ROBUST ACOUSTIC ECHO CANCELLATION IN THE SHORT-TIME FOURIER TRANSFORM DOMAIN USING ADAPTIVE CROSSBAND FILTERS

Jason Wung, Daniele Giacobello, Joshua Atkins

Beats Electronics, LLC, 1601 Cloverfield Blvd., Suite 5000N, Santa Monica, CA 90404, USA
{jason.wung, daniele.giacobello, josh.atkins}@beatsbydre.com

ABSTRACT

This paper presents a robust acoustic echo cancellation (AEC) system in the short-time Fourier transform (STFT) domain using adaptive crossband filters. The STFT-domain AEC allows for a simpler system structure compared to the traditional frequency-domain AEC, which normally requires several applications of the discrete Fourier transform (DFT) and the inverse DFT, while the robust AEC (RAEC) allows for continuous and stable filter updates during double talk without freezing the adaptive filter. The RAEC and the STFT-domain AEC have been investigated in the past in separate studies. In this work we propose a novel algorithm that combines the advantages of both approaches for robust update of the adaptive crossband filters even during double talk. Experimental results confirm the benefit of incorporating the robustness constraint for the adaptive crossband filters and show improved performance in terms of the echo reduction and the predicted sound quality.

Index Terms— Acoustic echo cancellation, adaptive filter, short-time Fourier transform (STFT), crossband filters, robustness

1. INTRODUCTION

The acoustic echo reduction system, which often consists of an acoustic echo cancellation (AEC) unit and a residual echo suppression (RES) unit as shown in Figure 1, is generally required for hands-free teleconferencing. The AEC unit cancels the linear part of the echo while the RES unit suppresses the tail or nonlinear part due to modeling mismatch of the adaptive filter and ambience noise. Traditionally, a least-mean-square (LMS) algorithm is used for the adaptation of the filter coefficients. However, the number of filter coefficients in the AEC can be several thousands, which increases the computation cost significantly and slows down the convergence rate. Frequency-domain adaptive filtering (FDAF) [1] has been proposed to reduce computational complexity and improve the convergence rate by taking advantage of the circular convolution property of the discrete Fourier transform (DFT), where time-domain filtering is achieved by multiplying frequency-domain coefficients. Large processing delay may be introduced due to the block processing nature of the FDAF, especially when the filter length is long. To reduce this processing delay, the multi-delay filter [2] was proposed to segment the adaptive filter into smaller blocks. However, the FDAF-type algorithms still require several applications of the DFT and the inverse DFT (IDFT) to enforce the gradient constraint.

System identification in the short-time Fourier transform (STFT) domain has recently been proposed in [3, 4] and applied to the AEC problem in [5]. Instead of relying on the DFT and the IDFT for the gradient constraint, a cross-multiplicative transfer function approximation is introduced, where data from adjacent frequency bins are used for the system identification. Compared to the FDAF-type algorithms, system identification in the STFT domain requires only one DFT and one IDFT for the analysis and the synthesis, respectively, of each signal. Furthermore, system identification in the STFT domain can be potentially integrated with a RES unit and/or a noise suppressor unit, which typically operates in the STFT domain [6–8]. However, during double-talk situations, a more traditional approach, i.e., the use of a double-talk detector (DTD) to freeze the filter adaptation, was often used to prevent divergence of the adaptive filter.

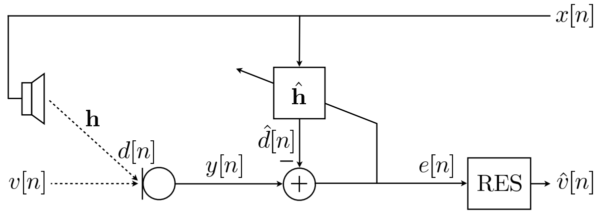

Fig. 1. An acoustic echo reduction system with an adaptive filter $\hat{\mathbf{h}}$ for acoustic echo cancellation and a residual echo suppressor.

