Two-Microphone Generalized Sidelobe Canceller with Post-Filter Based Speech Enhancement in Composite Noise

Jinsoo Park, Wooil Kim, David K. Han, and Hanseok Ko

This paper describes an algorithm to suppress composite noise in a two-microphone speech enhancement system for robust hands-free speech communication. The proposed algorithm has four stages. The first stage estimates the power spectral density of the residual stationary noise, which is based on the detection of nonstationary signal-dominant time-frequency bins (TFBs) at the generalized sidelobe canceller output. Second, speech-dominant TFBs are identified among the previously detected nonstationary signal-dominant TFBs, and power spectral densities of speech and residual nonstationary noise are estimated. In the final stage, the bin-wise output signal-to-noise ratio is obtained with these power estimates and a Wiener post-filter is constructed to attenuate the residual noise. Compared to the conventional beamforming and post-filter algorithms, the proposed speech enhancement algorithm shows significant performance improvement in terms of perceptual evaluation of speech quality.

Keywords: Two-microphone speech enhancement, nonstationary noise, generalized sidelobe canceller, spectral classification, beamforming, Wiener filter.

I. Introduction

Currently, there is an increasing trend for in-vehicle hands-free applications to be offered as standard in cars intended for Korean consumers. Such applications require hands-free speech communication; thus, among their many functions, the ability to perform high-quality speech communication is of the utmost importance. The present state of speech enhancement technology for speech communication permits a driver (the desired speech source) of a vehicle to control an in-vehicle hands-free application via voice commands. For example, a driver is able to select a song or make a phone call (from such an in-vehicle hands-free application) via a voice command. However, in real environments, a desired speech signal may be interfered with; for example, by the voice of a person other than the driver (from within the vehicle) or by the vehicle's air-conditioning system. In this situation, we assume that a background noise signal, such as that from the vehicle's engine or air-conditioning system, is *stationary*, and an interference noise signal, such as that from the voice of a person other than the driver, is *nonstationary*. That is, both the mean and power of a background noise signal are unchanging or very slowly changing over time, whereas those of the interference noise signal are rapidly changing.

To cope with acoustic noise, various signal processing techniques have been investigated over the past several decades. A single-channel technique is only able to exploit a received signal's spectral characteristics [1], [2]. A multi-microphone technique allows the additional use of spatial information [3], [4].

Manuscript received May 23, 2015; revised Nov. 18, 2015; accepted Dec. 9, 2015.

This work was supported by the Brain Korea 21Plus Project in 2015.

Jinsoo Park (jspark@ispl.korea.ac.kr) is with the Department of Biomicrosystem Technology, Korea University, Seoul, Rep. of Korea.

Wooil Kim (wikim@inu.ac.kr) is the School of Computer Science Engineering, Incheon National University, Rep. of Korea.

David K. Han (ctmkhan@gmail.com) is with the Office of Naval Research, Arlington, VA, USA.

Hanseok Ko (corresponding author, hsko@korea.ac.kr) is with the School of Electrical Engineering, Korea University, Seoul, Rep. of Korea.

http://dx.doi.org/10.4218/etrij.16.0115.0472

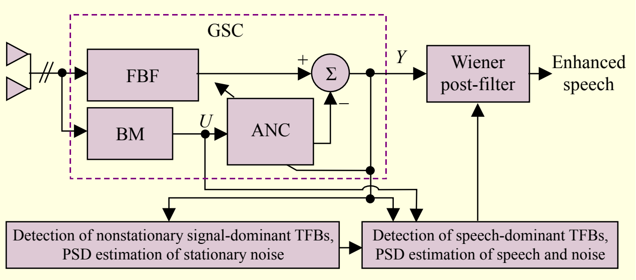
GSC
FBF
BM
ANC
U
Σ
+
−
Y
Wiener post-filter
Enhanced speech
Detection of nonstationary signal-dominant TFBs, PSD estimation of stationary noise
Detection of speech-dominant TFBs, PSD estimation of speech and noise

Fig. 1. Block diagram of proposed two-microphone speech enhancement.

A beamforming approach provides robust noise suppression performance by generating a distortionless beam toward a desired speech source while forming a null beam toward the interference noise source. Many beamforming techniques involve the generalized sidelobe canceller (GSC) algorithm of Griffiths and Jim [5]. As shown in Fig. 1, the GSC process is composed of three parts: a fixed beamformer (FBF) forms a beam in the *look* direction so that a desired speech signal is *passed* and all other signals are attenuated. A blocking matrix (BM) is applied to the input microphone signal to compute an estimate for a noise reference signal by blocking the components of a desired speech signal. The noise references at its output drive an adaptive noise canceller (ANC) whose coefficients are adapted to suppress the noise in the FBF output. However, the construction of the BM requires perfect knowledge of the direction of arrival of the desired speech signal for the time alignment of the microphone signals [6], [7]. Estimation errors in the direction of arrival and reflections of signals by objects and walls cause leakage of the desired speech signal into noise references, resulting in signal cancellation in the beamformer output.

This limitation was addressed by a transfer function–based GSC (TF-GSC) algorithm that specifies the FBF and BM by using relative transfer functions (RTFs) [8], [9]. It was also claimed that a BM containing RTFs from a source to individual microphones is sufficient to block a desired speech signal. Moreover, an FBF prevents the components of a desired speech signal from being distorted.

Recently, a pre-filtering algorithm for GSC (PF-GSC) has been developed to suppress both stationary and nonstationary noise by estimating RTFs along the acoustic paths from the interference noise to the microphones, and the power of the components of any background noise [10]. However, even if the RTFs are optimal, these algorithms always suffer from an inability to attenuate any residual noise, since reverberation in real environments is always lower than that of any coherence that exists within the residual noise. The RTFs alone do not produce a considerable improvement; thus, an additional post-filter is normally required to reduce any residual noise.

