Suppressing Acoustic Echo in a Spectral Envelope Space

Christof Faller and Jingdong Chen, *Member, IEEE*

***Abstract*—Full-duplex hands-free telecommunication systems employ an acoustic echo canceler (AEC) to remove the undesired echoes that result from the coupling between a loudspeaker and a microphone. Traditionally, the removal is achieved by modeling the echo path impulse response with an adaptive finite impulse response (FIR) filter and subtracting an echo estimate from the microphone signal. It is not uncommon that an adaptive filter with a length of 50–300 ms needs to be considered, which makes an AEC highly computationally expensive. In this paper, we propose an echo suppression algorithm to eliminate the echo effect. Instead of identifying the echo path impulse response, the proposed method estimates the spectral envelope of the echo signal. The suppression is done by spectral modification—a technique originally proposed for noise reduction. It is shown that this new approach has several advantages over the traditional AEC. Properties of human auditory perception are considered, by estimating spectral envelopes according to the frequency selectivity of the auditory system, resulting in improved perceptual quality. A conventional AEC is often combined with a post-processor to reduce the residual echoes due to minor echo path changes. It is shown that the proposed algorithm is insensitive to such changes. Therefore, no post-processor is necessary. Furthermore, the new scheme is computationally much more efficient than a conventional AEC.**

***Index Terms*—Acoustic echo cancellation, adaptive filter, echo suppression, spectral modification.**

I. Introduction

An *acoustic echo canceler* (AEC) is a necessary component for a full-duplex hands-free telecommunication system to eliminate undesired echo signals that result from acoustic coupling between a loudspeaker and a microphone. Traditionally, echo cancellation is accomplished by adaptively identifying the echo path impulse response and subtracting an estimate of the echo signal from the microphone signal. A typical AEC is illustrated in Fig. 1. The far-end talker signal $x(n)$ (loudspeaker signal) goes through the echo path, whose impulse response is modeled as a finite impulse response (FIR) filter, and adds to the microphone signal $y(n)$ together with the near-end talker signal $v(n)$ and the ambient noise $w(n)$:

$$\begin{aligned} y(n) &= u(n) + v(n) + w(n) \\ &= \mathbf{h}^T \mathbf{x}(n) + v(n) + w(n) \end{aligned} \tag{1}$$

Manuscript received August 4, 2003; revised July 13, 2004. This work was carried out at Agere Systems, Allentown, PA. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Futoshi Asano.

C. Faller is with the Audiovisual Communications Laboratory, School of Computer and Communication Sciences, EPFL Lausanne, Lausanne, Switzerland (e-mail: christof.faller@epfl.ch).

J. Chen is with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA (e-mail: jingdong@research.bell-labs.com).

Digital Object Identifier 10.1109/TSA.2005.852012

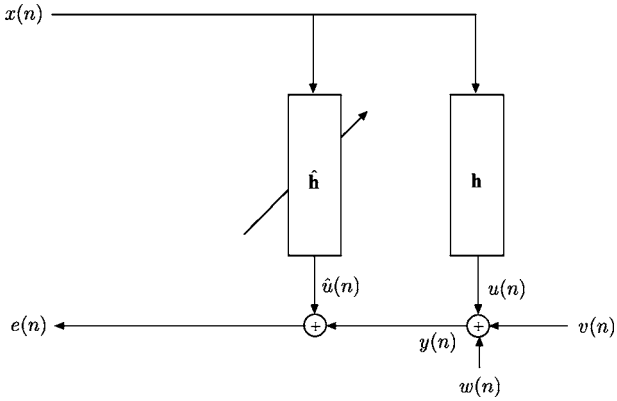

Fig. 1. Schematic diagram of an adaptive acoustic echo canceler.

where

$$\begin{aligned} \mathbf{x}(n) &= [x(n), x(n-1), \ldots, x(n-M+1)]^T \\ \mathbf{h} &= [h_0, h_1, \ldots, h_{M-1}]^T \end{aligned}$$

M is the length of the echo path impulse response, and ^{T} denotes the transpose of a vector or a matrix. To cancel the echo in the microphone signal, an echo estimate $\hat{u}(n)$ is needed, which is generated by passing the far-end talker signal through an FIR filter

$$\hat{\mathbf{h}} = [\hat{h}_0, \hat{h}_1, \ldots, \hat{h}_{L-1}]^T \tag{2}$$

of length L (generally less than M), i.e.,

$$\hat{u}(n) = [\hat{\mathbf{h}}^T, \mathbf{0}]\mathbf{x}(n). \tag{3}$$

The FIR filter coefficients are estimated adaptively in time. Subtracting $\hat{u}(n)$ from the microphone signal $y(n)$ yields the error signal

$$\begin{aligned} e(n) &= y(n) - \hat{u}(n) \\ &= [u(n) - \hat{u}(n)] + v(n) + w(n). \end{aligned} \tag{4}$$

The *mean square error* (MSE) can then be expressed as

$$\begin{aligned} E\{e^2(n)\} = E\left\{[u(n) - \hat{u}(n)]^2\right\} &+ E\{v^2(n)\} \\ &+ E\{w^2(n)\} + 2E\{[u(n) - \hat{u}(n)]\, v(n)\} \\ &+ 2E\{[u(n) - \hat{u}(n)]\, w(n)\} \\ &+ 2E\{v(n) w(n)\} \end{aligned} \tag{5}$$

where $E\{\cdot\}$ denotes mathematical expectation. If $u(n)$, $v(n)$, and $w(n)$ are assumed to be uncorrelated, then (5) can be simplified to

$$E\{e^2(n)\} = E\left\{[u(n) - \hat{u}(n)]^2\right\} + E\{v^2(n)\} + E\{w^2(n)\}. \tag{6}$$

Note that $E\{v^2(n)\}$ and $E\{w^2(n)\}$ are unaffected by the filter. Therefore, minimizing $E\{e^2(n)\}$ is equivalent to minimizing $E\{[u(n)-\hat{u}(n)]^2\}$. It is then obvious that the objective of AEC is to estimate an $\hat{\mathbf{h}}$ that minimizes $E\{e^2(n)\}$.

There is a vast literature addressing how to search for the optimum $\hat{\mathbf{h}}$ using adaptive techniques. Commonly used algorithms include normalized least-mean-square (NLMS), *recursive least-squares* (RLS), proportionate NLMS (PNLMS), *affine projection algorithm* (APA), etc. A good review of these algorithms can be found in [1], [2].

When the near-end talker is silent, i.e., $v(n) = 0$, and the signal-to-noise ratio (SNR) is high (e.g., SNR $\geq$ 30 dB), the adaptive filter $\hat{\mathbf{h}}$ can converge to a good estimate of the true echo path impulse response $\mathbf{h}$ and the echo will be canceled sufficiently, such that the far-end talker is not disturbed by returning echo signal components. In the presence of doubletalk, i.e., when the far-end and near-end talkers are active at the same time, the near-end signal $v(n)$ acts as a strong noise signal. This is likely to cause the adaptive filter to diverge, resulting in insufficient echo cancellation. To prevent this from happening, a doubletalk detector is used [3]–[7]. Whenever doubletalk is detected, the adaptive filter coefficients are frozen.

A commonly used measure to evaluate the convergence of the adaptive filter is the normalized misalignment, which is defined as

$$\epsilon = \frac{\left\|\mathbf{h} - [\hat{\mathbf{h}}^T, 0]^T\right\|}{\|\mathbf{h}\|} \tag{7}$$

where $\|\cdot\|$ denotes the l_2 norm. The normalized misalignment measures the mismatch between the echo path impulse response and the modeling filter. The smaller the misalignment is, the better is the echo cancellation performance. Other commonly used measures, such as the normalized MSE, will be discussed in Section IV.

In order to achieve acceptable performance, the length of the cancellation filter has to be long enough to capture most of the echo energy. In a small office environment, to achieve an even modest performance, for instance $\epsilon \leq -20$ dB, a cancellation filter of 50 milliseconds, which corresponds to 400 taps at 8-kHz sampling rate, is commonly considered [1]. For larger rooms and higher sampling rates the number of taps that need to be considered rises to several thousands. As a result, the computational complexity of an AEC is very high. The computational complexity can be reduced by implementing an AEC in the frequency domain, see e.g. [8]–[11]. But the computational cost remains high.

In this paper, we propose a novel algorithm for the purpose of eliminating the undesired echo effect, operating in a spectral envelope space. Instead of identifying the echo path impulse response, this new algorithm directly estimates the spectral envelope of the echo signal. The cancellation is done by spectral modification, a technique originally proposed for noise reduction [12], [13]. The spectral envelope is represented considering frequency selectivity properties of the human auditory system. For this reason, the proposed scheme is called *perceptual acoustic echo suppressor* (PAES).

Compared with conventional AECs, the proposed PAES offers several advantages.

- In the framework of PAES, perceptual aspects are easily incorporated, allowing optimization of the perceptual quality of the system.
- The spectral envelope contains no information from the phase spectrum or fine structure of the magnitude spectrum. Therefore, the PAES scheme is resilient against minor echo path changes that only affect the echo signal's phase spectrum or fine structure of its magnitude spectrum. As a result, no post-processor is necessary for suppressing residual echoes. AEC's usually require such a post-processor; see, e.g., [14] and [15].
- Fewer parameters need to be estimated, which makes the PAES algorithm computationally more efficient than a conventional AEC.

II. Proposed Acoustic Echo Suppression Algorithm

A. Notation and Variables

Before formulating the addressed problem and developing the proposed algorithm, we define the notation and variables used in this paper.

