PAPER Special Section on Acoustic Scene Analysis and Reproduction

A Two-Microphone Noise Reduction Method in Highly Non-stationary Multiple-Noise-Source Environments*

Junfeng LI[†a)], ***Nonmember***, **Masato AKAGI**[†b)], ***and*** **Yôiti SUZUKI**[††c)], ***Members***

SUMMARY In this paper, we propose a two-microphone noise reduction method to deal with non-stationary interfering noises in multiple-noise-source environments in which the traditional two-microphone algorithms cannot function well. In the proposed algorithm, multiple interfering noise sources are regarded as one virtually integrated noise source in each subband, and the spectrum of the integrated noise is then estimated using its virtual direction of arrival. To do this, we suggest a direction finder for the integrated noise using only two microphones that performs well even in speech active periods. The noise spectrum estimate is further improved by integrating a single-channel noise estimation approach and then subtracted from that of the noisy signal, finally enhancing the desired target signal. The performance of the proposed algorithm is evaluated and compared with the traditional algorithms in various conditions. Experimental results demonstrate that the proposed algorithm outperforms the traditional algorithms in various conditions in terms of objective and subjective speech quality measures.

key words: *noise reduction, non-stationary noise, multiple-noise-source environments, virtually integrated sound source*

1. Introduction

Performance and robustness of hands-free speech applications are dominantly degraded by acoustical noise, especially multiple highly non-stationary noises. Dealing with multiple non-stationary noise signals is today one of challenging research topics in speech research field. Compared with single-channel noise reduction technique, multi-channel technique has demonstrated the great potential in suppressing noise signals and enhancing speech quality using the spatial information of signals and acoustical environments [1]. In some applications, e.g., hearing aids, the number of microphones is strictly limited because of the practical requirements (e.g., the limited space and low energy capacity). Therefore, only small-size noise reduction systems are preferred to these applications. Two-microphone noise reduction, which is the smallest multi-microphone noise reduction system, has been a promising technique and attracted more attention recently [1], [2].

Among various multi-microphone approaches, beamforming technique has widely been used for noise reduction and speech enhancement [1]. The simplest beamforming technique is known as *delay-and-sum* (DAS) beamforming where the time-domain sensor signals are first delayed and then summed to give a single-channel output [3]. However, the low directivity of DAS beamformer results in low noise reduction performance. Another class of beamforming techniques is *superdirective* (SD) beamformer which calculates the channel filters by maximizing the array factor of directivity [4]. However, the beam pattern of SD beamformer is traditionally designed for diffuse noise condition and time-invariant, demonstrating degraded performance in non-diffuse or time-varying conditions. To deal with the problems (e.g., fixed directivity) of DAS and SD beamformers, therefore, adaptive beamforming techniques are called for [1]. A commonly used adaptive beamforming technique, referred to as *generalized sidelobe canceller* (GSC), was proposed to reduce the interfering noises [5]. Bitzer et al. theoretically analyzed the performance of the GSC algorithm and showed that the GSC algorithm is successful to suppress coherent noise when the number of noise source is less than that of microphones [6]. Moreover, to deal with time-varying noises, adaptive signal processing techniques (e.g., *least mean square* (LMS)) are normally exploited in the implementation of the GSC beamformer [5]. The problems associated with the traditional adaptive beamformer (e.g., GSC beamformer) are the inability in suppressing highly non-stationary noises due to the low convergence rate of adaptive signal processing and the low noise reduction performance when the number of the noise is larger than or equal to that of the microphones [5], [6], [9]. Recently, Kim et al. proposed a two-channel beamformer based on short-time spectral amplitude estimation [7]. However, its performance goes down for the multiple highly non-stationary interfering signals, because of the recursive estimation of transfer function using the long-time averaged spectrum of the observed signals and noise signals. More recently, Takahashi et al. introduced to estimate noise power spectrum based on *independent component analysis* (ICA) and further to suppress the spatially-distributed interfering signals [8]. This method involves the high computational cost due to the iterative estimation of the unmixing matrix and suffers from the inherent problems of the ICA algorithms, such as the violation of the independence assumption between sound sources in real environments.

Manuscript received August 3, 2007.
Manuscript revised November 19, 2007.

[†]The authors are with School of Information Science, Japan Advanced Institute of Science and Technology, Nomi-shi, 923-1292 Japan.
[††]The author is with Research Institute of Electrical Communication, Tohoku University, Sendai-shi, 980-8577 Japan.

*Part of this work was done when Junfeng Li was a post-doctoral researcher at Research Institute of Electrical Communication, Tohoku University, Sendai, Japan.

a) E-mail: junfeng@jaist.ac.jp
b) E-mail: akagi@jaist.ac.jp
c) E-mail: yoh@ais.riec.tohoku.ac.jp

DOI: 10.1093/ietfec/e91–a.6.1337

To deal with the problems of the traditional algorithms mentioned above, in this paper, we propose a two-microphone noise reduction algorithm to reduce highly non-stationary multiple-source interfering signals. In the proposed algorithm, the interfering signal is analytically estimated and subtracted from the observed noisy signal. No adaptive signal processing technique (e.g., LMS) is used, which avoids the problems suffered from the low convergence rate of adaptive signal processing and is expected to offer the high ability in reducing highly non-stationary interfering signals. Furthermore, to suppress the multiple-source interfering signals, we perform the interfering signal spectrum estimation in each subband based on its virtual DOA instead of the real DOAs of each interfering sources in the entire band. To do this, we develop a novel direction finder for the virtual interfering signal using only two microphones that is successful in estimating the DOA of interfering signal even in speech active periods. Moreover, the spectrum estimation accuracy of interfering signal is further improved by combining a single-channel noise estimation approach, which mitigates the sidelobe problems of the small-size two-microphone system. Compared with the traditional two-microphone arrays, the superiority of the proposed algorithm is finally confirmed in reducing non-stationary multiple interfering noises in various conditions.

2. Signal Model

Consider that in a noisy environment, two microphones are positioned arbitrarily with the inter-element spacing of d for noise reduction. The observed signal on each microphone consists of desired speech signal $s(t)$ coming from the direction such that the time delay between two microphones is ξ, and interfering noise signals $n_m(t), m = 1, 2, \ldots, M$ coming from the directions such that the time delays are $\delta_m, m = 1, 2, \ldots, M$. Hence, the signals $x_1(t)$ and $x_2(t)$ observed on two microphones can be represented as

$$x_1(t) = s(t) + \sum_{m=1}^{M} n_m(t), \tag{1}$$

$$x_2(t) = s(t - \xi) + \sum_{m=1}^{M} n_m(t - \delta_m). \tag{2}$$

The time delay of the desired speech signal can be compensated using the coherence based time delay estimation technique [10]. In this paper, we assume that the array has been calibrated and pre-steered to the direction of the desired speech source beforehand. Thus, the observed signals on two microphones can be reformulated by simply setting $\xi = 0$ in Eq. (2). This pre-steering may be omitted in some applications such as for hearing aids. For hearing aid users, the time delay compensation is generally accomplished by the fact that users unconsciously move their heads to the direction of the desired speech source before listening to the desired speech signal.

