3

Blind Source Separation for Speech Application Under Real Acoustic Environment

Hiroshi Saruwatari and Yu Takahashi
Nara Institute of Science and Technology
Japan

1. Introduction

A hands-free speech recognition system [1] is essential for the realization of an intuitive, unconstrained, and stress-free human-machine interface, where users can talk naturally because they require no microphone in their hands. In this system, however, since noise and reverberation always degrade speech quality, it is difficult to achieve high recognition performance, compared with the case of using a close-talk microphone such as a headset microphone. Therefore, we must suppress interference sounds to realize a noise-robust hands-free speech recognition system.

Source separation is one approach to removing interference sound source signals. Source separation for acoustic signals involves the estimation of original sound source signals from mixed signals observed in each input channel. Various methods have been presented for acoustic source signal separation. They can be classified into two groups: methods based on single-channel input, e.g., spectral subtraction (SS) [2], and those based on multichannel input, e.g., microphone array signal processing [3]. There have been various studies on microphone array signal processing; in particular, the delay-and-sum (DS) [4–6] array and adaptive beamformer (ABF) [7–9] are the most conventionally used microphone arrays for source separation and noise reduction. ABF can achieve higher performance than the DS array. However, ABF requires a priori information, e.g., the look direction and speech break interval. These requirements are due to the fact that conventional ABF is based on *supervised* adaptive filtering, which significantly limits its applicability to source separation in practical applications. Indeed, ABF cannot work well when the interfering signal is nonstationary noise.

Recently, alternative approaches have been proposed. Blind source separation (BSS) is an approach to estimating original source signals using only mixed signals observed in each input channel. In particular, BSS based on independent component analysis (ICA) [10], in which the independence among source signals is mainly used for the separation, has recently been studied actively [11–19]. Indeed, the conventional ICA could work, particularly in speech-speech mixing, i.e., all sources can be regarded as point sources, but such a mixing condition is very rare and unrealistic; real noises are often widespread sources. In this chapter, we mainly deal with generalized noise that cannot be regarded as a point source. Moreover, we assume this noise to be nonstationary noise that arises in many acoustical environments; however, ABF could not treat this noise well. Although ICA is not influenced by the nonstationarity of signals unlike ABF, this is still a very challenging task that can hardly be addressed by conventional ICA-based BSS because ICA cannot separate widespread sources.

To improve the performance of BSS, some techniques combining conventional ICA and beamforming have been proposed [18, 20]. However, these studies dealt with the separation of point sources, and the behavior of such methods under a non-point-source condition was not explicitly analyzed to our knowledge. Therefore, in this chapter, first, we analyze ICA under a non-point-source noise condition and point out that ICA is proficient in noise estimation rather than in speech estimation under such a noise condition. This analysis implies that we can still utilize ICA as an accurate noise estimator.

Next, we review blind spatial subtraction array (BSSA) [21], an improved BSS algorithm recently proposed in order to deal with real acoustic sounds. BSSA consists of an ICA-based noise estimator, and noise reduction in the proposed BSSA is achieved by subtracting the power spectrum of the estimated noise via ICA from the power spectrum of the noisy observations. This "power-spectrum-domain subtraction" procedure provides better noise reduction than conventional ICA with estimation-error robustness. The efficacy of BSSA can be determined in various experiments, including computer-simulation-based and real-recording-based experiments. This chapter shows strong evidence of BSSA providing promising speech enhancement results in a railway-station environment.

Finally, the real-time implementation issue of BSS is discussed. Several recent studies have dealt with the real-time implementation of ICA, but they still required high-speed personal computers. Consequently, BSS implementation on a small LSI still receives much attention in industrial applications. In this chapter, an example of hardware implementation of BSSA is introduced, which has yielded commercially available microphones adopted by the Japanese National Police Agency.

The rest of this chapter is organized as follows. In Sect. 2, the sound mixing model and conventional ICA are discussed. In Sect. 3, the analysis of ICA under a non-point-source condition is described in detail. In Sect. 4, BSSA is reviewed in detail. In Sect. 5, the experimental results are shown and compared with those of conventional methods. In Sect. 6, an example of hardware implementation of BSSA is introduced. Following the example, the chaper conclusions are given in Sect. 7.

2. Data model and conventional BSS method

2.1 Sound mixing model of microphone array

In this chapter, a straight-line array is assumed. The coordinates of the elements are designated $d_j (j = 1, \ldots, J)$, and the direction-of-arrivals (DOAs) of multiple sound sources are designated $\theta_k (k = 1, \ldots, K)$ (see Fig. 1). Then, we consider that only one target speech signal, some interference signals that can be regarded as point sources, and additive noise exist. This additive noise represents noises that cannot be regarded as point sources, e.g., spatially uncorrelated noises, background noises, and leakage of reverberation components outside the frame analysis. Multiple mixed signals are observed at microphone array elements, and a short-time analysis of the observed signals is conducted by frame-by-frame discrete Fourier transform (DFT). The observed signals are given by

$$\boldsymbol{x}(f,\tau) = \boldsymbol{A}(f)\left\{\boldsymbol{s}(f,\tau) + \boldsymbol{n}(f,\tau)\right\} + \boldsymbol{n}_a(f,\tau), \tag{1}$$

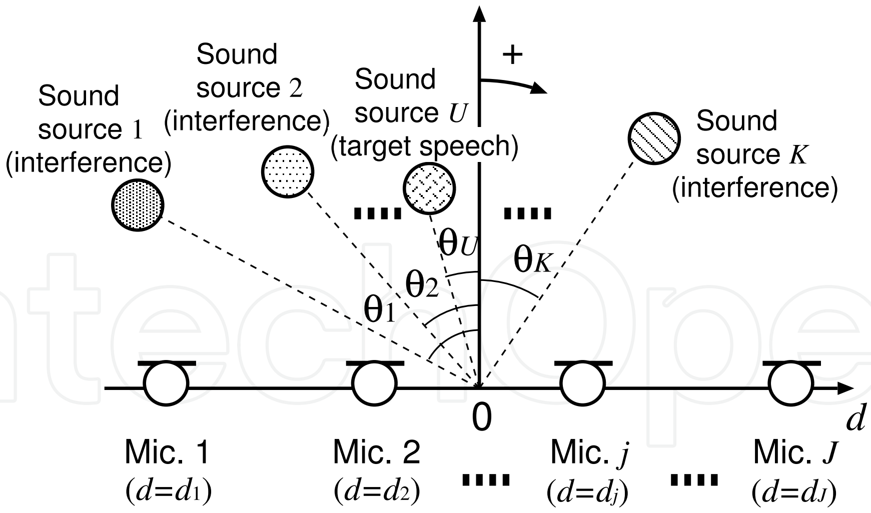
Sound source 1 (interference)
Sound source 2 (interference)
Sound source U (target speech)
Sound source K (interference)
+
θ1
θ2
θU
θK
0
d
Mic. 1 (d=d_1)
Mic. 2 (d=d_2)
Mic. j (d=d_j)
Mic. J (d=d_J)

Fig. 1. Configurations of microphone array and signals.

where f is the frequency bin and τ is the time index of DFT analysis. Also, $\boldsymbol{x}(f,\tau)$ is the observed signal vector, $\boldsymbol{A}(f)$ is the mixing matrix, $\boldsymbol{s}(f,\tau)$ is the target speech signal vector in which only the Uth entry contains the signal component $s_U(f,\tau)$ (U is the target source number), $\boldsymbol{n}(f,\tau)$ is the interference signal vector that contains the signal components except the Uth component, and $\boldsymbol{n}_a(f,\tau)$ is the nonstationary additive noise signal term that generally represents non-point-source noises. These are defined as

$$\boldsymbol{x}(f,\tau) = [x_1(f,\tau),\ldots,x_J(f,\tau)]^{\mathrm{T}}, \tag{2}$$

$$\boldsymbol{s}(f,\tau) = [\underbrace{0,\ldots,0}_{U-1}, s_U(f,\tau), \underbrace{0,\ldots,0}_{K-U}]^{\mathrm{T}}, \tag{3}$$

$$\boldsymbol{n}(f,\tau) = [n_1(f,\tau),\ldots,n_{U-1}(f,\tau),0,n_{U+1},\ldots,n_K(f,\tau)]^{\mathrm{T}}, \tag{4}$$

$$\boldsymbol{n}_a(f,\tau) = [n_1^{(a)}(f,\tau),\ldots,n_J^{(a)}(f,\tau)]^{\mathrm{T}}, \tag{5}$$

$$\boldsymbol{A}(f) = \begin{bmatrix} A_{11}(f) & \cdots & A_{1K}(f) \\ \vdots & & \vdots \\ A_{J1}(f) & \cdots & A_{JK}(f) \end{bmatrix}. \tag{6}$$

2.2 Conventional frequency-domain ICA

Here, we consider a case where the number of sound sources, K, equals the number of microphones, J, i.e., $J = K$. In addition, similarly to that in the case of the conventional ICA contexts, we assume that the additive noise $\boldsymbol{n}_a(f,\tau)$ is negligible in (1). In frequency-domain ICA (FDICA), signal separation is expressed as

$$\boldsymbol{o}(f,\tau) = [o_1(f,\tau),\ldots,o_K(f,\tau)]^{\mathrm{T}} = \boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{x}(f,\tau), \tag{7}$$

$$\boldsymbol{W}_{\mathrm{ICA}}(f) = \begin{bmatrix} W_{11}^{(\mathrm{ICA})}(f) & \cdots & W_{1J}^{(\mathrm{ICA})}(f) \\ \vdots & & \vdots \\ W_{K1}^{(\mathrm{ICA})}(f) & \cdots & W_{KJ}^{(\mathrm{ICA})}(f) \end{bmatrix}, \tag{8}$$

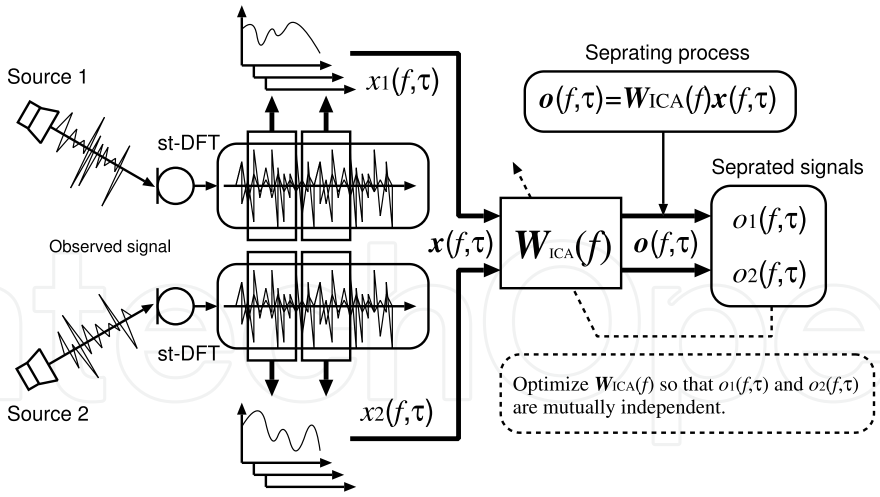

Fig. 2. Blind source separation procedure in FDICA in case of $J = K = 2$.

where $\boldsymbol{o}(f,\tau)$ is the resultant output of the separation and $\boldsymbol{W}_{\mathrm{ICA}}(f)$ is the complex-valued unmixing matrix (see Fig. 2).

The unmixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$ is optimized by ICA so that the output entries of $\boldsymbol{o}(f,\tau)$ become mutually independent. Indeed, many kinds of ICA algorithm have been proposed. In the second-order ICA (SO-ICA) [15, 17], the separation filter is optimized by the joint diagonalization of co-spectra matrices using the nonstationarity and coloration of the signal. For instance, the following iterative updating equation based on SO-ICA has been proposed by Parra and Spence [15]:

$$\boldsymbol{W}_{\mathrm{ICA}}^{[p+1]}(f) = -\mu \sum_{\tau_b} \chi(f)\,\text{off-diag}\left(\boldsymbol{R}_{oo}\left(f,\tau_b\right)\right)\boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f)\boldsymbol{R}_{xx}(f,\tau_b) + \boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f), \tag{9}$$

where μ is the step-size parameter, $[p]$ is used to express the value of the pth step in iterations, off-diag$[\boldsymbol{X}]$ is the operation for setting every diagonal element of matrix $\boldsymbol{X}$ to zero, and $\chi(f) = (\sum_{\tau_b} \|\boldsymbol{R}_{xx}(f,\tau_b)\|^2)^{-1}$ is a normalization factor ($\|\cdot\|$ represents the Frobenius norm). $\boldsymbol{R}_{xx}(f,\tau_b)$ and $\boldsymbol{R}_{oo}(f,\tau_b)$ are the cross-power spectra of the input $\boldsymbol{x}(f,\tau)$ and output $\boldsymbol{o}(f,\tau)$, respectively, which are calculated around multiple time blocks τ_b. Also, Pham et al. have proposed the following improved criterion for SO-ICA [17]:

$$\sum_{\tau_b}\left\{\frac{1}{2}\log\det\mathrm{diag}[\boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{R}_{oo}(f,\tau_b)\boldsymbol{W}_{\mathrm{ICA}}(f)^{\mathrm{H}}] - \log\det[\boldsymbol{W}_{\mathrm{ICA}}(f)]\right\}, \tag{10}$$

where the superscript H denotes Hermitian transposition. This criterion is to be minimized with respect to $\boldsymbol{W}_{\mathrm{ICA}}(f)$.

On the other hand, a higher-order-statistics-based approach exists. In higher-order ICA (HO-ICA), the separation filter is optimized on the basis of the non-Gaussianity of the signal. The optimal $\boldsymbol{W}_{\mathrm{ICA}}(f)$ in HO-ICA is obtained using the iterative equation

$$\boldsymbol{W}_{\mathrm{ICA}}^{[p+1]}(f) = \mu[\boldsymbol{I} - \langle\boldsymbol{\varphi}(\boldsymbol{o}(f,\tau))\boldsymbol{o}^{\mathrm{H}}(f,\tau)\rangle_{\tau}]\boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f) + \boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f), \tag{11}$$

where $\boldsymbol{I}$ is the identity matrix, $\langle\cdot\rangle_{\tau}$ denotes the time-averaging operator, and $\boldsymbol{\varphi}(\cdot)$ is the nonlinear vector function. Many kinds of nonlinear function $\boldsymbol{\varphi}(f,\tau)$ have been proposed.

Considering a batch algorithm of ICA, it is well-known that $\tanh(\cdot)$ or the sigmoid function is appropriate for super-Gaussian sources such as speech signals [22]. In this study, we define the nonlinear vector function $\boldsymbol{\varphi}(\cdot)$ as

$$\boldsymbol{\varphi}(\boldsymbol{o}(f,\tau)) \equiv [\varphi(o_1(f,\tau)), \ldots, \varphi(o_K(f,\tau))]^{\mathrm{T}}, \tag{12}$$

$$\varphi(o_k(f,\tau)) \equiv \tanh o_k^{(\mathrm{R})}(f,\tau) + i \tanh o_k^{(\mathrm{I})}(f,\tau), \tag{13}$$

where the superscripts (R) and (I) denote the real and imaginary parts, respectively. The nonlinear function given by (12) indicates that the nonlinearity is applied to the real and imaginary parts of complex-valued signals separately. This type of complex-valued nonlinear function has been introduced by Smaragdis [14] for FDICA, where it can be assumed for speech signals that the real (or imaginary) parts of the time-frequency representations of sources are mutually independent. According to Refs. [19, 23], the source separation performance of HO-ICA is almost the same as or superior to that of SO-ICA. Thus, in this chapter, HO-ICA is utilized as the basic ICA algorithm in the simulation (Sect. 3.4) and experiments (Sect. 5).

3. Analysis of ICA under non-point-source noise condition

In this section, we investigate the proficiency of ICA under a non-point-source noise condition. In relation to the performance analysis of ICA, Araki et al. have reported that ICA-based BSS has equivalence to parallel constructed ABFs [24]. However, this investigation was focused on separation with a nonsingular mixing matrix, and thus was valid for only point sources.

First, we analyze beamformers that are optimized by ICA under a non-point-source condition. In the analysis, it is clarified that beamformers optimized by ICA become specific beamformers that maximize the signal-to-noise ratio (SNR) in each output (so-called *SNR-maximize beamformers*). In particular, the beamformer for target speech estimation is optimized to be a DS beamformer, and the beamformer for noise estimation is likely to be a null beamformer (NBF) [16].

Next, a computer simulation is conducted. Its result also indicates that ICA is proficient in noise estimation under a non-point-source noise condition. Then, it is concluded that ICA is suitable for noise estimation under such a condition.

3.1 Can ICA separate any source signals?

Many previous studies on BSS provided strong evidence that conventional ICA could perform source separation, particularly in the special case of speech-speech mixing, i.e., all sound sources are point sources. However, such sound mixing is not realistic under common acoustic conditions; indeed the following scenario and problem are likely to arise (see Fig. 3):

- The target sound is the user's speech, which can be approximately regarded as a *point source*. In addition, the users themselves locate relatively *near the microphone array* (e.g., 1 m apart), and consequently the accompanying reflection and reverberation components are moderate.
- For the noise, we are often confronted with interference sound(s) which is *not a point source* but a widespread source. Also, the noise is usually far from the array and is heavily reverberant.

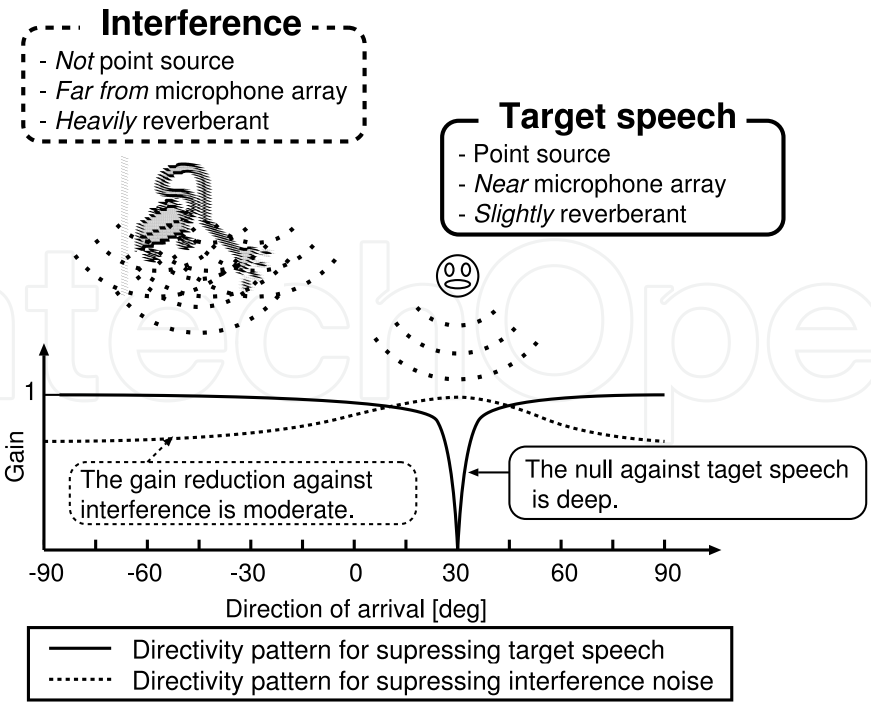

Fig. 3. Expected directivity patterns that are shaped by ICA.

In such an environment, can ICA separate the user's speech signal and a widespread noise signal? The answer is *no*. It is well expected that conventional ICA can suppress the user's speech signal to pick up the noise source, but ICA is very weak in picking up the target speech itself via the suppression of a distant widespread noise. This is due to the fact that ICA with small numbers of sensors and filter taps often provides only directional nulls against undesired source signals. Results of the detailed analysis of ICA for such a case are shown in the following subsections.

3.2 SNR-maximize beamformers optimized by ICA

In this subsection, we consider beamformers that are optimized by ICA in the following acoustic scenario: the target signal is the user's speech and the noise is not a point source. Then, the observed signal contains only one target speech signal and an additive noise. In this scenario, the observed signal is defined as

$$\boldsymbol{x}(f,\tau) = \boldsymbol{A}(f)\boldsymbol{s}(f,\tau) + \boldsymbol{n}_a(f,\tau). \tag{14}$$

Note that the additive noise $\boldsymbol{n}_a(f,\tau)$ cannot be negligible in this scenario. Then, the output of ICA contains two components, i.e., the estimated speech signal $y_s(f,\tau)$ and estimated noise signal $y_n(f,\tau)$; these are given by

$$[y_s(f,\tau), y_n(f,\tau)]^{\mathrm{T}} = \boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{x}(f,\tau). \tag{15}$$

Therefore, ICA optimizes two beamformers; these can be written as

$$\boldsymbol{W}_{\mathrm{ICA}}(f) = [\boldsymbol{g}_s(f), \boldsymbol{g}_n(f)]^{\mathrm{T}}, \tag{16}$$

where $\boldsymbol{g}_s(f) = [g_1^{(s)}(f), \ldots, g_J^{(s)}(f)]^{\mathrm{T}}$ is the coefficient vector of the beamformer used to pick up the target speech signal, and $\boldsymbol{g}_n(f) = [g_1^{(n)}(f), \ldots, g_J^{(n)}(f)]^{\mathrm{T}}$ is the coefficient vector of the beamformer used to pick up the noise. Therefore, (15) can be rewritten as

$$[y_s(f,\tau), y_n(f,\tau)]^{\mathrm{T}} = [\boldsymbol{g}_s(f), \boldsymbol{g}_n(f)]^{\mathrm{T}} \boldsymbol{x}(f,\tau). \tag{17}$$

In SO-ICA, the multiple second-order correlation matrices of distinct time block outputs,

$$\langle \boldsymbol{o}(f,\tau_b)\boldsymbol{o}^{\mathrm{H}}(f,\tau_b)\rangle_{\tau_b}, \tag{18}$$

are diagonalized through joint diagonalization.

