CROSSBAND FILTERING FOR STEREOPHONIC ACOUSTIC ECHO SUPPRESSION

Chul Min Lee[⋆], *Jong Won Shin*[†], *Yu Gwang Jin*[⋆], *Jeoung Hun Kim*[⋆] *and Nam Soo Kim*[⋆]

[⋆]Department of Electrical and Computer Engineering and INMC
Seoul National University, Seoul 151-742, Korea
[†]School of Information and Communications
Gwangju Institute of Science and Technology, Gwangju, Korea
E-mail: cmlee@hi.snu.ac.kr, jwshin@gist.ac.kr, {ygjin, jhkim}@hi.snu.ac.kr, nkim@snu.ac.kr

ABSTRACT

In this paper, we propose a novel stereophonic acoustic echo suppression (SAES) technique based on crossband filtering in the short-time Fourier transform (STFT) domain. The proposed algorithm considers spectral correlations among components in adjacent frequency bins, and estimates the extended power spectral density (PSD) matrices and cross PSD vectors from the signal statistics for more precise echo estimation. In the STFT domain, the echo spectra are estimated by performing the technique without any distinguishable double-talk detector. According to the experimental results, the proposed algorithm has been found to show better performances compared with the conventional SAES method.

Index Terms— Stereophonic acoustic echo suppression, crossband filtering, spectral correlations, signal-to-echo ratio, echo cancellation

1. INTRODUCTION

In the last few decades, much work has been dedicated to acoustic echo cancellation, which reduces effects of acoustic echo caused by the loudspeaker signals picked up by the microphones [1–9]. In particular, the increasing use of teleconferencing systems has led to the requirement of faster and more reliable acoustic echo cancellation algorithms. Most of the traditional stereo acoustic echo cancellation algorithms are based on an adaptive filters for tracking several echo paths [1–3]. However, because of the strong cross-correlation between the stereo signals, these approaches require some form of various de-correlation techniques [2–4] which demand substantial complexity as the pre-processing procedure and cause distortion of the reproduced signal.

To avoid the disadvantages of the de-correlation methods, a stereophonic acoustic echo suppression algorithms was presented recently [5]. This method estimates echo spectra and utilizes them to obtain *a priori* and *a posteriori* signal-to-echo ratio (SER) information which are exploited by various techniques for single-channel acoustic echo suppression in the short-time Fourier transform (STFT) domain [6–9]. The spectral gains were computed through Wiener filtering using estimated *a priori* SNR. Even though any distinguishable double-talk detector was not applied to this framework, the approach performed well in double-talk duration.

Inspired by the work [5], we propose a crossband filtering as an improved SAES algorithm in the STFT domain. The work by Avargel and Cohen [10] provided the theoretical background that crossband filters are needed when finite length windows are used. Taking accounts of spectral correlation among the adjacent frequency bins, we estimate the extended power spectral density (PSD) matrices and cross PSD vectors comprising those spectral components and obtain more accurate echo spectra. In addition, the echo overestimation control matrices are introduced to suppress the residual echoes. From the experimental results, the proposed algorithm showed better performances compared with the conventional SAES method.

2. PROPOSED SAES BASED ON CROSSBAND FILTERING

In Fig. 1, a stereophonic acoustic echo canceller is shown. To solve the stereophonic acoustic echo problem in this work, we concentrate on only one of the microphones because the technique can be applied on each microphone in parallel. Let $y(n)$ denote the input signal in the receiving room. Then it can be described as

$$y(n) = \sum_{i=1}^{2} h_i(n) * x_i(n) + s(n) \qquad (1)$$

where $h_i(n)$ represents the acoustic echo path from the i-th loudspeaker to the microphone and $s(n)$ is the near-end sig-

This research was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012R1A2A2A01045874) and by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2013-H0301-13-4005) supervised by the NIPA(National IT Industry Promotion Agency).

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

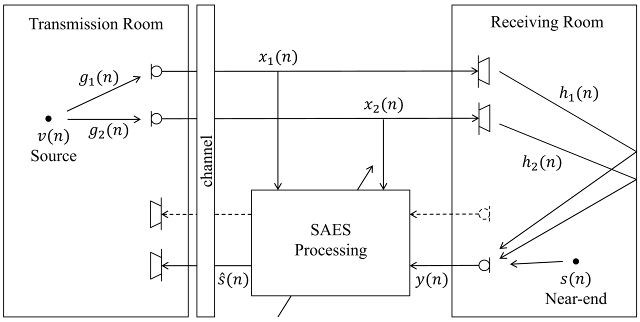


Fig. 1. Schematic diagram of a stereophonic acoustic echo canceler.

nal. Both $x_1(n)$ and $x_2(n)$ at time index n are the far-end signals which are yielded by the source signal $v(n)$ via the room impulse responses (RIRs) $g_1(n)$ and $g_2(n)$ in the transmission room.

