# Chapter 12
# Dual-Channel Speech Enhancement Using a Perceptual Filterbank for Hands-Free Communication

**Jongsung Yoon, Kihyeon Kim, Jounghoon Beh, Robert H. Baran, and Hanseok Ko**

**Abstract** We investigate a dual-channel speech enhancement method using perceptual adaptive noise suppressor, which improves perceptual quality of speech in automobile environment for hands-free communication. In particular, the perceptual adaptive noise suppressor, which is composed of a Mel-based perceptual filterbank, an adaptive filter, and a speech modification block, estimates the envelope of the desired speech by suppressing the nonspeech components. Experiments indicate that the proposed scheme shows 8.06 dB of NR improvement and 0.70 of PESQ score improvement compared to the Transfer Function Generalized Sidelobe Canceller structure alone.

**Keywords** Driver assistance • Dual-channel speech enhancement • Hands-free communication • In-vehicle speech technology

## 12.1 Introduction

Recently, the significance of multi-microphone-based speech enhancement has increased as the needs of hands-free communication systems grow, especially in in-vehicle situations. In this chapter, an efficient multichannel speech enhancement algorithm is presented, which improves the speech quality while minimizing the directional interference and ambient noise.

J. Yoon • K. Kim • R.H. Baran • H. Ko (✉)
Department of Electronics and Computer Engineering, Korea University, 5Ka-1 Anam-dong, Seongbuk-Gu, Seoul 136713, South Korea
e-mail: hsko@korea.ac.kr

J. Beh
Institute for Advanced Computer Studies, University of Maryland, College Park, MD, USA

J.H.L. Hansen et al. (eds.), *Digital Signal Processing for In-Vehicle Systems and Safety*,
DOI 10.1007/978-1-4419-9607-7_12, 

The conventional beamforming methods, such as the linearly constrained minimum variance (LCMV) [1] and the generalized sidelobe canceller (GSC), can reduce interference from undesired directions by exploiting the correlation among the noise signals of different sensors [2]. However, the beamformer cannot avoid suffering from high computational burden when the adaptive filter must be long enough to effectively suppress the noise. Hence, this aspect is not favorable for the system to be embedded on vehicular communication devices.

To solve this problem, we propose a novel algorithm which is based on spectral magnitude modification using the structure of the generalized sidelobe canceller. The envisioned algorithm applies an auditory filterbank on the primary signal, output of the fixed beamformer, and the noise reference signal, output of the blocking matrix, in order to estimate the spectral samples of noise components. Then, these samples are fed to the gain filter for spectral modification so that the optimal spectral envelope of the desired signal can be obtained. This structure provides unique advantages over traditional beamforming methods including improvement of the perceptual quality of speech, robustness against the stationary ambient noise, and high computational efficiency. We develop the envisioned algorithm on the basis of a dual-microphone array structure. In order to obtain the improved performance, we consider the optimal combination using conventional adaptive noise cancellation which is executed in general short-time Fourier transform domain.

## 12.2 Dual-Channel Speech Enhancement

### *12.2.1 Transfer Function Generalized Sidelobe Canceller (TFGSC)*

The basic GSC structure is composed of a fixed beamformer (FBF), a blocking matrix (BM), and a noise canceller filter (NC). The FBF forms a beam in the look direction so that the acoustic signal from the desired speaker is passed while interfering noises are suppressed. Then, the BM blocks the desired signal and produces a noise reference signal. The NC generates a replica of the component which is included in the FBF output and is correlated with the interference. An enhanced speech signal is obtained by subtracting the replica from the output of the FBF. Conventionally, these processes are often described in terms of sampled data representation. The broadband GSC expression, which is based on general transfer functions (TF) of room impulse responses (RIR), has recently been introduced [3]. Compared with the simple attenuation-and-delay assumption on RIRs, the TF-based BM forms a sharp null in the look direction so that the leakage signal of desired speech is more favorably attenuated. Ideally, the BM would convey a pure noise reference input to the NC. Moreover, use of TFs in the FBF provides the ability to keep the desired signal free from distortion in a highly reverberant room condition. Gannot et al. developed this concept based on the transfer function ratio (TFR) and constructed an adaptive GSC, so-called TFGSC [2]. In Fig. 12.1,

