LETTER

Pre-Filtering Algorithm for Dual-Microphone Generalized Sidelobe Canceller Using General Transfer Function

Jinsoo PARK[†], **Wooil KIM**[††], **David K. HAN**[†††], ***Nonmembers***, ***and*** **Hanseok KO**[†a)], ***Member***

SUMMARY We propose a new algorithm to suppress both stationary background noise and nonstationary directional interference noise in a speech enhancement system that employs the generalized sidelobe canceller. Our approach builds on advances in generalized sidelobe canceller design involving the transfer function ratio. Our system is composed of three stages. The first stage estimates the transfer function ratio on the acoustic path, from the nonstationary directional interference noise source to the microphones, and the powers of the stationary background noise components. Secondly, the estimated powers of the stationary background noise components are used to execute spectral subtraction with respect to input signals. Finally, the estimated transfer function ratio is used for speech enhancement on the primary channel, and an adaptive filter reduces the residual correlated noise components of the signal. These algorithmic improvements give consistently better performance than the transfer function generalized sidelobe canceller when input signal-to-noise ratio is 10 dB or lower.

key words: *speech enhancement, beamforming, adaptive signal processing, nonstationary noise*

1. Introduction

The generalized sidelobe canceller (GSC) has been an effective solution in many applications using two or more microphones to enhance speech quality in noisy environments [1]–[3]. Gannot *et al*. addressed this problem and derived a GSC solution based on the transfer function ratio (TFR) between sensors in response to a desired speech signal [4]–[6]. However, since the TFR estimation employs nonstationary noise characteristics of the desired speech signal under the assumption that the noise is stationary background noise (SBN) [7], the transfer function GSC (TFGSC) cannot be directly applied to nonstationary directional interference noise (NDIN). In real environments, for example, the desired speech signal may be interfered with by another person's voice or by TV sound in a moving vehicle or in an office where an air conditioner turns on. In that case, the TFR estimate of the NDIN can be obtained by exploiting its nonstationarity characteristics. To suppress both SBN and NDIN, this paper suggests a new pre-filtering algorithm using this additional TFR information along the acoustic paths from the NDIN source to the microphones.

Manuscript received February 6, 2014.
Manuscript revised May 2, 2014.

[†]The authors are with the School of Electrical Engineering, Korea University, Seoul, Korea.
[††]The author is with the School of Computer Science Engineering, Incheon National University, Incheon, Korea.
[†††]The author is with the Office of Naval Research, Arlington, VA, USA.
a) E-mail: hsko@korea.ac.kr

DOI: 10.1587/transinf.2014EDL8026

2. Review of Dual-Microphone GSC Using Transfer Function

$S(k,l)$ denote the desired speech signal and let $Z_i(k,l)$ and $N_{Si}(k,l)$ denote the observed signal and SBN components of the i-th microphone, respectively, in the time-frequency domain, where k is the index of the frequency bin and l is the frame number. Assuming that the NDIN $N_N(k,l)$ originated in a single source, the observed signals are given by

$$\begin{aligned}\mathbf{Z}(k,l) &= \begin{bmatrix} Z_1(k,l) \\ Z_2(k,l) \end{bmatrix} \\ &= \begin{bmatrix} A_1(k,l)S(k,l) + B_1(k,l)N_N(k,l) + N_{S1}(k,l) \\ A_2(k,l)S(k,l) + B_2(k,l)N_N(k,l) + N_{S2}(k,l) \end{bmatrix}, \end{aligned} \tag{1}$$

where A_i and B_i are the TFs of the acoustic path from the desired speech and NDIN source to the i-th microphone, respectively. The GSC solution is comprised of a fixed beamformer (FBF), a blocking matrix (BM) and an adaptive noise canceller (NC). Figure 1 is a schematic block diagram of the dual-microphone based GSC. The FBF forms a beam in the look direction, so that the desired speech signal is passed and all other signals are attenuated. The BM forms a null in the look direction so that the desired speech signal is suppressed and yields pure noise as the reference input to the NC. The NC generates a replica of the component correlated with the stationary noise in the primary signal and the enhanced speech signal is obtained by subtracting the replica from the primary signal [8]. Assuming that the acoustic path TFs are time invariant, the parameters H and R are defined as follows:

