Acoust. Sci. & Tech. **27**, 4 (2006)

PAPER

Noise reduction method based on generalized subtractive beamformer

Junfeng Li* and Masato Akagi†

School of Information Science, Japan Advanced Institute of Science and Technology, 1–1 Asahidai, Nomi, 923–1292 Japan

(*Received 21 July 2005, Accepted for publication 21 November 2005*)

Abstract: Combating noise signals is still a very important and challenging research topic. Previously, we presented a subtractive-beamformer-based noise reduction algorithm using paired microphones, which was shown to be effective in reducing directional noise. However, its basic assumption, namely a perfectly coherent noise field, is generally not satisfied in real-world environments. In this paper, we develop a general expression for the original algorithm we suggested earlier, on the basis of a generalized subtractive beamformer and by relaxing the strict assumption to that of an arbitrary noise field. Following ideas similar to those of the original algorithm, the generalized algorithm with a *generalized sidelobe canceller* (GSC)-like structure is derived. A theoretical analysis is then presented to show the linkage between this generalized algorithm and the original algorithm, and to show its noise reduction performance in theoretically defined noise fields. Finally, the superiority of the proposed noise reduction algorithm to other comparable algorithms was verified by experiments using multichannel recordings.

Keywords: Noise reduction, Generalized subtractive beamformer, Coherence function

PACS number: 43.72.Ew [doi:10.1250/ast.27.206]

1. INTRODUCTION

The problem of dealing with noise signals in hands-free telephones and teleconferences has been studied for several decades, and yet is still a challenge for researchers. In comparison with single-channel algorithms, multichannel algorithms have demonstrated a substantial superiority in reducing noise owing to their spatial filtering capability to suppress interfering signals arriving from directions other than the specified look direction [1]. Therefore, multichannel algorithms, e.g., beamformer-based algorithms, have attracted great research interest in recent years.

Various beamforming algorithms have been proposed [1–8]. The conventional beamformer, referred to as the *delay-and-sum beamformer* (DSBF), which enhances the desired speech signal by summing the in-phase microphone signals after the array is electronically steered in the look direction, was extensively studied. However, many microphones are needed to obtain an acceptable performance in real-world environments [1]. The linearly constrained adaptive beamformer, first presented by Frost, keeps the signals arriving from the desired look direction distortionless while suppressing signals from other directions by minimizing the power of the beamformer output [2]. A *generalized sidelobe canceller* (GSC) beamformer, first presented by Griffiths and Jim as an alternative implementation structure of the Frost beamformer, has also been extensively researched [3]. Recently, Gannot *et al.* [4] have extended the GSC beamformer to a *transfer function generalized sidelobe canceller* (TF-GSC) beamformer by considering transfer functions that relate the speech source and microphones. In the Frost and GSC beamformers, adaptive signal processing is generally used to prevent the cancellation of the desired speech signal [2–4]. The frequency-domain LMS algorithm and its two-dimensional extension were introduced and applied to the GSC beamformer to accelerate the convergence rate of adaptive beamformers [5,6]. However, adaptive signal processing systems still do not show a sufficiently high convergence rate and a high stability in practical environments.

To overcome the above drawbacks, a subtractive-beamformer-based noise reduction algorithm has recently been proposed by Akagi *et al.* [7,8]. In this algorithm, noises were analytically estimated based on the arrival time difference between paired microphones, instead of exploiting adaptive signal processing. Speech spectra are then enhanced by subtracting the estimated noise spectra from the observed noisy spectra. The superiority of this algorithm to other algorithms lies in its high ability to suppress

*e-mail: junfeng@jaist.ac.jp
†e-mail: akagi@jaist.ac.jp

directional noise, especially sudden noise, using only a small number of microphones. The main problem associated with this algorithm is the assumption that only directional noise sources exist in an environment, corresponding to a perfectly coherent noise field. A practical noise condition is generally not a coherent noise field, e.g., in a car or a reverberant room that can be approximately modelled as a diffuse noise field [9]. Therefore, the performance degradation of the noise reduction algorithm we previously presented is expected in these environments.

In this paper, we propose a noise reduction method based on a generalized subtractive beamformer under the assumption of an arbitrary noise field and on ideas similar to those of the original algorithm. This proposed method, which has a GSC-like structure, includes the algorithm we previously suggested [7,8] as a special case in a coherent noise field when only two microphones are available. The proposed algorithm have some advantages over traditional algorithms: exploiting no adaptive signal processing techniques (e.g., LMS); performing well under all noise conditions owing to the assumption of an arbitrary noise field; and offering an improved noise reduction ability since much spatial information is considered. The performance of the proposed method is then analyzed using coherence functions in theoretically defined noise fields. The superiority of the proposed method to other methods is further confirmed by experiments using real-world multichannel recordings.

2. REVIEW OF THE ORIGINAL METHOD

Considering an array with paired microphones in a noisy environment, the observed signals $x_1(t)$ and $x_2(t)$ on paired microphones are composed of two components: the desired speech signal $s(t)$ and the additive directional noise $n(t)$ arriving from a determinable direction. Thus, the observed signals on paired microphones can be represented as

$$x_1(t) = s(t) + n(t), \tag{1}$$

$$x_2(t) = s(t) + n(t - \delta), \tag{2}$$

where δ is the relative time delay between the paired microphones for a directional noise signal.

Using this signal model, the original noise reduction algorithm is accomplished in three steps, summarized as follows

(1) *Noise spectrum estimation*. To estimate the spectrum of directional noise, a subtractive beamformer is constructed using the signals $x_1(t)$ and $x_2(t)$ received by paired microphones. Two observed signals are first shifted $\pm\tau$ in the time domain, where τ is a certain constant ($\tau \neq 0$). Then, the subtractive beamformer output in the time domain $u_{12}(t)$ is defined as [7]

$$u_{12}(t) = \frac{1}{4}\big\{[x_1(t+\tau) - x_1(t-\tau)] - [x_2(t+\tau) - x_2(t-\tau)]\big\}. \tag{3}$$

Performing the *short-time Fourier transform* (STFT), in the frequency domain, we obtain

$$U_{12}(\omega) = N(\omega)e^{j\omega\frac{\delta}{2}} \sin\left(\omega\frac{\delta}{2}\right) \sin(\omega\tau), \tag{4}$$

where $U_{12}(\omega)$ and $N(\omega)$ are the STFTs of $u_{12}(t)$ and $n(t)$, respectively.