There have been several attempts to include post-filters in the noise suppression process to reduce any residual noise; for example, a general expression of a post-filter, based on an assumed knowledge of the complex coherence function of the noise field, thereby improving noise suppression performance by eliminating residual noise at the beamformer output [11], [12]. However, since the algorithms of [11] and [12] only depend on the stationariness of input noise signals, the post-filter degrades residual noise suppression performance in a nonstationary noise environment.

By assuming that a desired speech signal dominates over any interference and that W-disjoint orthogonality (WDO) exists between any speech and nonstationary noise [13], a time-frequency binary masking (TF-BM)–based method has the ability to distinguish time-frequency bins (TFBs) [14]. However, since a TF-BM proceeds by making some specific assumptions, it cannot be directly applied in a stationary noise environment. Consequently, conventional approaches such as beamforming and post-filtering have to apply a unique assumption with respect to noise characteristics. Therefore, it is difficult to handle composite noise that includes both stationary and nonstationary noise. To mitigate the difficulty in such a case, we propose the novel approach of combining a spectral classification–based Wiener post-filter within a GSC, which maintains the advantages of both the TF-BM algorithm and the GSC algorithm. The proposed algorithm introduces an effective power spectral density (PSD) estimation method of residual stationary noise, which is based on the detection of nonstationary signal-dominant TFBs (NTFBs) by using a spectral correlation between adjacent frequencies at the GSC output. Subsequently, speech-dominant TFBs are identified among the previously detected NTFBs by examining the power ratio between the GSC output and the BM output, and PSDs of speech and residual nonstationary noise are estimated. This idea is developed under a WDO assumption between any speech and nonstationary noise [13]. Finally, a bin-wise output signal-to-noise ratio (SNR) calculated from the PSD estimates is applied to a Wiener post-filter, which in turn attenuates any residual noise.

The remainder of this paper is organized as follows. Section II formulates the problem to be addressed. Section III provides a review of the power spectral classification model [13], from which the proposed post-filter is based. Section IV presents the proposed algorithm. Section V describes relevant experimental results along with discussions to validate the advantages of the proposed post-filter. Finally, conclusions are given in Section VI.

II. Problem Formulation

For application in mobile devices, a multiple-microphone noise suppression approach is required to be computationally efficient and easily installable. Generally, the noise suppression performance of such an approach increases as the number of microphones used increases. However, the ideal of having a large number of input microphones is difficult to implement in practice in mobile devices, due to the high computational burden associated with such an approach. Consequently, two-microphone speech enhancement is often the upper limit for practical approaches to mobile hands-free speech communication.

Let us consider a two-microphone array in a composite noisy environment. Under the assumption that the background noise signal is stationary and the interference noise signal is nonstationary, an observed input signal, $z_i(t)$, can be formulated by

$$z_i(t) = h_i(t) * s_0(t) + n_i(t) + a_i(t) \quad \text{for } i = 1, 2, \tag{1}$$

where $s_0(t)$ is the desired speech signal, $n(t)$ is the interference noise component coming from a fixed point source, $a(t)$ is a background noise component, and $h(t)$ is the transfer function from the source of the desired speech signal to one of the microphones; i is an index value indicating which of the two microphones is to be referred to. In a short-time Fourier transform (STFT) domain, (1) can be expressed as

$$Z_i(\tau, k) = H_i(k) S_0(\tau, k) + N_i(\tau, k) + A_i(\tau, k) \quad \text{for } i = 1, 2, \tag{2}$$

where τ and k are TFB indices.

In general, a beamformer such as a GSC is designed to reduce interference noise. However, such noise cannot be completely eliminated when a beam is roughly steered or the adaptation is not fast enough to track noise variation. Moreover, such a beamformer is not designed to handle background noise; hence, it normally suffers from residual noise at the output, $Y(\tau, k)$.

$$Y(\tau, k) = S(\tau, k) + \tilde{N}(\tau, k) + \tilde{A}(\tau, k), \tag{3}$$

where $S(\tau, k)$ is a desired speech signal distorted only by the first transfer function, $H_1(k)$, and $\tilde{\bullet}$ denotes the corresponding residual noise. The aim of noise reduction at this stage is to obtain $\hat{S}(\tau, k)$, which is an estimate of $S(\tau, k)$ for transforming it back to the time domain. A post-filter, $G(\tau, k)$, is developed to obtain $\hat{S}(\tau, k)$ as follows:

$$\hat{S}(\tau, k) = G(\tau, k) Y(\tau, k). \tag{4}$$

The optimal frequency-domain filter $G(\tau, k)$ should attenuate the residual noise components $\tilde{N}(\tau, k)$ and $\tilde{A}(\tau, k)$ while preserving $S(\tau, k)$. This can be accomplished by employing a Wiener filter [15], which is formulated by

$$G(k) = \frac{\Phi_{\mathrm{SY}}(k)}{\Phi_{\mathrm{YY}}(k)}, \tag{5}$$

where $\Phi_{\mathrm{YY}}(k)$ and $\Phi_{\mathrm{SY}}(k)$ denote, respectively, the auto-PSD of $Y(k)$ and the cross-PSD between $Y(k)$ and $S(k)$. However, due to the nonstationarity of speech signal $S(k)$ and $Y(k)$, the above Wiener filter should be formed in every frame as follows:

$$G(\tau, k) = \frac{\Phi_{\mathrm{SY}}(\tau, k)}{\Phi_{\mathrm{YY}}(\tau, k)}. \tag{6}$$

The challenge here is to estimate those spectral densities that produce a frame-based Wiener filter in short periods of Fourier analysis. Since $\Phi_{\mathrm{SY}}(\tau, k)$ and $\Phi_{\mathrm{YY}}(\tau, k)$ are unattainable in a given situation, we explore the method of estimating $\Phi_{SY}(\tau, k)$ and $\Phi_{\mathrm{YY}}(\tau, k)$ sufficient to enable $G(\tau, k)$ to suppress the residual noise while preserving the desired speech signal.