3. Proposed Noise Reduction Algorithm

The basic concept of the proposed noise reduction system is that noise components are first estimated in each temporal frame and each frequency subband, and then subtracted from the observed noisy signal. To estimate the interfering noise spectrum, we present a noise direction estimator using two microphones which shows good performance even when speech is present. To improve noise estimation accuracy, we suggest a hybrid noise estimation technique as well by combining a single-channel noise estimation approach.

The proposed two-microphone noise reduction system consists of four components: cancellation of the desired speech signal, noise direction estimation, noise spectrum estimation and noise reduction. The block diagram of this proposed algorithm is shown in Fig. 1.

3.1 Cancellation of the Desired Signal

Cancellation of the desired signal is achieved by subtracting the observed signal on the second microphone from that on the first microphone. The speech-cancelled signal $U(\omega)$ in the frequency domain is given by

$$\begin{aligned} U(\omega) &= X_1(\omega) - X_2(\omega) \\ &= 2j \sum_{m=1}^{M} N_m(\omega) e^{-j\omega\frac{\delta_m}{2}} \sin\left(\omega\frac{\delta_m}{2}\right), \end{aligned} \tag{3}$$

where $X_1(\omega)$ and $X_2(\omega)$ are the *short-time Fourier transform* (STFT) of the observed signals $x_1(t)$ and $x_2(t)$. $N_m(\omega)$ is the STFT of the m-th interfering noise $n_m(t)$. Note that after the microphone has been calibrated and steered to the desired direction, the speech-cancelled signal $U(\omega)$ does not include any desired speech component that has been cancelled.

Since that the sum of sinusoidal waves becomes a sinusoidal wave in a narrow subband (see Appendix for detail), we divide the full frequency band into several subbands [9]. Then we can further assume that the multiple noise sources can be regarded as one integrated interfering noise source in each subband [9]. Consequently, the speech-cancelled signal in each subband can be represented as

$$U(\tilde{\omega}) = 2jN_k(\tilde{\omega}) e^{-j\tilde{\omega}\frac{\delta_k}{2}} \sin\left(\tilde{\omega}\frac{\delta_k}{2}\right),$$

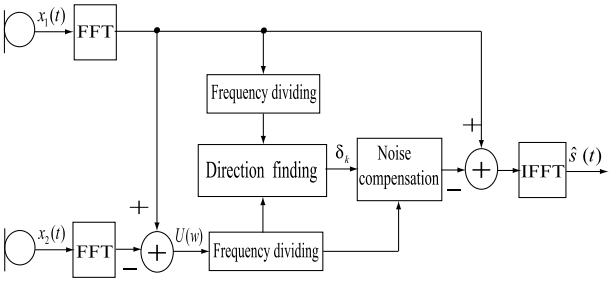

Fig. 1 Block diagram of the proposed two-microphone noise reduction system.

$$\omega_{k-1} \le |\tilde{\omega}| < \omega_k \ , \ k = 1, 2, \ldots, K \tag{4}$$

where $N_k(\tilde{\omega})$ is the spectrum of the integrated noise signal in the k-th subband, and δ_k is its virtual arrival time difference. K denotes the number of subband.

3.2 Noise Direction Estimation — A Two-Microphone Noise Direction Estimator

In order to derive the noise spectrum, as shown in Eq. (4), we have to first estimate the direction of the virtual integrated noise signal in each subband. To estimate noise direction, Mizumachi et al. [11] proposed a three-microphone noise direction finder by integrating two subtractive beamformers which were built using three microphones and the traditional cross-correlation DOA estimation method. However, this estimator fails to estimate the noise direction when only two microphones are available, such as in the proposed two-microphone noise reduction algorithm.

To provide noise direction information for estimating noise spectrum in the proposed system, we propose a noise direction estimator which exploits two microphones and is able to estimate noise direction even in speech active periods.

Let us assume that the desired speech signal and interfering noise signals are uncorrelated, which is a widely supposed assumption for speech enhancement [2], [4], [6], [12], [13]. With this assumption, the cross spectrum between the speech-cancelled signal and the observed signal on a reference microphone (e.g., the first microphone) in each subband can be written as

$$\begin{aligned} U(\tilde{\omega})X_1^*(\tilde{\omega}) &= N_k(\tilde{\omega})N_k^*(\tilde{\omega})(1 - e^{-j\tilde{\omega}\delta}) \\ &= 2jN_k(\tilde{\omega})N_k^*(\tilde{\omega})e^{-j\tilde{\omega}\frac{\delta_k}{2}} \sin\left(\tilde{\omega}\frac{\delta_k}{2}\right), \end{aligned} \tag{5}$$

where $X_1(\tilde{\omega})$ is the STFT of $x_1(t)$ in the k-th subband and the superscript * denotes conjugation operator. Note that this cross spectrum is independent of speech presence/absence state, since it includes noise-only components. With the cross spectrum in Eq. (5) and using the generalized cross-correlation based DOA estimation technique [14], the estimate of the virtual noise direction $\hat{\delta}_k$ of the k-th subband can be calculated as

$$\begin{aligned} \hat{\delta}_k = 2\hat{\delta}_k' = 2\arg\max_t \left[\mathrm{IFFT}\left[\frac{U(\tilde{\omega})X_1^*(\tilde{\omega})}{\left|U(\tilde{\omega})\|X_1^*(\tilde{\omega})\right|}\right]\right], \\ \omega_{k-1} \le |\tilde{\omega}| < \omega_k, \ \ k = 1, 2, \ldots, K \end{aligned} \tag{6}$$

where $\hat{\delta}_k'$ is the half of the difference estimate in the virtual arrival time between two microphones in the k-th subband.

It should be noted that the band width of each subband should be appropriately determined. This is because that very narrow subbands are needed for Eq. (4). However, Eq. (5) requires that the subbands cannot be very narrow since the assumption of zero correlation between speech and noise might be violated in too narrow subbands. Therefore, there is a trade-off in choosing the band width.

3.3 Noise Estimation — A Hybrid Noise Estimation Technique

Equation (4) indicates that the spectrum of the integrated noise can be estimated from that of the speech-cancelled signal $U(\tilde{\omega})$ with the help of the already estimated noise direction $\hat{\delta}_k$.