According to [10], a linear system in the STFT domain can be modeled more accurately by the crossband filtering. In order to take the spectral correlations among the frequency bins into account in the proposed algorithm, we introduce the augmented vector for each far-end signal as follows:

$$\begin{aligned}\underline{\mathbf{X}}_i(n,k) =& [X_i(n,k-K)\ \dots\ X_i(n,k)\\ &\dots\ X_i(n,k+K)] \quad (i=1,2)\end{aligned} \tag{2}$$

where $X_i(n,k)$ is the STFT coefficient of the far-end signal $x_i(n)$ for the k-th frequency bin at the n-th frame. The augmented vector defined in (2) consists of the $(2K+1)$ adjacent frequency bins. Then, the dimension of this augmented vector becomes $M = 2K+1$.

Performing the STFT on the both sides of (1) and considering the crossband filtering, $Y(n,k)$ can be modeled as

$$Y(n,k) = \sum_{i=1}^{2} \underline{\mathbf{H}}_i^{\#}(n,k)\underline{\mathbf{X}}_i(n,k) + S(n,k) \tag{3}$$

where $Y(n,k)$, $\underline{\mathbf{H}}_i(n,k)$, and $S(n,k)$ denote the STFT coefficient of $y(n)$, the crossband filters which represent the echo paths from $\underline{\mathbf{X}}_i(n,k)$ to $Y(n,k)$, and the STFT coefficient of the near-end signal $s(n)$, respectively and the superscript $^{\#}$ denotes the conjugate transpose.

We denote the echo spectra correlated with $x_1(n)$ as $D_1(n,k)$ and the component correlated with $x_2(n)$ but uncorrelated with $x_1(n)$ as $D_2(n,k)$. Then,

$$D_i(n,k) = \underline{\mathbf{H}}_i^{\#}(n,k)\underline{\mathbf{X}}_i(n,k) \quad (i=1,2) \tag{4}$$

The optimal weight vectors $\widehat{\underline{\mathbf{H}}}_1(n,k)$ and $\widehat{\underline{\mathbf{H}}}_2(n,k)$ can be obtained by jointly minimizing the mean-square error (MSE) criteria as

$$J_1 = E[|Y(n,k) - \underline{\mathbf{H}}_1^{\#}(n,k)\underline{\mathbf{X}}_1(n,k)|^2], \tag{5}$$

$$J_2 = E[|Y_1(n,k) - \underline{\mathbf{H}}_2^{\#}(n,k)\underline{\mathbf{X}}_2(n,k)|^2] \tag{6}$$

where $Y_1(n,k) = Y(n,k) - D_1(n,k)$ and $E[\cdot]$ denotes expectation. By minimizing (5) and (6) with respect to $\underline{\mathbf{H}}_1(n,k)$ and $\underline{\mathbf{H}}_2(n,k)$, we derive the optimal weight vectors as

$$\widehat{\underline{\mathbf{H}}}_1(n,k) = \mathbf{\Phi}^{-1}_{\underline{\mathbf{X}}_1\underline{\mathbf{X}}_1}(n,k)\mathbf{\Phi}_{\underline{\mathbf{X}}_1 Y}(n,k), \tag{7}$$

$$\widehat{\underline{\mathbf{H}}}_2(n,k) = \mathbf{\Phi}^{-1}_{\underline{\mathbf{X}}_2\underline{\mathbf{X}}_2}(n,k)\mathbf{\Phi}_{\underline{\mathbf{X}}_2 Y_1}(n,k) \tag{8}$$

where $\mathbf{\Phi}_{\underline{\mathbf{X}}\underline{\mathbf{X}}}(n,k)$ and $\mathbf{\Phi}_{\underline{\mathbf{X}}Y}(n,k)$ denote the extended PSD matrix and cross PSD vector defined by

$$\mathbf{\Phi}_{\underline{\mathbf{X}}\underline{\mathbf{X}}}(n,k) = E[\underline{\mathbf{X}}(n,k)\underline{\mathbf{X}}^{\#}(n,k)], \tag{9}$$

$$\mathbf{\Phi}_{\underline{\mathbf{X}}Y}(n,k) = E[\underline{\mathbf{X}}(n,k)Y^{*}(n,k)] \tag{10}$$

with the superscript * denoting complex conjugation.

