EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Single-channel acoustic echo cancellation in noise based on gradient-based adaptive filtering

Upal Mahbub[1*], Shaikh Anowarul Fattah[1], Wei-Ping Zhu[2] and M Omair Ahmad[2]

Abstract

In this paper, a two-stage scheme is proposed to deal with the difficult problem of acoustic echo cancellation (AEC) in single-channel scenario in the presence of noise. In order to overcome the major challenge of getting a separate reference signal in adaptive filter-based AEC problem, the delayed version of the echo and noise suppressed signal is proposed to use as reference. A modified objective function is thereby derived for a gradient-based adaptive filter algorithm, and proof of its convergence to the optimum Wiener-Hopf solution is established. The output of the AEC block is fed to an acoustic noise cancellation (ANC) block where a spectral subtraction-based algorithm with an adaptive spectral floor estimation is employed. In order to obtain fast but smooth convergence with maximum possible echo and noise suppression, a set of updating constraints is proposed based on various speech characteristics (e.g., energy and correlation) of reference and current frames considering whether they are voiced, unvoiced, or pause. Extensive experimentation is carried out on several echo and noise corrupted natural utterances taken from the TIMIT database, and it is found that the proposed scheme can significantly reduce the effect of both echo and noise in terms of objective and subjective quality measures.

Keywords: Adaptive filter; Convergence analysis; Echo cancellation; Least mean squares algorithm; Noise reduction; Spectral subtraction; Single-channel communication

1 Introduction

The phenomenon of acoustic echo occurs when the output speech signal from a loudspeaker gets reflected from different surfaces, like ceilings, walls, and floors and then fed back to the microphone. In its worst case, acoustic echo can cause howling of a significant portion of sound energy [1,2]. In real life applications, such as a lecture in a large conference hall or in the public address system of a trade fair, the presence of acoustic echo along with the environmental noise is a very common phenomenon, which degrades the speech quality even leading to complete loss of intelligibility.

In order to deal with the problem of acoustic echo cancellation (AEC), conventionally echo suppressors, earphones, and directional microphones have been used, which generally place restrictions on the talkers' movement [2]. As an alternate of such hardware-based solutions, adaptive filter algorithms are widely being applied where apart from the input channel, a separate echo-free reference channel is required [3-13]. Among different adaptive filter algorithms, the least mean squares (LMS) algorithm and its different variants are very popular for their satisfactory performances and less computational burden [4,10,12-14]. Besides these algorithms, the recursive least squares (RLS) algorithm is well-known for its fast convergence at the expense of computational complexity [13]. The adaptive filter algorithms have also been used for acoustic noise cancellation (ANC) [15].

There are some methods that deal with both acoustic echo and noise cancellation (AENC) [16-18]. The echo canceller used in [16] utilizes a sub-band noise cancellation scheme. In [17], echo cancellation is done by an adaptive LMS filter while a linear prediction error fil-

*Correspondence: omeecd@eee.buet.ac.bd
[1]Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka 1000, Bangladesh
Full list of author information is available at the end of the article

ter removes the residual echo and noise. In [18], a single Wiener filter is employed to simultaneously suppress the echo and noise. It is to be mentioned that all these AENC methods employ more than one microphone, while the solutions using single microphone are favorable in most of the real-life applications.

In this paper, an AENC scheme is proposed which can efficiently deal with the single-channel scenario. First, unlike conventional LMS algorithm, considering the delayed version of the previously echo- and noise-suppressed signal as reference, a gradient-based adaptive LMS algorithm is developed for single channel AEC. Preliminary results obtained by using this idea is reported in [19]. However, in the current paper, analytical proof of convergence towards the optimum Wiener-Hopf solution is presented. Next, a single-channel ANC algorithm based on spectral subtraction with an adaptive spectral floor estimation is developed, which reduces not only the effect of noise but also some residual echo. Finally, analyzing different speech characteristics of the reference and current frames, multiconditional updating constraints are proposed in order to obtain precise control on convergence characteristics. For performance evaluation, extensive experimentation is conducted on several real-life echo and noise corrupted speech signals at different acoustic environments.

2 Problem formulation

In order to formulate the problem of single-channel AENC, for a better understanding, first, a dual channel AENC scheme is presented in Figure 1 (according to [17]). Here, $s_1(n)$ and $s_2(n)$ are speech signals corresponding to near-end and far-end speakers, while $v_1(n)$ and $v_2(n)$ are additive noises, respectively. The noise corrupted far-end signal $(s_2(n) + v_2(n))$ is played through a loudspeaker at the near-end acoustic room environment and the echo signal $x_2(n)$ is generated. Thus, the input $y_1(n)$ to the near-end microphone is given by

$$y_1(n) = s_1(n) + v_1(n) + x_2(n). \tag{1}$$

The task of the adaptive filter-based AEC block placed at the near-end is to produce an estimate $\widehat{x}_2(n)$ of the echo $x_2(n)$ by minimizing the error

$$e_1(n) = y_1(n) - \widehat{x}_2(n). \tag{2}$$

Two major issues in dual channel system are (i) availability of a separate reference signal required for the adaptive filter, for example, here the delayed version of $(s_2(n) + v_2(n))$ and (ii) different speakers for input and echo signals. Moreover, use of the double talk detector (DTD) helps in controlling the update process. Unfortunately, these features are absent in single-channel scenario as shown in Figure 2. Instead of two speakers, in this case, the microphone receives the input $s(n)$ corrupted by noise $v(n)$ and echo generated from the same speaker.

In the presence of noise $v(n)$, the sole microphone input signal in single-channel scenario is given by

$$y(n) = s(n) + v(n) + x_s(n) + x_v(n), \tag{3}$$

where $x_s(n)$ and $x_v(n)$ denote the echo of the input speech and noise, respectively. The echo signals can be expressed as

$$x_s(n) = \mathbf{a}_n^T \mathbf{s}(n - k_0), \tag{4}$$

$$x_v(n) = \mathbf{a}_n^T \mathbf{v}(n - k_0), \tag{5}$$

where $\mathbf{s}(n-k_0) = [\, s(n-k_0-1), s(n-k_0-2), \ldots, s(n-k_0-p)]^T$ and $\mathbf{v}(n-k_0) = [\, v(n-k_0-1), v(n-k_0-2), \ldots, v(n-k_0-p)]^T$ with k_0 being a predefined flat delay and $\mathbf{a}_n = [\, a_n(1), a_n(2), \ldots, a_n(p)]^T$ consists of the coefficients corresponding to the acoustic room transfer function $A(z)$. The order p and coefficient values of $A(z)$ depend on the room characteristics. It is to be noted that in this case, there is no scope of obtaining a separate echo-free reference or a separate noise-only reference, which makes the single-channel AENC problem extremely difficult to handle.

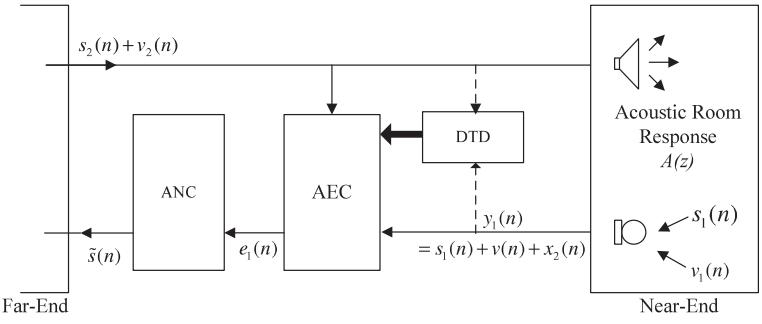


Figure 1 Adaptive filter-based echo and noise cancellation in dual channel communication system.

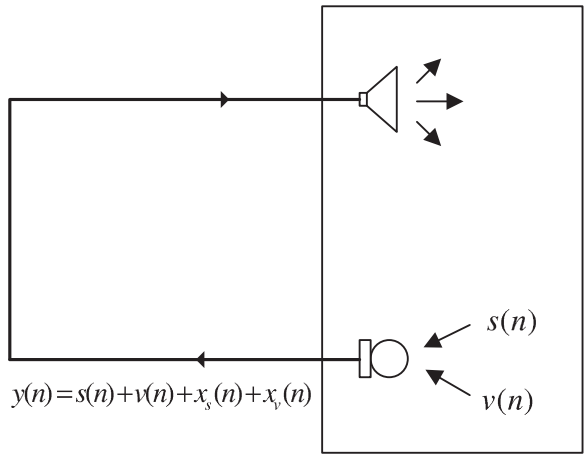


Figure 2 Single-channel acoustic echo generation in noisy room environment.

