Chapter 10
Statistical Analysis and Evaluation of Blind Speech Extraction Algorithms

Hiroshi Saruwatari and Ryoichi Miyazaki

Abstract In this chapter, a problem of blind source separation for speech applications operated under real acoustic environments is addressed. In particular, we focus on a blind spatial subtraction array (BSSA) consisting of a noise estimator based on independent component analysis (ICA) for efficient speech enhancement. First, it is theoretically and experimentally pointed out that ICA is proficient in noise estimation rather than in speech estimation under a nonpoint-source noise condition. Next, motivated by the above-mentioned fact, we introduce a structure-generalized parametric BSSA, which consists of an ICA-based noise estimator and post-filtering based on generalized spectral subtraction. In addition, we perform its theoretical analysis via higher-order statistics. Comparing a parametric BSSA and a parametric channelwise BSSA, we reveal that a channelwise BSSA structure is recommended for listening but a conventional BSSA is more suitable for speech recognition.

10.1 Introduction

A hands-free speech recognition system [1–3] 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.

H. Saruwatari (✉)
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
e-mail: hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp

R. Miyazaki
Nara Institute of Science and Technology, Nara 630-0192, Japan
e-mail: ryoichi-m@is.naist.jp

G. R. Naik and W. Wang (eds.), *Blind Source Separation*,
Signals and Communication Technology, DOI: 10.1007/978-3-642-55016-4_10,

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. There have been various studies on microphone array signal processing; in particular, the delay-and-sum (DS) [4–7] array and adaptive beamformer (ABF) [8–11] 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) [12], in which the independence among source signals is mainly used for the separation, has recently been studied actively [13–22]. 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.

In this chapter, first, we analyze ICA under a nonpoint-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) [23], an improved BSS algorithm recently proposed in order to deal with real acoustic sounds. BSSA consists of an ICA-based noise estimator and post-filtering such as spectral subtraction (SS) [24], where noise reduction in 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. However, BSSA always suffers from artificial distortion, so-called musical noise, owing to nonlinear signal processing. This leads to a serious tradeoff between the noise reduction performance and the amount of signal distortion in speech recognition.

In a recent study, two types of BSSA have been proposed (see Fig. 10.1) [25]. One is the conventional BSSA structure that performs SS after delay-and-sum (DS) (see Fig. 10.1a), and the other involves channelwise SS before DS (chBSSA; see Fig. 10.1b). Also, it has been theoretically clarified that chBSSA is superior to BSSA for the mitigation of the musical noise [26]. Therefore, in this chapter, we generalize

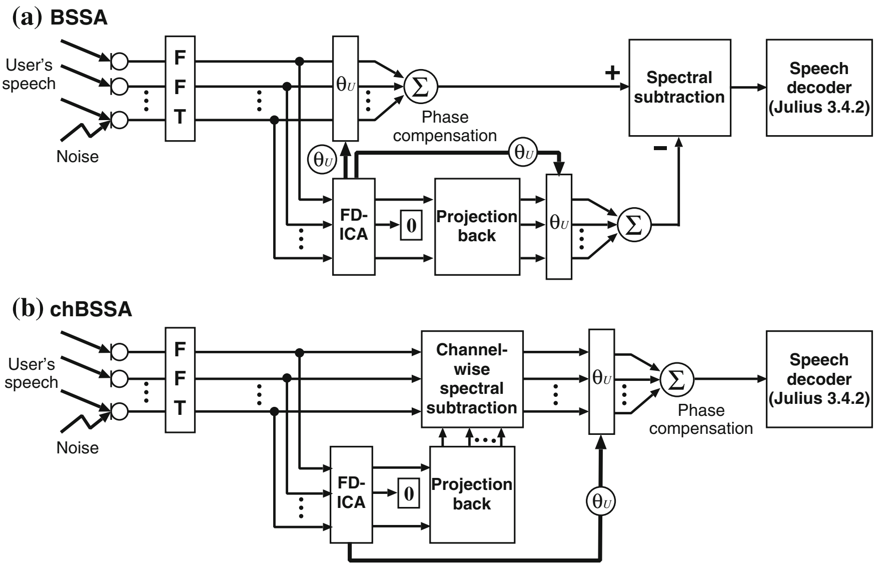


Fig. 10.1 Block diagrams of **a** SS after DS (BSSA) and **b** channelwise SS before DS (chBSSA)

the various types of BSSA as a *structure-generalized parametric BSSA* [27], and we provide a theoretical analysis of the amounts of musical noise and speech distortion generated in several types of methods using the structure-generalized parametric BSSA. From a mathematical analysis based on higher-order statistics, we prove the existence of a tradeoff between the amounts of musical noise and speech distortion in various BSSA structures. From experimental evaluations, it is revealed that the structure should be carefully selected according to the application, i.e., a chBSSA structure is recommended for listening but a conventional BSSA is more suitable for speech recognition.

The outline of this chapter is organized as follows. In Sect. 10.2, we provide a brief review of ICA used for speech applications [28, 29]. In Sect. 10.3, a theoretical analysis of ICA under nonpoint-source noise condition is given, and following this section, we give a review of BSSA and its generalized algorithms [23, 27] in Sect. 10.4. In Sect. 10.5, we describe a musical noise assessment method based on higher-order statistics [30–32]. Using the method, we give a theoretical analysis of musical noise generation and speech distortion for structure-generalized BSSA, where the authors can show that chBSSA is superior to BSSA in terms of less musical noise property, but BSSA is superior to chBSSA in terms of less speech distortion property [27]. In Sect. 10.6, we show results of experimental evaluation [27]. Following a discussion on the theoretical analysis and experimental results, we present our conclusions in Sect. 10.7.

10.2 Data Model and Conventional BSS Method

10.2.1 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. 10.2). Here, we assume the following sound sources: only one target speech signal, some interference signals that can be regarded as point sources, and additive noise. 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

$$\mathbf{x}(f, \tau) = \mathbf{A}(f)\,\{\mathbf{s}(f, \tau) + \mathbf{n}(f, \tau)\} + \mathbf{n}_a(f, \tau), \tag{10.1}$$

where f is the frequency bin and τ is the time index of DFT analysis. Also, $\mathbf{x}(f, \tau)$ is the observed signal vector, $\mathbf{A}(f)$ is the mixing matrix, $\mathbf{s}(f, \tau)$ 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), $\mathbf{n}(f, \tau)$ is the interference signal vector that contains the signal components except the Uth component, and $\mathbf{n}_a(f, \tau)$ is the nonstationary additive noise signal term that generally represents nonpoint-source noises. These are defined as

$$\mathbf{x}(f, \tau) = [x_1(f, \tau), \ldots, x_J(f, \tau)]^{\mathrm{T}}, \tag{10.2}$$

$$\mathbf{s}(f, \tau) = [\underbrace{0, \ldots, 0}_{U-1}, s_U(f, \tau), \underbrace{0, \ldots, 0}_{K-U}]^{\mathrm{T}}, \tag{10.3}$$

$$\mathbf{n}(f, \tau) = [n_1(f, \tau), \ldots, n_{U-1}(f, \tau), 0, n_{U+1}(f, \tau), \ldots, n_K(f, \tau)]^{\mathrm{T}}, \tag{10.4}$$

$$\mathbf{n}_a(f, \tau) = [n_1^{(a)}(f, \tau), \ldots, n_J^{(a)}(f, \tau)]^{\mathrm{T}}, \tag{10.5}$$

$$\mathbf{A}(f) = \begin{bmatrix} A_{11}(f) & \cdots & A_{1K}(f) \\ \vdots & & \vdots \\ A_{J1}(f) & \cdots & A_{JK}(f) \end{bmatrix}. \tag{10.6}$$

10.2.2 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 $\mathbf{n}_a(f, \tau)$ is negligible in (10.1). In frequency-domain ICA (FDICA), signal separation is expressed as

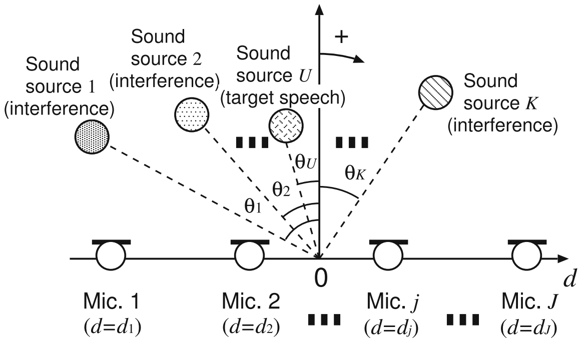

Sound source 1 (interference)
Sound source 2 (interference)
Sound source U (target speech)
Sound source K (interference)
+
θ_U
θ_K
θ_2
θ_1
0
d
Mic. 1 $(d=d_1)$
Mic. 2 $(d=d_2)$
Mic. j $(d=d_j)$
Mic. J $(d=d_J)$