Given the estimated echo spectra, we estimate $S(n,k)$, the near-end signal in the STFT domain, applying the Wiener gain $G(n,k)$ as follows:

$$\widehat{S}(n,k) = G(n,k)Y(n,k) \tag{11}$$

under the assumption that the near-end and the echo signals are uncorrelated.

3. ESTIMATION OF EXTENDED PSD MATRICES, CROSS PSD VECTORS AND ECHO SPECTRA

For the implementation of the proposed algorithm, we first obtain the extended PSD matrix and cross PSD vector related to $\underline{\mathbf{X}}_1(n,k)$ by first-order recursive averaging as follows:

$$\begin{aligned}\widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_1\underline{\mathbf{X}}_1}(n,k) &= \alpha_{\Phi}\widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_1\underline{\mathbf{X}}_1}(n-1,k)\\ &\quad + (1-\alpha_{\Phi})\underline{\mathbf{X}}_1(n,k)\underline{\mathbf{X}}_1^{\#}(n,k),\end{aligned} \tag{12}$$

$$\begin{aligned}\widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_1 Y}(n,k) &= \alpha_{\Phi}\widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_1 Y}(n-1,k)\\ &\quad + (1-\alpha_{\Phi})\underline{\mathbf{X}}_1(n,k)Y^{*}(n,k)\end{aligned} \tag{13}$$

where α_{Φ} is a smoothing factor. The estimated optimal weight vector $\widehat{\underline{\mathbf{H}}}_1$ is obtained by applying (12) and (13) to (7). For the improved estimate of $D_1(n,k)$, we introduce additional echo overestimation control matrices, $\mathbf{C}_i$, which extend the echo suppression level control factor in [5]:

$$\mathbf{C}_i = \mathrm{diag}\{c_1\ \dots\ c_M\} \ \ (i=1,2) \tag{14}$$

where c_m weights the m-th component of $\underline{\mathbf{X}}_i(n,k)$. These matrices are employed to further reduce the residual echo. In this work, considering that closer frequency components have more influence, each c_m is chosen as follows:

$$\begin{aligned}c_m &= \alpha_{C_i}\exp\left(-\beta_{f,i}|m-K-1|\right)\\ &\qquad (i=1,2,\ m=1,\dots,M)\end{aligned} \tag{15}$$

in which the parameters α_{C_i} and $\beta_{f,i}$ are experimentally determined constants. Then, we calculate the estimate of

$D_1(n,k)$ by using the echo overestimation control matrix, $\mathbf{C}_1$, as the following way:

$$\widehat{D}_1(n,k) = |\widehat{\underline{\mathbf{H}}}_1^{\#}(n,k)\mathbf{C}_1\underline{\mathbf{X}}_1(n,k)|. \tag{16}$$

With $\widehat{D}_1(n,k)$, the estimate of $Y_1(n,k)$ is obtained by performing the spectral subtraction method in [7].

In a similar manner, we estimate $D_2(n,k)$ by performing the following procedures:

$$\begin{aligned}\widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_2\underline{\mathbf{X}}_2}(n,k) &= \alpha_\Phi \widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_2\underline{\mathbf{X}}_2}(n-1,k), \\ &\quad + (1-\alpha_\Phi)\underline{\mathbf{X}}_2(n,k)\underline{\mathbf{X}}_2^{\#}(n,k), && (17)\\ \widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_2 Y_1}(n,k) &= \alpha_\Phi \widehat{\mathbf{\Phi}}_{\underline{\mathbf{X}}_2 Y_1}(n-1,k) \\ &\quad + (1-\alpha_\Phi)\underline{\mathbf{X}}_2(n,k)Y_1^*(n,k), && (18)\\ \widehat{D}_2(n,k) &= |\widehat{\underline{\mathbf{H}}}_2^{\#}(n,k)\mathbf{C}_2\underline{\mathbf{X}}_2(n,k)| && (19)\end{aligned}$$

where $\mathbf{C}_2$ is also the echo overestimation control matrix.

To estimate the near-end signal, $\widehat{S}(n,k)$, we obtain the gain function $G(n,k)$ based on the Wiener filter by introducing the *a priori* SER $\xi(n,k)$ and *a posteriori* SER $\gamma(n,k)$ as in [9],

$$G(n,k) = \frac{\xi(n,k)}{1+\xi(n,k)}, \tag{20}$$

$$\xi(n,k) \triangleq \frac{\lambda_S(n,k)}{\lambda_D(n,k)}, \tag{21}$$

$$\gamma(n,k) \triangleq \frac{|Y(n,k)|^2}{\lambda_D(n,k)} \tag{22}$$

where $\lambda_S(n,k)$ and $\lambda_D(n,k)$ denote the PSD of the near-end signal and stereo echo, respectively. We update the estimates of $\lambda_D(n,k)$, $\gamma(n,k)$, and $\xi(n,k)$ as [5]