3 Proposed single-channel AENC scheme

3.1 Proposed two-stage setup

In Figure 3a, a simple block diagram showing two stages of the proposed AENC scheme is presented and in Figure 3b, more detail of the adaptive filter-based AEC algorithm involved in the first stage is shown. Similar to Figure 2, the input to the microphone $y(n)$ can be described by (3). For the case of single-channel AEC, for example, while delivering a lecture in a large conference hall, the microphone in front of the speaker receives input speech $s(n)$ corrupted by $v(n)$. Once this noise-corrupted speech is transmitted through loudspeaker, echo signal is generated and thus the microphone after some initial time delay will receive noise-corrupted speech and echo of previously uttered speech. The task of AEC is to cancel the echo part from this input by using adaptive filter algorithm. In order to obtain adaptively an estimate $\widehat{x}_s(n)+\widehat{x}_v(n)$ of the echo signal, we propose to utilize delayed versions of the previously echo-suppressed samples of the noisy speech as reference signal [19]. A symbol hat on the variable is used to indicate estimated value. The error signal $e(n)$ thus obtained is given by

$$e(n) = y(n)-[\widehat{x}_s(n)+\widehat{x}_v(n)]. \tag{6}$$

The estimate of the echo signal can be expressed as

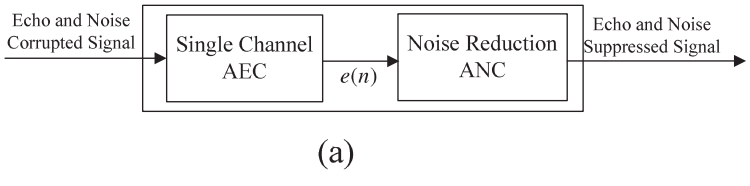
$$\widehat{x}_s(n)+\widehat{x}_v(n) = \widehat{\mathbf{w}}_n^T[\widehat{\mathbf{s}}(n-k_0)+\widehat{\mathbf{v}}(n-k_0)], \tag{7}$$

where $\widehat{\mathbf{w}}_n = [\widehat{w}_n(1), \widehat{w}_n(2)\ldots\widehat{w}_n(p)]^T$ is the estimated coefficient vector. The task of the adaptive filter is to obtain an optimum $\widehat{\mathbf{w}}_n$ by minimizing the error in (6) i.e.,

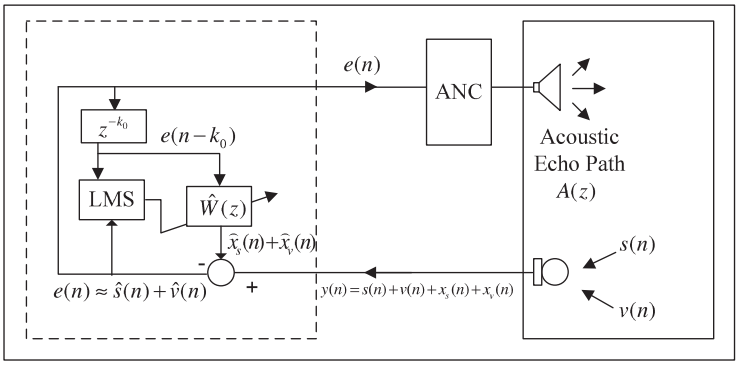
$$e(n) = s(n)+\{(v(n)+\delta_s(n))+\delta_v(n)\}, \tag{8}$$

where $\delta_s(n) = x_s(n)-\widehat{x}_s(n)$ and $\delta_v(n) = x_v(n)-\widehat{x}_v(n)$ are the residual echo of the speech and noise portions of the input signal, respectively, and it is assumed that these signals exhibit the properties of white Gaussian noise. Next, $e(n)$ is passed through a spectral subtraction-based single-channel ANC block which produces output $\widetilde{s}(n) \approx s(n)+\Psi(n)$ that closely resembles $s(n)$ provided that the residual echo-noise portion $\Psi(n)$ becomes very small.

It is to be noted that the task of noise reduction, unlike the proposed AENC scheme, may be carried out prior



(a)



(b)

Figure 3 Block diagram of proposed single-channel AENC scheme. (a) Two stages and **(b)** details of proposed adaptive filter-based AEC algorithm.

to the AEC block. However, because of possible nonlinearities introduced by the prior noise reduction block, no proper reference would be available for the single-channel AEC block [17]. Hence, the arrangement shown in Figure 3a is adopted, in which the noise reduction block also serves as a post-processor for attenuating the residual echo.

3.2 Development of proposed gradient-based single-channel LMS AEC scheme

A delayed version of the adaptive filter output $e(n)$ is proposed to use as the reference signal, and from (8), filter output $e(n)$ can be written as

$$e(n) = \widehat{s}(n) + \widehat{v}(n), \tag{9}$$

where $\widehat{s}(n) = s(n) + \delta_s(n)$ and $\widehat{v}(n) = v(n) + \delta_v(n)$. The objective function of the adaptive filter involves minimization of the mean square estimation of the error function and using (6) it can be written as

$$\begin{aligned} E\{e^2(n)\} &= E\{(s(n)+v(n))^2\} + E\{(x_s(n)+x_v(n) \\ &\quad -\widehat{x}_s(n)-\widehat{x}_v(n))^2\} + 2E\{(s(n)+v(n)) \\ &\quad \times (x_s(n)+x_v(n)-\widehat{x}_s(n)-\widehat{x}_v(n))\}, \end{aligned} \tag{10}$$

where $E\{.\}$ denotes the expectation operator. In (10), it is intended to use the basic definition of cross-correlation operation, for example, the cross-correlation function between $s(n)$ and $v(n)$ is defined as

$$r_{sv}(m) = E\{s(n)v(n-m)\}, \tag{11}$$

where m denotes the lag. Using (4), (5), (7), and the above definition, the last term of (10) can be expressed as

$$\begin{aligned} &2E\{[\,(s(n)+v(n))(x_s(n)+x_v(n)-\widehat{x}_s(n)-\widehat{x}_v(n))]\,\} \\ &= 2\sum_{k=1}^{k=p}\{(a_n(k)-\widehat{w}_n(k))(r_{ss}(k_0+k)+r_{sv}(k_0+k) \\ &\quad + r_{vs}(k_0+k)+r_{vv}(k_0+k)) - r_{s\delta_s}(k_0+k) \\ &\quad - r_{s\delta_v}(k_0+k) - r_{v\delta_s}(k_0+k) - r_{v\delta_v}(k_0+k)\}. \end{aligned} \tag{12}$$

Here, $r_{ss}(k_0+k)$ corresponds to the (k_0+k)th lag of the cross-correlation between $s(n)$ and its previous samples $s(n-k_0-k)$, and $r_{sv}(k_0+k)$ corresponds to the (k_0+k)th lag of the cross-correlation between $s(n)$ and $v(n-k_0-k)$. In a similar way, $r_{vs}(k_0+k)$, $r_{vv}(k_0+k)$, $r_{s\delta_s}(k_0+k)$, $r_{s\delta_v}(k_0+k)$, $r_{v\delta_s}(k_0+k)$, and $r_{v\delta_v}(k_0+k)$ can be defined. It is well known that the value of cross-correlation decreases rapidly with the increasing lags when two signals are uncorrelated. In ideal case, the cross-correlation function between two random noise signals would be nonzero only at the zero lag. Since $v(n)$ is assumed to be white Gaussian noise and, generally, the value of k_0 is very large, in (12), the effect of the terms $r_{sv}(k_0+k)$, $r_{vs}(k_0+k)$, and $r_{vv}(k_0+k)$ can be neglected. Moreover, because of noise-like characteristics of $\delta_s(n)$ and $\delta_v(n)$, in (12), one can neglect $r_{s\delta_v}(k_0+k)$, $r_{v\delta_s}(k_0+k)$, and $r_{v\delta_v}(k_0+k)$ too. Hence, it can easily be comprehended that optimal filter performance occurs when $r_{ss}(n)$ is minimum, i.e., the least possible correlation between $s(n-k_0-k)$ and $s(n)$ is desired. As a result, (10) reduces to

$$\begin{aligned} E\{e^2(n)\} &= E\{(s(n)+v(n))^2\} \\ &\quad + E\{[\,x_s(n)+x_v(n)-\widehat{x}_s(n)-\widehat{x}_v(n)]^2\,\} \\ &\quad + 2\sum_{k=1}^{k=p}(a_n(k)-\widehat{w}_n(k))r_{ss}(k_0+k). \end{aligned} \tag{13}$$

Here, the magnitude of $r_{ss}(k_0+k)$ strongly depends on speech characteristics and the amount of flat delay k_0. For a reasonably large k_0, the effect of $r_{ss}(k_0+k)$ in 13 can be neglected, and minimization of (13) results in

$$\begin{aligned} &\frac{\partial E\{e^2(n)\}}{\partial \widehat{\mathbf{w}}_n^T} = 0 \\ &E[\{x_s(n)+x_v(n)-\widehat{x}_s(n)-\widehat{x}_v(n)\}\{\widehat{\mathbf{s}}(n-k_0)+\widehat{\mathbf{v}}(n-k_0)\}]=0. \end{aligned} \tag{14}$$

Hence, we obtain

$$\begin{aligned} &E\{(x_s(n)+x_v(n))(\widehat{\mathbf{s}}(n-k_0)+\widehat{\mathbf{v}}(n-k_0))\} \\ &= \widehat{\mathbf{w}}_n^T E[\,\{\widehat{\mathbf{s}}(n-k_0)+\widehat{\mathbf{v}}(n-k_0)\}\{\widehat{\mathbf{s}}(n-k_0)+\widehat{\mathbf{v}}(n-k_0)\}]\,. \end{aligned} \tag{15}$$

The above equation is similar to Wiener-Hopf equation and its solution can be written as

$$\widehat{\mathbf{w}}_n^T = \mathbf{R}_{(s+v)(s+v)}(n-k_0)^{-1}\mathbf{r}_{(x_s+x_v)(s+v)}(n-k_0), \tag{16}$$

where $\mathbf{r}_{(x_s+x_v)(s+v)}(n-k_0)$ consists of different lags of cross-correlation between the echo signal $x_s(n)+x_v(n)$ and the noisy input signal $s(n)+v(n)$, while $\mathbf{R}_{(s+v)(s+v)}$ is the auto-correlation matrix of $s(n)+v(n)$. There is no doubt that $\widehat{\mathbf{w}}_n$ is the most optimum solution possible. Hence, it is shown that even for a single-channel noise corrupted AEC problem, the most optimum solution $\widehat{\mathbf{w}}_n$ can be achieved under the assumptions stated earlier.

