Contemporary Engineering Sciences, Vol. 7, 2014, no. 22, 1113 - 1120
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2014.49139

Acoustic Echo Cancellation Techniques for Far-End Telephony Speech Recognition in Barge-In Situations

Jong Han Joo

Dept. of Electronic Engineering
Seoul National University of Science and Technology
Seoul 139-743, Korea

Jung Hoon Lee, Young Sun Kim, Seung Ho Choi*

Dept. of Electronic and IT Media Engineering
Seoul National University of Science and Technology
Seoul 139-743, Korea
*Corresponding author

Se Jin Chang

Voiceware Co., Ltd.
Seoul 133-832, Korea

Abstract

In this paper, we present some techniques of acoustic echo cancellation for far-end (server-side) telephony speech recognition during barge-in situations. We develop a normalized least mean square algorithm for the adaptive filter of an echo canceller, and a double-talk detector for the online speech recognition services. In particular, we devise a voice activity detector for estimating the initial delay due to communication networks. In addition, we propose a hybrid method that uses the log-spectral distance measure, as well as the cross-correlation coeffi-

cients, to estimate the initial delay. From the simulation and the experiments in real environments, we conclude that the developed techniques can be successfully used for far-end telephony speech recognition services.

Keywords: Barge-in, speech recognition, acoustic echo cancellation, voice activity detector, delay estimation, double-talk detection

1. Introduction

In a far-end (server-side) or a near-end (terminal-side) 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 [1]. Therefore, we use echo cancellation to remove echoes from the message signal that is input into the microphone. The quality of the echo canceller depends on the speed of convergence and the accuracy of the adaptive filter [2].
Echo paths consist of an initial time delay with no echo signal, and active regions in which the echo signal is present. To save computational costs and increase echo cancellation performance, we use an adaptive filter to match the echo path impulse response only in the active region. To accomplish this, we must develop an algorithm to estimate the initial delay and to identify the active region.

First, we devised an automatic voice activity detector (VAD) to find the active region of the message prompt signal and make a reference segment. Then, we developed 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 [3]. In this research work, we propose a hybrid method that uses the log-spectral distance (LSD) measure as well as the CCC to estimate the initial delay.

Since the echo signal is usually modeled as the convolution of the transmitted message signal and an echo path impulse response, an adaptive filter is used to estimate the echo path impulse response. However, the characteristics of the echo path vary depending on the surrounding conditions, and therefore, the echo canceller generally updates the filter coefficients using an adaptive algorithm [4]. We utilize a normalized least mean square (NLMS) algorithm as the adaptive algorithm.

The remainder of this paper is organized as follows: We describe the barge-in situation and our developed echo cancellation techniques in Section 2. Section 3 is a description about performance evaluation. Finally, we conclude this manuscript in Section 4.

2. Barge-In Situation and Developed Echo Cancellation Techniques

Fig. 1 shows the echo cancellation system in our research work. In the figure, $x(n)$ is a message prompt signal and $y(n)$ is the echo signal that is a portion of $d(n)$ transmitted from the near-end microphone. The $h(n)$ represents the echo path impulse response, and $\hat{h}(n)$ is an estimated impulse response. The NLMS updates the adaptive filter coefficients using $x(n)$ and $d(n)$, and the estimated echo signal $\hat{y}(n)$ is obtained. If the NLMS algorithm converges, the synthesized echo signal $\hat{y}(n)$ is the estimate of the echo signal $y(n)$. In a barge-in situation, consequently, the difference signal $e(n) = d(n) - \hat{y}(n)$ is the estimated user's voice $v(n)$.

