A Single-Channel MVDR Filter for Acoustic Echo Suppression

Karim Helwani, Herbert Buchner, *Member, IEEE*, Jacob Benesty, and Jingdong Chen, *Senior Member, IEEE*

***Abstract*—Acoustic echo suppression techniques for full-duplex communication in low-complexity systems are commonly known to introduce distortion to the desired signal (i.e., near-end speech). Moreover, most traditional echo control techniques typically require accurately detecting the contribution of the near-end speaker to the microphone signal ("double talk"). In this letter, we propose a novel approach to acoustic echo suppression, which aims at extracting the near-end signal using a constraint for minimizing the distortion, and without requiring a double-talk detector.**

***Index Terms*—Acoustic echo suppression, adaptive filtering, minimum variance distortionless response filter.**

I. Introduction

In hands-free communication, disturbing echoes are produced by acoustic feedback from the loudspeaker to the microphone. Acoustic echo cancellation (AEC) aims at canceling the acoustic echoes from the microphone signal by filtering the far-end signal $x(t)$ with the estimated echo path (an adaptive FIR filter $\hat{\mathbf{g}}$ of length L), and subtracting the resulting signal from the near-end microphone signal. If the estimated echo path is equal to the true echo path $\mathbf{g}$, all disturbing echoes will be removed from the microphone signal [1].

In AEC, residual echo suppressors, originally introduced in a heuristic way, are typically employed after the actual system identification-based AEC in order to meet the requirements for a high attenuation of the echoes in practical applications including, e.g., rapidly-varying acoustic environments, microphone noise, and considerable network delay [2]. As an extreme case, under the assumption of a simplified echo path model consisting of delay and short-time spectral modification, a system purely based on the residual echo suppression stage (acoustic echo suppression, AES) has been proposed in [3]–[5].

The basic notion of AES is a spectral modification of the microphone signal, $d(t)$, in order to attenuate its echo component, which is caused by acoustical feedback of the loudspeaker signal, $x(t)$, along the unknown echo path. The core assumption that has been made in [4] is that the echo path (room impulse response) can be entirely modelled by a linear-phase filter, i.e., on its way to the microphone, the loudspeaker signal is shifted in time and its magnitude spectrum is shaped. The latter effect, also called coloration, is mostly caused by early reflections of the room. Hence, in this model the impact of late reflections is ignored. The suppression filter in [4] is designed as a weighting function for parametric spectral subtraction using the estimated echo signal based on the coloration filter.

Suppression techniques are commonly known to introduce distortion to the desired signal. Moreover, AEC as well as AES, briefly reviewed above, typically require accurately detecting the contribution of the near-end speaker to the microphone signal ("double talk"). This letter addresses both the distortion problem and the double-talk problem. In order to minimize the signal distortion in the AES systems, we present a novel two-stage approach in this paper that explicitly constrains the near-end signal. Using the interframe statistics of the signal and extending the work in [6], [7] allows us to derive a (single-channel) minimum variance distortionless response (MVDR) filter. Similar to our previous work [8], the presented echo control system does not require double-talk detection.

II. Problem Formulation and Proposed Approach

A. Signal Model

Let us consider the conventional signal model in which an acoustic echo is generated from the coupling between a loudspeaker and a microphone. The microphone signal at the time index t can be written as

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

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 $y(t) = g(t) * x(t)$ 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 signal, $x(t)$, and the microphone 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, which can then be transmitted to the far-end room. Using the short-time Fourier transform (STFT), (1) can be expressed in the time-frequency domain as

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

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

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

Manuscript received December 27, 2012; revised February 11, 2013; accepted February 14, 2013. Date of publication February 20, 2013; date of current version February 26, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Muhammad Zubair Ikram.

K. Helwani is with the Quality and Usability Lab, Deutsche Telekom Laboratories, 10587 Berlin, Germany (e-mail: karim.helwani@telekom.de).