Fig. 10.2 Configurations of microphone array and signals

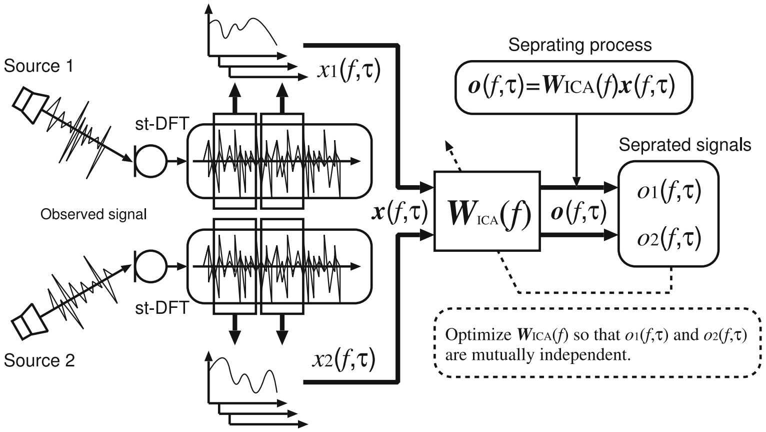


Fig. 10.3 Blind source separation procedure in FDICA in case of $J = K = 2$

$$\mathbf{o}(f,\tau) = [o_1(f,\tau),\ldots,o_K(f,\tau)]^{\mathrm{T}} = \mathbf{W}_{\mathrm{ICA}}(f)\mathbf{x}(f,\tau), \tag{10.7}$$

$$\mathbf{W}_{\mathrm{ICA}}(f) = \begin{bmatrix} W_{11}^{(\mathrm{ICA})}(f) & \cdots & W_{1J}^{(\mathrm{ICA})}(f) \\ \vdots & & \vdots \\ W_{K1}^{(\mathrm{ICA})}(f) & \cdots & W_{KJ}^{(\mathrm{ICA})}(f) \end{bmatrix}, \tag{10.8}$$

where $\mathbf{o}(f,\tau)$ is the resultant output of the separation and $\mathbf{W}_{\mathrm{ICA}}(f)$ is the complex-valued unmixing matrix (see Fig. 10.3).

The unmixing matrix $\mathbf{W}_{\mathrm{ICA}}(f)$ is optimized by ICA so that the output entries of $\mathbf{o}(f,\tau)$ become mutually independent. Indeed, many kinds of ICA algorithms have been proposed. In the second-order ICA (SO-ICA) [17, 19], 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 [17]:

$$\mathbf{W}_{\mathrm{ICA}}^{[p+1]}(f) = -\mu \sum_{\tau_b} \chi(f)\, \text{off-diag}\left(\mathbf{R}_{oo}(f, \tau_b)\right) \mathbf{W}_{\mathrm{ICA}}^{[p]}(f) \mathbf{R}_{xx}(f, \tau_b) + \mathbf{W}_{\mathrm{ICA}}^{[p]}(f), \tag{10.9}$$

where μ is the step-size parameter, $[p]$ is used to express the value of the pth step in iterations, off-diag[$\mathbf{X}$] is the operation for setting every diagonal element of matrix $\mathbf{X}$ to zero, and $\chi(f) = (\sum_{\tau_b} \|\mathbf{R}_{xx}(f, \tau_b)\|^2)^{-1}$ is a normalization factor ($\|\cdot\|$ represents the Frobenius norm). $\mathbf{R}_{xx}(f, \tau_b)$ and $\mathbf{R}_{oo}(f, \tau_b)$ are the cross-power spectra of the input $\mathbf{x}(f, \tau)$ and output $\mathbf{o}(f, \tau)$, respectively, which are calculated around multiple time blocks τ_b. Also, Pham et al. have proposed the following improved criterion for SO-ICA [19]:

$$\sum_{\tau_b} \left\{ \frac{1}{2} \log \det \text{diag}\left[\mathbf{W}_{\mathrm{ICA}}(f) \mathbf{R}_{oo}(f, \tau_b) \mathbf{W}_{\mathrm{ICA}}(f)^{\mathrm{H}}\right] - \log \det \left[\mathbf{W}_{\mathrm{ICA}}(f)\right] \right\}, \tag{10.10}$$

where the superscript H denotes Hermitian transposition. This criterion is to be minimized with respect to $\mathbf{W}_{\mathrm{ICA}}(f)$. Another possible way to achieve SO-ICA has been proposed as the direct joint diagonalization based on the linear algebraic procedure [33, 34].

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 $\mathbf{W}_{\mathrm{ICA}}(f)$ in HO-ICA is obtained using the iterative equation

$$\mathbf{W}_{\mathrm{ICA}}^{[p+1]}(f) = \mu[\mathbf{I} - \langle \boldsymbol{\varphi}(\mathbf{o}(f, \tau)) \mathbf{o}^{\mathrm{H}}(f, \tau) \rangle_\tau] \mathbf{W}_{\mathrm{ICA}}^{[p]}(f) + \mathbf{W}_{\mathrm{ICA}}^{[p]}(f), \tag{10.11}$$

where $\mathbf{I}$ 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 [35, 36]. In this study, we define the nonlinear vector function $\boldsymbol{\varphi}(\cdot)$ as

$$\boldsymbol{\varphi}(\mathbf{o}(f, \tau)) \equiv [\varphi(o_1(f, \tau)), \ldots, \varphi(o_K(f, \tau))]^{\mathrm{T}}, \tag{10.12}$$

$$\varphi(o_k(f, \tau)) \equiv \tanh o_k^{(\mathrm{R})}(f, \tau) + i \tanh o_k^{(\mathrm{I})}(f, \tau), \tag{10.13}$$

where the superscripts (R) and (I) denote the real and imaginary parts, respectively. The nonlinear function given by (10.12) 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 [16] 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. [21, 37], 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 hereafter.

FDICA has the inherent problem so-called *permutation problem*, i.e., difficulty in removing the ambiguity of the source order in each frequency subband. In the context of the permutation problem in the ICA study, there exist many methods for solving the permutation problem, such as the source DOA-based method [38], subband correlation-based method [15], and their combination method [39]. The definite way to avoid the permutation problem is to use time-domain ICA (TDICA), which has, however, other problems like relatively slow convergence and complex implementation. Several literatures can be available for understanding the difference and comparison between TDICA and FDICA [40–42].

10.3 Analysis of ICA Under Nonpoint-source Noise Condition

In this section, we investigate the proficiency of ICA under a nonpoint-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 [43, 44]. 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 nonpoint-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) [18].

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

10.3.1 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. 10.4):

- 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.

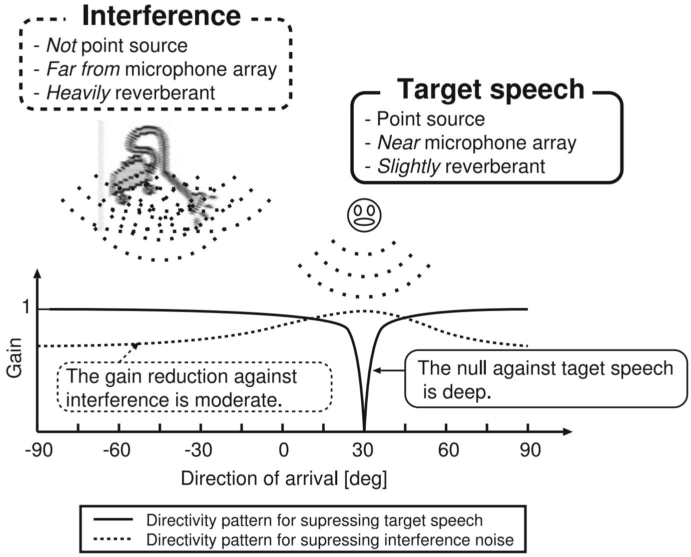


Fig. 10.4 Expected directivity patterns that are shaped by ICA

- 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.

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.

10.3.2 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

$$\mathbf{x}(f,\tau) = \mathbf{A}(f)\mathbf{s}(f,\tau) + \mathbf{n}_a(f,\tau). \tag{10.14}$$

Note that the additive noise $\mathbf{n}_a(f,\tau)$ 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

$$[y_s(f,\tau), y_n(f,\tau)]^{\mathrm{T}} = \mathbf{W}_{\mathrm{ICA}}(f)\mathbf{x}(f,\tau). \tag{10.15}$$

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

$$\mathbf{W}_{\mathrm{ICA}}(f) = [\mathbf{g}_s(f), \mathbf{g}_n(f)]^{\mathrm{T}}, \tag{10.16}$$

where $\mathbf{g}_s(f) = \left[g_1^{(s)}(f), \ldots, g_J^{(s)}(f)\right]^{\mathrm{T}}$ is the coefficient vector of the beamformer used to pick up the target speech signal, and $\mathbf{g}_n(f) = \left[g_1^{(n)}(f), \ldots, g_J^{(n)}(f)\right]^{\mathrm{T}}$ is the coefficient vector of the beamformer used to pick up the noise. Therefore, (10.15) can be rewritten as

$$[y_s(f,\tau), y_n(f,\tau)]^{\mathrm{T}} = [\mathbf{g}_s(f), \mathbf{g}_n(f)]^{\mathrm{T}}\mathbf{x}(f,\tau). \tag{10.17}$$

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

$$\langle \mathbf{o}(f,\tau_b)\mathbf{o}^{\mathrm{H}}(f,\tau_b)\rangle_{\tau_b}, \tag{10.18}$$

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

$$\begin{aligned}
\varphi(o_k(f,\tau)) &= \tanh o_k^{(\mathrm{R})}(f,\tau) + i \tanh o_k^{(\mathrm{I})}(f,\tau),\\
&= \left\{ o_k^{(\mathrm{R})}(f,\tau) - \frac{\left(o_k^{(\mathrm{R})}(f,\tau)\right)^3}{3} + \cdots \right\}\\
&\quad + i \left\{ o_k^{(\mathrm{I})}(f,\tau) - \frac{\left(o_k^{(\mathrm{I})}(f,\tau)\right)^3}{3} + \cdots \right\},\\
&= o_k(f,\tau) - \left(\frac{\left(o_k^{(\mathrm{R})}(f,\tau)\right)^3}{3} + i \frac{\left(o_k^{(\mathrm{I})}(f,\tau)\right)^3}{3} \right) + \cdots .
\end{aligned} \tag{10.19}$$