$$H(k) = A_2(k)/A_1(k), \quad R(k) = B_2(k)/B_1(k). \tag{2}$$

The primary and the reference signals are given by

$$\gamma(k,l) = \mathbf{W}^{\mathrm{H}}(k) + \mathbf{Z}(k,l) = A_1(k)S(k,l) + \frac{1}{1+|H(k)|^2} \cdot$$

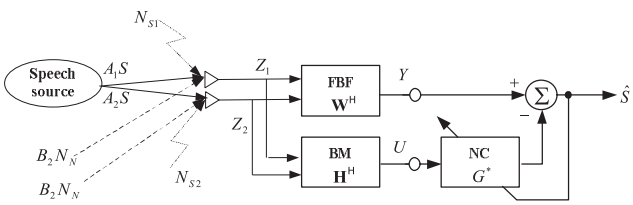

Fig. 1 Schematic block diagram of dual-microphone based GSC.

Copyright © 2014 The Institute of Electronics, Information and Communication Engineers

$$\left\{\left[1 + H^*(k)R(k)\right]B_1(k)N_N(k,l) + N_{S1}(k,l) + H^*(k)N_{S2}(k,l)\right\} \quad (3)$$

and

$$\begin{aligned} U(k,l) &= \mathbf{H}^{\mathrm{H}}(k)\mathbf{Z}(k,l) \\ &= [R(k) - H(k)]B_1(k)N_N(k,l) - H(k)N_{S1}(k,l) + N_{S2}(k,l) \end{aligned} \quad (4)$$

where

$$\begin{aligned} \mathbf{W}(k) &= \frac{1}{\left[1 + |H(k)|^2\right]} \cdot [1 \;\; H(k)]^T, \\ \mathbf{H}(k) &= [-H^*(k) \;\; 1] \end{aligned} \quad (5)$$

represent FBF and BM filters, respectively. The TFR H is estimated by using the nonstationary characteristics of the desired speech signal when the NDIN is absent [7]. The NC executes normalized least mean square (NLMS) algorithm to reduce the stationary noise component of the primary signal correlated to the reference signal. The NC filter G is updated only when there is no active signal and the update rule is

$$G(k, l+1) = G(k,l) + \mu \frac{U(k,l)\hat{S}(k,l)}{P_{est}(k,l)} \quad (6)$$

where P_{est} controls the adaptation term by using the power of the input sensor signals [4]. Consequently, the system output is

$$\hat{S}(k,l) = Y(k,l) - G^*(k,l)U(k,l). \quad (7)$$

The GSC algorithm may be an appropriate solution to suppress noises since the estimated blocking matrix provides a sharp null in the desired direction of the speech signal and attenuates the leakage signal efficiently in the reference input. In NLMS adaptation, P_{est} makes the system be free from considering voice activity detection (VAD) also. However, the algorithm has the limits that it is only applicable to the stationary noise environment because the adaptive filter does not have the ability to trace transient changes of the noise even though there is large correlation between noise signals. Moreover, the nonstationarity impedes convergence of the adaptive filter to the stationary case solution.