III. Spectral Classification

Assuming that each of the signal components in (3) is mutually uncorrelated, the power spectrum of the beamformer output in each of the TFBs is given by

$$|Y(\tau, k)|^2 = |S(\tau, k)|^2 + |\tilde{N}(\tau, k)|^2 + |\tilde{A}(\tau, k)|^2 . \tag{7}$$

If the interference noise is nonstationary, such as in a human voice or music, and the background noise is stationary, then the TFBs can be classified into one of the following two cases:

- Case 1: nonstationary signal–dominant TFBs, where $|S(\tau, k)|^2 + |\tilde{N}(\tau, k)|^2 \geq |\tilde{A}(\tau, k)|^2$.
- Case 2: stationary noise–dominant TFBs, otherwise.

Then, $\tilde{N}(\tau, k)$ has a sparse time-frequency representation; thus, the WDO condition can be applied to both the desired speech signal and the nonstationary noise, which supports the following mutually disjoint expression in the time-frequency domain [13]:

$$S(\tau, k) \tilde{N}(\tau, k) = 0 \quad \text{for } \forall (\tau, k). \tag{8}$$

Equation (8) essentially states that the energy of the interference noise is approximately zero when the desired speech signal is dominant at a TFB. That is, a TFB of Case 1 does not include both S and $\tilde{N}$ components at the same time [13]. Furthermore, if most of the energy of S or $\tilde{N}$ is concentrated in several narrow bands satisfying the condition of Case 1, then a TFB of Case 2 has a very small number of nonstationary components. Eventually, the power spectrum of the beamformer output, $|Y(\tau, k)|^2$, can be classified as follows:

$$|Y(\tau, k)|^2 \sim \begin{cases} |S(\tau, k)|^2 + |\tilde{A}(\tau, k)|^2 & \text{where } (\tau, k) \in \text{Class } 1, \\ |\tilde{N}(\tau, k)|^2 + |\tilde{A}(\tau, k)|^2 & \text{where } (\tau, k) \in \text{Class } 2, \\ |\tilde{A}(\tau, k)|^2 & \text{where } (\tau, k) \in \text{Class } 3, \end{cases} \tag{9}$$

where Class 1, Class 2, and Class 3 denote the speech-dominant, nonstationary noise–dominant, and stationary noise–dominant TFB groups, respectively. This expression makes it possible to directly estimate the power of each output component with a simple TFB classification. In the next section, a reliable spectral classification method is introduced to construct an optimal Wiener post-filter based on the output power model of (9).

IV. Proposed Algorithm

The conventional beamforming and post-filtering algorithms have some problems reducing both stationary background noise and nonstationary interference noise. Beamforming algorithms such as GSC, TF-GSC, and PF-GSC cannot avoid suffering from residual noise. Moreover, the nonstationary characteristics of the residual noise cause a conventional post-filter to become ineffective by disturbing the estimation of the stationary noise power.

This section proposes algorithmic improvements that suppress both the stationary and nonstationary noise signals. The proposed speech enhancement algorithm adds four stages to the conventional two-microphone GSC structure. We initially estimate residual stationary noise by detecting NTFBs at the GSC output. Subsequently, speech-dominant TFBs are identified among the previously detected NTFBs, and PSDs of speech and residual nonstationary noise are then estimated. Finally, a bin-wise output SNR calculated from the PSD estimates is applied to a Wiener post-filter, which then attenuates any residual noise. Figure 1 shows a schematic diagram of the proposed algorithm.

1. Detection of NTFBs and PSD Estimation of Stationary Noise

Generally, nonstationary signals, such as the human voice or music, have a significant spectral correlation between adjacent frequencies. Also, since the nonstationary signals show strong correlations, the cross-PSD of the GSC output is significantly larger in NTFBs than in stationary noise–only TFBs. Hence, we define a detection statistic, D, for the nonstationary signal component of the GSC output [16] as

$$D(\tau,k)=\sum_{m=1}^{M}\left[\left|Y(\tau,k)Y^{*}(\tau,k+m)\right|+\left|Y(\tau,k)Y^{*}(\tau,k-m)\right|\right], \tag{10}$$

where M is the index of the frequency bin near k. The two terms in the expression "[| |]" represent the cross-periodogram between $Y(\tau, k)$ and $Y(\tau, k-m)$, and $Y(\tau, k)$ and $Y(\tau, k+m)$, respectively. The cross-periodogram represents an instantaneous estimate of the spectral correlation function, which is evaluated in the frequency range from $k-M$ to $k+M$. By comparing the detection statistic with the threshold value of ξ_D, a reliable indicator function, I_{T}, for nonstationary signal–dominant bins can be given as

$$I_T(\tau,k)=\begin{cases}1 & \text{if } D(\tau,k)>c\cdot\xi_D(\tau,k),\\ 0 & \text{otherwise},\end{cases} \tag{11}$$

where ξ_D is the minimum statistics of D obtained with the method of [17], and c (>1) is a pre-determined margin to reduce false alarms. Then, the PSD of the output stationary noise, $\Phi_{\tilde{A}\tilde{A}}$, is easily estimated by recursively averaging the complementary set of the detected bins as follows:

$$\begin{aligned}\hat{\Phi}_{\tilde{A}\tilde{A}}(\tau,k)=\hat{\Phi}_{\tilde{A}\tilde{A}}(\tau-1,k)+(1-\alpha)&\\ \times\bar{I}_{\mathrm{T}}(\tau,k)\times\left[\left|Y(\tau,k)\right|^{2}-\hat{\Phi}_{\tilde{A}\tilde{A}}(\tau-1,k)\right]&,\end{aligned} \tag{12}$$

where α ($0<\alpha<1$) is a smoothing parameter and $\bar{I}$ denotes a negation indicator of I. Equation (12) calculates the $\hat{\Phi}_{\tilde{A}\tilde{A}}$ at all TFBs where the nonstationary signal component is not dominant; thus, it provides a robust estimate even if the nonstationary signal is activated.

