INTERSPEECH 2009 BRIGHTON

Soft Decision-Based Acoustic Echo Suppression in a Frequency Domain

Yun-Sik Park, Ji-Hyun Song, Jae-Hun Choi, Joon-Hyuk Chang

School of Electronic Engineering
Inha University
Incheon, Korea
{yspark,jhsong,jhchoi}@dsp.inha.ac.kr,changjh@inha.ac.kr

Abstract

In this paper, we propose a novel acoustic echo suppression (AES) technique based on soft decision in a frequency domain. The proposed approach provides an efficient and unified framework for such procedures as AES gain computation, AES gain modification using soft decision, and estimation of relevant parameters based on the same statistical model assumption of the near-end and far-end signal instead of the conventional strategies requiring the additional residual echo suppression (RES) step. Performances of the proposed AES algorithm are evaluated by objective tests under various environments and better results compared with the conventional AES method are obtained.

Index Terms: Acoustic Echo Suppression, Soft Decision

10.21437/Interspeech.2009-684

1. Introduction

Acoustic echoes often arise due to acoustic coupling between a loudspeaker and a microphone. Because these echoes can adversely affect conversation in hands-free telecommunication systems, acoustic echo suppressors have been employed to eliminate undesired echoes. To date, much work has been dedicated to the problem of reducing the effects of acoustic echo [1]-[5]. Most of the traditional acoustic echo suppression (AES) algorithms are based on an adaptive finite impulse response (FIR) filter for estimating the echo path response. The FIR filter in turn is based on normalized least-mean-square (NLMS) in the time domain, wherein an echo estimate is directly subtracted from the microphone input signal based on the estimated impulse response for the echo path [3], [4].

Recently, AES algorithms based on a spectral subtraction rule were proposed and impressive performance resulting from their flexibility and low computational complexity was demonstrated. In particular, Avandano proposed an efficient acoustic echo suppressor using spectral subtraction based on an interference estimation [4]. A low complexity AES algorithm was presented, in the method of Faller and Tourney [5], based on a spectral modification algorithm such as a Wiener filter by incorporating a gain filter mimicking the echo path [6].

In this paper, we propose a novel approach to the AES based on soft decision, where the near-end speech absence probability (NSAP) is derived in each frequency component to modify the AES gain derived from minimum mean square estimation (MMSE) for further echo reduction [7]-[9]. Based on this, the proposed method can efficiently suppress the acoustic echo without the help of an additional residual echo suppressor (RES). The performance of the proposed algorithm is evaluated by echo return loss enhancement (ERLE), speech attenuation (SA) and speech spectrogram tests and is demonstrated to be better than that of the conventional method.

2. Review of echo path response

Estimating the echo path response is a crucial component for AES operation due to the possible mismatch between the actual echo path and the estimated adaptive filter, which results in residual echo in many practical applications [10]. In this section, we briefly review the estimation for the echo path response in the discrete Fourier transform (DFT) domain as given by [5]. In Fig. 1, which presents an overall block diagram of the AES system, $|\hat{Y}(i,k)|$ denotes an estimate of the magnitude spectrum of the echo signal compared to the far-end speech signal $X(i,k)$ with a time index i and frequency index k. The estimated echo magnitude spectrum is then given by

$$|\hat{Y}(i,k)| = H(i,k)|X_d(i,k)|. \quad (1)$$

The gain filter $H(i,k)$ mimicking the response of the echo path is obtained by the magnitude of the least squares estimator [5]

$$H(i,k) = \left| \frac{E[X_d^*(i,k)Y(i,k)]}{E[X_d^*(i,k)X_d(i,k)]} \right| \quad (2)$$

where $*$ denotes the complex conjugate and d indicates d samples delay. Since the echo path is time varying, $H(i,k)$ is estimated iteratively as [5]

$$\hat{H}(i,k) = \frac{C(i,k)}{R(i,k)} \quad (3)$$

where

$$C(i,k) = \zeta_C C(i-1,k) + (1-\zeta_C)|X_d^*(i,k)Y(i,k)| \quad (4)$$

$$R(i,k) = \zeta_R R(i-1,k) + (1-\zeta_R)|X_d^*(i,k)X_d(i,k)|, \quad (5)$$

and $\zeta_C (= 0.998)$ and $\zeta_R (= 0.998)$ are smoothing parameters. It should be noted that the near-end speech may cause the echo path response filter $H(i,k)$ to diverge in a double talk situation. To prevent this problem, in this paper, we follow the cross-correlation coefficients-based double talk detection method proposed by [11] in the frequency domain.