For iterative estimation of optimal filter coefficients, the adaptive LMS algorithm is very popular. It is fast and efficient, and it does not require any correlation measurements or matrix inversion [13]. The update equation of the LMS adaptive algorithm is generally expressed as

$$\widehat{\mathbf{w}}_{n+1}^T = \widehat{\mathbf{w}}_n^T - \mu\nabla\xi(n), \tag{17}$$

where μ is the step factor controlling the stability and rate of convergence, $\xi(n)$ is the cost function, and ∇ is the gradient operator. The LMS algorithm simply approximates the mean square error by the square of the instantaneous

error, i.e., $\xi(n) = e^2(n)$, and therefore, from (6) and (7) , the gradient of $\xi(n)$ can be expressed as

$$\nabla\xi(n) = \frac{\partial\xi(n)}{\partial\widehat{\mathbf{w}}_n^T} = -2e(n)(\widehat{\mathbf{s}}(n-k_0)+\widehat{\mathbf{v}}(n-k_0)).$$

Thus, the update equation for the proposed single-channel LMS adaptive scheme can be written as

$$\widehat{\mathbf{w}}_{n+1}^T = \widehat{\mathbf{w}}_n^T + 2\mu e(n)(\widehat{\mathbf{s}}(n-k_0)+\widehat{\mathbf{v}}(n-k_0)). \quad (18)$$

3.3 Convergence analysis of the proposed AEC scheme

Considering expectation operation on both sides of the update Eq. 18, one can obtain

$$\underline{\widehat{\mathbf{w}}}_{n+1}^T = \underline{\widehat{\mathbf{w}}}_n^T + 2\mu E\{e(n)(\widehat{\mathbf{s}}(n-k_0)+\widehat{\mathbf{v}}(n-k_0))\}. \quad (19)$$

Here, an underline beneath $\widehat{\mathbf{w}}_n$ is introduced to represent the expected value $E\{\widehat{\mathbf{w}}_n\}$. For the kth unknown weight vector (where $k = 1,2,\ldots,p$), using (6) and neglecting the effect of $r_{ss}(n)$ that has already been discussed in the previous subsection, the last term of (19) can be written as

$$\begin{aligned}&2\mu E\{e(n)(\widehat{\mathbf{s}}(n-k_0)+\widehat{\mathbf{v}}(n-k_0))\}\\ &= 2\mu E\{[x_s(n)+x_v(n)-\widehat{x}_s(n)-\widehat{x}_v(n)]\times(\widehat{\mathbf{s}}(n-k_0)\\ &\quad +\widehat{\mathbf{v}}(n-k_0))\}. \end{aligned} \quad (20)$$

Based on the assumptions on cross-correlation terms stated in the previous subsection, one can obtain

$$\begin{aligned}E\{e(n)(\widehat{\mathbf{s}}(n-k_0)+\widehat{\mathbf{v}}(n-k_0))\} &= \mathbf{r}_{(x_s+x_v)(s+v)}(n-k_0)\\ &\quad -\mathbf{R}_{(s+v)(s+v)}(n-k_0)\widehat{\mathbf{w}}_n^T.\end{aligned} \quad (21)$$

Using (21), the update Eq. 19 can be written as

$$\begin{aligned}\underline{\widehat{\mathbf{w}}}_{n+1}^T &= \underline{\widehat{\mathbf{w}}}_n^T - 2\mu\mathbf{R}_{(s+v)(s+v)}(n-k_0)\underline{\widehat{\mathbf{w}}}_n^T\\ &\quad +2\mu\mathbf{r}_{(x_s+x_v)(s+v)}(n-k_0).\end{aligned} \quad (22)$$

Evaluating the homogeneous and particular solutions of (22), the total solution can be obtained as (see Appendix)

$$\underline{\widehat{w}}_{n+1}^U(k) = C_k(1-2\mu\lambda(k))^n + \frac{1}{\lambda(k)}r^U(n-k_0-k), \quad (23)$$

where $\lambda(k)$ is the kth diagonal element of the eigenvalue matrix obtained by eigenvalue decomposition of $\mathbf{R}_{(s+v)(s+v)}(n-k_0)$ and $r^U(n-k_0-k)$ is the kth element of $\mathbf{U}^T\mathbf{r}_{(x_s+x_v)(s+v)}(n-k_0) = \mathbf{r}^U_{(x_s+x_v)(s+v)}(n-k_0)$ with the matrix $\mathbf{U}$ consisting of eigenvectors corresponding to eigenvalues. Since in the iterative update procedure, the homogeneous part $(1-2\mu\lambda(k))^n$ diminishes with iterations, (23) in a matrix form can be expressed as

$$\begin{aligned}\underline{\widehat{\mathbf{w}}}^T &= \mathbf{U}\Lambda^{-1}\mathbf{U}^T\mathbf{r}_{(x_s+x_v)(s+v)}(n-k_0)\\ &= \mathbf{R}^{-1}_{(s+v)(s+v)}(n-k_0)\mathbf{r}_{(x_s+x_v)(s+v)}(n-k_0).\end{aligned} \quad (24)$$

Thus, it is found that the average value of the weight vector converges to the Wiener-Hopf solution, which is the optimum solution with increasing number of iteration.

3.4 Noise reduction in spectral domain

In the proposed AENC scheme, the operation of the ANC block is processed frame by frame for noise reduction based on single-channel spectral subtraction algorithm [20-22]. According to (9), for the ith frame, the error signal for the duration of a frame length can be written as

$$e_i(n) = \widehat{s}_i(n) + \widehat{v}_i(n). \quad (25)$$

Corresponding frequency domain representation is given by

$$E_i(\omega) = \widehat{S}_i(\omega) + \widehat{V}_i(\omega). \quad (26)$$

The magnitude squared spectrum of $\widehat{s}_i(n)$ can be written as

$$|\widehat{S}_i(\omega)|^2 = |E_i(\omega)|^2 - |\widehat{V}_i(\omega)|^2 - \widehat{V}_i(\omega)\widehat{S}_i^*(\omega) - \widehat{S}_i(\omega)\widehat{V}_i^*(\omega). \quad (27)$$

It is desired to choose an estimate $\widetilde{S}_i(\omega)$ that will minimize

$$Err_i(\omega) = ||\widetilde{S}_i(\omega)|^2 - |\widehat{S}_i(\omega)|^2|. \quad (28)$$

Since the noise is assumed to be zero mean and uncorrelated with the signal, the expected values of the last two terms of (27) can be neglected. Thus, (28) can be expressed as

$$Err_i(\omega) = |\widetilde{S}_i(\omega)|^2 - |E_i(\omega)|^2 + E\{|\widehat{V}_i(\omega)|^2\}. \quad (29)$$

This expression of $Err_i(\omega)$ can be minimized by choosing

$$|\widetilde{S}_i(\omega)|^2 = |E_i(\omega)|^2 - E\{|\widehat{V}_i(\omega)|^2\}. \quad (30)$$

With an estimate of noise spectrum $E\{|\widehat{V}_i(\omega)|^2\}$, signal spectrum $\widetilde{S}_i(\omega)$ can be computed as

$$\widetilde{S}_i(\omega) = |\widetilde{S}_i(\omega)|\, e^{\mathrm{j}\arg[E_i(\omega)]}, \quad (31)$$

where the phase $(\arg[E_i(\omega)])$ is generally assumed to be the phase of the noise corrupted signal without causing significant degradation in terms of loss of intelligibility of the speech signal [20]. It can be seen that an estimate of the magnitude spectrum $|\widetilde{S}_i(\omega)|$ of the signal can be obtained provided an estimate of noise spectrum $E\{|\widehat{V}_i(\omega)|^2\}$ is available, which is generally computed during the periods when speech is known *a priori* not to be present.

Final output of the AENC system is the speech frame $(\widetilde{s}_i(n))$, which consists of the original speech $s_i(n)$ and a negligible amount of noise-like signal $\Psi_i(n)$. The signal $\Psi_i(n)$, although very weak, may contain some signature of the input noise $v(n)$, the residual echo $\delta_s(n)$, and the residual noise $\delta_v(n)$. In order to overcome the problem of

musical noise and to avoid the speech distortion caused by speech subtraction, in (31), an over estimate of the noise power spectrum can be subtracted carefully such that the spectral floor is preserved [21]. Thus, (31) can be modified as

$$\begin{aligned} |\,\widetilde{S}_i(\omega)\,|^2 &= |\,\widehat{E}_i(\omega)\,|^2 - \alpha_{ss} E\{|\,\widehat{V}_i(\omega)\,|^2\}, \\ &\quad \text{if } |\,\widetilde{S}_i(\omega)\,|^2 > \beta_{ss}\{|\,\widehat{V}_i(\omega)\,|^2\} \\ &= \beta_{ss}\{|\,\widehat{V}_i(\omega)\,|^2\}, \text{ otherwise.} \end{aligned} \tag{32}$$

Here, α_{ss} is the subtraction factor and β_{ss} is the spectral floor parameter with $\alpha_{ss} \geq 1$ and $0 \leq \beta_{ss} \leq 1$. The task of noise power spectral density estimation is carried out based on the minimum statistics noise estimator proposed in [23] which can handle the time-varying nature of the noise.

4 Development of adaptive update constraints

The AEC part of the proposed AENC scheme may suffer from some common problems of adaptive filter-based algorithms, such as slow convergence rate and fluctuation around the desired estimates, especially in practical cases where the assumption on negligibility of cross-correlation terms (stated in the previous section) may not strictly hold. In order to overcome such problems, some updating constraints are proposed based on the following speech characteristics:

(i) The level of cross-correlation
(ii) The amount of signal power
(iii) The mean square error (MSE) between consecutive estimates of the unknown filter coefficients.

Through extensive experimentation on different speech frames, it is found that the negligibility of the cross-correlation terms $r_{ss}(n)$, $r_{s\delta_v}(n)$, $r_{v\delta_s}(n)$, and $r_{v\delta_v}(n)$ (as described after (12)) strongly depends on the voicing characteristics of speech frames and the input noise. Because of inherent periodicity of the voiced speech frame, the degree of cross-correlation between two voiced speech frames of a person becomes higher in comparison to that between two unvoiced speech frames which are random in nature. Regarding signal power, the ratio of power of a voiced speech frame and an unvoiced speech frame is found to be higher in comparison to that of the two voiced speech frames. As white Gaussian noise is considered, the degree of cross-correlation between the speech and noise is found to be negligible and the noise powers in two different frames may not differ significantly. As a result, the effect of input noise is found to be negligible on the power ratio.