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

$$\langle \boldsymbol{\varphi}(\mathbf{o}(f,\tau))\mathbf{o}^{\mathrm{H}}(f,\tau)\rangle_\tau = \langle \mathbf{o}(f,\tau)\mathbf{o}^{\mathrm{H}}(f,\tau)\rangle_\tau + \Psi(f), \tag{10.20}$$

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

$$C = \frac{\left|\langle y_s(f,\tau)y_n^*(f,\tau)\rangle_\tau\right|}{\sqrt{\langle |y_s(f,\tau)|^2\rangle_\tau}\sqrt{\langle |y_n(f,\tau)|^2\rangle_\tau}}, \tag{10.21}$$

$$y_s(f,\tau) = \hat{s}(f,\tau) + r_s\hat{n}(f,\tau), \tag{10.22}$$

$$y_n(f,\tau) = \hat{n}(f,\tau) + r_n\hat{s}(f,\tau), \tag{10.23}$$

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

$$\Gamma_s = \langle |\hat{s}(f,\tau)|^2\rangle_\tau / (|r_s|^2\langle |\hat{n}(f,\tau)|^2\rangle_\tau), \tag{10.24}$$

$$\Gamma_n = \langle |\hat{n}(f,\tau)|^2\rangle_\tau / (|r_n|^2\langle |\hat{s}(f,\tau)|^2\rangle_\tau), \tag{10.25}$$

where Γ_s is the SNR of $y_s(f,\tau)$ and Γ_n is the SNR of $y_n(f,\tau)$. Using (10.22)–(10.25), we can rewrite (10.21) as

$$C = \frac{\left|1/\sqrt{\Gamma_s}\cdot e^{j\arg r_s} + 1/\sqrt{\Gamma_n}\cdot e^{j\arg r_n^*}\right|}{\sqrt{1+1/\Gamma_s}\sqrt{1+1/\Gamma_n}} = \frac{\left|1/\sqrt{\Gamma_s} + 1/\sqrt{\Gamma_n}\cdot e^{j(\arg r_n^* - \arg r_s)}\right|}{\sqrt{1+1/\Gamma_s}\sqrt{1+1/\Gamma_n}}, \tag{10.26}$$

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 $\mathbf{g}_s(f)$ and $\mathbf{g}_n(f)$.

In Ref. [23], the following has been proved.

- 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., $\mathbf{g}_s(f)$ and $\mathbf{g}_n(f)$, so that the SNR of the output of each beamformer becomes maximum.

10.3.3 What Beamformers Are Optimized Under Nonpoint-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 nonpoint-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 nonpoint-source noise. First, we focus on the beamformer $\mathbf{g}_s(f)$ that picks up the target speech signal. The SNR-maximize beamformer for $\mathbf{g}_s(f)$ minimizes the undesired signal's power under the condition that the target signal's gain is kept constant. Thus, the desired beamformer should satisfy

$$\min_{\mathbf{g}_s(f)} \mathbf{g}_s^{\mathrm{T}}(f)\mathbf{R}(f)\mathbf{g}_s(f) \quad \text{subject to } \mathbf{g}_s^{\mathrm{T}}(f)\mathbf{a}(f, \theta_s) = 1, \tag{10.27}$$

$$\mathbf{a}(f, \theta_s(f)) = [\exp(i2\pi(f/M)f_{\mathrm{s}}d_1 \sin\theta_s/c), \ldots, \exp(i2\pi(f/M)f_{\mathrm{s}}d_J \sin\theta_s/c)]^{\mathrm{T}}, \tag{10.28}$$

where $\mathbf{a}(f, \theta_s(f))$ 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 $\mathbf{R}(f) = \langle \mathbf{n}_a(f, \tau)\mathbf{n}_a^{\mathrm{H}}(f, \tau)\rangle_\tau$ is the correlation matrix of $\mathbf{n}_a(f, \tau)$. 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 (10.27) is

$$\mathbf{g}_s(f)^{\mathrm{T}} = \frac{\mathbf{a}(f, \theta_s(f))^{\mathrm{H}}\mathbf{R}^{-1}(f)}{\mathbf{a}(f, \theta_s(f))^{\mathrm{H}}\mathbf{R}^{-1}(f)\mathbf{a}(f, \theta_s(f))}. \tag{10.29}$$

This beamformer is called a minimum variance distortionless response (MVDR) beamformer [45]. 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 nonpoint-source noises.

Although the correlation matrix is often not diagonalized in lower frequency subbands [45], 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 $\mathbf{R}(f) = \delta^2(f) \cdot \mathbf{I}$. Therefore, the inverse of the correlation matrix $\mathbf{R}^{-1}(f) = \mathbf{I}/\delta^2(f)$ and (10.29) can be rewritten as

$$\mathbf{g}_s(f)^{\mathrm{T}} = \frac{\mathbf{a}(f, \theta_s(f))^{\mathrm{H}}}{\mathbf{a}(f, \theta_s(f))^{\mathrm{H}}\mathbf{a}(f, \theta_s(f))}. \tag{10.30}$$

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

$$\begin{aligned}&\mathbf{g}_s(f)\\ &= \frac{1}{J}[\exp(-i2\pi(f/M)f_{\mathrm{s}}d_1 \sin\theta_s(f)/c), \ldots, \exp(-i2\pi(f/M)f_{\mathrm{s}}d_J \sin\theta_s(f)/c)]^{\mathrm{T}}.\end{aligned} \tag{10.31}$$

This filter $\mathbf{g}_s(f)$ is approximately equal to a DS beamformer [4]. Note that the filter $\mathbf{g}_s(f)$ 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 nonpoint-source noise condition is proportional to $10\log_{10} J$ [dB]; this performance is not particularly good.

Next, we consider the other beamformer $\mathbf{g}_n(f)$, 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 [18]. 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

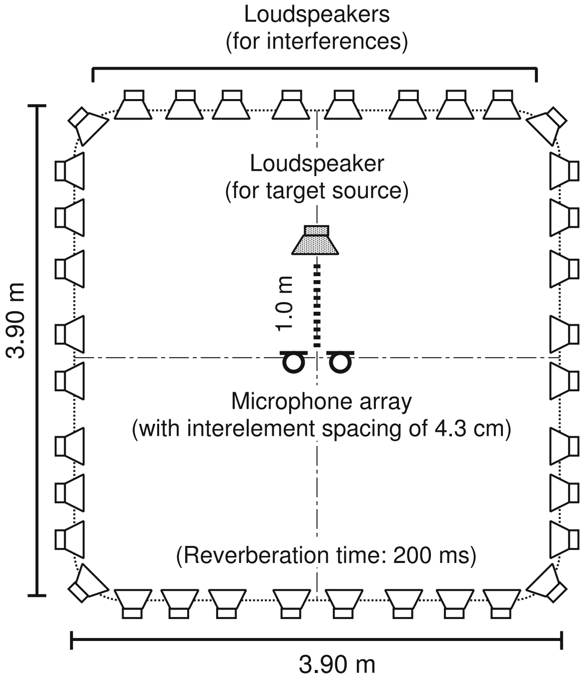
$$\begin{aligned}&\mathbf{g}_n(f)\\ &= [\exp(-i2\pi(f/M)f_{\mathrm{s}}d_1 \sin\theta_s(f)/c), -\exp(-i2\pi(f/M)f_{\mathrm{s}}d_2 \sin\theta_s(f)/c)]^{\mathrm{T}} \cdot \sigma(f),\end{aligned} \tag{10.32}$$

where $\sigma(f)$ is the gain compensation parameter. This beamformer surely satisfies $\mathbf{g}_n^{\mathrm{T}}(f) \cdot \mathbf{a}(f, \theta_s(f)) = 0$. The steering vector $\mathbf{a}(f, \theta_s(f))$ expresses the wavefront of the plane wave arriving from $\theta_s(f)$. Thus, $\mathbf{g}_n(f)$ 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 $\mathbf{g}_n(f)$ 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.

