A STUDY OF THE MVDR FILTER FOR ACOUSTIC ECHO SUPPRESSION

Hai Huang[1]*, Jacob Benesty*[2]*, Jingdong Chen*[1]*, Karim Helwani*[3]*, and Herbert Buchner*[4]

[1]: Northwestern Polytechnical University
127 Youyi West Road
Xi'an, Shaanxi 710072, China

[2]: INRS-EMT, University of Quebec
800 de la Gauchetiere Ouest, Suite 6900
Montreal, QC H5A 1K6, Canada

[3]: Telekom Innovation Laboratories
Ernst-Reuter-Platz 7
10587 Berlin, Germany

[4]: Technische Universität Berlin
Franklinstr. 28/29,
10587 Berlin, Germany

ABSTRACT

This paper studies an echo suppression approach to reducing the undesired echoes that result from the acoustic coupling between a loudspeaker and a microphone in duplex voice communication. The approach consists of four basic steps. First, both the loudspeaker and microphone signals are partitioned into small overlapping frames. Second, each frame is transformed into the short-time Fourier transform (STFT) domain. Third, a minimum variance distortionless response (MVDR) filter is designed in each subband by explicitly using the interframe signal correlation. This MVDR filter is then used to estimate the echo signal and the obtained estimate is subsequently subtracted from the microphone signal. Finally, the time-domain processed signal is constructed using the overlap-add technique with the inverse STFT. Experiments are performed and the results demonstrate that this proposed method can achieve significant amount of echo suppression in practical room environments.

Index Terms— Acoustic echoes, echo suppression, minimum variance distortionless response (MVDR) filter.

1. INTRODUCTION

How to control the detrimental acoustic echo effect in voice communication systems such as audio bridging and teleconferencing has become a more and more important problem as hands-free mode is more often used. In the literature, two fundamental approaches to reducing or eliminating the undesired echoes were developed, i.e., acoustic echo cancellation (AEC) and acoustic echo suppression (AES).

AEC assumes that the acoustic echo path from the loudspeaker to the microphone is linear and can be modeled with a finite-impulse-response (FIR) filter. Echo cancellation is then achieved by adaptively identifying the echo path impulse response and subtracting an echo estimate from the microphone signal [1]–[7]. Since it does not apply any filter to the microphone signal, this approach cancels echo without introducing any distortion to the desired near-end signal. As a result, AEC has the potential to achieve high-fidelity and full-duplex voice communication. However, if the adaptive filter does not converge to the true echo impulse response (which happens often in practice) or there is some nonlinearity in the echo path, an AEC may not cancel the echo completely, leaving some residual echo that can be unpleasant to listen, thereby affecting the quality of voice communication.

In comparison, AES directly applies a suppression filter to the microphone signal to attenuate echo without explicitly modeling the acoustic echo path [8]–[16]. AES can be either used in conjunction with AEC to further attenuate the residual echo after AEC (in this case, it is generally called residual echo suppressor) or it can be used in an independent way to deal with the echo problem. The advantage of this method is that it is generally robust to nonlinear effect as well as small changes in the echo path. The disadvantage of AES as compared to AEC is that some distortion to the desired near-end signal is generally unavoidable since the suppression filter affects the echo and near-end signals at the same time. Therefore, careful attention has to be paid to the design of the suppression filter in order to significantly reduce the echo signal without much distorting the near-end speech.

In this paper, we present an approach to echo suppression in the short-time Fourier transform (STFT) domain, which consists of four basic steps. First, both the loudspeaker and microphone signals are partitioned into small overlapping frames with a frame length ranging from a few to a few tens of milliseconds. Second, each frame is transformed into the STFT domain. Third, a minimum variance distortionless response (MVDR) filter is designed to estimate the echo signal and subtract it from the microphone signal in each subband. Finally, the time-domain processed signal is constructed using the overlap-add technique with the inverse STFT. Like the traditional AES methods, this presented approach achieves echo reduction by filtering the microphone signal. However, it differs from the traditional techniques in that it attempts to obtain an echo estimate and then subtract this estimate from the microphone signal. In this sense, it also resembles, in a certain degree, the AEC technique.

