ACOUSTIC-COUPLING LEVEL ESTIMATION FOR PERFORMANCE IMPROVEMENT OF ECHO REDUCTION

[1]*Masahiro Fukui, Suehiro Shimauchi, Akira Nakagawa, Yoichi Haneda, and* [2]*Akitoshi Kataoka*

[1]NTT Cyber Space Laboratories, NTT Corporation, Tokyo 180-8585, Japan
[2]Faculty of Science and Technology, Ryukoku University, Shiga 520-2914, Japan

ABSTRACT

The echo-reduction process suppresses the undesired echo signal that results from acoustic coupling by multiplying a echo reduction gain in a frequency domain. However, its process cannot easily remove an echo signal immediately after an echo-path change is caused. For the performance improvement of the echo reduction, we propose a fast and accurate acoustic-coupling level estimation method. The performance of this method is demonstrated by simulation results in which the echo is rapidly suppressed.

Index Terms**—** Acoustic echo canceller, Echo reduction, Wiener filtering, Acoustic-coupling level estimation

1. INTRODUCTION

An acoustic echo canceller (AEC) is indispensable for hands-free telecommunications to eliminate an undesired echo signal that results from acoustic coupling between a loudspeaker and a microphone. In the AEC, adaptive filter (ADF) [1] process is commonly used to remove the echo signal. The ADF process achieves echo removal by modeling the unknown echo path using the ADF and subtracting an echo estimate from the microphone signal. However, the ADF process cannot easily eliminate the echo signal after an echo path change is caused because of the slow convergence speed of the ADF. If the AEC is used, usually the echo reduction (ER) [2, 3] process is also used in series after the ADF process as a post filter to suppress the residual echo signal. The ER process suppresses the echo signal by a multiplicative gain in the frequency domain. Multiplying the known received speech power spectrum by the acoustic-coupling level (ACL) estimate, the ER obtains the echo power spectral estimate and suppresses the echo signal. The echo-reduction performance depends on the speed and accuracy of estimating the ACL.

Two different ACL estimation methods have been proposed: one is a direct and simple estimation method [2] based on an averaged power spectral ratio of the microphone signal to the received speech signal (Method A: Amplitude method); the other is a cross-spectral method [4] based ACL estimation method [3] (Method C: Cross-spectral method). The significant problem with Method A is that the ACL cannot be estimated during double-talk situations. On the other hand, Method C estimates the ACL even during double-talk situations because the near-end speech component can be removed by calculating an averaged cross spectrum between received speech and microphone signals. However, Method C cannot estimate the ACL accurately after the echo path change occurs because the averaged cross spectrum calculation needs a very long time constant. In addition, if a frame employed in a FFT is not long enough compared to an echo path length, a bias error in the ACL estimate is generated.

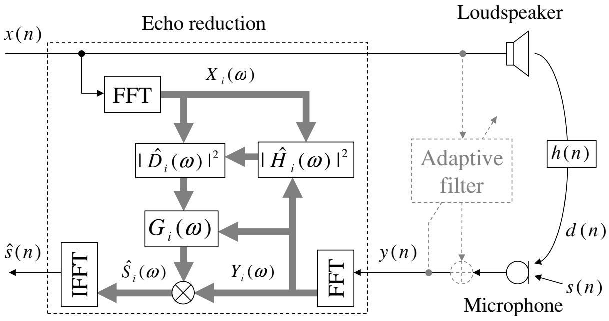

Fig. 1. Structure of echo reduction process.

In this paper, we propose a fast and accurate ACL estimation method to achieve effective echo reduction even if a rapid echo-path change is caused. For example, a near-end speaker presses a power switch to the OFF position on a front microphone. A second near-end speaker also presses a power switch to the ON position on another front microphone. The proposed method, which is based on Method C, rapidly estimates the ACL by focusing on time and frequency spectral domains for an averaged cross-spectral calculation. In addition, this method also compensates for the bias error by obtaining a magnitude-squared coherence (MSC) between received speech and echo signals using Method A.