2. Detection of Speech-Dominant TFBs

To discriminate speech-dominant bins among the previously detected bins, we introduce another indicator function, $\mathrm{I_S}$, which uses the power ratio between the GSC output, Y, and the BM output, U, as shown in the following expression:

$$I_{\mathrm{S}}(\tau,k)=\begin{cases}1 & \text{if } I_{\mathrm{T}}(\tau,k)\cdot R(\tau,k)>\zeta(\tau,k),\\ 0 & \text{otherwise},\end{cases} \tag{13}$$

where $R(\tau,k)=|Y(\tau,k)|^2/|U(\tau,k)|^2$. In the GSC structure, the BM blocks the desired speech signal and produces a reference noise signal, $U(\tau, k)$, for the ANC input [9]. The threshold ζ is given by its recursive averaging over TFBs of Class 3, as follows:

$$\begin{aligned}\zeta(\tau,k)=\zeta(\tau-1,k)&\\ +(1-\alpha)\cdot\bar{I}_T(\tau,k)\cdot\left[R(\tau,k)-\zeta(\tau-1,k)\right]&.\end{aligned} \tag{14}$$

Let $U_S(\tau, k)$, $U_N(\tau, k)$, and $U_A(\tau, k)$ denote speech leakage, nonstationary noise, and stationary noise components at the BM output, respectively. By categorizing the power model of $U(\tau, k)$ into three classes as in (9), the power ratio R is given by

$$R\sim\begin{cases}(|S|^2+|\tilde{A}|^2)/(|U_S|^2+|U_A|^2) & \text{at Class } 1,\\ (|\tilde{N}|^2+|\tilde{A}|^2)/(|U_N|^2+|U_A|^2) & \text{at Class } 2,\\ |\tilde{A}|^2/|U_A|^2 & \text{at Class } 3,\end{cases} \tag{15}$$

where (τ, k) is omitted for simplicity. When the BM is appropriately designed, the reference noise includes a small amount of the speech leakage, while the GSC outputs enhance speech at Class 1. Moreover, it is assured that the interference signal is more attenuated than the background noise by the GSC at Class 2. Consequently, the three classes have the following relations:

$$\frac{|S(\tau,k)|^2}{|U_S(\tau,k)|^2} > \frac{|\tilde{A}(\tau,k)|^2}{|U_A(\tau,k)|^2} > \frac{|\tilde{N}(\tau,k)|^2}{|U_N(\tau,k)|^2}. \tag{16}$$

Therefore, (17) can be easily derived from (15) and (16). Then, $R(\tau, k)$ values have the following relations according to the TFB class of (9):

$$R(\tau,k)|_{\text{Class 1}} > R(\tau,k)|_{\text{Class 3}} > R(\tau,k)|_{\text{Class 2}}\ . \tag{17}$$

3. PSD Estimation of Speech and Interference Noise

Previously (Sections IV-1 and IV-2), the method to obtain nonstationary signal–dominant bins and speech-dominant bins as well as to estimate the PSD of stationary noise was introduced. These may be valuable information when estimating the PSD of a nonstationary signal. By eliminating only the power of the stationary noise from the power of nonstationary signal–dominant bins, it is possible to accurately estimate the PSD of a nonstationary signal.

In particular, in the case where nonstationary signals include both speech and interference noise, it is necessary to estimate both speech and interference noise, as in the proposed method of this section.

By means of a speech indicator function, speech-dominant bins are selected. Then, the PSD of the desired speech signal is estimated by subtracting the power of the stationary noise from the power of these bins. The PSD of interference noise is obtained similarly by those bins that are not indicated by the speech indicator function. A spectral subtraction (SS) algorithm [18] essentially conducts subtraction of stationary noise to avoid the effect of any musical noise that is caused by the stationary noise itself.

We propose a method of estimating $\Phi_{\text{SS}}(\tau, k)$ based on a speech indicator function (see (13)). The proposed estimation of $\Phi_{\text{SS}}(\tau, k)$ is as follows:

1) Spectral cues resembling the PSD of speech at (τ, k) are obtained with the speech indicator function $I_{\text{S}}(\tau, k)$ based on SS.

$$\hat{\Phi}'_{\text{SS}}(\tau,k) = \max\left\{ I_{\text{S}}(\tau,k)\cdot\left[|Y(\tau,k)|^2 - \hat{\Phi}_{\tilde{\text{A}}\tilde{\text{A}}}(\tau,k)\right],\ \beta\,\hat{\Phi}_{\tilde{\text{A}}\tilde{\text{A}}}(\tau,k)\right\}. \tag{18}$$

2) Obtained spectral cues are smoothed via a 9th-order Hamming window along the frequency bin at the τth frame. As a result, we adopt a smoothed cue spectrum as the estimate of the PSD of the desired speech signal.

$$\hat{\Phi}_{\text{SS}}(\tau,k) = \sum_{m=k-4}^{k+4} h_{\text{w}}(m-k)\hat{\Phi}'_{\text{SS}}(\tau,m), \tag{19}$$

where $h_{\text{w}}(m)$ denotes a 9th-order non-causal Hamming window.