2. SIGNAL MODEL AND PROBLEM FORMULATION

Let us consider the conventional signal model in which acoustic echoes are generated from the coupling between a loudspeaker and a microphone [1]. The microphone signal at the time index t can be written as

$$d(t) = g(t) * x(t) + u(t) = y(t) + u(t), \tag{1}$$

where $x(t)$ is the loudspeaker (or far-end) signal, $g(t)$ is the impulse response from the loudspeaker to the microphone, $u(t)$ is the near-end signal, and $y(t)$ is the echo signal. We assume that $y(t)$ and $u(t)$ are uncorrelated. All signals are considered to be real, zero mean, and broadband. Our objective is to estimate the echo, $y(t)$, given the far-end (or input) signal, $x(t)$, and the microphone (or output) signal, $d(t)$. When this echo is correctly estimated, it can be subtracted from the output signal to get an estimate of the near-end signal, $u(t)$, which can then be transmitted to the far-end room.

Using the STFT, the signal model given in (1) can be expressed in the time-frequency domain as

$$D(k, n) = Y(k, n) + U(k, n), \tag{2}$$

Effort of the first and third authors is partially supported by the Anhui Science and Technology Project (11010202191).

where $D(k,n)$, $Y(k,n)$, and $U(k,n)$ are the STFTs of $d(t)$, $y(t)$, and $u(t)$, respectively, at the frequency bin $k \in \{0,1,\ldots,K-1\}$ and the time frame n. A bit later, the approximation:

$$Y(k,n) \approx G(k)X(k,n) \tag{3}$$

will be used, where $G(k)$ and $X(k,n)$ are the STFTs of $g(t)$ and $x(t)$, respectively. Since $y(t)$ and $u(t)$ are uncorrelated by assumption, the variance of $D(k,n)$ is

$$\phi_D(k,n) = E\left[|D(k,n)|^2\right] = \phi_Y(k,n) + \phi_U(k,n), \tag{4}$$

where $E[\cdot]$ denotes mathematical expectation, and $\phi_Y(k,n) = E\left[|Y(k,n)|^2\right]$ and $\phi_U(k,n) = E\left[|U(k,n)|^2\right]$ are the variances of $Y(k,n)$ and $U(k,n)$, respectively.

We propose to estimate the echo signal by applying an FIR filter to the microphone signal at different time frames [17], i.e.,

$$\begin{aligned}\widehat{Y}(k,n) &= \sum_{l=0}^{L-1} H_l^*(k,n)D(k,n-l)\\ &= \mathbf{h}^H(k,n)\mathbf{d}(k,n),\end{aligned} \tag{5}$$

where $\widehat{Y}(k,n)$ is supposed to be the estimate of $Y(k,n)$, the superscripts * and ^{H} are the complex-conjugation and transpose-conjugation operators, respectively, L is the number of consecutive time frames,

$$\begin{aligned}\mathbf{h}(k,n) &= \begin{bmatrix} H_0(k,n) & \cdots & H_{L-1}(k,n) \end{bmatrix}^T,\\ \mathbf{d}(k,n) &= \begin{bmatrix} D(k,n) & \cdots & D(k,n-L+1) \end{bmatrix}^T\end{aligned}$$

are vectors of length L, and the superscript ^{T} denotes transposition. We can rewrite (5) as

$$\begin{aligned}\widehat{Y}(k,n) &= \mathbf{h}^H(k,n)\mathbf{y}(k,n) + \mathbf{h}^H(k,n)\mathbf{u}(k,n)\\ &= Y_{\mathrm{f}}(k,n) + U_{\mathrm{rn}}(k,n),\end{aligned} \tag{6}$$

where $\mathbf{y}(k,n)$ and $\mathbf{u}(k,n)$ are defined in a similar way to $\mathbf{d}(k,n)$,

$$Y_{\mathrm{f}}(k,n) = \mathbf{h}^H(k,n)\mathbf{y}(k,n) \tag{7}$$

is a filtered version of the echo signal at L consecutive time frames, and

$$U_{\mathrm{rn}}(k,n) = \mathbf{h}^H(k,n)\mathbf{u}(k,n) \tag{8}$$

is the residual near-end signal, which is incoherent with $Y_{\mathrm{f}}(k,n)$.