For a flat delay of k_0 samples, the initial k_0 samples of the utterance $s(n) + v(n)$ can be treated as a reference signal (echo-free signal) responsible for the generation of echo signal that corrupts the current samples at or after k_0 samples. Considering a window of M samples with $M \ll K_0$, power of the reference signal $(\widehat{s}(n-k_0) + \widehat{v}(n-k_0))$ can be computed as

$$P_{\text{ref}}(n) = \frac{1}{M} \sum_{i=-\frac{M}{2}}^{\frac{M}{2}-1} [\widehat{s}(n-k_0+i) + \widehat{v}(n-k_0+i)]^2 . \tag{33}$$

For a window of last M samples of the echo-suppressed speech signal $\widehat{s}(n)$, the average power $P_{\text{sup}}(n)$ can be computed as

$$P_{\text{sup}}(n) = \frac{1}{M} \sum_{j=0}^{M-1} [\widehat{s}(n-j) + \widehat{v}(n-j)]^2 . \tag{34}$$

The ratio of $P_{\text{ref}}(n)$ and $P_{\text{sup}}(n)$ is denoted as the power ratio $P_{\text{rs}}(n)$ and considered as one of the control characteristics.

Another important characteristic criterion is the correlation coefficient $C_{\text{rs}}(n)$ between a frame of the noisy reference signal $(\widehat{s}(n-k_0) + \widehat{v}(n-k_0))$ and a frame of the current noisy signal $(\widehat{s}(n) + \widehat{v}(n))$. For a frame length of M samples, correlation coefficient $C_{\text{rs}}(n)$ is defined as

$$\begin{aligned} C_{\text{rs}}(n) &= \frac{1}{\sigma_{\widehat{s}(n-k_0+i)+\widehat{v}(n-k_0+i)} \sigma_{\widehat{s}(n-j)+\widehat{v}(n-j)}} \\ &\quad \times \{\text{cov}((\,\widehat{s}\,(n-k_0+i) + \widehat{v}(n-k_0+i)) \\ &\quad \times (\,\widehat{s}\,(n-j) + \widehat{v}(n-j)))\} \end{aligned} \tag{35}$$

where $-M/2 \leq i \leq M/2 - 1$ and $0 \leq j \leq (M-1)$.

Finally, the parameter estimation accuracy is also considered for the purpose of analyzing the convergence property. In this regard, the mean square error $\text{MSE}_{\text{ideal}}(n)$ between the values of estimated coefficients $\widehat{w}_n$ and those of true coefficients a_n is computed as

$$\text{MSE}_{\text{ideal}}(n) = \frac{1}{p} \sum_{k=1}^{p} [\,\widehat{w}_n(k) - a_n(k)]^2 . \tag{36}$$

In Figure 4, considering a real-life speech utterance of 250 ms corrupted by echo and noise, behavior of the control parameters obtained by using (33), (34), (35), and (36) is shown. The speech utterance (*/iy/ − /ix/*) contains a voiced phoneme followed by another voiced phoneme [24]. Here $k_0 = 1,000$, $M = 100$, $N_f = 1002$, sampling frequency 16 kHz and SNR = 15 db is used.

In a similar fashion, in Figure 5, a speech utterance consisting of a voiced phoneme /ih/ followed by an unvoiced phoneme /sh/ and, in Figure 6, a voiced phoneme /ih/ followed by pause are considered. It is observed that the characteristic parameters vary depending on the nature of reference and current frames. When the current frame

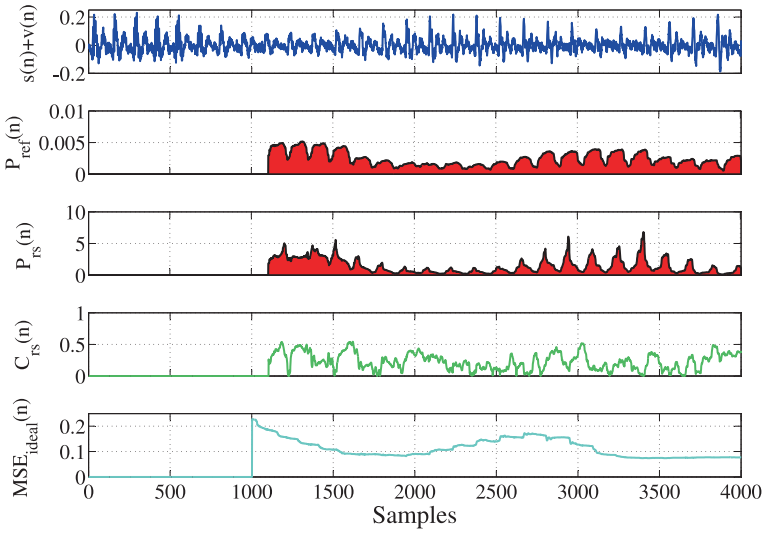


Figure 4 Characteristics of controlling factors - a voiced phoneme followed by another voiced frame.

is a pause or weakly unvoiced, the power ratio becomes higher in comparison to the case when the current frame is a voiced one. On the contrary, the correlation coefficient becomes smaller when measured between a voiced and an unvoiced frame, but it becomes quite larger when measured between two voiced frames. It is also found that the presence of voiced frame as a reference strongly governs the rate of convergence and the estimation error of the proposed LMS algorithm. In Figure 4, because of all through presence of the voiced frame as the reference as well as the current frame, it is found that the convergence performance is not very satisfactory and the estimation error is relatively higher. On the other hand, in Figure 6, it is observed that when the current frame is pause, even in the presence of voiced reference frame, a very fast convergence is obtained with a little estimation error. In Figure 5, as the current frame is unvoiced instead of pause, a comparatively slower convergence is observed with higher estimation error.

Next, in Figures 7, 8, 9, the reference frame is considered unvoiced, and in Figures 10, 11, 12, it is considered pause. When the reference frame is considered unvoiced because of the existence of a little correlation between the current and reference frames, the convergence performance of the proposed LMS algorithm is found quite satisfactory irrespective of the power of the reference

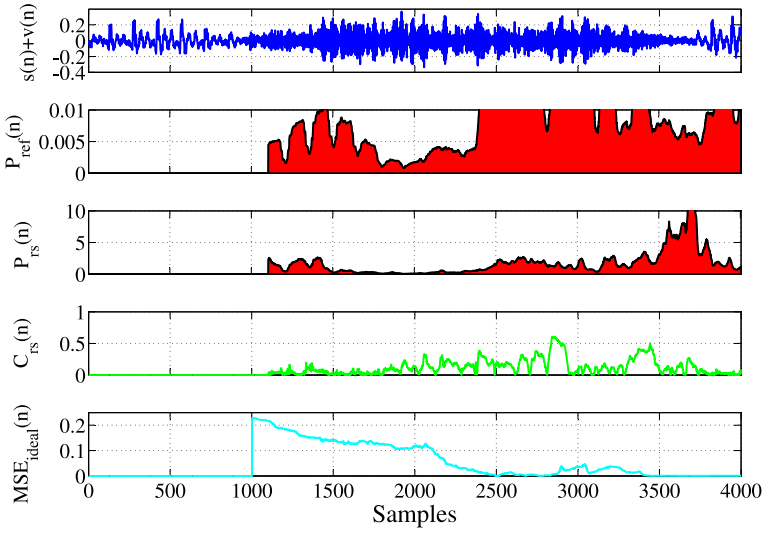


Figure 5 Characteristics of controlling factors - a voiced phoneme followed by an unvoiced phoneme.

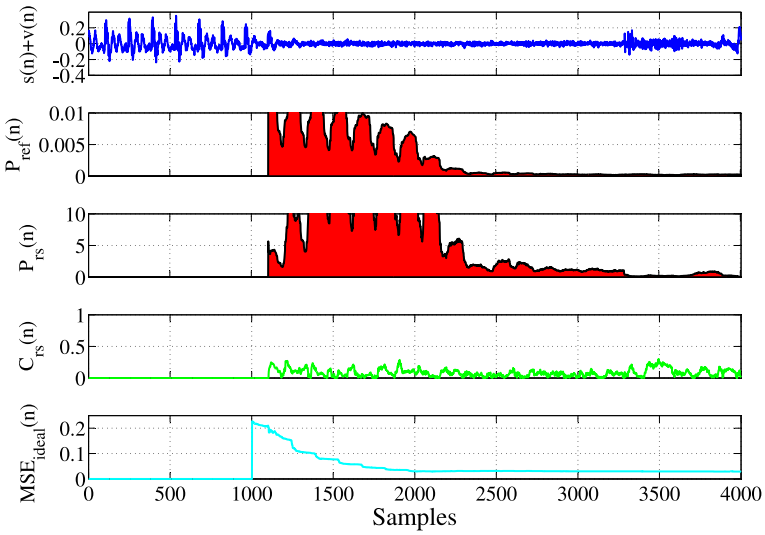


Figure 6 Characteristics of controlling factors - a voiced phoneme followed by pause frame.

signal (strong unvoiced or weakly unvoiced). In the case when the current frame is pause, no matter whether the reference frame is voiced or unvoiced, a fast convergence with high estimation accuracy is achieved using the proposed LMS algorithm. The reasons behind are (i) negligible cross-correlation between reference frame and current frame and (ii) a comparatively higher power ratio. In Figures 10, 11, 12, it is observed that even the reference frame is a pause or stop because of the presence of additive white noise, the reference frame may contain significant energy. In these cases, a reasonable estimation of the room response can be obtained given that the noise power is quite high. Findings in the above cases are summarized in Table 1.