3. Proposed acoustic echo suppression method based on soft decision

From the previous section, it is noted that the estimated echo magnitude spectrum is obtained using the least squares in the Fourier domain. Based on this, we propose a novel AES algorithm by taking advantage of soft decision. We consider near-end speech absence or presence, where the probability of near-end speech absence is introduced to modify the echo suppression spectral gain. For this, we first assume that two hypotheses,

Copyright © 2009 ISCA

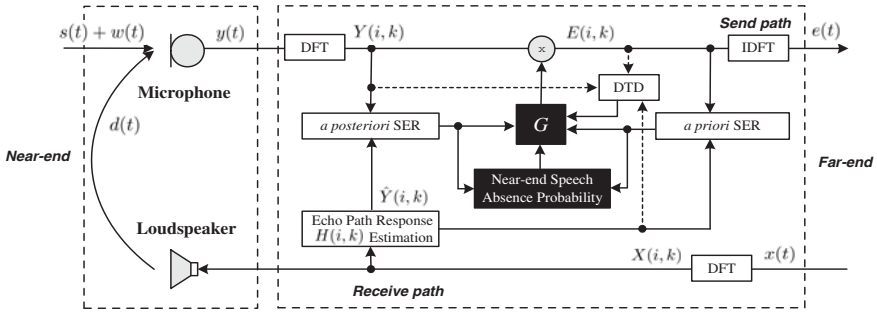


Figure 1: Block diagram of the proposed AES algorithm.

H_0 and H_1, indicate near-end speech absence and presence as follows:

$$\begin{aligned} H_0 &: \text{near-end speech absent} : Y(i,k) = D(i,k) \\ H_1 &: \text{near-end speech present} : Y(i,k) = D(i,k) + S(i,k) \end{aligned} \quad (6)$$

where $D(i,k)$, $S(i,k)$ and $Y(i,k)$, respectively, represent the Fourier domain spectra of the echo signal, the near-end speech and the signal picked up by the microphone. This time, the background noise is not taken into account since we assume that near-end speech absence is not correlated with the background noise.

Under the assumption that $D(i,k)$ and $S(i,k)$ are characterized by separate zero-mean complex Gaussian distributions, the following is obtained [7]-[9].

$$p(Y(i,k)|H_0) = \frac{1}{\pi\lambda_d(i,k)}\exp\Big[-\frac{|Y(i,k)|^2}{\lambda_d(i,k)}\Big] \quad (7)$$

$$p(Y(i,k)|H_1) = \frac{1}{\pi(\lambda_s(i,k)+\lambda_d(i,k))} \exp\Big[-\frac{|Y(i,k)|^2}{\lambda_s(i,k)+\lambda_d(i,k)}\Big] \quad (8)$$

where $\lambda_s(i,k)$ and $\lambda_d(i,k)$ are the variance of the near-end speech and estimated echo, respectively. The near-end speech absence probability (NSAP) $p(H_0|Y(i,k))$ for each frequency band is derived from Bayes' rule, such that [7]:

$$\begin{aligned} &p(H_0|Y(i,k)) \quad (9)\\ &= \frac{p(Y(i,k)|H_0)p(H_0)}{p(Y(i,k)|H_0)p(H_0)+p(Y(i,k)|H_1)p(H_1)} \\ &= \frac{1}{1+q\Lambda(Y(i,k))} \end{aligned}$$

where $q = p(H_1)/p(H_0)$ and $p(H_0)$ represents the *a priori* probability of near-end speech absence. Substituting (7) and (8) into (9), the likelihood ratio $\Lambda(Y(i,k))$ can be computed as follows:

$$\Lambda(Y(i,k)) = \frac{p(Y(i,k)|H_1)}{p(Y(i,k)|H_0)} = \frac{1}{1+\xi(i,k)}\exp\Big[\frac{\gamma(i,k)\xi(i,k)}{1+\xi(i,k)}\Big] \quad (10)$$

where the *a posteriori* signal-to-echo ratio (SER) $\gamma(i,k)$ and the *a priori* SER $\xi(i,k)$ are defined by

$$\gamma(i,k) \equiv \frac{|Y(i,k)|^2}{\lambda_d(i,k)} \quad (11)$$

$$\xi(i,k) \equiv \frac{\lambda_s(i,k)}{\lambda_d(i,k)}. \quad (12)$$

In time, $\xi(i,k)$ is estimated with the help of the well-known decision-directed approach with $\alpha_{DD} = 0.6$ [8]. Then

