t)$ is then normalized in block 140 by first determining a certain gain factor, and then multiplying $y(t)$ by the gain factor in block 145.

The processed form of the input $x(t)$ is next processed in blocks 150 through 165; the processed form of the input $y(t)$ is next processed 151
25 through 166. Because both channels are processed in an identical manner

which is well known and understood, only a brief description will be provided. In blocks 150 and 151, the well-known Hanning window is applied to the processed inputs. They are then transformed into their respective frequency domain using the well-known fast Fourier transform (FFT), blocks 155 and 156, and then to Bark Scales, $P_x(b)$ and $P_y(b)$, using the Bark Frequency Warping technique in blocks 165 and 166.

In block 170, the transfer function $H(b)$ is estimated. The transfer function is then used in block 175 to normalize $P_y(b)$, which is then used to estimate the gain $G(b)$ which will be used to suppress the echo. In block 180, the Bark Scales are unwarped. The gain function is then used to suppress the echo from the input spectrum in block 185. The well-known inverse FFT (IFFT) is performed in block 190 and the overlap add in block 195 to yield an echo-free signal.

Using the flow diagrams of FIGs. 3a through 3c and the circuit diagram of FIG. 2, the echo cancellation process employed by the present invention will now be described in greater detail.

Referring now to FIG. 3a, M samples (in this case 256 overlap 128, though other values are possible) are taken from the inputs $x(t)$ and $y(t)$ in step 205 at 16 ms block intervals (8 KHz sampling rate). Sometimes a dc component exists with the inputs and so it is removed, step 210, using a common procedure well known to those skilled in the art. The next step, 220, is to compute the normalized cross-correlation as represented by a value C where,

$$C = \bar{x}^T \bar{y} \quad \text{and} \quad \bar{x} = \frac{x}{\sqrt{x^T x}}, \quad \bar{y} = \frac{y}{\sqrt{y^T y}}$$

where T denotes the transpose of a vector. A number of C values will result from this calculation so in step 220, the maximum value representing C , or $C_{\max}$, is chosen.

Once $C_{\max}$ is found, the amount of delay between the two inputs, or D_n , is estimated in step 230. A comparison is made in step 232 to determine if $C_{\max} > \rho_{\text{new}}$ where ρ_{new} initially has a value of 0. If the condition is met, i.e., $C_{\max} > \rho_{\text{new}}$, then ρ_{new} is updated following the formula $\rho_{\text{new}} = \gamma C_{\max}$ where a value for γ is empirically chosen to be 0.8. The delay D_n is then updated based on the most current value of $C_{\max}$. On the other hand, if the condition $C_{\max} > \rho_{\text{new}}$ is not met, then ρ_{new} is updated using the formula $\rho_{\text{new}} = \gamma \rho_{\text{old}}$ where ρ_{old} simply represents the previous ρ_{new} , and the delay D_n from the previous sample block is used. Whether or not the delay D_n is updated or not, the two inputs, $x(t)$ and $y(t)$, are aligned by delaying the $y(t)$ by the amount D_n in step 245.

It is important to note here that while the updating of the delay D_n is a process included in the preferred embodiment of present invention, it is not crucial. For instance, the present canceller can still function, though not as optimally, even if steps 232, 234, 236, and 240 were eliminated, and step 245 were to be performed immediately after 230 using the same D_n each time.

After the alignment of the inputs in step 245, in step 250, an amplitude normalization is performed on the inputs using a gain normalization factor, Z which is initially set at 1, but which is continually

updated in step 269 when the stated condition is met. In step 255, the well-known Hanning Window is applied and the FFT is computed as follows:

$$\begin{aligned} \mathbf{X}(f) &= x_r(f) + j x_i(f) \quad \text{where} \quad \mathbf{X}(f) = FFT(x) \\ \mathbf{Y}(f) &= y_r(f) + j y_i(f) \quad \text{where} \quad \mathbf{Y}(f) = FFT(y) \end{aligned}$$

Coherence estimation is performed in step 257, where the coherence factor, Φ , is computed is as follows:

10

$$\Phi = \text{Real} \left(\frac{\mathbf{X}(f) \text{conj}(\mathbf{Y}(f))}{\text{norm}(\mathbf{X}(f)) \text{norm}(\mathbf{Y}(f))} \right)$$

It can be seen from this formula that if $\mathbf{X}(f)$ and $\mathbf{Y}(f)$ are coherent, Φ will be near to 1 which indicates that only the echo is present. However, if Φ is near to 0, that indicates either a double-talk or only near-end speech or only silence. The coherence factor, Φ , is used together with a non-linear energy function described in step 267 (see below) to further control the echo suppression.