On the other hand, in HO-ICA, the higher-order correlation matrix is also diagonalized. Using the Taylor expansion, we can express the factor of the nonlinear vector function of HO-ICA, $\varphi(o_k(f,\tau))$, as

$$\begin{aligned}
\varphi(o_k(f,\tau)) &= \tanh o_k^{(\mathrm{R})}(f,\tau) + i \tanh o_k^{(\mathrm{I})}(f,\tau),\\
&= \left\{ o_k^{(\mathrm{R})}(f,\tau) - \frac{\left(o_k^{(\mathrm{R})}(f,\tau)\right)^3}{3} + \cdots \right\} + i \left\{ o_k^{(\mathrm{I})}(f,\tau) - \frac{\left(o_k^{(\mathrm{I})}(f,\tau)\right)^3}{3} + \cdots \right\},\\
&= o_k(f,\tau) - \left(\frac{\left(o_k^{(\mathrm{R})}(f,\tau)\right)^3}{3} + i \frac{\left(o_k^{(\mathrm{I})}(f,\tau)\right)^3}{3} \right) + \cdots .
\end{aligned} \tag{19}$$

Thus, the calculation of the higher-order correlation in HO-ICA, $\boldsymbol{\varphi}(\boldsymbol{o}(f,\tau))\boldsymbol{o}^{\mathrm{H}}(f,\tau)$, can be decomposed to a second-order correlation matrix and the summation of higher-order correlation matrices of each order. This is shown as

$$\langle \boldsymbol{\varphi}(\boldsymbol{o}(f,\tau))\boldsymbol{o}^{\mathrm{H}}(f,\tau)\rangle_\tau = \langle \boldsymbol{o}(f,\tau)\boldsymbol{o}^{\mathrm{H}}(f,\tau)\rangle_\tau + \Psi(f), \tag{20}$$

where $\Psi(f)$ is a set of higher-order correlation matrices. In HO-ICA, separation filters are optimized so that all orders of correlation matrices become diagonal matrices. Then, at least the second-order correlation matrix is diagonalized by HO-ICA. In both SO-ICA and HO-ICA, at least the second-order correlation matrix is diagonalized. Hence, we prove in the following that ICA optimizes beamformers as SNR-maximize beamformers focusing on only part of the second-order correlation. Then the absolute value of the normalized cross-correlation coefficient (off-diagonal entries) of the second-order correlation, C, is defined by

$$C = \frac{|\langle y_s(f,\tau) y_n^*(f,\tau)\rangle_\tau|}{\sqrt{\langle |y_s(f,\tau)|^2\rangle_\tau}\sqrt{\langle |y_n(f,\tau)|^2\rangle_\tau}}, \tag{21}$$

$$y_s(f,\tau) = \hat{s}(f,\tau) + r_s \hat{n}(f,\tau), \tag{22}$$

$$y_n(f,\tau) = \hat{n}(f,\tau) + r_n \hat{s}(f,\tau), \tag{23}$$

where $\hat{s}(f,\tau)$ is the target speech component in ICA's output, $\hat{n}(f,\tau)$ is the noise component in ICA's output, r_s is the coefficient of the residual noise component, r_n is the coefficient of the target-leakage component, and the superscript $*$ represents a complex conjugate. Therefore, the SNRs of $y_s(f,\tau)$ and $y_n(f,\tau)$ can be respectively represented by

$$\Gamma_s = \langle|\hat{s}(f,\tau)|^2\rangle_\tau / (|r_s|^2 \langle|\hat{n}(f,\tau)|^2\rangle_\tau), \tag{24}$$

$$\Gamma_n = \langle|\hat{n}(f,\tau)|^2\rangle_\tau / (|r_n|^2 \langle|\hat{s}(f,\tau)|^2\rangle_\tau), \tag{25}$$

where Γ_s is the SNR of $y_s(f,\tau)$ and Γ_n is the SNR of $y_n(f,\tau)$. Using (22), (23), (24), and (25), we can rewrite (21) as

$$C = \frac{\left|1/\sqrt{\Gamma_s}\cdot e^{j\arg r_s} + 1/\sqrt{\Gamma_n}\cdot e^{j\arg r_n^*}\right|}{\sqrt{1+1/\Gamma_s}\sqrt{1+1/\Gamma_n}} = \frac{\left|1/\sqrt{\Gamma_s} + 1/\sqrt{\Gamma_n}\cdot e^{j(\arg r_n^* - \arg r_s)}\right|}{\sqrt{1+1/\Gamma_s}\sqrt{1+1/\Gamma_n}}, \tag{26}$$

where $\arg r$ represents the argument of r. Thus, C is a function of only Γ_s and Γ_n. Therefore, the cross-correlation between $y_s(f,\tau)$ and $y_n(f,\tau)$ only depends on the SNRs of beamformers $\boldsymbol{g}_s(f)$ and $\boldsymbol{g}_n(f)$.

Now, we consider the minimization of C, which is identical to the second-order correlation matrix diagonalization in ICA. When $|\arg r_n^* - \arg r_s| > \pi/2$, where $-\pi < \arg r_s \leq \pi$ and $-\pi < \arg r_n^* \leq \pi$, it is possible to make C zero or minimum independently of Γ_s and Γ_n. This case is appropriate for the orthogonalization between $y_s(f,\tau)$ and $y_n(f,\tau)$, which is related to principal component analysis (PCA) unlike ICA. However, SO-ICA requires that all correlation matrices in the different time blocks are diagonalized (joint diagonalization) to maximize independence among all outputs. Also, HO-ICA requires that all order correlation matrices are diagonalized, i.e., not only $\langle \boldsymbol{o}(f,\tau)\boldsymbol{o}^{\mathrm{H}}(f,\tau)\rangle_\tau$ but also $\Psi(f)$ in (20) is diagonalized. These diagonalizations result in the prevention of the orthogonalization of $y_s(f,\tau)$ and $y_n(f,\tau)$; consequently, hereafter, we can consider only the case of $|\arg r_n^* - \arg r_s| \leq \pi/2$. Then, the partial differential of C^2 with respect to Γ_s is given by

$$\frac{\partial C^2}{\partial \Gamma_s} = \frac{(1-\Gamma_s)}{(\Gamma_s+1)^2(\Gamma_n+1)} + \frac{\Gamma_s\sqrt{\Gamma_s\Gamma_n}(1-\Gamma_s)}{(\Gamma_s+1)^2(\Gamma_n+1)} \cdot 2\mathrm{Re}\left[e^{j(\arg r_n^* - \arg r_s)}\right] < 0, \tag{27}$$

where $\Gamma_s > 1$ and $\Gamma_n > 1$. Similarly to the partial differential of C^2 with respect to Γ_n, we can also prove that $\partial C^2/\partial\Gamma_n < 0$, where $\Gamma_s > 1$ and $\Gamma_n > 1$ in the same manner. Therefore, C is a monotonically decreasing function of Γ_s and Γ_n. The above-mentioned fact indicates the following in ICA.

- The absolute value of cross-correlation only depends on the SNRs of the beamformers spanned by each row of an unmixing matrix.
- The absolute value of cross-correlation is a monotonically decreasing function of SNR.
- Therefore, the diagonalization of a second-order correlation matrix leads to SNR maximization.

Thus, it can be concluded that ICA, in a parallel manner, optimizes multiple beamformers, i.e., $\boldsymbol{g}_s(f)$ and $\boldsymbol{g}_n(f)$, so that the SNR of the output of each beamformer becomes maximum.

3.3 What beamformers are optimized under non-point-source noise condition?

In the previous subsection, it has been proved that ICA optimizes beamformers as SNR-maximize beamformers. In this subsection, we analyze what beamformers are optimized by ICA, particularly under a non-point-source noise condition, where we assume a two-source separation problem. The target speech can be regarded as a point source, and the noise is a non-point-source noise. First, we focus on the beamformer $g_s(f)$ that picks up the target speech signal. The SNR-maximize beamformer for $g_s(f)$ minimizes the undesired signal's power under the condition that the target signal's gain is kept constant. Thus, the desired beamformer should satisfy

$$\min_{g_s(f)} g_s^{\mathrm{T}}(f)\boldsymbol{R}(f)g_s(f) \quad \text{subject to } g_s^{\mathrm{T}}(f)\boldsymbol{a}(f,\theta_s) = 1, \tag{28}$$

$$\boldsymbol{a}(f,\theta_s(f)) = [\exp(i2\pi(f/M)f_{\mathrm{s}}d_1 \sin\theta_s/c), \ldots, \exp(i2\pi(f/M)f_{\mathrm{s}}d_J \sin\theta_s/c)]^{\mathrm{T}}, \tag{29}$$

where $\boldsymbol{a}(f,\theta_s(f))$ is the steering vector, $\theta_s(f)$ is the direction of the target speech, M is the DFT size, f_{s} is the sampling frequency, c is the sound velocity, and $\boldsymbol{R}(f) = \langle \boldsymbol{n}_a(f,\tau)\boldsymbol{n}_a^{\mathrm{H}}(f,\tau)\rangle_\tau$ is the correlation matrix of $\boldsymbol{n}_a(f,\tau)$. Note that $\theta_s(f)$ is a function of frequency because the DOA of the source varies in each frequency subband under a reverberant condition. Here, using the Lagrange multiplier, the solution of (28) is

$$g_s(f)^{\mathrm{T}} = \frac{\boldsymbol{a}(f,\theta_s(f))^{\mathrm{H}}\boldsymbol{R}^{-1}(f)}{\boldsymbol{a}(f,\theta_s(f))^{\mathrm{H}}\boldsymbol{R}^{-1}(f)\boldsymbol{a}(f,\theta_s(f))}. \tag{30}$$

This beamformer is called a minimum variance distortionless response (MVDR) beamformer [25]. Note that the MVDR beamformer requires the true DOA of the target speech and the noise-only time interval. However, we cannot determine the true DOA of the target source signal and the noise-only interval because ICA is an *unsupervised* adaptive technique. Thus, the MVDR beamformer is expected to be the upper limit of ICA in the presence of non-point-source noises.

Although the correlation matrix is often not diagonalized in lower-frequency subbands [25], e.g., diffuse noise, we approximate that the correlation matrix is almost diagonalized in subbands in the entire frequency. Then, regarding the power of noise signals as approximately $\delta^2(f)$, the correlation matrix results in $\boldsymbol{R}(f) = \delta^2(f) \cdot \boldsymbol{I}$. Therefore, the inverse of the correlation matrix $\boldsymbol{R}^{-1}(f) = \boldsymbol{I}/\delta^2(f)$ and (30) can be rewritten as

$$g_s(f)^{\mathrm{T}} = \frac{\boldsymbol{a}(f,\theta_s(f))^{\mathrm{H}}}{\boldsymbol{a}(f,\theta_s(f))^{\mathrm{H}}\boldsymbol{a}(f,\theta_s(f))}. \tag{31}$$

Since $\boldsymbol{a}(f,\theta_s(f))^{\mathrm{H}}\boldsymbol{a}(f,\theta_s(f)) = J$, we finally obtain

$$g_s(f) = \frac{1}{J}[\exp\left(-i2\pi(f/M)f_{\mathrm{s}}d_1 \sin\theta_s(f)/c\right), \ldots, \exp\left(-i2\pi(f/M)f_{\mathrm{s}}d_J \sin\theta_s(f)/c\right)]^{\mathrm{T}}. \tag{32}$$

This filter $g_s(f)$ is approximately equal to a DS beamformer [4]. Note that the filter $g_s(f)$ is not a simple DS beamformer but a *reverberation-adapted DS beamformer* because it is optimized for a distinct $\theta_s(f)$ in each frequency bin. The resultant noise power is $\delta^2(f)/J$ when the noise is spatially uncorrelated and white Gaussian. Consequently the noise-reduction performance of the DS beamformer optimized by ICA under a non-point-source noise condition is proportional to $10 \log_{10} J$ [dB]; this performance is not particularly good.