Note that the output of this beamformer does not contain any desired speech components that have been blocked successfully. Given the *direction of arrival* (DOA) of the directional noise signal δ, the noise spectrum can be easily estimated from the output of this beamformer, given by

$$N(\omega) = \underbrace{\frac{1}{e^{j\omega\frac{\delta}{2}} \sin\left(\omega\frac{\delta}{2}\right) \sin(\omega\tau)}}_{\text{weight factor}} U_{12}(\omega). \tag{5}$$

(2) *Noise direction estimation*. As shown in Eq. (5), the DOA information of the directional noise signal δ is a "must" for estimating the noise spectrum. To do this, a robust direction finder integrating two subtractive beamformers with the traditional cross-correlation DOA estimation method has been presented by Mizumachi *et al.* [10].

(3) *Noise reduction*. After estimating the noise spectrum, nonlinear spectral subtraction is employed to reduce the noise estimate from the noisy signal received by one microphone [7].

This proposed method has some advantages over other traditional noise reduction algorithms: it can deal with various types of directional noise by estimating the noise spectrum frame by frame; other algorithms, however, are poor at eliminating non-stationary noise, such as sudden noise [8].

3. PROPOSED NOISE REDUCTION METHOD

In practical environments, the performance of the original noise reduction algorithm will decrease since its basic assumption is generally not satisfied in those conditions. Performance improvement is expected if a more reasonable noise model is assumed or estimated in the noise reduction algorithm.

3.1. Problem Formulation

To simplify the following explanation without losing of generality, let us assume that a microphone array with M sensors has been pre-calibrated to achieve an in-phase identical speech signal on each microphone. The observed

noisy signal $x_k(t)$ on the k-th microphone is composed of the desired speech signal $s(t)$ and additive noise $n_k(t)$, described as

$$x_k(t) = s(t) + n_k(t). \quad k = 1, 2, \ldots, M \tag{6}$$

In the frequency domain, we have what in vector form as

$$\boldsymbol{X}(\omega) = \boldsymbol{S}(\omega) + \boldsymbol{N}(\omega), \tag{7}$$

where

$$\boldsymbol{X}(\omega) = [X_1(\omega), X_2(\omega), \cdots, X_M(\omega)]^{\mathrm{T}}, \tag{8}$$

$$\boldsymbol{N}(\omega) = [N_1(\omega), N_2(\omega), \cdots, N_M(\omega)]^{\mathrm{T}}. \tag{9}$$

and the superscript $^{\mathrm{T}}$ represents the transpose operator, and $S(\omega)$, $X_k(\omega)$ and $N_k(\omega)$ are the STFTs of the respective signals.

Note that, compared with the original algorithm, our proposed algorithm has two generalizations: (1) the additive noise signal on each microphone includes all undesired signals, which might be composed of directional and nondirectional components, not only directional noise as assumed in the original algorithm; (2) the number of microphones is M, not only two as in the original algorithm.

3.2. Derivation of Proposed Noise Reduction Method

The proposed noise reduction method based on a generalized subtractive beamformer, which has a GSC-like structure, is shown in Fig. 1. This proposed method is composed of three components: a *fixed beamformer* (FBF) that constructs the speech reference signal in the upper path, a *blocking matrix* (BM) that blocks the desired speech signal and constructs the noise reference signal, and a *noise canceller* (NC) that suppresses residual noise by minimizing the power of the system output. The three components of the proposed noise reduction algorithm are implemented as follows

(1) *Fixed beamformer*. To be consistent with the original algorithm and make the implementation simple, the FBF of the proposed algorithm is an all-pass filter for the signal on the reference channel (e.g., the first microphone) and blocks the signals from other microphones. Thus, the output of FBF $Y_{\mathrm{FBF}}(\omega)$, which is the speech reference signal, is given by

$$Y_{\mathrm{FBF}}(\omega) = X_1(\omega) = \boldsymbol{W}^{\dagger}\boldsymbol{X}(\omega), \tag{10}$$

where † denotes conjugation transpose and $\boldsymbol{W}^{\dagger} = [1, 0, \cdots, 0]$.

Note that, comparatively, in the original GSC beamformer [3], the FBF was usually implemented by the DSBF which introduced some additional "NULLs" in the beam pattern of this beamformer, as shown in detail in [7].

(2) *Blocking matrix*. Since the beamformer we previously constructed successfully blocks desired speech components, the BM part of the proposed algorithm is implemented using the same mechanism, defined as ($\tau \neq 0$)

$$u_{1k}(t) = \frac{1}{4}\Big\{[x_1(t+\tau) - x_1(t-\tau)] - \big[x_k(t+\tau) - x_k(t-\tau)\big]\Big\}. \quad k = 2, 3, \ldots, M \tag{11}$$

With the generalized signal model shown in Eq. (6), the corresponding representation of this beamformer in the frequency domain can be described as

$$\begin{aligned} U_{1k}(\omega) &= \frac{1}{2} j \sin(\omega\tau)(N_1(\omega) - N_k(\omega)) \\ &= \frac{1}{2} j \sin(\omega\tau)(X_1(\omega) - X_k(\omega)). \end{aligned} \tag{12}$$

That is, we have what in vector form as

$$\boldsymbol{U}(\omega) = \boldsymbol{B}^{\dagger}(\omega)\boldsymbol{X}(\omega), \tag{13}$$

where $\boldsymbol{U}(\omega)$ and $\boldsymbol{B}^{\dagger}(\omega)$ are

$$\boldsymbol{U}(\omega) = [U_{12}(\omega), U_{13}(\omega), \cdots, U_{1M}(\omega)]^{\mathrm{T}}, \tag{14}$$

$$\begin{aligned} \boldsymbol{B}^{\dagger}(\omega) &= \frac{1}{2} j \sin(\omega\tau) \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & -1 \end{bmatrix} \\ &\triangleq \frac{1}{2} j \sin(\omega\tau)\boldsymbol{B}_1^{\dagger}. \end{aligned} \tag{15}$$

Note, that Eq. (12) cannot be converted to Eq. (4) as in the original algorithm, since the noise signals $n_1(t)$ and $n_k(t)$ on the two microphones are not directly related and no priori assumption between them is made here. Moveover, in the original GSC beamformers [3], the BM part was implemented by the difference between the observed signals on adjacent sensors given by

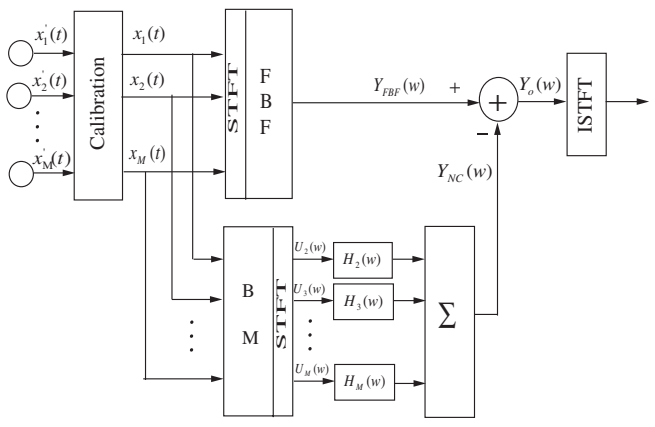

Fig. 1 Block diagram of proposed noise reduction method.

$$\boldsymbol{B}_2^{\dagger} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix}, \tag{16}$$

which indicates that only limited spatial information was used. Comparatively, the proposed algorithm considers spatial information not only between adjacent sensors but also between other sensor pairs, as shown in Eqs. (12) and (15).