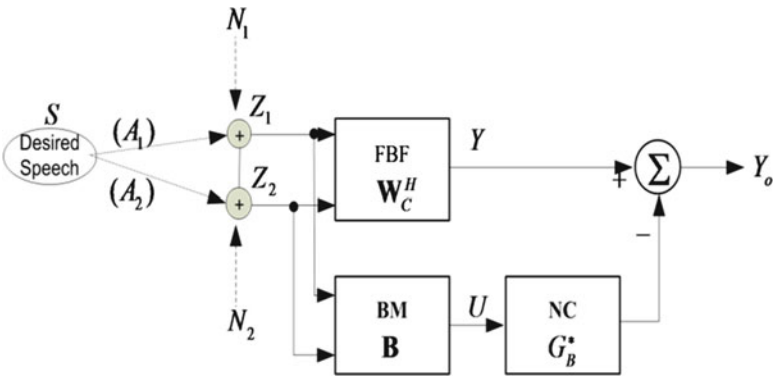


**Fig. 12.1** Schematic diagram of TFGSC

a schematic diagram of the dual-channel TFGSC is shown with the signal propagation model in the frequency domain.

The transfer function ratio $H$ is defined by

$$H = \frac{A_2}{A_1}. \tag{12.1}$$

Through the FBF, primary signal is given by

$$Y = \mathbf{W}_C^H \mathbf{Z} = \frac{1}{1+|H|^2}\begin{bmatrix}1 & H^*\end{bmatrix}\begin{bmatrix}Z_1\\ Z_2\end{bmatrix} = A_1 S + \frac{1}{1+|H|^2}[N_1 + H^* N_2]. \tag{12.2}$$

The FBF forms a beam in the look direction to pass speech and outputs a signal consisting of the distortionless speech and noise components including both the directional interference and the in-vehicle ambient noise. Next, BM forms a null beam to block speech and produces the noise reference signal:

$$U = \mathbf{B}^H \mathbf{Z} = \begin{bmatrix}-H & 1\end{bmatrix}\begin{bmatrix}Z_1\\ Z_2\end{bmatrix} = -HN_1 + N_2. \tag{12.3}$$

The noise reference signal generated goes to an NC block, and it constructs a filter, $\hat{G}_B^*$, to estimate and eliminate the noise component in FBF output via the general wiener filter solution as [4]

$$\hat{G}_B^* = \frac{E[UY_c]^H}{E[UU^H]} = \frac{\Phi_{UY}^*}{\Phi_{UU}} \tag{12.4}$$

$$Y_o = Y - \hat{G}_B^* U. \tag{12.5}$$

Normalized least mean squares (NLMS) algorithm is implemented for adaptive noise canceller [5, 6]:

$$\hat{G}_B(k,t+1) = \hat{G}_B(k,t) + \mu \frac{U(k,t)Y_o(k,t)}{P_{est}(k,t)}, \tag{12.6}$$

in which the time–frequency index returns to describe the update in short-time Fourier transform domain. In (12.6), the adaptation term is controlled by the power estimate of the input sensor signals:

$$P_{est}(k,t) = \alpha P(k,t-1) + (1-\alpha)\sum_{i=1}^{2} |Z_i|^2, \tag{12.7}$$

where $\alpha$is a forgetting factor. Then, the resulting system output is given by

$$Y(k,l) = Y_C(k,l) - \hat{G}_B^*(k,l)U(k,l). \tag{12.8}$$

A high computational burden occurs in the TFGSC when the number of adaptive filter coefficients is large enough to cover the signal path in a reverberant chamber. A save/add method is applied to perform a linear convolution using FFT. It necessitates a computationally efficient adaptive noise suppression filter while keeping the advantage of TFGSC.

