ADAPTATION MODE CONTROL WITH RESIDUAL NOISE ESTIMATION FOR BEAMFORMER-BASED MULTI-CHANNEL SPEECH ENHANCEMENT

Seon Man Kim[1], *Hong Kook Kim*[1], *Sung Joo Lee*[2], *and Yun Keun Lee*[2]

[1] School of Information and Communications
Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea
[2] Speech/Language Information Research Center
Electronics and Telecommunications Research Institute (ETRI), Daejeon 305-700, Korea
[1]{kobem30002, hongkook}@gist.ac.kr, [2]{lee1862, yklee}@etri.re.kr

ABSTRACT

In this paper, we propose a new adaptation mode controller (AMC) for a generalized sidelobe canceller (GSC) having prior knowledge of the direction-of-arrival (DOA) of a desired speech source. In order to optimize the adaptation mode of a GSC, the residual noise remaining in the GSC output must be employed for adapting the AMC. The residual noise in the GSC output is estimated by using a short-time Fourier transform (STFT)-based Wiener filter, where *a priori* signal-to-noise ratio (SNR) and *a posteriori* target-to-non-target-directional signal ratio (TNR) are estimated based on a decision-directed approach and a DOA-based approach, respectively. The estimated residual noise is finally incorporated as a control parameter into the adaptive filters in the AMC. The performance of the proposed AMC is evaluated by measuring the perceptual evaluation of speech quality (PESQ) scores and cepstral distortion in car noise environments with SNRs from 0 to 20 dB. Experimental results show that the proposed AMC performs better than the conventional AMCs.

Index Terms— Speech enhancement, array signal processing, adaptive beamformer, generalized sidelobe canceller, adaptation mode controller

1. INTRODUCTION

The generalized sidelobe canceller (GSC) is one of the most popular adaptive beamformers due to its structural simplicity and ease of implementation [1][2]. The GSC mainly consists of a fixed beamformer (FBF), blocking matrix (BM), and noise canceller (NC). The FBF provides a fixed sound beam in the target direction so that non-target-directional signals are attenuated. In contrast, the BM blocks the target-directional signal so that only non-target-directional signals can pass through. The NC generates an enhanced target-directional signal from the FBF and BM output signals. However, the output signal processed by the BM includes a part of the target-directional signal, which is called residual target signal. Thus, the residual target signal can degrade the performance of the NC. To overcome such a problem, an adaptation mode controller (AMC) has been used for the NC, where the filter adaptation is selectively performed depending on the target signal activity estimation [3].

The AMC can be classified into two categories: hard-decision based scheme and soft- decision based one. The hard-decision AMC considers two cases: presence and absence of the target signal in each observed frame signal [3]. On the other hand, a soft-decision AMC provides probabilistic values ranging from zero to one for the presence of the target signal to control the adaption mode. It has been reported in [4][5] that a soft-decision AMC performed better than a hard-decision AMC. In fact, a hard- or a soft-decision AMC can be used to estimate the residual noise components of a GSC output, which is used in the adaptive algorithm of the NC. Thus, the reliability of the adaptive weight is strongly dependent on the estimate of the residual noise of the GSC output, which is explained in detail in Section 2.

In this paper, we propose a new AMC driven by the residual noise estimate of a GSC output. In particular, we only consider the adaptation mode in the NC. Note that because an AMC aims to minimize the target-signal cancellation of a GSC output, originating from the target-signal leakage of the BM, the extent to which the noise reduction performance of an AMC can be improved is limited. While several techniques for the performance improvement of noise reduction have been introduced in [6][7], we only consider the performance for the adaptation mode of the AMC.

This paper is organized as follows. Following this introduction, in Section 2, we review the conventional AMCs for GSCs from the viewpoint of the residual noise of a GSC output. In Section 3, we describe the overall procedure of the GSC with the proposed AMC, where the estimation process of the residual noise of a GSC output is explored. Section 4 describes the target-speech enhancement experiments performed using 3-channel audio signals recorded in car noise environments. The performance of the proposed AMC is evaluated by measuring the perceptual evaluation of the speech quality (PESQ) scores and cepstral distortion. Finally, we summarize our findings in Section 5.