(3) *Noise canceller*. The noise canceller output $Y_{\mathrm{NC}}(\omega)$, which is an estimate of noise in the speech reference signal $Y_{\mathrm{FBF}}(\omega)$, is constructed by filtering the BM outputs $\boldsymbol{U}(\omega)$ with the filters $\boldsymbol{H}(\omega)$ and is given by

$$Y_{\mathrm{NC}}(\omega) = \boldsymbol{H}^{\dagger}(\omega)\boldsymbol{U}(\omega), \tag{17}$$

where

$$\boldsymbol{H}(\omega) = [H_2(\omega), H_3(\omega), \cdots, H_M(\omega)]^{\mathrm{T}}. \tag{18}$$

With the assumption of a zero correlation between speech and noise, minimizing the mean square error between the speech reference signal $Y_{\mathrm{FBF}}(\omega)$ and the NC output $Y_{\mathrm{NC}}(\omega)$ and considering the Wiener theory, the optimal filters $\hat{\boldsymbol{H}}_{\mathrm{opt}}(\omega)$ is given by [1,11]

$$\hat{\boldsymbol{H}}_{\mathrm{opt}}(\omega) = \Phi_{\boldsymbol{UU}}^{-1}(\omega)\Phi_{\boldsymbol{UY}}(\omega), \tag{19}$$

where $\Phi_{\boldsymbol{UU}}(\omega)$ is the *cross-spectral density matrix* of the BM output signals $\boldsymbol{U}(\omega)$, $\Phi_{\boldsymbol{UY}}(\omega)$ the *cross-spectral density* between the BM output signals $\boldsymbol{U}(\omega)$ and the FBF output signal $Y_{\mathrm{FBF}}(\omega)$. They are defined as

$$\Phi_{\boldsymbol{UU}}(\omega) = E\big[\boldsymbol{U}(\omega)\boldsymbol{U}^{\dagger}(\omega)\big], \tag{20}$$

$$\Phi_{\boldsymbol{UY}}(\omega) = E\big[\boldsymbol{U}(\omega)Y_{\mathrm{FBF}}^{*}(\omega)\big], \tag{21}$$

where $E[.]$ is the expectation operator.

After determining the three components of the proposed algorithm, the output of this algorithm $Y_{\mathrm{o}}(\omega)$ is calculated as the difference between the FBF output $Y_{\mathrm{FBF}}(\omega)$ in the upper path and the NC output $Y_{\mathrm{NC}}(\omega)$ in the lower path, that is

$$Y_{\mathrm{o}}(\omega) = \boldsymbol{W}^{\dagger}(\omega)\boldsymbol{X}(\omega) - \boldsymbol{H}^{\dagger}(\omega)\boldsymbol{B}^{\dagger}(\omega)\boldsymbol{X}(\omega). \tag{22}$$

Note that the performance of the proposed algorithm should only be dependent on the characteristics of noise field since the optimal filters $\hat{\boldsymbol{H}}_{\mathrm{opt}}(\omega)$ are only determined by input noise signals under the assumption of a zero correlation between the desired speech signal and the noise signal.

3.3. Theoretical Analysis of Proposed Method

In this subsection, we first define a measure used to show the theoretical noise reduction performance of the proposed algorithm. Then its performance is examined on the basis of the coherence functions in theoretically defined noise fields.

3.3.1. Performance evaluation measure

To examine the performance of the proposed noise reduction algorithm, we define and use a measure referred to as *noise reduction performance* (NR). NR is defined as the ratio of the *power spectral density* (PSD) of the system input $\phi_{XX}^{(n)}(\omega)$ and that of the system output $\phi_{Y_{\mathrm{o}}Y_{\mathrm{o}}}^{(n)}(\omega)$ when no desired speech signal is present, and is given by [12]

$$\mathrm{NR}(\omega) = \frac{\phi_{XX}^{(n)}(\omega)}{\phi_{Y_{\mathrm{o}}Y_{\mathrm{o}}}^{(n)}(\omega)}, \tag{23}$$

where $\phi_{XX}^{(n)}(\omega) = E[N(\omega)N^{*}(\omega)]$ and $\phi_{Y_{\mathrm{o}}Y_{\mathrm{o}}}^{(n)}(\omega) = E[Y_{\mathrm{o}}(\omega)Y_{\mathrm{o}}^{*}(\omega)]$.

Under the assumptions that (1) the desired speech and noise are uncorrelated, (2) the PSDs of noises on all microphone are identical, we can respectively rewrite $\hat{\boldsymbol{H}}_{\mathrm{opt}}(\omega)$ and NR as (see Appendix A for detail)

$$\hat{\boldsymbol{H}}_{\mathrm{opt}}(\omega) = \big(\boldsymbol{B}^{\dagger}(\omega)\boldsymbol{\Gamma}(\omega)\boldsymbol{B}(\omega)\big)^{-1}\boldsymbol{B}^{\dagger}(\omega)\boldsymbol{\Gamma}(\omega)\boldsymbol{W}(\omega), \tag{24}$$

and

$$\begin{aligned}\mathrm{NR}(\omega) = \Big[&\boldsymbol{W}^{\dagger}(\omega)\boldsymbol{\Gamma}(\omega)\boldsymbol{W}(\omega) - \boldsymbol{W}^{\dagger}(\omega)\boldsymbol{\Gamma}(\omega)\boldsymbol{B}_1 \\ &\big(\boldsymbol{B}_1^{\dagger}\boldsymbol{\Gamma}(\omega)\boldsymbol{B}_1\big)^{-1}\boldsymbol{B}_1^{\dagger}\boldsymbol{\Gamma}(\omega)\boldsymbol{W}(\omega)\Big]^{-1},\end{aligned} \tag{25}$$

where $\boldsymbol{\Gamma}(\omega)$ is the coherence function matrix of noise signals on all microphones, given by

$$\boldsymbol{\Gamma}(\omega) = \begin{bmatrix} 1 & \Gamma_{N_1N_2}(\omega) & \cdots & \Gamma_{N_1N_M}(\omega) \\ \Gamma_{N_2N_1}(\omega) & 1 & \cdots & \Gamma_{N_2N_M}(\omega) \\ \vdots & \ddots & \ddots & \vdots \\ \Gamma_{N_MN_1}(\omega) & \Gamma_{N_MN_2}(\omega) & \cdots & 1 \end{bmatrix} \tag{26}$$

and $\Gamma_{N_kN_l}(\omega)$ is the complex coherence function between the noises $N_k(\omega)$ and $N_l(\omega)$, defined as

$$\Gamma_{N_kN_l}(\omega) = \frac{\phi_{N_kN_l}(\omega)}{\sqrt{\phi_{N_kN_k}(\omega)\phi_{N_lN_l}(\omega)}}. \tag{27}$$

Note that, as Eqs. (24) and (25) show, the optimal NC filters and noise reduction performance are only determined by the coherence function matrix $\boldsymbol{\Gamma}(\omega)$ of noise signals, corresponding to the characteristics of noise fields.