H. Buchner is with the Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany.

J. Benesty is with INRS-EMT, University of Quebec, Montreal, QC H5A 1K6 Canada.

J. Chen is with Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China.

Digital Object Identifier 10.1109/LSP.2013.2247998

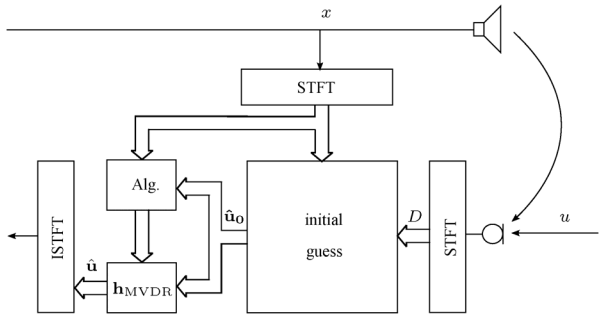

Fig. 1. Block diagram of the proposed system.

will be used, where $G(k)$ and $X(k,n)$ are the STFTs of $g(t)$ and $x(t)$, respectively. Hence, the microphone signal can be described as

$$D(k,n) = \begin{bmatrix} G(k) & 1 \end{bmatrix} \begin{bmatrix} X(k,n) \\ U(k,n) \end{bmatrix}. \tag{4}$$

Further, we assume that the near-end and echo signals are uncorrelated such that

$$\hat{\mathcal{E}}\left\{U(k,n)X^*(k,n)\right\} = 0, \tag{5}$$

where $\hat{\mathcal{E}}\{\cdot\}$ denotes empirical expectation (time average) and superscript * is the complex-conjugate operator.

B. System Architecture of Proposed Echo Suppressor

We now introduce a solution based on the previous assumptions, which is composed of two processing stages as depicted in Fig. 1. In the first stage, an initial guess of the near-end signal is obtained. The estimated signal is then post-processed to minimize distortion. As detailed in the following subsections C and D, respectively, both stages follow directly from (3)–(5).

C. Initial Guess of Near-End Signal

For simultaneous estimation of $G(k)$ and the near-end signal, $U(k,n)$, we set up the system of (6), shown at the bottom of the page, by combining (4) and (5), where $\hat{U}_0(k,n)$ denotes an estimate of $U(k,n)$ and $\hat{G}(k,n)$ is an estimate of $G(k,n)$.

Note that the matrix on the right-hand side exclusively depends on the loudspeaker signal $X(\cdot)$, while the left-hand side exclusively depends on the microphone signal $D(\cdot)$. Further, for the estimation in (5) we build a circular matrix using the conjugate of the vector $\mathbf{x}(k,n) := [X(k,n), \ldots, X(k,n-M_1+1)]^{\mathrm{T}}$.

Let us define

$$\hat{\mathbf{u}}_{\mathbf{0}}(k,n) := \left[\hat{U}_0(k,n), \ldots, \hat{U}_0(k,n-M_2+1)\right]^{\mathrm{T}},$$

which is an estimate of $\mathbf{u}(k,n) := [U(k,n), \ldots, U(k,n-M_2+1)]^{\mathrm{T}}$. The estimate $\hat{\mathbf{u}}_{\mathbf{0}}(k,n)$ can be obtained from (6) using the pseudoinverse. The solution can be interpreted as an explicit block-online version of [8], which explains how this approach works without additional double talk detection. Note also that the pseudoinverse can be carried out efficiently due to the sparse matrix structure in (6), although this is outside the scope of this letter.

The elements $\hat{U}_0(k,n)$ could still contain both a residual echo component that is considered as an interference and a part of the desired near-end signal. For suppression of the residual echo signal, we consider further decomposing the estimated near-end signal as

$$\hat{\mathbf{u}}_{\mathbf{0}}(k,n) = \mathbf{u}_{\mathrm{c}}(k,n) + \mathbf{u}_{\mathrm{i}}(k,n) + \mathbf{r}(k,n), \tag{7}$$

where $\mathbf{u}_{\mathrm{c}}(k,n)$ is the component of the estimated near-end signal vector that is coherent with $U(k,n)$, $\mathbf{u}_{\mathrm{i}}(k,n)$ is the incoherent component that is orthogonal to the coherent component $\mathbf{u}_{\mathrm{c}}(k,n)$, and $\mathbf{r}$ denotes the residual echo. In the next section, we show how this decomposition can be done in practice.