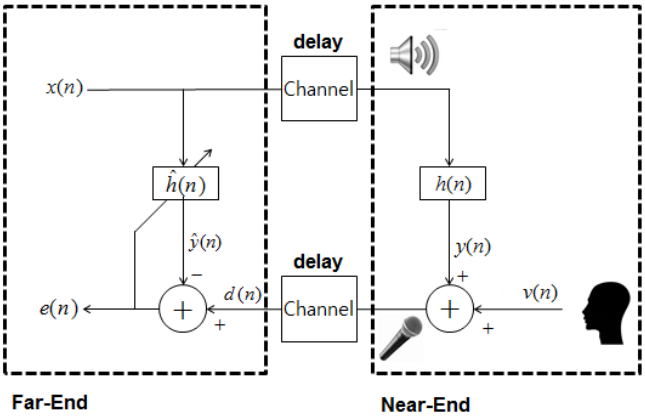


Fig. 1. System block diagram of the echo canceller

2.1. 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 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. To figure out that, we use the cross-correlation coefficient (CCC) denoted as

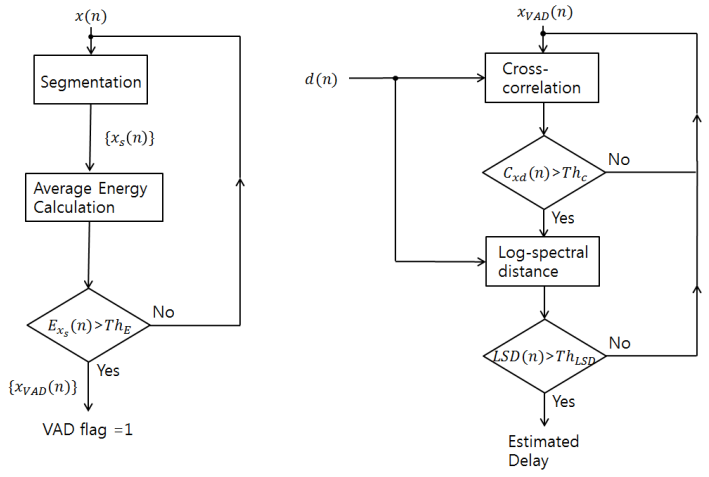
$$\mathrm{CCC} = \frac{\sum_{n=0}^{N-1} x_{VAD}(n)\, d_s(n)}{\sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x_{VAD}{}^2(n)}\sqrt{\frac{1}{N}\sum_{n=0}^{N-1} d_s{}^2(n)}} \quad . \tag{1}$$

The CCC has a peak value when the segment $\{d_s(n)\}$ is in accord with the

segment { $x_{\mathrm{VAD}}(n)$ }, 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

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

where $X_{VAD}(k)$ and $D_s(k)$ represent the discrete Fourier transform (DFT) spectra of $x_{\mathrm{VAD}}(n)$ and $d_{\mathrm{s}}(n)$, respectively. 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.



(a) (b)

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

2.2. NLMS Algorithm

The NLMS algorithms are a class of adaptive filter used to mimic the desired filter by finding the filter coefficients [5, 6]. In this research work, the NLMS algorithm is adopted to estimate the impulse response $h(n)$ of the room echo path, as shown in Fig.1. The filter coefficient $\hat{h}(n)$ is continuously updated for each sample using Eq. (3), where μ is the step size that determines the convergence speed [7].

$$\vec{\hat{h}}(n+1) = \vec{\hat{h}}(n) + \mu\frac{e(n)\vec{x}(n)}{\|\vec{x}(n)\|} \tag{3}$$

where $\vec{x}(n) = [x(n)x(n-1)\cdots x(n-p)]^T$.

2.3. Double-Talk Detection

In a barge-in phenomenon, when the near-end speech and far-end speech occur simultaneously, the so-called double-talk (DT) mode, the adaptation of the adaptive filter will be severely disturbed by the near-end signal [8]. If the recognition system stops transmitting the message signal when DT is detected, recognition performance will be greatly improved. In addition, the DT detection (DTD) result can be the criterion if the filter coefficients need to be updated.