10.3.4 Computer Simulations

We conduct computer simulations to confirm the performance of ICA under a nonpoint-source noise condition. Here, we used HO-ICA [16] as the ICA algorithm. We used the following 8 kHz-sampled signals as the ICA's input; the original target

Fig. 10.5 Layout of reverberant room in our simulation



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 1,600 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 10.5 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. 10.6 and 10.7. Figure 10.6 shows the result for the average noise reduction rate (NRR) [18] 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

$$\text{NRR [dB]} = \frac{1}{J}\sum_{j=1}^{J}(\text{OSNR} - \text{ISNR}_j), \qquad (10.33)$$

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

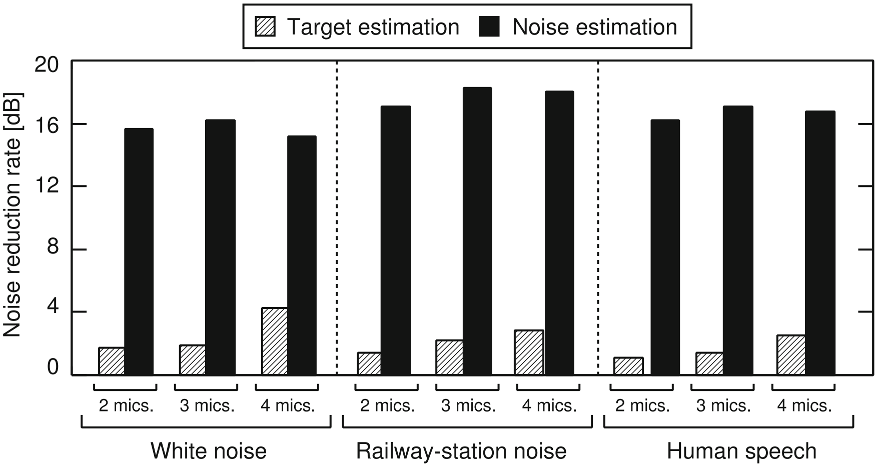
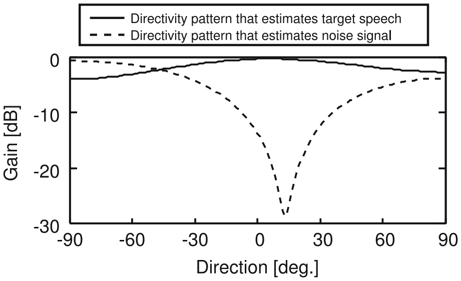


Fig. 10.6 Simulation-based separation results under nonpoint-source noise condition

Fig. 10.7 Typical directivity patterns under nonpoint-source noise condition shaped by ICA at 2 kHz and two-element array for case of white Gaussian noise



is consistent with the previously stated theory. Moreover, Fig. 10.7 shows directivity patterns shaped by the beamformers optimized by ICA in the simulation. It is clearly indicated that beamformer $\mathbf{g}_s(f)$, which picks up the target speech, resembles the DS beamformer, and that beamformer $\mathbf{g}_n(f)$, 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 nonpoint-source noise condition are DS and NBF, is valid.

10.4 Blind Spectral Subtraction Array

10.4.1 Motivation and Strategy

As clearly shown in Sects. 10.3.3 and 10.3.4, ICA is proficient in noise estimation rather than in target speech estimation under a nonpoint-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 [23]. BSSA consists of a DS-based primary path and a reference path including ICA-based noise estimation (see Fig. 10.1a). 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.

10.4.2 Partial Speech Enhancement in Primary Path

We again consider the generalized form of the observed signal as described in (10.1). The target speech signal is partly enhanced in advance by DS. This procedure can be given as

$$\begin{aligned} y_{\mathrm{DS}}(f,\tau) &= \mathbf{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\mathbf{x}(f,\tau) \\ &= \mathbf{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\mathbf{A}(f)\mathbf{s}(f,\tau) + \mathbf{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\mathbf{A}(f)\mathbf{n}(f,\tau) + \mathbf{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\mathbf{n}_a(f,\tau), \end{aligned} \tag{10.34}$$

$$\mathbf{w}_{\mathrm{DS}} = [w_1^{(\mathrm{DS})}(f), \ldots, w_J^{(\mathrm{DS})}(f)]^{\mathrm{T}}, \tag{10.35}$$

$$w_j^{(\mathrm{DS})}(f) = \frac{1}{J}\exp\left(-i2\pi(f/M)f_{\mathrm{s}}d_j \sin\theta_U/c\right), \tag{10.36}$$

where $y_{\mathrm{DS}}(f,\tau)$ is the primary path output that is a slightly enhanced target speech, $\mathbf{w}_{\mathrm{DS}}(f)$ 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. 10.4.3. In (10.34), the second and third terms on the right-hand side express the remaining noise in the output of the primary path.

10.4.3 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 $\mathbf{o}(f,\tau)$ as

$$\mathbf{o}(f,\tau) = \mathbf{W}_{\mathrm{ICA}}(f)\mathbf{x}(f,\tau), \tag{10.37}$$

$$\mathbf{o}(f,\tau) = [o_1(f,\tau), \ldots, o_{K+1}(f,\tau)]^{\mathrm{T}}, \tag{10.38}$$

$$\mathbf{W}_{\mathrm{ICA}}(f) = \begin{bmatrix} W_{11}^{(\mathrm{ICA})}(f) & \cdots & W_{1J}^{(\mathrm{ICA})}(f) \\ \vdots & & \vdots \\ W_{(K+1)1}^{(\mathrm{ICA})}(f) & \cdots & W_{(K+1)J}^{(\mathrm{ICA})}(f) \end{bmatrix}, \tag{10.39}$$

where the unmixing matrix $\mathbf{W}_{\mathrm{ICA}}(f)$ is optimized by (10.11) . 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 $\mathbf{n}_a(f, \tau)$ 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 $\mathbf{n}_a(f, \tau)$ 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. 10.3. Therefore, the estimated noise signal is still beneficial.

Next, we estimate DOAs from the unmixing matrix $\mathbf{W}_{\mathrm{ICA}}(f)$ [18]. This procedure is represented by

$$\theta_u = \sin^{-1} \frac{\arg\left(\frac{[\mathbf{W}_{\mathrm{ICA}}^{-1}(f)]_{ju}}{[\mathbf{W}_{\mathrm{ICA}}^{-1}(f)]_{j'u}}\right)}{2\pi f_{\mathrm{s}} c^{-1}(d_j - d_{j'})}, \tag{10.40}$$

where θ_u is the DOA of the uth sound source. Then, we choose the Uth source signal, which is nearest to 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 to the known target speech DOA.
- For an interaction robot system [46], 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 [47].

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 $\mathbf{o}(f, \tau)$. This can be written as

$$\mathbf{q}(f, \tau) = \left[o_1(f, \tau), ..., o_{U-1}(f, \tau), 0, o_{U+1}(f, \tau), ..., o_{K+1}(f, \tau)\right]^{\mathrm{T}}, \tag{10.41}$$

where $\mathbf{q}(f, \tau)$ is the "noise-only" signal vector that contains only noise components. Next, we apply the projection back (PB) [15] method to remove the ambiguity of amplitude. This procedure can be represented as

$$\hat{\mathbf{q}}(f,\tau) = \left[\hat{q}_1(f,\tau), ..., \hat{q}_J(f,\tau)\right]^{\mathrm{T}} = \mathbf{W}_{\mathrm{ICA}}^{+}(f)\mathbf{q}(f,\tau), \tag{10.42}$$

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

$$\hat{\mathbf{q}}(f,\tau) \simeq \mathbf{A}(f)\mathbf{n}(f,\tau) + \mathbf{W}_{\mathrm{ICA}}^{+}(f)\hat{\mathbf{n}}_a(f,\tau), \tag{10.43}$$

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

$$z_{\mathrm{DS}}(f,\tau) = \mathbf{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\hat{\mathbf{q}}(f,\tau) \simeq \mathbf{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\mathbf{A}(f)\mathbf{n}(f,\tau) + \mathbf{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\mathbf{W}_{\mathrm{ICA}}^{+}(f)\hat{\mathbf{n}}_a(f,\tau). \tag{10.44}$$

This equation means that $z_{\mathrm{DS}}(f,\tau)$ 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 (10.34)). Of course this noise estimation is not perfect, but we can still enhance the target speech signal via *oversubtraction* in the amplitude or power spectrum domain, where the overestimated noise component is subtracted from the observed noisy speech component with an allowance of speech distortion, as described in Sect. 10.4.4. Note that $z_{\mathrm{DS}}(f,\tau)$ is a function of the frame index τ, unlike the constant noise prototype in the traditional SS method [24]. Therefore, the proposed BSSA can deal with *nonstationary* noise.

10.4.4 Formulation of Structure-Generalized Parametric BSSA

In a recent study, two types of BSSA have been proposed (see Fig. 10.1). One is the conventional BSSA structure that performs SS after DS (see Fig. 10.1a), and the other involves channelwise SS before DS (chBSSA; see Fig. 10.1b). Also, it has been theoretically clarified that chBSSA is superior to BSSA for the mitigation of the musical noise generation [26]. In this chapter, we generalize the various types of BSSA as a *structure-generalized parametric BSSA* [27].

First, parametric BSSA is described. Using (10.34) and (10.44), we perform generalized SS (GSS) [48] and obtain the enhanced target speech signal as

$$\begin{aligned} & y_{\mathrm{BSSA}}(f,\tau) \\ & = \begin{cases} \sqrt[2n]{|y_{\mathrm{DS}}(f,\tau)|^{2n} - \beta\overline{|z_{\mathrm{DS}}(f,\tau)|^{2n}}}\, e^{i\,\arg(y_{\mathrm{DS}}(f,\tau))} \\ (\text{if} \quad |y_{\mathrm{DS}}(f,\tau)|^{2n} - \beta\overline{|z_{\mathrm{DS}}(f,\tau)|^{2n}} > 0), \\ 0 \quad (\text{otherwise}), \end{cases} \end{aligned} \tag{10.45}$$

where $y_{\text{BSSA}}(f, \tau)$ is the final output of the parametric BSSA, β is an oversubtraction parameter, n is an exponent parameter, and $\overline{|z_{\text{DS}}(f, \tau)|^{2n}}$ is the smoothed noise component within a certain time frame window.