2. BEAMFORMER-BASED MULTI-CHANNEL SPEECH ENHANCEMENT

In this section, a beamformer-based multi-channel speech enhancement system is briefly reviewed and a conventional AMC is introduced. To begin with, let $T_k(\ell)$, $E_k(\ell)$, and $Y_k(\ell)$ be target speech, residual noise, and GSC output at the k-th frequency bin $(k = 0,1,\cdots,K-1)$ and ℓ-th frame $(\ell = 0,1,\cdots)$. Then,

$$Y_k(\ell) = T_k(\ell) + E_k(\ell). \quad (1)$$

Next, the FBF output $B_k(\ell)$ is defined as a mixture of the target speech $T_k(\ell)$ and the non-target noise $N_k(\ell)$ as

$$B_k(\ell) = T_k(\ell) + N_k(\ell). \quad (2)$$

Let $w_k(\ell)$ and $Z_{m,k}(\ell)$ denote the adaptive weights in the NC and the BM output of the *m*-th channel, respectively. Then, the GSC output estimate $Y_k(\ell)$ is obtained as

$$Y_k(\ell) = B_k(\ell) - R_k(\ell), \quad (3)$$

where

$$R_k(\ell) = \sum_{m=1}^{M-1} w_{m,k}(\ell) Z_{m,k}(\ell). \quad (4)$$

In Eq. (4), $w_{m,k}(\ell)$ is usually updated by minimizing $Y_k(\ell)$ using a normalized least mean square (NLMS) algorithm as

$$w_{m,k}(\ell+1) = w_{m,k}(\ell) + \mu_w \frac{Z_{m,k}(\ell)}{\| Z_{m,k}(\ell) \|} Y_k(\ell), \quad (5)$$

where μ_w is the step size. In Eq. (5), because the weight $w_{m,k}(\ell)$ is updated by minimizing $\hat{Y}_k(\ell) = T_k(\ell) + E_k(\ell)$ without distortion of $T_k(\ell)$, $Y_k(\ell)$ can also be replaced with $E_k(\ell)$, i.e.,

$$w_{m,k}(\ell+1) = w_{m,k}(\ell) + \mu_w \frac{Z_{m,k}(\ell)}{\| Z_{m,k}(\ell) \|} E_k(\ell). \quad (6)$$

Let $G_k(\ell)$ be a productive value ranging from zero to one according to the AMC strategy, which is applied to adapting the NLMS algorithm in Eq. (5). Therefore, we have

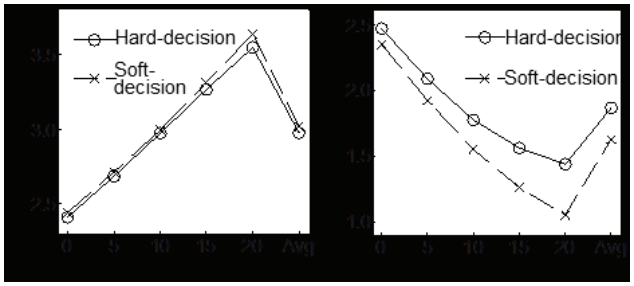
$$w_{m,k}(\ell+1) = w_{m,k}(\ell) + \mu_w \frac{Z_{m,k}(\ell)}{\| Z_{m,k}(\ell) \|} Y_k(\ell)\big(1 - G_k(\ell)\big), \quad (7)$$

where $G_k(\ell)$ is a binary value, i.e., $G_k(\ell) = 0$ or $G_k(\ell) = 1$, in the target signal absence or presence interval, respectively, according to the hard decision AMC strategy [3][8][9]. In fact, $G_k(\ell)$ corresponds to a target signal presence probability, according to the soft-decision AMC strategy [4][5]. Here, when $V_k(\ell)$ is defined as the GSC output weighted by the probability of the target signal, i.e., $V_k(\ell) = Y_k(\ell) \cdot \big(1 - G_k(\ell)\big)$, $V_k(\ell)$ becomes the approximate of the residual noise component $E_k(\ell)$. Thus, the adaptation mode control by AMC is linked to the estimate of the residual noise $\hat{E}_k(\ell)$. In other words, the reliability of the adaptive weight $w_{m,k}(\ell)$ is strongly dependent on the estimate of $\hat{E}_k(\ell)$.