Next, we consider the other beamformer $\boldsymbol{g}_n(f)$, which picks up the noise source. Similar to the noise signal, the beamformer that removes the target signal arriving from $\theta_s(f)$ is the SNR-maximize beamformer. Thus, the beamformer that steers the directional null to $\theta_s(f)$ is the desired one for the noise signal. Such a beamformer is called NBF [16]. This beamformer compensates for the phase of the signal arriving from $\theta_s(f)$, and carries out subtraction. Thus, the signal arriving from $\theta_s(f)$ is removed. For instance, NBF with a two-element array is designed as

$$\boldsymbol{g}_n(f) = [\exp(-i2\pi(f/M)f_\mathrm{s}d_1 \sin\theta_s(f)/c), -\exp(-i2\pi(f/M)f_\mathrm{s}d_2 \sin\theta_s(f)/c)]^\mathrm{T} \cdot \sigma(f), \quad (33)$$

where $\sigma(f)$ is the gain compensation parameter. This beamformer surely satisfies $\boldsymbol{g}_n^\mathrm{T}(f) \cdot \boldsymbol{a}(f, \theta_s(f)) = 0$. The steering vector $\boldsymbol{a}(f, \theta_s(f))$ expresses the wavefront of the plane wave arriving from $\theta_s(f)$. Thus, $\boldsymbol{g}_n(f)$ actually steers the directional null to $\theta_s(f)$. Note that this always occurs regardless of the number of microphones (at least two microphones). Hence, this beamformer achieves a reasonably high, ideally infinite, SNR for the noise signal. Also, note that the filter $\boldsymbol{g}_n(f)$ is not a simple NBF but a *reverberation-adapted NBF* because it is optimized for a distinct $\theta_s(f)$ in each frequency bin. Overall, the performance of enhancing the target speech is very poor but that of estimating the noise source is good.

3.4 Computer simulations

We conduct computer simulations to confirm the performance of ICA under a non-point-source noise condition. Here, we used HO-ICA [14] as the ICA algorithm. We used the following 8-kHz-sampled signals as the ICA's input; the original target speech (3 s) was convoluted with impulse responses that were recorded in an actual environment, and to which three types of noise from 36 loudspeakers were added. The reverberation time (RT_{60}) is 200 ms; this corresponds to mixing filters with 1600 taps in 8 kHz sampling. The three types of noise are an independent Gaussian noise, actually recorded railway-station noise, and interference speech by 36 people. Figure 4 illustrates the reverberant room used in the simulation. We use 12 speakers (6 males and 6 females) as sources of the original target speech, and the input SNR of test data is set to 0 dB. We use a two-, three-, or four-element microphone array with an interelement spacing of 4.3 cm.

The simulation results are shown in Figs. 5 and 6. Figure 5 shows the result for the average noise reduction rate (NRR) [16] of all the target speakers. NRR is defined as the output SNR in dB minus the input SNR in dB. This measure indicates the objective performance of noise reduction. NRR is given by

$$\mathrm{NRR}\ [\mathrm{dB}] = \frac{1}{J} \sum_{j=1}^{J} (\mathrm{OSNR} - \mathrm{ISNR}_j), \quad (34)$$

where OSNR is the output SNR and ISNR_j is the input SNR of microphone j.

From this result, we can see an imbalance between the target speech estimation and the noise estimation in every noise case; the performance of the target speech estimation is significantly poor, but that of noise estimation is very high. This result is consistent with

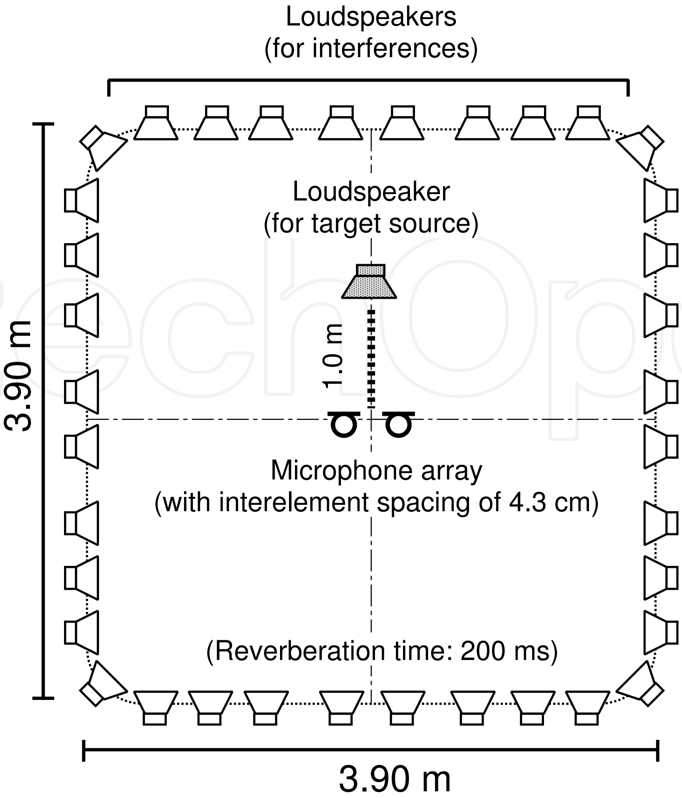

Fig. 4. Layout of reverberant room in our simulation.

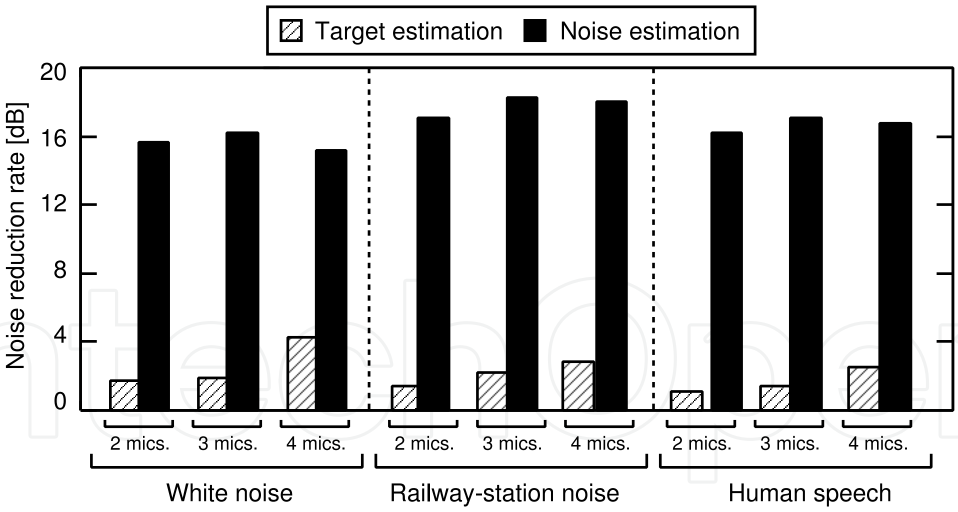

Fig. 5. Simulation-based separation results under non-point-source noise condition.

the previously stated theory. Moreover, Fig. 6 shows directivity patterns shaped by the beamformers optimized by ICA in the simulation. It is clearly indicated that beamformer $g_s(f)$, which picks up the target speech, resembles the DS beamformer, and that beamformer $g_n(f)$, which picks up the noise, becomes NBF. From these results, it is confirmed that the previously stated theory, i.e., the beamformers optimized by ICA under a non-point-source noise condition are DS and NBF, is valid.

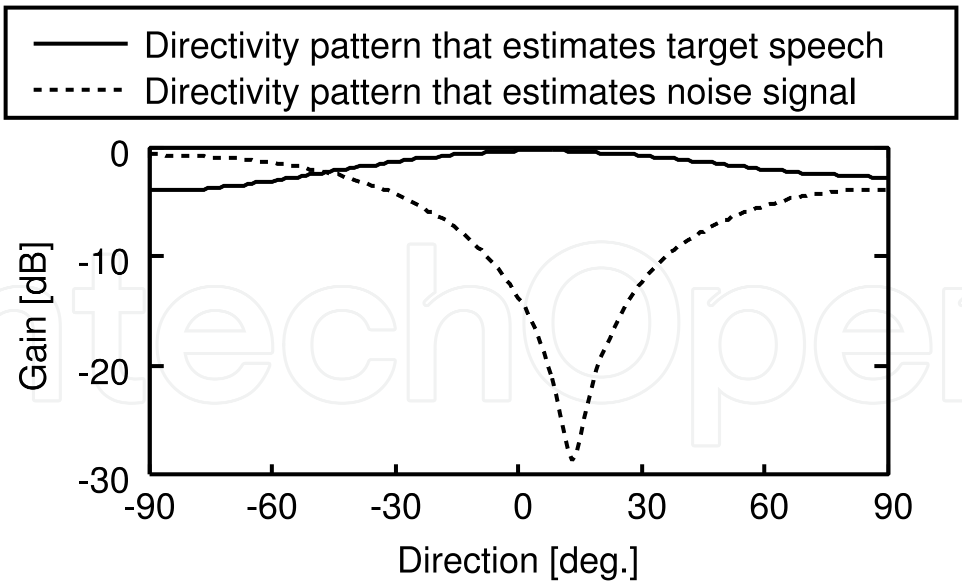

Fig. 6. Typical directivity patterns under non-point-source noise condition shaped by ICA at 2 kHz and two-element array for case of white Gaussian noise.

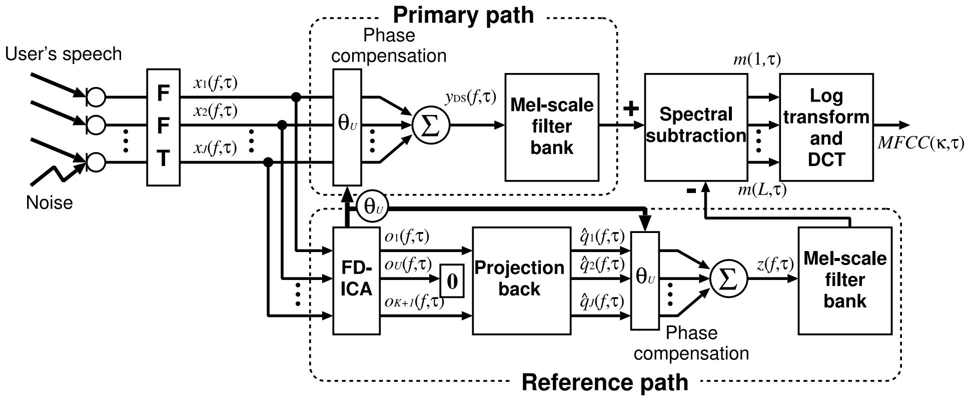

Fig. 7. Block diagram of blind spatial subtraction array.

4. Blind spectral subtraction array

4.1 Motivation and strategy

As clearly shown in Sects. 3.3 and 3.4, ICA is proficient in noise estimation rather than in target-speech estimation under a non-point-source noise condition. Thus, we cannot use ICA for direct target estimation under such a condition. However, we can still use ICA as a noise estimator. This motivates us to introduce an improved speech-enhancement strategy, i.e., BSSA [21]. BSSA consists of a DS-based primary path and a reference path including ICA-based noise estimation (see Fig. 7). The estimated noise component in ICA is efficiently subtracted from the primary path in the power-spectrum domain without phase information. This procedure can yield better target-speech enhancement than simple ICA, even with the additional benefit of estimation-error robustness in speech recognition applications. The detailed process of signal processing is shown below.