$x(n)$	Far-end signal/speech.
$v(n)$	Near-end signal/speech (doubletalk).
$w(n)$	Ambient noise.
$y(n)$	Microphone signal including echo, ambience noise, and possibly near-end signal.
$\mathbf{h}$	$= [h_0, h_1, \ldots, h_{M-1}]^T$, true echo path.
M	Length of the echo path impulse response.
$\hat{\mathbf{h}}$	$= [\hat{h}_0, \hat{h}_1, \ldots, \hat{h}_{L-1}]^T$, estimated echo path.
L	Length of the estimated echo path impulse response.
$\mathbf{x}(n)$	$= [x(n), x(n-1), \ldots, x(n-M+1)]^T$, excitation vector.
$u(n)$	$= \mathbf{h}^T\mathbf{x}(n)$, echo signal.
$z(n)$	$= v(n) + w(n)$.
$\mathbf{x}_k$	$= [x_k(0), x_k(1), \ldots, x_k(W-1)]^T = [x(kN), x(kN+1), \ldots, x(kN+W-1)]^T$, a frame of the far-end signal at time index k; ($\mathbf{y}_k$, $\mathbf{v}_k$, $\mathbf{w}_k$, $\mathbf{u}_k$, and $\mathbf{z}_k$ are defined similarly).
W	*Short-time Fourier transform* (STFT) window size.
N	STFT window hop size.
$X_k(j\omega)$	$= \sum_{m=0}^{W-1} \delta(m)x_k(m)e^{-j\omega m}$, STFT of $\mathbf{x}_k$.
$\delta(m)$	Analysis window.
ω	Radial frequency.
$Y_k(j\omega)$	STFT of $\mathbf{y}_k$; [$V_k(j\omega)$, $W_k(j\omega)$, $U_k(j\omega)$, and $Z_k(j\omega)$ are defined similarly].

B. Problem Formulation

With the defined variables and notations, the signal model given in (1) can be rewritten in a vector form as

$$\mathbf{y}_k = \mathbf{u}_k + \mathbf{z}_k. \tag{8}$$

Taking STFT on both sides of (8) yields

$$Y_k(j\omega) = U_k(j\omega) + Z_k(j\omega). \tag{9}$$

The echo cancellation can then be formulated as an estimation problem in the time-frequency domain, which aims to estimate $Z_k(j\omega)$ from the observed signal $Y_k(j\omega)$. This can be done by obtaining a replica of $U_k(j\omega)$, and then subtracting it from $Y_k(j\omega)$.

A complex spectrum can be written as

$$\begin{aligned} Y_k(j\omega) &= |Y_k(j\omega)|\, e^{j\Psi_{Y_k}(\omega)} \\ Z_k(j\omega) &= |Z_k(j\omega)|\, e^{j\Psi_{Z_k}(\omega)}. \end{aligned} \tag{10}$$

The echo cancellation problem becomes now equivalent to the design of two signal estimators that make decisions separately on the spectral magnitude, $|Z_k(j\omega)|$, and the phase component, $\Psi_{Z_k}(\omega)$.

It has been shown that human perception is relatively insensitive to phase distortion [16]–[18]. Therefore, $\Psi_{Y_k}(\omega)$ can be used as an estimate of $\Psi_{Z_k}(\omega)$ for echo suppression purpose. Keeping this in mind, the echo cancellation problem can be simplified to only estimating $|Z_k(j\omega)|$ based on $Y_k(j\omega)$. This serves as the basis for the proposed echo suppression algorithm.

C. Spectral Magnitude Modification Based Echo Suppressor (SMMES)

Given $Y_k(j\omega)$, $|Z_k(j\omega)|$ can be estimated through spectral modification. By assuming that $\mathbf{u}_k$ and $\mathbf{z}_k$ are uncorrelated, it follows from (9) that $|Y_k(j\omega)|^2$ can be approximated with [13], [19]

$$|Y_k(j\omega)|^2 \approx |U_k(j\omega)|^2 + |Z_k(j\omega)|^2 . \tag{11}$$

Therefore, the instantaneous power spectrum of the signal $\mathbf{z}_k$, *viz.* $|Z_k(j\omega)|^2$, can be recovered by subtracting an estimate of $|U_k(j\omega)|^2$ from $|Y_k(j\omega)|^2$, i.e.,

$$\begin{aligned} \left|\hat{Z}_k(j\omega)\right|^2 &= |Y_k(j\omega)|^2 - \left|\hat{U}_k(j\omega)\right|^2 \\ &= |Z_k(j\omega)|^2 + \left[|U_k(j\omega)|^2 - \left|\hat{U}_k(j\omega)\right|^2\right]. \end{aligned} \tag{12}$$

The corresponding spectral magnitude of z_k is computed as

$$\begin{aligned} \left|\hat{Z}_k(j\omega)\right| &= \sqrt{\left|\hat{Z}_k(j\omega)\right|^2} \\ &= G_{\mathrm{P}}(\omega)\, |Y_k(j\omega)| \end{aligned} \tag{13}$$

where

$$G_{\mathrm{P}}(\omega) = \left[\frac{|Y_k(j\omega)|^2 - \left|\hat{U}_k(j\omega)\right|^2}{|Y_k(j\omega)|^2}\right]^{\frac{1}{2}} \tag{14}$$

is called a *gain filter*.

A similar gain filter can be formulated in the spectral magnitude domain [13], [19], [20]. In general, $|Z_k(j\omega)|$ can be recovered through

$$\left|\hat{Z}_k(j\omega)\right| = G(\omega)\, |Y_k(j\omega)| \tag{15}$$

where

$$G(\omega) = \left[\frac{|Y_k(j\omega)|^\alpha - \beta\left|\hat{U}_k(j\omega)\right|^\alpha}{|Y_k(j\omega)|^\alpha}\right]^{\frac{1}{\alpha}}.$$

α is an exponent, and β is a parameter introduced to control the amount of echo to be suppressed in case it is under (or over) estimated. Combined with the phase spectrum, an estimate of the spectrum of $\mathbf{z}_k$ is

$$\begin{aligned} \hat{Z}_k(j\omega) &= G(\omega)\, |Y_k(j\omega)|\, e^{j\Psi_{Y_k}(\omega)} \\ &= G(\omega) Y_k(j\omega). \end{aligned} \tag{16}$$

This is often referred to as the spectral modification technique (or sometimes parametric Wiener filtering technique, or parametric spectral subtraction). It has been widely adopted for the purpose of additive noise suppression and speech enhancement [12], [13], [16], [19]. It was also investigated in [21] for the purpose of echo suppression. A diagram of the *spectral magnitude modification based echo suppressor* (SMMES) is shown in Fig. 2. It eliminates echo signals in the time-frequency domain on a frame-by-frame basis. First, the incoming microphone signal is partitioned into successive frames. The frame length is typically selected between 10 and 40 ms. A window function (e.g. Hann window) is applied to the frame signal for a better estimation. Then, the short-time Fourier spectrum is obtained by applying STFT to the windowed frame signal. Next, the echo components are estimated by modeling the echo path with an adaptive filter in each STFT frequency bin [21]. The gain filter is then computed based on the estimated spectral magnitudes (or instantaneous power spectra) of both the echo signal and the microphone signal. Given the gain filter, the STFT spectra of the microphone signal are modified such that the echo components are suppressed while maintaining the near-end talker signal, enabling duplex communication. Finally, the echo-suppressed output signal is constructed using the overlap-add technique with inverse STFT.

Although it is shown in [21] that the SMMES approach is computationally cheaper than a time-domain AEC, our investigation indicates that it is not significantly more efficient than an AEC based on a frequency-domain adaptive algorithm, since the number of parameters that need to be estimated is not significantly reduced. Spectral modification for the purpose of noise reduction often results in a perceptually annoying phenomenon called "musical noise" due to the isolated spectral peaks resulting from the nonlinear gain manipulation [12]. SMMES has a similar problem and often suffers from audible artifacts.

D. Perceptual Acoustic Echo Suppressor (PAES)

Auditory properties [22] have been widely incorporated into speech and audio processing techniques. For instance, in the areas of speech/audio coding and speech enhancement important progress has been achieved by employing masking effects and other auditory principles [23]–[25]. Masking has also been explored in combined systems and noise and residual echo suppression [15].

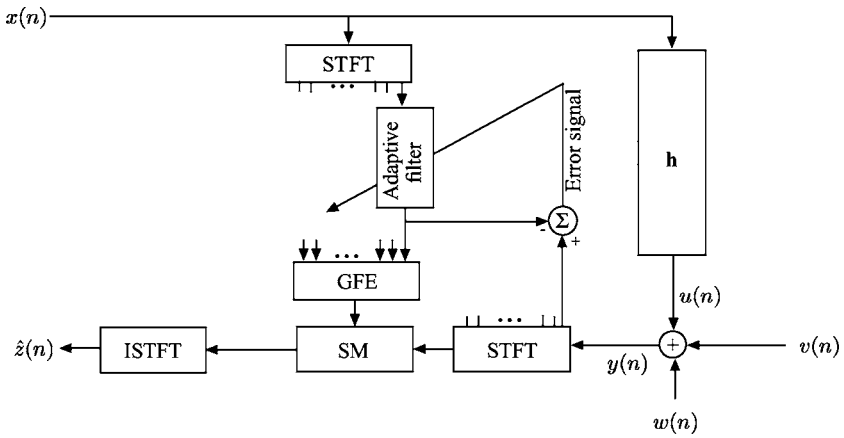

Fig. 2. Block diagram of the echo suppression algorithm by modifying the spectral magnitude, where STFT, GFE, SM, and ISTFT stand for short-time Fourier transform, gain filter estimation, spectral modification, and inverse short-time Fourier transform, respectively.

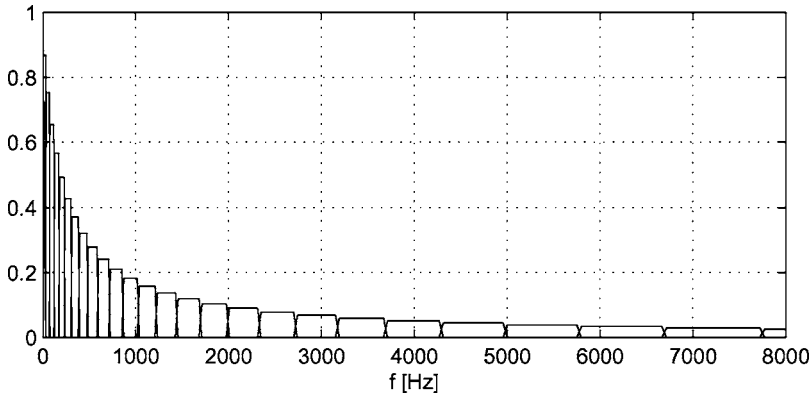

Fig. 3. Frequency response of an auditory filterbank following the ERB scale.