To estimate the noise spectrum on the first microphone, the speech-cancelled signal has to be compensated. To do this, with the use of the estimated noise direction $\hat{\delta}_k$, we construct a noise compensator for matching noise components on the first microphone as

$$H_k(\tilde{\omega}) = \frac{e^{j\tilde{\omega}\frac{\hat{\delta}_k}{2}}}{2j\sin\left(\tilde{\omega}\frac{\hat{\delta}_k}{2}\right)}. \tag{7}$$

Consequently, the spectral estimate of the integrated noise $\hat{N}_{m,k}(\tilde{\omega})$ in the k-th subband on the first microphone using this multi-channel (i.e., two-channel) estimation approach, can be obtained by weighting the speech-cancelled signal $U(\tilde{\omega})$ with the noise compensator $H_k(\tilde{\omega})$, given by

$$\hat{N}_{m,k}(\tilde{\omega}) = U(\tilde{\omega})H_k(\tilde{\omega}) \tag{8}$$

Note that as $\tilde{\omega}\hat{\delta}$ approaches 2π, Eq. (7) approximates infinity because of the too small value of the denominator in Eq. (7). In this case, the multi-channel estimation approach in Eq. (8) overestimates the noise spectrum, which corresponds to the sidelobe problem of the small-size multi-microphone systems. To mitigate this problem and improve the noise estimation accuracy, we present a hybrid noise estimation technique by combining a single-channel noise estimation approach. Using the hybrid noise estimation method, the spectral estimate of the integrated noise in the k-th subband is given by [15]

$$\left|\hat{N}_k(\tilde{\omega})\right| = \begin{cases} \left|\hat{N}_{m,k}(\tilde{\omega})\right|, & \left|\sin\left(\tilde{\omega}\frac{\hat{\delta}_k}{2}\right)\right| \ge \varepsilon \\ \left|\hat{N}_{s,k}(\tilde{\omega})\right|, & \text{otherwise} \end{cases} \tag{9}$$

where ε is a small positive value; $\hat{N}_{m,k}(\omega)$ and $\hat{N}_{s,k}(\omega)$ are the estimated noise spectrum by the multi-channel approach in Eq. (8) and the soft decision based single-channel noise estimation approach, given by [16]

$$\begin{aligned} &|\hat{N}_{s,k}(\tilde{\omega})|^2 \\ &\quad = \mu|\hat{N}_{s,k}^{pre}(\tilde{\omega})|^2 + (1-\mu)E\left[|N_{s,k}(\tilde{\omega})|^2\big||X_1(\tilde{\omega})|^2\right], \end{aligned} \tag{10}$$

where μ $(0 < \mu < 1)$ is a forgetting factor controlling the update rate of noise estimation, $\hat{N}_{s,k}^{pre}(\tilde{\omega})$ indicates the estimated noise spectrum in the previous frame and $E[\cdot]$ is the expectation operator. Under speech presence uncertainty, the second term in the right side of Eq. (10) can be estimated as

$$\begin{aligned} &E\left[|N_{s,k}(\tilde{\omega})|^2\big||X_1(\tilde{\omega})|^2\right] \\ &\quad = q_k(\tilde{\omega})|X_1(\tilde{\omega})|^2 + (1-q_k(\tilde{\omega}))|\hat{N}_{s,k}^{pre}(\tilde{\omega})|^2, \end{aligned} \tag{11}$$

where $q_k(\tilde{\omega})$ denotes the speech absence probability that is calculated as in [16].

Finally, the spectral estimate of the integrated noise $\hat{N}(\omega)$ on the first microphone can be calculated over the entire frequency region as

$$\left|\hat{N}(\omega)\right| = \sum_{k=1}^{K} \left|\hat{N}_k(\tilde{\omega})\right|. \ \omega_{k-1} \le |\tilde{\omega}| < \omega_k \tag{12}$$

3.4 Noise Reduction

After estimating the noise spectrum, the desired speech signals are enhanced by subtracting the estimated noise from the noisy observation on the first microphone by non-linear spectral subtraction, given by [13]

$$|\hat{S}(\omega)| = \begin{cases} |X_1(\omega)| - \alpha|\hat{N}(\omega)|, & |X_1(\omega)| \ge \alpha|\hat{N}(\omega)|, \\ \beta|X_1(\omega)|, & \text{otherwise}, \end{cases} \tag{13}$$

where α is the subtraction factor and β is the flooring factor.

4. Experiments and Results

The performance of the proposed two-microphone noise reduction algorithm was examined in various acoustic environments and further compared to that of the conventional algorithms, including the *delay-and-sum* (DAS) beamformer [3], the *superdirective* (SD) beamformer [4] and the standard *generalized sidelobe canceller* (GSC) algorithm [5].

4.1 Objective Evaluation Measures

To evaluate the studied noise reduction methods for speech enhancement, two objective speech quality measures were used: *perceptual evaluation of subjective quality* (PESQ) and *log-spectral distance* (LSD).

The first measure is *perceptual evaluation of speech quality* (PESQ) [17], which is able to predict subjective quality with good correlation in a very wide range of conditions specified by the ITU-T as recommendation P.862 [17]. Note that a higher PESQ means the higher speech quality of the enhanced signal.

The other measure is *log-spectral distance* (LSD), which is often used to assess the distortion of the desired speech signal [18], [19]. LSD is defined as the difference between the log spectrum of clean speech and that of the noisy signal or enhanced signal by the studied algorithms, given by

$$\text{LSD} = \frac{10}{L}\sum_{\ell=0}^{L-1}\left(\frac{1}{K}\sum_{\omega=0}^{W-1}\left[\log_{10}\mathcal{A}S_\ell(\omega) - \log_{10}\mathcal{A}\hat{S}_\ell(\omega)\right]^2\right)^{\frac{1}{2}}, \tag{14}$$

where $\mathcal{A}S_\ell(\omega) \triangleq \max\{|S_\ell(\omega)|^2, \delta\}$ is the clipped spectral power in the ω-th frequency bin and the ℓ-th frame such that the log-spectrum dynamic range is confined to about 50 dB (that is, $\delta = 10^{-50/10}\max_{\omega,\ell}\{|S_\ell(\omega)|^2\}$) [20]. Note that a lower LSD level indicates the lower speech distortion.

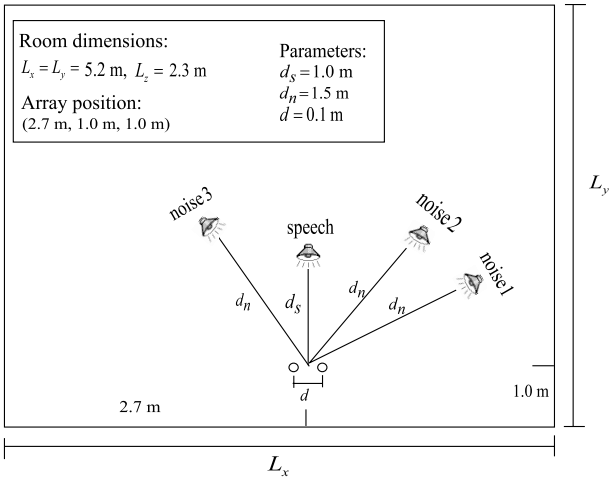

Fig. 2 Geometrical configuration of experiments in a soundproof room environment.