3. PROPOSED AMC

Let $X_{m,k}(\ell)$ be a spectral component of the *m*-th channel input signal at the *k*-th frequency bin $(k = 0,1,\cdots,K-1)$ and ℓ-th frame. Fig. 1 shows a block diagram of a GSC with the proposed AMC. First, the *a posteriori* target-to-non-target signal ratio (TNR) $\tilde{\eta}_k(\ell)$ is computed from the phase differences among multiple signals. Then, $\tilde{\eta}_k(\ell)$ and the FBF output $B_k(\ell)$ are used to design a Wiener filter in the frequency domain, which also includes a statistical voice activity detector (VAD) [10] by using $B_k(\ell)$. The designed Wiener filter is then used to estimate the residual noise components $\hat{E}_k(\ell)$ of the GSC output $Y_k(\ell)$. Finally, $\hat{E}_k(\ell)$ is used as the adaptation control parameter for the AMC in the NC.

Fig.1. Block diagram of a GSC with the proposed soft-decision AMC based on a residual non-target noise estimate.



(a) (b)

Fig.2. Performance comparison of (a) PESQ scores and (b) cepstral distortion (dB) of hard- and soft-decision AMC approaches with known residual noise.

3.1. Adaptation mode control with known residual noise

The residual noise $E_k(\ell)$ of the GSC output $Y_k(\ell)$ is represented as the difference between $N_k(\ell)$ and $R_k(\ell)$ from Eqs. (1), (2), and (3). That is,

$$E_k(\ell) = N_k(\ell) - R_k(\ell), \quad (8)$$

where $R_k(\ell)$ is determined by $w_{m,k}(\ell)$ and $Z_{m,k}(\ell)$ as described in Eq. (4), thus the residual noise estimate $\hat{E}_k(\ell)$ is dependent on the non-target noise estimate $\hat{N}_k(\ell)$ of the FBF output $B_k(\ell)$.

Fig. 2 compares the PESQ scores and cepstral distortion of the GSCs with the hard and soft decision AMCs. In particular, for the hard decision AMC, the target signal is detected in each frequency bin by using the target clean signal. On the other hand, for the soft decision AMC, the pure noise component $N_k(\ell)$ of the FBF output is employed in Eqs. (6) and (8). As shown in Fig. 2, for both PESQ scores and cepstral distortion and for all input signal-to-noise ratios

(SNRs), the soft decision AMC performs better than the binary decision AMC.

3.2 Adaptation mode control with unknown residual noise

In order to optimize the adaptation mode in Eq. (6), $\hat{N}_k(\ell)$ should be first estimated, because it is necessary for obtaining the residual noise estimate $\hat{E}_k(\ell) = \hat{N}_k(\ell) - R_k(\ell)$. Here, a short-time Fourier transform (STFT) based Wiener filter is used to estimate $\hat{N}_k(\ell)$. The statistical model-based VAD is employed to provide the target signal activity information, where the VAD is driven by two situations $H_0 : B_k(\ell) = N_k(\ell)$ and $H_1 : B_k(\ell) = T_k(\ell) + N_k(\ell)$. The non-target noise spectral variance $\hat{\lambda}_{N,k}(\ell)$ is estimated by a recursive procedure executed only when H_0 is determined to be true. In other words, we have

$$\hat{\lambda}_{N,k}(\ell) = \zeta_N \cdot \hat{\lambda}_{N,k}(\ell-1) + (1-\zeta_N) \cdot |B_k(\ell)|^2, \tag{9}$$

where ζ_N is a smoothing parameter. Then, the *a priori* SNR estimate $\hat{\xi}_k(\ell)$ and *a posteriori* $\hat{\varphi}_k(\ell)$ are obtained by the decision-directed (DD) approach [11]. Thus, we have

$$\hat{\xi}_k(\ell) = \zeta_{DD} \cdot \frac{|\hat{T}_k(\ell-1)|^2}{\hat{\lambda}_{N,k}(\ell-1)} + (1-\zeta_{DD}) \cdot \max\left(\hat{\varphi}_k(\ell) - 1,\ 0\right), \tag{10}$$

$$\hat{\varphi}_k(\ell) = \frac{|B_k(\ell)|^2}{\hat{\lambda}_{N,k}(\ell)}. \tag{11}$$