$$\hat{\xi}(i,k) = \alpha_{DD}\frac{|\hat{S}(i-1,k)|^2}{\lambda_d(i-1,k)} + (1-\alpha_{DD})P[\gamma(i,k)-1] \quad (13)$$

where $P[z] = z$ if $z \geq 0$, and $P[z] = 0$ otherwise. Also, $\hat{\lambda}_d(i,k)$ is the estimate for $\lambda_d(i,k)$ and the power spectrum of the echo signal can be estimated when the near-end speech signal is not present in the observation, as given by

$$\hat{\lambda}_d(i,k) = \zeta_{\lambda_d}\hat{\lambda}_d(i-1,k) + (1-\zeta_{\lambda_d})|\hat{Y}(i,k)|^2 \quad (14)$$

where ζ_{λ_d} is a smoothing parameter and $E[|\hat{Y}(i,k)|^2]$ is given by (1).

Under the assumption that the spectral components of the input signal at the microphone are statistically independent, we employ the MMSE estimator obtained for the echo suppression from the observation $Y(i,k)$ as follows:

$$|\hat{S}(i,k)| = E[|S(i,k)||Y(i,k)]. \quad (15)$$

It is useful to consider the estimated near-end speech spectrum $\hat{S}(i,k)$ as being achieved from $Y(i,k)$ by a multiplicative gain function G_{MMSE} as given below

$$\hat{S}(i,k) = G_{MMSE}\Big(\xi(i,k),\gamma(i,k)\Big)\cdot Y(i,k) \quad (16)$$

in which the gain $G_{MMSE}(\cdot,\cdot)$ of the MMSE estimator is given by [8]

$$\begin{aligned} G\Big(\xi(i,k),\gamma(i,k)\Big) &= \frac{\sqrt{\pi\upsilon(i,k)}}{2\gamma(i,k)}\exp\Big(-\frac{\upsilon(i,k)}{2}\Big) \quad (17)\\ &\cdot\Big[\Big(1+\upsilon(i,k)\Big)I_0\Big(\frac{\upsilon(i,k)}{2}\Big)+\upsilon(i,k)I_1\Big(\frac{\upsilon(i,k)}{2}\Big)\Big] \end{aligned}$$

and I_0 and I_1 are modified Bessel functions of zero and first order. Also, $\upsilon(i,k)$ is defined as follows:

$$\upsilon(i,k) = \frac{\xi(i,k)}{1+\xi(i,k)}\gamma(i,k). \quad (18)$$

Finally, in the proposed AES scheme, the echo suppression gain derived from MMSE estimation is combined with the NSAP depending on soft decision for robust performance, as given by

$$\hat{S}(i,k) = \Big(1-p\big(H_0|Y(i,k)\big)\Big)G_{MMSE}\Big(\hat{\xi}(i,k),\hat{\gamma}(i,k)\Big)Y(i,k) \quad (19)$$

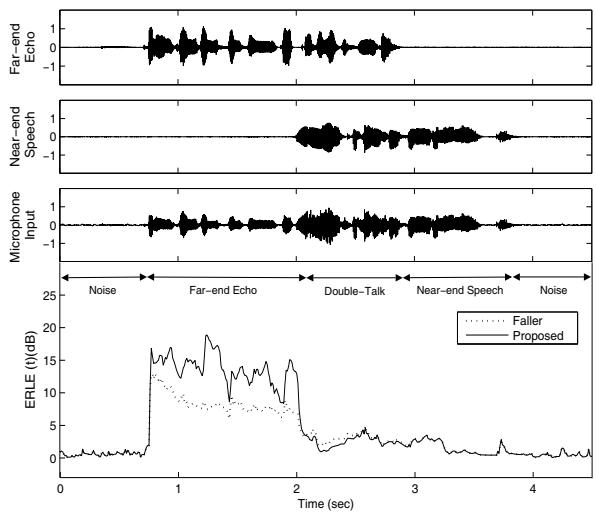


Figure 2: Time variation of $ERLE(t)$ (vehicular noise, SNR=20 dB)