At the time frame n, the echo signal is $Y(k,n)$, which needs to be estimated. To achieve such an estimation, we decompose the vector $\mathbf{y}(k,n)$ into two orthogonal components: one coherent and the other incoherent with the echo signal, i.e.,

$$\begin{aligned}\mathbf{y}(k,n) &= Y(k,n)\boldsymbol{\gamma}_Y(k,n) + \mathbf{y}_{\mathrm{i}}(k,n)\\ &= \mathbf{y}_{\mathrm{c}}(k,n) + \mathbf{y}_{\mathrm{i}}(k,n),\end{aligned} \tag{9}$$

where

$$\mathbf{y}_{\mathrm{c}}(k,n) = Y(k,n)\boldsymbol{\gamma}_Y(k,n) \tag{10}$$

is the coherent echo signal vector,

$$\mathbf{y}_{\mathrm{i}}(k,n) = \begin{bmatrix} Y_{\mathrm{i}}(k,n) & \cdots & Y_{\mathrm{i}}(k,n-L+1) \end{bmatrix}^T$$

is the incoherent echo signal vector that satisfies

$$E\left[\mathbf{y}_{\mathrm{i}}(k,n)Y^*(k,n)\right] = \mathbf{0}, \tag{11}$$

and

$$\boldsymbol{\gamma}_Y(k,n) = \frac{E\left[\mathbf{y}(k,n)Y^*(k,n)\right]}{\phi_Y(k,n)} \tag{12}$$

is the (normalized) interframe correlation vector. By using (3) in (12), it is easy to see that

$$\boldsymbol{\gamma}_Y(k,n) = \boldsymbol{\gamma}_X(k,n) = \frac{E\left[\mathbf{x}(k,n)X^*(k,n)\right]}{\phi_X(k,n)}, \tag{13}$$

where $\phi_X(k,n)$ is the variance of $X(k,n)$. This form is of great importance in practice since now the correlation vector $\boldsymbol{\gamma}_Y(k,n)$ that is needed to estimate the echo suppression filter can easily be estimated from the loudspeaker signal.

Substituting (9) into (6), we get

$$\widehat{Y}(k,n) = Y_{\mathrm{fe}}(k,n) + Y_{\mathrm{ri}}(k,n) + U_{\mathrm{rn}}(k,n), \tag{14}$$

where

$$Y_{\mathrm{fe}}(k,n) = Y(k,n)\mathbf{h}^H(k,n)\boldsymbol{\gamma}_Y(k,n) \tag{15}$$

is the filtered echo signal and

$$Y_{\mathrm{ri}}(k,n) = \mathbf{h}^H(k,n)\mathbf{y}_{\mathrm{i}}(k,n) \tag{16}$$

is the residual incoherent echo signal. We observe that the estimate of the echo signal is the sum of three terms that are mutually incoherent. Therefore, the variance of $\widehat{Y}(k,n)$ is

$$\phi_{\widehat{Y}}(k,n) = \phi_{Y_{\mathrm{fe}}}(k,n) + \phi_{Y_{\mathrm{ri}}}(k,n) + \phi_{U_{\mathrm{rn}}}(k,n), \tag{17}$$

where

$$\begin{aligned}\phi_{Y_{\mathrm{fe}}}(k,n) &= \phi_Y(k,n)\left|\mathbf{h}^H(k,n)\boldsymbol{\gamma}_Y(k,n)\right|^2\\ &= \mathbf{h}^H(k,n)\boldsymbol{\Phi}_{\mathbf{y}_{\mathrm{c}}}(k,n)\mathbf{h}(k,n),\end{aligned} \tag{18}$$