Next, in the parametric chBSSA, we first perform GSS independently in each input channel and derive multiple enhanced target speech signals by channelwise GSS using (10.2) and (10.42). This procedure can be given by

$$y_j^{(\text{chGSS})}(f, \tau) = \begin{cases} \sqrt[2n]{|x_j(f, \tau)|^{2n} - \beta\overline{|\hat{q}_j(f, \tau)|^{2n}}}\, e^{i \arg(x_j(f,\tau))} \\ (\text{if} \quad |x_j(f, \tau)|^{2n} - \beta\overline{|\hat{q}_j(f, \tau)|^{2n}} > 0), \\ 0 \quad (\text{otherwise}), \end{cases} \tag{10.46}$$

where $y_j^{(\text{chGSS})}(f, \tau)$ is the enhanced target speech signal obtained by GSS at a specific channel j. Finally, we obtain the resultant-enhanced target speech signal by applying DS to $\mathbf{y}_{\text{chGSS}} = [y_1^{(\text{chGSS})}(f, \tau), \ldots, y_J^{(\text{chGSS})}(f, \tau)]^{\text{T}}$. This procedure can be expressed by

$$y_{\text{chBSSA}}(f, \tau) = \mathbf{w}_{\text{DS}}^{\text{T}}(f)\mathbf{y}_{\text{chGSS}}(f, \tau), \tag{10.47}$$

where $y_{\text{chBSSA}}(f, \tau)$ is the final output of the parametric chBSSA.

10.5 Theoretical Analysis of Structure-Generalized Parametric BSSA

10.5.1 Motivation and Strategy

In general, BSSA can achieve good noise reduction performance but always suffers from artificial distortion, so-called musical noise, owing to its nonlinear signal processing. This leads to a serious tradeoff between the noise reduction performance and the amount of signal distortion in speech recognition. Therefore, in this chapter, we provide a theoretical analysis of the amounts of musical noise and speech distortion generated in several types of methods using the structure-generalized parametric BSSA. From a mathematical analysis based on higher-order statistics, we prove the existence of a tradeoff between the amounts of musical noise and speech distortion in various BSSA structures. From experimental evaluations, we reveal that the structure should be carefully selected according to the application, i.e., a chBSSA structure is recommended for listening but a conventional BSSA is more suitable for speech recognition.

In this chapter, we assume that the input signal x in the power spectral domain can be modeled by the gamma distribution as [49, 50]

$$P_{\mathrm{GM}}(x) = \frac{x^{\alpha-1}\exp(-\frac{x}{\theta})}{\theta^{\alpha}\Gamma(\alpha)}, \tag{10.48}$$

where α is the shape parameter corresponding to the type of the signal, θ is the scale parameter of the gamma distribution. In addition, $\Gamma(\alpha)$ is the *gamma function*, defined as

$$\Gamma(\alpha) = \int_{0}^{\infty} t^{\alpha-1} \exp(-t)\mathrm{d}t. \tag{10.49}$$

If the input signal is Gaussian, its complex-valued DFT coefficients also have the Gaussian distributions in the real and imaginary parts. Therefore, the p.d.f. of its power spectra obeys the chi-square distribution with two degrees of freedom, which corresponds to the gamma distribution with $\alpha = 1$. Also, if the input signal is super-Gaussian, the p.d.f. of its power spectra obeys the gamma distribution with $\alpha < 1$.

10.5.2 Analysis of Amount of Musical Noise

10.5.2.1 Metric of Musical Noise Generation: Kurtosis Ratio

We speculate that the amount of musical noise is highly correlated with the number of isolated power spectral components and their level of isolation (see Fig. 10.8). In this chapter, we call these isolated components *tonal components*. Since such tonal components have relatively high power, they are strongly related to the weight of the tail of their probability density function (p.d.f.). Therefore, quantifying the tail of the p.d.f. makes it possible to measure the number of tonal components. Thus, we adopt kurtosis, one of the most commonly used higher-order statistics, to evaluate the percentage of tonal components among all components. A larger kurtosis value indicates a signal with a heavy tail, meaning that the signal has many tonal components. Kurtosis is defined as

$$\mathrm{kurt} = \frac{\mu_4}{\mu_2^2}, \tag{10.50}$$

where “kurt” is the kurtosis and μ_m is the mth-order moment, given by

$$\mu_m = \int_{0}^{\infty} x^m P(x)\mathrm{d}x, \tag{10.51}$$

where $P(x)$ is the p.d.f. of the random variable X. Note that μ_m is not a central moment but a raw moment. Thus, (10.50) is not kurtosis in the mathematically strict

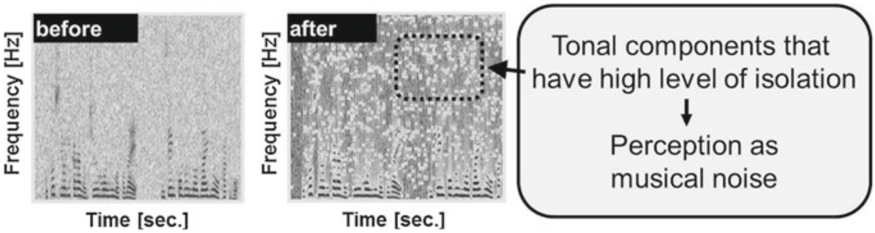


Fig. 10.8 Example of generation of tonal component after signal processing, where input signal is speech with white Gaussian noise and output is processed signal by GSS

definition but a modified version; however, we still refer to (10.50) as kurtosis in this chapter.

In this study, we apply such a kurtosis-based analysis to a time–frequency period of subject signals for the assessment of musical noise. Thus, this analysis should be conducted during, for example, periods of silence in speech when we evaluate the degree of musical noise arising in remaining noise. This is because we aim to quantify the tonal components arising in the noise-only part, which is the main cause of musical noise perception, and not in the target speech-dominant part.

Although kurtosis can be used to measure the number of tonal components, note that the kurtosis itself is not sufficient to measure the amount of musical noise. This is obvious since the kurtosis of some unprocessed noise signals, such as an interfering speech signal, is also high, but we do not recognize speech as musical noise. Hence, we turn our attention to the change in kurtosis between before and after signal processing to identify only the musical noise components. Thus, we adopt the *kurtosis ratio* as a measure to assess musical noise [30–32]. This measure is defined as

$$\text{kurtosis ratio} = \frac{\text{kurt}_{\text{proc}}}{\text{kurt}_{\text{org}}}, \tag{10.52}$$

where $\text{kurt}_{\text{proc}}$ is the kurtosis of the processed signal and kurt_{org} is the kurtosis of the original (unprocessed) signal. This measure increases as the amount of generated musical noise increases. In Ref. [30], it was reported that the kurtosis ratio is strongly correlated with the human perception of musical noise. Figure 10.9 shows an example of the relation between the kurtsis ratio (in log scale) and a human-perceptual score of degree of musical noise generation, where we can confirm the strong correlation.

10.5.2.2 Analysis in the Case of Parametric BSSA

In this section, we analyze the kurtosis ratio in a parametric BSSA. First, using the shape parameter of input noise α_{n}, we express the kurtosis of a gamma distribution, $\text{kurt}_{\text{in}}^{(\text{n})}$, as [51]

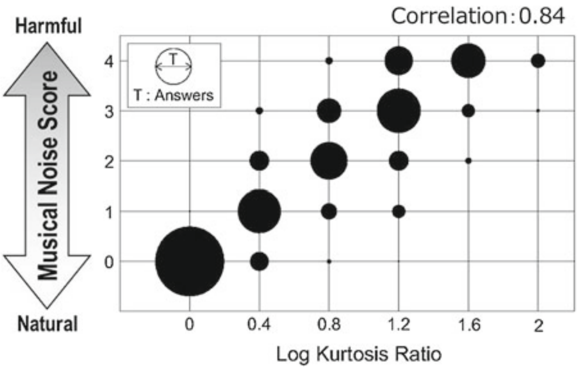


Fig. 10.9 Relation between kurtsis ratio (in log scale) and human-perceptual score of degree of musical noise generation [30]

$$\mathrm{kurt}_{\mathrm{in}}^{(\mathrm{n})} = \frac{\int_0^\infty x^4 P_{\mathrm{GM}}(x)\mathrm{d}x}{\left(\int_0^\infty x^2 P_{\mathrm{GM}}(x)\mathrm{d}x\right)^2} \tag{10.53}$$

$$= \frac{(\alpha_{\mathrm{n}}+2)(\alpha_{\mathrm{n}}+3)}{\alpha_{\mathrm{n}}(\alpha_{\mathrm{n}}+1)}. \tag{10.54}$$

The kurtosis in the power spectral domain after DS is given by [26]

$$\mathrm{kurt}_{\mathrm{DS}}^{(\mathrm{n})} \simeq J^{-0.7} \cdot (\mathrm{kurt}_{\mathrm{in}}^{(\mathrm{n})} - 6) + 6. \tag{10.55}$$

Similarly to (10.53), the shape parameter α_{DS} corresponding to the kurtosis after DS, $\mathrm{kurt}_{\mathrm{DS}}$, is given by solving the following equation in α_{DS}:

$$\mathrm{kurt}_{\mathrm{DS}}^{(\mathrm{n})} = \frac{(\alpha_{\mathrm{DS}}+2)(\alpha_{\mathrm{DS}}+3)}{\alpha_{\mathrm{DS}}(\alpha_{\mathrm{DS}}+1)}. \tag{10.56}$$