First of all, it is observed that a better convergence in terms of iterations and estimation error is obtained when the current frame is a pause (P) or stop and the reference frame is either voiced (V) or unvoiced (U), namely, V-P and U-P. This fact leads to a decision that the updating needs to be carried out at high level of power ratio, i.e.,

$$P_{\mathrm{rs}}(n) = \frac{P_{\mathrm{ref}}(n)}{P_{\mathrm{sup}}(n)} \geq \zeta, \tag{37}$$

where $P_{\mathrm{ref}}(n)$ and $P_{\mathrm{sup}}(n)$ are defined in (33) and (34), respectively. If the value of the lower bound ζ is chosen too large, the updating would be postponed for most of the instances resulting in very slow convergence. On the other hand, a very small value of ζ may cause more frequent

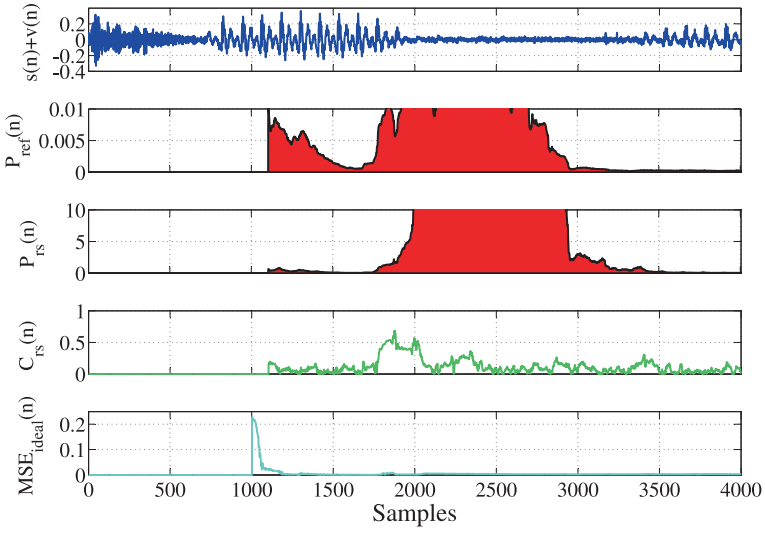


Figure 7 Characteristics of controlling factors - an unvoiced frame followed by a voiced frame.

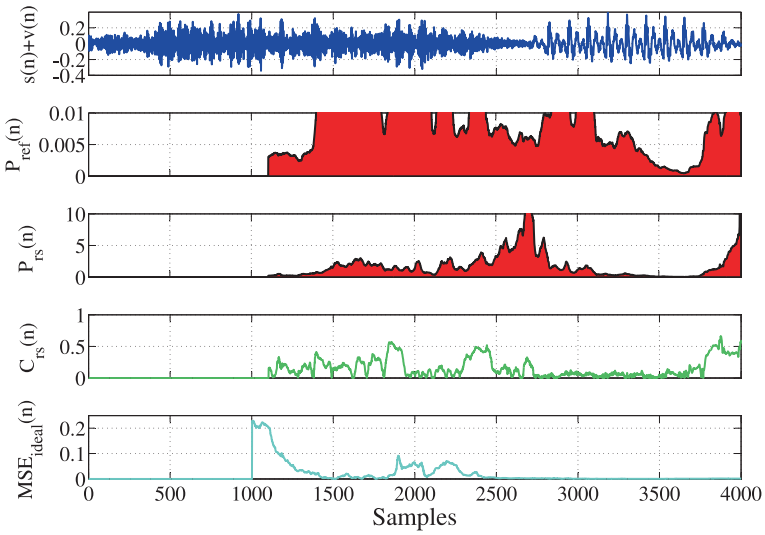


Figure 8 Characteristics of controlling factors - an unvoiced frame followed by another unvoiced frame.

updates where possibility of wrong estimations of filter coefficients would be higher, especially in V-P, U-P, and P-P cases. It is to be noted that considering only a lower bound of $P_{\mathrm{rs}}(n)$ may not always be sufficient to ensure that the reference frame possesses significant energy. For example in Figure 13, it is shown that high value of $P_{\mathrm{rs}}(n)$ may arise (marked block in the figure) from an initial silence frame where only a very little amount of noise is present. In order to prevent the updating in these situations, a lower bound β on the power of the reference frame is employed, i.e., $P_{\mathrm{ref}}(n) \geq \beta$. The value of β should surpass the power of speech pauses and ensure that the LMS update is postponed even if a frame of speech containing a partial pause is available as the reference. Hence, the first constraint for updating the algorithm is proposed as

Condition I: $P_{\mathrm{rs}}(n) \geq \zeta$ and $P_{\mathrm{ref}}(n) \geq \beta$.

In some cases, it is observed that though the power ratio is very small, quite satisfactory updating is obtained, such as the U-V case shown in Figure 7. Another characteristic observed here is lower value of correlation coefficient $C_{\mathrm{rs}}(n)$ with higher value of $P_{\mathrm{ref}}(n)$. It is to be mentioned that the proposed AEC algorithm is developed on the assumption of negligibility of the cross correlation between current frame and reference frame. However,

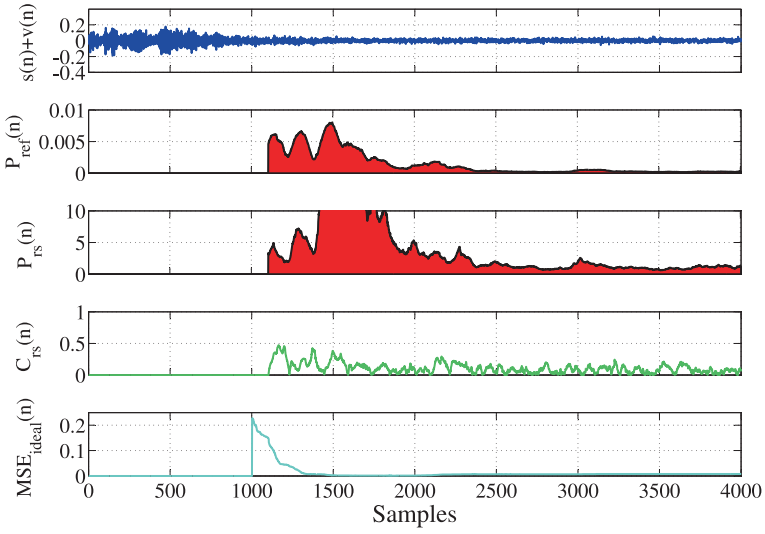


Figure 9 Characteristics of controlling factors - an unvoiced frame followed by a pause.

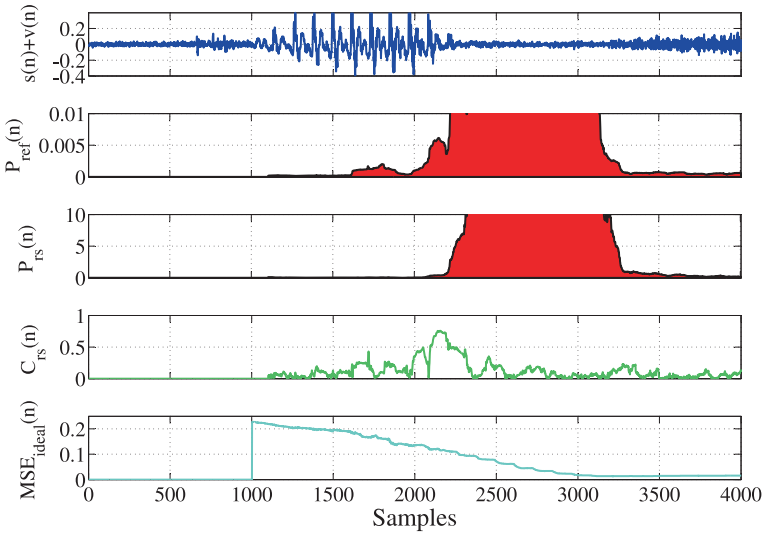


Figure 10 Characteristics of controlling factors - pause followed by a voiced frame.

since both reference and current frame may belong to the same person, in case of high degree of correlation, the adaptive algorithm would try to suppress portion from the echo-corrupted signal resulting in unusual degradation = in convergence performance. Hence, introducing an upper bound on $C_{\mathrm{rs}}(n)$, the second condition is proposed as

Condition II: $C_{\mathrm{rs}}(n) \leq \Upsilon 1$ and $P_{\mathrm{ref}} \geq \beta$.

The presence of a certain level of noise can be utilized as an advantage in pause instances where generally the updating is not performed. Since noise is considered uncorrelated to itself, updating at frames where only noise is present would be quite satisfactory. In this case, the value of $C_{\mathrm{rs}}(n)$ must be very small and thus another condition on updating is proposed as

Condition III: $C_{\mathrm{rs}}(n) \leq \Upsilon 2 \leq \Upsilon 1$.

Another important factor is the MSE of the estimations of successive iterations, which is defined as

$$e_{\mathrm{coeff}}(n) = \sum_{K=1}^{p} (\widehat{w}_n(k) - \widehat{w}_{n-1}(k))^2 / p. \tag{38}$$

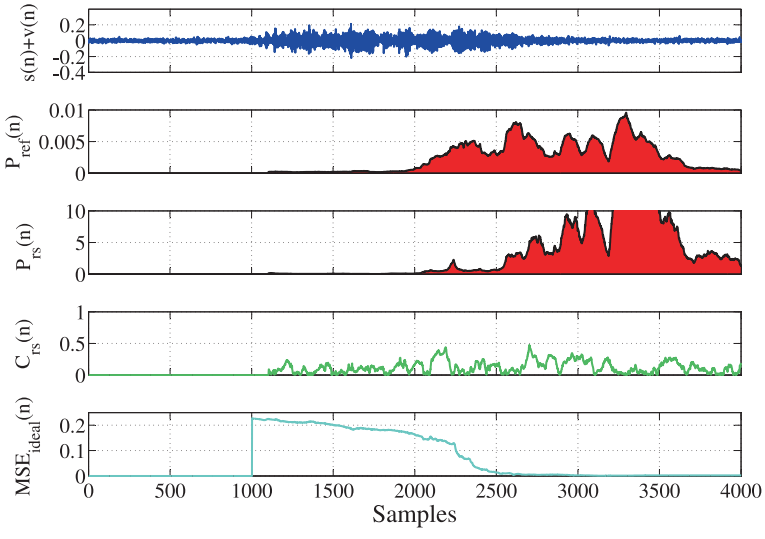


Figure 11 Characteristics of controlling factors - pause followed by an unvoiced frame.