2. ECHO REDUCTION (ER)

In this section, we derive the ER process [2, 3] on the basis of a short-time spectral amplitude (STSA) estimation [3]. The ER process is illustrated in Fig. 1. The microphone signal

$y(n)$ can be written in terms of the received speech signal $x(n)$ convoluted by the echo path $h(n)$ plus the desired (target) near-end speech signal $s(n)$:

$$y(n) = x(n) * h(n) + s(n), \tag{1}$$

where $*$ denotes convolution. The short-time spectrum of $y(n)$ is represented as

$$Y_i(\omega) = D_i(\omega) + S_i(\omega), \tag{2}$$

where ω is the discrete frequency index, i is the discrete time-frame index, $D_i(\omega)$ and $S_i(\omega)$ are the short-time spectrums of echo signal $d(n) = x(n) * h(n)$ and $s(n)$, respectively. The echo reduction can be expressed as

$$\hat{S}_i(\omega) = G_i(\omega) Y_i(\omega), \tag{3}$$

where $\hat{S}_i(\omega)$ is the estimate of $S_i(\omega)$, and

$$G_i(\omega) = \frac{|Y_i(\omega)|^2 - |\hat{D}_i(\omega)|^2}{|Y_i(\omega)|^2} \tag{4}$$

is the echo-reduction gain based on Wiener filtering [3]. The estimate of $|D_i(\omega)|^2$ is usually calculated as

$$|\hat{D}_i(\omega)|^2 = |\hat{H}_i(\omega)|^2 |X_i(\omega)|^2 + \beta |\hat{D}_{i-1}(\omega)|^2, \tag{5}$$

where $|\hat{H}_i(\omega)|^2$ denotes the estimate of the ACL $|H_i(\omega)|^2$, which is the power spectrum of $h(n)$, $X_i(\omega)$ is the short-time spectrum of $x(n)$, and β is a design parameter to control the reverberation time; $0 \leq \beta < 1$ is used. Multiplying $|X_i(\omega)|^2$ by the ACL estimate $|\hat{H}_i(\omega)|^2$, we can obtain the power spectrum of the echo estimate. The echo-estimation accuracy depends on the speed and accuracy to estimate the ACL.

3. CONVENTIONAL METHODS

In this section, we describe two different standard ACL estimation methods used by the ER in Sec. 2 as follows: the direct and simple estimation method [2] based on the averaged power spectral ratio of the microphone signal to the received speech signal (Method A) and the cross-spectral method based estimation method [3] (Method C). We also discuss their problems.

3.1. Acoustic-coupling level (ACL) estimations

Since the ACL is likely to vary in time, the ACL is estimated iteratively. The ACL estimate in Method A is given by

$$\left|\hat{H}_i(\omega)\right|_{\mathrm{A}}^2 = \frac{\sum_{k=0}^{N_{\mathrm{S}}-1} |Y_{i-k}(\omega)|^2}{\sum_{k=0}^{N_{\mathrm{S}}-1} |X_{i-k}(\omega)|^2}, \tag{6}$$

where N_{S} indicates the number of frames required for the averaged power spectral calculation. On the other hand, the ACL estimate in Method C is given as

$$\left|\hat{H}_i(\omega)\right|_{\mathrm{C}}^2 = \left| \frac{\sum_{k=0}^{N_{\mathrm{L}}-1} X_{i-k}^*(\omega) Y_{i-k}(\omega)}{\sum_{k=0}^{N_{\mathrm{L}}-1} |X_{i-k}(\omega)|^2} \right|^2, \tag{7}$$

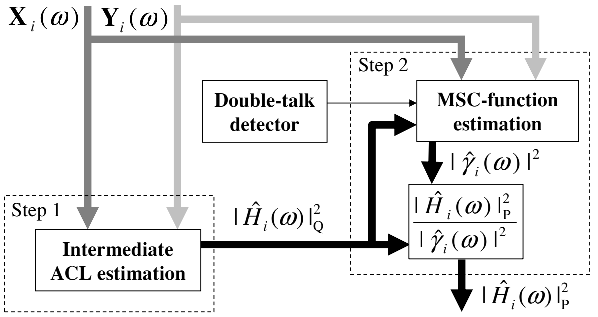
$\mathbf{X}_i(\omega)$
$\mathbf{Y}_i(\omega)$
Step 2
Double-talk detector
MSC-function estimation
$|\hat{\gamma}_i(\omega)|^2$
Step 1
$|\hat{H}_i(\omega)|_{\mathrm{Q}}^2$
$\frac{|\hat{H}_i(\omega)|_{\mathrm{P}}^2}{|\hat{\gamma}_i(\omega)|^2}$
Intermediate ACL estimation
$|\hat{H}_i(\omega)|_{\mathrm{P}}^2$

Fig. 2. Structure of proposed ACL estimation method.

where $X_i^*(\omega)$ is the complex conjugate of $X_i(\omega)$, N_{L} denotes the number of frames required to calculate the cross correlation, and $N_{\mathrm{L}} \gg N_{\mathrm{S}}$ is used.