D. MVDR Processing Stage and Residual Echo Suppression

In the following, we show how to derive the MVDR filter for acoustic echo suppression. The idea is to estimate a distortionless version $\hat{U}(k,n)$ of the near-end signal starting from the initial estimate $\hat{\mathbf{u}}_{\mathbf{0}}(k,n)$. Coherence between $U(k,n)$ and the estimate $\hat{U}(k,n)$ occurs if the following condition is fulfilled:

$$\hat{\mathcal{E}}\left\{\hat{U}(k,n)U^*(k,n)\right\} = \phi_U(k,n), \tag{8}$$

$$\begin{bmatrix} D(k,n) \\ D(k,n-1) \\ \vdots \\ D(k,n-M_2+1) \\ \hline 0 \\ \vdots \\ 0 \end{bmatrix} = \left[\begin{array}{c|cccccc} X(k,n) & 1 & 0 & & & \ldots & 0 \\ X(k,n-1) & 0 & 1 & 0 & & \ldots & 0 \\ \vdots & \vdots & & & & \ddots & \vdots \\ X(k,n-M_2+1) & 0 & & & & \ldots & 1 \\ \hline 0 & X^*(k,n) & X^*(k,n-1) & \ldots & X^*(k,n-M_1+1) & 0 \ \ldots & 0 \\ 0 & X^*(k,n-M_1+1) & X^*(k,n) & \ldots & X^*(k,n-M_1+2) & 0 \ \ldots & 0 \\ \vdots & \vdots & \ddots & & & \vdots & \vdots \\ 0 & X^*(k,n-1) & X^*(k,n-2) & \ldots & X^*(k,n) & 0 \ \ldots & 0 \end{array}\right] \cdot \begin{bmatrix} \hat{G}(k) \\ \hline \hat{U}_0(k,n) \\ \vdots \\ \hat{U}_0(k,n-M_1+1) \\ \vdots \\ \hat{U}_0(k,n-M_2+1) \end{bmatrix}, \tag{6}$$

where $\phi_U(k,n)$ is defined as

$$\phi_U(k,n) := \hat{\mathcal{E}}\left\{U(k,n)U^*(k,n)\right\}. \tag{9}$$

Using

$$\begin{aligned}\hat{U}(k,n) &= \mathbf{h}^{\mathrm{H}}(k,n)\hat{\mathbf{u}}_{\mathbf{0}}(k,n)\\ &= \mathbf{h}^{\mathrm{H}}(k,n)\left[\mathbf{u}_{\mathrm{c}}(k,n)+\mathbf{u}_{\mathrm{i}}(k,n)+\mathbf{r}(k,n)\right],\end{aligned} \tag{10}$$

we obtain with $\boldsymbol{\gamma}_{\mathbf{u}}(k,n) := \mathbf{u}_{\mathrm{c}}(k,n)/U(k,n)$ and (8):

$$\begin{aligned}\hat{\mathcal{E}}\left\{\hat{U}(k,n)U^*(k,n)\right\} &= \mathbf{h}^{\mathrm{H}}(k,n)\hat{\mathcal{E}}\left\{\mathbf{u}_{\mathrm{c}}(k,n)U^*(k,n)\right\}\\ &= \mathbf{h}^{\mathrm{H}}(k,n)\boldsymbol{\gamma}_{\mathbf{u}}(k,n)\hat{\mathcal{E}}\left\{U(k,n)U^*(k,n)\right\}.\end{aligned} \tag{11}$$

Since $\mathbf{u}_{\mathrm{c}}(k,n)$ is in general unknown, we derive for determining $\boldsymbol{\gamma}_{\mathbf{u}}(k,n)$

$$\begin{aligned}\hat{\mathcal{E}}\left\{\mathbf{u}_{\mathrm{c}}(k,n)U^*(k,n)\right\} &= \hat{\mathcal{E}}\left\{\mathbf{u}(k,n)U^*(k,n)\right\}\\ &= \boldsymbol{\gamma}_{\mathbf{u}}(k,n)\hat{\mathcal{E}}\left\{U(k,n)U^*(k,n)\right\},\end{aligned} \tag{12}$$

$$\boldsymbol{\gamma}_{\mathbf{u}}(k,n) = \frac{\hat{\mathcal{E}}\left\{\mathbf{u}(k,n)U^*(k,n)\right\}}{\phi_U(k,n)}. \tag{13}$$

Now, from condition (11), we immediately obtain the following important constraint for $\mathbf{h}$:

$$\mathbf{h}^{\mathrm{H}}\boldsymbol{\gamma}_{\mathbf{u}}(k,n) = 1. \tag{14}$$

In a practical implementation, we determine $\boldsymbol{\gamma}_{\mathbf{u}}(k,n)$ using the initial guess $\hat{\mathbf{u}}_{\mathbf{0}}$.