A robust acoustic echo cancellation (RAEC) system [9, 10] has recently been proposed to avoid freezing the adaptive filter during double talk. The RAEC utilizes the error recovery nonlinearity (ERN), which "enhances" the filter estimation error prior to the adaptation, and the noise-robust adaptive step-size [11] with block-iterative adaptation that enables the recovery of lost convergence speed due to the aggressive step-size control. Therefore, the RAEC allows for continuous and stable adaptation of the filter even during double talk without requiring a DTD. However, the RAEC approach was previously applied only to the FDAF-type algorithms.

In this paper, we propose a novel algorithm for RAEC in the STFT domain using adaptive crossband filters. The proposed algorithm uses the STFT-domain AEC framework and incorporates components from the RAEC algorithm to allow for robust update of the adaptive crossband filters during double talk. Even though the processing is done in the frequency domain, traditional FDAF-type algorithms eventually output a time-domain signal that has to be transformed again to the STFT domain for subsequent nonlinear processing, i.e., RES and noise suppression. The STFT-domain AEC framework has the advantage of a unified system architecture when combining both the AEC and the RES into a single STFT-domain processing framework, while the RAEC components enforce robustness of the algorithm under noisy conditions without a DTD.

This paper is organized as follows. We review the STFT-domain AEC with adaptive crossband filters in Section 2. The STFT-domain RAEC is proposed in Section 3. Experimental evaluation and the conclusion are discussed in Section 4 and 5.

2. PROBLEM BACKGROUND

2.1. Short-Time Fourier Transform Domain Processing

The STFT-domain processing framework can be formulated as follows. Given a signal $x[n]$, the signal is transformed to the STFT domain by

$$X_k[m] = \sum_{n=0}^{N-1} x[n + mR] w_A[n] \omega_N^{kn},$$

where k is the frequency index, m is the frame index, N is the frame size, R is the frame shift size, $w_A[n]$ is an analysis window of size N, and $\omega_N \equiv \exp(-j\frac{2\pi}{N})$. The frequency-domain coefficients can be synthesized back by applying the inverse STFT (ISTFT)

$$x[n] = \sum_{m} \sum_{k=0}^{N-1} X_k[m] w_S[n - mR] \omega_N^{-k(n-mR)},$$

where $w_S[n]$ is a synthesis window. For perfect reconstruction of the signal $x[n]$, the analysis and synthesis windows must satisfy the so-called *completeness condition*, i.e.,

$$\sum_{m} w_A[n + mR] w_S[n + mR] = 1, \quad \forall n.$$

2.2. Acoustic Echo Cancellation with Crossband Filters

A single-channel AEC system operating in the STFT domain is shown in Figure 2. Let $y[n]$ be the near-end microphone signal, which consists of the near-end speech and/or noise $v[n]$ mixed with the acoustic echo $d[n] = h[n] * x[n]$, where $h[n]$ is the impulse response of the system, $x[n]$ is the far-end reference signal, and $*$ is the convolution operator. Let $\mathbf{x}[m] = [x[mR], \ldots, x[mR + N - 1]]^{\mathrm{T}}$ be the m^{th} reference signal vector, $\mathbf{w}_A = [w_A[0], \ldots, w_A[N-1]]^{\mathrm{T}}$ be the analysis window vector, $(\mathbf{F})_{k+1,n+1} = \omega_N^{kn}$, $k, n = 0, \ldots, N-1$ be the $N \times N$ discrete Fourier transform (DFT) matrix, and $\underline{\mathbf{x}}[m] = \mathbf{F}(\mathbf{w}_A \circ \mathbf{x}[m]) = [X_0[m], \ldots, X_{N-1}[m]]^{\mathrm{T}}$ be the DFT of the windowed reference signal vector, where $\circ$ is the Hadamard (element-wise) product operator and $\{\cdot\}^{\mathrm{T}}$ is the transpose operator. The acoustic echo signal can be modeled in the STFT domain as [3]