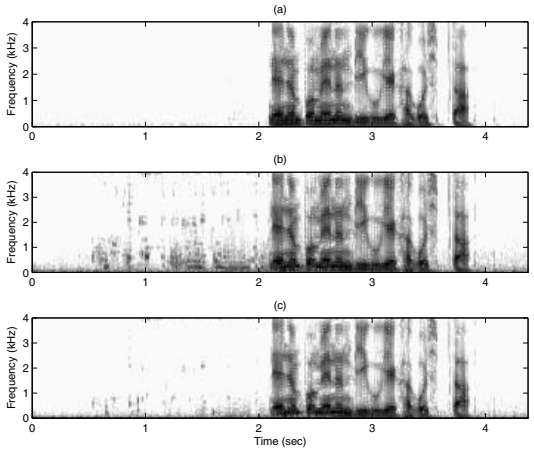


Figure 3: Speech spectrograms (vehicular noise, SNR=20 dB) (a) Spectrogram of the clean and near-end speech signal (b) Spectrogram of output signal obtained by Faller [5] (c) Spectrogram of output signal obtained by the proposed method.

where we can see that a better echo suppression rule is formulated to apply greater attenuation for a given frequency bin consisting of echo alone.

In addition, the proposed echo suppression gain-modification is further improved in that distinct values of q's in (9) are estimated for different frames and frequency bins such as $q(i,k)$ that can be tracked in time [12]. Therefore, the proposed algorithm employs a decision rule to decide whether the near-end speech signal is present in the kth bin, as given by

$$q(i,k) = \alpha_q q(i-1,k) + (1-\alpha_q) I(i,k) \qquad (20)$$

in which the smoothing parameter α_q is set as 0.7 and $I(i,k)$ denotes an indicator function for the result in (21), i.e., $I(i,k) = 1$ if $\gamma(i,k) > \gamma_{th}$ and $I(i,k) = 0$ otherwise. The value of $q(i,k)$ can be easily updated using the *a posteriori* SER $\gamma(i,k)$ as follows:

$$\gamma(i,k) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma_{th} \qquad (21)$$

where the threshold γ_{th} is set to 3.0 considering the desired significance level.

4. Experimental results

In order to evaluate the performance of the proposed AES algorithm, we conducted objective comparison experiments under various noise conditions. Twenty test phrases, spoken by seven speakers and sampled at 8 kHz, were used as the experimental data. For assessing the performance of the proposed method, we artificially created 20 data files, where each file was obtained by mixing the far-end signal with the near-end signal. Each frame of the windowed signal was transformed into its corresponding spectrum through 128-point DFT after zero padding. We then constructed 16 frequency bands employing the sub-band combining to cover whole frequency ranges ($\sim$4 kHz) of the narrow band speech signal which is analogous to that of the IS-127 noise suppression algorithm [13]. The far-end speech signal was passed through a filter simulating the acoustic echo path before being mixed [14], [15]. The simulation environment was designed to fit a small office room having a size of $5\times4\times3\ m^3$. The echo level measured at the input microphone was 3.5 dB lower than that of the input near-end speech on average. In order to create noisy conditions, white, babble and vehicular noises from the NOISEX-92 database were added to clean near-end speech signals at signal-to-noise ratios (SNRs) of 10, 15, 20 and 25 dB. For the purpose of an objective comparison, we evaluated the performance of the proposed scheme and that of the conventional AES algorithm proposed by Faller *et al.* [5], [6]. The performance of the approach was measured in terms of echo return loss enhancement (ERLE), speech attenuation (SA), which are defined by [15]

$$ERLE(t) = 10\log_{10}\Big[\frac{E[y^2(t)]}{E[e^2(t)]}\Big] \qquad (22)$$

$$SA(t) = \frac{1}{N}\sum^{N} 10\log_{10}\Big[\frac{E[s^2(t)]}{E[\tilde{s}^2(t)]}\Big] \qquad (23)$$

where t is a sample index, $E[\cdot]$ denotes the expected value, N is the number of samples during the double-talk periods and $\tilde{s}^2(t)$ denotes the near-end speech component in the output signal $e(t)$.