In (7), $\mathbf{r}(k,n)$ in turn can be decomposed into coherent and incoherent orthogonal components relative to the echo signal. With the assumption that the residual echo signal is coherent with the loudspeaker signal, we derive

$$\mathbf{r}(k,n) = \alpha(k,n)X(k,n)\boldsymbol{\gamma}_{\mathbf{x}}(k,n) + \mathbf{r}_{\mathrm{i}}(k,n), \tag{15}$$

where $\alpha(k,n)$ models the initial suppression, done using (6), and, analogous to (13), we calculate

$$\boldsymbol{\gamma}_{\mathbf{x}}(k,n) = \frac{\hat{\mathcal{E}}\left\{\mathbf{x}(k,n)X^*(k,n)\right\}}{\phi_X(k,n)}. \tag{16}$$

It is preferable to estimate the near-end signal with no distortion while minimizing the residual echo. Therefore, we have two constraints, (14) and

$$\mathbf{h}^{\mathrm{H}}\boldsymbol{\gamma}_{\mathbf{x}}(k,n) = 0. \tag{17}$$

1) Minimum Variance: Based on the minimum variance criterion, we aim at minimizing the cost function

$$\begin{aligned}J_0(\mathbf{h}) &:= \hat{\mathcal{E}}\left\{\hat{U}(k,n)\hat{U}^*(k,n)\right\}\\ &= \mathbf{h}^{\mathrm{H}}\boldsymbol{\Phi}_{\hat{\mathbf{u}}_{\mathbf{0}}\hat{\mathbf{u}}_{\mathbf{0}}}(k,n)\mathbf{h},\end{aligned} \tag{18}$$

where

$$\boldsymbol{\Phi}_{\hat{\mathbf{u}}_{\mathbf{0}}\hat{\mathbf{u}}_{\mathbf{0}}}(k,n) = \hat{\mathcal{E}}\left\{\hat{\mathbf{u}}_{\mathbf{0}}(k,n)\hat{\mathbf{u}}_{\mathbf{0}}^{\mathrm{H}}(k,n)\right\}. \tag{19}$$

By applying Tikhonov regularization, we obtain one more constraint on the ℓ_2-norm of $\mathbf{h}$ so that the regularized cost function reads

$$J_1(\mathbf{h}) := \mathbf{h}^{\mathrm{H}}\boldsymbol{\Phi}_{\hat{\mathbf{u}}_{\mathbf{0}}\hat{\mathbf{u}}_{\mathbf{0}}}(k,n)\mathbf{h} + \delta\mathbf{h}^{\mathrm{H}}\mathbf{h}, \tag{20}$$

where δ is a Lagrange multiplier.

2) Distortionless Response: The constraints in (14) and (17) can be combined into a system of equations

$$\boldsymbol{\Gamma}(k,n)\mathbf{h} = \begin{bmatrix}1\\0\end{bmatrix}, \tag{21}$$

where

$$\boldsymbol{\Gamma}(k,n) := \begin{bmatrix}\boldsymbol{\gamma}_{\mathbf{u}}^{\mathrm{H}}(k,n)\\ \boldsymbol{\gamma}_{\mathbf{x}}^{\mathrm{H}}(k,n)\end{bmatrix}. \tag{22}$$

Adding these constraints in (21) to the cost function (20) and using the Lagrangian multiplier technique, we obtain the final cost function

$$J(\mathbf{h}) := \mathbf{h}^{\mathrm{H}}\boldsymbol{\Phi}_{\hat{\mathbf{u}}_{\mathbf{0}}\hat{\mathbf{u}}_{\mathbf{0}}}(k,n)\mathbf{h} + \delta\mathbf{h}^{\mathrm{H}}\mathbf{h} + \boldsymbol{\lambda}^{\mathrm{T}}\left(\begin{bmatrix}1\\0\end{bmatrix} - \boldsymbol{\Gamma}(k,n)\mathbf{h}\right), \tag{23}$$

where $\boldsymbol{\lambda}$ is an additional 2×1 vector of Lagrange multipliers.