4.2 Experimental Configuration

To evaluate the performance of the studied algorithms, we have done the experiments in both simulated conditions and real environments with the same configuration shown in Fig. 2. Two microphones with the inter-element spacing of 10 cm were placed in a room whose dimension is 5.2 m × 5.2 m × 2.3 m. Four loudspeakers were used for one target sound source and three interfering sound sources. Both loud speakers and microphones were set 1.0 m from the floor. The target source was 1.0 m in the front of two microphones (e.g., 0 degree), and three interfering noise sources were 1.5 m apart from the microphones with DOAs of 60, 40, −40 degrees, respectively.

In the simulated environments, the impulse responses between sound sources and microphones were simulated using the image method [21] with the reverberation time of 0.1 ms to simulate the anechoic room. The observed target and interfering signals on each microphone were generated by convoluting the "dry" target and interfering signals with the simulated impulse responses. In the real acoustic environments, we recorded the sound data in a soundproof room with the same configuration shown in Fig. 2 and the reverberation time of about 0.25 s, at Research Institute of Electrical communication, Tohoku University.

For the simulated and real acoustic environments, two noise acoustic conditions, one-noise-source and three-noise-source conditions, were generated and used to examine the performance of the studied algorithms. In the one-noise-source condition, the noisy signals were obtained by summing the interfering signals with DOA of 60 degrees and the target signals at different global SNR levels [−5, 20] dB with the step of 5 dB. In the three-noise-source condition,

the integrated interfering signals at two microphones were first generated by mixing three interfering signals, and the observed noisy signals were finally created by adding the integrated interfering signals into the target signals at two microphones at different global SNR levels, as in the one-noise-source condition. Note that since the duration of target signals might be different from that of interfering signals, the signals having longer durations were truncated to the one of the shorter duration when generating the observed noisy signals.

All evaluations in both simulated and real environments were done under the following condition. The sampling frequency is set to 12 kHz. The frame length was 42.6 ms (512 samples) with the frame shift of 21.3 ms, and the window function was Hamming. Other implementation parameters were experimentally optimized in the simulated acoustic conditions.

4.3 Performance Evaluation in Simulated Conditions

In the first experiment, we optimize the implementation parameters (e.g., ε and *bandwidth*) used in our proposed noise reduction algorithm and evaluate the effectiveness of the studied algorithms in the simulated anechoic room with the configuration shown in Fig. 2. In this simulated condition, both target signals and interfering signals were selected from TIMIT speech database [22]. For target signal, we chose 200 utterances spoken by 20 speakers, and for three interfering signals, different 600 utterances spoken by 60 speakers (20 speakers with 200 utterance for each interfering source). The simulated target and interfering signals were first re-sampled to 12 kHz and then mixed to obtain the observed noisy mixture signals.

4.3.1 Parameter Optimization

As described in Sect. 3, the proposed two-microphone noise reduction algorithm involves several implementation parameters, that is, ε in Eq. (9), *bandwidth* of subband in frequency dividing, α and β in Eq. (13). Since the noise spectrum is analytically estimated, it is expected to give much more accurate noise spectral estimate. Therefore, the parameters α is set to 1.0 and β is empirically set to a small value 0.001. In this subsection, we thus optimize the parameters: *bandwith* of each subband and ε in Eq. (9).

To optimize the parameters *bandwidth* and ε, half data set (100 utterances for target signals and 300 utterances for interfering signals) was used and the other half data set was exploited for evaluating the effectiveness of the tested algorithms in the one- and three-noise-source conditions. In the optimization procedure, the objective measure PESQ is calculated for each utterance, and then averaged across all utterances and all SNR levels. The optimized *bandwidth* and ε are derived as the ones which result in the highest average PESQ results.

The averaged PESQ results as a function of *bandwidth* in the one- and three-noise-source environments are plotted in Fig. 3. As Fig. 3 shows, the PESQ results demonstrate a high degree of dependence on the *bandwidth* of subband in two noise conditions. The highest PESQ results are achieved when the *bandwith* is set to 2 samples, corresponding to approximately 47 Hz, which is used in the following evaluations.

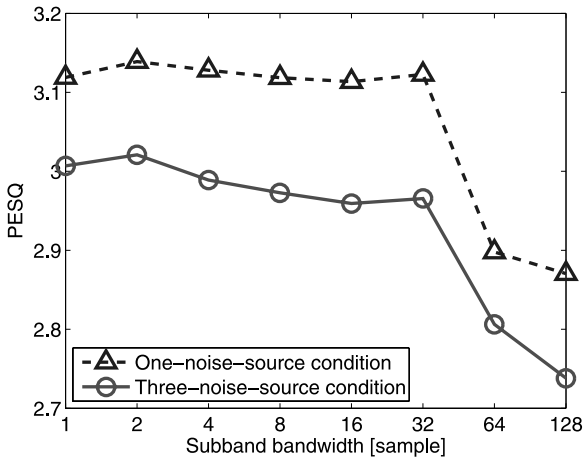

Fig. 3 Average PESQ results as a function of different *bandwidth* in the one- and three-noise-source conditions.

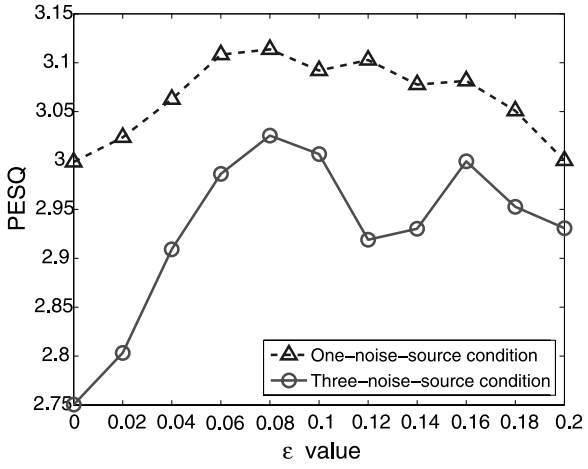

Fig. 4 Average PESQ results as a function of different ε in the one- and three-noise-source conditions.

The averaged PESQ results as a function of ε in the one- and three-noise-source environments are plotted in Fig. 4. From these results, we can see that the ε of 0.08 leads to the highest average PESQ results in both one- and three-noise-source conditions. In the later experiments, therefore, the ε is set to 0.08.

As a result, the following evaluations were done with the following optimized implementation parameters. The factors α and β for spectral subtraction were set to 1.0 and 0.001, respectively. The bandwidth in the proposed algorithm was set to 47 Hz.