$$\begin{aligned}\phi_{Y_{\mathrm{ri}}}(k,n) &= \mathbf{h}^H(k,n)\boldsymbol{\Phi}_{\mathbf{y}_{\mathrm{i}}}(k,n)\mathbf{h}(k,n)\\ &= \mathbf{h}^H(k,n)\boldsymbol{\Phi}_{\mathbf{y}}(k,n)\mathbf{h}(k,n) -\\ &\qquad \phi_Y(k,n)\left|\mathbf{h}^H(k,n)\boldsymbol{\gamma}_Y(k,n)\right|^2,\end{aligned} \tag{19}$$

$$\phi_{U_{\mathrm{rn}}}(k,n) = \mathbf{h}^H(k,n)\boldsymbol{\Phi}_{\mathbf{u}}(k,n)\mathbf{h}(k,n), \tag{20}$$

$\boldsymbol{\Phi}_{\mathbf{y}_{\mathrm{c}}}(k,n) = \phi_Y(k,n)\boldsymbol{\gamma}_Y(k,n)\boldsymbol{\gamma}_Y^H(k,n)$ is the correlation matrix (whose rank is equal to 1) of $\mathbf{y}_{\mathrm{c}}(k,n)$, and $\boldsymbol{\Phi}_{\mathbf{z}}(k,n) = E\left[\mathbf{z}(k,n)\mathbf{z}^H(k,n)\right]$ is the correlation matrix of $\mathbf{z}(k,n) \in \{\mathbf{y}(k,n), \mathbf{y}_{\mathrm{i}}(k,n), \mathbf{u}(k,n)\}$.

3. THE MVDR FILTER FOR ECHO SUPPRESSION

In this section, we show how to derive the MVDR filter for acoustic echo suppression. This optimal filter is similar to the one derived in [18] for single-channel noise reduction.

We define the subband error signal between the echo and its estimate as

$$\begin{aligned}\mathcal{E}(k,n) &= Y(k,n) - \widehat{Y}(k,n)\\ &= D(k,n) - U(k,n) - \widehat{Y}(k,n)\\ &= \widehat{U}(k,n) - U(k,n),\end{aligned} \tag{21}$$

where

$$\widehat{U}(k,n) = D(k,n) - \widehat{Y}(k,n) \tag{22}$$

is the estimate of the near-end signal that will be transmitted to the far-end room. Clearly, this estimate is obtained from the estimate of the echo signal. The subband mean-square error (MSE) is then

$$\begin{aligned} J\left[\mathbf{h}(k,n)\right] &= E\left[|\mathcal{E}(k,n)|^2\right] \\ &= \phi_Y(k,n) + \mathbf{h}^H(k,n)\mathbf{\Phi}_{\mathbf{d}}(k,n)\mathbf{h}(k,n) \\ &\quad - \phi_Y(k,n)\mathbf{h}^H(k,n)\boldsymbol{\gamma}_Y(k,n) \\ &\quad - \phi_Y(k,n)\boldsymbol{\gamma}_Y^H(k,n)\mathbf{h}(k,n), \end{aligned} \tag{23}$$

where $\mathbf{\Phi}_{\mathbf{d}}(k,n) = E\left[\mathbf{d}(k,n)\mathbf{d}^H(k,n)\right]$ is the correlation matrix of $\mathbf{d}(k,n)$.

It is preferable to estimate the echo with no distortion such that it is correctly subtracted from the microphone signal. This can be achieved by minimizing $J\left[\mathbf{h}(k,n)\right]$ with the constraint:

$$\mathbf{h}^H(k,n)\boldsymbol{\gamma}_Y(k,n) = 1. \tag{24}$$

Solving the above constrained problem, we find the MVDR filter:

$$\begin{aligned} \mathbf{h}_{\mathrm{MVDR}}(k,n) &= \frac{\mathbf{\Phi}_{\mathbf{d}}^{-1}(k,n)\boldsymbol{\gamma}_Y(k,n)}{\boldsymbol{\gamma}_Y^H(k,n)\mathbf{\Phi}_{\mathbf{d}}^{-1}(k,n)\boldsymbol{\gamma}_Y(k,n)} \\ &= \frac{\mathbf{\Phi}_{\mathbf{d}}^{-1}(k,n)\boldsymbol{\gamma}_X(k,n)}{\boldsymbol{\gamma}_X^H(k,n)\mathbf{\Phi}_{\mathbf{d}}^{-1}(k,n)\boldsymbol{\gamma}_X(k,n)}, \end{aligned} \tag{25}$$

The statistics forming this filter are easy to estimate since the microphone and loudspeaker signals, i.e., $D(k,n)$ and $X(k,n)$, are available.