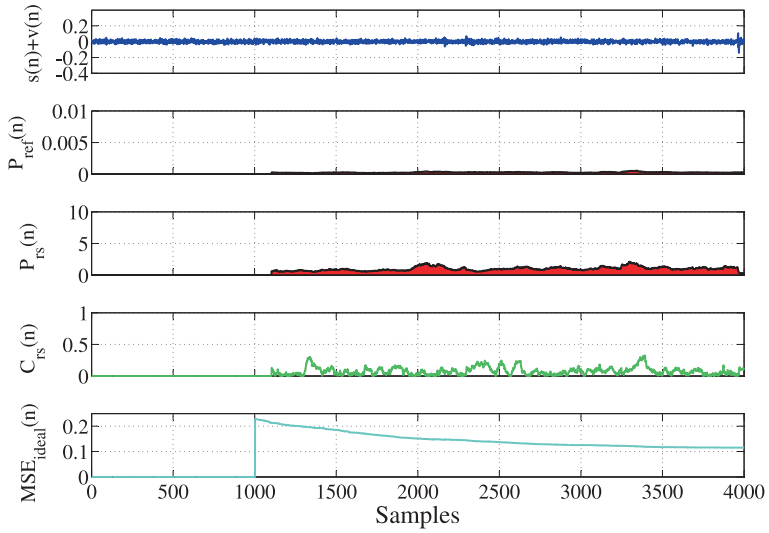


Figure 12 Characteristics of controlling factors - pause followed by another pause.

In order to continue the updating, an upper bound on the variation of successive estimates is set as following condition:

Condition IV: $e_{coeff}(n) \leq \aleph$.

Considering smaller values of $e_{\text{coeff}}(n)$ allows to avoid updating at those instances where abrupt and significant changes occur in the estimated coefficients. In the proposed method, in order to carry out the LMS update, at least one of the above four conditions must be fulfilled.

5 Simulation results and comments

Performance of the proposed algorithm is investigated in different echo-generating environments at various input noise levels considering several male and female utterances available in the TIMIT database [24]. An acoustic room environment is simulated using an FIR filter of length N_f, where as per conventional approaches, filter coefficients during the flat delay portion are assumed to be zero. The flat delay time (k_0) can be pre-calculated based on the distance between the microphone and the speaker [25]. Because of the implicit zeros corresponding to the flat delay, it is evident that a few number ($N_f - k_0$) of unknown coefficients has to be determined. In the proposed method, a smaller step size is used to obtain a smooth convergence.

Table 1 Variation of LMS updating performance due to various characteristics of reference and current speech frame

Reference speech sample	Current noise- and echo-corrupted speech sample	LMS update performance
Voiced	Voiced	Poor
Voiced	Unvoiced	Unsatisfactory
Voiced	Pause	Satisfactory/Excellent
Unvoiced	Voiced	Excellent
Unvoiced	Unvoiced	Excellent
Unvoiced	Pause	Excellent
Pause	Voiced	Poor
Pause	Unvoiced	Poor
Pause	Pause	Poor

First, a subjective evaluation is carried out based on the feedback about the quality of the echo- and noise-suppressed signal provided by five individual listeners at different noisy echo-generating environments. From the overall response of the listeners in terms of mean objective score (MOS), a very satisfactory performance of the proposed method is obtained even under severe echo-generating conditions in noise.

Next, two objective measures, namely, echo return loss enhancement (ERLE) and signal-to-distortion ratio (SDR) are employed. The ERLE is defined as the ratio of the instantaneous power of the residual echo signal $\eta_\varsigma(n)$ and that of the input echo signal $\eta_x(n)$ and expressed in dB as [1]

$$\text{ERLE}(n) = -10\log\frac{\eta_\varsigma(n)}{\eta_x(n)}. \tag{39}$$

The average value of ERLE(n) over time is considered. The input and output SDRs in dB are respectively defined as

$$\text{SDR}_{\text{in}} = 10\log\frac{P_s}{P_{x+v}} \tag{40}$$

$$\text{SDR}_{\text{out}} = 10\log\frac{P_s}{P_{\widehat{s+v}-s}}, \tag{41}$$

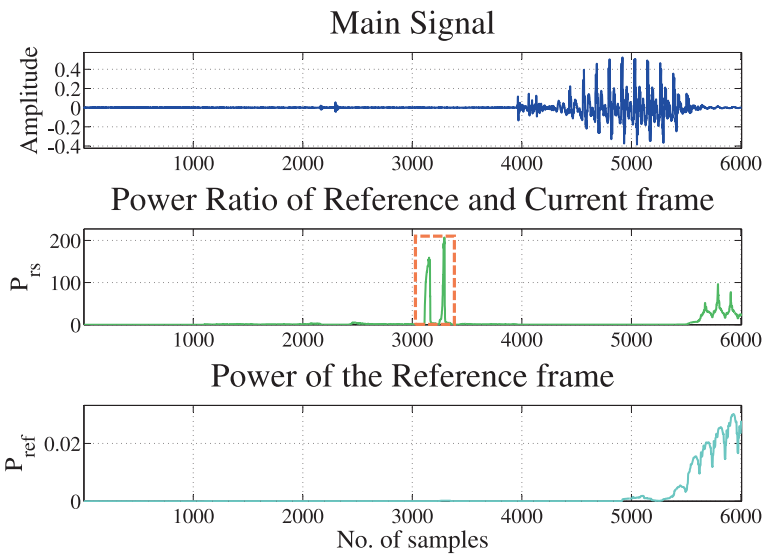


Figure 13 Example of high power ratio during initial silence frame.

where P_s is the power of original signal $s(n)$, P_{x+v} is the power of microphone input, and $P_{\widehat{s+v}-s}(n)$ is the power of distortion present in the echo-suppressed output signal. The SDR improvement is given by

$$\text{SDRI} = \text{SDR}_{\text{out}} - \text{SDR}_{\text{in}}, \tag{42}$$

which indicates the overall distortion removal.

The proposed algorithm has been tested on several different sentences taken from the TIMIT database. In order to demonstrate the principle of selecting different threshold values required in the proposed updating constraints, as a typical example, a sample utterance 'Good service should be rewarded by big tips' is shown in Figure 14 [24]. Voicing decisions are marked in the figure as 'P' for pause, 'V' for voiced, and 'U' for unvoiced. Considering white Gaussian noise with SNR = 15 dB, $N_f = 1,002$, $k_0 = 1,000$, and $M = 100$ in Figure 14b,c,d,e, $P_{\text{rs}}(n)$, $P_{\text{ref}}(n)$, $C_{\text{rs}}(n)$, and $\text{MSE}_{\text{ideal}}(n)$ are shown, respectively. Note that

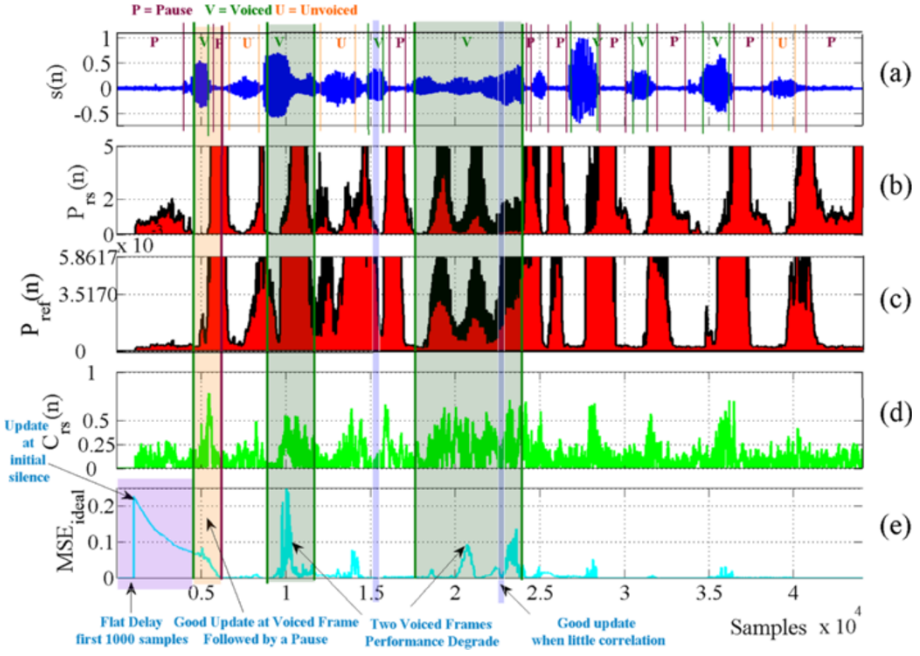


Figure 14 Plots of (a) utterance $s(n)$ and update parameters (b) $P_{rs}(n)$, (c) $P_{ref}(n)$, (d) $C_{rs}(n)$, and (e) MSE (without using constraint).