$$\underline{\mathbf{d}}[m] = \sum_{i=0}^{M-1} \mathbf{H}_i[m-1] \underline{\mathbf{x}}[m-i], \qquad (1)$$

where $\underline{\mathbf{d}}[m]$ is the DFT of the m^{th} frame echo signal, $\mathbf{H}_i$ is the i^{th} impulse response matrix, and M is the filter length in the STFT domain. If the impulse response matrix is diagonal, (1) reduces to the multiplicative transfer function approximation [4] which may not be accurate due to the finite analysis window length. The modeling accuracy can be improved by adding $2K$ cross-terms, or $2K$ off-diagonal bands, around the main diagonal terms of $\mathbf{H}$ without significantly increasing the computational complexity.

Let $\hat{\mathbf{H}}$ be the adaptive filter matrix using $2K+1$ diagonal bands. The estimated echo can be written as $\hat{\underline{\mathbf{d}}}[m] = \sum_{i=0}^{M-1} \hat{\mathbf{H}}_i[m-1]\underline{\mathbf{x}}[m-i]$, and the adaptive filter matrix can be updated using

$$\hat{\mathbf{H}}_i[m] = \hat{\mathbf{H}}_i[m-1] + \mathbf{G} \circ \Delta\hat{\mathbf{H}}_i[m], \quad i = 0, \ldots, M-1,$$

where $\Delta\hat{\mathbf{H}}_i[m]$ is an update matrix for the filter coefficients matrix and $\mathbf{G} = \sum_{k=-K}^{K} \mathbf{P}^k$ is a matrix that selects the $2K+1$ diagonal bands with $\mathbf{P}$ being a permutation matrix defined as

$$\mathbf{P} \equiv \begin{bmatrix} 0 & 0 & \ldots & 0 & 1 \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 \\ 0 & \ldots & 0 & 1 & 0 \end{bmatrix}.$$

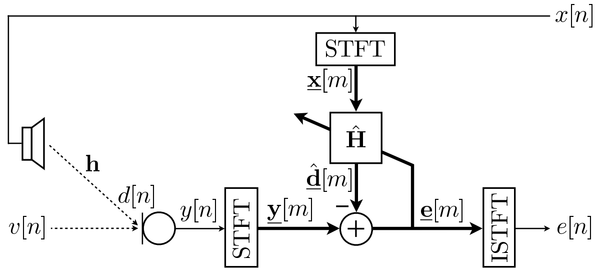
$x[n]$
STFT
$\underline{\mathbf{x}}[m]$
$\hat{\mathbf{H}}$
$\mathbf{h}$
$\hat{\underline{\mathbf{d}}}[m]$
$d[n]$
$y[n]$
$-$
$v[n]$
STFT
$\underline{\mathbf{y}}[m]$
$\underline{\mathbf{e}}[m]$
ISTFT
$e[n]$

Fig. 2. The STFT-domain AEC, where the STFT block represents windowing and transforming to the frequency domain. Note that a RES block, omitted here for simplicity, can potentially be inserted before the ISTFT for combined AEC and RES in the STFT domain.

The matrix $\mathbf{G}$ limits the number of crossband filters that are useful for system identification in the STFT domain since increasing the number of crossband filters does not necessarily imply a lower steady-state error [3, 5].

The update matrix based on the LMS algorithm is given by

$$\Delta\hat{\mathbf{H}}_i^{\mathrm{LMS}}[m] = \mu \underline{\mathbf{e}}[m] \underline{\mathbf{x}}^{\mathrm{H}}[m-i], \qquad (2)$$

where $\underline{\mathbf{e}}[m] = \underline{\mathbf{y}}[m] - \hat{\underline{\mathbf{d}}}[m]$ is the error signal vector in the STFT domain, $\mu > 0$ is a step-size, and $\{\cdot\}^{\mathrm{H}}$ is the Hermitian transpose operator. Instead of using only the diagonal terms, which is normally done in the FDAF-type algorithms, (2) takes into account the contribution of the cross-frequency components of the reference signal without relying on the DFT and the IDFT for canceling the aliased components, allowing for a simplified system architecture. Detailed analysis of acoustic echo cancellation with adaptive crossband filters using the LMS algorithm can be found in [5].