3.3.2. Theoretical performance analysis

In the following, we examine the performance of the proposed algorithm in theoretically defined noise fields.

(1) *Coherent noise field*. In a coherent noise field, e.g., a point sound source in the far field of a microphone array, the coherence function $\Gamma_{N_kN_l}(\omega)$ is given by [12,13]

$$\Gamma_{N_kN_l}(\omega) = e^{-j\omega\delta_{kl}}, \tag{28}$$

where δ_{kl} denotes the time delay between the k-th and l-th microphones. To determine the relationship between this proposed algorithm and the original algorithm [7,8], let us assume that only two microphones are available and the time delay between them is δ. The optimal solution for the NC filter can be derived as (see Appendix B)

$$\hat{H}^{*}_{\mathrm{opt}}(\omega) = \frac{1}{e^{j\omega\frac{\delta}{2}}\sin(\omega\tau)\sin\left(\omega\frac{\delta}{2}\right)}, \qquad (29)$$

where the superscript * is the conjugation operator.

Comparing this optimal NC filter in Eq. (29) with the "weight factor" in Eq. (5), note that they are exactly same, which indicates the following

a. The proposed noise reduction algorithm reduces to the previously presented original algorithm in a perfectly coherent noise field.
b. The original algorithm is also an optimal solution in the *minimum mean square error* (MMSE) sense for reducing coherent noise.

Putting Eq. (28) into (25), we can see that the noise reduction performance of this proposed algorithm reaches infinity at all frequencies in a coherent noise field.

(2) *Incoherent noise field*. In an incoherent noise field, e.g., the sensor self-noise, the coherence function is zero for all frequencies, $\Gamma_{n_k n_l}(\omega) = 0, \forall\omega$. In this noise field, the noise reduction performance amounts to M, the number of microphones.

(3) *Diffuse noise field*. A diffuse noise field has been shown to be a reasonable model of many practical noise environments, such as reverberant rooms and car environments [9,14]. A diffuse noise field is characterized by the coherence function [12,13,15]

$$\Gamma(\omega) = \frac{\sin(\omega d/c)}{\omega d/c}, \qquad (30)$$

where d and c represent the inter-element spacing and the velocity of sound, respectively. Putting Eq. (30) into (25), we can find that noise reduction performance depends on the inter-element spacing d and the number of microphones M. Figures 2 and 3 plot the noise reduction performance as a function of the frequency for different inter-element spacings d and different numbers of microphones M. Figures 2 and 3 show that the proposed algorithm achieves a high noise reduction performance at moderate and high frequencies, with a relatively low ability at very low frequencies (especially, when the inter-element spacing d is small).

Moreover, under the assumption of identical noise PSD on each microphone, we can derive the same noise reduction performance for our proposed algorithm and the original GSC beamformer [3,14], as shown in Figs. 2 and 3. However, in practical environments, the noise PSDs on different microphones are generally not equivalent, which makes it difficult to represent noise reduction as a function of noise coherence function; this is the failure of Eq. (25). In this case, the performance of the two algorithms is examined by experiments using real-world recordings in the following section.

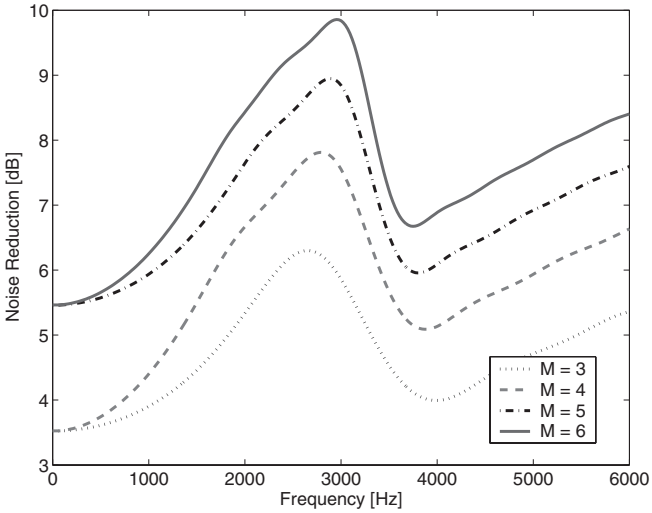

Fig. 2 Noise reduction performance in a diffuse noise field for different numbers of microphones ($d = 10$ cm).

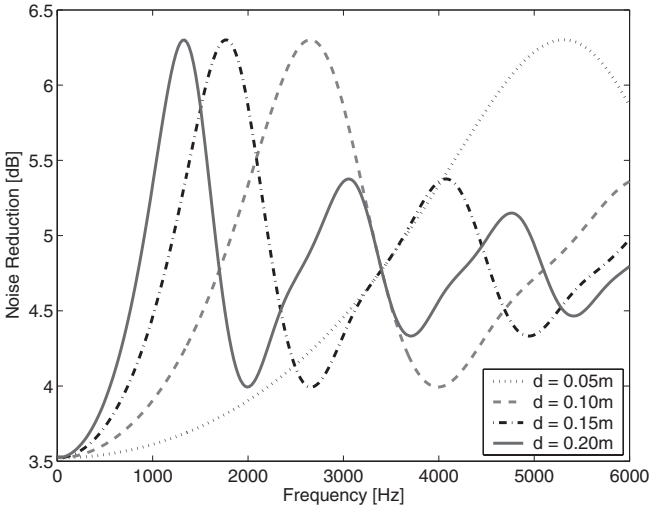

Fig. 3 Noise reduction performance in a diffuse noise field for different distances between adjacent microphones ($M = 3$).

4. EXPERIMENTS AND RESULTS

The performance of the proposed noise reduction algorithm based on a generalized subtractive beamformer

(PRO-GSBF) was evaluated using multichannel recordings and its performance was further compared with that of other traditional algorithms, delay-and-sum beamformer (DSBF), the original subtractive-beamformer-based algorithm (ORG-SBF) [8] and the original GSC beamformer (ORG-GSC) [3,14,15], in terms of both objective and subjective evaluation measures.