Finally, we derive

$$\begin{aligned}\mathbf{h}_{\mathrm{MVDR}}(k,n) &= \left(\boldsymbol{\Phi}_{\hat{\mathbf{u}}_{\mathbf{0}}\hat{\mathbf{u}}_{\mathbf{0}}}(k,n)+\delta\mathbf{I}\right)^{-1}\boldsymbol{\Gamma}^{\mathrm{H}}(k,n)\\ &\times\left[\boldsymbol{\Gamma}(k,n)\left(\boldsymbol{\Phi}_{\hat{\mathbf{u}}_{\mathbf{0}}\hat{\mathbf{u}}_{\mathbf{0}}}(k,n)+\delta\mathbf{I}\right)^{-1}\boldsymbol{\Gamma}^{\mathrm{H}}(k,n)\right]^{-1}\begin{bmatrix}1\\0\end{bmatrix},\end{aligned} \tag{24}$$

where $\mathbf{I}$ is the identity matrix. For the special case $\boldsymbol{\Phi}_{\mathbf{u}_{\mathbf{0}}\mathbf{u}_{\mathbf{0}}}(k,n) = \mathbf{I}$, we obtain the desired filter as

$$\mathbf{h}_{\mathrm{MVDR}}k,n) = \boldsymbol{\Gamma}^{\dagger}(k,n)\begin{bmatrix}1\\0\end{bmatrix}. \tag{25}$$

III. Experimental Results

A. Performance Measures

The two most important means to evaluate acoustic echo suppression performance are the attenuation of the acoustic echo, and the distortion of the near-end signal. We define the fullband acoustic echo reduction factor at time frame n as

$$\xi(n) := \frac{\sum_{k=0}^{K-1}\phi_Y(k,n)}{\sum_{k=0}^{K-1}\phi_{\hat{U}}(k,n)}, \tag{26}$$

where $\phi_Y(k,n)$ and $\phi_{\hat{U}}(k,n)$ are defined analogously to (9). This definition is equivalent to the echo-return loss enhancement (ERLE) in the single-talk case [1]. The ERLE should be greater than or equal to 1. When $\xi = 1$, there is no echo reduction, and the higher the value of ξ, the more the echo is reduced. Note that in contrast to AEC, we cannot easily give a bound on ERLE (in AEC the upper bound is given by the echo-to-background noise ratio [1]) as we do not have a fixed set of system parameters, but have to estimate an optimum signal-dependent parameter set in each signal block. Further, we define the fullband near-end signal distortion index at time frame n as

$$v(n) := \frac{\sum_{k=0}^{K-1}\hat{\mathcal{E}}\left\{\left|\hat{U}(k,n)-U(k,n)\right|^2\right\}}{\sum_{k=0}^{K-1}\phi_U(k,n)}. \tag{27}$$

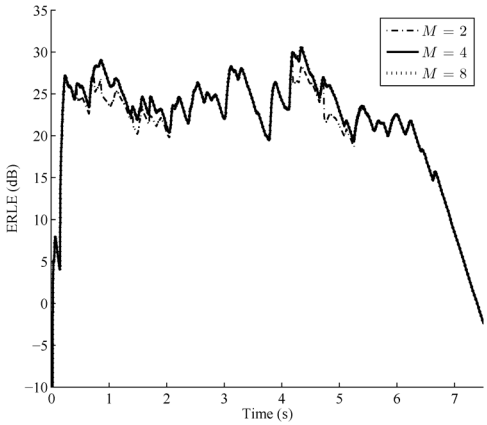

Fig. 2. Achieved echo-return loss enhancement (ERLE) of the proposed system in the single-talk period.

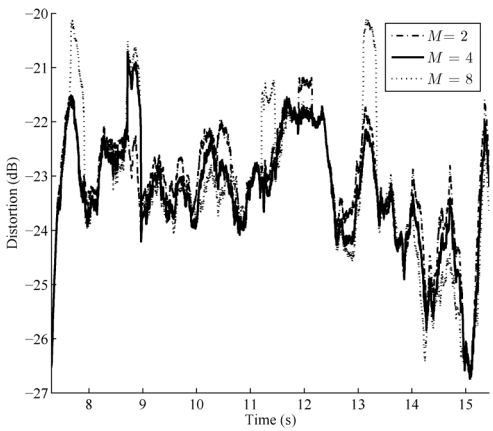

Fig. 3. Achieved distortion of the near-end signal during the double-talk period.