This can be expanded as

$$\alpha_{\mathrm{DS}}^2(\mathrm{kurt}_{\mathrm{DS}}^{(\mathrm{n})} - 1) + \alpha_{\mathrm{DS}}(\mathrm{kurt}_{\mathrm{DS}}^{(\mathrm{n})} - 5) - 6 = 0, \tag{10.57}$$

and we have

$$\alpha_{\mathrm{DS}} = \frac{-\mathrm{kurt}_{\mathrm{DS}} + 5 + \sqrt{\mathrm{kurt}_{\mathrm{DS}}^2 + 14\,\mathrm{kurt}_{\mathrm{DS}} + 1}}{2\,\mathrm{kurt}_{\mathrm{DS}} - 2}. \tag{10.58}$$

Then, using (10.53) and (10.55), α_{DS} can be expressed in terms of α_{n} as

$$\begin{aligned}\alpha_{\mathrm{DS}} = &\left[2J^{-0.7}\cdot\left\{\frac{(\alpha_{\mathrm{n}}+2)(\alpha_{\mathrm{n}}+3)}{\alpha_{\mathrm{n}}(\alpha_{\mathrm{n}}+1)}-6\right\}+10\right]^{-1}\\ &\cdot\left[\left\{\left(J^{-0.7}\cdot\left\{\frac{(\alpha_{\mathrm{n}}+2)(\alpha_{\mathrm{n}}+3)}{\alpha_{\mathrm{n}}(\alpha_{\mathrm{n}}+1)}-6\right\}+6\right)^{2}\right.\right.\\ &\left.+14J^{-0.7}\cdot\left\{\frac{(\alpha_{\mathrm{n}}+2)(\alpha_{\mathrm{n}}+3)}{\alpha_{\mathrm{n}}(\alpha_{\mathrm{n}}+1)}-6\right\}+85\right\}^{0.5}\\ &\left.-\left(J^{-0.7}\cdot\left\{\frac{(\alpha_{\mathrm{n}}+2)(\alpha_{\mathrm{n}}+3)}{\alpha_{\mathrm{n}}(\alpha_{\mathrm{n}}+1)}-6\right\}\right)-1\right]. \qquad (10.59)\end{aligned}$$

Next, we calculate the change in kurtosis after parametric BSSA. With the shape parameter after DS, α_{DS}, the resultant kurtosis after the parametric BSSA is represented as

$$\mathrm{kurt}_{\mathrm{BSSA}}^{(\mathrm{n})} = \mathscr{M}(\alpha_{\mathrm{DS}}, \beta, 4, n)/\mathscr{M}^{2}(\alpha_{\mathrm{DS}}, \beta, 2, n), \qquad (10.60)$$

where $\mathscr{M}(\alpha, \beta, m, n)$ is referred to as *normalized moment function* that represents the resultant mth-order moment after GSS in the case that the oversubtraction parameter is β, the exponent parameter is n and the input signal's shape parameter is α. This can be expressed as [52]

$$\begin{aligned}\mathscr{M}(\alpha, \beta, m, n) = &\sum_{l=0}^{m/n}\frac{(-\beta)^{l}\Gamma^{l}(\alpha+n)\Gamma(m/n+1)}{\Gamma^{l+1}(\alpha)\Gamma(l+1)\Gamma(m/n-l+1)}\\ &\Gamma\left(\alpha+m-nl, \left(\beta\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}\right)^{\frac{1}{n}}\right), \qquad (10.61)\end{aligned}$$

where $\Gamma(\alpha, z)$ is the upper incomplete gamma function

$$\Gamma(\alpha, z) = \int_{z}^{\infty} t^{\alpha-1}\exp(-t)\mathrm{d}t. \qquad (10.62)$$

Finally, using (10.52), (10.53), and (10.60),

we can determine the resultant kurtosis ratio through a parametric BSSA as

$$\text{kurtosis ratio}_{\mathrm{BSSA}}^{(\mathrm{n})} = \mathrm{kurt}_{\mathrm{BSSA}}^{(\mathrm{n})}/\mathrm{kurt}_{\mathrm{in}}^{(\mathrm{n})}. \qquad (10.63)$$

10.5.2.3 Analysis in the Case of Parametric chBSSA

In this section, we analyze the kurtosis ratio in a parametric chBSSA. First, we calculate the change in kurtosis after channelwise GSS. Using (10.60) with the shape parameter of input noise α_n, we can express the resultant kurtosis after channelwise GSS as

$$\mathrm{kurt}_{\mathrm{chGSS}}^{(\mathrm{n})} = \mathscr{M}(\alpha_\mathrm{n}, \beta, 4, n)/\mathscr{M}^2(\alpha_\mathrm{n}, \beta, 2, n). \tag{10.64}$$

Next, using (10.55) and (10.64), we can derive the change in kurtosis after a parametric chBSSA as

$$\mathrm{kurt}_{\mathrm{chBSSA}}^{(\mathrm{n})} \simeq J^{-0.7} \cdot (\mathrm{kurt}_{\mathrm{chGSS}}^{(\mathrm{n})} - 6) + 6. \tag{10.65}$$

Finally, we can obtain the resultant kurtosis ratio through a parametric chBSSA as

$$\text{kurtosis ratio}_{\mathrm{chBSSA}}^{(\mathrm{n})} = \mathrm{kurt}_{\mathrm{chBSSA}}^{(\mathrm{n})}/\mathrm{kurt}_{\mathrm{in}}^{(\mathrm{n})}. \tag{10.66}$$

10.5.3 Analysis of Amount of Speech Distortion

10.5.3.1 Analysis in the Case of BSSA

In this section, we analyze the amount of speech distortion on the basis of the kurtosis ratio in speech components. Hereafter, we define $s(f, \tau)$ and $n(f, \tau)$ as the observed speech and noise components at each microphone, respectively. Assuming that speech and noise are disjoint, i.e., there is no overlap in the time–frequency domain, speech distortion is caused by subtracting the average noise from the pure speech component.

Thus, the distorted speech after BSSA is given by

$$\begin{aligned} |s_{\mathrm{BSSA}}(f, \tau)| &= \sqrt[2n]{|s(f, \tau)|^{2n} - \beta\overline{|z_{\mathrm{DS}}(f, \tau)|^{2n}}} \\ &= \sqrt[2n]{|s(f, \tau)|^{2n} - \beta C_{\mathrm{BSSA}}\overline{|s(f, \tau)|^{2n}}}, \end{aligned} \tag{10.67}$$

where $s_{\mathrm{BSSA}}(f, \tau)$ is the output speech component in BSSA. Also, calculating the nth-order moment of the gamma distribution, C_{BSSA} is given by

$$\begin{aligned} C_{\mathrm{BSSA}} &= \overline{|z_{\mathrm{DS}}(f, \tau)|^{2n}}/\overline{|s(f, \tau)|^{2n}} \\ &= J^{-n}\overline{|n(f, \tau)|^{2n}}/\overline{|s(f, \tau)|^{2n}} \\ &= J^{-n}\left(\frac{\alpha_\mathrm{s}}{\alpha_\mathrm{n}}\right)^n \frac{\Gamma(lpha_\mathrm{n} + n)/\,\Gamma(\alpha_\mathrm{n})}{\Gamma(\alpha_\mathrm{s} + n)/\,\Gamma(\alpha_\mathrm{s})} \left(\frac{\overline{|n(f, \tau)|^2}}{\overline{|s(f, \tau)|^2}}\right)^n, \end{aligned} \tag{10.68}$$

where α_s is the shape parameter of the input speech. Equation (10.68) indicates that the speech distortion increases when the input SNR, $\overline{|s(f, \tau)|^2}/\overline{|n(f, \tau)|^2}$, and/or the number of microphones, J, decreases. Using (10.61) and (10.68) with the input speech shape parameter α_s, we can obtain the speech kurtosis ratio through BSSA as

$$\begin{aligned}&\text{kurtosis ratio}^{(\mathrm{s})}_{\mathrm{BSSA}}\\&\quad=\frac{\mathscr{M}(\alpha_\mathrm{s},\beta C_\mathrm{BSSA},4,n)}{\mathscr{M}^2(\alpha_\mathrm{s},\beta C_\mathrm{BSSA},2,n)}\frac{\alpha_\mathrm{s}(\alpha_\mathrm{s}+1)}{(\alpha_\mathrm{s}+2)(\alpha_\mathrm{s}+3)}.\end{aligned}\tag{10.69}$$

10.5.3.2 Analysis in the Case of chBSSA

In chBSSA, since channelwise GSS is performed before DS, C_BSSA is therefore replaced with

$$\begin{aligned}C_\mathrm{chBSSA}&=(\overline{|n(f,\tau)|^{2n}}/\overline{|s(f,\tau)|^{2n}})\\&=\left(\frac{\alpha_\mathrm{s}}{\alpha_\mathrm{n}}\right)^n\frac{\Gamma(\alpha_n+n)/\Gamma(\alpha_\mathrm{n})}{\Gamma(\alpha_\mathrm{s}+n)/\Gamma(\alpha_\mathrm{s})}\left(\frac{\overline{|n(f,\tau)|^2}}{\overline{|s(f,\tau)|^2}}\right)^n.\end{aligned}\tag{10.70}$$

Equation (10.70) indicates that the speech distortion increases only when the input SNR decreases, regardless of the number of microphones. Thus, the distortion does not change even if we prepare many microphones, unlike the case of a parametric BSSA. Using (10.61) and (10.70) with α_s, we can obtain the speech kurtosis ratio through chBSSA as

$$\begin{aligned}&\text{kurtosis ratio}^{(\mathrm{s})}_{\mathrm{chBSSA}}\\&\quad=\frac{\mathscr{M}(\alpha_\mathrm{s},\beta C_\mathrm{chBSSA},4,n)}{\mathscr{M}^2(\alpha_\mathrm{s},\beta C_\mathrm{chBSSA},2,n)}\frac{\alpha_\mathrm{s}(\alpha_\mathrm{s}+1)}{(\alpha_\mathrm{s}+2)(\alpha_\mathrm{s}+3)}.\end{aligned}\tag{10.71}$$