We note that in the presence of near-end speech/noise $v[n]$, the error signal vector is given by

$$\underline{\mathbf{e}}[m] = \underline{\mathbf{v}}[m] + \underline{\mathbf{d}}[m] - \hat{\underline{\mathbf{d}}}[m] = \underline{\mathbf{v}}[m] + \underline{\mathbf{b}}[m], \qquad (3)$$

where $\underline{\mathbf{v}}[m]$ and $\underline{\mathbf{b}}[m]$ is the noise vector and the noise-free error signal vector, respectively, in the STFT domain. Since the error signal vector $\underline{\mathbf{e}}[m]$ deviates from the true, noise-free, residual echo signal vector $\underline{\mathbf{b}}[m]$, the adaptive filter may diverge from the optimal solution due to the near-end interference, in which case a DTD is often required to freeze the adaptation of the filter and may lead to suboptimal AEC performance.

3. PROPOSED ALGORITHM

3.1. Robust Acoustic Echo Cancellation

The RAEC system [9, 10] has recently been proposed to allow for robust update of the adaptive filter coefficients even under heavy near-end interference. The RAEC utilizes ERN, which tries to recover the true error signal prior to the adaptive filter update and can be expressed as a nonlinear clipping function [12], i.e.,

$$\phi(E_k[m]) = \begin{cases} \frac{T_k[m]}{|E_k[m]|} E_k[m], & |E_k[m]| \geq T_k[m], \\ E_k[m], & \text{otherwise}, \end{cases} \tag{4}$$

that limits the error signal when its magnitude is above a certain threshold $T_k[m]$. The threshold is estimated based on the near-end signal statistics and is approximated by $T_k[m] = \sqrt{S_{ee,k}[m]}$ with

$$S_{ee,k}[m] \equiv \mathrm{E}\{|E_k[m]|^2\} \approx \beta S_{ee,k}[m-1] + (1-\beta)|E_k[m]|^2,$$

where $S_{ee,k}[m]$ is the power spectral density (PSD) of the error signal, $\mathrm{E}\{\cdot\}$ is the expectation operator, and $0 \ll \beta < 1$ is a forgetting factor. The nonlinear clipping function (4) is one of the several nonlinear functions investigated in [10] that gives the best performance, where the residual echo signal $b[n]$ and the near-end signal $v[n]$ are assumed to be Gaussian distributed and Laplace distributed, respectively. Detailed discussion of the convergence behavior with different choices of nonlinearity functions under different signal model assumptions can be found in [10, 13, 14].

The regularization parameter plays an important role in adaptive algorithms [15]. Without good regularization, an adaptive algorithm may not behave properly under noisy conditions. A fixed regularization term is traditionally applied to the step-size of the normalized LMS (NLMS) algorithm to stabilize the filter update

$$\mu_k^{\mathrm{NLMS}}[m] = \mu \frac{1}{S_{xx,k}[m] + \delta}, \tag{5}$$

where δ is the fixed regularization term. In conjunction with the ERN, the RAEC incorporates a noise-robust adaptive step-size from [11] that is given in the frequency domain as [9]

$$\mu_k[m] = \mu \frac{S_{xx,k}[m]}{S_{xx,k}^2[m] + \gamma S_{ee,k}^2[m]} = \mu \frac{1}{S_{xx,k}[m] + \delta_k[m]}, \tag{6}$$

where γ is a tuning parameter and $\delta_k[m] = \gamma \frac{S_{ee,k}^2[m]}{S_{xx,k}[m]}$ is a frequency-dependent regularization term. The adaptive step-size in (6), similar in form to (5), can be viewed as using the frequency-dependent regularization term that scales down the step-size automatically when the near-end signal $v[n]$ is large. Detailed implementation of the RAEC system and its improved version through the system approach can be found in [9, 10, 12].