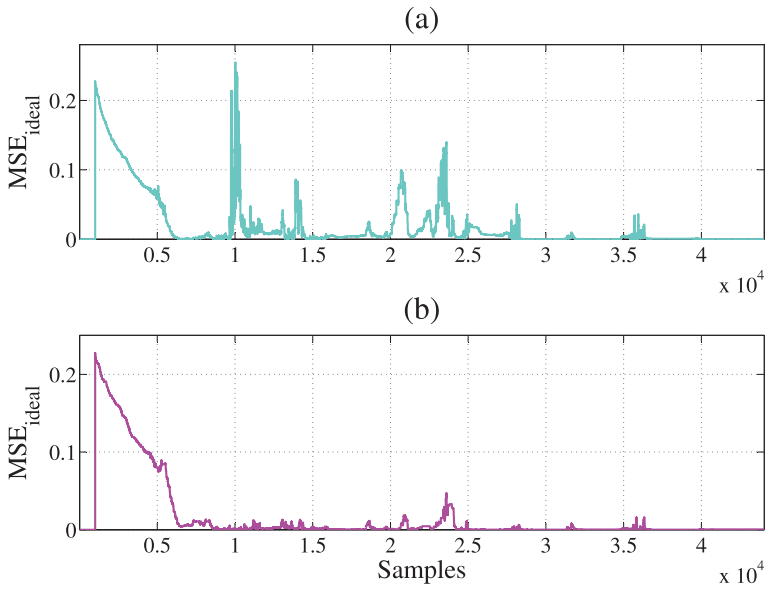


Figure 15 MSEs for the utterance shown in Figure 14. (a) Without conditions and **(b)** with conditions.

in this case, the proposed algorithm is used without the update constraints, and thus, the $\text{MSE}_{\text{ideal}}(n)$ exhibits some higher values. The comments provided in Table 1 can be better visualized from different marked zones of this figure. From extensive experimentations, it is found that a better update requires $P_{\text{ref}}(n)$ to be at least twice of $P_{\text{supp}}(n)$ and a small percentage (1% to 5%) of the power of a regular voiced frame can be chosen as the lower bound of β for $P_{\text{ref}}(n)$. Analyzing $C_{\text{rs}}(n)$ in different speech frames, $\Upsilon 1$ in condition 2 is chosen as 0.25 to ensure that no speech is being suppressed during the update procedure by confusing it with the echo and $\Upsilon 2$ is kept very small, i.e, $\Upsilon 2 \approx 0.1$ to allow updating for cases where there exists no correlation or extremely low correlation

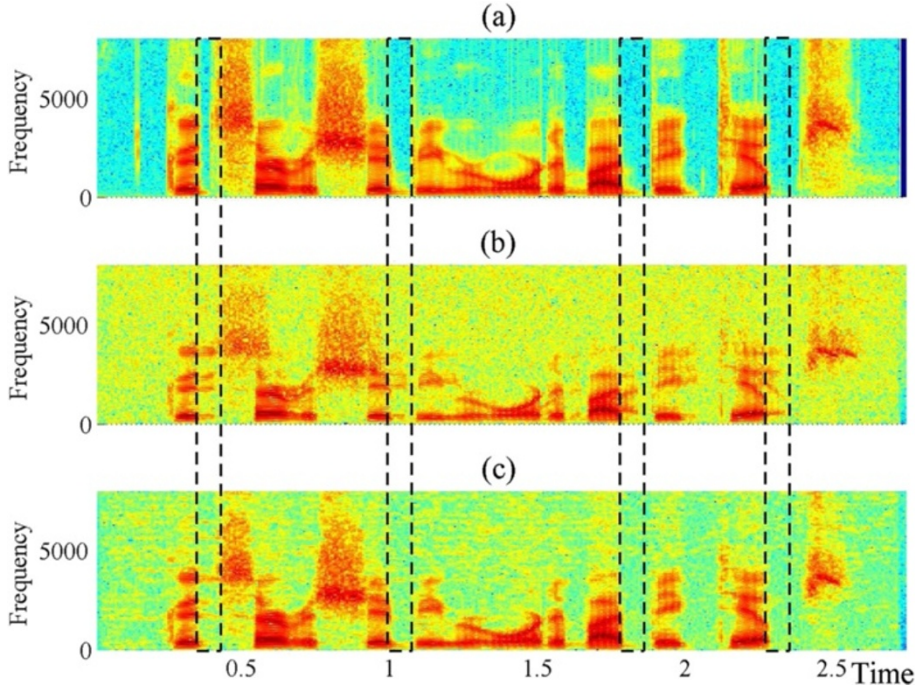


Figure 16 Spectrograms of (a) original signal, (b) echo- and noise-corrupted input and (c) enhanced output.

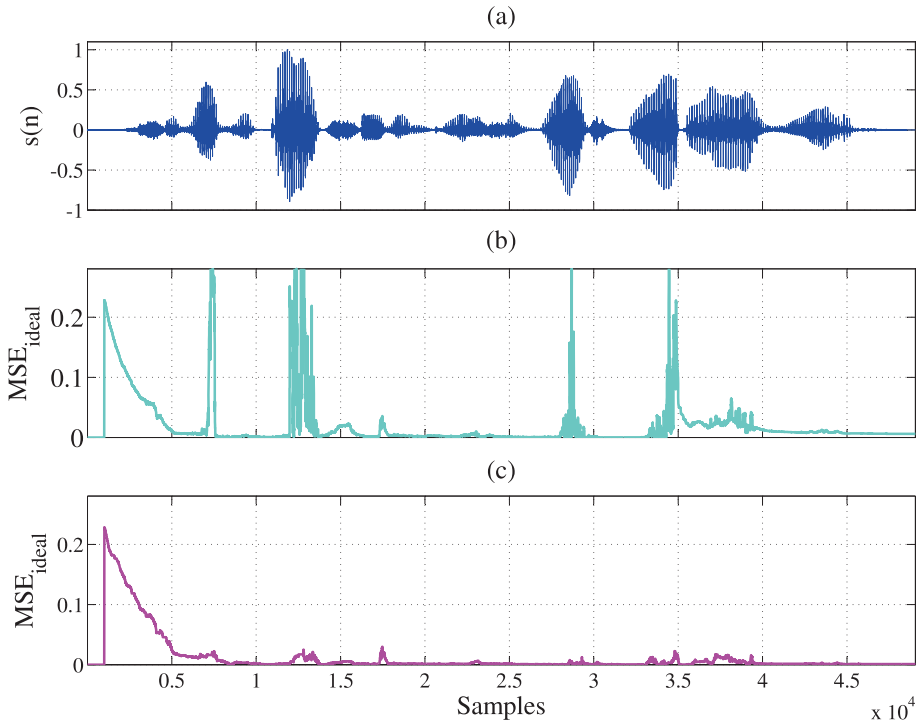


Figure 17 Another TIMIT utterance (a). MSEs of LMS estimations: **(b)** without conditions and **(c)** with conditions.

between the reference signal and echo-suppressed signal. The value of the threshold $\aleph$ for $e_{\text{coeff}}(n)$ in condition IV is chosen to be very small (0.7×10^{-4}) such that there will be no update of the LMS algorithm when the magnitude of $e_{\text{coeff}}(n)$ is comparatively much larger.

In Figure 15, the effect of incorporating the proposed conditions is shown. It is vividly observed from Figure 15 that by employing the proposed conditions, the convergence is improved to a greater extent. Moreover, in order to demonstrate the performance in frequency domain, spectrograms of the original signal, echo- and noise-corrupted signal, and the output of the proposed AENC block are depicted in Figure 16a,b, respectively. For convenience, some zones are marked on the spectrograms where significant reduction in echo and noise can easily be observed

In order to show the effectiveness of the proposed conditions, the $\text{MSE}_{\text{ideal}}(n)$ obtained in Figure 14e is redrawn in Figure 15. In Figure 15, the effect of incorporating the conditions is shown. It is vividly observed from Figure 15 that by employing the proposed conditions, the convergence is improved to a greater extent. Moreover, in order to demonstrate the performance in frequency domain, spectrograms of the original signal, echo- and noise-corrupted signal, and the output of the proposed AENC block are depicted in Figure 16a,b, respectively. For convenience, some zones are marked on the spectrograms where significant reduction in echo and noise can easily be observed. For a better understanding, another TIMIT utterance 'She had your dark suit in greasy wash water all year', under similar acoustic environment as used in Figure 14, is considered and corresponding echo- and noise-corrupted speech signal is shown in Figure 17a. The MSEs obtained by using the proposed method with and without the conditions are presented in Figure 17b,c, which clearly demonstrate the performance improvement in the later case.

In Table 2, the performance of the proposed algorithm with and without applying the conditions is shown in

Table 2 Performance comparison with varying room acoustics

	No condition		With conditions	
$N_f - k_0$	SDRI (dB)	Avg. ERLE (dB)	SDR (dB)	Avg. ERLE (dB)
2	4.9921	8.8496	6.9848	10.6772
4	4.9027	2.0696	5.7731	2.2787
6	8.391	4.6507	9.2744	5.0313
8	6.4551	2.4214	6.5558	2.6797
10	6.0507	2.6341	6.1730	2.854
12	6.7127	3.0277	7.0978	3.2048
14	7.8763	3.7481	8.2515	3.8909

Table 3 Performance comparison with noise level variation

	No condition		With conditions	
Input noise Level (dB)	SDRI (dB)	Avg. ERLE (dB)	SDR (dB)	Avg. ERLE (dB)
25	7.4065	3.183	7.8189	3.2759
20	7.613	3.5382	7.9346	3.6171
15	7.8763	3.7481	8.2515	3.8909
10	8.2085	3.5999	8.386	3.6064
5	8.2434	3.0533	8.8839	3.0765
0	8.7968	2.4493	9.4557	2.542
-5	8.2259	2.0032	10.5136	2.2912

terms of the SDR improvement (dB) and ERLE (dB) for utterance 1. In order to evaluate the performance under different room environments, length (N_f) and parameter values of the room response filter are varied while keeping the input SNR constant to 15 dB. Considering $k_0 = 1,000$, $N_f - k_0$ is varied from 2 to 14. Results shown in the table clearly demonstrate the effectiveness of using the conditions on performance measures; in all cases, higher values of SDR and ERLE are obtained.