4. PERFORMANCE MEASURES

The two most important means to evaluate the acoustic echo suppression performance are: 1) the attenuation of the acoustic echo and 2) the distortion of the near-end signal.

To evaluate the amount of echo attenuation, we can examine the so-called echo-return loss enhancement (ERLE) [2], [12], which can be defined in our scenario as

$$\xi_{\mathrm{ERLE}}(t) = \frac{\mathrm{LPF}\{|y(t)|^2\}}{\mathrm{LPF}\{|y(t) - \widehat{y}(t)|^2\}}, \tag{26}$$

where $\widehat{y}(t)$ is the time-domain signal reconstructed from $\widehat{Y}(k,n)$, and LPF$\{\cdot\}$ denotes a lowpass filter operation.

We can also assess the level of echo attenuation in the STFT domain by examining the subband and fullband acoustic echo reduction factors at the time frame n, which are defined as

$$\xi\left[\mathbf{h}(k,n)\right] = \frac{\phi_Y(k,n)}{E\left[|Y(k,n) - \mathbf{h}^H(k,n)\mathbf{y}(k,n)|^2\right]} \tag{27}$$

and

$$\xi\left[\mathbf{h}(:,n)\right] = \frac{\sum_{k=0}^{K-1}\phi_Y(k,n)}{\sum_{k=0}^{K-1} E\left[|Y(k,n) - \mathbf{h}^H(k,n)\mathbf{y}(k,n)|^2\right]}. \tag{28}$$

The acoustic echo reduction factors with an optimal echo suppression filter should be greater than or equal to 1. If $\xi = 1$, there is no echo reduction. The higher is the value of ξ, the more the echo is reduced. The definition of $\xi\left[\mathbf{h}(:,n)\right]$ is equivalent to the ERLE.

To evaluate the amount of distortion introduced by the suppression filter to the near-end signal, we can examine the so-called speech-distortion index [19], [20], which can be defined in our application as

$$\upsilon_{\mathrm{sd}}(t) = \frac{\mathrm{LPF}\{|u(t) - u_{\mathrm{rn}}(t)|^2\}}{\mathrm{LPF}\{|u(t)|^2\}}, \tag{29}$$

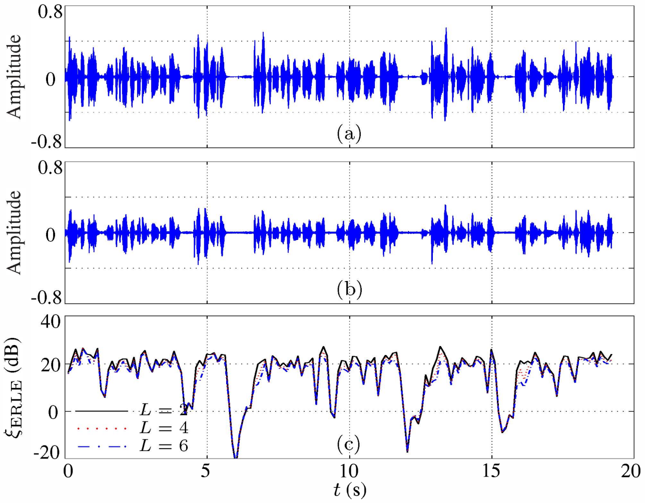

Fig. 1. Echo suppression performance: (a) waveform of the far-end speech, (b) waveform of the microphone signal, and (c) ERLE of the MVDR echo suppression filter for three different filter lengths, i.e., $L = 2, 4$, and 6.