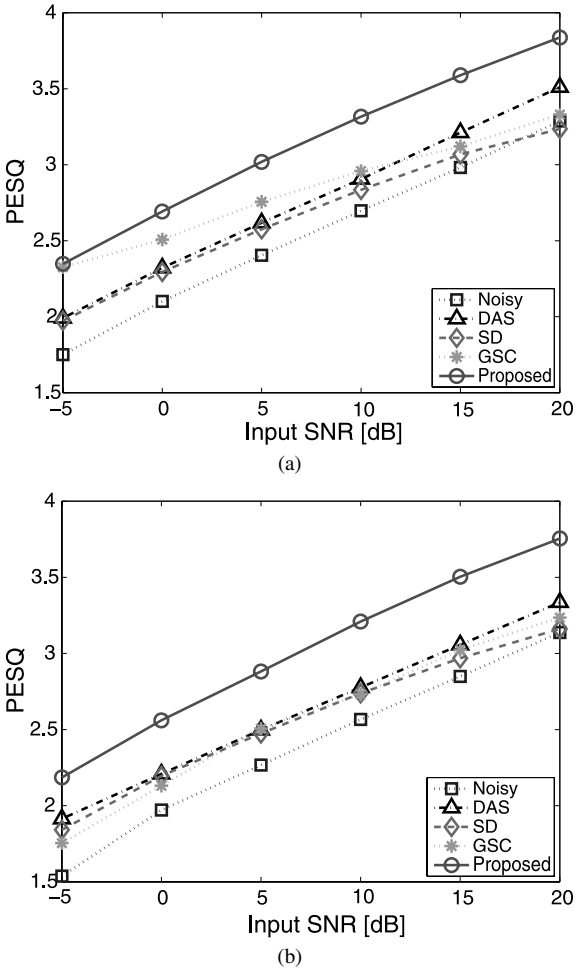

Fig. 5 Perceptual evaluation of speech quality (PESQ) of the noisy input (□), the delay-and-sum (DAS) beamformer (△), the superdirective (SD) beamformer (◊), the standard GSC beamformer (∗) and the proposed algorithm (○), in the simulated one-noise-source acoustic condition (a) and the simulated three-noise-source acoustic condition (b).

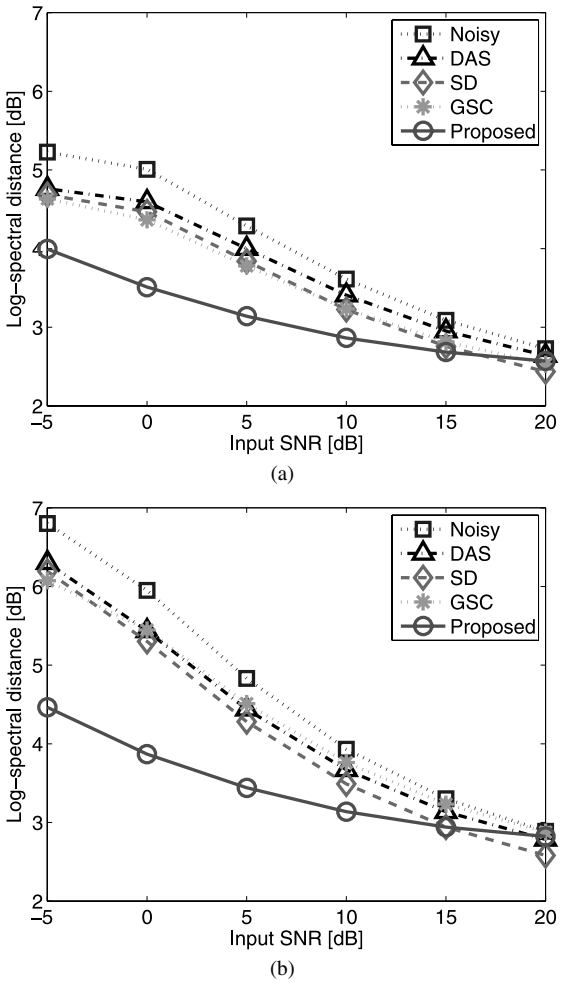

Fig. 6 Log-spectral distance (LSD) of the noisy input (□), the delay-and-sum (DAS) beamformer (△), the superdirective (SD) beamformer (◊), the standard GSC beamformer (∗) and the proposed algorithm (○), in the simulated one-noise-source acoustic condition (a) and the simulated three-noise-source acoustic condition (b).

4.3.2 Evaluation Results

The averaged results of PESQ and LSD are shown in Figs. 5 and 6. The performance was evaluated at the first microphone, the traditional noise reduction algorithms (i.e., DAS, SD and GSC) output and the proposed algorithm output. The average PESQ results across remaining 100 utterances in two simulated conditions, shown in Fig. 5, indicates that our proposed noise reduction algorithm produces the highest PESQ results, corresponding to the highest quality of enhanced signal, in comparison of the tested traditional algorithms in both the simulated one-noise-source condition and three-noise-source condition.

The average LSD results are plotted in Fig. 6. From Fig. 6, we can observe that the traditional noise reduction algorithms provide the relatively small degree of LSD decrease. Among the tested noise reduction algorithms, our proposed algorithm leads to the lowest LSDs especially in the low SNR conditions. This further indicates that the proposed algorithm produces the lowest speech distortion compared with the traditional algorithms.

4.4 Performance Evaluation in Real Environments

To evaluate the performance of the studied algorithms in real acoustic environments, we used the recorded sound data for target and interfering signals. Both target signals and interfering signals were selected from ATR speech database [23]. Specifically, eight Japanese sentences uttered by one female and one male were used as target signals. Other different twelve Japanese sentences uttered by one female and two males were used as interfering noise signals. The durations of these utterances were ranged from about 5 s to 10 s. In our experiments, target signals were played back through the loudspeaker *speech*, and interfering signals were played back through the loudspeakers *noise1, noise2, noise3*, as shown in Fig. 2. To facilitate the following objective evaluations, either target signal or interfering signal was recorded