B. Simulation

To evaluate how successful the described algorithm is in suppressing the echo signal, a dialogue sequence of roughly 25 seconds is simulated. The sequence is split into three parts of approximately equal duration. The first consists of only the (German) far-end talker, the second both talkers simultaneously ("double talk"), and the third part of only the (English) talker at the near-end. Far-end and near-end speech signals have been adjusted manually to exhibit roughly equal loudness. For simulation of the echo path, we employ a time-invariant impulse response (IR) recorded in a meeting room with a reverberation time (T_{60}) of approximately 300 ms.

To make the setting more realistic, Gaussian white noise is added to the microphone signal with an SNR of 35 dB [measured as root-mean-square (RMS) value]. The sampling frequency of the signals is 8 kHz. The chosen FFT length is 256 with an overlap factor of 50%. Fig. 2 shows the achieved ERLE in the single-talk period as defined in (26) for different filter lengths $M_1 = M_2 = M \in \{2, 4, 8\}$. The distortion in the double-talk period is given in Fig. 3. The choice of the optimal filter order depends on the length of the impulse response of the near-end room and the signal statistics.

The simulation has proven the efficiency of the proposed algorithm. The achieved echo suppression is comparable to typical ERLE requirements in AEC and increasing the FIR filter order leads to improved echo suppression. The distortion of the near-end signal in the double-talk period is upper limited to -20 dB. The simulation also shows that filters with low orders offer high tracking performance. As we may expect, the near-end signal is even less distorted when only the near-end speaker is active, as shown in Fig. 4. Note that less distortion can be achieved with large values of M. Hence, the choice of the filter length is also a trade-off between the desired tracking performance and the required suppression.

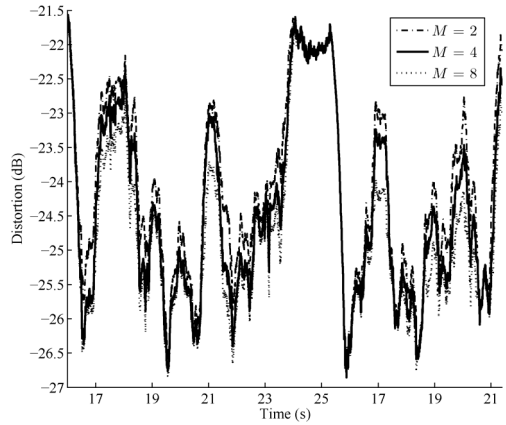

Fig. 4. Achieved distortion of the near-end signal during the period where only the near-end speaker is active.

Preliminary psychoacoustic experiments by the authors have shown that the proposed approach satisfies the expected quality of an AES. In future work, psychoacoustic experiments will be conducted to evaluate the performance of the proposed approach from an end user's point of view.

IV. Conclusions

In this letter, we have presented an approach to acoustic echo suppression that extracts a distortionless near-end signal from the microphone signal without requiring a double-talk detector. Simulation results show that the new approach offers a high degree of flexibility and is scalable and highly efficient.

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] R. Martin and J. Altenhoner, "Coupled adaptive filters for acoustic echo control and noise reduction," in *Proc. IEEE ICASSP*, 1995, vol. 5, pp. 3043–3043.

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

[4] C. Faller and C. Tournery, "Estimating the delay and coloration effect of the acoustic echo path for low complexity echo suppression," in *Proc. IWAENC*, 2005, pp. 1–4.

[5] C. Faller and C. Tournery, "Stereo acoustic echo control using a simplified echo path model," in *Proc. IWAENC*, 2006, pp. 1–4.

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

[7] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*. Berlin, Germany: Springer-Verlag, 2011.

[8] H. Buchner and W. Kellermann, "A fundamental relation between blind and supervised adaptive filtering illustrated for blind source separation and acoustic echo cancellation," in *Proc. HSCMA*, 2008, pp. 17–20.