The proposed algorithm and other traditional algorithms were performed using the *overlap-and-add* (OLA) technique. The window length was 42.6 ms (512 samples) with an overlap of 21.3 ms (256 samples). In our implementations, for the PRO-GSBF and ORG-GSC [14], the estimated noise component (the output of the NC filter) was subtracted from the output of the upper path in a spectral magnitude domain, not in a complex spectral domain considered in the theoretical analysis. This is same as the ORG-SBF [8] and different from the ORG-GSC discussed in detail in [14]. We performed this implementation for the following reasons: (1) phase information for speech quality is relatively unimportant in speech enhancement applications [16]; (2) amplitude spectra are only important for speech recognition systems [17].

To assess the performance of the studied noise reduction algorithms, an equally spaced linear array consisting of three microphones with an inter-element spacing of 10 cm was mounted on the roof near the driver's sun visor in a car. The array was about 50 cm away from and directly in front of the driver. Multichannel speech recordings were performed across all channels when the car is stationary. Speech signals, consisting of 100 Japanese city names, were uttered by two speakers (one male and one female) at the driver's position. Multichannel noise recordings were performed across all channels when the car was running under two conditions: (1) at a speed of 50 km/h without air-conditioner noise (the air conditioner is off), (2) at a speed of 100 km/h with a high-level air-conditioner noise (the air conditioner is on). Both speech and noise signals were first resampled to 12 kHz at a 16 bit accuracy. We generated multichannel noisy signals by artificially mixing multichannel speech recordings and multichannel car noise recordings at different global SNRs $[-5, 15]$ dB. (The calculation of global SNR is detailed in [18].)

4.1. Objective Evaluation Measures

The objective evaluation measures used in our experiments include *Segmental SNR* (SEGSNR) and *Mel-Frequency Cepstral Coefficient* (MFCC) Distance.

Segmental SNR (SEGSNR) is a widely used objective evaluation criterion for speech enhancement or noise reduction algorithms since it highly correlates to subjective results [18]. SEGSNR is defined as the ratio of the power of an "ideal" clean speech to that of the noise signal embedded in a noisy signal or an enhanced speech signal processed by tested algorithms over all frames, given by

$$\text{SEGSNR} = \frac{1}{L}\sum_{l=0}^{L-1} 10\log_{10}\left(\frac{\sum_{m=0}^{W-1}[s(lW+m)]^2}{\sum_{m=0}^{W-1}[\hat{s}(lW+m)-s(lW+m)]^2}\right), \quad (31)$$

where $s(.)$ and $\hat{s}(.)$ are the reference speech signal and noisy signal or enhanced signals processed by the tested algorithms; L and W represent the number of frames in the signal and the number of samples per frame (equal to the length of STFT), respectively. Note, that a higher SEGSNR means a higher speech quality of the enhanced signal.

A second evaluation measure, MFCC distance, is defined as the distance between MFCCs of a clean speech signal and those of a noisy signal or an enhanced signal, and is represented as

$$d_{\text{mfcc}} = \frac{1}{|\Phi|}\sum_{l\in\Phi}\sum_{i}\left(c_i - c_i'\right)^2, \quad (32)$$

where Φ represents the set of frames in which speech is present and $|\Phi|$ its cardinality; c_i and c_i' are the 12-order MFCCs of the clean speech signal and noisy signal or enhanced speech signals, respectively. Note, that a shorter MFCC distance indicates a lower speech distortion, corresponding to a higher speech quality.

4.2. Objective Evaluation Results

Experimental results of SEGSNR, averaged across all sentences under two noise conditions at various SNRs, are plotted in Fig. 4. The results demonstrate that the DSBF provides a very limited SEGSNR improvement since only three microphones were used in our experiments, and that the ORG-SBF does not show sufficient performance improvement due to its unpractical assumption of a coherent noise field. The ORG-GSC beamformer shows higher SEGSNR improvements compared with the DSBF and ORG-SBF. Furthermore, the PRO-GSBF offers the highest SEGSNR improvements, corresponding to the highest speech quality, among the studied algorithms under all test conditions.

Experimental results of MFCC distance under two noise conditions at various SNRs are plotted in Fig. 5. Compared with the noisy inputs, the DSBF and ORG-SBF algorithms decrease MFCC distances under all conditions, particularly at low SNRs. The ORG-GSC beamformer shows a further decrease under all noise conditions. Moreover the PRO-GSBF method offers the shortest MFCC distance, corresponding to the lowest speech distortion, compared with other algorithms under all conditions.

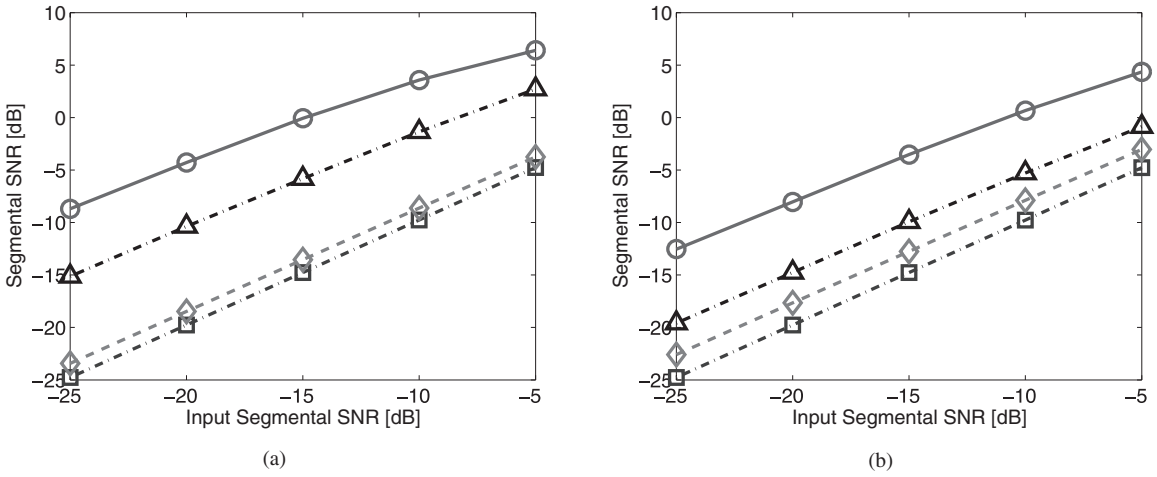

Fig. 4 Average segmental SNR (SEGSNR) at delay-and-sum beamformer (DSBF) output (□), original GSC beamformer (ORG-GSC) (△), original subtractive-beamformer-based (ORG-SBF) algorithm output (◇) and proposed generalized subtractive-beamformer-based (PRO-GSBF) algorithm output (○), under various noise conditions: speeds of 50 km/h (a) and 100 km/h (b).