$$\begin{aligned}\widehat{\lambda}_D(n,k) &= \alpha_\lambda \widehat{\lambda}_D(n-1,k) \\ &\quad + (1-\alpha_\lambda)(|\widehat{D}_1(n,k)|^2 + |\widehat{D}_2(n,k)|^2), && (23)\\ \widehat{\gamma}(n,k) &= \frac{|Y(n,k)|^2}{\widehat{\lambda}_D(n,k)}, && (24)\\ \widehat{\xi}(n,k) &= \alpha_{DD}\widehat{\gamma}(n-1,k)G^2(n-1,k) \\ &\quad + (1-\alpha_{DD})\max(\widehat{\gamma}(n,k)-1,0) && (25)\end{aligned}$$

where α_λ and α_{DD} are forgetting factors. Given the gain function from the estimate of $\xi(n,k)$, the estimated near-end $\widehat{S}(n,k)$ is calculated from (11).

4. EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed SAES method, we designed computer simulations under various conditions. For performance assessment, we created 22 data sets from the TIMIT database such that each set consisted of the source signal $v(n)$ and near-end signal $s(n)$. The data sets were sampled at 16 kHz. The length of each data set ranged from $10\,s$ to $18\,s$ and the total length of the data was $332\,s$. The duration of double-talk interval was between $5\,s$ to $10\,s$. Both the transmission room and the receiving room were designed to fit a small office room of a size $4\,m \times 4\,m \times 3\,m$. All of the RIRs were generated with the reverberation time $T_{60} = 200\,ms$ by the image method [11]. The length of the RIRs was set to 512. The echo level measured at the input microphone was 3.5 dB lower than that of the near-end speech on average. A white noise was added to the microphone signals with $SNR = 30, 20$, and 10 dB. We applied a 7/8-overlapping Hamming window of length 2048 for taking the STFT. In the proposed algorithm, the following parameters were chosen; $\alpha_\Phi = 0.999$, $\alpha_\lambda = 0.001$, $\alpha_{DD} = 0.001$, $\alpha_{C_1} = 1.35$, $\alpha_{C_2} = 1.2$, $\beta_{f,1} = 0.3$, $\beta_{f,2} = 0.32$, and $N = 2048$.

Table 1. ERLE and PESQ Scores in Noiseless Condition with Different Values of K

K	0	1	2	3	4
ERLE	12.76	21.85	25.54	27.11	27.90
PESQ	2.57	2.91	2.99	3.01	3.02

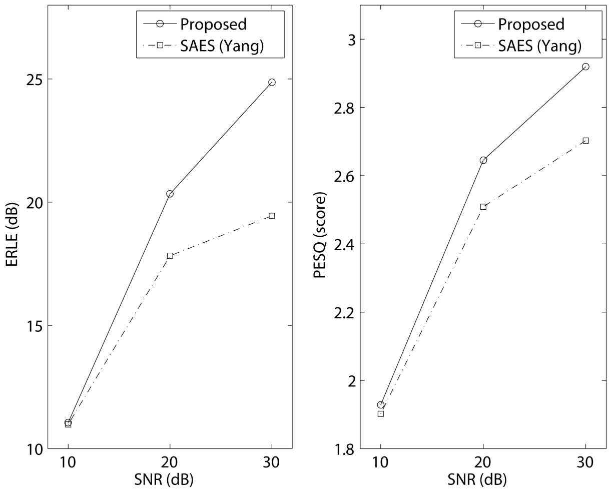
To test the performance of the proposed method, we evaluated the ITU-T Recommendation P. 862 perceptual evaluation of speech quality (PESQ) score [12] and the echo return loss enhancement (ERLE) measure which is defined by [8]

$$ERLE(n) = 10\log_{10}\left[\frac{E[y^2(n)]}{E[\hat{s}^2(n)]}\right](dB) \tag{26}$$

where $\hat{s}(n)$ denotes the near-end signal estimate at time index n after suppressing echoes in single-talk case, which is the residual echo components.

In Table 1, the overall results of the ERLE and PESQ scores measured in clean conditions are shown for the different values of K. Note that the proposed algorithm using the augmented vector with $K = 0$ corresponds to the conventional SAES algorithm [5]. We can remark that as more spectral correlations among adjacent components were considered, the higher ERLE and PESQ scores we obtained. It is found that the crossband filtering incorporating the correlations was beneficial to cancel the echo signals effectively. On the other hand, as the number of crossbands increases, the improvement of the ERLE scores for each increase of K decreases and the gap between the PESQ scores also diminishes. Thus, we need to take the adequate value of K for the effectiveness of the proposed algorithm.