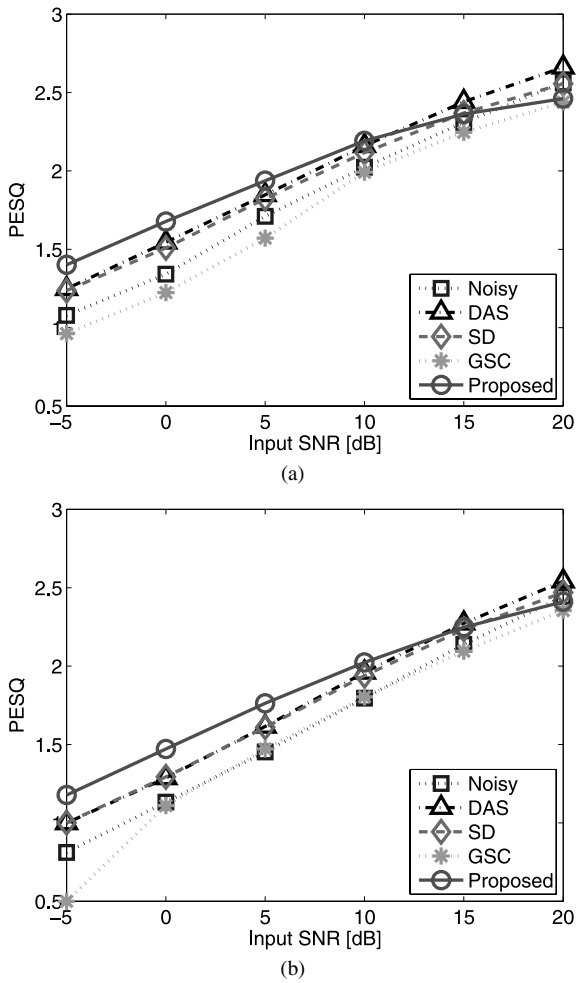

Fig. 7 Perceptual evaluation of speech quality (PESQ) of the noisy input (□), the delay-and-sum (DAS) beamformer (△), the superdirective (SD) beamformer (◊), the standard GSC beamformer (∗) and the proposed algorithm (○), in real one-noise-source acoustic condition (a) and real three-noise-source acoustic condition (b).

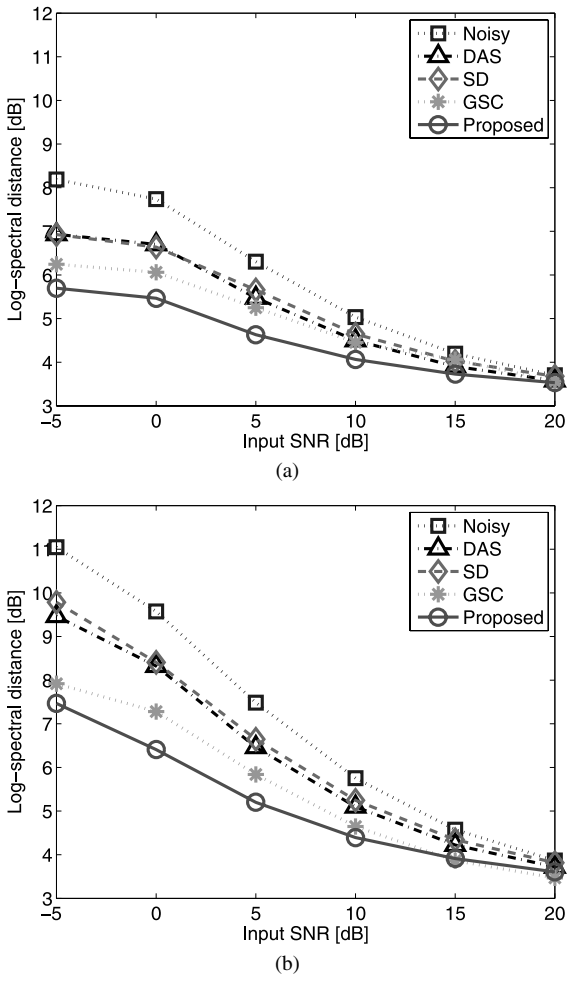

Fig. 8 Log spectral distance (LSD) of the noisy input (□), the delay-and-sum (DAS) beamformer (△), the superdirective (SD) beamformer (◊), the standard GSC beamformer (∗) and the proposed algorithm (○), in real one-noise-source acoustic condition (a) and real three-noise-source acoustic condition (b).

by two microphones separately with the sampling frequency of 48 kHz at 16 bit accuracy. Thus, the target utterance and the interfering utterances might not start simultaneously. The target and interfering signals were first re-sampled to 12 kHz and then mixed to obtain the observed noisy mixture signals.

4.4.1 Evaluation Results

The experimental results of PESQ and LSD are shown in Fig. 7 and Fig. 8, respectively. As shown in Fig. 7, all studied noise reduction algorithms provide consistent PESQ improvements compared to the noisy inputs in both one-noise-source and three-noise-source conditions. Furthermore, with respect to the traditional DAS, SD and GSC, the proposed two-microphone noise reduction algorithm offers much higher PESQ improvements in both one- and three-noise-source conditions, especially at the low SNR levels. The highest PESQ improvements indicate the proposed algorithm offers the enhanced signal with the highest speech quality.

Concerning the results of LSD, shown in Fig. 8, it is seen that all tested noise reduction algorithms show some degree of LSD decreases, especially in low SNR conditions. Compared to the traditional algorithms (i.e., DAS, SD and GSC), the proposed two-microphone noise reduction algorithm gives the markedly decreased LSD in all tested noise conditions at all SNR levels. The lowest LSDs achieved by the proposed algorithm reveal that the proposed algorithm involves the lowest speech distortion with respect to the tested traditional algorithms. Moreover, it also can be seen that the performance improvements in terms of LSD decreases as the input signals become "clean."

4.4.2 Subjective Evaluations

All the tested noise reduction algorithms were also assessed with the listening tests. Four sentences were selected and used to evaluate the tested algorithms at three SNR levels

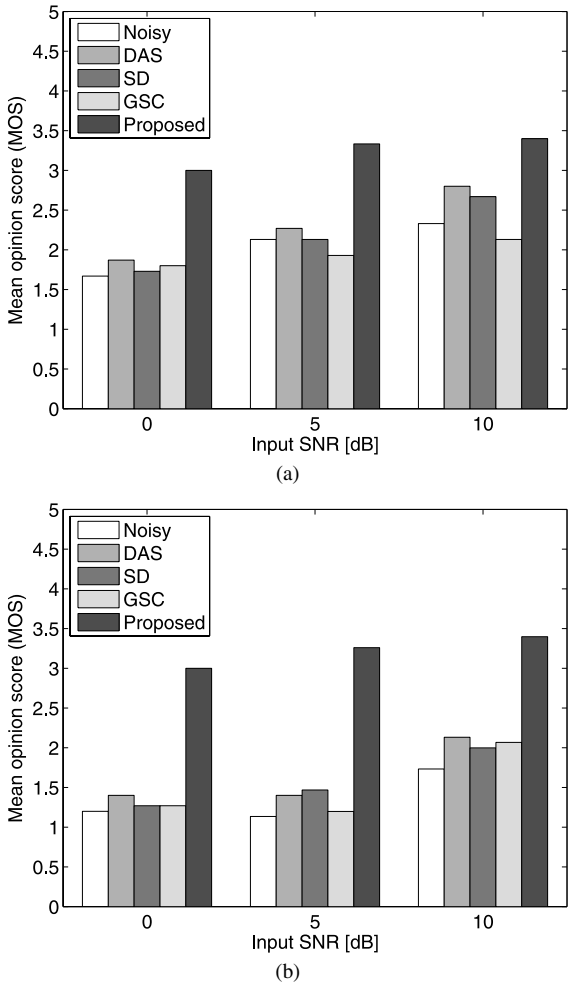

Fig. 9 Mean opinion score (MOS) of the noisy input, the delay-and-sum (DAS) beamformer, the superdirective (SD) beamformer, the standard GSC beamformer and the proposed algorithm, in one-noise-source acoustic condition (a) and three-noise-source acoustic condition (b).