Because recursive smoothing approaches in the DD estimator rely on a stationary condition between successive spectral magnitudes, their performance is limited when noise is non-stationary. In order to overcome the performance limitation of the conventional DD approach, a DOA-based *a posteriori* TNR in [5] is employed for the *a priori* SNR estimation. That is, if the target-directional enhanced and rejected powers are denoted as $|\tilde{T}e_{m,k}(\ell)|^2$ and $|\tilde{T}r_{m,k}(\ell)|$, respectively, we can compute the DOA-based TNR $\eta_{m,k}(\ell)$ as

$$\eta_{m,k}(\ell) = \frac{|\tilde{T}e_k(\ell)|^2}{|\tilde{T}r_k(\ell)|^2} = \frac{1+\cos\phi_{m,k}(\ell)}{4\cdot\left(1-\cos\phi_{m,k}(\ell)\right)}, \tag{12}$$

where m refers to the counterpart channel of a reference channel among a pair of channels and $\phi_{m,k}(\ell)$ is the phase difference between the m-channel and reference-channel signals. Then, the *a posteriori* TNR $\tilde{\eta}_k(\ell)$ is estimated as the average value of $\eta_{m,k}(\ell)$ over $m = (2,3,\cdots M)$ from all microphone pairs including the reference channel. That is, we have

$$\tilde{\eta}_k(\ell) = \frac{1}{M-1} \sum_{m=2}^{M} \eta_{m,k}(\ell). \tag{13}$$

Let $\Re(\tilde{\eta}_k(\ell))$ be an amplitude reduction function for non-stationary noise residual components, which are not estimated by the DD approach. The non-stationary noise reduced version of $|\hat{T}_k(\ell)|$, $|\hat{T}'_k(\ell)| = |\hat{T}_k(\ell)| \cdot \Re(\tilde{\eta}_k(\ell))$, leads to the enhanced *a priori* SNR $\hat{\xi}'_k(\ell)$ via non-stationary noise, which is denoted as

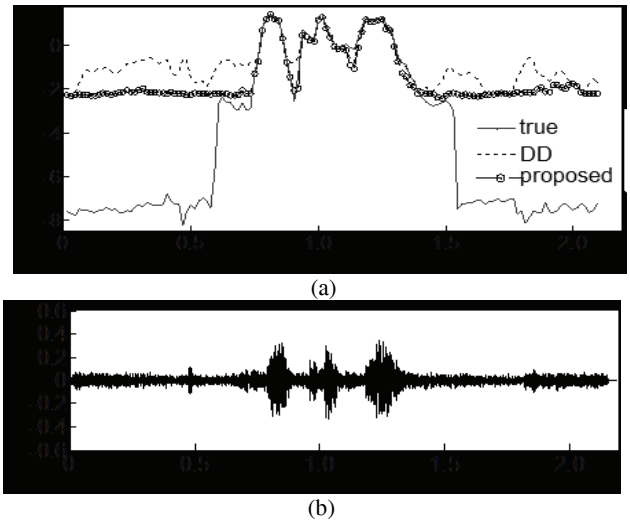


Fig.3. Illustration of (a) SNR estimated by conventional DD and proposed approach in conjunction with (b) a noisy speech having the SNR of 5 dB.

$$\begin{aligned}\hat{\xi}'_k(\ell) &= \frac{|\hat{T}'_k(\ell)|^2}{\hat{\lambda}_{N,k}(\ell)} = \frac{|\hat{T}_k(\ell)\cdot\Re(\tilde{\eta}_k(\ell))|^2}{\hat{\lambda}_{N,k}(\ell)} \\ &= \frac{|\hat{T}_k(\ell)|^2}{\hat{\lambda}_{N,k}(\ell)} \cdot |\Re(\tilde{\eta}_k(\ell))|^2 = \hat{\xi}_k(\ell) \cdot |\Re(\tilde{\eta}_k(\ell))|^2,\end{aligned} \tag{14}$$

where $\Re(\tilde{\eta}_k(\ell))$ can be estimated using the Wiener-Hoff equation with $\tilde{\eta}_k(\ell)$ as

$$\Re(\tilde{\eta}_k(\ell)) \approx \frac{1}{1+\mu\cdot(1/\eta_k(\ell))}, \tag{15}$$

where μ is a control parameter.