4.2 Partial speech enhancement in primary path

We again consider the generalized form of the observed signal as described in (1). The target speech signal is partly enhanced in advance by DS. This procedure can be given as

$$y_{\mathrm{DS}}(f,\tau)=\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{x}(f,\tau)=\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{A}(f)\boldsymbol{s}(f,\tau)+\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{A}(f)\boldsymbol{n}(f,\tau)+\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{n}_a(f,\tau), \quad (35)$$

$$\boldsymbol{w}_{\mathrm{DS}} = [w_1^{(\mathrm{DS})}(f),\ldots,w_J^{(\mathrm{DS})}(f)]^{\mathrm{T}}, \quad (36)$$

$$w_j^{(\mathrm{DS})}(f) = \frac{1}{J}\exp\left(-i2\pi(f/M)f_{\mathrm{s}}d_j\sin\theta_U/c\right), \quad (37)$$

where $y_{\mathrm{DS}}(f,\tau)$ is the primary-path output that is a slightly enhanced target speech, $\boldsymbol{w}_{\mathrm{DS}}(f)$ is the filter coefficient vector of DS, and θ_U is the estimated DOA of the target speech given by the ICA part in Sect. 4.3. In (35), the second and third terms on the right-hand side express the remaining noise in the output of the primary path.

4.3 ICA-based noise estimation in reference path

BSSA provides ICA-based noise estimation. First, we separate the observed signal by ICA and obtain the separated signal vector $\boldsymbol{o}(f,\tau)$ as

$$\boldsymbol{o}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{x}(f,\tau), \quad (38)$$

$$\boldsymbol{o}(f,\tau) = [o_1(f,\tau),\ldots,o_{K+1}(f,\tau)]^{\mathrm{T}}, \quad (39)$$

$$\boldsymbol{W}_{\mathrm{ICA}}(f) = \begin{bmatrix} W_{11}^{(\mathrm{ICA})}(f) & \cdots & W_{1J}^{(\mathrm{ICA})}(f) \\ \vdots & & \vdots \\ W_{(K+1)1}^{(\mathrm{ICA})}(f) & \cdots & W_{(K+1)J}^{(\mathrm{ICA})}(f) \end{bmatrix}, \quad (40)$$

where the unmixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$ is optimized by (11). Note that the number of ICA outputs becomes $K+1$, and thus the number of sensors, J, is more than $K+1$ because we assume that the additive noise $\boldsymbol{n}_a(f,\tau)$ is not negligible. We cannot estimate the additive noise perfectly because it is deformed by the filter optimized by ICA. Moreover, other components also cannot be estimated perfectly when the additive noise $\boldsymbol{n}_a(f,\tau)$ exists. However, we can estimate at least noises (including interference sounds that can be regarded as point sources, and the additive noise) that do not involve the target speech signal, as indicated in Sect. 3. Therefore, the estimated noise signal is still beneficial.

Next, we estimate DOAs from the unmixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$ [16]. This procedure is represented by

$$\theta_u = \sin^{-1}\frac{\arg\left(\frac{[\boldsymbol{W}_{\mathrm{ICA}}^{-1}(f)]_{ju}}{[\boldsymbol{W}_{\mathrm{ICA}}^{-1}(f)]_{j'u}}\right)}{2\pi f_{\mathrm{s}}c^{-1}(d_j - d_{j'})}, \quad (41)$$

where θ_u is the DOA of the uth sound source. Then, we choose the Uth source signal, which is nearest the front of the microphone array, and designate the DOA of the chosen source signal as θ_U. This is because almost all users are expected to stand in front of the microphone array in a speech-oriented human-machine interface, e.g., a public guidance system. Other strategies for choosing the target speech signal can be considered as follows.

- If the approximate location of a target speaker is known in advance, we can utilize the location of the target speaker. For instance, we can know the approximate location of the target speaker at a hands-free speech recognition system in a car navigation system in advance. Then, the DOA of the target speech signal is approximately known. For such systems, we can choose the target speech signal, selecting the specific component in which the DOA estimated by ICA is nearest the known target-speech DOA.
- For an interaction robot system [26], we can utilize image information from a camera mounted on a robot. Therefore, we can estimate DOA from this information, and we can choose the target speech signal on the basis of this estimated DOA.
- If the only target signal is speech, i.e., none of the noises are speech, we can choose the target speech signal on the basis of the Gaussian mixture model (GMM), which can classify sound signals into voices and nonvoices [27].

Next, in the reference path, no target speech signal is required because we want to estimate only noise. Therefore, we eliminate the user's signal from the ICA's output signal $\boldsymbol{o}(f,\tau)$. This can be written as

$$\boldsymbol{q}(f,\tau) = [o_1(f,\tau), ..., o_{U-1}(f,\tau), 0, o_{U+1}(f,\tau), ..., o_{K+1}(f,\tau)]^{\mathrm{T}}, \tag{42}$$

where $\boldsymbol{q}(f,\tau)$ is the "noise-only" signal vector that contains only noise components. Next, we apply the projection back (PB) [13] method to remove the ambiguity of amplitude. This procedure can be represented as

$$\hat{\boldsymbol{q}}(f,\tau) = \boldsymbol{W}^{+}_{\mathrm{ICA}}(f)\boldsymbol{q}(f,\tau), \tag{43}$$

where $\boldsymbol{M}^{+}$ denotes the Moore-Penrose pseudo-inverse matrix of $\boldsymbol{M}$. Thus, $\hat{\boldsymbol{q}}(f,\tau)$ is a good estimate of the noise signals received at the microphone positions, i.e.,

$$\hat{\boldsymbol{q}}(f,\tau) \simeq \boldsymbol{A}(f)\boldsymbol{n}(f,\tau) + \boldsymbol{W}^{+}_{\mathrm{ICA}}(f)\hat{\boldsymbol{n}}_a(f,\tau), \tag{44}$$

where $\hat{\boldsymbol{n}}_{\boldsymbol{a}}(f,\tau)$ contains the deformed additive noise signal and separation error due to an additive noise. Finally, we construct the estimated noise signal $z(f,\tau)$ by applying DS as

$$z(f,\tau) = \boldsymbol{w}^{\mathrm{T}}_{\mathrm{DS}}(f)\hat{\boldsymbol{q}}(f,\tau) \simeq \boldsymbol{w}^{\mathrm{T}}_{\mathrm{DS}}(f)\boldsymbol{A}(f)\boldsymbol{n}(f,\tau) + \boldsymbol{w}^{\mathrm{T}}_{\mathrm{DS}}(f)\boldsymbol{W}^{+}_{\mathrm{ICA}}(f)\hat{\boldsymbol{n}}_a(f,\tau). \tag{45}$$

This equation means that $z(f,\tau)$ is a good candidate for noise terms of the primary path output $y_{\mathrm{DS}}(f,\tau)$ (see the 2nd and 3rd terms on the right-hand side of (35)). Of course this noise estimation is not perfect, but we can still enhance the target speech signal via oversubtraction in the power-spectrum domain, as described in Sect. 4.4. Note that $z(f,\tau)$ is a function of the frame index τ, unlike the constant noise prototype in the traditional spectral subtraction method [2]. Therefore, the proposed BSSA can deal with *nonstationary* noise.

4.4 Noise reduction processing in BSSA

In BSSA, noise reduction is carried out by subtracting the estimated noise power spectrum (45) from the partly enhanced target speech signal power spectrum (35). This procedure is given as

$$y_{\mathrm{BSSA}}(f,\tau) = \begin{cases} \{|y_{\mathrm{DS}}(f,\tau)|^2 - \beta \cdot |z(f,\tau)|^2\}^{\frac{1}{2}} \\ \quad (\text{ if } |y_{\mathrm{DS}}(f,\tau)|^2 - \beta \cdot |z(f,\tau)|^2 \geq 0 \text{ }), \\ \gamma \cdot |y_{\mathrm{DS}}(f,\tau)| \qquad (\text{otherwise}), \end{cases} \tag{46}$$

where $y_{\mathrm{BSSA}}(f,\tau)$ is the final output of BSSA, β is the oversubtraction parameter, and γ is the flooring parameter. Their appropriate setting, e.g., $\beta > 1$ and $\gamma \ll 1$, results in efficient noise reduction. For example, a larger oversubtraction parameter ($\beta \gg 1$) leads to a larger SNR improvement. However, the target signal would be distorted. On the other hand, a smaller oversubtraction parameter ($\beta \ll 1$) gives a less-distorted target signal. However, the SNR improvement is decreased. In the end, a trade-off between SNR improvement and the distortion of the output signal exists with respect to the parameter β; $1 < \beta < 2$ is usually used.

The system switches between two equations depending on the conditions in (46). If the calculated noise components using ICA in (45) are underestimated, i.e., $|y_{\mathrm{DS}}(f,\tau)|^2 > \beta|z(f,\tau)|^2$, the resultant output $y_{\mathrm{BSSA}}(f,\tau)$ corresponds to power-spectrum-domain subtraction among the primary and reference paths with an oversubtraction rate of β. On the other hand, if the noise components are overestimated in ICA, i.e., $|y_{\mathrm{DS}}(f,\tau)|^2 < \beta|z(f,\tau)|^2$, the resultant output $y_{\mathrm{BSSA}}(f,\tau)$ is floored with a small positive value to avoid a negative-valued unrealistic spectrum. These *oversubtraction* and *flooring* procedures enable error-robust speech enhancement in BSSA rather than a simple linear subtraction. Although the nonlinear processing in (46) often generates an artificial distortion, so-called *musical noise*, it is still applicable in the speech recognition system because the speech decoder is not very sensitive to such a distortion. BSSA involves mel-scale filter bank analysis and directly outputs the mel-frequency cepstrum coefficient (MFCC) [28] for speech recognition. Therefore, BSSA requires no transformation into the time-domain waveform for speech recognition.

In BSSA, DS and SS are processed in addition to ICA. In HO-ICA or SO-ICA, to calculate the correlation matrix, at least hundreds of product-sum operations are required in each frequency subband. On the other hand, in DS, at most J product-sum operations are required in each frequency subband. A mere 4 or 5 products are required for SS. Therefore, the complexity of BSSA does not increase by as much as 10% compared with ICA.

4.5 Variation and extension in noise reduction processing

As mentioned in the previous subsection, the noise reduction processing of BSSA is mainly based on SS, and therefore it often suffers from the problem of musical noise generation due to its nonlinear signal processing. This becomes a big problem in any audio applications aimed for human hearing, e.g., hearing-aids, teleconference systems, etc.