3. Proposed Pre-Filtering SBN and NDIN

A pre-filtering algorithm that suppresses both SBN and NDIN is proposed in order to overcome the disadvantages of GSC mentioned in Sect. 2. We focus on establishing a reasonable strategy using additional TFR information along the acoustic paths from, the NDIN source to the microphones. The TFR R is used as the core component of the blocking matrix for nonstationary directional interference noise (BMN) which blocks NDIN. The proposed algorithm can be separated in to the following three stages. Estimation of the TFR R and the BMN is executed at first stage. The BMN

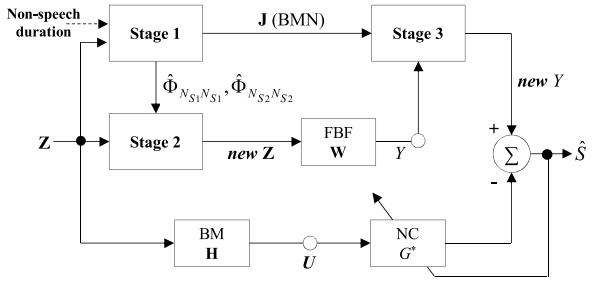

Fig. 2 Dual-microphone GSC incorporating proposed algorithm.

outputs only SBN components by blocking NDIN when desired speech signal is absent, and thus a simple linear transformation of the input and the BMN output can provide estimates of the SBN power. Secondly, suppression of the SBN is performed at the input by conventional spectral subtraction (SS) using the SBN power estimates. This stage produces a new input signal which is composed of desired speech and NDIN signals. The third stage provides the primary signal and the noise reference by propagating the new input through the FBF and the BMN respectively. This noise reference information helps to estimate the BMN which is used to suppress the NDIN in the primary signal. Finally, the NLMS based adaptive NC filter decreases residual correlated noise components from the enhanced primary signal. In fact, this sequential approach is nearly inevitable since the BMN can not be correctly estimated without elimination of the SBN at the input. Figure 2 describes how these processes are combined in the GSC.

3.1 Estimation of BMN and SBN Power

Both the TFR H and R are estimated by applying the nonstationarity based algorithm. The TFR H of the desired speech source is given in advance by using only the clean desired speech signal while the TFR R of the NDIN is continuously estimated when the desired speech signal is absent. We determined the absence of the desired speech signal by inspecting the ratio between the power of the GSC output signal which includes desired speech, and that of the noise reference signal U which is obtained by blocking matrix (BM) $\mathbf{H}$ from the TFR H. The power ratio value approaches ∞ in the presence of desired speech signal while it approach zero in the absence of desired speech signal. The parameter R is given by

$$R(k) = \frac{E_L\left[\hat{\Phi}_{Z_{N1}Z_{N1}}(k,l)\hat{\Phi}_{Z_{N2}Z_{N1}}(k,l)\right] - E_L\left[\hat{\Phi}_{Z_{N1}Z_{N1}}(k,l)\right]E_L\left[\hat{\Phi}_{Z_{N2}Z_{N1}}(k,l)\right]}{E_L\left[\hat{\Phi}^2_{Z_{N1}Z_{N1}}(k,l)\right] - E_L\left[\hat{\Phi}_{Z_{N1}Z_{N1}}(k,l)\right]^2} \quad (8)$$

where L is the number of frames within the analysis interval and $E_L[\cdot]$ refers to the time average over the frames in which S is zero. Thus, the input signals of microphones are given by

$$\mathbf{Z}_N(k,l) = \begin{bmatrix} Z_{N1}(k,l) \\ Z_{N2}(k,l) \end{bmatrix}$$

$$= \begin{bmatrix} B_1(k)N_N(k,l) + N_{S1}(k,l) \\ R(k,l)B_1(k)N_N(k,l) + N_{S2}(k,l) \end{bmatrix}. \quad (9)$$