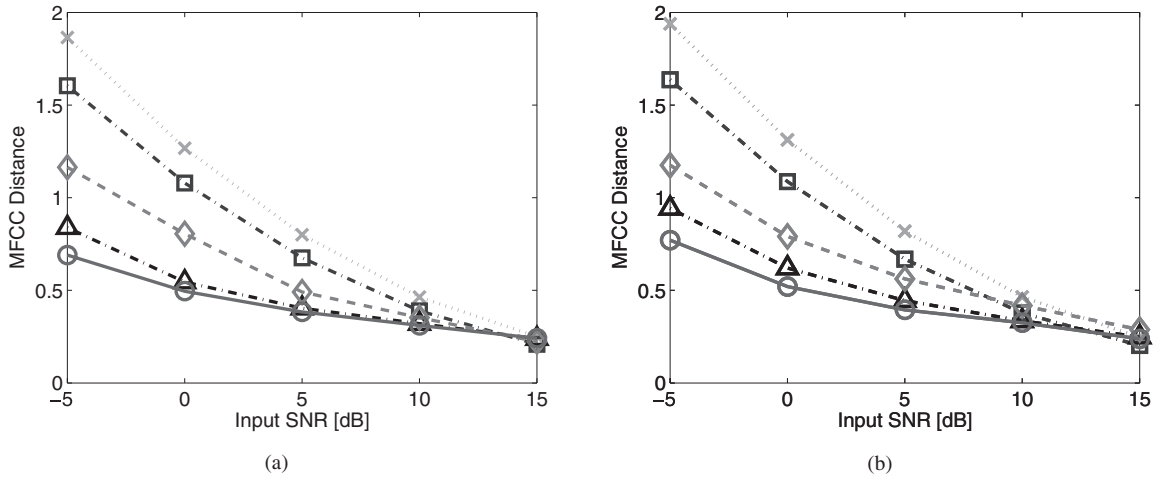

Fig. 5 Average MFCC distance (MCD) at first microphone (×), delay-and-sum beamformer (DSBF) output (□), original GSC beamformer (ORG-GSC) (△), original subtractive-beamformer-based (ORG-SBF) algorithm output (◇) and proposed generalized subtractive-beamformer-based (PRO-GSBF) algorithm output (○), under various noise conditions: speeds of 50 km/h (a) and 100 km/h (b).

4.3. Subjective Evaluation Results

Subjective evaluations of the studied algorithms were performed using speech spectrograms. Typical examples of speech spectrograms, corresponding to the Japanese sentence "hatinohe kesennuma yukuhasi," are plotted in Fig. 6, in a car environment at a speed of 100 km/h. As Fig. 6(c) shows, the output of the DSBF is characterized by high-level noise since only a small number (3ch) of microphones were used. The ORG-GSC does not have sufficient suppression ability for low-frequency noise, as shown in Fig. 6(d). As plotted in Fig. 6(e), the ORG-SBF algorithm still shows very limited performance improvement, especially in the low frequency region. Comparatively, Fig. 6(f) demonstrates that the PRO-GSBF algorithm provides a much higher performance improvement, particularly in low-frequency region, compared with the other studied algorithms.

4.4. Discussions

From the experimental results presented in the last subsection, the superiorities of the proposed generalized method to the other algorithms are discussed below.

The proposed PRO-GSBF outperforms the DSBF. For the DSBF, many microphones are needed to obtain an acceptable performance, whereas for the proposed method, few microphones are sufficient to achieve the same noise

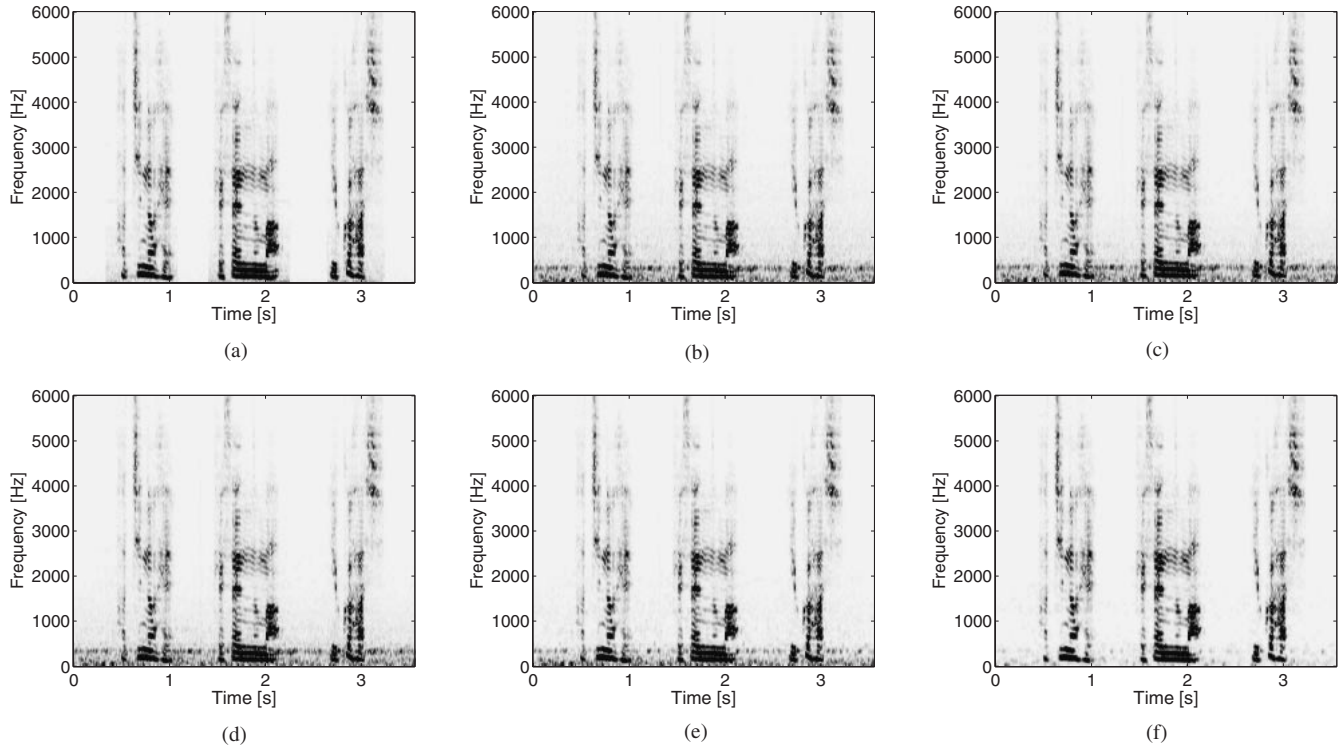

Fig. 6 Speech spectrograms. (a) original clean speech signal at first microphone: "hatinohe kesennuma yukuhasi"; (b) noisy signal at first microphone (SNR = 5 dB); (c) delay-and-sum beamformer (DSBF) output; (d) original GSC beamformer (ORG-GSC) output; (e) original subtractive-beamformer-based (ORG-SBF) algorithm output; (f) proposed generalized subtractive-beamformer-based (PRO-GSBF) algorithm output.

reduction performance.