3.2. Robust Adaptive Crossband Filters

The application of ERN to the STFT-domain AEC is straightforward since the function of ERN is to limit the effect of noise on the true error signal. From (3) we note that the error enhancement procedure is not changed in the STFT domain since the near-end interference is still additive in the STFT domain regardless of the echo cancellation framework. The additive noise assumption is often seen in the signal enhancement community. All derivation of the ERNs, or noise-suppressing nonlinearities, in [10] is still valid in the STFT-domain processing framework, where the nonlinearity is applied to the error signal in each frequency bin. No cross-frequency component of the error signal needs to be considered when applying the ERN to the update of the adaptive crossband filters.

Application of the noise-robust adaptive step-size to the STFT-domain AEC is more complicated since the step-size in (6) depends on the PSDs of both the reference signal and the error signal. To simplify the problem, we first consider the step-size for the NLMS algorithm in (5), which can be written in the vector form as

$$\underline{\mathbf{n}}[m] = (\underline{\mathbf{s}}_{xx}[m] + \delta \mathbf{1}_{N\times 1})^{\circ(-1)},$$

where $\{\cdot\}^{\circ(-1)}$ is the Hadamard (element-wise) inverse operator, $\mathbf{1}_{N\times 1} = [1, \ldots, 1]^{\mathrm{T}}$, and $\underline{\mathbf{s}}_{xx}[m] = \mathrm{E}\{\underline{\mathbf{x}}[m] \circ \underline{\mathbf{x}}^*[m]\} \equiv [S_{xx,0}[m], \ldots, S_{xx,N-1}[m]]^{\mathrm{T}}$ is the PSD vector of the reference signal with $\{\cdot\}^*$ being the element-wise complex conjugate operator. The LMS update matrix in (2) can be rewritten for the NLMS update as

$$\Delta\hat{\mathbf{H}}_i^{\mathrm{NLMS}}[m] = \mu \mathbf{e}[m](\underline{\mathbf{n}}[m] \circ \underline{\mathbf{x}}[m-i])^{\mathrm{H}}, \tag{7}$$

where the reference signal is normalized by its signal power before being multiplied by the error signal. Note that each element of the NLMS update matrix in (7) is given by

$$(\Delta\hat{\mathbf{H}}_i^{\mathrm{NLMS}}[m])_{k+1,l+1} = \mu \frac{E_k[m] X_l^*[m-i]}{S_{xx,l}[m] + \delta}. \tag{8}$$

Given (5), (6), and (8), the extension of the noise-robust adaptive step-size to the STFT-domain crossband filters can be viewed as adding a *cross-frequency dependent* regularization term $\delta_{k,l}[m] = \gamma \frac{S_{ee,k}^2[m]}{S_{xx,l}[m]}$ instead of the fixed regularization in (8), and the update can be modified as (with the ERN applied)

$$(\Delta\hat{\mathbf{H}}_i[m])_{k+1,l+1} = \mu \frac{\phi(E_k[m]) X_l^*[m-i]}{S_{xx,l}[m] + \delta_{k,l}[m]}.$$

Therefore, we propose the noise-robust adaptive step-size for the STFT-domain AEC in the matrix form as

$$(\mathbf{M}[m])_{k+1,l+1} = \frac{S_{xx,l}[m]}{S_{xx,l}^2[m] + \gamma S_{ee,k}^2[m]},$$

and the update matrix for the STFT-domain RAEC is given by

$$\Delta\hat{\mathbf{H}}_i[m] = \mu \mathbf{M}[m] \circ \{\phi(\underline{\mathbf{e}}[m]) \underline{\mathbf{x}}^{\mathrm{H}}[m-i]\},$$

where $\phi(\underline{\mathbf{e}}[m]) \equiv [\phi(E_0[m]), \ldots, \phi(E_{N-1}[m])]^{\mathrm{T}}$ is the estimate of the true error signal vector after applying ERN. The proposed STFT-domain RAEC algorithm with adaptive crossband filters is summarized in Table 1.