3.2. Problems of conventional methods

The significant problem with Method A is that the ACL cannot be estimated during double-talk situations because the power spectrum of the microphone signal is increased by an amount equal to the power spectrum of the near-end speech signal. On the other hand, Method C estimates the ACL even during double-talk situations when received and near-end speech signals are uncorrelated because the near-end speech component can be removed by calculating an averaged cross spectrum between received speech and microphone signals. However, Method C cannot estimate the ACL accurately after the echo path change occurs because the averaged cross and power spectral calculations in Eq. (7) need a large number of frames N_{L}. Thus, Method C cannot accurately estimate the ACL over a long time in which N_{L} frames pass after the echo path change occurs.

In addition, in Method C, if a frame employed in a FFT is not long enough compared to an echo path length, however, a bias error in the ACL estimate due to the insufficiency of the frame length (FFT size) is generated. The cause of the bias error is that a signal in a microphone frame does not contain the whole response of the echo path to a signal in a received speech frame and contains an extraneous response to the preceding received speech signal. The other problem with Method C is that the estimation accuracy is degraded by the bias error.

4. PROPOSED METHOD

For resolving these problems and improving ER performance, we propose a fast and accurate hybrid ACL estimation method, which expands the conventional Method C and combines that with Method A. The structure of our proposed method is shown in Fig. 2. This method has two steps as follows: one is the ACL estimation focused on time and frequency spectral domains for the averaged cross-spectral calculation, and the other is the compensation of the ACL estimation error caused by the bias error using Method A.

4.1. Acoustic-coupling level (ACL) estimation

The difference between our proposed ACL estimation method and the conventional Method C is that the ratio of the averaged cross-spectrum between received speech and echo signals to the averaged near-end speech power spectrum is calculated in time and frequency spectral domains to sufficiently shorten the time constant determined by the number of frames as follows:

$$\left|\hat{H}_i(\omega)\right|_{\mathrm{Q}}^2 = \left[\frac{\sum_{r=-R_\omega}^{R_\omega}\left|\sum_{k=0}^{N_\mathrm{S}-1} X_{i-k}^*(\omega+r)Y_{i-k}(\omega+r)\right|}{\sum_{r=-R_\omega}^{R_\omega}\sum_{k=0}^{N_\mathrm{S}-1}|X_{i-k}(\omega+r)|^2}\right]^2. \tag{8}$$

The proposed method calculates the ACL estimate even during double-talk situations in a small number of frames, N_S. However, a problem remains because the ACL estimation accuracy is degraded as the bias error due to the shorter frame length than the echo path length.

4.2. Estimation-error compensation

A bias error ratio is expressed as a ratio of the cross-spectral method based ACL estimate to the power spectral ratio based estimate during single-talk situations at $N_\mathrm{S} = N_\mathrm{L}$ given by

$$\begin{aligned}\frac{\left|\hat{H}_i(\omega)\right|_{\mathrm{C}}^2}{\left|\hat{H}_i(\omega)\right|_{\mathrm{A}}^2} &= \frac{\left|\sum_{k=0}^{N_\mathrm{S}-1} X_{i-k}^*(\omega)D_{i-k}(\omega)\right|^2}{\sum_{k=0}^{N_\mathrm{S}-1}|X_{i-k}(\omega)|^2\sum_{k=0}^{N_\mathrm{S}-1}|D_{i-k}(\omega)|^2}\\ &= |\gamma_i(\omega)|_\mathrm{S}^2 \leq 1,\end{aligned} \tag{9}$$

where $|\gamma_i(\omega)|_\mathrm{S}^2$ is the MSC between near-end speech and echo signals. The ACL estimate $|\hat{H}_i(\omega)|_\mathrm{C}^2$ with the bias error is always less than $|\hat{H}_i(\omega)|_\mathrm{A}^2$, so we compensate for the ACL estimate in Eq. (8) as follows

$$\begin{aligned}&\left|\hat{H}_i(\omega)\right|^2 = \frac{1}{|\gamma_i(\omega)|_\mathrm{S}^2}\left|\hat{H}_i(\omega)\right|_\mathrm{C}^2\\ \Longrightarrow &\left|\hat{H}_i(\omega)\right|_\mathrm{P}^2 = \frac{1}{|\hat{\gamma}_i(\omega)|_\mathrm{S}^2}\left|\hat{H}_i(\omega)\right|_\mathrm{Q}^2,\end{aligned} \tag{10}$$

where $|\hat{\gamma}_i(\omega)|_\mathrm{S}^2$ is the MSC estimate. According to the talk situation, we calculate its MSC estimate as follows.