To compare the performance of the proposed technique with that of the conventional SAES method, we also evaluated the ERLE and PESQ performance under various SNR conditions. For the low complexity, we tested using the augmented vector with $K = 2$ in the proposed algorithm. We denote the



(a) ERLE Scores

(b) PESQ Scores

Fig. 2. ERLE and PESQ Scores in Different SNR Conditions.

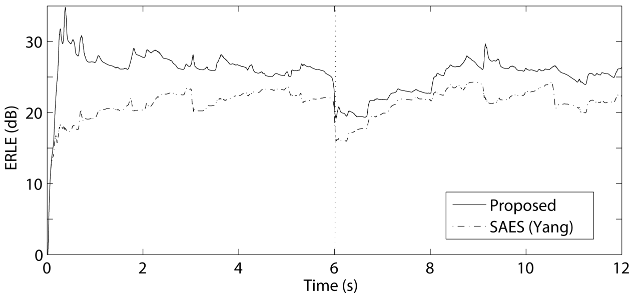


Fig. 3. Temporal variation of ERLE for comparison of convergence speeds and tracking performances between the proposed $(K = 2)$ and the SAES (Yang) algorithm in single-talk case. The microphone in the receiving room moved at $6\,s$, $SNR = 30\,\text{dB}$ and $T_{60} = 200\,ms$.

conventional method by SAES (Yang) and used the same parameters ($\beta_1 = 1.35$, $\beta_2 = 1.2$, $\alpha_\lambda = 0.6$, $\alpha_{DD} = 0.6$, $\alpha_\phi = 0.975$, $N = 2048$) as in [5]. From the whole results in Fig. 2, the proposed approach showed better performance than SAES (Yang) algorithm. Especially, in the high SNR conditions, we observed that the proposed method preserved the near-end signal and suppressed the stereo echo signal significantly better compared with SAES (Yang).

In Fig. 3, we compared the convergence speeds and tracking performances of the proposed and conventional SAES algorithms in single-talk situation using temporal variation of ERLE. In this experiment, the microphone in the receiving room changed its location at $6\,s$. The proposed crossband filtering for SAES always outperformed the conventional SAES method and we did not find any substantial tracking issue in the test environments.

5. CONCLUSIONS

In this paper, we have proposed a crossband filtering for SAES incorporating spectral correlations. The proposed algorithm estimates the extended PSD matrices and cross PSD vectors based on the correlations and introduces the echo overestimation control matrices to track and suppress the stereo echo signal. The proposed technique showed better performances in both single-talk and double-talk cases than the conventional SAES method. We conclude that the proposed algorithms can be seen as an effective way for more accurate echo estimation.

6. REFERENCES

[1] H. I. Rao and B. Farhang-Boroujeny, “Fast LMS/Newton algorithms for stereophonic acoustic echo cancelation,” *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 2919-2930, Aug. 2009.

[2] H. Buchner, J. Benesty, T. Gansler, and W. Kellermann, “Robust extended multidelay filter and double-talk detector for acoustic echo cancellation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1633-1644, Sep. 2006.

[3] J. Benesty, D. R. Morgan, and M. M. Sondhi, “A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 6, no. 2, pp. 156-165, Mar. 1998.

[4] L. Romoli, S. Cecchi, P. Peretti, and F. Piazza, “A mixed decorrelation approach for stereo acoustic echo cancellation based on the estimation of the fundamental frequency,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 690-698, Feb. 2012.

[5] F. Yang, M. Wu, and J. Yang, “Stereophonic acoustic echo suppression based on Wiener filter in the short-time Fourier transform domain,” *IEEE Signal Process. Lett.*, vol. 19, no. 4, pp. 227-230, Apr. 2012.

[6] C. Faller and C. Tournery, “Robust acoustic echo control using a simple echo path model”, in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing.*, May. 2006, vol. 5, pp. 281-284.

[7] 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. Sep. 2005.

[8] S. Y. Lee and N. S. Kim, “A statistical model based residual echo suppression,” *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 758-761, Oct. 2007.

[9] Y. S. Park and J. H. Chang, “Frequency domain acoustic echo suppression based on soft decision,” *IEEE Signal Process. Lett.*, vol. 16, no. 1, pp. 53-56, Jan. 2009.

[10] 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, May. 2007.

[11] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Amer.*, vol. 65, pp. 943-950, Apr. 1979.

[12] ITU-T, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs”, ITU-T Rec. P. 862, 2000.