4. EXPERIMENTAL EVALUATION

The impulse responses were measured through two *Beats Pill™* speakers that were spaced 1 meter apart. The sound pressure level (SPL) of the two speakers were calibrated to 85 $\mathrm{dB_C}$ at 1 meter away with a -20 dBFS narrowband (500 Hz to 2 kHz) pink noise. One of the speakers was used for playing the near-end speech while the other one playing the far-end speech. The microphone was placed closely to the original microphone position of one of the speakers to measure the room impulse response $\mathbf{h}$ and the impulse response from the other speaker to the microphone. With the measured impulse response, the SPL of the echo signal was about 20 dB stronger than the near-end speech.

Speech files and noise files from the ITU-T P.501 test signals [16] were randomly selected for the near-end speech plus noise and the far-end speech to generate the database for our experiment. The noise was added to the near-end speech with a segmental signal-to-noise ratio (SSNR) of -5, 0, 5, and 10 dB. The near-end speech plus noise and the far-end speech were constantly overlapped from

Table 1. RAEC in the STFT domain with adaptive crossband filters.

Definitions

$$(\mathbf{F})_{k+1,n+1} = \omega_N^{kn} \equiv e^{-j\frac{2\pi}{N}kn}, \quad k, n = 0, \ldots, N-1$$

$$\mathbf{G} = \sum_{k=-K}^{K} \mathbf{P}^k, \quad \mathbf{P} \equiv \begin{bmatrix} \mathbf{0}_{1\times N-1} & 1 \\ \mathbf{I}_{N-1\times N-1} & \mathbf{0}_{N-1\times 1} \end{bmatrix}$$

$$\phi(\underline{\mathbf{e}}[m]) \equiv [\phi(E_0[m]), \ldots, \phi(E_{N-1}[m])]^{\mathrm{T}}$$

Echo cancellation

$$\underline{\mathbf{x}}[m] = \mathbf{F}(\mathbf{w}_A \circ [x[mR], \ldots, x[mR+N-1]]^{\mathrm{T}})$$

$$\underline{\mathbf{y}}[m] = \mathbf{F}(\mathbf{w}_A \circ [y[mR], \ldots, y[mR+N-1]]^{\mathrm{T}})$$

$$\underline{\mathbf{e}}[m] = \underline{\mathbf{y}}[m] - \sum_{i=0}^{M-1} \hat{\mathbf{H}}_i[m-1]\underline{\mathbf{x}}[m-i]$$

Filter adaptation

$$\underline{\mathbf{s}}_{xx}[m] = \beta\underline{\mathbf{s}}_{xx}[m-1] + (1-\beta)(\underline{\mathbf{x}}[m] \circ \underline{\mathbf{x}}^*[m])$$

$$\underline{\mathbf{s}}_{ee}[m] = \beta\underline{\mathbf{s}}_{ee}[m-1] + (1-\beta)(\underline{\mathbf{e}}[m] \circ \underline{\mathbf{e}}^*[m])$$

$$\phi(E_k[m]) = \begin{cases} \frac{\sqrt{S_{ee,k}[m]}}{|E_k[m]|} E_k[m], & |E_k[m]| \geq \sqrt{S_{ee,k}[m]} \\ E_k[m], & \text{otherwise} \end{cases}$$

$$(\mathbf{M}[m])_{k+1,l+1} = \frac{S_{xx,l}[m]}{S_{xx,l}^2[m] + \gamma S_{ee,k}^2[m]}, \quad k, l = 0, \ldots, N-1$$

$$\Delta\hat{\mathbf{H}}_i[m] = \mu\mathbf{M}[m] \circ \{\phi(\underline{\mathbf{e}}[m])\underline{\mathbf{x}}^{\mathrm{H}}[m-i]\}, \quad i = 0, \ldots, M-1$$

$$\hat{\mathbf{H}}_i[m] = \hat{\mathbf{H}}_i[m-1] + \mathbf{G} \circ \Delta\hat{\mathbf{H}}_i[m], \quad i = 0, \ldots, M-1$$

the very beginning to simulate the continuous double-talk scenario. 100 utterances were generated for the simulation with an averaged length of about 40 seconds for each utterance.