The proposed PRO-GSBF outperforms the ORG-SBF algorithm. The basic assumption of the ORG-SBF algorithm, namely a perfectly coherent noise field, is seldom satisfied in real-world environments. On the other hand, no priori assumption on noise signals is made in the PRO-GSBF method. That is, the PRO-GSBF algorithm is a natural extension of the ORG-SBF in which the unpractical assumption of an coherent noise field to that of an arbitrary noise field is relaxed. Therefore, the improved noise reduction performance can be achieved for the PRO-GSBF algorithm. Moreover, the high performance in reducing unstable noise (sudden noise) is expected for the PRO-GSC beamformer because the PRO-GSC is derived based on the same ideas as those of the ORG-SBF which has the ability of reducing sudden noise.

The proposed PRO-GSBF outperforms the ORG-GSC algorithm. In theory, with the assumption of identical noise PSD on each microphone, both PRO-GSBF and ORG-GSC show the same noise reduction performance. In practice, noise PSDs on different microphones are generally different. The PRO-GSBF provides improved noise reduction performance, particularly in reducing low-frequency noise, due to the fact that different inter-element spacings (more spatial information) are used, as shown by Eq. (15). On the other hand, the ORG-GSC beamformer achieves limited performance owing to the use of limited spatial information, shown by Eq. (16). However, if the desired speech signals received by different microphones are greatly different, both PRO-GSBF and ORG-GSC will introduce some speech distortion, particularly for PRO-GSBF which is also because of the use of sensor pairs with larger spacings.

Thus, the proposed algorithm provided the highest noise reduction performance among the studied algorithms under all experimental conditions, as shown in Sections 4.2 and 4.3.

5. CONCLUSIONS

In this paper, we developed a noise reduction method based on a generalized subtractive beamformer and in which the strict assumption of a coherent noise field to that of an arbitrary noise field is relaxed. This presented algorithm is an extension of the original noise reduction algorithm we previously presented. The theoretical results showed that the proposed generalized algorithm includes the original algorithm as a special case in a perfectly coherent noise field. The performance limits of this generalized algorithm were also examined in three theoretically defined noise fields (coherent, incoherent and diffuse noise fields). Compared with other traditional algorithms, given the noise coherent functions, the pro-

posed algorithms can deal with all types of interfering noise signal (e.g., sudden noise) with a small number of microphones and with no adaptive signal processing techniques. Experimental results using real-world recordings demonstrated that the proposed algorithm outperforms other traditional algorithms.

REFERENCES

[1] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications* (Springer-Verlag, Berlin, 2001).

[2] O. L. Frost, "An algorithm for linearly constrained adaptive array processing," *Proc. IEEE*, **60**, 926–935 (1972).

[3] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propag.*, **AP-30**, 27–34 (1982).

[4] S. Gannot, D. Burshtein and E. Weinstein, "Signal enhancement using beamforming and nonstationary with application to speech," *IEEE Trans. Signal Process.*, **49**, 1614–1626 (2001).

[5] Y. H. Chen and H. D. Fang, "Frequency-domain implementation of Griffiths-Jim adaptive beamformer," *J. Acoust. Soc. Am.*, **6**, 3354–3366 (1992).

[6] J. An and B. Champagne, "GSC realisations using the two-dimensional transform-domain LMS algorithm," *IEE Proc. Radar Sonar Navig.*, **141**, pp. 270–278 (1994).

[7] M. Akagi and M. Mizumachi, "Noise Reduction by Paired Microphones," *Proc. EUROSPEECH 97*, pp. 335–338 (1997).

[8] M. Akagi and T. Kago, "Noise reduction using a small-scale microphone array in multi noise source environment," *Proc. ICASSP 2002*, pp. 909–912 (2002).

[9] J. Li, X. Lu and M. Akagi, "Noise reduction system in arbitrary noise environments and its applications to speech enhancement and speech recognition," *Proc. ICASSP 2005*, pp. 277–280 (2005).

[10] M. Mizumachi and M. Akagi, "Noise reduction method that is equipped for a robust direction finder in adverse environments," *Proc. IEEE Workshop Robust Method for Speech Recognition in Adverse Conditions*, pp. 179–182 (1999).

[11] I. Cohen, "Multi-channel post-filtering in non-stationary noise environments," *IEEE Trans. Signal Process.*, **52**, 1149–1160 (2004).

[12] J. Meyer, K. U. Simmer and K. D. Kammeyer, "Comparison of one- and two-channel noise estimation techniques," *Proc. IWAENC 97*, pp. 17–20 (1997).

[13] J. Chen, L. Shue, K. Phua and H. Sun, "Theoretical comparison of dual microphone systems," *ICASSP 2004*, pp. VI 73–76 (2004).

[14] J. Bitzer, K. U. Simmer and K. D. Kammeyer, "Theoretical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement," *Proc. ICASSP 99*, pp. 2965–2968 (1999).

[15] J. Bitzer, K. D. Kammeyer and K. U. Simmer, "An alternative implementation of the superdirective beamformer," *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, pp. 7–10 (1999).

[16] D. L. Wang, J. S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust. Speech Signal Process.*, **30**, 679–681 (1982).

[17] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, 1993).

[18] S. R. Quackenbush, T. P. Barnwell and M. A. Clements, *Objective Measures of Speech Quality* (Prentice-Hall, Englewood Cliffs, 1988).

APPENDIX A: DERIVATION OF *NOISE REDUCTION PERFORMANCE* (NR)

For simplification, we omit the frequency index ω in the following derivation.