where $u_{\mathrm{rn}}(t)$ is the time-domain signal reconstructed from $U_{\mathrm{rn}}(k,n)$. Alternatively, we can also assess the distortion of the near-end signal in the STFT domain, where we define the subband and fullband near-end signal distortion indices at the time frame n as

$$\upsilon\left[\mathbf{h}(k,n)\right] = \frac{E\left[|U(k,n) - U_{\mathrm{rn}}(k,n)|^2\right]}{\phi_U(k,n)} \tag{30}$$

and

$$\upsilon\left[\mathbf{h}(:,n)\right] = \frac{\sum_{k=0}^{K-1} E\left[|U(k,n) - U_{\mathrm{rn}}(k,n)|^2\right]}{\sum_{k=0}^{K-1}\phi_U(k,n)}. \tag{31}$$

The near-end signal distortion indices are always greater than or equal to 0. The higher their value, the more the near-end signal is distorted. Therefore, we want to keep these indices as small as possible.

5. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of the presented MVDR filter for echo suppression. Two sets of experiments were performed: one set concerns the performance in a scenario where there is no doubletalk while the other pertains to a case where there is doubletalk.

5.1. Echo Suppression Performance without Doubletalk

The first simulation studies the case where there is no doubletalk, i.e, there is only the far-end signal but no near-end speech. The far-end signal is a speech signal recorded in a quiet room from a female talker with a sampling rate of 8 kHz. The microphone signal is generated by convolving the far-end signal with an impulse response measured in a room with a reverberation time (T_{60}) of approximately 300 ms. To make the setting more realistic, white Gaussian noise is added to the microphone signal with an SNR of 30 dB. The overlap-add technique is employed in the implementation with an FFT length of $K = 1024$ and 75% overlap between neighboring frames. To minimize the aliasing effect, a Kaiser window of size 1024 is applied both before the STFT and after the inverse STFT.

To implement the MVDR filter given in (25), we need to know the correlation matrix $\mathbf{\Phi}_{\mathbf{d}}(k,n)$ and the normalized interframe correlation vector $\boldsymbol{\gamma}_X(k,n)$. Since both the microphone and far-end signals are accessible, we directly compute $\mathbf{\Phi}_{\mathbf{d}}(k,n)$ and $\gamma_X(k,n)$ from the corresponding signal with a short-time average using the most recent 20 frames.

The results of this simulation is plotted in Fig. 1. As seen, the MVDR filter can achieve more than 20-dB echo attenuation, which shows the effectiveness of the developed MVDR filter for echo suppression. It is interesting to see that $L = 2$ is sufficient to achieve good suppression performance. Notice that obviously the ERLE, ξ_{ERLE}, is low during the absence of the far-end speech. During the silence periods of the far-end speech, there is no echo in the microphone signal. If we still estimate and apply the suppression filter, it adds some noise into the microphone signal and as a result, ξ_{ERLE} becomes negative in decibel. One easy way to circumvent this issue is to apply a voice activity detector (VAD) to the far-end signal. If a processing frame is detected as silence, this frame is simply passed through without being filtered.

5.2. Echo Suppression Performance with Doubletalk

The second experiment evaluates the echo suppression performance of the MVDR filter in the situation where there is doubletalk. The far-end signal and room impulse response are same as used in the previous experiment. The near-end signal is a voice signal recorded in a quiet room from a male talker. Again, we add some white Gaussian noise to the microphone signal to make the experimental configuration realistic and the SNR is 30 dB. The results of this experiment are shown in Fig. 2. Note that we only plotted ξ_{ERLE} and υ_{sd} in the figure and did not show the subband and fullband echo reduction factors $\xi\left[\mathbf{h}(k,n)\right]$ and $\xi\left[\mathbf{h}(:,n)\right]$ and the distortion indices $\upsilon\left[\mathbf{h}(k,n)\right]$ and $\upsilon\left[\mathbf{h}(:,n)\right]$ due to space limitations.

From Fig. 2(d), one can see that more than 20-dB ERLE is achieved when there is no doubletalk. During doubletalk, the amount of ERLE yielded is less than that in the absence of doubletalk. As in the traditional AEC, the near-end speech behaves like noise and affects the estimation of the suppression filter during doubletalk. Again, the length of the MVDR filter does not seem to affect the ERLE much.