The proposed estimation of $\Phi_{\tilde{\text{N}}\tilde{\text{N}}}(\tau,k)$ is obtained via the same way of (18) and (19) as follows:

1) Spectral cues resembling the PSD of the interference noise at (τ, k) are obtained with the negation of the speech indicator function $I_{\text{S}}(\tau, k)$; that is, $\overline{I}_{\text{S}}(\tau,k)$. Hence,

$$\hat{\Phi}'_{\tilde{\text{N}}\tilde{\text{N}}}(\tau,k) = \max\left\{ \overline{I}_{\text{S}}(\tau,k)\cdot\left[|Y(\tau,k)|^2 - \hat{\Phi}_{\tilde{\text{A}}\tilde{\text{A}}}(\tau,k)\right],\ \beta\,\hat{\Phi}_{\tilde{\text{A}}\tilde{\text{A}}}(\tau,k)\right\}. \tag{20}$$

2) Obtained spectral cues of the interference noise are smoothed in the same way as in (19), as follows:

$$\hat{\Phi}_{\tilde{\text{N}}\tilde{\text{N}}}(\tau,k) = \sum_{m=k-4}^{k+4} h_{\text{w}}(m-k)\hat{\Phi}'_{\tilde{\text{N}}\tilde{\text{N}}}(\tau,m)\,. \tag{21}$$

Ideally, $\Phi'_{\text{SS}}(\tau,k)$ and $\Phi'_{\tilde{\text{N}}\tilde{\text{N}}}(\tau,k)$, respectively, become only the PSD estimate of the desired speech signal and that of the interference noise, as shown in (18) and (20). However, these powers cannot avoid including approximation errors due to SS. A nonlinear mapping of the spectral estimate produces spectral discontinuity in the STFT domain. To correct this problem, in (19) and (21), the PSDs of the speech and of the interference noise are smoothed in the frequency domain, thus mitigating any spectral discontinuity at the band edge of the nonstationary signal.

4. Construction of Wiener Post-Filter

By using the assumption of non-correlation between speech and noise signal, the numerator of the Wiener filter (see (6)), $\Phi_{\text{SY}}(\tau, k)$, is expanded as

$$\begin{aligned}\Phi_{\text{SY}}(\tau,k) &= \text{E}\left[S(\tau,k)\cdot Y^*(\tau,k)\right]\\ &= \text{E}\left[S(\tau,k)\cdot \text{gsc}\{Z_1(\tau,k), Z_2(\tau,k)\}^*\right]\\ &= \text{E}\left[S(\tau,k)\cdot S^*(\tau,k)\right]\\ &= \Phi_{\text{SS}}(\tau,k),\end{aligned} \tag{22}$$

where gsc$\{Z_1(\tau, k), Z_2(\tau, k)\}$ represents the GSC process with input signals $Z_1(\tau, k)$ and $Z_2(\tau, k)$. Thus, an estimation of $\Phi_{\text{SY}}(\tau, k)$ yields an estimation of $\Phi_{\text{SS}}(\tau, k)$.

Now, consider the estimation of $\Phi_{\text{YY}}(\tau, k)$; that is, the denominator of the Wiener filter (see (6)), which is expanded as

$$\begin{aligned}\Phi_{\text{YY}}(\tau,k) &= \text{E}\left[Y(\tau,k)\cdot Y^*(\tau,k)\right]\\ &= \Phi_{\text{SS}}(\tau,k) + \Phi_{\tilde{\text{N}}\tilde{\text{N}}}(\tau,k) + \Phi_{\tilde{\text{A}}\tilde{\text{A}}}(\tau,k).\end{aligned} \tag{23}$$

In (23), $\Phi_{\text{SS}}(\tau, k)$, $\Phi_{\tilde{\text{N}}\tilde{\text{N}}}(\tau,k)$, and $\Phi_{\tilde{\text{A}}\tilde{\text{A}}}(\tau,k)$ are already estimated by (19), (21), and (12), respectively.

Finally, the estimated PSDs are used to construct an optimal Wiener post-filter, H_{post}, as follows:

$$H_{\text{post}}(\tau,k) = \frac{\hat{\Phi}_{\text{SS}}(\tau,k)}{\hat{\Phi}_{\text{SS}}(\tau,k) + \hat{\Phi}_{\tilde{\text{N}}\tilde{\text{N}}}(\tau,k) + \hat{\Phi}_{\tilde{\text{A}}\tilde{\text{A}}}(\tau,k)}. \tag{24}$$

Then, the enhanced output signal in the STFT domain is given by

$$\hat{S}(\tau,k) = H_{\text{post}}(\tau,k)Y(\tau,k). \tag{25}$$

The enhanced output (25) is transformed back to the time-domain via an inverse fast Fourier transform.

V. Experimental Setup and Results

1. Speech Database

The proposed algorithm was evaluated in stationary and nonstationary noise environments. To evaluate the performance of the proposed algorithm, we used signals recorded at a sampling rate of 16 kHz in a typical vehicle space with dimensions 2 m × 3 m × 1.5 m. The reverberation time of the room is about 250 ms. Two microphones were located 15 cm apart in the middle of the enclosure. The desired speaker was located at a distance of 50 cm in the forward direction (90°) from the microphone array. The nonstationary interference noise source and the stationary background noise source are located at 80 cm and 150 cm from the array center, along 30° and 120° lines, respectively. In this configuration, sketched in Fig. 2, the stationary background noise source is far from the microphone array so that the assumption of weak correlation between stationary noise inputs is satisfied.

The desired speech source was that of a male voice pronouncing the digits one through five in English. The nonstationary interference noise source was that of a different male voice reading an arbitrary Korean sentence composed of seven continuous words without pause. A stationary background noise source was added comprising the noise of a vehicle's engine (performing speeds of between 70 km/h and 90 km/h on a road surface) and wind noise emitting from the air-conditioner inside the vehicle. These signals were recorded separately and were mixed to generate input signals at various SNR levels ranging from −5 dB to 20 dB. A time-frequency analysis was performed with a Hamming window of 32 ms in length; a 256-point fast Fourier transform (FFT) was used for every 16 ms.