### 12.2.2 *Perceptually Adaptive Noise Suppressor (PANS) Based on TFGSC*

The structure of the PANS based on TFGSC is shown in Fig. 12.2. The PANS is composed of three blocks: a fixed beamformer (FBF), a blocking matrix (BM), and a perceptually adaptive noise suppressor (PANS). It is used to estimate the spectral envelope (SE) of the desired speech signal. As shown in Fig. 12.3, an auditory filterbank such as the Mel-filterbank or the equivalent rectangular bandwidth characterizes the PANS [7, 8]. The filterbank is composed of band-pass filters imaging the effect of auditory masking. Accordingly, specific frequency resolution of a human auditory system is provided. As shown in Fig. 12.2, the filterbank outputs the auditory SE of the primary signal $\tilde{Y}$ and that of the reference noise $\tilde{U}$. Then, an adaptive filter estimates the noise SE $\tilde{N}$ of the primary signal with the input $\tilde{U}$. Given the estimate $\hat{\tilde{N}}$, the spectral modification is executed to obtain the desired speech as

$$\hat{S} = F_{itp}\left(\left[1 - \alpha\hat{\tilde{\xi}}\right]^{0.5}\right)Y, \tag{12.9}$$

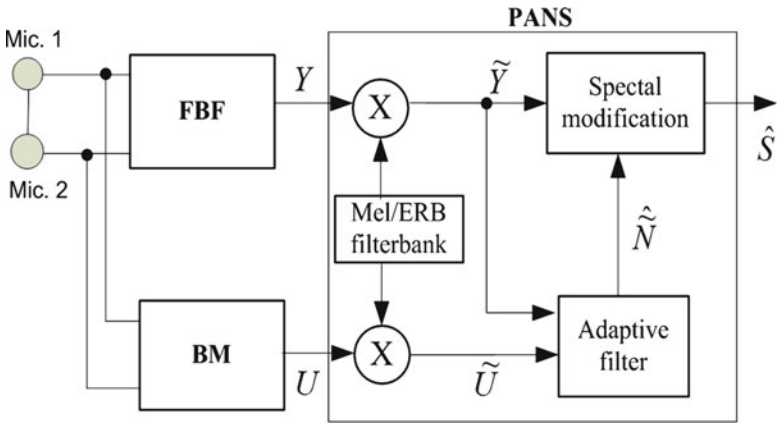


**Fig. 12.2** Schematic diagram of PANS based on TFGSC

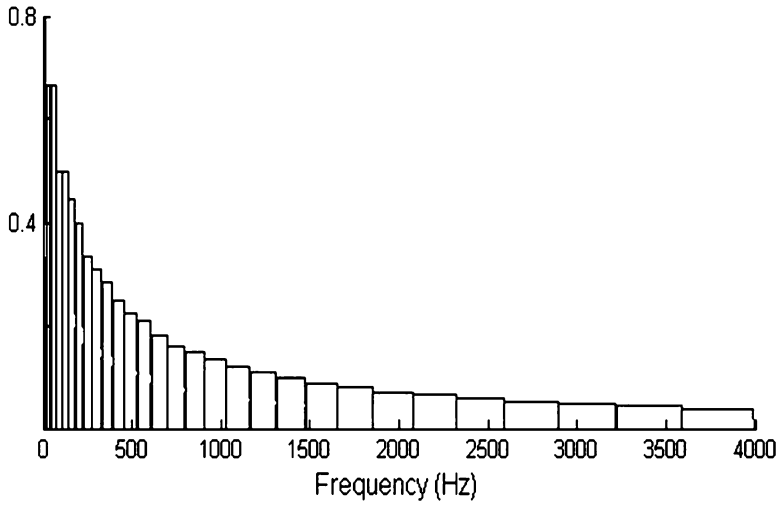


**Fig. 12.3** Frequency response of an ERB filterbank [10]

where $\alpha$ is a parameter to control the noise suppression, and the power ratio $\hat{\tilde{\xi}}$ is defined by $\hat{\tilde{N}}/\tilde{Y}$. Since the SE samples only appear at center frequencies of the filterbank, the function $F_{itp}$ is used to interpolate the power ratio samples $\hat{\tilde{\xi}}$ in the frequency domain. With the auditory filterbank and spectral modification, the proposed structure has the improved perceptual quality of an enhanced speech while minimizing the number of coefficients in the adaptive filter. Moreover, the system also promises to have the robustness against the in-vehicle ambient noise. This is based on the fact that the adaptive filter provides an SE estimate including overall noise components.