In Eq. (8), and the instantaneous the cross-power spectral density (CPSD) between Z_{N1} and Z_{N2} is used as an estimate of the CPSD $\hat{\Phi}_{Z_{N1}Z_{N1}}(k,l)$ in the l–th frame. This can be a reasonable choice to decorrelate successive frames while keeping quasi-stationarity within frames because the correlation length of the nonstationary noise is generally shorter than that of the analysis window and each update term uses only non-overlapped frames (only an odd or even number of frames). The BMN $\mathbf{J}$ is estimated from the TFR R:

$$\mathbf{J}(k) = \begin{bmatrix} -R^*(k,l) & 1 \end{bmatrix}^T. \quad (10)$$

Let V denote the signal at the output of the BMN ($\mathbf{J}$), which is given by

$$V(k,l) = \mathbf{J}^{\mathrm{H}}(k)\mathbf{Z}_N(k,l) = -R(k)N_{S1}(k,l) + N_{S2}(k,l). \quad (11)$$

Assuming weak correlations between SBN components, it is possible to express the power spectral density (PSD) of the signals V and $Z_{N\mathrm{i}}$'s in terms of the SBN and NDIN components from (9) and (11),

$$\begin{aligned} \Phi_{VV}(k) &= |R(k)|^2\Phi_{N_{S1}N_{S1}}(k) + \Phi_{N_{S2}N_{S2}}(k) \\ \Phi_{Z_{N1}Z_{N1}}(k) &= |B_1(k)|^2\Phi_{N_NN_N}(k) + \Phi_{N_{S1}N_{S1}}(k) \\ \Phi_{Z_{N2}Z_{N2}}(k) &= |R(k)|^2|B_1(k)|^2\Phi_{N_NN_N}(k) + \Phi_{N_{S2}N_{S2}}(k). \end{aligned} \quad (12)$$

Hence the PSD of the SBN noise components in the input can be estimated by

$$\begin{bmatrix} \hat{\Phi}_{N_{S1}N_{S1}}(k) \\ \hat{\Phi}_{N_{S2}N_{S2}}(k) \end{bmatrix} = \begin{bmatrix} |R(k)|^2 & 1 \\ |R(k)|^2 & -1 \end{bmatrix}^{-1} \begin{bmatrix} \Phi_{VV}(k,l) \\ |R(k)|^2\Phi_{Z_{N1}Z_{N1}}(k,l) - \Phi_{Z_{N2}Z_{N2}}(k,l) \end{bmatrix}. \quad (13)$$

3.2 Suppression of SBN

With the estimated powers of the SBN in (13), the second stage executes SS on the input signals [9]. In the SS, the amount of subtraction (α) changes according to the frequency. The new input signals are given by

$$\begin{aligned} |newZ_i(k,l)|^2 &= \begin{cases} P_{SS_i}(k,l), & \text{if } P_{SS_i}(k,l) > \beta|Z_i(k,l)|^2 \\ \beta|Z_i(k,l)|^2, & \text{otherwise} \end{cases} \\ \angle newZ_i(k,l) &= \angle Z_i(k,l) \end{aligned} \quad (14)$$

where

$$\begin{aligned} P_{SS_i}(k,l) &= Z_i(k,l)Z_i^*(k,l) - \alpha(k)\hat{\Phi}_{N_{Si}N_{Si}}(k), \\ \alpha(k) &= 1 + \mathrm{A}\cdot\exp(-0.1\cdot k). \end{aligned} \quad (15)$$

To prevent musical noises after SS, α should be larger than 1 and generally selects 3 to 5. However, the over subtraction may give result in the reduction of the desired speech signals, especially in the high frequency domain, because the absolute amount of the power is very small. Equation (15) makes SS preserve the desired speech power in high frequency by lowering α to 1. Typically we use A = 4 and $\beta = 0.1$.