10.5.4 Comparison of Amounts of Musical Noise and Speech Distortion Under Same Amount of Noise Reduction

According to the previous analysis, we can compare the amounts of musical noise and speech distortion among a parametric BSSA and a parametric chBSSA under the same NRR (output SNR–input SNR in dB). Figure 10.10 shows the theoretical behaviors of the noise kurtosis ratio and speech kurtosis ratio. In Fig. 10.10a, b, the shape parameter of input noise, α_n, is set to 0.95 and 0.83, corresponding to almost white Gaussian noise and railway station noise, respectively. Also, in Fig. 10.10c, d, the shape parameter of input speech, α_s, is set to 0.1, and the input SNR is set to 10 and 5 dB, respectively. Here, we assume an eight-element array with the interelement spacing of 2.15 cm. The NRR is varied from 11 to 15 dB, and the oversubtraction parameter β is adjusted so that the target speech NRR is achieved. In the parametric

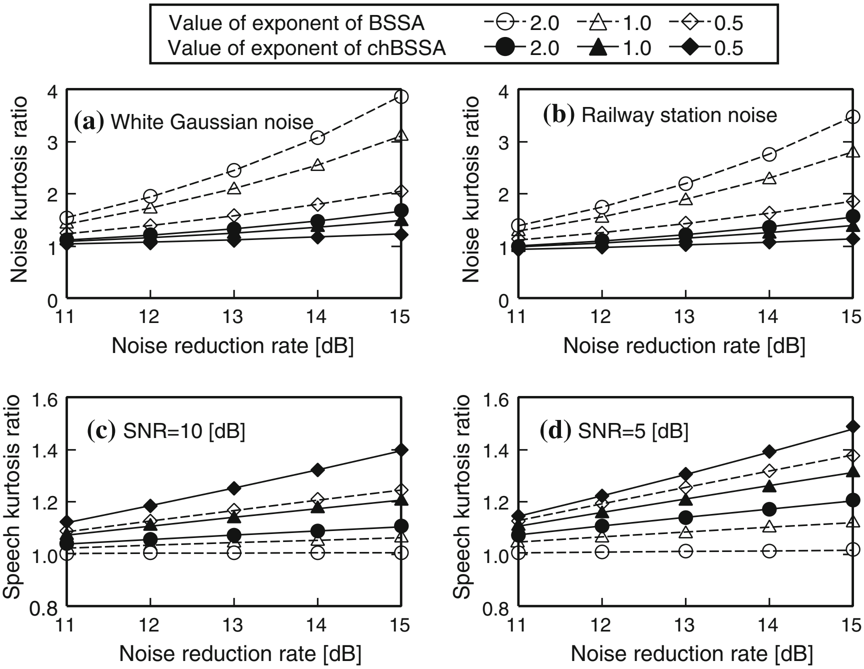


Fig. 10.10 **a** and **b** are theoretical behaviors of noise kurtosis ratio in structure-generalized parametric BSSA. **a** is for white Gaussian noise and **b** is for railway station noise. **c** and **d** are theoretical behaviors of speech kurtosis ratio in structure-generalized parametric BSSA, where the input SNR is set to 10 and 5 dB, respectively

BSSA and parametric chBSSA, the signal exponent parameter $2n$ is set to 2.0, 1.0, and 0.5.

Figure 10.10a, b indicates that the noise kurtosis ratio of chBSSA is smaller than that of BSSA, i.e., less musical noise is generated in a parametric chBSSA than in a parametric BSSA, and a smaller amount of musical noise is generated when a lower exponent parameter is used, regardless of the type of noise and NRR. However, Fig. 10.10c, d shows that speech distortion is lower in a parametric BSSA than in a parametric chBSSA, and a small amount of speech distortion is generated when a higher exponent parameter is used, regardless of the type of noise and NRR. These results theoretically prove the existence of a tradeoff between the amounts of musical noise and speech distortion in BSSA and chBSSA.

10.6 Experiment

10.6.1 Experimental Setup

In this study, we conducted a speech recognition experiment. We used an eight-element microphone array with an interelement spacing of 2.15 cm, and the direction of the target speech was set to be normal to the array. The size of the experimental room is $4.2 \times 3.5 \times 3.0\,\mathrm{m}^3$ and the reverberation time is approximately 200 ms. All the signals used in this experiment are sampled at 16 kHz with 16-bit accuracy. The observed signal consists of the target speech signal of 200 speakers (100 males and 100 females) and two types of diffuse noise (white Gaussian noise and railway station noise) emitted from eight surrounding loudspeakers. The input SNR of the test data is set to 3, 5, and 10 dB. The FFT size is 1,024, and the frame shift length is 256 in BSSA. The speech recognition task is a 20k-word Japanese newspaper dictation, where we used Julius 3.4.2 [53] as the speech decoder. The acoustic model is a phonetic-tied mixture [53], and we use 260 speakers (150 sentences/speaker) to train the acoustic model. In this experiment, the NRR, i.e., the target SNR improvement, is set to 10 dB for white Gaussian noise and 5 dB for railway station noise, the exponent parameter $2n$ is set to 1.0 and 0.5, and the oversubtraction parameter β is adjusted so that the target NRR is achieved.

10.6.2 Evaluation of Speech Recognition Performance and Discussion

Figure 10.11 shows the results of word accuracy in the parametric BSSA and parametric chBSSA, which reveal that better speech recognition performance can be obtained in a parametric BSSA when the input SNR is low (e.g., 3 dB).

This result is of considerable interest because Takahashi et al. [26] reported a contradictory result, i.e., the sound quality of chBSSA is always superior to that of BSSA. Indeed, we conducted a subjective evaluation. We presented 56 pairs of signals processed by a parametric BSSA and a parametric chBSSA, selected from sentences used in the speech recognition experiment, in random order to 10 examinees, who selected which signal they preferred. The result is shown in Fig. 10.12, confirming that chBSSA is preferred by humans, in contrast to the speech recognition results. This is partially true regarding noise distortion, i.e., the amount of musical noise generated, as theoretically shown in Fig. 10.10a, b. Thus, the human evaluation is strongly affected by noise distortion.

However, as shown in Fig. 10.10c, d, the speech distortion in chBSSA is larger than that in BSSA; this leads to the degradation of speech recognition performance. In summary, we should carefully select the structure of signal processing in BSSA, i.e., chBSSA is recommended for listening but BSSA is suitable for speech recognition under a low-input SNR condition.

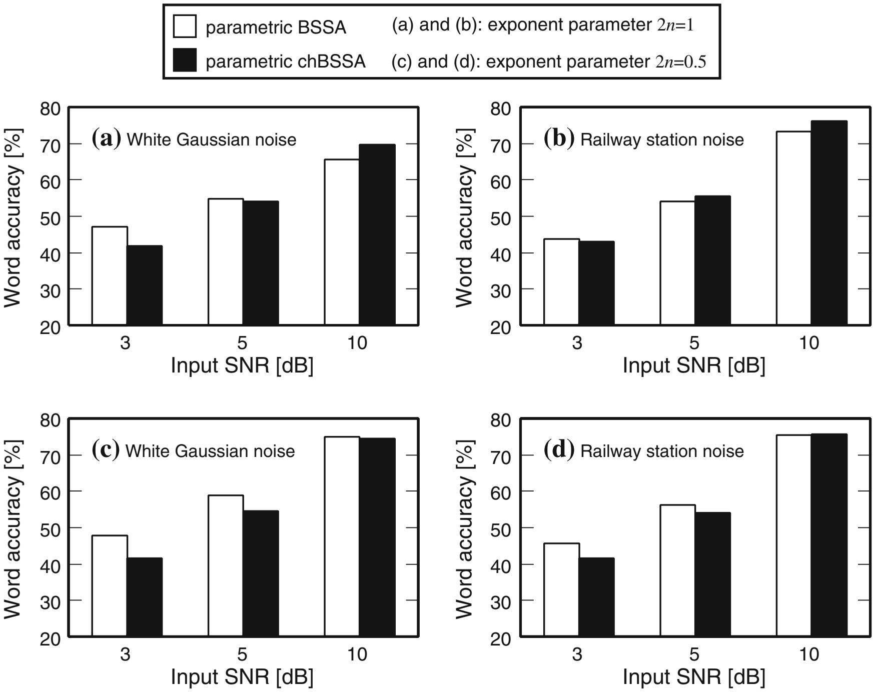


Fig. 10.11 Results of word accuracy in parametric BSSA and parametric chBSSA

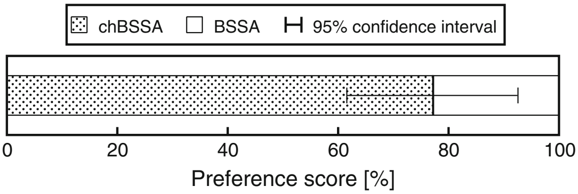


Fig. 10.12 Subjective evaluation results: BSSA versus chBSSA

10.7 Conclusions and Remarks

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 nonpoint-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 nonpoint-source noise condition. Therefore, it is valid to use ICA as a noise estimator.