In the early stages of a human auditory system, the acoustic signal is decomposed into spectral components. This spectral decomposition is often modeled with an auditory filterbank, which consists of bandpass filters with nonuniform bandwidths [26]. An auditory filterbank can be viewed as a nonlinear mapping from the linear frequency to a warped frequency since the filterbank outputs are nonuniformly distributed along the frequency axis [27]. Commonly used nonlinear frequency scales describing such a mapping are the Bark scale [22] and the *equivalent rectangular bandwidth* (ERB) scale [28]. The frequency responses of an auditory filterbank with rectangular bandpass filters following the ERB scale are illustrated in Fig. 3. Note that with increasing frequency the frequency resolution of the auditory filterbank decreases. Speech and audio processing algorithms often take advantage of the specific frequency resolution of the auditory system for improving their performance, see, e.g., [29]–[31]. For example, in [31] spectral magnitude modification is applied to audio signals. More smoothing is applied at higher frequencies where the frequency resolution of the auditory system is lower for reducing artifacts.

Here, we propose to take into account the frequency resolution of an auditory system for the purpose of echo suppression. This is done by considering the spectral envelope, rather than the STFT magnitude or power spectra directly as SMMES does. The spectral envelope is computed such that they reflect the frequency resolution of the auditory system and is denoted as *auditory spectral envelope*. It will be shown that the gain filter computed in the domain of auditory spectral envelope changes as a function of frequency as smoothly as permitted by the frequency resolution of the auditory system. Particularly at higher frequencies, this results in a very smoothed gain filter. Compared with a nonsmoothed gain filter used in SMMES, this smoothed gain filter will introduce less artifacts to the outgoing signal. In addition, the auditory spectral envelope is represented with less parameters than a corresponding magnitude or power spectrum. Thus, the number of parameters that PAES needs to estimate is smaller than the number of parameters estimated by SMMES, resulting in a lower computational complexity.

PAES is illustrated in Fig. 4. Comparing Fig. 4 with Fig. 2, one can see the difference between SMMES and PAES. In brief, PAES estimates the echo and gain filter in an auditory spectral envelope space, while the SMMES approach estimates echo in the complex spectral domain and the gain filter in the spectral magnitude domain.

The key features of the proposed PAES are the estimation of auditory spectral envelope of the microphone signal, the adaptive estimation of the auditory spectral envelope of the echo signal, and the computation of the gain filter. In the following, these processing steps are described in detail.

1) Auditory Spectral Envelope Estimation: There are mainly two approaches to estimate the auditory spectral envelope. One is to estimate the spectral envelope using either

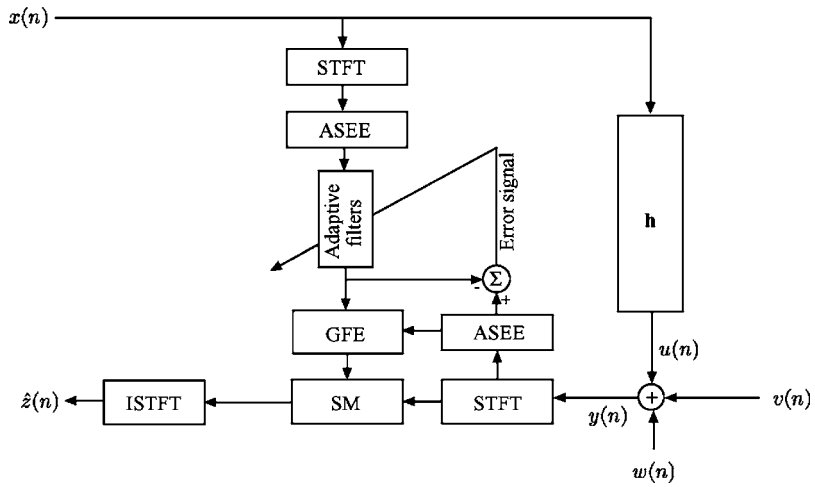

Fig. 4. Block diagram of the proposed PAES algorithm, where STFT, ASEE, GFE, SM, and ISTFT stand for short-time Fourier transform, auditory spectral envelope estimation, gain filter estimation, spectral modification, and inverse short-time Fourier transform, respectively.

the *linear prediction* (LP) technique [32] or the standard smoothing technique [33], and then project it to a nonuniform auditory scale, such as the Bark scale or the ERB scale. The other is to directly smooth the instantaneous power or magnitude spectrum over frequency with an auditory filterbank. The second approach usually has a lower computational complexity, and is the choice that we have taken here. The speech signal is transformed using STFT and the magnitude square is taken. The magnitude-square coefficients are then binned by correlating them with the frequency response of each bandpass filter of an auditory filterbank. Here binning means that each magnitude-square coefficient is multiplied by the corresponding bandpass filter gain and the results are accumulated. Similar processing has been widely used in speech and audio processing, see e.g. [29]–[31]. If we denote the frequency response of the i^{th} bandpass filter centered at ω_i as $\mathcal{W}_{\omega_i}(\omega)$ [$\mathcal{W}_{\omega_i}(\omega) \geq 0$], its output can be expressed as

$$\tilde{Y}_k(\omega_i) = \int_0^\pi \mathcal{W}_{\omega_i}(\omega) \left|Y_k(j\omega)\right|^2 d\omega \tag{17}$$

where the nonzero span of $\mathcal{W}_{\omega_i}(\omega)$ is centered around ω_i. The values obtained by (17), i.e., $\tilde{Y}_k(\omega_i)$ ($1 \leq i \leq I$, $I =$ number of bandpass filters), are frequency-domain samples representing the auditory spectral envelope. Substituting (11) into (17) yields

$$\tilde{Y}_k(\omega_i) = \tilde{U}_k(\omega_i) + \tilde{Z}_k(\omega_i). \tag{18}$$

It follows from the previous section on echo suppression that (15) can be used to recover the auditory spectral envelope of the signal $z(n)$, if an estimate of $\tilde{U}_k(\omega_i)$ can be obtained.

III. Adaptive Estimation of the Spectral Envelope of the Echo Signal

It can easily be derived from the given notation (the finite window-length effect is neglected) that

$$U_k(j\omega) = H(j\omega) X_k(j\omega) \tag{19}$$

and

$$\left|U_k(j\omega)\right|^2 = \left|H(j\omega)\right|^2 \left|X_k(j\omega)\right|^2 \tag{20}$$

where $H(j\omega)$ is the transfer function of the echo path. The i^{th} auditory spectral envelope sample can be expressed as

$$\begin{aligned}\tilde{U}_k(\omega_i) &= \int_0^\pi \mathcal{W}_{\omega_i}(\omega) \left|U_k(j\omega)\right|^2 d\omega \\ &= \int_0^\pi \mathcal{W}_{\omega_i}(\omega) \left|H(j\omega)\right|^2 \left|X_k(j\omega)\right|^2 d\omega. \end{aligned} \tag{21}$$

From (21), it can be shown that

$$\tilde{U}_k(\omega_i) = \mathcal{H}_i(k) \tilde{X}_k(\omega_i) \tag{22}$$

where $\mathcal{H}_i(k) = |H(j\xi_i)|^2$, $\tilde{X}_k(\omega_i) = \int_0^\pi \mathcal{W}_{\omega_i}(\omega) |X_k(j\omega)|^2 d\omega$, and $\xi_i \in [0, \pi]$.

Since the far-end signal $\mathbf{x}_k$ is available, $\tilde{X}_k(\omega_i)$ is known. If an estimate of $\mathcal{H}_i(k)$ can be obtained, then $\tilde{U}_k(\omega_i)$ can be computed using (22). Therefore, the estimation of $\tilde{U}_k(\omega_i)$ is essentially a matter of estimating $\mathcal{H}_i(k)$. Different estimation theories may be applied to measure $\mathcal{H}_i(k)$. In what follows, we describe several estimators for obtaining $\mathcal{H}_i(k)$.

A. Single-Tap Least Squares Estimator

The LS estimator is widely used in practice because it is easy to implement. It is derived from the minimization of a least-squares error criterion. Let us assume that the echo path does not change during P frames and define the error signal for the k^{th} frame and i^{th} auditory subband as

$$e_k(\omega_i) = \tilde{Y}_k(\omega_i) - \hat{\mathcal{H}}_i(k) \tilde{X}_k(\omega_i) \tag{23}$$

where $\hat{\mathcal{H}}_i(k)$ is a trial value of $\mathcal{H}_i(k)$. Consider the following cost function which is the arithmetic mean of $e_k^2(\omega_i)$ over P frames

$$J(\hat{\mathcal{H}}_i) = \sum_{n=k-P+1}^{k} e_n^2(\omega_i). \tag{24}$$

Minimization of (24) with respect to $\hat{\mathcal{H}}_i(k)$ gives the LS estimator

$$\hat{\mathcal{H}}_i(k) = \frac{\sum_{n=k-P+1}^{k} \tilde{Y}_n(\omega_i)\tilde{X}_n(\omega_i)}{\sum_{n=k-P+1}^{k} \tilde{X}_n^2(\omega_i)}. \quad (25)$$

Based on this estimator, the estimated spectral envelope samples of the echo signal are

$$\hat{\tilde{U}}_k(\omega_i) = \hat{\mathcal{H}}_i(k)\tilde{X}_k(\omega_i). \quad (26)$$

B. Multitap Least Squares Estimator

For the single-tap LS estimator, in each auditory subband, the echo path is modeled with a single coefficient. Its accuracy may not be sufficient due to the limited window length effect. The estimation accuracy, however, can be improved by considering a multitap LS estimator which involves an FIR filter per subband. Another benefit with a multitap estimator is that the channel estimate has a smaller variation than that achieved with a single-tap estimator, reducing the artifacts introduced during spectral manipulation. For a multitap estimator, the error signal for the k^{th} frame and i^{th} auditory subband is

$$e_k(\omega_i) = \tilde{Y}_k(\omega_i) - \sum_{n=0}^{Q-1} \hat{\mathcal{H}}_{i,n}(k)\tilde{X}_{k-n}(\omega_i). \quad (27)$$

