International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:1, 2015

Environmentally Adaptive Acoustic Echo Suppression for Barge-in Speech Recognition

Jong Han Joo, Jeong Hun Lee, Young Sun Kim, Jae Young Kang, Seung Ho Choi

Abstract**—**In this study, we propose a novel technique for acoustic echo suppression (AES) during speech recognition under barge-in conditions. Conventional AES methods based on spectral subtraction apply fixed weights to the estimated echo path transfer function (EPTF) at the current signal segment and to the EPTF estimated until the previous time interval. However, the effects of echo path changes should be considered for eliminating the undesired echoes. We describe a new approach that adaptively updates weight parameters in response to abrupt changes in the acoustic environment due to background noises or double-talk. Furthermore, we devised a voice activity detector and an initial time-delay estimator for barge-in speech recognition in communication networks. The initial time delay is estimated using log-spectral distance measure, as well as cross-correlation coefficients. The experimental results show that the developed techniques can be successfully applied in barge-in speech recognition systems.

Keywords**—**Acoustic echo suppression, barge-in, speech recognition, echo path transfer function, initial delay estimator, voice activity detector.

I. Introduction

IN a far-end or a near-end speech recognition service, if the server and the user speak simultaneously, the user's voice and the server's message signal will reach the near-end microphone at the same time. This barge-in situation results in the serious degradation of speech recognition performance due to the undesired echoes of the message signal from the loudspeaker [1]. Therefore, usually acoustic echo suppression (AES) is used for removing the acoustic echoes [2]-[5]. Echo paths consist of an initial time delay with no echo signal, and active regions in which the echo signal is present. Therefore, the AES is only effective when the adaptive filter constantly adapts to the current acoustic echo path. Furthermore, to save computational costs and increase the AES performance, it is necessary to match the echo path transfer function (EPTF) only in the active region.

In this paper, to accomplish this, we develop an algorithm to estimate the initial delay and to identify the active region. We devise an automatic voice activity detector (VAD) to find the active region of the message prompt signal and make a reference segment. Then, we develop methods to estimate the initial time delay. Cross-correlation coefficients (CCCs) between the reference segment and an input signal segment are computed in a conventional manner, and the initial delay is estimated as a function of the index of the peak value of the cross-correlation lags. However, the CCC-based method may exhibit poor performance when used with colored input signals such as speech signals [6]. In this research work, we develop a hybrid method that uses the log-spectral distance (LSD) measure as well as the CCC to estimate the initial time-delay for barge-in speech recognition in communication network services.

It has recently been shown that there are various advantages in adopting spectral subtraction-based methods in a short-time Fourier transform (STFT) domain [3], [4]. These methods remove the acoustic echoes by means of short-time spectral modification in the frequency domain. In conventional spectral subtraction-based AES methods, for updating the EPTF, fixed weights are applied to the estimated EPTF at the current signal segment and to the EPTF updated until the previous time interval [5]. In real environments, the room impulse response between a loudspeaker and microphone varies depending on many factors, such as body movements, temperature changes [7], etc. We propose an adaptive algorithm that updates the weight parameters to consider abrupt changes in the acoustic environment due to background noises or double-talk (DT).

The remainder of this paper is organized as follows: Section II describes the VAD, delay estimation, spectral subtraction-based AES, environmentally adaptive AES methods for barge-in speech recognition. The performance of these techniques is evaluated in Section III. Finally, we summarize the conclusions that can be drawn from this study in Section IV.

II. Acoustic Echo Suppression (AES) Techniques for Barge-In Speech Recognition

A. Echo Suppression System

In the echo suppression system as shown in Fig. 1, $x(n)$ is a message prompt signal and $y(n)$ is an echo signal that is a portion of $d(n)$ transmitted from the near-end microphone. The $h(n)$ represents an echo path impulse response. In a barge-in situation, the difference signal $e(n) = d(n) - \hat{y}(n)$ is the estimate of user's voice $v(n)$.

B. Voice Activity Detection and Delay Estimation

First, as shown in Fig. 2 (a), we developed a VAD algorithm to detect active regions in the message signal. The VAD flag is

J. H. Joo is with the Department of Electronic Engineering, Seoul National University of Science and Technology, Nowon-gu, Seoul, 139-743, Korea (e-mail: whdgks37@naver.com).