The frame size of the STFT was $N = 256$ samples with a 50% overlap. The analysis and synthesis windows were chosen to be the square root of a Hann window. The parameters for the STFT-domain RAEC were chosen to be $M = 10$, $\beta = 0.98$, and $\gamma = 1$, $\mu = 0.03/(1 + K)$, where $K = 0, 1, 2, 4$. To compare the echo cancellation performance with and without the robustness constraint, the traditional NLMS step-size (5) without ERN was also used. The regularization parameter for the NLMS step-size was $\delta = 10^{-6}$.

Figure 3 shows a box plot of the mean opinion score (MOS) measured by the Perceptual Evaluation of Speech Quality (PESQ) [17]. Evaluation was done using the last 10 seconds to ensure the adaptive filters were stabilized. To evaluate the AEC performance only, no noise suppression was applied, and all processed signals were compared to the clean near-end speech signal. The label "None" represents the microphone signal itself without any processing. The label "CB" represents the crossband filters with the traditional NLMS update, and the number appended at the end of each label represents K. Similarly, the label "R" represents the RAEC in the STFT domain. In general, the mean and median MOS's with the robustness contraint outperform the traditional NLMS update with a fixed regularization term. We observe that with increasing number of crossbands, the MOS of the traditional NLMS decreases while that of the RAEC increases. We note that the RAEC performs the best around $K = 2$, which is consistent with [5] where it was found that increasing the cross-terms past the optimal value degrades the AEC performance. A similar evaluation using the Perceptual Objective Listening Quality Assessment (POLQA) [18] was also conducted with similar results to the PESQ.

Figure 4 shows the true echo return loss enhancement (TERLE) obtained from one of the utterances processed with R2, CB0, and CB2, as well as the microphone signal and the near-end speech plus noise signal with −5 dB SSNR. R2 and CB0 were compared since they were the best performing settings for the STFT-domain AEC with and without the robustness constraint, respectively. The TERLE is defined as

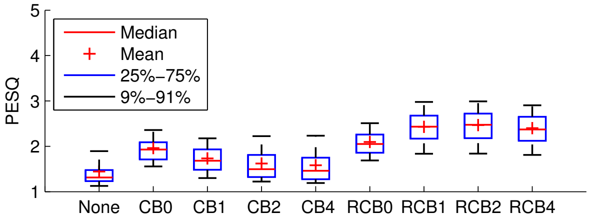

Fig. 3. MOS results using PESQ.

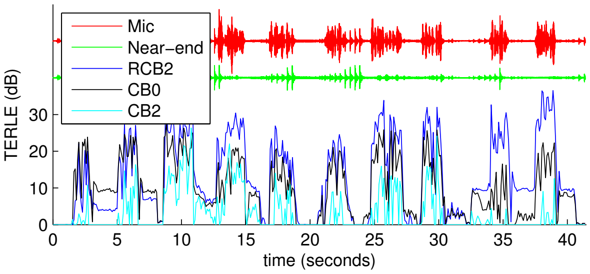

Fig. 4. TERLE plot comparing the STFT-domain AEC with and without the robustness constraint.

$$\text{TERLE (dB)} \equiv 10\log_{10}\left(\frac{\sum_n |y[n] - v[n]|^2}{\sum_n |e[n] - v[n]|^2}\right),$$

i.e., the ERLE measured after the near-end speech and noise are subtracted from both the microphone signal and the error signal. We note that with more crossbands, the initial convergence rate is slightly slower than the one without any crossband filter, as reported in [5]. However, the steady-state performance of R2 is much better than CB0, providing as much as 15 dB more echo reduction. We note that while CB2 is still able to cancel the echo in the beginning, the TERLE fluctuates throughout the utterance and decreases dramatically after 30 seconds, indicating that the adaptive crossband filters are diverging during double talk. R2 on the other hand provides much stabler TERLE.