2. Results of Speech Enhancement

First, the proposed algorithm is separately evaluated for stationary- and nonstationary-noise environments. Then, the proposed algorithm is evaluated for a composite environment including both noise environments. The evaluation results are listed in Tables 1 and 2, respectively. The performance of the proposed algorithm is compared with that of the conventional PF-GSC [10]; TF-BM [14]; and TF-GSC with post-filtering by McCowan (TP-MC) [11] using a state-of-the-art speech quality evaluation standard: Perceptual Evaluation of Speech Quality (PESQ) scores [19]. PESQ is a psychoacoustics-based objective measure and has better correlation with subjective tests than other alternative objective measures. It was selected by the International Telecommunication Union Telecommunication standardization sector as recommendation P.862 to evaluate the speech quality of a test signal. In computation of the PESQ score for an enhanced speech signal (or a noisy input signal), the clean and enhanced speech signals are initially level-equalized to a standard listening level and then filtered by a filter with response similar to a standard telephone handset. The clean and enhanced speech signals are

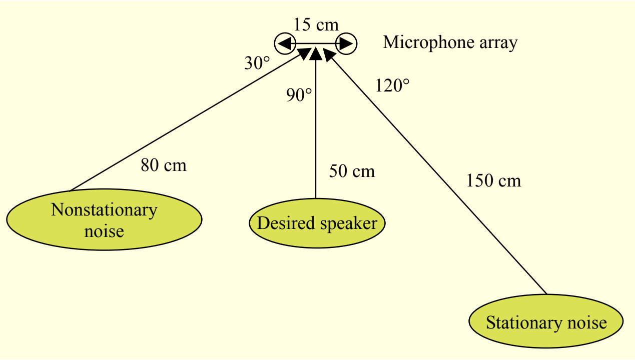

Fig. 2. Experimental setting for microphones and data recording.

Table 1. PESQ scores for proposed and conventional algorithms in nonstationary- and stationary-noise environments.

Environment	Nonstationary noise			Stationary noise		
Input SNR	–5 dB	7.5 dB	20 dB	–5 dB	7.5 dB	20 dB
Input	1.60	2.57	3.40	2.11	3.04	3.82
PF-GSC	2.13	3.04	3.75	2.78	3.59	4.05
TF-BM	2.33	3.33	3.85	2.60	3.50	4.02
TP-MC	2.22	3.10	3.74	**3.16**	3.80	3.74
Proposed	**2.57**	**3.52**	**3.99**	3.14	**3.95**	**4.11**

Table 2. PESQ scores for proposed and conventional algorithms in composite noise environment.

Environment	Composite noise					
Input SNR	–5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
Input	1.47	1.87	2.24	2.60	2.95	3.27
PF-GSC	1.84	2.23	2.58	2.94	3.25	3.62
TF-BM	1.94	2.44	2.81	3.18	3.42	3.84
TP-MC	2.23	2.57	2.86	3.14	3.43	3.63
Proposed	**2.34**	**2.60**	**2.94**	**3.26**	**3.60**	**3.90**

aligned in the time domain to correct the time delays between these two signals. Hence, these two signals are processed through an auditory transform to obtain the loudness spectra. The disturbance, obtained by computing the difference between the loudness spectra for both the clean and the enhanced speech signals, is computed and averaged over time and frequency to produce a predicted subjective mean opinion score. PESQ scores range from –0.5 (for the worst case) to 4.5 (for the best case) — higher values represent better quality.

Tables 1 and 2 demonstrate the improved performance of the proposed algorithm with respect to the PESQ measure. The proposed algorithm shows improved PESQ scores of 27.17% from the PF-GSC, 20.62% from the TF-BM, and 4.93% from the TP-MC for an input SNR of −5 dB in a composite noise environment. The PF-GSC alone has the worst performance among the considered algorithms in a composite noise environment, but particularly in an environment where there is a high level of nonstationary noise. In this case, we observe greater residual nonstationary noise at the output of the PF-GSC. The TP-MC shows better PESQ results only in an environment where there is a high level of stationary noise, but also shows considerable performance degradation in PESQ in an environment where there is a high level of nonstationary noise. On the other hand, the TF-BM has a high PESQ in an environment where there is a high level of nonstationary noise, but also shows considerable performance degradation in PESQ in an environment where there is a high level of stationary noise. This is due to the fact that the post-filter depends on the stationarity of the input noise signals. The proposed speech enhancement algorithm demonstrates a robust performance by controlling stationary and nonstationary noises effectively at most of the considered input SNRs. The experimental results imply that the proposed algorithm is effective at both types of noise while preserving the desired speech signal.

Figure 3 presents example waveforms and their corresponding spectrograms at an input SNR of –5 dB. It shows that the PF-GSC alone did not support equivalent performance improvements in a composite noise environment. The McCowan's post-filter supported the TF-GSC quite well by attenuating stationary noise under some environments. However, noise estimates became contaminated by the residual nonstationary noise during non-speech intervals and TP-MC did not correctly reduce the nonstationary noise components. The TF-BM shows a comparable noise reduction (see Fig. 3(e)), yet it also suppressed the components of the desired speech signal that propagated to distort the output signal during those portions of the signal that included desired speech. In contrast, the proposed algorithm shows a superior capability to control noise in the overall frequency range while preserving the desired speech signal.