We see from From Fig. 2(e) that some distortion is added into the near-end signal particularly during the transition periods between no doubletalk and doubletalk. The amount of distortion is also dramatically affected by the filter length where a larger filter length causes more near-end speech distortion. The underlying reasons could be multiple. First, we use a short-time average (with 20 most recent frames) to estimate the signal statistics that are needed to implement the MVDR filter. These signal estimates are not accurate during the transition periods between no doubletalk and doubletalk. Second, we need to compute the inverse of the $\mathbf{\Phi}_{\mathbf{d}}(k,n)$ matrix in implementation. With only 20 frames to estimate it, this matrix can be ill-conditioned, particularly when the filter length is large. Finally, during the formulation of the MVDR filter, we did not add any constraint on the near-end signal distortion. Work is in progress to study how to achieve better estimates of the correlation matrix $\mathbf{\Phi}_{\mathbf{d}}(k,n)$ and the interframe correlation vector $\gamma_X(k,n)$, particularly when there is doubletalk. We are also working to improve the MVDR filter by including some constraints on signal distortion.

6. CONCLUSION AND DISCUSSION

In this paper, we studied an echo suppression approach to reducing the undesired echoes that result from acoustic coupling between a loudspeaker and a microphone in duplex voice communication. The process is divided into four basic steps. The first step partitions both the microphone and loudspeaker signals into short and overlapping frames. Each frame is then transformed into the STFT domain. Next, an MVDR filter is designed in each subband by explicitly using the interframe signal correlation and an estimate of the echo signal is estimated and subsequently subtracted from the microphone signal, thereby removing the echo signal. Finally, the time-domain echo-suppressed signal is constructed using the inverse STFT. Experiments were carried out to evaluate the performance in both the absence and presence of doubletalk. The results demonstrated that the presented method can achieve significant amount of echo attenuation. It is observed that some distortion is added into the near-end signal, primarily due to inaccurate estimation of the signal statistics during doubletalk. Work is in progress to improve the MVDR filter to not only optimize the amount of echo suppression, but also manage the level of near-end speech distortion.

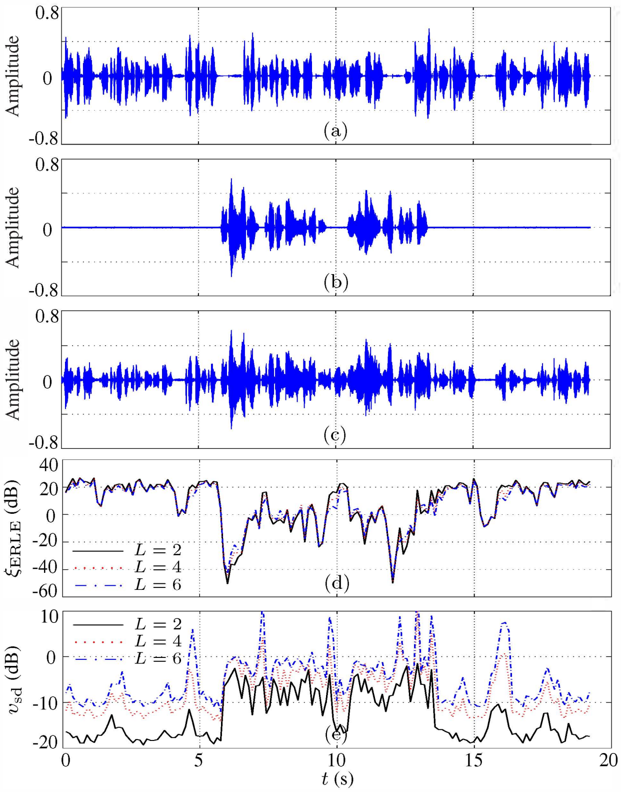

Fig. 2. Echo suppression performance: (a) waveform of the far-end speech, (b) waveform of the near-end speech, (c) waveform of the microphone signal, (d) ERLE of the MVDR echo suppression filter for three different filter lengths, i.e., $L = 2, 4, 6$, and (e) the distortion index of the near-end signal for the three different filter lengths.