Fig. 3 shows the behavior of the overall SNR estimated from the DOA-based TNR with the DD approach for noisy speech. It is shown from Fig. 3(a) that the SNR estimated by the proposed approach is significantly better matched with the true SNR than the conventional DD does.

Next, the Wiener filter coefficient $F_k(\ell)$ is obtained to estimate the non-target noise $\hat{N}_k(\ell)$ of the FBF output as

$$F_k(\ell) = 1 - \frac{\hat{\xi}'_k(\ell)}{\hat{\xi}'_k(\ell)+1}. \tag{16}$$

Then, the residual noise component $\hat{E}_k(\ell)$ of the GSC output is estimated using the following rules:

$$H_0 : \hat{E}_k(\ell) = B_k(\ell) - R_k(\ell), \tag{17a}$$

$$H_1 : \hat{E}_k(\ell) = B_k(\ell)\cdot F_k(\ell) - R_k(\ell). \tag{17b}$$

Finally, the adaptation algorithm driven by residual noise is represented as

$$w_{m,k}(\ell+1) = w_{m,k}(\ell) + \mu_w \frac{Z_{m,k}(\ell)}{\|Z_{m,k}(\ell)\|}\hat{E}_k(\ell). \tag{18}$$

Table 1
Comparison of PESQ scores according to different AMCs

SNR (dB)	No	Wiener	[3]	[8]	[9]	[5]	Proposed
20	3.26	3.49	3.39	3.52	3.47	3.42	3.58
15	2.91	3.23	3.14	3.24	3.18	3.19	3.24
10	2.63	2.87	2.88	2.93	2.87	2.92	2.94
5	2.37	2.52	2.62	2.64	2.60	2.65	2.66
0	2.0	2.22	2.38	2.38	2.35	2.39	2.39
Avg	2.63	2.87	2.88	2.94	2.90	2.92	2.96

Table 2
Comparison of cepstral distortion according to different AMCs

SNR (dB)	No	Wiener	[3]	[8]	[9]	[5]	Proposed
20	1.94	2.01	2.00	1.76	1.79	1.91	1.32
15	2.46	2.52	2.31	2.07	2.12	2.18	1.74
10	3.19	3.29	2.80	2.59	2.65	2.64	2.33
5	4.04	4.18	3.40	3.26	3.30	3.26	3.08
0	4.98	5.11	4.12	4.05	4.07	4.02	3.93
Avg	3.32	3.42	2.92	2.74	2.79	2.80	2.48

4. PERFORMANCE EVALUATION

A triple-microphone array database with a 4-cm equivalent space was used in a car noise environment. The detail description was already described in [12]. The desired speech signal was played from an acoustic speaker mounted below the headrest of the driver's seat, and background music was played from the car audio system in drive mode with an average velocity ranging from 60 to 80 km/h. Ten speech (five males and five females) and ten music audio signals were separately recorded at a sampling rate of 16 kHz and artificially mixed with different SNRs ranging from 0 to 20 dB in a step of 5 dB. Subsequently, 100 noisy speech signals were prepared as a test database for each SNR. The test signals were segmented by a half overlapping rectangular window whose length was 32 ms, and the GSC processed signal was synthesized by a cosine window. In this experiment, the frame size was set to 512 samples, which corresponded to 32 ms. In addition, $\zeta_N = 0.95$ in Eq. (9), $\zeta_{DD} = 0.8$ in Eq. (10), $\mu = 10$ in Eq. (15), and $\mu_w = 0.06$ in Eq. (18).