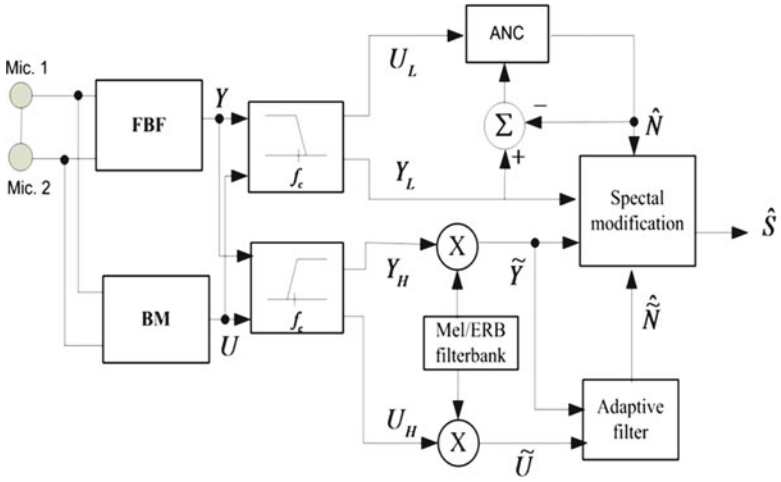


**Fig. 12.4** Schematic diagram of combination of the PANS and ANC

However, this approach cannot avoid speech distortion due to the interpolation process and the adaptive power estimation without using any phase information. Speech distortion becomes notable, especially, in a low frequency range where the energy of the speech is concentrated.

To overcome this degradation, a combination of PANS and the conventional adaptive noise canceller (ANC) is also considered. In a low frequency range, the ANC filter is applied to produce an accurate power estimate of the directional interference. Then, the spectral modification uses the noise estimate in order to enhance the speech without distortion. In a high frequency range, however, the PANS still applies the spectral modification with the auditory SE.

### *12.2.3 Combination of the PANS with Adaptive Noise Canceller (ANC)*

The structure of the PANS based on TFGSC is shown in Fig. 12.4. At low frequency range, noise is estimated by a conventional ANC filter. At high frequency range, the noise is estimated by the PANS filter. Enhanced speech is obtained by spectral modification. Since the energy of voiced speech is concentrated in low frequency range, in order to prevent the speech distortion, the spectral power of directional interference is estimated for each frequency bin rather than using filterbank. The PANS is applied to high frequency range so that the perceptual quality of speech is preserved and it also saves the computational load compared to the conventional ANC approach as well.

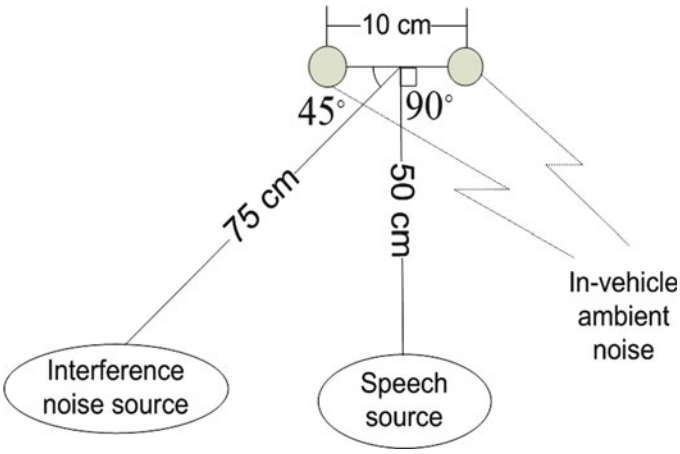


**Fig. 12.5** Microphone array aperture and location of signal sources for the RIR measurement

## 12.3 Experiments

To generate dual-channel speech signal and nonstationary interference signals, room impulse responses (RIRs) were measured in a vehicular chamber which has a reverberation time, $T_{60} = 250$ ms. The desired speech source was modeled to be located 50 cm from the microphone array along the broadside direction (90°) and the nonstationary interference source to be 75 cm along the 45° line. The array was located in front of the speech source with a 10-cm aperture. Figure 12.5 describes the experimental setup for signal generation.