7. RELATION TO PRIOR WORK

The detrimental acoustic echo effect can be controlled through two different ways, i.e., AEC [1]–[7] and AES [8]–[16]. The former achieves echo cancellation by adaptively identifying the echo path impulse response and then subtracting an echo estimate from the microphone signal while the latter attenuates echoes by directly applying a suppression filter to the microphone signal. In this paper, we presented an MVDR filter to reduce echoes in the STFT domain. Same as the traditional AES techniques, the presented method directly applies a suppression filter to the microphone signal and, therefore, introduces near-end distortion. As a result, it is called a echo suppression approach. However, the approach resembles the AEC technique in the sense that it also attempts to achieve an estimate of the echo signal and then subtract the estimate from the microphone signal.

8. REFERENCES

[1] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*. Berlin, Germany: Springer-Verlag, 2001.

[2] C. Paleologu, J. Benesty, and S. Ciochină, *Sparse Adaptive Filters for Echo Cancellation*. Morgan & Claypool Publishers, Synthesis Lectures on Speech and Audio Processing, 2010.

[3] M. M. Sondhi and D. R. Morgan, "Stereophoinc acoustic echo cancellation–an overview of the fundamental problem," *IEEE Signal Process. Lett.*, vol. 2, pp. 148–151, Aug. 1995.

[4] M. M. Sondhi and A. J. Presti, "A Self-adaptive echo canceller," *Bell Syst. Tech. J.*, 1966, pp, 1851–1854.

[5] C. Breining, P. Dreiseitel, E. Hänsler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic echo control-an application of very-high-order adpative filters, " *IEEE Signal Process. Mag.*, vol. 16, pp. 42–69, July 1999.

[6] Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing*. Berlin, Germany: Springer-verlag, 2006.

[7] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 156–165, Mar. 1998.

[8] C. Avendano and G. Garcia, "STFT-based multi-channel acoustic interference suppressor," in *Proc. IEEE ICASSP*, 2001, pp. 625–628.

[9] C. Faller and J. Chen, "Suppressing acoustic echo in a spectral envelope Space," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 1048–1062, Sept. 2005.

[10] C. Avendano, "Acoustic echo suppression in the STFT domain," in *Proc. IEEE WASPAA*, 2001, pp. 175–178.

[11] J. D. Gordy and R. A. Goubran, "Postfiltering for suppression of residual echo from vocoder distortion in packet-based telephony," in *Proc. IEEE ICME*, 2006, pp. 1953–1956.

[12] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control–A Practical Approach*. Hoboken, NJ: Wiley, 2004.

[13] A. S. Chhetri, A. C. Surendran, J. W. Stokes, and J. C. Platt, "Regression based residual acoustic echo suppression," in *Proc. IWAENC*, 2005, pp. 201–204.

[14] N. Madhu, I. Tashev, and A. Acero "An EM-based probabilistic approach for acoustic echo suppression," in *Proc. IEEE ICASSP*, 2008, pp. 265–268.

[15] P. Yun-Sik, and C. Joon-Hyuk, "Frequency domain acoustic echo suppression based on soft decision," *IEEE Signal Process. Lett.*, vol. 16, pp. 53–56, Jan. 2009.

[16] C. Faller and C. Tournery, "Robust acoustic echo control using a simple echo path model," in *Proc. IEEE ICASSP*, 2006, pp. V-281–V-284.

[17] J. Benesty, J. Chen, and E. Habets, *Speech Enhancement in the STFT Domain*. Berlin, Germany: Springer Briefs in Electrical and Computer Engineering, 2011.

[18] J. Benesty and Y. Huang, "A single-channel noise reduction MVDR filter," in *Proc. IEEE ICASSP*, 2011, pp. 273–276.

[19] J. Chen, J. Benesty, Y. Huang, and S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 1218–1234, July 2006.

[20] J. Benesty, J. Chen, Y. Huang, and S. Doclo, "Study of the Wiener filter for noise reduction," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Berlin, Germany: Springer-Verlag, 2005, ch. 2, pp. 9-41.