3.3 Enhancement of the Primary Signal

In applying the BMN (10) when speech is active, the desired speech signal can be kept without distortion, while eliminating the NDIN component, unless the estimation of the BMN is inaccurate.

$$Y'(k,l) = A_1(k)S(k,l) + \frac{-R(k,l)N_{S1}(k,l) + N_{S2}(k,l)}{H(k) - R(k,l)} \quad (16)$$

Though the NDIN component can theoretically be diminished, it does not always ensure the performance of the speech enhancement such as signal-to-noise ratio (SNR), because of the denominator of the noise components in (16). If the difference $|H - R|$ in the denominator is too small, then the stationary component may be amplified, and elimination of the NDIN may be meaningless even though much of the SBN is eliminated by the stage 2. Therefore, the enhancement of the primary signal is accomplished only when the power is decreased by applying the BMN, that is, when the total amount of noise is decreased. The output of the stage 3 process is

$$|newY(k,l)|^2 = \begin{cases} |Y(k,l)|^2 - a(k,l)(|Y(k,l)|^2 - |Y'(k,l)|^2), \\ \qquad \text{if } |Y'(k,l)|^2 < |Y(k,l)| \\ |Y(k,l)|^2, \qquad \text{otherwise} \end{cases} \quad (17)$$

where

$$\begin{aligned} \angle newY(k,l) &= \angle Y(k,l), \\ a(k,l) &= \log(|U(k,l)|^2/|Y(k,l)|^2). \end{aligned} \quad (18)$$

In (17), $a(k,l)$ controls noise suppression according to the power ratio of the reference signal to the original primary signal. The power ratio has the information about how much the primary signal may be contaminated by the NDIN signal. That is, the bigger $a(k,l)$ gives rise to greater noise power suppression. Finally, the NLMS based adaptive NC filter decreases residual correlated noise components from the enhanced primary signal.

4. Experimental Results

To evaluate the performance of the proposed algorithm, we used signals recorded at a rate of 8 kHz and 16 bits per a sample in a typical office room with dimensions of 4 m×3.5 m×2 m. Two microphones are located with 15 cm spacing in the middle. The desired speaker is located at 50 cm from the center of the arrays along the vertical line (90°) to the array's axis. The NDIN and SBN sources are located at 1 m, and 1.5 m from the array center, along 30°, and 60° lines, respectively. The SBN source is the furthest from the microphone array so that the assumption of weak correlation between SBN inputs is satisfied. The desired speech signal

Table 1 Performance comparison of speech enhancement methods.

Input SNR	NR (dB)		SNR (dB)		LSD (dB)	
(dB)	GSC	Proposed	GSC	Proposed	GSC	Proposed
0*	1.96	**4.75**	7.75	**8.92**	20.51	**18.50**
-5	4.47	**9.52**	1.21	**3.87**	29.07	**24.82**
0	4.47	**9.52**	5.76	**7.60**	22.73	**20.21**
5	4.47	**9.52**	9.83	**10.76**	17.91	**17.11**
10	4.47	**9.52**	13.00	**13.09**	**14.57**	15.07

(*does not include nonstationary interference)

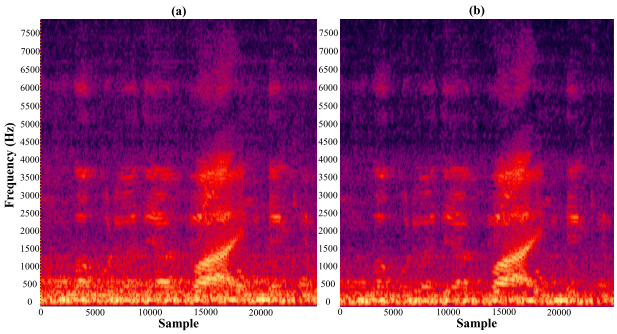

Fig. 3 Output spectrograms of baseline GSC (a) and proposed algorithm (b) with input SNR of 0 dB.

consists of male pronunciation of 5 digits from one to five in English and the NDIN signal is another male voice reading an arbitrary Korean sentence composed of seven continuous words without pause. The SBN source emits car engine noise. These signals were recorded separately and were mixed to generate input signals at various SNR levels ranging from −5 to +10 dB. The amount of the NDIN is approximately adjusted to be the same as that of the SBN object. The time-frequency analysis was performed with a Hamming window of 32 ms in length and a 256-point FFT was used for every 16 ms. The performance of the proposed algorithm was compared with that of the baseline GSC, with three object quality measures: noise reduction (NR), SNR and log-spectral distance (LSD), respectively [10].