Thereafter, in step 260, P_x and P_y are computed as follows:

20

$$\begin{aligned} P_x &= |x_r(f)| + |x_i(f)| + \varepsilon * |x_r(f)| * |x_i(f)| \\ P_y &= |y_r(f)| + |y_i(f)| + \varepsilon * |y_r(f)| * |y_i(f)| \end{aligned}$$

where ε is a scaling factor which controls the amount of echo to be suppressed and is a trade-off between speech quality and echo suppression. In step 265, P_x and P_y are converted to Bark Scales $P_x(b)$ and $P_y(b)$ using the well-known Bark Frequency Warping technique.

A non-linear energy computation is performed in step 267 where the energy, E_n , is computed as follows:

$$E_n = \sqrt{\sum_{b=1}^L P_y(b)(b)^2}$$

5

where L represents the number of Bark frequency band. In the preferred embodiment $L = 18$ is used.

In step 269, the gain normalization factor, Z , is updated if the following conditions are met: $\Phi > \tau$ and $E_n > T_n$. The gain

10

normalization factor, Z , is computed as follows:

$$Z = Z\sigma + (1 - \sigma) \frac{\sum_{i=1}^N |x_i|}{\sum_{i=1}^N |y_i|}$$

where $\sigma < 1$. It is important to note that while this is the preferred method other gain normalization methods may be used.

In step 271, it is determined if the condition $E_n < T_n$ is met. If yes, T_n is updated in step 273. T_n is computed as follows:

15

$$T_n = VT_n' + (1 - V)E_n$$

where T_n' is the T_n from the previous run where $V < 1$. It is important to note here that the noise threshold, T_n , is initially estimated during the silence period. It is computed as follows:

20

$$T_n = E_n * \theta$$

where θ is chosen between the range 1.125 and 1.25.

In step 275, it is determined whether $\Phi > \tau$ and $E_n > T_n$. In the preferred embodiment, $\tau = 0.65$, though a different value may be optimal

for τ under different configuration, e.g., different microphone set-up. If the condition in step 275 is met, the transfer function $\mathbf{H}(b)$ is updated from its initial value of 1 in step 280. If the condition in step 275 is not met, then the step 285 is performed without updating the $\mathbf{H}(b)$. The $\mathbf{H}(b)$ is calculated
 5 as follows:

$$\mathbf{H}(b) = \alpha \mathbf{H}(b) + (1 - \alpha) \left(\frac{P_x(b)}{P_y(b)} \right)$$

where $\alpha < 1$

In step 285, $P_y(b)$ is normalised by $\mathbf{H}(b)$ as follows:

$$\tilde{P}_y(b) = \mathbf{H}(b) * P_y(b)$$

10 A buffer is provided to store M old values of $\tilde{P}_y(b)$. In step 295, the total echo power is computed as follows:

$$\tilde{P}_y(b) = \sum_{i=1}^m W_i * \tilde{P}_y^i(b)$$

W_i is a weighting fraction and its value depends on the echo path characteristics. The typical values are listed in Table 3. In step 300, the
 15 value of G_s is found by referring to lookup Table 1 using the current value for Φ .

In step 310, $R_{pr}(b)$ is computed as follows:

$$R_{pr}(b) = \frac{R_{pr}(b)}{(1 + R_{pr}(b))}$$

where $R_{pr}(b) = \gamma P_{rpo}(b) + (1 - \gamma) \bar{P}_{rpo} * G_s$

20 where γ is a smooth factor with $\gamma < 1$ ($\gamma \approx 0.02$) and $\bar{P}_{rpo}$ initially has a value of 0.

and $P_{rpo}(b) = R_{po}(b) - 1$

where if $P_{rpo}(b) < 0$ then $P_{rpo}(b) = 0$

and $R_{po}(b) = \frac{P_x(b)}{\bar{P}_y(b)} * G_s$

5

In step 315, $L(b)$ is computed as follows:

$$L_b = R_{rpr}(b) * R_{po}(b)$$

In step 320, a look-up Table 2 is used to find a value for $\bar{L}_b(b)$. In step 325, the gain $G(b)$ is computed as follows:

10 $G(b) = R_{rpr}(b) * \bar{L}_b(b)$

In step 330, $\bar{P}_{rpo}$ is computed as follows:

$$\bar{P}_{rpo} = \frac{G(b) * P_x(b)}{\bar{P}_y(b)}$$

After the step 330, the steps 310, 315, 320, 325 and 330 are repeatedly performed for each sample block of input, each loop producing an updated
15 value for the parameters involved.

In step 335, the $G(b)$ is unwarped to produce $G(f)$. The output spectrum is then computed in step 340 as follows:

$$\tilde{X}(f) = G(f) * X(f)$$

In step 345 the well-known inverse FFT (IFFT) and overlap add are
20 performed on $\tilde{X}(f)$ and to produce an echo-free signal $\tilde{X}(t)$.