Each RIR was convoluted with a single-channel clean speech signal to produce a dual-channel speech signal, and with an interfering human voice for a dual-channel nonstationary interference noise at a sampling rate of 8 kHz. Brownian noise was added as the in-vehicle ambient noise. Next, the interference plus the ambient noise was combined with the speech signal to simulate various signals with interference and noise ratios (SINR) ranging from −5 to 20 dB. The speech signal in experiments was formed from Korean digit strings and a nonstationary interference noise generated by using arbitrary Korean words.

To evaluate the performance of the noise suppression and the perceptibility of the enhanced speech signal, the noise reduction (NR) in log-domain and the perceptual evaluation of the speech (PESQ) are used as measures [9], respectively. Table 12.1 shows the performance of the proposed dual-channel speech enhancement system, where "PANS" and "PANS+ANC" denote the usage of PANS only and PANS with the ANC to estimate the desired spectral envelopes, respectively. The findings of the proposed algorithms is compared with the conventional transfer function–based GSC (TFGSC) method [2].

As shown in Table 12.1, the proposed PANS and PANS with ANC show superior performance over that of the TFGSC. Although PANS shows similar speech quality with TFGSC in adverse noise environment, this problem is solved by combining it with an ANC.

**Table 12.1** Experimental results of proposed algorithm

| Input SINR (dB) | | −5 | 0 | 5 | 10 | Avg |
|---|---|---|---|---|---|---|
| NR (dB) | TFGSC | 13.74 | 13.74 | 13.73 | 13.70 | 13.73 |
| | PANS | 22.45 | 22.43 | 22.23 | 21.72 | 22.21 |
| | ANC+PANS | 22.00 | 21.94 | 21.79 | 21.43 | 21.79 |
| PESQ | TFGSC | 2.02 | 2.44 | 2.64 | 3.02 | 2.53 |
| | PANS | 2.05 | 2.51 | 3.04 | 3.45 | 2.76 |
| | ANC+PANS | 2.76 | 3.12 | 3.41 | 3.62 | 3.23 |

## 12.4 Conclusions

We have proposed a dual-channel speech enhancement method using perceptual adaptive noise suppressor, which has improved the perceptual quality of speech for hands-free communication inside the auto chamber. The proposed method used an auditory filterbank based adaptive filter to estimate the noise SE and combined it with the ANC. This method resulted in reduced number of adaptive filter coefficients and has improved perceptual quality of speech. The usage of auditory SE was demonstrated to ensure robust noise suppression in the presence of in-vehicle ambient noise without additional postprocessing as demonstrated by the experimental results.

## References

1. Frost OL III (1972) An algorithm for linearly constrained adaptive array processing. Proc IEEE 60:926–935
2. Gannot S, Burshtein D, Weinstein E (2001) Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Trans Signal process 49(8):1614–1626
3. Brandstein M, Ward D (2001) Microphone arrays, Signal processing techniques and applications. New York, Springer
4. Meyer J, Simmer K U (1997) Multichannel speech enhancement in a car environment using Wiener filtering and spectral subtraction. In: Proceedings of ICASSP, IEEE Computer Society Washington DC, 1997, pp 1167–1170
5. Widrow B, Stearns S (1985) Adaptive signal processing. Prentice Hall, Englewood Cliffs, N.J.
6. Haykin S (2002) Adaptive filter theory, 4th edn. Prentice Hall, Upper Saddle River, N.J.
7. Faller C, Chen J (2005) Suppressing acoustic echo in a spectral envelope space. IEEE Trans Speech Audio Process 13(5):1048–1062
8. Wallin F, Faller C (2004) Perceptual quality of hybrid echo canceler/suppressor. ICASSP 4:157–160
9. Rix AW, Beerends JG, Hollier MP, Hekstra AP (2001) Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech coders. ITU-T Recommendation, pp 862, Feb 2001
10. Malcolm Slaney. Auditory toolbox, version 2. Technical Report #1998-010, Interval Research Corporation, 1998