This can be written in a vector-matrix form as

$$e_k(\omega_i) = \tilde{Y}_k(\omega_i) - \hat{\mathbf{H}}_i^T(k)\tilde{\mathbf{X}}_k(\omega_i) \quad (28)$$

where

$$\hat{\mathbf{H}}_i(k) = \left[\hat{\mathcal{H}}_{i,Q-1}(k), \hat{\mathcal{H}}_{i,Q-2}(k), \cdots, \hat{\mathcal{H}}_{i,0}(k)\right]^T$$
$$\tilde{\mathbf{X}}_k(\omega_i) = \left[\tilde{X}_{k-Q+1}(\omega_i), \tilde{X}_{k-Q+2}(\omega_i), \cdots, \tilde{X}_k(\omega_i)\right]^T \quad (29)$$

and Q is the order of the FIR filter. Again, if we assume that the echo path does not change during P frames, minimizing the cost function

$$J(\hat{\mathbf{H}}_i) = \sum_{n=k-P+1}^{k} e_n^2(\omega_i) \quad (30)$$

yields the multitap LS estimator:

$$\hat{\mathbf{H}}_i(k) = \mathbf{R}_i^{-1}\mathbf{r}_i \quad (31)$$

where

$$\mathbf{R}_i = \sum_{n=k-P+1}^{k} \tilde{\mathbf{X}}_n(\omega_i)\tilde{\mathbf{X}}_n^T(\omega_i)$$

and

$$\mathbf{r}_i = \sum_{n=k-P+1}^{k} \tilde{Y}_k(\omega_i)\tilde{\mathbf{X}}_n(\omega_i).$$

C. Adaptive Estimators

With the error signal defined in (28) [(23) is a particular case of (28)], an adaptive algorithm can be applied to search for the optimum $\hat{\mathbf{H}}_i(n)$. For example, the NLMS algorithm can be expressed as

$$\hat{\mathbf{H}}_i(k+1) = \hat{\mathbf{H}}_i(k) + \mu\frac{e_k(\omega_i)}{\tilde{\mathbf{X}}_k^T(\omega_i)\tilde{\mathbf{X}}_k(\omega_i)}\tilde{\mathbf{X}}_k(\omega_i) \quad (32)$$

where μ is the normalized step-size.

Once the adaptive filter converges, the spectral envelope sample of the echo signal can be computed as

$$\hat{\tilde{U}}_k(\omega_i) = \hat{\mathbf{H}}_i^T(k)\tilde{\mathbf{X}}_k(\omega_i). \quad (33)$$

D. Doubletalk Detection

In an AEC system, when there is presence of doubletalk, the near-end signal acts as uncorrelated noise, which is likely to cause the adaptive filter to diverge, resulting in insufficient echo cancellation. The most commonly used method to deal with doubletalk is to use a doubletalk detector. Whenever the presence of doubletalk is detected, the adaptive filter is frozen. Similarly, in the PAES algorithm, when there is doubletalk, the estimate of the auditory spectral envelope deviates from its true value, resulting incorrect amount of echo suppression. Therefore, it is important that we have a doubletalk controller operating in the sampled spectral envelope domain. We have investigated various doubletalk detection algorithms [3]–[7], [39] and found that the method presented in [39] is more straightforward to implement in the sampled spectral envelope domain, and therefore is adopted here. The detection accuracy of this approach may not necessarily be higher than those of the algorithms presented in [3]–[7]. However, it is out of the scope of this paper to discuss the accuracy of doubletalk detection.

E. Gain Filter Estimation

Given the estimated samples of the echo signal spectral envelope, i.e., $\hat{\tilde{U}}_k(\omega_i)$, it is easy to derive the corresponding gain filter at time instant k according to the parametric Wiener filtering technique described in Section II-C, i.e.,

$$G_k(\omega_i) = \left[\frac{\tilde{Y}_k^{\frac{\alpha}{2}}(\omega_i) - \beta\hat{\tilde{U}}_k^{\frac{\alpha}{2}}(\omega_i)}{\tilde{Y}_k^{\frac{\alpha}{2}}(\omega_i)}\right]^{\frac{1}{\alpha}}. \quad (34)$$

If we define the ratio between $\hat{\tilde{U}}_k(\omega_i)$ and $\tilde{Y}_k(\omega_i)$ as the a posteriori echo-to-signal ratio (ESR):

$$\gamma(\omega_i) = \sqrt{\frac{\hat{\tilde{U}}_k(\omega_i)}{\tilde{Y}_k(\omega_i)}} \quad (35)$$

it follows that

$$G_k(\omega_i) = \left[1 - \beta\gamma^{\alpha}(\omega_i)\right]^{\frac{1}{\alpha}}. \quad (36)$$

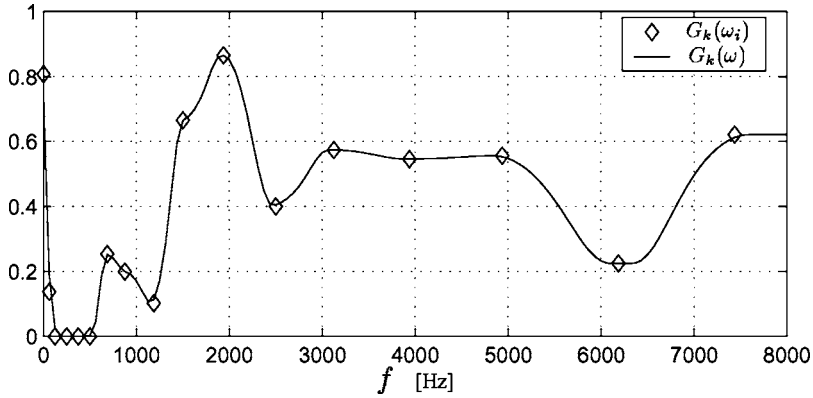

Fig. 5. Gain filter $G_k(\omega)$ (solid) is obtained by interpolating the sampled auditory spectral envelope gains $G_k(\omega_i)$ (diamonds).

By tuning α and β, we can control the amount of echo to be eliminated. It should be pointed out here that there are other ways to improve the gain filter [13]. Although important, finding the optimal gain filter is beyond the scope of this paper. Note that $G_k(\omega_i)$ is only a sampled version of the gain filter. The gain filter, $G_k(\omega)$, which is applied to modify the STFT spectrum, is computed by interpolating the estimated samples of the gain filter [i.e., $G_k(\omega_i)$] using an interpolation algorithm. Fig. 5 shows a numerical example of $G_k(\omega_i)$ and $G_k(\omega)$, where $G_k(\omega_i)$ is estimated according to (36) and $G_k(\omega)$ is obtained by interpolating $G_k(\omega_i)$ in the ERB-scale domain. Due to both the smoothing process and the multitap estimator, the estimate of $G_k(\omega)$ is found to change smoothly with respect to time and frequency. This makes artifacts (such as musical noise) resulting from the suppression algorithm less noticeable as compared to the SMMES method.

IV. Simulations and Results

Commonly used measures for assessing the performance of conventional AECs are the normalized misalignment given in (7) and the (normalized) *mean square error* (MSE), which is defined as

$$\text{MSE}(n) = \frac{\text{LPF}\left\{e^2(n)\right\}}{\text{LPF}\left\{y^2(n)\right\}} \tag{37}$$

where LPF denotes a lowpassfilter operation. This criterion can be directly used to evaluate the PAES algorithm, if $e(n)$ is replaced with the output signal of PAES. The convergence of the adaptive filters in PAES is assessed by examining the estimation mean square error of the echo spectral envelope, which is given as

$$\varepsilon(k) = \frac{\text{LPF}\left\{\sum_{i=1}^{I}\left|\tilde{U}_k(\omega_i) - \hat{\mathbf{H}}_i^T(k)\tilde{\mathbf{X}}_k(\omega_i)\right|^2\right\}}{\text{LPF}\left\{\sum_{i=1}^{I}\left|\tilde{U}_k(\omega_i)\right|^2\right\}} \tag{38}$$

or for the ith subband

$$\varepsilon_i(k) = \frac{\text{LPF}\left\{\left|\tilde{U}_k(\omega_i) - \hat{\mathbf{H}}_i^T(k)\tilde{\mathbf{X}}_k(\omega_i)\right|^2\right\}}{\text{LPF}\left\{\left|\tilde{U}_k(\omega_i)\right|^2\right\}}. \tag{39}$$

It is trivial to show that in the single-tap case (38) is equivalent to the misalignment criterion.

All simulations presented in this paper use the following common parameters: Sampling rate is 16 kHz; STFT window is a Hann window of size $W = 256$ (16 ms) with 50% overlap ($N = 128$); ambient noise $w(n)$ is a computer generated zero-mean white Gaussian process; SNR $= 30$ dB unless otherwise noted; near-end signal $v(n) = 0$ except in doubletalk simulation. SNR is defined as the ratio between the power of the near-end signal plus echo and that of the ambience noise. In case when $v(n) = 0$, it is the ratio between the power of the echo and that of the ambience noise. For representing the auditory spectral envelope $I = 17$ are used. This corresponds to using an auditory filterbank with bandpass filters being approximately 2-ERB wide. Informal listening revealed that choosing a higher frequency resolution does not notably improve performance. The bandpass filters are nonoverlapping. The bandwidths of the 17 bandpass filters expressed in STFT bins are: 1, 1, 2, 2, 2, 2, 3, 4, 5, 6, 8, 9, 12, 14, 18, 22, and 18, respectively. The last subband is less wide than the second last one because it is pruned at the Nyquist frequency. Impulse responses measured in the Bell Labs Varechoic Chamber [35], [36] are used as the true echo paths for the simulations.

A. Convergence of the Adaptive Estimator

The first experiment is carried out to assess the convergence properties of the adaptive estimator. The far-end signal $x(n)$ is a white Gaussian process. The near-end signal is zero [$v(n) = 0$], i.e., there is no doubletalk. The step-size of the NLMS algorithm for each auditory subband is chosen to be $\mu = 0.02$. Other simulation parameter values used are: $M = 4096$ and $Q = 2$. Fig. 6 shows ε and some arbitrarily selected ε_i, all as a function of time. We observe from Fig. 6 that the adaptive filters for all auditory subbands experience a similar convergence rate though they may have different steady-state ε_i. With the selected μ, they converge in approximately half a second. A faster convergence rate can be achieved by choosing a larger μ. However, this may result in a larger steady-state MSE.