J. H. Lee, Y. S. Kim, J. Y. Kang, S. H. Choi are with the Department of Electronic and IT Media Engineering, Seoul National University of Science and Technology, Nowon-gu, Seoul, 139-743, Korea (corresponding author to provide phone: 82-2-970-6461; fax: 82-2-979-7903; e-mail: shchoi@seoultech.ac.kr).

set if the average energy of some of the frames exceeds the predefined threshold. Then, a speech segment of the message signal is saved for delay estimation. Fig. 2 (b) shows the block diagram when the segment of microphone input signal $\{d_s(n)\}$ matches closely with the saved message segment.

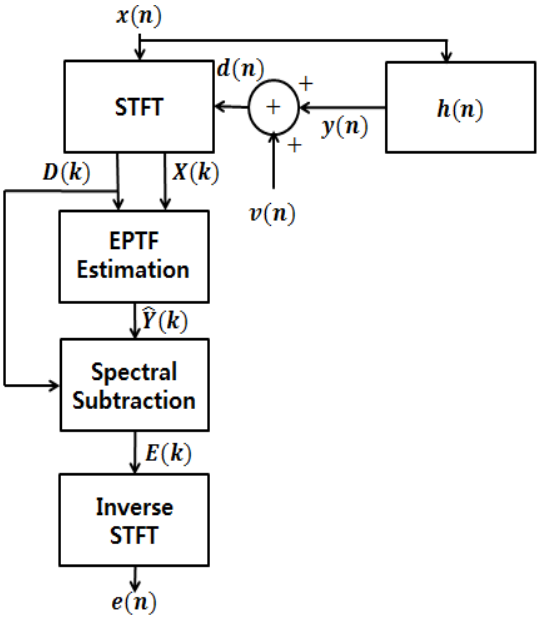
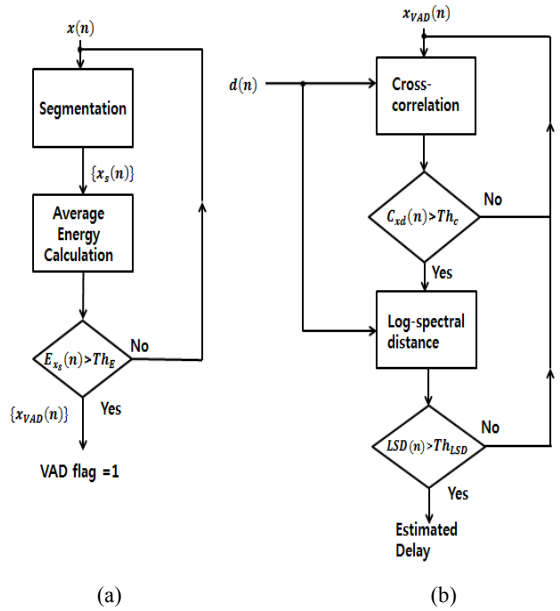


Fig. 1 Block diagram of the spectral subtraction-based acoustic echo suppression system



(a) (b)

Fig. 2 Delay estimation: (a) Voice activity detection. (b) LSD and CCC

To figure out that, we use the CCC denoted as

$$CCC = \frac{\sum_{n=0}^{N-1} x_{VAD}(n)\, d_S(n)}{\sqrt{\frac{1}{N}\sum_{n=0}^{N-1} {x_{VAD}}^2(n)}\ \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} {d_s}^2(n)}}. \quad (1)$$

The CCC has a peak value when the segment $\{d_s(n)\}$ is in accord with the segment $\{x_{VAD}(n)\}$, and the CCC flag is set if the peak CCC value is greater than a predefined threshold. Moreover, we propose a hybrid method that uses the LSD measure as well as the CCC to estimate the initial delay. The LSD is a distance measure between two spectra, and is obtained by

$$LSD = \sqrt{\frac{1}{N}\sum_{k=0}^{N-1} (10\log_{10}\frac{|X_{VAD}(k)|^2}{|D_s(k)|^2})^2} \quad (2)$$

where $X_{VAD}(k)$ and $D_s(k)$ represent the discrete Fourier transform (DFT) spectra of $x_{VAD}(n)$ and $d_s(n)$, respectively, and the k indicates the frame index. If the LSD value is lower than a preset threshold, the LSD flag is set. We consider the time point as the initial delay if both CCC and LSD conditions are satisfied.