Table 1 shows the improved performance of the proposed algorithm with respect to NR, SNR, and LSD measures, respectively. The performance evaluation is conducted in the frequency range of 250 Hz ∼ 8 KHz to ensure reliability of the evaluation over the entire frequency range, since slight changes of the SBN in the low frequency range can considerably affect performance due to concentrated energy to that range. The results demonstrate improved performance of the proposed algorithm. In every case, the proposed algorithm provides better NR, SNR and reduced LSD values except that LSD of 10 dB input is increased slightly. NR is a measure that is intended to compare the noise power in the enhanced output with the noise power recorded by the first microphone during desired speech signal is absent. Unlike SNR and LSD which contain speech components, NR is intended to compare noise power only. Therefore NR is not affected by the input SNR levels which were from −5 to +10 dB in the experiments.

Figure 3 compares the output of the proposed algorithm (b) with that of the baseline GSC (a). Desired speech is active in the range of [14000, 19000] samples. Though the difference is not clear, a decrease of the SBN over the range of 300 Hz and the diminished NDIN can be observed in the range of [6000, 12000] samples.

5. Conclusion

In this paper, a pre-filtering algorithm is proposed to suppress both SBN and NDIN by using a blocking matrix for NDIN and power estimates of the SBN components. It shows better performances compared to the baseline GSC in every case where SNRs under 10 dB are applied to the system. Although the algorithm alone does not show considerable improvement, it provides new approaches, using pre-filters to handle noise. It is also possible to improve the performance of the existing speech enhancement system by combining the pre-filtering algorithm with post-filters.

Acknowledgments

This research was supported by the Seoul R&BD (WR080951) Program.

References

[1] L.J. Griffiths and C.W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” IEEE Trans. Antennas Propag., vol.ASSP-30, no.1, pp.27–34, Jan. 1982.
[2] O. Hoshuyama, A. Sugiyama, and A. Hirano, “A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters,” IEEE Trans. Signal Process., vol.47, no.10, pp.2677–2684, Oct. 1999.
[3] W. Herbordt and W. Kellermann, “Analysis of blocking matrices for generalized sidelobe cancellers for non-stationary broadband signals,” IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.4, pp.IV-4187, May 2002.
[4] S. Gannot, D. Burshtein, and E. Weinstein, “Signal enhancement using beamforming and nonstationarity with applications to speech,” IEEE Trans. Signal Process., vol.49, pp.1614–1626, Aug. 2001.
[5] S. Gannot, D. Burshtein, and E. Weinstein, “Theoretical analysis of the general transfer function GSC,” Proc. Int. Workshop Acoustic Echo Noise Control, 2001, pp.103–106.
[6] K. Kim and H. Ko, “Robust relative transfer function estimation for dual microphone-based generalized sidelobe canceller,” IEICE Trans. Inf. & Syst., vol.E92-D, no.9, pp.1794–1797, Sept. 2009.
[7] O. Shalvi and E. Weinstein, “System identification using nonstationary signals,” IEEE Trans. Signal Process., vol.44, pp.2055–2063, Aug. 1996.
[8] O. Hoshuyama and A. Sugiyama, “Robust adaptive beamforming,” in Microphone Arrays, ed. M. Brandstein and D. Ward, pp.87–109, Springer, Berlin, 2001.
[9] S.V. Vaseghi, “Spectral subtraction” in Advanced Digital Signal Processing and Noise Reduction, second ed. ch. 11, pp.333–352, John Wiley & Sons, Chichester, 2002.
[10] I. Cohen, “Analysis of two-channel generalized sidelobe canceller (GSC) with post-filtering,” IEEE Trans. Speech Audio Process., vol.11, pp.684–699, Nov. 2003.