Ambient noise, which is uncorrelated with the far-end signal, manifests as an offset in the power spectral domain [see (11)]. Therefore, it is expected to have some negative effect on the estimator. Fig. 7 shows ε of the adaptive estimator in different SNR conditions. Note that the noise effect does not severely

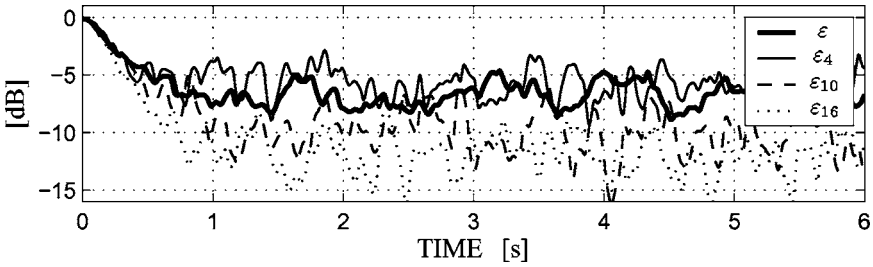

Fig. 6. ε and three arbitrarily selected ε_i. In this case, $i = 4$, 10, and 16, respectively.

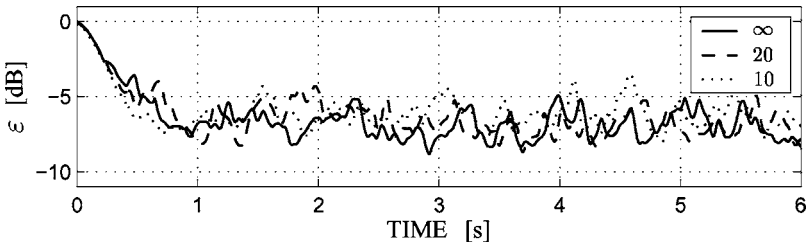

Fig. 7. ε in different SNR conditions: SNR $=$ 10 dB, 20 dB, and ∞ dB, respectively.

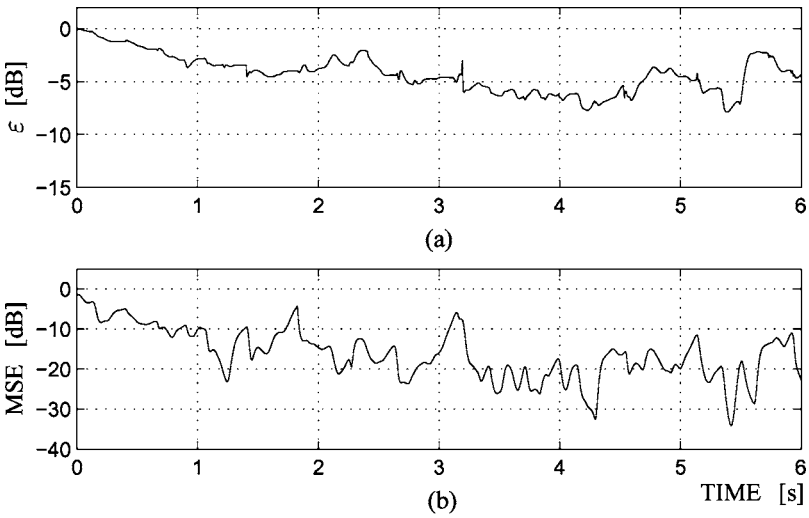

Fig. 8. Echo suppression performance. (a) ε versus time. (b) MSE versus time.

degrade the performance of the adaptive estimator when SNR is moderately high (e.g., SNR $>$ 10 dB). This indicates that the proposed algorithm is reasonably robust with respect to ambient noise.

B. Echo Suppression Performance

The second experiment evaluates the performance of PAES in a more realistic situation, where the far-end signal is a speech from a male talker. The simulation is conducted in the absence of doubletalk, i.e., $v(n) = 0$. Other simulation parameters are: $M = 4096$, $Q = 2$, and $\alpha = 1.0$. We compute the gain filter in the spectral magnitude domain. Our investigation shows that in most cases the echo is slightly underestimated. To have an effective echo suppression, we choose $\beta = 1.2$. The α and β parameters may be further optimized for a better performance. However, as we mentioned earlier, optimizing the gain filter is beyond the scope of this paper.

The results are plotted in Fig. 8, with Fig. 8(a) showing ε and Fig. 8(b) showing MSE, both as a function of time. We observe that when the adaptive filter converges, the MSE defined in (37) is about -20 dB or less. Informal listening test with our real-time PAES implementation (Section IV-F) shows that with such a degree of suppression, we do not hear residual echo. It should be pointed out here that in case more echo suppression is required, it can be achieved by controlling the α and β parameters. However, with stronger suppression, it may introduce a stronger distortion into the outgoing signal. Therefore, the selection of α and β is a tradeoff between echo attenuation and degree of distortion.

C. Doubletalk Situation

This simulation examines the performance of PAES in a doubletalk situation, assuming ideal doubletalk detection. To do so, a speech signal from a male talker is used as the near-end signal and a speech signal from another male talker is used as

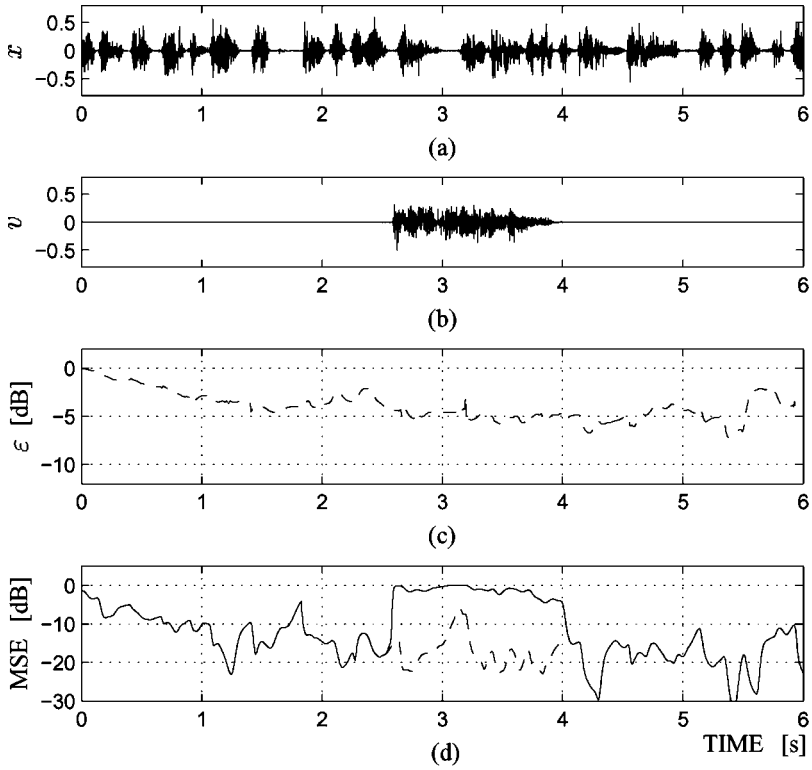

Fig. 9. Performance in the presence of doubletalk. (a) Far-end talker signal. (b) Near-end talker signal. (c) ε. (d) MSE.

the far-end signal. The doubletalk is active during the time interval from 2.5 to 4.0 s. The other parameters are the same as were used for the previous experiment. The adaptive filters are frozen in the time interval when the doubletalk is active.

The results are presented in Fig. 9. From Fig. 9(c), we notice that during the doubletalk period, even though the coefficients of the adaptive filters are not updated, the estimates of the echo spectral envelopes are still reasonably accurate since the estimation error does not notably increase. Fig. 9(d) shows two curves. The dashed line plots the MSE computed from (37) but the near-end signal (doubletalk) is not considered. This demonstrates the degree of echo suppression during doubletalk. We see that the curve does not increase, indicating that the echo component is successfully suppressed also during doubletalk. The solid line shows the same MSE, but this time the near-end signal is included. Note that during the doubletalk, the curve increases significantly, indicating that the doubletalk was not suppressed, as anticipated.

D. Comparison With AEC

1) Performance versus the Length of the Modeling Filter: For a conventional AEC, in order to achieve a reasonably good performance, the length of the modeling filter has to be long enough to capture most of the echo energy. If the true echo path impulse response is known a priori (e.g., in a simulation situation), the length of the modeling filter can be determined by examining the misalignment. Fig. 10 shows the misalignment as a function of the length of the modeling filter. We assume that the modeling filter $\hat{\mathbf{h}}$ is a perfect estimate of the true echo path $\mathbf{h}$ only ignoring its tail. This indeed shows the lower bound of the misalignment, which is achievable for a given length of the modeling filter. As can be seen, the misalignment decreases as the length of the modeling filter increases. It diminishes when the length of the adaptive filter approaches that of the true echo path. From Fig. 10, one can tell how many taps are needed to obtain a certain degree of echo cancellation. For instance, if $\epsilon < -20$ dB is to be achieved, at least $L = 3500$ taps have to be used for the modeling filter.

For the PAES algorithm, it is not obvious how many taps should be used. To find out how many taps are needed in practice, we performed an experiment to assess the effect of the number of taps on the mean square error of the echo estimates. The true echo path impulse response is the same as in the previous experiment (4096 taps). The far-end signal is a speech from a male talker. Fig. 11 shows ε for different numbers of taps. We observe that the performance of a 2-tap filter for each auditory subband is significantly better than that of a single-tap filter. Further increasing the number of taps yields some, but limited improvement over 2 taps. Several simulations in different environments were performed, the results confirm the above observation. Therefore, in the subsequent experiments, we use a 2-tap adaptive filter for each auditory subband.

2) Echo Suppression Performance: This experiment was carried out to compare a conventional NLMS-based AEC to the proposed PAES algorithm. Again, speech from a male talker is used as the far-end signal. A measured impulse response truncated to 512 taps is used as the true echo path, such that the NLMS-based conventional AEC can converge relatively fast.