5. CONCLUSION (RELATION TO PRIOR WORK)

The STFT-domain AEC framework [3–5] provides an alternative to the traditional FDAF algorithms [1, 2] and has the potential to simplify the system architecture when combined with RES [6–8]. On the other hand the robustness constraint [9–14] stabilizes the adaptive filter update of the traditional FDAF algorithms without the requirement of a DTD. We present in this paper a novel algorithm that combines the advantages of both the simplicity of the STFT-domain AEC framework and the robustness constraint for the adaptive crossband filter update even during continuous double talk. By correctly modifying the update equation for the adaptive crossband filters, we achieve superior echo reduction and stabler steady-state performance as verified by TERLE and PESQ.

6. REFERENCES

[1] J. J. Shynk, “Frequency-domain and multirate adaptive filtering,” *IEEE Signal Process. Mag.*, vol. 9, no. 1, pp. 14–37, 1992.

[2] J. S. Soo and K. K. Pang, “Multidelay block frequency domain adaptive filter,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 2, pp. 373–376, 1990.

[3] Y. Avargel and I. Cohen, “System identification in the short-time Fourier transform domain with crossband filtering,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1305–1319, 2007.

[4] Y. Avargel and I. Cohen, “On multiplicative transfer function approximation in the short-time Fourier transform domain,” *IEEE Signal Process. Lett.*, vol. 14, no. 5, pp. 337–340, 2007.

[5] Y. Avargel and I. Cohen, “Adaptive system identification in the short-time Fourier transform domain using cross-multiplicative transfer function approximation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 162–173, 2008.

[6] W. L. B. Jeannes, 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, 2001.

[7] 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, 2002.

[8] C. Faller and J. Chen, “Suppressing acoustic echo in a spectral envelope space,” *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1048–1062, 2005.

[9] T. S. Wada and B.-H. Juang, “Acoustic echo cancellation based on independent component analysis and integrated residual echo enhancement,” in *Proc. IEEE WASPAA*, pp. 205–208, 2009.

[10] T. S. Wada and B.-H. Juang, “Enhancement of residual echo for robust acoustic echo cancellation,” *IEEE Trans. Speech Audio Process.*, vol. 20, no. 1, pp. 175–189, 2012.

[11] A. Hirano and A. Sugiyama, “A noise-robust stochastic gradient algorithm with an adaptive step-size suitable for mobile hands-free telephones,” in *Proc. IEEE ICASSP*, vol. 2, pp. 1392–1395, 1995.

[12] J. Wung, T. S. Wada, B.-H. Juang, B. Lee, M. Trott, and R. W. Schafer, “A system approach to acoustic echo cancellation in robust hands-free teleconferencing,” in *Proc. IEEE WASPAA*, pp. 101–104, 2011.

[13] T. S. Wada and B.-H. Juang, “Enhancement of residual echo for improved frequency-domain acoustic echo cancellation,” in *Proc. IEEE WASPAA*, pp. 175–178, 2007.

[14] T. S. Wada and B.-H. Juang, “Towards robust acoustic echo cancellation during double-talk and near-end background noise via enhancement of residual echo,” in *Proc. IEEE ICASSP*, pp. 253–256, 2008.

[15] J. Benesty, C. Paleologu, and S. Ciochina, “On regularization in adaptive filtering,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1734–1742, 2011.

[16] ITU-T Recommendation P.501, “Test signals for use in telephonometry,” 2012.

[17] ITU-T Recommendation P.862, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs,” 2001.

[18] ITU-T Recommendation P.863, “Perceptual objective listening quality assessment,” 2011.