The variable $e(n)$ has relatively lower and higher values before and after the user's voice occurs, respectively, as shown in Eqs. (4) and (5).

$$d(n) = v(n) + y(n), \tag{4}$$

$$e(n) = d(n) - \hat{y}(n) = v(n) + y(n) - \hat{y}(n). \tag{5}$$

We determine the DTD point when the normalized error energy $E_{err}(n)$ in Eq. (6) exceeds a predefined threshold, as shown in Fig. 3.

$$E_{err}(n) = \sum_{k=0}^{K-1} e_s^{\,2}(n-k) / \sum_{k=0}^{K-1} d_s^{\,2}(n-k) \tag{6}$$

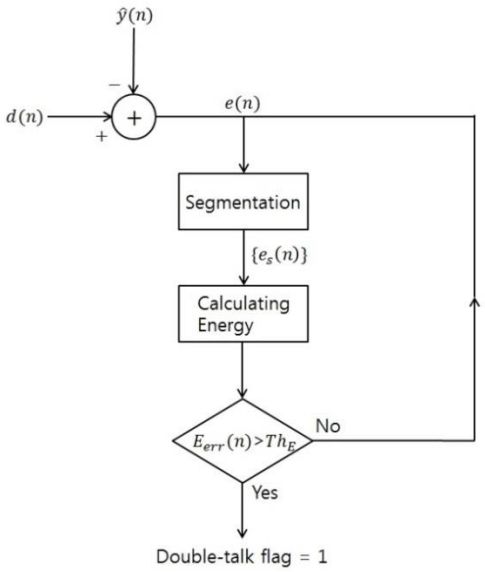


Fig. 3. Double-talk detection method

3. Performance Evaluation

3.1. Simulation

In the simulation, we used a woman's voice as the message signal $x(n)$, and the voice signals of five men and five women as the input signal $v(n)$. The audio files were sampled at 16 kHz. We assumed that the distance between the micro-

phone and speaker was 10 cm. In addition, we generated an artificial echo path impulse response for the simulation experiments.

First, we verified the proposed hybrid method that uses both the LSD measure and the CCC. Fig. 4 shows the CCC and LSD values obtained using two speech samples of the message and microphone input. We found that the maximum value of the CCCs and the minimum LSD value were at approximately the same position. Therefore, we concluded that the message signal matched closely with the microphone input signal at that position.

In order to find the appropriate filter length and step size, and to evaluate how well the acoustic echo signal was removed, we utilized the echo-return loss enhancement (ERLE) [9], which is defined as

$$ERLE(dB) = 10\log_{10}\left(\frac{E[y^2(n)]}{E[(y(n)-\hat{y}(n))^2]}\right) \tag{7}$$

where $\hat{y}(n)$ represents the estimated value of $y(n)$.

Fig. 5 shows the ERLE results when the filter lengths are 128, 256, and 512, and the step size belongs to (0.01, 0.1]. From the results, we determined that the proper filter length and step size were 256 and 0.01 – 0.02, respectively.

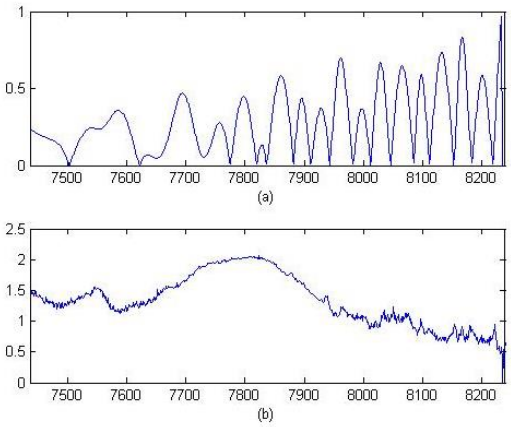


Fig. 4. Values of (a) CCCs and (b) LSD at the message signal, in accordance with the microphone input signal

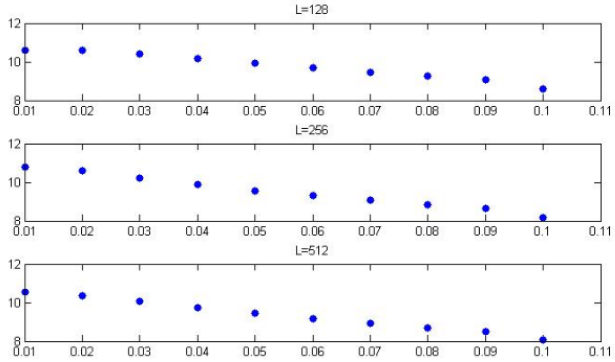


Fig. 5. ERLE results

3.2. Experiments in Real Environments

We integrated the developed techniques: VAD, delay estimation, echo cancellation, and double-talk detection. Then, we implemented the integrated program in a real environment. As shown in Fig. 6, each technique works well, and the implemented echo cancellation technology successfully operates in a real online environment.