(0 dB, 5 dB and 10 dB) in the one-noise-source real condition and the three-noise-source real condition. The resulting $24(4 \times 3 \times 2)$ noisy speech sentences were then processed by the four algorithms: DAS, SD, GSC and the proposed algorithm. Eight graduate students with normal hearing attended the listening tests. The tested speech materials were randomly presented to each listener through a headphone at a comfortable loudness level in a sound-proof room. The listeners were instructed to rate the quality of the enhanced output signals based on their preference in terms of *mean opinion score* (MOS): 1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent.

The MOS results, plotted in Fig. 9, demonstrate that all studied noise reduction algorithms result in higher MOS rates compared with the noisy input signals in all tested conditions except at 5 and 10 dB in the one-noise-source real conditions where the GSC beamformer shows slightly lower MOS values. The achieved higher MOS rates indicate that all the studied algorithms produce the enhanced signal of the higher speech quality which is preferred by the listeners. Furthermore, among the tested algorithms, the proposed noise reduction algorithm results in the significantly improved and highest MOS rates in all conditions, corresponding to the "cleanest" output signal with the highest quality. Moreover, comparing the MOS rates in the one- and three-noise-source real conditions, we can find that the performance of the traditional algorithms (DAS, SD and GSC) markedly degrades when the number of interfering signals increases from one to three, while the MOS results of the proposed algorithm are only slightly different in the one- and three-noise-source real conditions. That is, the proposed noise reduction algorithms is successful in reducing the multiple interfering signals even if for the highly non-stationary interference (e.g., speech signal as used in our experiments).

5. Discussions

From the experimental results presented in the last section, the superiorities of the proposed noise reduction method to the other traditional methods are discussed in the following paragraphs.

The proposed method outperforms the DAS beamformer. For the DAS beamformer, the acceptable noise reduction performance can be obtained only when a number of microphones are available. In contrast, the proposed method exploits only two microphones and still achieves the high noise reduction performance. This superiority can be attributed to the low directivity of DAS beamformer and the highly accurate noise estimation and high noise reduction performance of the proposed method.

The proposed method outperforms the SD beamformer. The SD beamformer was implemented with an assumption of a diffuse noise field. However, the noise field in which multiple noise sources are present cannot be regarded as a diffuse noise field. This inconsistence results in the low performance of the SD beamformer. On the other hand, the proposed method estimates the spectrum for the multiple noises dependent on the virtual DOA in each subband, which leads to high noise-estimation accuracy and further high noise reduction performance in such adverse conditions. This superiority can be attributed to the diffuse noise field assumption of the SD beamformer and the analytical noise estimation approach of the proposed noise reduction method.

The proposed method outperforms the GSC beamformer. The traditional GSC beamformer suffers from the low noise reduction performance in dealing with the non-stationary noises, e.g., the interfering speech signals, due to the utilization of adaptive signal processing technique (e.g., LMS). In contrast, the proposed method is able to reduce the non-stationary noises by analytically estimating the spectrum of the non-stationary interfering noises and then subtracting it from that of the noisy observations. The use of the analytical scheme instead of the adaptive signal processing technique provides the proposed noise reduction method the ability in dealing with the non-stationary noises. Furthermore, the performance of the GSC beamformer significantly decreases when the number of noise sources is larger

than that of the microphones. While, it is possible for the proposed method to suppress the multiple interfering noises because of the use of the subband signal processing.

Additionally, the proposed two-microphone noise reduction algorithm should be effective in suppressing the interfering signals even when more than three noises exist and the interfering noises are changed. This is because that the noises can still be considered as one integrated virtual noise signal in each subband when the number of noises is more than three, as derived in Appendix. Moreover, the integration of multiple interfering signals is regardless of their "real" positions, e.g., for the case where the positions of the interfering signals change.

As a result, the proposed noise reduction method provides the highest performance (e.g., suppressing the noise signals as much as possible while keeping speech distortionless) among the studied noise reduction algorithms under all tested experimental conditions, as shown in Sects. 4.3.2, 4.4.1 and 4.4.2.

6. Conclusions

This paper presented a two-microphone noise reduction method to reduce highly non-stationary multiple interfering noises. The proposed method analytically estimates the spectrum of interfering signal in each subband upon its virtual DOA and a single-channel estimation approach, and subtracts the spectral estimate from that of the noisy observation, enhancing the target signal. To determine the DOA of interfering signal, we developed a noise direction estimator using two microphones that shows good performance even when speech is present. The effectiveness and superiority of the proposed noise reduction algorithm were confirmed by experiments in real room acoustic environments. The small physical size, computational efficiency and practical effectiveness of the proposed algorithm might be the preferable points for many applications, such as, hearing aids.

Though the proposed two-microphone noise reduction algorithm has in this paper been proven to be effective in dealing with highly non-stationary multiple-source interfering signals, its performance will be degraded in the presence of reverberation in practical conditions, especially when the reverberation time is long. Therefore, in the future work along this research, we will study on further improving this two-microphone noise reduction algorithm by integrating a dereverberation technique to cope with the reverberation effects in real-world environments, finally building up an integrated noise-reduction/dereverberation algorithm for real-world applications.

Acknowledgement

This study was supported by Sendai Intelligent Knowledge Cluster and the Grant-in-Aid for Young Scientists (B) (No. 19700156) from the Ministry of Education, Science, Sports and Culture of Japan.