C. Spectral Subtraction-Based AES

In this subsection, we briefly review the estimation of the EPTF in the spectral subtraction-based AES. $X(i,k)$, $D(i,k)$, $\hat{H}(i,k)$ denote the spectra of far-end speech signal, microphone input signal and estimated EPTF with frequency index i and frame index k, respectively. The estimated EPTF $\hat{H}(i,k)$ is iteratively computed as follows:

$$\hat{H}(i,k) = \frac{H_{num}(i,k)}{H_{den}(i,k)} \quad (3)$$

where

$$H_{num}(i,k) = \lambda H_{num}(i,k-1) + (1-\lambda)X^*(i,k)D(i,k) \quad (4)$$

$$H_{den}(i,k) = \lambda H_{den}(i,k-1) + (1-\lambda)X^*(i,k)X(i,k) \quad (5)$$

and λ is a weight parameter [8]. Then, the estimated echo magnitude spectrum $\hat{Y}(i,k)$ is given by

$$\hat{Y}(i,k) = \hat{H}(i,k)|X(i,k)|\ . \quad (6)$$

and $|\hat{E}(i,k)|$, denoting the estimated short-time magnitude spectrum of residual echo signal, is given by

$$|\hat{E}(i,k)| = (|D(i,k)|^{\alpha} - \beta|\hat{Y}(i,k)|^{\alpha})^{\frac{1}{\alpha}}. \quad (7)$$

Subsequently, short-time phase of the microphone input, $\angle D(i,k)$, is used as the phase of $\hat{E}(i,k)$, i.e.

$$\hat{E}(i,k) = |E(i,k)|\, e^{j\angle D(i,k)}. \quad (8)$$

D. Environmentally Adaptive AES

Conventionally, spectral subtraction-based AES methods apply fixed-weights to the estimated EPTF at the current signal segment and to the EPTF estimated until the previous time interval, as in (4) and (5). We replace the λ with a time-varying $\lambda(k)$ that is adaptively updated in response to an abrupt change in acoustic environment due to background noises or double-talk, as follows:

$$H_{num}(i,k) = \lambda(k) H_{num}(i,k-1) + (1-\lambda(k)) X^*(i,k) D(i,k) \quad (9)$$

$$H_{den}(i,k) = \lambda(k) H_{den}(i,k-1) + (1-\lambda(k)) X^*(i,k) X(i,k). \quad (10)$$

The parameter $\lambda(k)$ is controlled by the CCC $\rho(k)$ between a $|\hat{Y}(k)|$ and $|\hat{D}(k)|$,

$$\rho(k) = \frac{\frac{1}{N}\sum_{i=0}^{N-1} |D(i,k)|\,|\hat{Y}(i,k)|}{\|D(k)\|\,\|\hat{Y}(k)\|}. \quad (11)$$

Then, we let $\lambda(k)$ decrease linearly with $\rho(k)$:

$$\lambda(i) = \alpha\rho(k) + b,\ a < 0. \quad (12)$$

In the proposed method, $H_{num}(i,k)$ and $H_{den}(i,k)$ are updated in the DT interval as well as single-talk interval. Fig. 3 shows the block diagram of our proposed method.

Fig. 3 Block diagram of the proposed EPTF estimation method

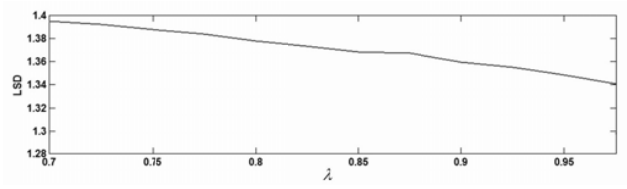
III. Performance Evaluation

To evaluate the performance of the proposed AES, we used a woman's voice as the message signal $x(n)$, and four voices of two men and two women as the input signal $v(n)$. The audio files were sampled at 16 kHz. We generated an artificial echo path impulse response for the experiments. The message signal $x(n)$ was convolved with the impulse response before being mixed. For each frame of the Hamming-windowed signal, $x(n)$ and $d(n)$ were transformed into their spectra through 256-point DFT after zero padding. To obtain an objective comparison, we evaluated the performance of echo return loss enhancement (ERLE) [9] which is defined by