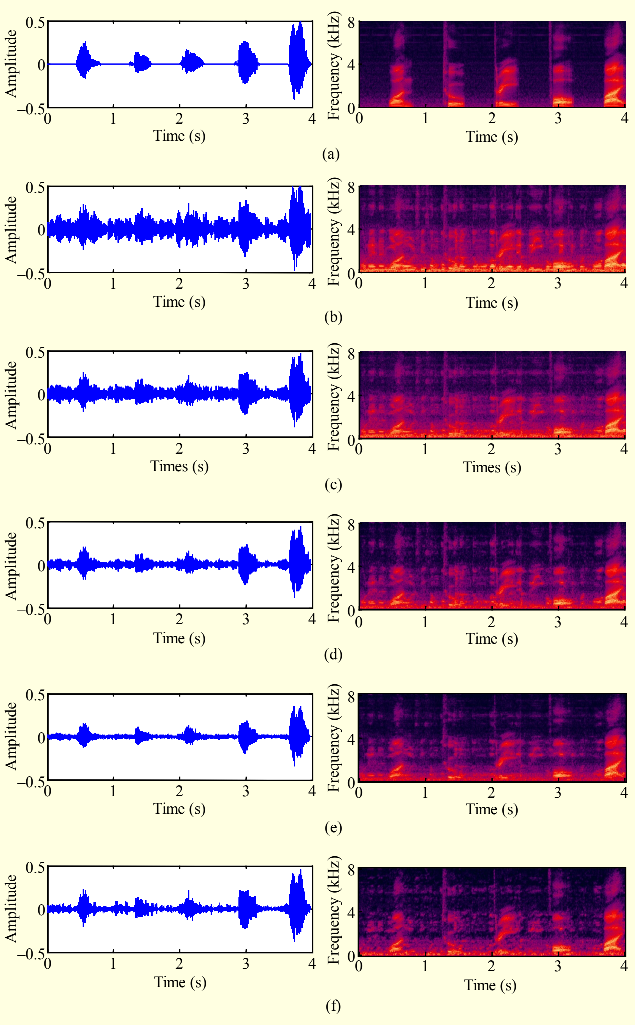

Fig. 3. Experimental speech waveforms and spectrogram: (a) original clean speech signal at first microphone, (b) noisy input signal at first microphone, (c) output of PF-GSC, (d) output of TP-MC, (e) output of TF-BM, and (f) output of proposed algorithm.

3. Evaluation of Spatial Scenarios

In this section, performance of the proposed speech enhancement algorithm and conventional algorithms is evaluated based on a speech database that was acquired under various experimental conditions. In practical applications of hands-free speech communication, the location of the driver (desired speech signal) as well as another person's relative location may affect the performance. In addition, depending on product specifications such as the distance between the microphones, the performance may also vary. Therefore, to assure robustness of the proposed method in practical vehicular situations, four representative spatial scenarios were defined based on various vehicular environments.

As listed in Table 3, the configuration given in Fig. 2 was

Table 3. List of spatial scenarios according to speech- and noise-source locations, and distance between microphones.

Scenarios	DM (cm)	DNM (°)			DSM (cm)		
		DS	NN	SN	DS	NN	SN
Scenario 1	5	90	30	120	50	80	150
Scenario 2	10	90	30	150	80	150	80
Scenario 3	15	90	60	120	150	80	50
Scenario 4	20	90	60	150	150	50	80

Table 4. Comparison of PESQ scores averaged over high or low SNRs according to proposed and conventional algorithms for four spatial scenarios.

Scenarios	Input SNR	Input	PF-GSC	TF-BM	TP-MC	Proposed
Scenario 1	High	2.96	3.21	3.38	3.39	3.53
	Low	1.90	2.19	2.35	2.50	2.59
Average		2.43	2.7	2.87	2.95	**3.06**
Scenario 2	High	2.95	3.24	3.41	3.42	3.56
	Low	1.88	2.22	2.36	2.55	2.62
Average		2.42	2.73	2.89	2.99	**3.09**
Scenario 3	High	2.93	3.29	3.48	3.41	3.60
	Low	1.86	2.22	2.40	2.57	2.65
Average		2.40	2.76	2.94	2.99	**3.13**
Scenario 4	High	2.93	3.29	3.49	3.41	3.61
	Low	1.86	2.23	2.40	2.59	2.66
Average		2.40	2.76	2.95	3.00	**3.13**

expanded to consider the influence of geometry between the microphone array, source of desired speech (DS), source of nonstationary noise (NN), and source of stationary noise (SN). Specifically, the representative spatial scenarios were defined according to various parameters such as (a) distance between microphones (DM), (b) direction between noise sources and microphone array (DNM), and (c) distance between each source and microphone array (DSM). In addition to the configuration, the experimental setup was created in the same way as described in Section V-1.

Table 4 compares the PESQ scores of the desired speech signals averaged over high or low SNRs according to the four spatial scenarios. In the experiments, a high SNR contained 20 dB, 15 dB, and 10 dB signals, while a low SNR contained 5 dB, 0 dB, and −5 dB signals. It was shown from the table that the desired speech signals with the proposed Wiener post-filter still attained the best performance among the considered algorithms under every scenario. This result demonstrates that the PESQ scores did not seem to be dependent on the spatial scenario. As we have already mentioned, detected NTFBs can estimate the PSD of noise well when both stationary and nonstationary noise exist in an input signal. Moreover, the extracted noise components need to be smoothed before determining the final PSD of nonstationary noise. A smoothing factor is helpful for minimizing the performance degradation caused by the spectral discontinuity. The results indicate that the proposed algorithm is very effective in real environments.

VI. Conclusion

This paper described an effective technique to suppress composite noise of nonstationary and stationary types by combining a spectral classification–based Wiener post-filter with a two-microphone GSC. The proposed Wiener post-filter algorithm estimated the PSDs of a desired speech signal and residual nonstationary and stationary noises by detecting the target signal–dominant TFBs.

By providing the waveform and spectrogram of the desired speech signal, we showed that the desired speech signal was similar to that of the original signal. Experimental results confirmed the effectiveness of the proposed algorithm in attenuating both types of noise simultaneously while minimizing speech distortion. Conducted representative speech-enhancement experiments confirmed that the proposed algorithm was superior to the conventional algorithms by an average of 8.13% in terms of PESQ for all the considered environments.

As a future work, it is possible to improve the performance of the existing speech enhancement system by combining the proposed post-filtering algorithm with pre-filtering.

References

[1] R. Martin, "Statistical Methods for the Enhancement of Noisy Speech," *Int. Workshop Acoust. Echo Noise Contr.*, Kyoto, Japan, Sept. 8–11, 2003, pp. 1–6.

[2] J. Beh and H. Ko, "Spectral Subtraction Using Spectral Harmonics for Robust Speech Recognition in Car Environments," *Lecture Notes Comput. Sci.*, vol. 2660, June 2003, pp. 1109–1116.

[3] J. Benesty, J. Chen, and Y. Huang, "*Microphone Array Signal Processing*," Berlin, Germany: Springer-Verlag, 2008, pp. 1–222.