The step-size parameter for the NLMS algorithm of the conventional AEC is $\mu = 0.2$, and for the proposed scheme $\mu = 0.02$ for all auditory subbands. Other PAES parameters are: $Q = 2$, $\alpha = 1.0$, and $\beta = 1.2$. The results are presented in

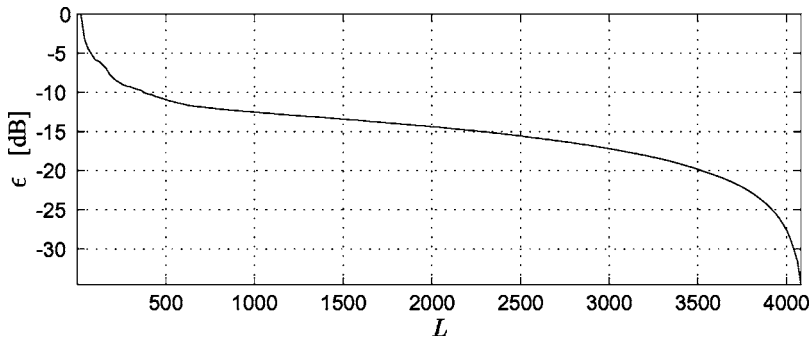

Fig. 10. Lower bound for the normalized misalignment ϵ defined in (7) as a function of the length of the modeling filter. This plot was computed using a measured room impulse response that has $M = 4096$ taps.

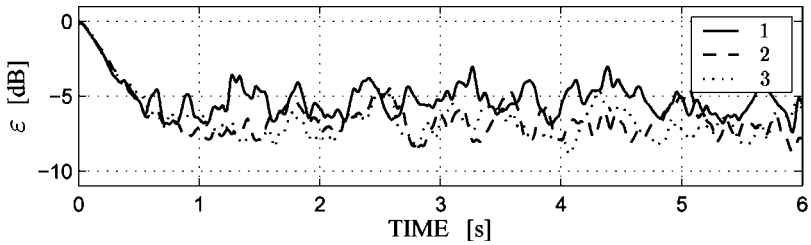

Fig. 11. ε versus time for different adaptive filter lengths: $Q = 1$, 2, and 3, respectively.

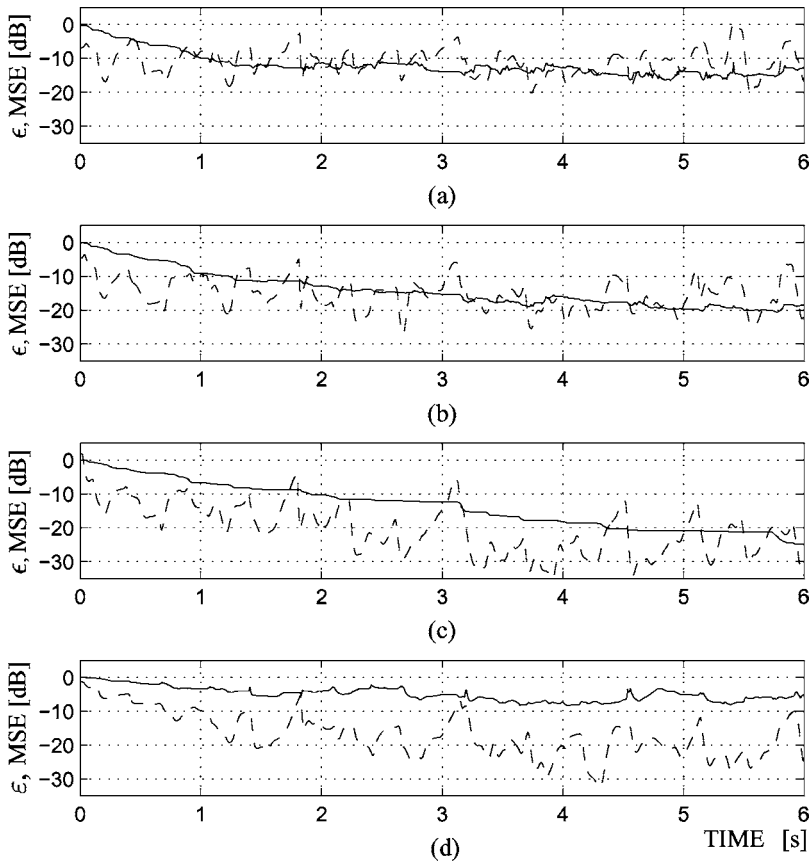

Fig. 12. Comparison between PAES and a conventional NLMS-based AEC: (a) ϵ and MSE for AEC with $L = 128$ and $\mu = 0.2$; (b) ϵ and MSE for AEC with $L = 256$ and $\mu = 0.2$; (c) ϵ and MSE for AEC with $L = 512$ and $\mu = 0.2$; (d) ε and MSE for PAES with $I = 17$, $Q = 2$, and $\mu = 0.02$.

Fig. 12, where Fig. 12(a), (b), and (c) shows the performance of the conventional AEC for adaptive filters of length 128, 256, and 512, respectively, and Fig. 12(d) shows the performance for the proposed PAES algorithm.

As can be seen, the echo cancellation performance of the conventional AEC is improved as the length of the modeling filter L increases. The proposed PAES performs as good as the conventional AEC considering a modeling filter with the same length

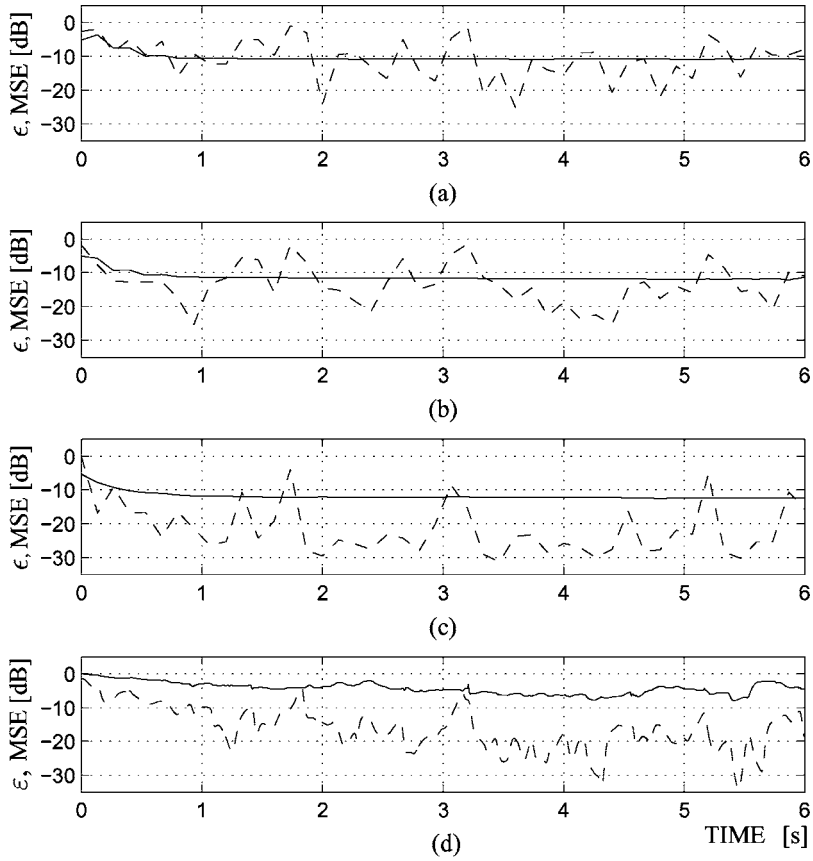

Fig. 13. Comparison between PAES and a conventional FLMS-based AEC. (a) ϵ and MSE for AEC with $L = 512$ and $\mu = 0.3$. (b) ϵ and MSE for AEC with $L = 1024$ and $\mu = 0.3$. (c) ϵ and MSE for AEC with $L = 2048$ and $\mu = 0.3$. (d) ε and MSE for PAES with $I = 17$, $Q = 2$, and $\mu = 0.02$.

as the echo path impulse response, i.e., $L = 512$. This indicates the effectiveness of the proposed scheme.

The previous simulation was repeated, using a self-orthogonalizing frequency-domain LMS (FLMS) algorithm [9]. The same measured room impulse response of length 4096 as used in previous experiments was used. The simulation was carried out with modeling filter lengths of 512, 1024, and 2048. Other FLMS parameters are: Step-size $\mu = 0.3$, and exponential forgetting factor for spectral density estimation $\gamma = 0.7$. All the other parameters, including PAES parameters, were the same as in the previous simulation.

The results are shown in Fig. 13, where Fig. 13(a), (b), and (c) show the performance of the FLMS-based AEC for adaptive filters of length 512, 1024, and 2048, respectively, and Fig. 13(d) shows the performance for the proposed PAES algorithm. It can be seen that initially the FLMS-based AEC converges faster than PAES. However, once converged, the PAES algorithm has as good echo suppression performance as the FLMS algorithm with a 2048-tap modeling filter. The FLMS algorithm, with a shorter modeling filter (1024 and 512) perform worse than PAES. We notice that PAES performs similarly for both short (Fig. 12) and long (Fig. 13) echo path with the same parameters and number of estimation filter taps.

3) Robustness: We also compared a conventional NLMS-based AEC and PAES for their robustness with respect to echo path changes. We repeated the previous simulations for PAES and NLMS-based AEC with the same parameters and adaptive

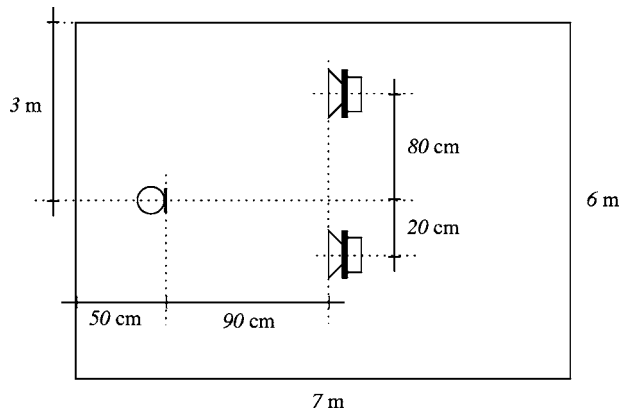

Fig. 14. Echo path changes are modeled by toggling between two echo path impulse responses measured with the shown setup in the Bell Labs Varechoic chamber.

modeling filter for the case when the conventional AEC has 512 taps. In an attempt to simulate echo path changes, we toggled $\mathbf{h}$ every 1.5 s between two echo path impulse responses that were measured [35] in the Bell Labs Varechoic chamber with a geometrical setup as shown in Fig. 14. Note that the measuring setup is nonsymmetric and thus the delay of the direct path is different for the two echo path impulse responses. The normalized misalignment given in (7) between the two impulse responses is 2.9 dB.