(i) **Case of no double-talk situation** The short-time spectrum of the microphone signal can be considered as $Y_i(\omega) = D_i(\omega)$ during single-talk situations because the microphone is only picking up the received speech signal. Then, using the ACL estimate in Eq. (6), the MSC between the received speech and echo signals can readily be estimated as

$$\begin{aligned}|\hat{\gamma}_i(\omega)|_\mathrm{S}^2 &= \frac{\left|\hat{H}_i(\omega)\right|_\mathrm{Q}^2}{\left|\hat{H}_i(\omega)\right|_\mathrm{A}^2}\\ &= \left|\hat{H}_i(\omega)\right|_\mathrm{Q}^2\frac{\sum_{k=0}^{N_\mathrm{S}-1}|X_{i-k}(\omega)|^2}{\sum_{k=0}^{N_\mathrm{S}-1}|D_{i-k}(\omega)|^2}.\end{aligned} \tag{11}$$

When there is no received speech signal, the echo-reduction gain is given by $G_i(\omega) = 1$ because the echo signal is not contained in the microphone signal.

(ii) **Case of double-talk situation** The proposed method estimates the MSC between the received speech and echo signals even during double-talk situations only at frequency components without a near-end speech component because speech signals are sparse in the frequency domain. We give an MSC candidate $\rho_i(\omega)$ by

$$\rho_i(\omega) = \left|\hat{H}_i(\omega)\right|_\mathrm{Q}^2\frac{\sum_{k=0}^{N_\mathrm{S}-1}|X_{i-k}(\omega)|^2}{\sum_{k=0}^{N_\mathrm{S}-1}|Y_{i-k}(\omega)|^2}. \tag{12}$$

From Eq. (12), we readily find that $\rho_i(\omega)$ is large when there is no near-end speech component in the frequency component. Thus, we calculate the MSC estimate as

$$|\hat{\gamma}_i(\omega)|_\mathrm{S}^2 = \begin{cases}\rho_i(\omega), & \text{if } \rho_i(\omega) > |\hat{\gamma}_{i-1}(\omega)|_\mathrm{S}^2\\ |\hat{\gamma}_{i-1}(\omega)|_\mathrm{S}^2, & \text{otherwise}\end{cases} \tag{13}$$

to avoid MSC estimation in the frequency component with the near-end speech component.

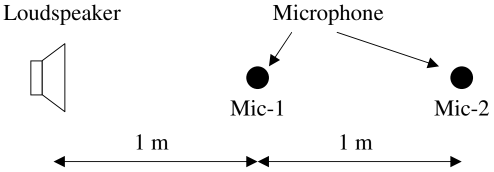

Fig. 3. Locations of loudspeaker and microphones.

Table 1. Experimental conditions

Sampling rate	16 kHz
Frame length	256 samples
Frame shift	128 samples
Reverberation time	300 ms

5. EXPERIMENTAL RESULTS

To compare the proposed ACL estimation method to the conventional method (Method C), some experiments were performed by simulations. The arrangement for a loudspeaker

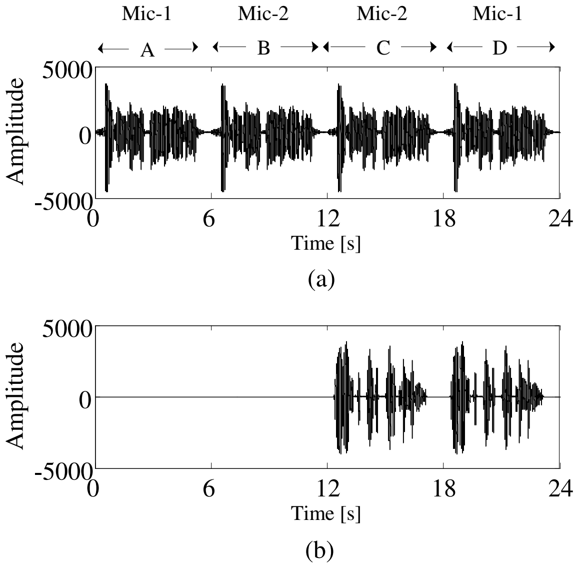

Fig. 4. (a) received speech signal (female speech) and (b) near-end speech signal (male speech).