To avoid the cancellation of the desired speech signal, we calculate the optimal NC filters $\hat{\boldsymbol{H}}_{\text{opt}}$ when desired speech is absent, that is, $\boldsymbol{X} = \boldsymbol{N}$. Thus, Eqs. (10) and (13) can be respectively rewritten as

$$Y_{\text{FBF}} = \boldsymbol{W}^{\dagger}\boldsymbol{N}, \tag{33}$$

$$\boldsymbol{U} = \boldsymbol{B}^{\dagger}\boldsymbol{N}. \tag{34}$$

Using Eqs. (33) and (34), the PSDs $\boldsymbol{\Phi}_{\boldsymbol{UU}}$ and $\boldsymbol{\Phi}_{\boldsymbol{UY}}$ are respectively calculated as

$$\boldsymbol{\Phi}_{\boldsymbol{UU}} = \boldsymbol{B}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}\boldsymbol{B} \tag{35}$$

and

$$\boldsymbol{\Phi}_{\boldsymbol{UY}} = \boldsymbol{B}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}\boldsymbol{W}, \tag{36}$$

where $\Phi_{\boldsymbol{NN}} = E[\boldsymbol{N}\boldsymbol{N}^{\dagger}]$. Substituting Eqs. (35) and (36) into Eq. (19), the optimal NC filters $\hat{\boldsymbol{H}}_{\text{opt}}$ are

$$\hat{\boldsymbol{H}}_{\text{opt}} = \left(\boldsymbol{B}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}\boldsymbol{W}. \tag{37}$$

With Eqs. (22) and (37), the PSD of the output signal Y_{o} is given by

$$\begin{aligned}\phi_{Y_{\text{o}}Y_{\text{o}}} = {} & \boldsymbol{W}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{XX}}\boldsymbol{W} - \boldsymbol{W}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{XX}}\boldsymbol{B}\left(\boldsymbol{B}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}\boldsymbol{W} \\ & - \boldsymbol{W}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}^{\dagger}\boldsymbol{B}\left(\boldsymbol{B}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}^{\dagger}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{XX}}\boldsymbol{W} + \boldsymbol{W}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}^{\dagger}\boldsymbol{B} \\ & \left(\boldsymbol{B}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}^{\dagger}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{XX}}\boldsymbol{B}\left(\boldsymbol{B}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}\boldsymbol{W}.\end{aligned} \tag{38}$$

To determine theoretical noise reduction performance, we consider speech-absent periods. In this case, the output PSD $\Phi_{Y_{\text{o}}Y_{\text{o}}}$ reduces to

$$\phi_{Y_{\text{o}}Y_{\text{o}}}^{(n)} = \boldsymbol{W}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}\boldsymbol{W} - \boldsymbol{W}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}\boldsymbol{B}\left(\boldsymbol{B}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^{\dagger}\boldsymbol{\Phi}_{\boldsymbol{NN}}\boldsymbol{W}, \tag{39}$$

Under the assumption of identical noise PSDs on all microphones, $\boldsymbol{\Phi}_{\boldsymbol{NN}}$ should be $\boldsymbol{\Phi}_{\boldsymbol{NN}} = \phi_{NN}\boldsymbol{\Gamma}$, where $\boldsymbol{\Gamma}$ denotes the complex coherence function given by Eq. (26), and the PSD of input $\phi_{\boldsymbol{XX}}$ reduces to

$$\phi_{XX}^{(n)} = \phi_{NN} \tag{40}$$

Using Eqs. (19), (23), (39) and (40), we can respectively rewrite the optimal NC filters $\hat{\boldsymbol{H}}$ and *Noise Reduction Performance* (NR) as

$$\hat{\boldsymbol{H}}_{\text{opt}} = \left(\boldsymbol{B}^{\dagger}\boldsymbol{\Gamma}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^{\dagger}\boldsymbol{\Gamma}\boldsymbol{W}, \tag{41}$$

and

$$\text{NR} = \left(\boldsymbol{W}^{\dagger}\boldsymbol{\Gamma}\boldsymbol{W} - \boldsymbol{W}^{\dagger}\boldsymbol{\Gamma}\boldsymbol{B}_1\left(\boldsymbol{B}_1^{\dagger}\boldsymbol{\Gamma}\boldsymbol{B}_1\right)^{-1}\boldsymbol{B}_1^{\dagger}\boldsymbol{\Gamma}\boldsymbol{W}\right)^{-1}. \tag{42}$$

APPENDIX B: DERIVATION OF OPTIMAL NC FILTER FOR COHERENT NOISE FIELD

Assuming only two microphones are available, the BM output is a one-channel signal, given by

$$U(\omega) = \frac{1}{2} j \sin(\omega\tau)(N_1(\omega) - N_2(\omega)). \tag{43}$$

With the assumption of identical noise PSDs on all microphones, the PSDs of $\Phi_{UU}(\omega)$ and $\Phi_{UY}(\omega)$ is respectively given by

$$\Phi_{UU}(\omega) = \frac{1}{2}\phi_{NN}(\omega)\sin^2(\omega\tau)\big(1 - \Re\{\Gamma_{N_1N_2}(\omega)\}\big), \tag{44}$$

$$\Phi_{UY}(\omega) = \frac{1}{2} j\phi_{NN}(\omega)\sin(\omega\tau)\big(1 - \Gamma_{N_2N_1}(\omega)\big). \tag{45}$$

Substituting Eqs. (44) and (45) into Eq. (19), the optimal NC filter $\hat{H}_{\text{opt}}(\omega)$ is obtained as

$$\hat{H}^*_{\text{opt}}(\omega) = \frac{-j\big(1 - \Gamma_{N_1N_2}(\omega)\big)}{\sin(\omega\tau)\big(1 - \Re\{\Gamma_{N_1N_2}(\omega)\}\big)}. \tag{46}$$

In a coherent noise field, substituting Eqs. (28) into (46), the optimal NC filter $\hat{H}_{\text{opt}}(\omega)$ in this field is obtained as

$$\hat{H}^*_{\text{opt}}(\omega) = \frac{1}{e^{j\omega\frac{\delta}{2}}\sin(\omega\tau)\sin\left(\omega\frac{\delta}{2}\right)}. \tag{47}$$

Obviously, this optimal filter is exactly identical to the "weight factor" in Eq. (5) in our original algorithm.

Junfeng Li received his B.E. degree from Zhengzhou University of Technology and his M.S. degree from Xidian University, China, both in Computer Science, in 2000 and 2003, respectively. He is currently pursuing his Ph.D. degree in Information Science at Japan Advanced Institute of Science and Technology, Ishikawa, Japan. His current research interests include noise reduction and speech enhancement using microphone arrays, robust speech recognition in adverse environments.

Masato Akagi received his B.E. degree in Electronic Engineering from Nagoya Institute of Technology in 1979, and his M.E. and Dr. Eng. degrees in Computer Science from Tokyo Institute of Technolgoy in 1981 and 1984, respectively. In 1984, he joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation (NTT). From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992, he has been with the School of Information Science, Japan Advanced Institute of Science and Technology, where he is currently a professor. His research interests include speech perception mechanisms of humans and speech signal processing. Dr. Akagi received the IEICE Excellent Paper Award from an IEICE in 1987, and the Sato Prize for Outstanding Paper from ASJ in 1998.