In Fig. 2, an example of $ERLE(t)$ variation over time demonstrates that the proposed algorithm attenuates the echo signal more efficiently than the conventional AES technique while preserving the near-end signal quite well during the double-talk periods. Also, the speech spectrograms are presented in Fig. 3. Figs. 3(b) and 3(c) show the spectrograms obtained with the conventional and proposed algorithm, respectively. In the proposed method, the residual echo is further reduced compared to the conventional technique during the active far-end echo period. Finally, given the three types of noise environments, overall results for the aforementioned 20 data files are shown in Table 1 and Table 2. ERLE and SAs scores were averaged to yield final mean score results for the case of three types-noise sources. From Table 1, it is evident that in most noisy conditions, the proposed AES algorithm based on soft decision yielded a higher ERLE compared to the conventional technique. The SAs of the proposed method during double-talk periods are shown in Table 2, where we can observe that the SAs of the proposed scheme based on soft decision were better than

Table 1: Comparison of ERLE results obtained from the proposed AES algorithm with respect to the conventional Faller's method during the far-end echo period.

Environments		ERLE (dB)	
Noise	SNR (dB)	Faller	Proposed
White	10	5.72	**7.16**
	15	6.82	**9.65**
	20	7.45	**11.53**
	25	7.73	**12.70**
Babble	10	5.50	**6.77**
	15	6.60	**9.25**
	20	7.31	**11.19**
	25	7.67	**12.31**
Vehicle	10	5.14	**6.68**
	15	6.49	**9.35**
	20	7.29	**11.38**
	25	7.66	**12.62**
Clean speech	∞	7.90	**13.70**

Table 2: Comparison of SA results obtained from the proposed AES algorithm with respect to the conventional Faller's method during double-talk.

Environments		SA (dB)	
Noise	SNR (dB)	Faller	Proposed
White	10	1.31	**1.05**
	15	1.45	**1.21**
	20	1.51	**1.29**
	25	1.53	**1.32**
Babble	10	1.36	**1.12**
	15	1.48	**1.25**
	20	1.52	**1.30**
	25	1.54	**1.32**
Vehicle	10	1.21	**0.99**
	15	1.41	**1.18**
	20	1.50	**1.28**
	25	1.53	**1.32**
Clean speech	∞	1.54	**1.33**

that of the previous scheme in all the tested conditions. Summarizing the overall results, the proposed approach is found to be effective in the AES technique.

5. Conclusions

In this paper, we have proposed a novel AES algorithm based on a soft decision scheme in the frequency domain. The MMSE estimator-based filter is applied to the AES gain in conjunction with the soft decision scheme considering the probability of near-end speech absence for effective echo suppression. The performance of the proposed algorithm has been found to be superior to that of the conventional technique through objective evaluation tests.

6. Acknowledgements

This work was partly supported by ETRI SoC Industry Promotion Center and This work was partly supported by the IT R&D program of MKE/IITA [2008-F-045-01].

7. References

[1] S. Gustafsson, R. Martin and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony," *Signal Processing*, vol. 64, no. 1, pp. 21-32, Jan. 1998.

[2] C. Beaugeant, V. Turbin, P. Scalart and A. Gilloire, "New optimal filtering approaches for hands-free telecommunication terminals," *Signal Processing*, vol. 64, no. 1, pp. 33-47, Jan. 1998.

[3] P. S. R. Diniz, Adaptive Filtering: Algorithm and Practical Implementation. Norwell, MA: Kluwer, 1997.

[4] C. Avendano, "Acoustic echo suppression in the STFT domain," in *Proc. IEEE Workshop on Appl. of Sig. Proc. to Audio and Acoust.*, Oct. 2001.

[5] C. Faller and C. Tournery, "Robust echo control using a simple echo path model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Processing*, vol. 5, pp. V281-V284, 2006.

[6] 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, Sept. 2005.

[7] R. J. McAualy and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 137-145, Apr. 1980.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984.

[9] N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Processing Letters*, vol. 7, no. 5, pp. 108-110, May 2000.

[10] V. Turbin, A. Gilloire and P. Scalart, "Comparison of three post-filtering algorithms for residual acoustic echo reduction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Processing*, pp. 307-310, 1997.

[11] S. J. Park, C. G. Cho, C. Lee and D. H. Youn, "Integrated echo and noise canceler for hands-free applications," *IEEE Trans. on Circuits and Systems II*, vol. 49, issue 3, pp. 186-195, Mar. 2002.

[12] D. Malah, R. Cox and A. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Processing*, pp. 789-792, 1999.

[13] TIA/EIA/IS-127, "Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems," 1996.

[14] S. McGovern, A Model for Room Acoustics, 2003 [Online]. Available: http://2pi.us/rir.html

[15] S. Y. Lee and N. S. Kim, "A statistical model based residual echo suppression," *IEEE Signal Processing Letters*, vol. 14, no. 10, pp. 758-761, Oct. 2007.