To improve the sound quality of BSSA, many kinds of variations have been proposed and implemented in the post-processing part in (46). Generalized SS and parametric Wiener filtering algorithms [29] have been introduced to successfully mitigate musical

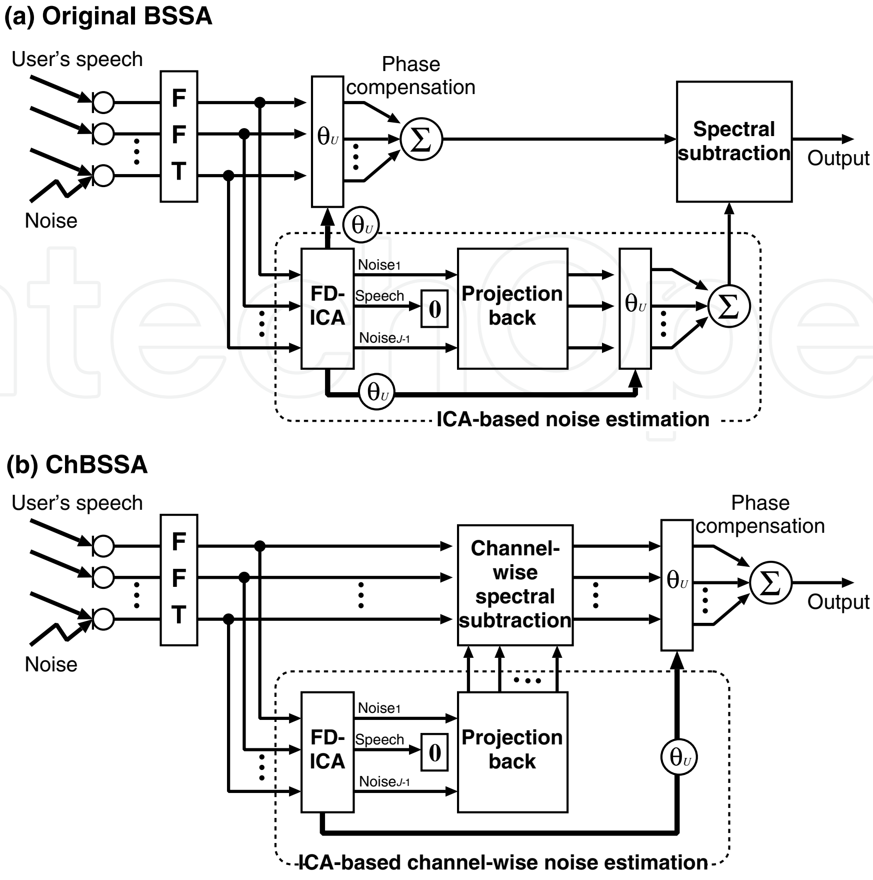

Fig. 8. Configurations of (a) original BSSA and (b) chBSSA.

noise generation [30]. Furthermore, the minimum mean-square error (MMSE) short-time spectral amplitude (STSA) estimator [31] can be used for achieving low-distortion speech enhancement in BSSA [32]. In addition, this MMSE-STSA estimator with ICA-based noise estimation has been modified to deal with binaural signal enhancement, where the spatial cue of the target speech signal can be maintained in the output of BSSA [33].

In recent studies, an interesting extension in the signal processing structure has been addressed [34, 35]. Two types of the BSSA structures are shown in Fig. 8. One is the original BSSA structure that performs SS after DS (see Fig. 8(a)), and another is that SS is channelwisely performed before DS (chBSSA; see Fig. 8(b)). It has been theoretically clarified that chBSSA is superior to BSSA in the mitigation of the musical noise generation via higher-order statistics analysis.

5. Experiment and evaluation

5.1 Experiment in reverberant room

In this experiment, we present a comparison of typical blind noise reduction methods, namely, the conventional ICA [14] and the traditional SS [2] cascaded with ICA (ICA+SS). We utilize the HO-ICA algorithm as conventional ICA [14]. Hereafter, 'ICA' simply indicates HO-ICA. For ICA+SS, we first obtain the estimated noise from the speech pause interval in the target speech estimation by ICA. The noise reduction achieved by SS is

$$y_{\mathrm{ICA+SS}}(f,\tau) = \begin{cases} \{|o_U(f,\tau)|^2 - \beta|\hat{n}_{\mathrm{remain}}(f)|^2\}^{\frac{1}{2}} & (\text{where } |o_U(f,\tau)|^2 - \beta|\hat{n}_{\mathrm{remain}}(f,\tau)|^2 \geq 0), \\ \gamma|o_U(f,\tau)| & (\text{otherwise}), \end{cases} \tag{47}$$

where $\hat{n}_{\mathrm{remain}}(f)$ is the noise signal from the speech pause in the target speech estimated by ICA. Moreover, a DOA-based permutation solver[16] is used in conventional ICA and in the ICA part in BSSA.

We used 16-kHz-sampled signals as test data; the original speech (6 s) was convoluted with impulse responses recorded in an actual environment, to which cleaner noise or a male's interfering speech recorded in an actual environment was added. Figure 9 shows the layout of the reverberant room used in the experiment. The reverberation time of the room is 200 ms; this corresponds to mixing filters of 3200 taps in 16 kHz sampling. The cleaner noise is not a simple point source signal but consists of several *nonstationary* noises emitted from a motor, an air duct, and a nozzle. Also, the male's interfering speech is not a simple point source but is slightly moving. In addition, these interference noises involve background noise. The SNR of the background noise (power ratio of target speech to background noise) is about 28 dB. We use 46 speakers (200 sentences) as the source of the target speech. The input SNR is set to 10 dB at the array. We use a four-element microphone array with an interelement spacing of 2 cm. The DFT size is 512. The oversubtraction parameter β is 1.4 and the flooring coefficient γ is 0.2. Such parameters were experimentally determined. The speech recognition task and conditions are shown in Table 1.

Regarding the evaluation index, we calculate NRR described in (34), cepstral distortion (CD), and speech recognition, which is the final goal of BSSA, in which the separated sound quality is fully considered. CD [36] is a measure of the degree of distortion via the cepstrum domain. It indicates the distortion among two signals, which is defined as

$$\mathrm{CD}\ [\mathrm{dB}] \equiv \frac{1}{T}\sum_{\tau=1}^{T} D_b \sqrt{\sum_{\rho=1}^{B} 2(C_{\mathrm{out}}(\rho;\tau) - C_{\mathrm{ref}}(\rho;\tau))^2}, \tag{48}$$

$$D_b = \frac{20}{\log 10}, \tag{49}$$

where T is the frame length, $C_{\mathrm{out}}(\rho;\tau)$ is the ρth cepstrum coefficient of the output signal in the frame τ, $C_{\mathrm{ref}}(\rho;\tau)$ is the ρth cepstrum coefficient of the speech signal convoluted with the impulse response, and D_b is a constant that transforms the measure into dB. Moreover, B is the number of dimensions of the cepstrum used in the evaluation. Moreover, we use the word accuracy (WA) score as a speech recognition performance. This index is defined as

$$\mathrm{WA}\ [\%] \equiv \frac{W_{\mathrm{WA}} - S_{\mathrm{WA}} - D_{\mathrm{WA}} - I_{\mathrm{WA}}}{W_{\mathrm{WA}}} \times 100, \tag{50}$$

where W_{WA} is the number of words, S_{WA} is the number of substitution errors, D_{WA} is the number of dropout errors, and I_{WA} is the number of insertion errors.

First, actual separation results obtained by ICA for the case of cleaner noise and interference speech are shown in Fig. 10. We can confirm the imbalanced performance between target estimation and noise estimation, similar to the simulation-based results (see Sect. 3.4).

Database	JNAS [37], 306 speakers (150 sentences/speaker)
Task	20 k newspaper dictation
Acoustic model	phonetic tied mixture (PTM) [37], clean model
Number of training speakers for acoustic model	260 speakers (150 sentences/speaker)
Decoder	JULIUS [37] ver 3.5.1

Table 1. Conditions for Speech Recognition

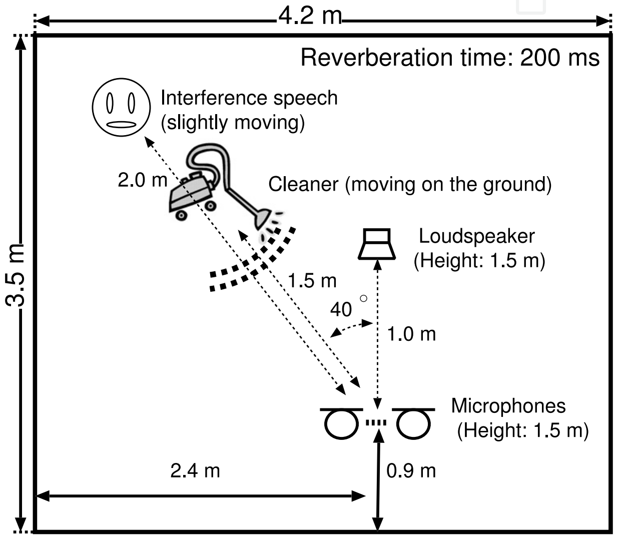

Fig. 9. Layout of reverberant room used in our experiment.

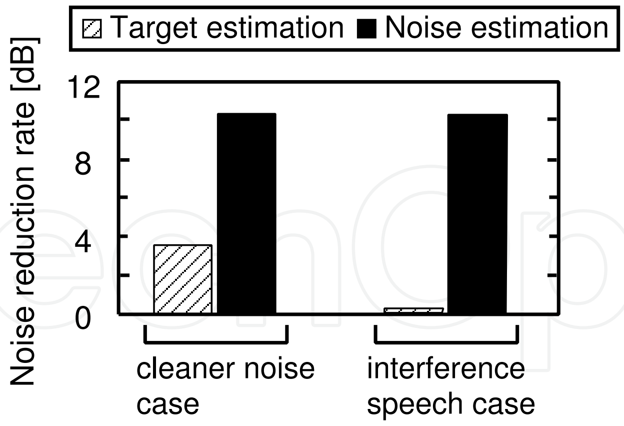

Fig. 10. NRR-based separation performance of conventional ICA in environment shown in Fig. 9.

Next, we discuss the NRR-based experimental results shown in Figs. 11(a) and 12(a). From the results, we can confirm that the NRRs of BSSA are more than 3 dB greater than those of conventional ICA and ICA+SS. However, we can see that the distortion of BSSA is slightly higher from Figs. 11(b) and 12(b). This is due to the fact that the noise reduction of BSSA

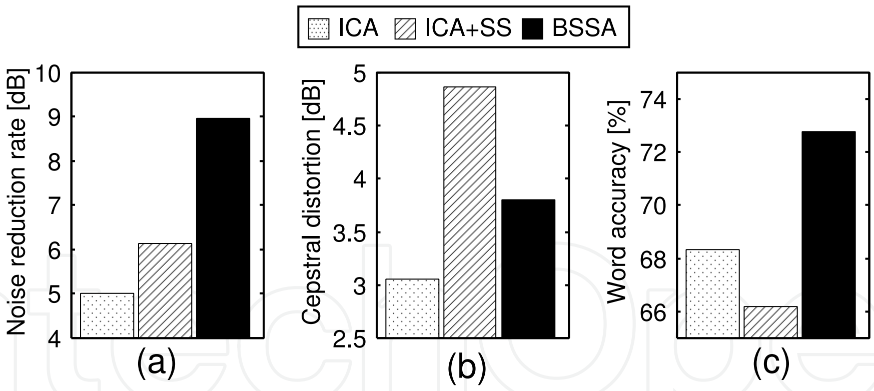

Fig. 11. Results of (a) noise reduction rate, (b) cepstral distortion, and (c) speech recognition test for each method (cleaner noise case).

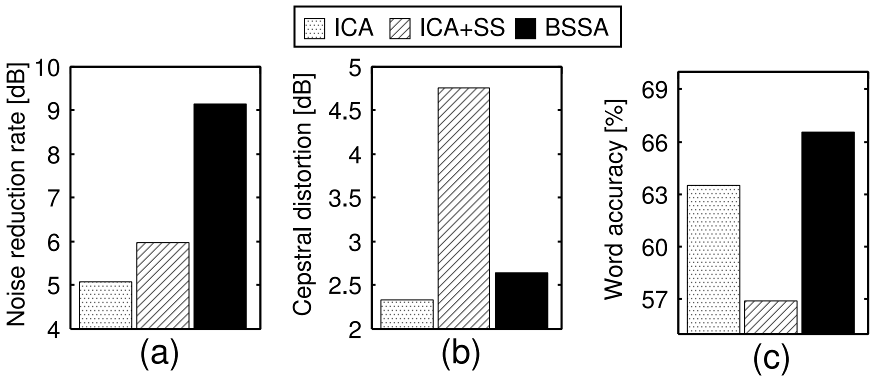

Fig. 12. Results of (a) noise reduction rate, (b) cepstral distortion, and (c) speech recognition test using each method (interference speech case).

is performed on the basis of spectral subtraction. However, the increase in the degree of distortion is expected to be negligible.