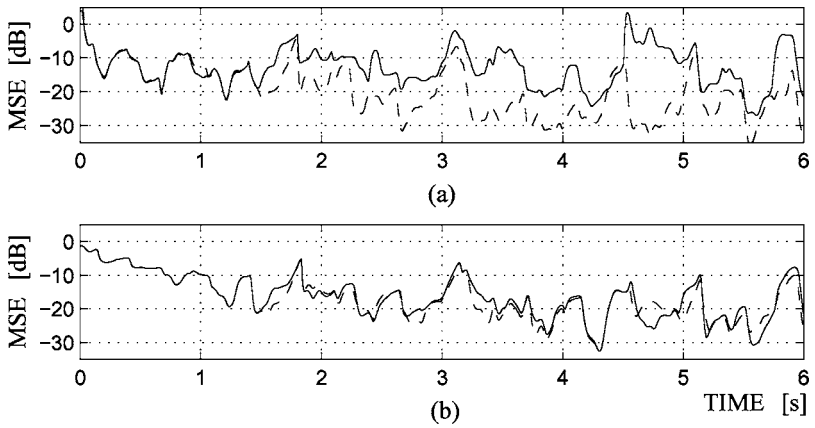

Fig. 15. Comparison between PAES and AEC for their robustness with respect to echo path changes: (a) MSE of the conventional AEC (dashed line: no echo path change; solid line: two echo paths are toggled every 1.5 s.); (b) MSE of the PAES algorithm (dashed line: no echo path change; solid line: two echo paths are toggled every 1.5 s).

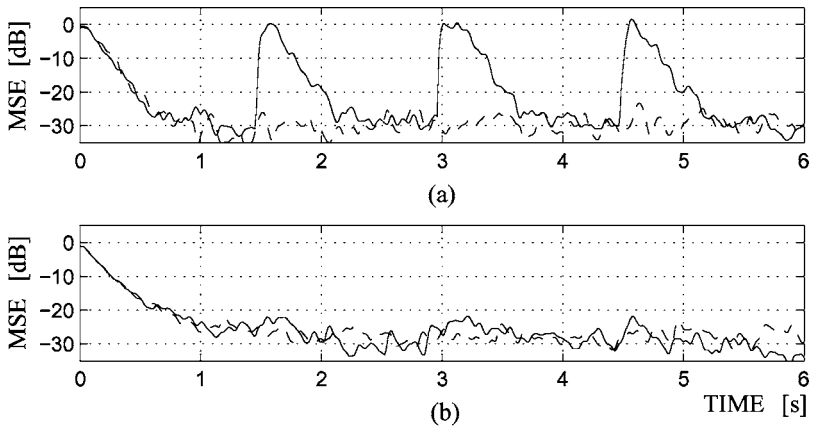

Fig. 16. Comparison between (a) AEC and (b) PAES for their robustness with respect to echo path changes when the far-end signal is a wide-band Gaussian process.

The results are presented in Fig. 15, where Fig. 15(a) shows the performance of the conventional AEC with and without the echo path changes and Fig. 15(b) shows the corresponding performance of PAES. As opposed to the conventional AEC, the two MSE curves for PAES are close to each other, indicating that the performance of the PAES algorithm is nearly unaffected by the echo path changes.

Fig. 16 shows the results of a similar simulation, but this time the far-end signal is a white Gaussian random process. We see that once the echo path changes, the MSE of AEC increases significantly until the adaptive filter reconverges. The performance of the PAES algorithm does not change much when the echo path changes, indicating that the PAES method is more robust to echo path changes than the conventional AEC. Due to its robustness, PAES does not need a post-processor for eliminating residual echoes, whereas AEC needs such a post-processor.

E. Computational Complexity

In this section we compare PAES with AEC in terms of their computational complexity. The time- and frequency-domain NLMS and fast RLS (FRLS) adaptive algorithms are considered for AEC. We also include the complexity for the SMMES scheme. Table I summarizes the number of multiplications needed for each algorithm. For brevity, the detailed calculation of the complexity is omitted.

TABLE I
NUMBER OF MULTIPLICATION OPERATIONS NEEDED BY DIFFERENT ALGORITHMS. FD DENOTES FREQUENCY DOMAIN. THE LAST ROW SHOWS THE COMPLEXITY OF ONLY THE STFT THAT IS USED FOR THE FREQUENCY-DOMAIN ALGORITHMS. THE RIGHT COLUMN SHOWS A NUMERICAL EXAMPLE (SEE TEXT FOR THE SPECIFIC PARAMETERS USED)

Algorithm	multiplies/sample	multiplies/sample
NLMS-based AEC	$2L$	2048
FRLS-based AEC	$32L$	32768
NLMS-based FD-AEC	$\text{STFT} + \frac{4WQ}{N}$	112
FRLS-based FD-AEC	$\text{STFT} + \frac{64WQ}{N}$	1072
NLMS-based SMMES	$\text{STFT} + \frac{4WQ}{N} + \frac{4W}{N}$	120
FRLS-based SMMES	$\text{STFT} + \frac{64WQ}{N} + \frac{4W}{N}$	1080
NLMS-based PAES	$\text{STFT} + \frac{2IQ}{N} + \frac{4W}{N}$	57
STFT	$\frac{3W[\log_2(W)+1]}{N}$	48

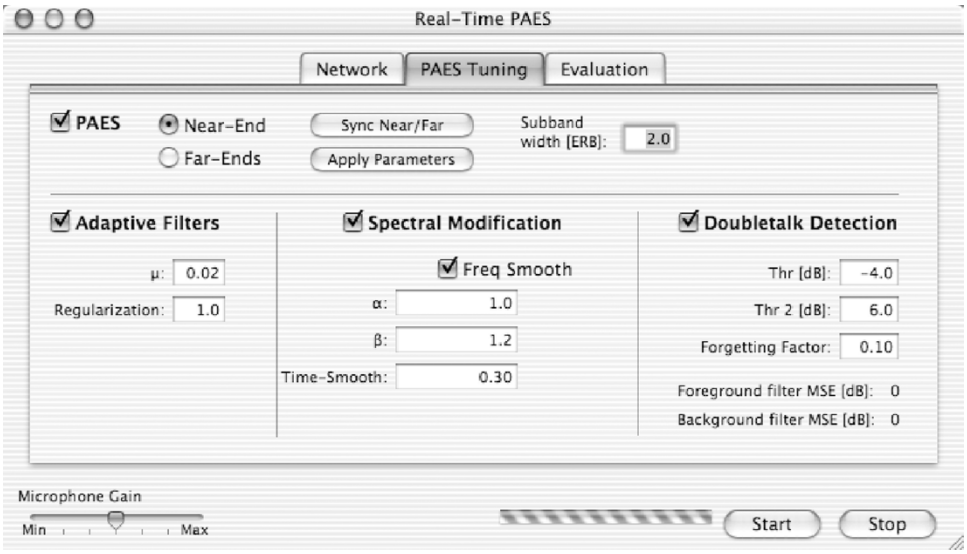

Fig. 17. Client software of the PAES real-time implementation. Parameters can be tuned in real-time.

The time-domain methods perform echo cancellation on a sample-by-sample basis. Their complexities depend on the length of the modeling filter, and the complexity of the real-valued adaptive algorithms [37] as well. The frequency-domain approaches carry out cancellation/suppression on a frame-by-frame basis. The computational burden depends on parameters such as the window size (W), the window hop size (N), the length of the adaptive filter in each subband or frequency bin (Q), and the complexity of the complex-valued adaptive algorithms [38]. The second column of Table I shows the average number of multiplications per sample. Here we assume that all frequency-domain algorithms use the same STFT, whose complexity is shown in the last row of Table I.

Compared with an AEC using a frequency-domain adaptive algorithm, the SMMES method requires the same number of multiplications to estimate the echo signal; but for each frame it needs $3W$ additional multiplications to compute the power spectra of the microphone signal, the loudspeaker signal, and the estimated echo signal; another W multiplications are required for applying the gain filter. Here we assume that the gain filter computation is implemented with a lookup table, which does not require any multiplications.

The complexity of PAES consists of the multiplications necessary for the STFT and $2W$ multiplications for computing the power spectra of the microphone signal and the loudspeaker signal. Additionally, PAES applies I real-valued NLMS algorithms every frame. The interpolation applied for obtaining the gain filter to compute the spectral envelope samples requires an additional $2W$ multiplications. Note that the number of adaptive filter taps Q is chosen smaller for PAES than for frequency-domain AECs or SMMES.

The third column of Table I shows a numerical example for the computational complexity of various algorithms. The parameter values used include: $L = 1024$, $W = 256$, $N = 128$, and $Q = \lfloor L/N \rfloor = 8$ for frequency-domain AECs (FD-AEC) and the SMMES method with the same Q as suggested by [21], and $Q = 2$ and $I = 17$ for the PAES algorithm. As seen, the PAES scheme has lower computational complexity than any of the other studied methods.

F. Real-Time Implementation

A real-time system using the proposed PAES algorithm was developed to control the echo effect occurring in VoIP and other voice communication systems. A graphical user interface as shown in Fig. 17 enables users to adjust such parameters as the step-size, regularization in the NLMS algorithm [1], [2], and α and β for computing the gain filter. Furthermore, the user can set a time-smoothing factor, which in turn controls a lowpass filter to smooth the gain filter. Such an operation can further reduce artifacts resulting from the spectral modification. By switching between the near-end and far-end buttons, the user can change the parameters for both the near-end and the far-end systems.