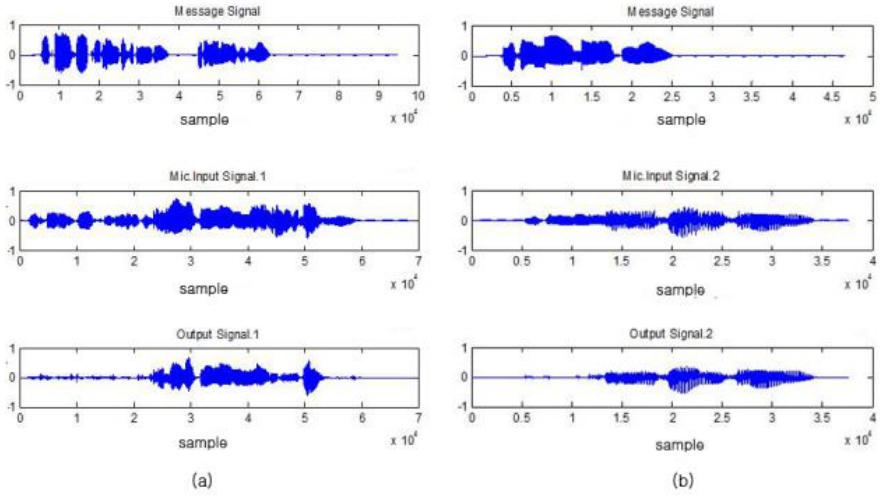


Fig. 6. Experimental results in real-time online environment: (a) man's voice, (b) woman's voice

4. Conclusion

We described some developed techniques of acoustic echo cancellation for far-end telephony speech recognition in a barge-in situation. We developed the adaptive filter and the double-talk detector for online speech recognition services. In particular, we devised a VAD to estimate the initial time delay in communication networks. Furthermore, we proposed a hybrid method that uses the LSD measure as well as the CCCs to estimate the initial delay. The simulation results and the experiment results in real environments showed that the developed techniques can be used successfully for far-end telephony speech recognition services.

Acknowledgements. This work was partly supported by ICT R&D program of MSIP/IITP [100252, Development of dialog-based spontaneous speech interface technology on mobile platform] and ICT R&D program of MSIP/IITP [2014-044-055-022, Loudness based broadcasting loudness and stress assessment of indoor environment noises].

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* (2007), Volume 2007, Article ID 57470.
[2] Y. Lu, R. Fowler, W. Tian, and L. Thompson, Enhancing echo cancellation via estimation of delay, *IEEE Transactions on Signal Processing* (2005), Vol. 53, No. 11, pp. 4159-4168.
[3] T. Aboulnasr and K. Mayyas, A robust variable step-size LMS-type algorithms: analysis and simulations, *IEEE Transactions on Signal Processing* (1997), Vol. 45, No. 3, pp. 631-639.
[4] K. Murano, S. Unagami, and F. Amano, Echo cancellation and applications, *IEEE Communications Magazine* (1990), Vol. 28, No. 1, pp. 49-55.
[5] S. Haykin, *Adaptive Filter Theory* (2002), Prentice Hall, New Jersey.
[6] B. Widrow and S. D. Stearns, *Adaptive Signal Processing* (1985), Prentice Hall, New Jersey.
[7] D. L. Duttweiler, Proportionate normalized least-mean-squares adaptation in echo cancelers, *IEEE Transactions on Speech and Audio Signal Processing* (2000), Vol. 8, No. 5, pp. 508-518.
[8] J. Lee and H. C. Huang, A robust double-talk detector for acoustic echo cancellation, *Proc. of International MultiConference of Engineers and Computer Scientists*, Vol. 2, (2010).
[9] J. Benesty, D. R. Morgan, and J. H. Cho, A new class of doubletalk detectors based on cross-correlation, *IEEE Transactions on Speech and Audio Processing* (2000), Vol. 8, No. 2, pp. 168-172.

Received: August 7, 2014