In order to evaluate the performance of the proposed AMC, we used two objective quality measures such as the PESQ [13] score and cepstral distortion [14]. Table 1 compares the PESQ scores of speech signals processed by the GSCs with different AMCs. In particular, the performance of the Wiener filter alone was presented, which was also used to estimate the residual noise components of the proposed AMC. All the AMC approaches in [3], [8], [9], and [5] were used for the adaptation process in the NC to carry out a fair comparison. It was shown from Table 1 that compared to other AMC approaches, the proposed AMC achieved the highest scores for all the input SNRs. On one hand, Table 2 compares the cepstral distortion of speech signals processed by the GSCs with different AMCs. It was also shown from the table that the proposed AMC provided the lowest distortion for all the input SNRs among all the AMCs. Thus, it can be concluded here that the GSC with the proposed AMC outperformed those with conventional AMCs for all the input SNRs in terms of PESQ score and cepstral distortion.

5. CONCLUSION

In this paper, we proposed a new AMC for GSCs. The proposed AMC could provide an optimal adaptation mode by exploring the characteristics of the residual noise in a GSC output. In other words, the residual noise component of the GSC output was estimated using a Wiener filter in order to optimize the AMC. Finally, the residual noise component was employed as an adaptation parameter in the NC. It was shown from performance comparison using PESQ score and cepstral distortion that the GSC with the proposed AMC outperformed those with conventional AMCs.

6. ACKNOWLEDGEMENTS

This work was supported in part by the ISTD program, 10035252, funded by the MKE, Korea, by the Basic Research Project through a grant provided by GIST in 2011, and by the BSR program through the NRF of Korea funded by the MEST, 2011-0026201.

7. REFERENCES

[1] M. Brandstein and D. Ward, *Microphone Arrays*, Springer, Heidelberg, 2001.

[2] L. J. Griffiths and C. W. Jim, "An alternative approach for linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagation*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[3] O. Hoshuyama, B. Begasse, A. Sugiyama, and A. Hirano, "A real time robust adaptive microphone array controlled by an SNR estimate," in *Proc. of ICASSP*, pp. 3605-3608, May 1998.

[4] M. S. Choi, C. H. Baik, Y. C. Park, and H. G. Kang, "A soft-decision adaptation mode controller for an efficient frequency-domain generalized sidelobe canceller," in *Proc. of ICASSP*, pp. 893-896, Apr. 2007.

[5] S. M. Kim and H. K. Kim, "Hybrid probabilistic adaptation model controller for generalized sidelobe canceller-based target-directional speech enhancement," in *Proc. of ICASSP*, pp. 2532-2535, May 2011.

[6] A. Spriet, M. Moonen, and J. Wouters, "Spatially pre-processed speech distortion weighted multi-channel Wiener filtering for noise reduction," *Signal Processing*, vol. 84, no. 12, pp. 2367-2387, July 2004.

[7] I. Cohen, "Analysis of two-channel generalized sidelobe canceller (GSC) with post-filtering," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 684–699, Nov. 2004.

[8] S. Han, J. Hong, S. Jeong, and M. Hahn, "Probabilistic adaption mode control algorithm for GSC-based noise reduction," *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol. E93-A, no. 3, pp. 627-630, Mar.2010.

[9] Y. Jung, H, Kang, C. Lee, D. Youn, C. Choi, and J. Kim, "Adaptive microphone array system with two-stage adaptation mode controller," *IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences*, vol. E88-A, no. 4, pp.972-977, Apr. 2005.

[10] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1-3, 1999.

[11] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, pp. 1109-1121, Dec. 1984.

[12] J. Hong, S. Han, S. Jeong, and M. Hahn, "Adaptive microphone array processing for high-performance speech recognition in car environment," *IEEE Trans. Consumer Electronics*, vol. 57, no. 1, pp. 260-266, Feb. 2011.

[13] ITU-T Recommendation P.862, *Perceptual Evaluation of Speech Quality (PESQ), and Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Coders*, Feb. 2001.

[14] A. M. Peinado and J. C. Segura, *Speech Recognition Over Digital Channels: Robustness and Standards*, Wiley, Chichester, 2006.