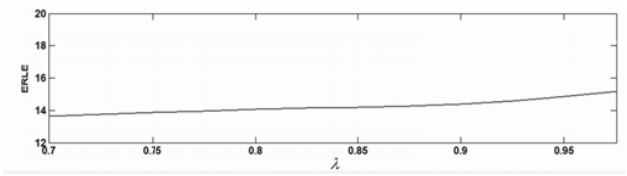
$$ELRE\ (dB) = 10\log_{10}\left[\frac{E[d^2(n)]}{E[e^2(n)]}\right]. \quad (13)$$

We calculated the LSD values during the DT period, and the ERLE is used when DT is not included. Table I compares the values of LSD and ERLE obtained.

As shown in Fig. 4 and Table I, the conventional AES technique usually obtained the best values of LSD and ERLE it can when λ was 0.925 — 0.975; however the proposed AES method obtained better ERLE and LSD value. It is evident that the performance of the proposed AES is superior to that of conventional AES.



(a)



(b)

Fig. 4 LSD and ERLE results of conventional method: (a) LSD. (b) ERLE.

TABLE I
Comparison of LSD and ERLE Results

λ	Average LSD	Average ERLE
0.7	1.3950	13.6500
0.725	1.3922	13.7691
0.75	1.3878	13.8809
0.775	1.3840	13.9798
0.8	1.3781	14.0651
0.825	1.3737	14.1389
0.85	1.3687	14.2099
0.875	1.3647	14.2926
0.9	1.3595	14.4133
0.925	1.3549	14.6041
0.95	1.3483	14.8803
0.975	1.3408	15.1858
Proposed $\lambda(n)$	1.2974	18.1502

IV. Conclusion

In this paper, we have described new AES techniques for speech recognition under barge-in situations. We have proposed an adaptive approach that updates weight parameters of echo path transfer function estimator in response to an abrupt change in the acoustic environment due to background noises or double talk. Furthermore, we have devised a voice activity detector and an initial time-delay estimator for barge-in speech recognition in communication networks. The initial time delay was estimated using log-spectral distance measure, as well as cross-correlation coefficients. Result of the objective

evaluation tests showed that the developed technique performs better than the conventional method.

ACKNOWLEDGMENT

This study was supported by the Research Program funded by the Seoul National University of Science and Technology.

REFERENCES

[1] S. Miyabe, Y. Hinamoto, H. Saruwatari, K. Shikano, and Y. Tatekura, "Interface for barge-in free spoken dialogue system based on sound field reproduction and microphone array," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 57470, 13 pages.

[2] M. M. Sondhi, "An adaptive echo canceler," *Bell Syst. Tech. J.,* vol. 46, pp. 497-510, Mar. 1967.

[3] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE trans. Acoust. Speech Sig. Processing*, vol. 27, no. 2, pp. 113–120, Nov. 1979.

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

[5] C. Faller and J. Chen, "Suppressing acoustic echo in a spectral envelope space," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 1048-1062, Sep. 2006.

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

[7] G. W. Elko, E. Diethorn, and T. G¨ansler, "Room impulse response variation due to thermal fluctuation and its impact on acoustic echo cancellation," *Proc. Intl. Workshop on Acoust. Echo and Noise Control (IWAENC)*, Kyoto, Japan, pp. 67-70, Sep. 2003.

[8] C. Faller and C. Tournery, "Robust echo control using a simple echo path model," *In Proc. IEEE Int. Conf. Acous., Speech Signal Processing*, vol. 5, pp. 281-284, 2006.

[9] T. Aboulnasr and K. Mayyas, "A robust variable step-size LMS-type algorithms: analysis and simulations," *IEEE Trans. on Signal Processing*, vol.45, no.3, pp.631-639, Mar. 1997.