Finally, we show the speech recognition result in Figs. 11(c) and 12(c). It is evident that BSSA is superior to conventional ICA and ICA+SS.

5.2 Experiment in real world

An experiment in an actual railway-station environment is discussed here. Figure 13 shows the layout of the railway-station environment used in this experiment, where the reverberation time is about 1000 ms; this corresponds to mixing filters of 16000 taps in 16 kHz sampling. We used 16-kHz-sampled signals as test data; the original speech (6 s) was convoluted with impulse responses recorded in the same railway-station environment, to which a real-recorded noise was added. We use 46 speakers (200 sentences) as the original source of the target speech. The noise in the environment is nonstationary and is almost a non-point-source; it consists of various kinds of interference noise, namely, background noise and the sounds of trains, ticket-vending machines, automatic ticket gates, footsteps, cars, and wind. Figure 14 shows two typical noises, i.e., noises 1 and 2, which are recorded in distinct time periods and used in this experiment. A four-element array with an interelement spacing of 2 cm is used.

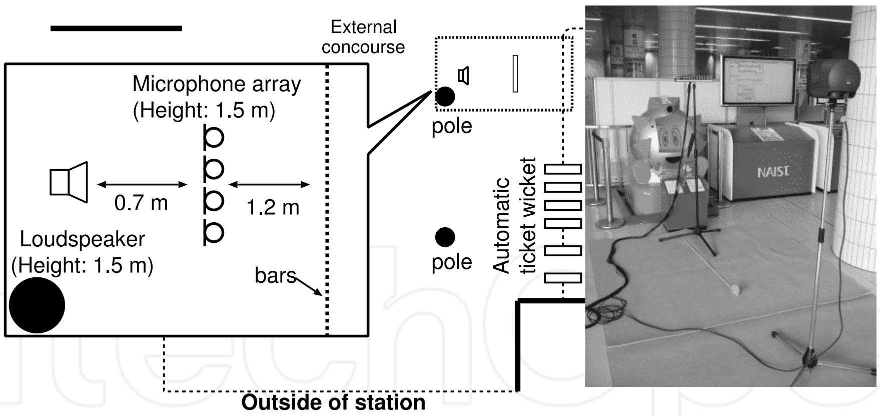

Fig. 13. Layout of railway-station environment used in our experiment.

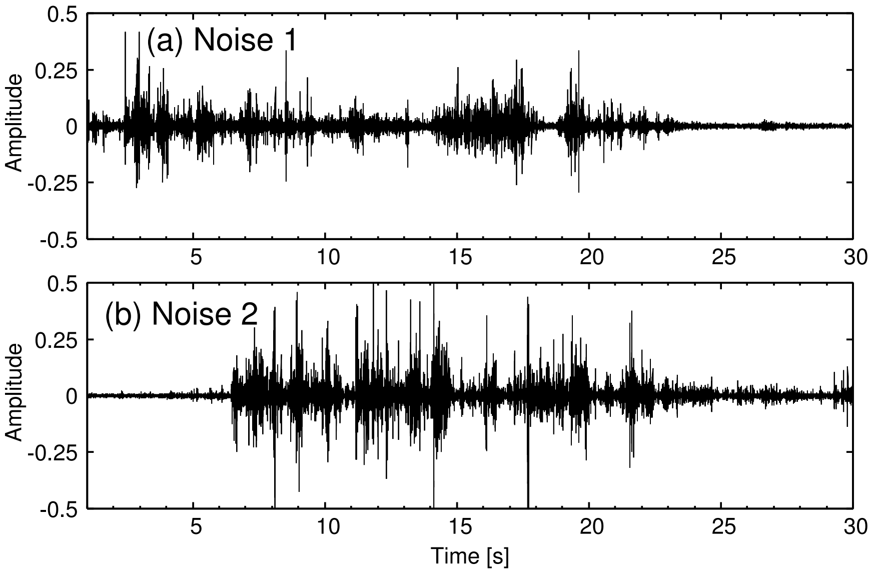

Fig. 14. Two typical noises in railway-station environment.

Figure 15 shows the real separation results obtained by ICA in the railway-station environment. We can ascertain the imbalanced performance between target estimation and noise estimation, similar to the simulation-based results (see Sect. 3.4).

In the next experiment, we compare conventional ICA, ICA+SS, and BSSA in terms of NRR, cepstral distortion, and speech recognition performance. Figure 16(a) shows the results of the average NRR for whole sentences. From these results, we can see that the NRR of BSSA that utilizes ICA as a noise estimator is superior to those of conventional methods. However, we find that the cepstral distortion in BSSA is greater than compared with that in ICA from Fig. 16(b).

Finally, we show the results of speech recognition, where the extracted sound quality is fully considered, in Fig. 16(c). The speech recognition task and conditions are the same as those in Sect. 5.1, as shown in Table 1. From this result, it can be concluded that the target-enhancement performance of BSSA, i.e., the method that uses ICA as a noise estimator, is evidently superior to the method that uses ICA directly as well as ICA+SS.

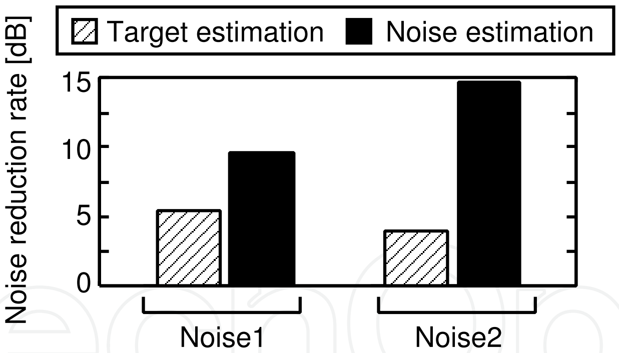

Fig. 15. NRR-based noise reduction performance of conventional ICA in railway-station environment.

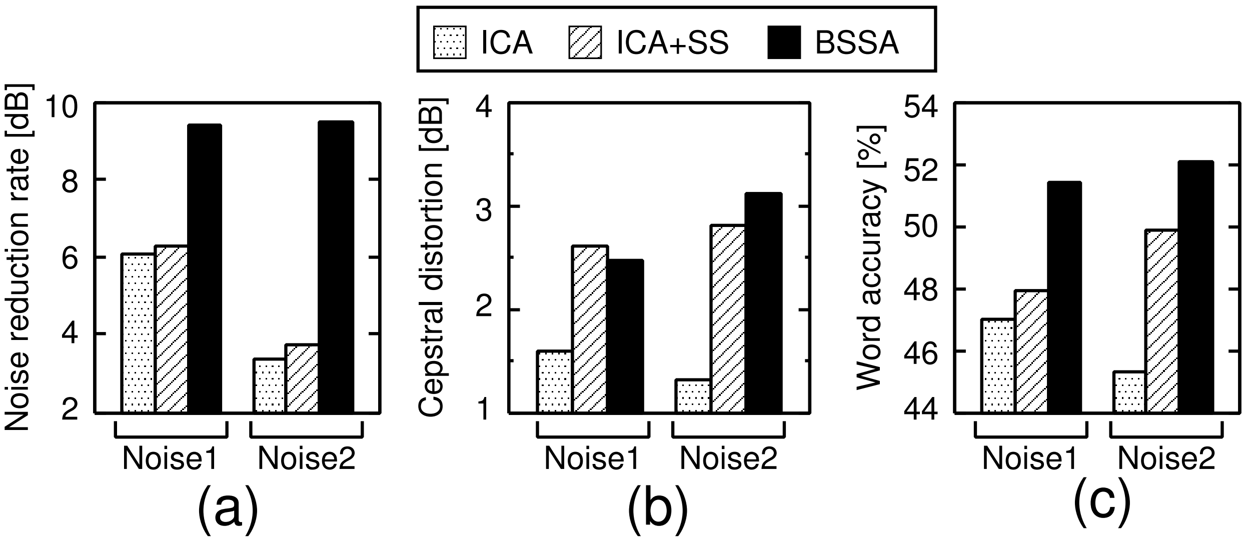

Fig. 16. Experimental results of (a) noise reduction rate, (b) cepstral distortion, and (c) speech recognition test in railway-station environment.

6. Real-time implementation of BSS

Several recent studies [19, 38, 39] have dealt with the issue of real-time implementation of ICA. The methods used, however, require high-speed personal computers, and BSS implementation on a small LSI still receives much attention in industrial applications. As a recent example of the implementation of real-time BSS, a real-time BSSA algorithm and its development are described in the following.

In BSSA's signal processing, the DS, SS, and separation filtering parts are possible to work in real-time. However, it is toilsome to optimize (update) the separation filter in real-time because the optimization of the unmixing matrix by ICA consumes huge amount of computations. Therefore, we should introduce a strategy in which the separation filter optimized by using the past time period data is applied to the current data. Figure 17 illustrates the configuration of the real-time implementation of BSSA. Signal processing in this implementation is performed as follows.

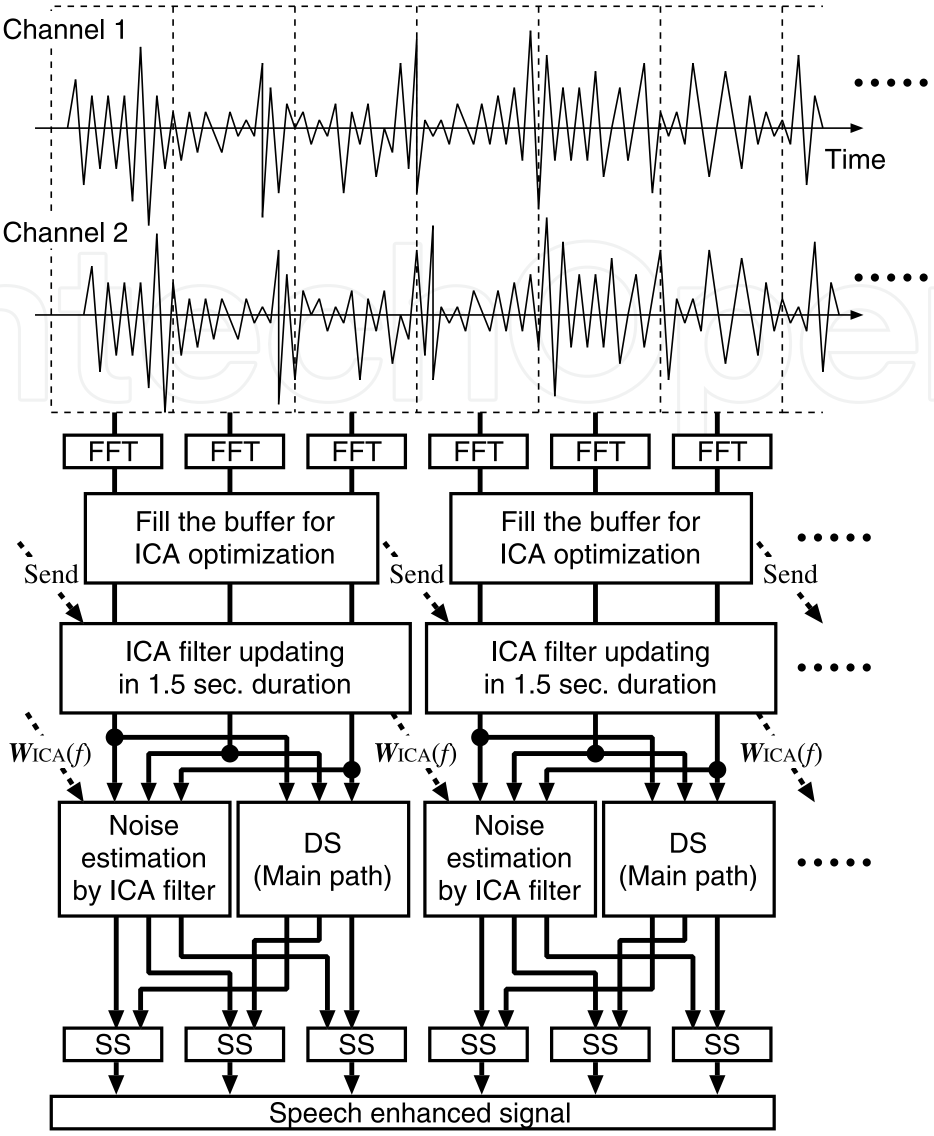
Channel 1
Time
Channel 2
FFT
Fill the buffer for ICA optimization
Send
ICA filter updating in 1.5 sec. duration
$W_{\mathrm{ICA}}(f)$
Noise estimation by ICA filter
DS (Main path)
SS
Speech enhanced signal

Fig. 17. Signal flow in real-time implementation of BSSA.