and microphones is shown in Fig. 3, and Table 1 gives experimental conditions. Values of N_S and N_L are set to 50 and 500, respectively. The proposed method uses a Geigel algorithm [5] for double-talk detection. The conventional ACL estimate compensates for the bias error with a constant, which is set to 6 dB, in order to sufficiently suppress the echo signal. In simulations, we assume that the echo path rapidly changes by switching between Mic-1 and Mic-2, as shown in Fig. 3. The received and near-end speech signals are shown in Figs. 4 (a) and (b), respectively. Periods A and B are single-talk situations in Mic-1 and Mic-2, respectively. Periods C and D are double-talk situations in Mic-2 and Mic-1, respectively.

The microphone input signal $y(n)$ is shown in Fig. 5 (a). The sent signals after processing by conventional and proposed methods are shown in (b) and (c), respectively. As seen in Figs. 5 (a), (b), and (c), the proposed and conventional methods sufficiently suppress echo signals over the entire period. Echo-suppression levels are 36.2 dB in the conventional method and 37.0 dB in the proposed method. However, the conventional method suffers from near-end speech distortion during double-talk situations though the number of frames of the conventional method is ten times as large as the number of frames of the proposed method. The near-end speech distortion was improved by using the proposed method, and the subjective quality was good.

6. CONCLUSION

An ACL estimation method for the echo reduction was proposed. The proposed method is focused on time and frequency spectral domains. That method rapidly tracks the ACL and improves estimation accuracy by estimating the MSC between received speech and echo signals based on the talk situation. According to experimental results, we confirmed that the proposed ACL estimation method achieves better echo reduction performance than that of the conventional method.

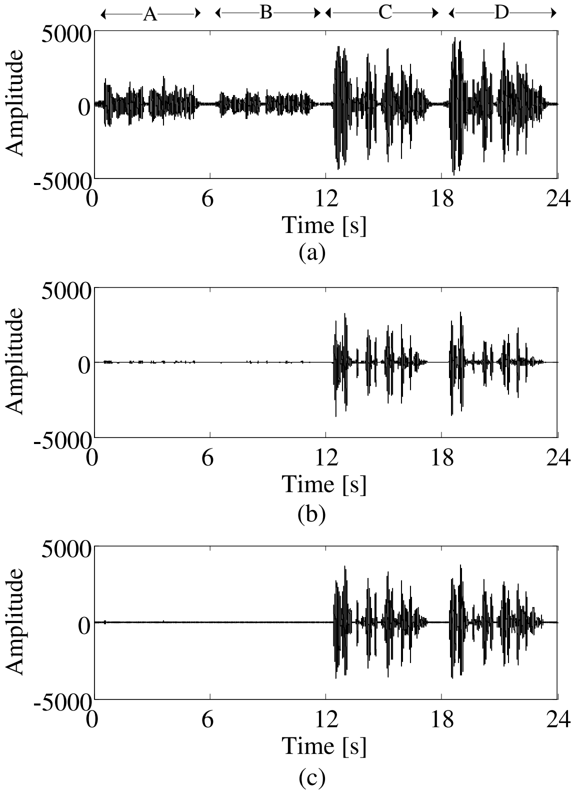

Fig. 5. (a) microphone input signal, (b) sent signal in conventional method, and (c) sent signal in proposed method.

7. REFERENCES

[1] S. Haykin, “Adaptive filter theory third edition,” *Prentice-Hall, Inc.*, New Jersey, 1996.

[2] C. Avendano, “Acoustic echo suppression in the STFT domain,” *IEEE Workshop Sig. Proc. to Audio and Acoust.*, vol. 21, no. 24, pp. 175–178, Oct. 2001.

[3] C. Faller and C. Tournery, “Estimating the delay and coloration effect of the acoustic echo path for low complexity echo suppression,” *Proc. IWAENC2005*, pp. 53–56, Oct. 2005.

[4] J. S. Bendat and A. G. Piersol, “Engineering Applications of Correlation and Spectral Analysis,” *John Wiley and Sons*, New York, 1993.

[5] D. L. Duttweiler, “A twelve-channel digital echo canceler,” *IEEE Trans. Commun.*, vol. COM-26, no. 5, pp. 647–653, May 1978.