[4] J. Beh, R.H. Baran, and H. Ko, "Dual Channel Based Speech Enhancement Using Novelty Filter for Robust Speech Recognition in Automobile Environment," *IEEE Trans. Consum. Electron.*, vol. 52, no. 2, May 2006, pp. 583–589.

[5] L.J. Griffiths and C.W. Jim, "An Alternative Approach to Linearly Constrained Adaptive Beamforming," *IEEE Trans. Antennas Propag.*, vol. 30, no. 1, Jan. 1982, pp. 27–34.

[6] 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, Oct. 1999, pp. 2677–2684.

[7] W. Herbordt and W. Kellermann, "Analysis of Blocking Matrices for Generalized Sidelobe Cancellers for Non-stationary Broadband Signals," *IEEE Int. Conf. Acoust. Speech Signal Process.*, Orlando, FL, USA, May 13–17, 2002, pp. IV–4187.

[8] S. Gannot, D. Burshtein, and E. Weinstein, "Signal Enhancement Using Beamforming and Nonstationarity with Applications to Speech," *IEEE Trans. Signal Process.*, vol. 49, no. 8, Aug. 2001, pp. 1614–1626.

[9] S. Gannot, D. Burshtein, and E. Weinstein, "Theoretical Analysis of the General Transfer Function GSC," *Int. Workshop Acoust. Echo Noise Contr.*, Darmstadt, Germany, Sept. 10–13, 2001, pp. 103–106.

[10] J. Park et al., "Pre-filtering Algorithm for Dual-Microphone Generalized Sidelobe Canceller Using General Transfer Function," *IEICE Trans. Inf. Syst.*, vol. E97–D, no. 9, Sept. 2014, pp. 2533–2566.

[11] I.A. McCowan and H. Bourlard, "Microphone Array Post-Filter Based on Noise Field Coherence," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, Nov. 2003, pp. 709–716.

[12] H. Yoon and H. Ko, "Microphone Array Post-Filter Using Input Output Ratio of Beamformer Noise Power Spectrum," *Electron. Lett.*, vol. 43, no. 18, Aug. 2007, pp. 1003–1005.

[13] O. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, July 2004, pp. 1830–1847.

[14] S. Jeong, S. Lee, and M. Hahn, "Dual Microphone-Based Speech Enhancement by Spectral Classification and Wiener Filtering," *Electron. Lett.*, vol. 44, no. 3, July 2008, pp. 253–254.

[15] S.V. Vaseghi, "Wiener Filters," in *Advanced Digital Signal Processing and Noise Reduction*, Chichester, UK: John Wiley & Sons Ltd., 2002, pp. 178–202.

[16] K.W. Baugh and K.R. Hardwicke, "On the Detection of Transient Signals Using Spectral Correlation," *Circuits Syst. Signal Process.*, vol. 13, no. 4, Dec. 1994, pp. 467–479.

[17] R. Martin, "Spectral Subtraction Based on Minimum Statistics," *Proc. European Signal Process. Conf.*, Edinburgh, UK, Sept. 13–16, 1994, pp. 1182–1185.

[18] S.V. Vaseghi, "Spectral Subtraction," in *Advanced Digital Signal Processing and Noise Reduction*, Chichester, UK: John Wiley & Sons Ltd., 2002, pp. 333–352.

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

Jinsoo Park received his BS degree in electrical engineering from Kyung Hee University, Suwon, Rep. of Korea, in 2008. Since 2008, he has been participating in an integrated MS and PhD course at Korea University, Seoul, Rep. of Korea. He received his PhD degree in electrical engineering from Korea University, in 2016. His research interests include voice activity detection; acoustic echo cancellation; sound-source localization and tracking; and multi-microphone-based noise reduction.

Wooil Kim received his BS, MS, and PhD degrees in electronics engineering from Korea University, Seoul, Rep. of Korea, in 1996, 1998, and 2003, respectively. He has been an assistant professor with the School of Computer Science and Engineering, Incheon National University, Rep. of Korea, since 2012. Previously, he was a research assistant professor and a research associate at the Erik Jonsson school of Engineering and Computer Science, University of Texas at Dallas, Richardson, TX, USA. His research interests include human–computer interfaces, automatic speech recognition, human behavior signal processing, multimedia information retrieval, and pattern recognition.

David K. Han received his BS degree in electrical engineering from Carnegie Mellon University, Pittsburgh, PA, USA and his MSE and PhD degrees in electrical engineering from Johns Hopkins University, Baltimore, MD, USA. He served as a naval nuclear engineer at Pearl Harbor Naval Shipyard, Honolulu, HI, USA, from 1981 to 1987. From 1987 to 1995, he was with the Naval Surface Warfare Center at White Oak, MD, USA, as a research engineer in the underwater weapons program. In 1995, he became a program officer for the Office of Naval Research (ONR), Arlington, VA, USA, directing research programs in mine countermeasure technologies. In 1998, he joined Johns Hopkins University's applied physics laboratory as a senior member of the professional staff. In 2005, he joined the University of Maryland, College Park, USA, as both a visiting associate professor and the deputy director of the Center for Energetic Concepts Development. He returned to ONR in 2009, serving as a program officer in the Ocean Engineering and Marine Systems Team. He was appointed as the deputy director of research of ONR in 2012. His professional interests include signal processing for pattern recognition.

http://dx.doi.org/10.4218/etrij.16.0115.0472

Hanseok Ko received his BS degree in electrical engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 1982; his MS degree in electrical engineering from Johns Hopkins University, Baltimore, MD, USA, in 1988; and his PhD degree in electrical engineering from the Catholic University of America, Washington, DC, USA, in 1992. From the onset of his career, he was with the White Oak Laboratory, MD, USA, where his work involved signal and image processing. In March of 1995, he joined the faculty of the School of Electrical Engineering, Korea University, Seoul, Rep. of Korea, where he is currently working as a professor. He is the president of the Korea Acoustical Society and a fellow of the Institution of Engineering and Technology. His professional interests include signal processing for pattern recognition, multi-modal analysis, and intelligent data fusion.