It is very important for one of ordinary skilled in the art to understand that many of the steps and/or components of the preferred embodiment of

the echo canceller of the present invention are included as a way of optimizing the performance of the canceller, and, therefore, may be substituted or even eliminated in some instances without negating the function and the purpose of the present invention. In addition, although the preferred embodiment of the present invention was described in the context of a teleconferencing system, it is clear that the present echo canceller may be used in other telecommunications systems where echoes are present in the similar manner as the scenarios described herein. While one skilled in the art could certainly appreciate these principles, some examples will be given for illustration purposes.

For instance, in referring to FIG. 2 and FIG. 3, the cross-correlation estimation technique employed here may be substituted with other techniques for determining the correction between two signals. Also, the amplitude normalisation, the use of Hanning Window and Bark Scales, while contributing to the effectiveness of the preferred embodiment of the present invention, may be eliminated under some circumstances without completely negating the function of the present invention. The Bark Scales, for instance, are used in this case as way of reducing computation time and, therefore, may not unduly effect the performance of the present echo canceller. In addition, although the Hanning Window was found to be optimal in this case, it may be replaced with other windows. Similarly, while the choice to take 256 overlap 128 samples in 16 ms intervals was found to be optimal in this case, other sample sizes and intervals may be chosen. The presently disclosed embodiments are, therefore, to be considered in all respects as illustrative and not restrictive, the scope of the invention being indicated by the appended claims and all changes which come within the

meaning and range of equivalency of the claims are, therefore, to be embraced therein.



CLAIMS

I Claim:

- 1 1. An acoustic echo canceller for a telecommunications system adapted
 2 for communication between a first speaker and a second speaker, said
 3 system having a microphone and a loud-speaker, said microphone
 4 receiving a speech signal from said first speaker and an echo from a speech
 5 of said second speaker emanating from said loud-speaker, comprising:
 6 a means for collecting samples of two inputs from said
 7 telecommunications system, $x(t)$ and $y(t)$, said $y(t)$ being a speech signal
 8 from a first speaker, said $x(t)$ being a combination of a speech signal from a
 9 second speaker and an echo signal of said $y(t)$;
 10 a means for estimating a correlation between the $x(t)$ and $y(t)$;
 11 a means for estimating a delay between the $x(t)$ and $y(t)$;
 12 a means for aligning $x(t)$ and $y(t)$;
 13 a means for transforming said signals into a frequency domain and
 14 back to time domain;
 15 a means for estimating a transfer function;
 16 a means for estimating a gain function; and
 17 a means for multiplying said input signals by said gain function.
- 1 2. The echo canceller as claimed in claim 1 further comprising a means
 2 for applying a Henning Window to said input signals.
- 1 3. The echo canceller as claimed in claim 1 further comprising a means
 2 for Bark Scale Warping and means for Bark Scale Unwarping.

This Sheet Has Been Replaced





1 4. The echo canceller as claimed in claim 2 further comprising a means
2 for Bark Scale Warping and means for Bark Scale Unwarping.

1 5. The echo canceller as claimed in claim 1 wherein 256 samples are
2 collected overlapping 128 samples in 16 ms block intervals.

1 6. The echo canceller as claimed in claim 5 further comprising a means
2 for overlap add.

1 7. A method of cancelling an acoustic echo in a telecommunications
2 system adapted for communication between a first speaker and a second
3 speaker, said system having a microphone and a loud-speaker, said
4 microphone receiving a speech signal from said first speaker and an echo
5 from a speech of said second speaker emanating from said loud-speaker,
6 said method comprising:

7 a) collecting samples of two inputs from said telecommunications
8 system, $x(t)$ and $y(t)$, said $y(t)$ being a speech signal from a first speaker, said
9 $x(t)$ being a combination of a speech signal from a second speaker and an
10 echo signal of said $y(t)$;

11 b) estimating a correlation between the $x(t)$ and $y(t)$;

12 c) estimating a delay between the $x(t)$ and $y(t)$;

13 d) aligning $x(t)$ and $y(t)$;

14 e) transforming said signals into their spectrums;

15 f) estimating a transfer function;

16 g) estimating a gain function; and

17 h) multiplying the spectrum of $x(t)$ by said gain function;

18 i) transforming said resulting spectrum of from step h) into a
19 time domain signal._

This Sheet Has Been Replaced





- 1 8. The method as claimed in claim 7 further comprising applying a
2 Henning Window to said input signals after step d).
- 1 9. The method as claimed in claim 7 further Bark Scale Warping said
2 spectrums from step e) and Bark Scale Unwarping after step g).
- 1 10. The method as claimed in claim 7 further Bark Scale Warping said
2 spectrums from step e) and Bark Scale Unwarping after step g).
- 1 11. The method as claimed in claim 7 wherein said 256 samples are
2 collected overlapping 128 samples in 16 ms block intervals.
- 1 12. The method as claimed in claim 11 further comprising performing
2 overlap add.