Step 1: Inputted signals are converted into time-frequency domain series by using a frame-by-frame fast Fourier transform (FFT).

Step 2: ICA is conducted using the past 1.5-s-duration data for estimating the separation filter while the current 1.5 s. The optimized separation filter is applied to the next (*not current*) 1.5 s samples. This staggered relation is due to the fact that the filter update in ICA requires substantial computational complexities and cannot provide an optimal separation filter for the current 1.5 s data.

Step 3: Inputted data is processed in two paths. In the primary path, the target speech is partly enhanced by DS. In the reference path, ICA-based noise estimation is conducted. Again, note that the separation filter for ICA is optimized by using the past time period data.

Step 4: Finally, we obtain the target-speech-enhanced signal by subtracting the power spectrum of the estimated noise signal in the reference path from the power spectrum of the primary path's output.

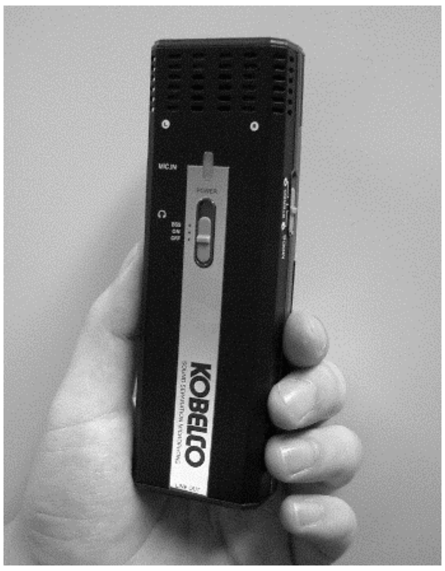

Fig. 18. BSS microphone (SSM-001 by KOBELCO Ltd., Japan) based on BSSA algorithm [40].

Although the update of the separation filter in the ICA part is not real-time processing, but involves a total latency of 3.0 s, the entire system still seems to run in real-time because DS, SS, and separation filtering can be carried out in the current segment with no delay. In the system, the performance degradation due to the latency problem in ICA is mitigated by oversubtraction in spectral subtraction.

Figure 18 shows an example of the hardware implementation of BSSA, which was developed by KOBELCO Ltd., Japan [40]. They have fabricated a pocket-size real-time BSS microphone, where the BSSA algorithm can work on a general-purpose DSP (TEXAS INSTRUMENTS TMS320C6713; 200 MHz clock, 100 kB program size, 1 MB working memory). This microphone was made commercially available in 2007 and has been adopted for the purpose of surveillance by the Japanese National Police Agency.

7. Conclusion

This chapter addressed the BSS problem for speech applications under real acoustic environments, particularly focusing on BSSA that utilizes ICA as a noise estimator. Under a non-point-source noise condition, it was pointed out that beamformers optimized by ICA are a DS beamformer for extracting the target speech signal that can be regarded as a point source and NBF for picking up the noise signal. Thus, ICA is proficient in noise estimation under a non-point-source noise condition. Therefore, it is valid to use ICA as a noise estimator. In experiments involving computer-simulation-based and real-recording-based data, the SNR improvement and speech recognition results of BSSA are superior to those of conventional methods. These results indicate that the ICA-based noise estimation is beneficial for speech enhancement in adverse environments. Also, the hardware implementation of BSS was discussed with a typical example of a real-time BSSA algorithm.

8. References

[1] B. H. Juang and F. K. Soong, "Hands-free telecommunications," *Proc. International Conference on Hands-Free Speech Communication*, pp. 5–10, 2001.

[2] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[3] G. W. Elko, “Microphone array systems for hands-free telecommunication,” *Speech Communication*, vol. 20, pp. 229–240, 1996.

[4] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, “Computer-steered microphone arrays for sound transduction in large rooms,” *Journal of the Acoustical Society of America*, vol. 78, no. 5, pp. 1508–1518, 1985.

[5] M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, “Microphone array based speech recognition with different talker-array positions,” *Proc. ICASSP’97*, pp. 227–230, 1997.

[6] H. F. Silverman and W. R. Patterson, “Visualizing the performance of large-aperture microphone arrays,” *Proc. ICASSP’99*, pp. 962–972, 1999.

[7] O. Frost, “An algorithm for linearly constrained adaptive array processing,” *Proceedings of the IEEE*, vol. 60, pp. 926–935, 1972.

[8] L. J. Griffiths and C. W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[9] Y. Kaneda and J. Ohga, “Adaptive microphone-array system for noise reduction,” *IEEE Trans. Acoust. Speech, Signal Process.*, pp. 2109–2112, 1986.

[10] P. Comon, “Independent component analysis, a new concept?” *Signal Processing*, vol. 36, pp. 287–314, 1994.

[11] J. F. Cardoso, “Eigenstructure of the 4th-order cumulant tensor with application to the blind source separation problem,” *Proc. ICASSP’89*, pp. 2109–2112, 1989.

[12] C. Jutten and J. Herault, “Blind separation of sources part i: An adaptive algorithm based on neuromimetic architecture,” *Signal Processing*, vol. 24, pp. 1–10, 1991.

[13] S. Ikeda and N. Murata, “A method of ICA in the frequency domain,” *Proc. International Workshop on Independent Component Analysis and Blind Signal Separation*, pp. 365–371, 1999.

[14] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[15] L. Parra and C. Spence, “Convolutive blind separation of non-stationary sources,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 320–327, 2000.

[16] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, and T. Nishikawa, “Blind source separation combining independent component analysis and beamforming,” *EURASIP Journal on Applied Signal Processing*, vol. 2003, pp. 1135–1146, 2003.

[17] D.-T. Pham, C. Serviere, and H. Boumaraf, “Blind separation of convolutive audio mixtures using nonstationarity,” *International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pp. 975–980, 2003.

[18] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, “Blind source separation based on a fast-convergence algorithm combining ICA and beamforming,” *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 2, pp. 666–678, 2006.

[19] Y. Mori, H. Saruwatari, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, Y. Ikeda, H. Hashimoto, and T. Morita, “Blind separation of acoustic signals combining SIMO-model-based independent component analysis and binary masking,” *EURASIP Journal on Applied Signal Processing*, vol.2006, Article ID 34970, 17 pages, 2006.

[20] B. Sallberg, N. Grbic, and I. Claesson, “Online maximization of subband kurtosis for blind adaptive beamforming in realtime speech extraction,” *Proc. IEEE Workshop DSP 2007*, pp. 603–606, 2007.

[21] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, “Blind spatial subtraction array for speech enhancement in noisy environment,” *IEEE Transactions on Audio, Speech and Language Processing*, vol.17, no.4, pp.650–664, 2009.

[22] T.-W. Lee, *Independent Component Analysis*. Norwell, MA: Kluwer Academic, 1998.

[23] S. Ukai, T. Takatani, T. Nishikawa, and H. Saruwatari, “Blind source separation combining SIMO-model-based ICA and adaptive beamforming,” *Proc. ICASSP2005*, vol. III, pp. 85–88, 2005.

[24] S. Araki, S. Makino, Y. Hinamoto, R. Mukai, T. Nishikawa, and H. Saruwatari, “Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming for convolutive mixtures,” *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1157–1166, 2003.

[25] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Springer-Verlag, 2001.

[26] H. Saruwatari, N. Hirata, T. Hatta, R. Wakisaka, K. Shikano, and T. Takatani, "Semi-blind speech extraction for robot using visual information and noise statistics," *Proc. of the 11th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT2011)*, pp.238–243, 2011.

[27] A. Lee, K. Nakamura, R. Nishimura, H. Saruwatari, and K. Shikano, “Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs,” *Proc. 8th International Conference on Spoken Language Processing (ICSLP2004)*, vol.I, pp.173–176, 2004.

[28] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol. ASSP-28, no. 4, pp. 357–366, 1982.

[29] T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano, and K. Kondo, "Theoretical analysis of musical noise in generalized spectral subtraction based on higher-order statistics," *IEEE Transactions on Audio, Speech and Language Processing*, vol.19, no.6, pp.1770–1779, 2011.

[30] R. Miyazaki, H. Saruwatari, R. Wakisaka, K. Shikano, and T. Takatani, "Theoretical analysis of parametric blind spatial subtraction array and its application to speech recognition performance prediction," *Proc. of Joint Workshop on Hands-free Speech Communication and Microphone Arrays 2011 (HSCMA2011)*, pp.19–24, 2011.

[31] Y. Ephraim, and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, Signal Processing*, vol.32, no.6, pp.1109–1121, 1984.

[32] R. Okamoto, Y. Takahashi, H. Saruwatari, and K. Shikano, "MMSE STSA estimator with nonstationary noise estimation based on ICA for high-quality speech enhancement," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2010)*, pp.4778–4781, 2010.

[33] H. Saruwatari, M. Go, R. Okamoto, and K. Shikano, "Binaural hearing aid using sound-localization-preserved MMSE STSA estimator with ICA-based noise estimation," *Proc. of International Workshop on Acoustic Echo and Noise Control (IWAENC2010)*, 2010.

[34] Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Musicalnoise analysis in methods of integrating microphone array and spectral subtraction based on higher-order statistics," *EURASIP Journal on Advances in Signal Processing*, vol.2010, Article ID 431347, 25 pages, 2010.

[35] R. Miyazaki, H. Saruwatari, and K. Shikano, "Theoretical analysis of amounts of musical noise and speech distortion in structure-generalized parametric spatial subtraction

array," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol.95-A, no.2, pp.586–590, 2012.

[36] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ: Prentice Hall PTR, 1993.

[37] A. Lee, T. Kawahara, and K. Shikano, “Julius – an open source real-time large vocabulary recognition engine,” *European Conference on Speech Communication and Technology*, pp. 1691–1694, 2001.

[38] R. Mukai, H. Sawada, S. Araki, and S. Makino, “Blind source separation for moving speech signals using blockwise ICA and residual crosstalk subtraction,” *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol.E87-A, no.8, pp.1941–1948, 2004.

[39] H. Buchner, R. Aichner, and W. Kellermann, “A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics,” *IEEE Transactions on Speech and Audio Processing*, vol.13, no.1, pp.120–134, 2005

[40] T. Hiekata, Y. Ikeda, T. Yamashita, T. Morita, R. Zhang, Y. Mori, H. Saruwatari, and K. Shikano, “Development and evaluation of pocket-size real-time blind source separation microphone,” *Acoustical Science and Technology*, vol.30, no.4, pp.297–304, 2009.