References

[1] M.S. Brandstein and D.B. Ward, eds., Microphone Arrays: Signal Processing Techniques and Applications, Springer-Verlag, Berlin, 2001.
[2] J. Chen, L. Shue, K. Phua, and H. Sun, "Theoretical comparison of dual microphone systems," Proc. ICASSP2004, pp. IV-73–75, 2004.
[3] D.H. Jonhson and D.E. Dungeon, Array Signal Processing: Concepts and Techniques, Prentice-Hall, 1993.
[4] H. Cox, R.M. Zeskind, and M.M. Owen, "Robust adaptive beamforming," IEEE Trans. Acoust. Speech Signal Process., vol.35, no.10, pp.1365–1375, Oct. 1987.
[5] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," IEEE Trans. Antennas Propag., vol.AP-30, no.1, pp.27–34, 1982.
[6] J. Bitzer, K.U. Simmer, and K.D. Kammeyer, "Multichannel noise reduction — Algorithms and theoretical limits," Proc. EUSIPCO1998, pp.105–108, 1998.
[7] H.Y. Kim, F. Asano, Y. Suzuki, and T. Sone, "Speech enhancement based on short-time spectral amplitude estimation with two-channel beamformer," IEICE Trans. Fundamentals, vol.E79-A, no.12, pp.2151–2158, Dec. 1996.
[8] Y. Takahashi, T. Takatani, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array with indepdent component analysis for hands-free speech recognition," Proc. IWAENC2006, 2006.
[9] M. Akagi and T. Kago, "Noise reduction using a small-scale microphone array in multi noise source environment," Proc. ICASSP2002, pp.I909–912, 2002.
[10] G.C. Carter, "Coherence and time delay estimation," Proc. IEEE, vol.75, no.2, pp.236–255, 1987.
[11] M. Mizumachi and M. Akagi, "Noise reduction method that is equipped for a robust direction finder in adverse environments," Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions, pp.179–182, 1999.
[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoust. Speech Signal Process., vol.ASSP-32, no.6, pp.1109–1121, 1984.
[13] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," Proc. ICASSP1979, pp.208–211, 1979.
[14] M. Omologo and P. Svaizer, "Acoustic source localization in noisy and reverberant environment using CSP analysis," Proc. ICASSP1996, pp.921–924, 1996.
[15] J. Li and M. Akagi, "Noise reduction using hybrid noise estimation techniques and post-filtering," Proc. ICSLP2004, pp.2705–2708, 2004.
[16] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," Signal Process., vol.81, no.11, pp.2403–2418, 2001.
[17] ITU-T Recommendation P.862, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of 3.1 kHz handset telephony (narrow-band) networks and speech codecs," Feb. 2001.
[18] S.R. Quackenbush, T.P. Barnwell, and M.A. Clements, Objective Measures of Speech Quality, Prentice-Hall, Englewood Cliffs, NJ, 1988.
[19] J.H.L. Hansen and B. Pellom, "An effective quality evaluation protocal for speech enhancement algorithms," Proc. Int. Conf. on Spoken Language Processing, vol.7, pp.2819–2822, 1998.
[20] I. Cohen, "Multi-channel post-filtering in non-stationary noise environments," IEEE Trans. Signal Process., vol.52, no.5, pp.1149–1160, 2004.
[21] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am., vol.65, no.4, pp.943–950, 1979.

[22] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "Timit acoustic-phonetic continous speech corpus," NTIS order number PB91-100354, 1993.
[23] K. Takeda, Y. Sagisaka, S. Katagiri, M. Abe, and H. Kuwabara, "Speech database user's manual," ATR Technical Report, TR-I-0028, 1988.

Appendix

In this appendix, we give the theoretical evidences for the assumption, the sum of sinusoidal waves becomes a sinusoidal wave in a narrow subband, used in the procedure of virtual noise direction estimation.

Let us assume two sinusoidal waves, $y_1(\tilde{\omega})$ and $y_2(\tilde{\omega})$, are given by

$$y_1(\tilde{\omega}) = A_1 \sin(\tilde{\omega}t + \theta_1); \quad (A\cdot 1)$$

$$y_2(\tilde{\omega}) = A_2 \sin(\tilde{\omega}t + \theta_2), \quad (A\cdot 2)$$

where A_1 and θ_1 denote the amplitude and phase of $y_1(\tilde{\omega})$ at the frequency $\tilde{\omega}$, $\omega_{k-1} \le |\tilde{\omega}| < \omega_k$, $k = 1, 2, \ldots, K$; and A_2 and θ_2 are those of $y_2(\tilde{\omega})$.

Thus, the sum of these two sinusoidal waves will be

$$\begin{aligned} y_1(\tilde{\omega}) + y_2(\tilde{\omega}) &= A_1 \sin(\tilde{\omega}t + \theta_1) + A_2 \sin(\tilde{\omega}t + \theta_2) \\ &= A_1\big(\sin(\tilde{\omega}t)\cos\theta_1 + \cos(\tilde{\omega}t)\sin\theta_1\big) \\ &\quad + A_2\big(\sin(\tilde{\omega}t)\cos\theta_2 + \cos(\tilde{\omega}t)\sin\theta_2\big) \\ &= \sin(\tilde{\omega}t)\big(A_1\cos\theta_1 + A_2\cos\theta_2\big) \\ &\quad + \cos(\tilde{\omega}t)\big(A_1\sin\theta_1 + A_2\sin\theta_2\big) \\ &= A\sin(\tilde{\omega}t) + B\cos(\tilde{\omega}t) \\ &= \sqrt{A^2 + B^2}\sin\left(\tilde{\omega}t + \arctan\left(\frac{B}{A}\right)\right), \end{aligned} \quad (A\cdot 3)$$

where A and B are defined as

$$A = A_1\cos\theta_1 + A_2\cos\theta_2; \quad (A\cdot 4)$$

$$B = A_1\sin\theta_1 + A_2\sin\theta_2. \quad (A\cdot 5)$$

Here, we can see that the amplitude and phase of the summed signal are given by $\sqrt{A^2 + B^2}$ and $\arctan\left(\frac{B}{A}\right)$, respectively. As a result, the sum of two sinusoidal wave signals has been proven to be still a sinusoidal wave signal. Following the similar deviation procedure, this result can be further generalized as: the sum of several sinusoidal waves becomes a sinusoidal wave signal in a narrow sunbband, which was assumed in our method.

Junfeng Li received the B.E. degree from Zhengzhou University and the M.S. degree from Xidian University, China, both in Computer Science, in 2000 and 2003, respectively. He received the Ph.D. degree in Information Science from Japan Advanced Institute of Science and Technology (JAIST) in March 2006. From April 2006 to March 2007, he was a post-doctoral researcher at Research Institute of Electrical Communication, Tohoku University. Since April, 2007, he has been an Assistant Professor in School of Information Science, JAIST. His research interests include speech signal processing and intelligent hearing aids. Dr. Li received the Best Student Award in Engineering Acoustics First Prize from the Acoustical Society of America in 2006, and the Best Paper Award from the JCA2007 in 2007.

Masato Akagi received the B.E. degree in Electronic Engineering from Nagoya Institute of Technology in 1979, the M.E. and the 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 Graduate 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, speech signal processing. Dr. Akagi received the IEICE Excellent Paper Award from the IEICE in 1987, and the Sato Prize for Outstanding Paper from the ASJ in 1998 and 2006. Dr. Akagi is now a vice President of the Acoustical Society of Japan.

Yôiti Suzuki graduated from Tohoku University, Sendai, Japan in 1976 from which he also received the Ph.D. degree in electrical and communication engineering in 1981, He is currently a Professor with the Research Institute of Electrical Communication, Tohoku University. His research interests include psychoacoustics, high-definition auditory display, and digital signal processing of acoustic signals. Dr. Suzuki received the Takenaka and RCA David Sarnoff Scholarships, the Awaya Kiyoshi Award, and the Sato Prize from the Acoustical Society of Japan. Dr. Suzuki served as a vice President of the Acoustical Society of Japan and the society's Editor-in-Chief from 2001 to 2003, and the President of the Acoustical Society of Japan from 2005 to 2007.