In Table 3, the performance of the proposed algorithm with and without applying the conditions is evaluated for different levels of input SNR ranging from 25 to −5 dB for the first utterance considering white Gaussian noise and $N_f = 1014$. It can be seen that the proposed method provides satisfactory performance at all SNR levels. Especially, the use of proposed conditions exhibits comparatively better performance.

6 Conclusion

The problem of echo cancellation in the presence of noise, especially in single-channel environment, is a very challenging task, which has been efficiently tackled in this paper. First, the single-channel AEC block is designed based on the gradient-based adaptive LMS filter where to overcome the problem of getting a separate reference signal, we propose to use the delayed version of the echo-suppressed signal. Such a unique proposal of getting the reference signal is justified by presenting a detailed mathematical proof of achieving the most optimum Wiener-Hopf solution of the estimated filter coefficients, and a convergence analysis is carried out. Moreover, in order to achieve fast and smooth convergence, a set of updating constraints is proposed by analyzing the speech characteristics of different types of speech frames, such as voiced, unvoiced, and pause. In the ANC block, a modified single-channel spectral subtraction method is considered for its robust performance. It is shown that the proposed AENC scheme with updating constraints provides a very satisfactory performance in different echo-generating conditions and various levels of SNR in terms of SDR and ERLE.

Appendix

Derivation of the solution of the LMS update

In order to obtain a homogeneous solution of the update Eq. 22, one may consider

$$\underline{\widehat{\mathbf{w}}}_{n+1}^{T} = \underline{\widehat{\mathbf{w}}}_{n}^{T} - 2\mu \mathbf{R}_{(s+v)(s+v)}(n-k_0)\underline{\widehat{\mathbf{w}}}_{n}^{T}. \tag{43}$$

Eigenvalue decomposition of the correlation matrix $\mathbf{R}_{(s+v)(s+v)}(n-k_0)$ results in

$$\mathbf{R}_{(s+v)(s+v)}(n-k_0) = \mathbf{U}\Lambda\mathbf{U}^{T}, \tag{44}$$

where each column of the matrix $\mathbf{U}$ consists of eigenvectors corresponding to eigenvalues constituting the diagonal elements of the matrix Λ and $\mathbf{U}^{T}\mathbf{U} = \mathbf{I}$. Forward multiplication by $\mathbf{U}^{T}$ on both sides of (43) results in

$$\underline{\widehat{\mathbf{w}}}_{n+1}^{T^U} = \underline{\widehat{\mathbf{w}}}_{n}^{T^U} - 2\mu\Lambda\underline{\widehat{\mathbf{w}}}_{n}^{T^U}, \tag{45}$$

where $\mathbf{U}^{T}\underline{\widehat{\mathbf{w}}}_{n}^{T} = \underline{\widehat{\mathbf{w}}}_{n}^{T^U}$. The kth coefficient of the weight vector can be expressed as

$$\underline{\widehat{w}}_{n+1}^{U}(k) = (1 - 2\mu\lambda(k))\underline{\widehat{w}}_{n}^{U}(k), \tag{46}$$

where $\lambda(k)$ is the kth diagonal element of the eigenvalue matrix obtained by eigenvalue decomposition of $\mathbf{R}_{(s+v)(s+v)}(n-k_0)$. Hence, the homogeneous solution can be obtained as

$$\widehat{w}_{h.s} = C_k(1 - 2\mu\lambda(k))^{n}, \tag{47}$$

where C_k is a constant. Next, in order to obtain the particular solution for the kth coefficient, based on (22) one can get

$$\widehat{w}_{p.s} = \widehat{w}_{p.s} - 2\mu\lambda(k)\widehat{w}_{p.s} + 2\mu r^{U}(n-k_0-k). \tag{48}$$

Here, $r^{U}(n-k_0-k)$ is the kth element of $\mathbf{U}^{T}\mathbf{r}_{(x_s+x_v)(s+v)}(n-k_0) = \mathbf{r}^{U}_{(x_s+x_v)(s+v)}(n-k_0)$. For a particular solution $\widehat{w}_{p.s} = K_p r^{U}(n-k_0-k)$, (48) can be written as

$$\begin{aligned} K_p r^{U}(n-k_0-k) &= K_p r^{U}(n-k_0-k) \\ &\quad - 2\mu\lambda(k)K_p r^{U}(n-k_0-k) \\ &\quad + 2\mu r^{U}(n-k_0-k), \end{aligned} \tag{49}$$

which leads to $K_p = \frac{1}{\lambda(k)}$ and the particular solution

$$\widehat{w}_{p.s} = \frac{1}{\lambda(k)} r^{U}(n-k_0-k). \tag{50}$$

Competing interests

The authors declare that they have no competing interests.

Author details
[1]Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology, Dhaka 1000, Bangladesh. [2]Department of Electrical and Computer Engineering, Concordia University, Montreal, Quebec H3G 1M8, Canada.

Received: 12 November 2013 **Accepted:** 25 March 2014
Published: 3 May 2014

References

1. SV Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 2nd edn. (Wiley, Chichester, 2000)
2. SM Kuo, BH Lee, *Real-Time Digital Signal Processing*. (Wiley, 2001)
3. C Breining, P Dreiseitel, E Hänsler, A Mader, B Nitsch, H Puder, T Schertler, G Schmidt, J Tilp, Acoustic echo control - an application of very-high-order adaptive filters. IEEE Signal Process. Mag. **16**(4), 42–69 (1999)
4. E Hänsler, The hands-free telephone problem: an annotated bibliography. Signal Process. **27**(3), 259–271 (1992)
5. AWH Khong, PA Naylor, Stereophonic acoustic echo cancellation employing selective-tap adaptive algorithms. IEEE Trans. Audio, Speech, Lang. Process. **14**(3), 785–796 (2006)
6. F Lindstrom, C Schuldt, I Claesson, An improvement of the two-path algorithm transfer logic for acoustic echo cancellation. IEEE Trans. Audio, Speech, Lang. Process. **15**(4), 1320–1326 (2007)
7. S Wu, X Qiu, M Wu, Stereo acoustic echo cancellation employing frequency-domain preprocessing and adaptive filter. IEEE Trans. Audio, Speech, Lang. Process. **19**(3), 614–623 (2011)
8. R Nath, Adaptive echo cancellation based on a multipath model of acoustic channel. Circuits, Syst. Signal Process., Springer US. **32**(4), 1673–1698 (2013)
9. M Yukawa, RC de Lamare, R Sampaio-Neto, Efficient acoustic echo cancellation with reduced-rank adaptive filtering based on selective decimation and adaptive interpolation. IEEE Trans. Audio, Speech, Lang. Process. **16**(4), 696–710 (2008)
10. E Hänsler, G Schmidt, *Acoustic Echo and Noise Control: a Practical Approach*. (Wiley, New York, 2004)
11. V Myllylä, Residual echo filter for enhanced acoustic echo control. Signal Process. **86**(6), 1193–1205 (2006)
12. R Topa, I Muresan, BS Kirei, I Homana, A digital adaptive echo-canceller for room acoustics improvement. Adv. Electrical Comput. Eng. **10**, 450–453 (2004)
13. S Haykin, *Adaptive Filter Theory*, 3rd edn. (Prentice-Hall, Inc., Upper Saddle River, NJ, 1996)
14. G Schmidt, Applications of acoustic echo control: an overview, in *Proc. Eur. Signal Process. Conf.* (EUSIPCO Vienna, 2004), pp. 9–16
15. B Widrow, JRJ Glover, JM McCool, J Kaunitz, CS Williams, RH Hearn, JR Zeidler, JE Dong, RC Goodlin, Adaptive noise cancelling: principles and applications. Proc. IEEE. **63**(12), 1692–1716 (1975)
16. H Yasukawa, An acoustic echo canceller with sub-band noise cancelling. IEICE Trans. Fundamentals Electron. Commun. Comput. Sci. **E75–A**(11), 1516–1523 (1992)
17. SJ Park, CG Cho, C Lee, DH Youn, Integrated echo and noise canceller for hands-free applications. IEEE Trans. Circuits Syst.-II: Analog Digital Signal Process. **49**(3) (2002)
18. C Beaugeant, V Turbin, P Scalart, A Gilloire, New optimal filtering approaches for hands-free telecommunication terminals. Signal Process. **64**(1), 33–47 (1998)
19. U Mahbub, SA Fattah, Gradient based adaptive filter algorithm for single channel acoustic echo cancellation in noise, in *Proc. Int. Conf. Electrical Computer Engineering (ICECE), 2012 7th International Conference On* (Dhaka, 688 Bangladesh, 2012), pp. 880–883
20. S Boll, A spectral subtraction algorithm for suppression of acoustic noise in speech, in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process. (ICASSP) '79*, (1979), pp. 200–203
21. M Berouti, R Schwartz, J Makhoul, Enhancement of speech corrupted by acoustic noise. IEEE Conf. Acoust. Speech Signal Process. (ICASSP), 208–211 (1979)
22. JS Lim, Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise. IEEE Trans. Acoust. Speech Signal Process. **26**(5), 471–472 (1978)
23. R Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**(5), 504–512 (2001)
24. JS Garofolo, LF Lamel, WM Fisher, JG Fiscus, DS Pallett, NL Dahlgren, V Zue, *Timit acoustic-phonetic continuous speech corpus*. (Linguistic Data Consortium, Philadelphia, 1993)
25. F Guangzeng, L Feng, A new echo caneller with the estimation of flat delay, in *IEEE Region Ten Conf. TENCON 92* (Melbourne, Australia, 1992). vol. 1, pp. 1–5, Print ISBN 0-7803-0849-2, DOI- 10.1109/TENCON.1992.271995

doi:10.1186/1687-4722-2014-20
Cite this article as: Mahbub *et al.*: **Single-channel acoustic echo cancellation in noise based on gradient-based adaptive filtering.** *EURASIP Journal on Audio, Speech, and Music Processing* 2014 **2014**:20.