We set up a tele-conferencing system with PCs and external loudspeakers. Several tele-conferencing sessions were conducted and a number of participants were invited to the sessions to evaluate the PAES system. The PAES system performs well in various environments and was judged favorably by all participants.

V. Conclusions

Conventional acoustic echo cancelers eliminate the undesired echoes by modeling the echo path impulse response with an adaptive FIR filter. They generally involve extensive computations due to the fact that a large number of taps are required for the modeling filter. In this paper, the problem of echo cancellation was studied in a spectral envelope space from a practical point of view. Aiming at eliminating the undesired echo effect and achieving a low complexity, we proposed a PAES. Three issues were addressed, which include representation of spectral envelopes, adaptive estimation of the spectral envelope of the echo signal, and suppressing echo using spectral modification. Compared with the conventional AEC, the proposed PAES algorithm offers several advantages. It has much lower complexity since fewer parameters need to be estimated. It is

more robust with respect to minor echo path changes. In addition, since some human auditory aspects are incorporated in this new framework, it has the potential for improved perceptual quality. We also compared PAES with a suppression algorithm performed in the magnitude spectral domain. The new algorithm is not only more computationally efficient, but also suffers from fewer artifacts due to its smooth gain filter.

Extensive numerical studies were conducted. Various impulse responses of different lengths, measured from the Bell Labs Varechoic Chamber, together with a male speech signal and white Gaussian noise were used to evaluate the performance of the proposed PAES. The results support the appealing features we claimed for the proposed algorithm. A real-time system based on PAES was implemented and tested in various conditions. Informal subjective listening indicates that the proposed algorithm yields as good echo suppression performance as an AEC when the near-end talker does not move. When the near-end talker moves, which causes some minor changes to the echo path, the proposed algorithm delivers a better performance than the AEC since the latter often has audible residual echos.

ACKNOWLEDGMENT

The authors thank T. Gänsler, P. Kroon, F. Wallin, and the two anonymous reviewers for their valuable suggestions for improvement of this manuscript.

REFERENCES

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

[2] S. Haykin, *Adaptive Filter Theory (Third Edition)*. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[3] D. L. Duttweiler, "A twelve-channel digital echo canceler," *IEEE Trans. Commun.*, vol. 26, pp. 647–653, May 1978.

[4] H. Ye and B.-X. Wu, "A new double-talk detection algorithm based on the orthogonality theorem," *IEEE Trans. Commun.*, vol. 39, no. 11, pp. 1542–1545, Nov. 1991.

[5] T. Gänsler, M. Hansson, C.-J. Ivarsson, and G. Salomonsson, "A double-talk detector based on coherence," *IEEE Trans. Commun.*, vol. 44, no. 11, pp. 1421–1427, Nov. 1996.

[6] G.-T. Ryu, D.-W. Kim, J.-G. Choe, D.-S. Kim, S.-H. Kim, and H.-D. Bae, "Double talk detection in adaptive echo canceller using fuzzy logic," in *Proc. ICSP*, 1996, pp. 1643–1646.

[7] J. Benesty, D. R. Morgan, and J. H. Cho, "A new class of doubletalk detectors based on cross-correlation," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 2, pp. 168–172, Mar. 2000.

[8] M. Dentino, J. McCool, and W. Widrow, "Adaptive filtering in the frequency domain," *Proc. IEEE*, vol. 66, pp. 1658–1659, Dec. 1978.

[9] E. R. Ferrara, "Fast implementation of LMS adaptive filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, pp. 474–475, Aug. 1980.

[10] J.-S. Soo and K. K. Pang, "Multidelay block frequency domain adaptive filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-38, pp. 373–376, Feb. 1990.

[11] D. Mansour and A. H. Gray Jr., "Unconstrained frequency-domain adaptive filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-30, pp. 726–734, Oct. 1982.

[12] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Nov. 1979.

[13] W. Etter and G. S. Moschytz, "Noise reduction by noise-adaptive spectral magnitude expansion," *J. Audio Eng. Soc.*, vol. 42, pp. 341–349, May 1994.

[14] R. Le Bouquin Jeannès, P. Scalart, G. Faucon, and C. Beaugeant, "Combined noise and echo reduction in hands-free systems: a survey," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 808–820, Nov. 2001.

[15] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 245–256, Jul. 2002.

[16] P. Vary, "Noise suppression by spectral magnitude estimation-mechanism and theoretical limits," *Signal Proces.*, vol. 8, pp. 387–400, Jul. 1985.

[17] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[18] H. Pobloth and W. B. Kleijn, "On phase perception in speech," in *Proc. IEEE ICASSP*, vol. 1, Mar. 1999, pp. 29–32.

[19] E. J. Diethorn, "Noise reduction techniques with a single microphone," in *Acoustic Signal Processing for Telecommunication*, S. L. Gay and J. Benesty, Eds. Boston, MA: Kluwer, 2000.

[20] J. Chen, Y. Huang, and J. Benesty, "Filtering techniques for noise reduction and speech enhancement," in *Adaptive Signal Processing: Applications to Real World Problems*, Y. Huang and J. Benesty, Eds. Berlin: Springer, 2003.

[21] C. Avendano, "Acoustic echo suppression in the STFT domain," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2001.

[22] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*. New York: Springer, 1999.

[23] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech, Audio Process.*, vol. 2, pp. 115–132, Jan. 1994.

[24] B. Carnero and A. Drygajlo, "Perceptual speech coding and enhancement using frame-synchronized fast wavelet packet transform algorithms," *IEEE Trans. Signal Process.*, vol. 47, pp. 1622–1635, Jun. 1999.

[25] D. Sinha, J. D. Johnston, S. Dorward, and S. Quackenbush, "The perceptual audio coder (PAC)," in *The Digital Signal Processing Handbook*, V. Madisetti and D. B. Williams, Eds. Boca Raton, FL: CRC, IEEE Press, 1997, ch. 42.

[26] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception, Proc. 9th Int. Symp. Hearing*, 1992, pp. 429–446.

[27] J. O. Smith and J. S. Abel, "Bark and ERB bilinear transform," *IEEE Trans. Speech, Audio Process.*, vol. 7, pp. 697–708, Nov. 1999.

[28] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 1990.

[29] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[30] S. Young, "A review of large-vocabulary continuous-speech recognition," *IEEE Signal Process. Mag.*, vol. 13, pp. 45–57, Sep. 1996.

[31] C. Faller and F. Baumgarte, "Binaural cue coding—part II: schemes and applications," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, Nov. 2003.

[32] A. El-Jaroudi and J. Makhoul, "Discrete all pole modeling," *IEEE Trans. Signal Process.*, vol. 39, pp. 411–423, Feb. 1991.

[33] D. B. Paul, "The spectral envelope estimation vocoder," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-29, pp. 786–7941, Aug. 1981.

[34] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Wetterling, *Numerical Recipes in C*. Cambridge, MA: Cambridge Univ. Press, 1988.

[35] A. Härmä, T. Lokki, and V. Pulkki, "Drawing quality maps of the sweet spot and its surroundings in multichannel reproduction and coding," in *Proc. AES 21st Conf. on Architectural Acoustics and Sound Reinforcement*, Jun. 2002, pp. 317–325.

[36] W. C. Ward, G. W. Elko, R. A. Kubli, and W. C. McDougald, "The new varechoic chamber at AT&T Bell Labs," in *Proc. Wallance, Clement, Sabine Centennial Symposium*, Woodbury, NY, 1994, pp. 343–346.

[37] J. Benesty, F. Amand, A. Gilloire, and Y. Grenier, "Adaptive filtering algorithms for stereophonic acoustic echo cancellation," in *Proc. IEEE ICASSP*, May 1995, pp. 3099–3102.

[38] P. Eneroth, S. L. Gay, T. Gänsler, and J. Benesty, "A real-time implementation of a stereophonic acoustic echo canceler," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, July 2001.

[39] K. Ochiai, T. Araseki, and T. Ogihara, "Echo canceler with two echo path models," *IEEE Trans. Commun.*, vol. 25, no. 6, pp. 589–595, Jun. 1977.

Christof Faller received the M.S. (Ing.) degree in electrical engineering from ETH, Zurich, Switzerland, in 2000 and the Ph.D. degree in computer and communication sciences from EPFL, Lausanne, Switzerland, in 2004.

During his studies, he was an independent consultant for Swiss Federal Labs, applying neural networks to process parameter optimization of sputtering processes. He spent one year at the Czech Technical University (CVUT), Prague. In 2000, he became a Consultant for the Speech and Acoustics Research Department, Bell Laboratories, Lucent Technologies. After one and a half years of consulting, partially from Europe, he became a Member of Technical Staff, focusing on new techniques for audio coding applied to digital satellite radio broadcasting. At the Lucent spin-off, Agere Systems, he developed algorithms for parametric coding of multichannel audio signals, echo control, and other communications-related audio applications. He is currently with the Audiovisual Communications Laboratory at EPFL Lausanne.

Jingdong Chen (M'99) received the B.S. degree in electrical engineering and the M.S. degree in array signal processing from the Northwestern Polytechnic University in 1993 and 1995, respectively, and the Ph.D. degree in pattern recognition and intelligence control from the Chinese Academy of Sciences in 1998. His Ph.D. research focused on speech recognition in noisy environments, involving the study and proposal of several techniques covering speech enhancement and HMM adaptation by signal transformation.

From 1998 to 1999, he was with ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan, where he conducted research on speech synthesis and speech analysis, as well as objective measurements for evaluating speech synthesis. He then joined the Griffith University, Brisbane, Australia, as a Research Fellow, where he engaged in research in robust speech recognition, signal processing, and discriminative feature representation. From 2000 to 2001, he was with ATR Spoken Language Translation Research Laboratories, Kyoto, where he conducted research in robust speech recognition and speech enhancement. He joined Bell Laboratories, Murray Hill, NJ, as a Member of Technical Staff in July 2001. His current research interests include adaptive signal processing, speech enhancement, adaptive noise/echo cancellation, microphone array signal processing, signal separation, and source localization.

Dr. Chen is the recipient of a 1998–1999 research grant from the Japan Key Technology Center and the 1996–1998 President's Award from the Chinese Academy of Sciences.