Motivated by the above-mentioned fact, we introduced a structure-generalized parametric BSSA, which consists of an ICA-based noise estimator and GSS-based post-filtering. In addition, we performed its theoretical analysis via higher-order statistics. Comparing a parametric BSSA and parametric chBSSA, we revealed that a channelwise BSSA structure is recommended for listening but a conventional BSSA is more suitable for speech recognition.

In this chapter, the SS-based BSSAs, which involve SS-based post-filtering, were mainly addressed. Recent studies have provided the further extended methods that include other types of post-filtering, such as Wiener filtering [54, 55], the minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator [56, 57], and the combination method with cepstral smoothing for mitigating musical noise [58]. Also, the theoretical analysis based on the higher-order statistics for these methods is available in several literatures [59–63]. In addition, thanks to the same higher-order statistics analysis, *musical-noise-free* post-filtering [64], in which no musical noise is perfectly generated, has been proposed, and successfully introduced into the channelwise BSSA architecture [65, 66].

BSS implementation on a small hardware still receives much attention in industrial applications. Due to the limitation of space, however, the authors skip the discussion on this issue. Instead, several studies [21, 67, 68] have dealt with the issue of real-time implementation of ICA and BSSA, which would be helpful for the readers.

References

1. Juang, B.H., Soong, F.K.: Hands-free telecommunications. In: Proceedings of International Conference on Hands-Free, Speech Communication, pp. 5–10 (2001)
2. Prasad, R., Saruwatari, H., Shikano, K.: Robots that can hear, understand and talk. Adv. Robot. **18**(5), 533–564 (2004)
3. Saruwatari, H., Kawanami, H., Takeuchi, S., Takahashi, Y., Cincarek, T., Shikano, K.: Hands-free speech recognition challenge for real-world speech dialogue systems. In: Proceedings of 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2009), pp. 3729–3782 (2009)
4. Flanagan, J.L., Johnston, J.D., Zahn, R., Elko, G.W.: Computer-steered microphone arrays for sound transduction in large rooms. J. Acoust. Soc. Am. **78**(5), 1508–1518 (1985)
5. Omologo, M., Matassoni, M., Svaizer, P., Giuliani, D.: Microphone array based speech recognition with different talker-array positions. In: Proceedings of ICASSP'97, pp. 227–230 (1997)
6. Silverman, H.F., Patterson, W.R.: Visualizing the performance of large-aperture microphone arrays. In: Proceedings of ICASSP'99, pp. 962–972 (1999)
7. Saruwatari, H., Kajita, S., Takeda, K., Itakura, F.: Speech enhancement using nonlinear microphone array based on complementary beamforming. IEICE Trans. Fundam. **E82-A**(8), 1501–1510 (1999)
8. Frost, O.: An algorithm for linearly constrained adaptive array processing. Proc. IEEE **60**, 926–935 (1972)
9. Griffiths, L.J., Jim, C.W.: An alternative approach to linearly constrained adaptive beamforming. IEEE Trans. Antennas Propag. **30**(1), 27–34 (1982)
10. Kaneda, Y. Ohga, J.: Adaptive microphone-array system for noise reduction. IEEE Trans. Acoust. Speech Signal Process. **34**(6),1391–1400 (1986)
11. Saruwatari, H., Kajita, S., Takeda, K., Itakura, F.: Speech enhancement using nonlinear microphone array based on noise adaptive complementary beamforming. IEICE Trans. Fundam. **E83-A**(5), 866–876 (2000)

12. Comon, P.: Independent component analysis, a new concept? Signal Process. **36**, 287–314 (1994)
13. Cardoso, J.F.: Eigenstructure of the 4th-order cumulant tensor with application to the blind source separation problem. In: Proceedings of ICASSP’89, pp. 2109–2112 (1989)
14. Jutten, C., Herault, J.: Blind separation of sources Part I: an adaptive algorithm based on neuromimetic architecture. Signal Process. **24**, 1–10 (1991)
15. Ikeda, S., Murata, N.: A method of ICA in the frequency domain. In: Proceedings of International Workshop on Independent Component Analysis and Blind, Signal Separation, pp. 365–371 (1999)
16. Smaragdis, P.: Blind separation of convolved mixtures in the frequency domain. Neurocomputing **22**(1–3), 21–34 (1998)
17. Parra, L., Spence, C.: Convolutive blind separation of non-stationary sources. IEEE Trans. Speech Audio Process. **8**, 320–327 (2000)
18. Saruwatari, H., Kurita, S., Takeda, K., Itakura, F., Nishikawa, T.: Blind source separation combining independent component analysis and beamforming. EURASIP J. Appl. Signal Process. **2003**, 1135–1146 (2003)
19. Pham, D.-T., Serviere, C., Boumaraf, H.: Blind separation of convolutive audio mixtures using nonstationarity. In: International Symposium on Independent Component Analysis and Blind, Signal Separation (ICA2003), pp. 975–980 (2003)
20. Saruwatari, H., Kawamura, T., Nishikawa, T., Lee, A., Shikano, K.: Blind source separation based on a fast-convergence algorithm combining ICA and beamforming. IEEE Trans. Speech Audio Process. **14**(2), 666–678 (2006)
21. Mori, Y., Saruwatari, H., Takatani, T., Ukai, S., Shikano, K., Hiekata, T., Ikeda, Y., Hashimoto, H., Morita, T.: Blind separation of acoustic signals combining SIMO-model-based independent component analysis and binary masking. EURASIP J. Appl. Signal Process. **2006**, ArticleID 34970, 17 (2006)
22. Prasad, R., Saruwatari, H., Shikano, K.: Enhancement of speech signals separated from their convolutive mixture by FDICA algorithm. Digit. Signal Process. **19**(1), 127–133 (2009)
23. Takahashi, Y., Takatani, T., Osako, K., Saruwatari, H., Shikano, K.: Blind spatial subtraction array for speech enhancement in noisy environment. IEEE Trans. Audio Speech Lang. Process. **17**(4), 650–664 (2009)
24. Boll, S.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. **ASSP-27**(2), 113–120 (1979)
25. Saruwatari, H., Takahashi, Y., Shikano, K., Kondo, K.: Blind speech extraction combining ICA-based noise estimation and less-musical-noise nonlinear post processing. In: Proceedings of 2010 Asilomar Conference on Signals, Systems, and Computers, pp. 1415–1419 (2010)
26. Takahashi, Y., Saruwatari, H., Shikano, K., Kondo, K.: Musical-noise analysis in methods of integrating microphone array and spectral subtraction based on higher-order statistics. EURASIP J. Adv. Signal Process. **2010**, Article ID 431347, 25 (2010)
27. Miyazaki, R., Saruwatari, H., Shikano, K.: Theoretical analysis of amount of musical noise and speech distortion in structure-generalized parametric blind spatial subtraction array. IEICE Trans. Fundam. **95-A**(2), 586–590 (2011)
28. Saruwatari, H., Takatani, T., Shikano, K.: SIMO-model-based blind source separation -principle and its applications. In: Makino, S., et al. (eds.) Blind Speech Separation, pp. 149–168. Springer, New York (2007). ISBN 978-1-4020-6479-1
29. Saruwatari, H., Takahashi, Y.: Blind source separation for speech application under real acoustic environment. In: Naik, G. (ed.) Independent Component Analysis for Audio and Biosignal Applications, pp. 41–66. InTech Publishing, Rijeka (2012). ISBN 978-953-51-0782-8
30. Uemura, Y., Takahashi, Y., Saruwatari, H., Shikano, K., Kondo, K.: Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics. In: Proceedings of 2008 International Workshop on Acoustic Echo and Noise, Control (IWAENC2008) (2008)

31. Uemura, Y., Takahashi, Y., Saruwatari, H., Shikano, K., Kondo, K.: Musical noise generation analysis for noise reduction methods based on spectral subtraction and MMSE STSA estimation. In: Proceedings of 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2009), pp. 4433–4436 (2009)
32. Takahashi, Y., Miyazaki, R., Saruwatari, H., Kondo, K.: Theoretical analysis of musical noise in nonlinear noise reduction based on higher-order statistics. In: Proceedings of 2012 APSIPA Annual Summit and Conference (APSIPA2012) (2012)
33. Tachibana, K., Saruwatari, H., Mori, Y., Miyabe, S., Shikano, K. Tanaka, A.: Efficient blind source separation combining closed-form second-order ICA and nonclosed-form higher-order ICA. In: Proceedings of 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2007), vol. 1, pp. 45–48 (2007)
34. Saruwatari, H., Takahashi, Y., Tachibana, K., Mori, Y., Miyabe, S., Shikano, K., Tanaka, A.: Fast and versatile blind separation of diverse sounds using closed-form estimation of probability density functions of sources. In: Proceedings of 3rd International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP2009), pp. 249–252 (2009)
35. Lee, T.-W.: Independent Component Analysis. Kluwer Academic, Norwell (1998)
36. Prasad, R., Saruwatari, H., Shikano, K.: Probability distribution of time-series of speech spectral components. IEICE Trans. Fundam. **E87-A**(3), 584–597 (2004)
37. Ukai, S., Takatani, T., Nishikawa, T., Saruwatari, H.: Blind source separation combining SIMO-model-based ICA and adaptive beamforming. In: Proceedings of 2005 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2005), vol. 3, pp. 85–88 (2005)
38. Kurita, S., Saruwatari, H., Kajita, S., Takeda, K., Itakura, F.: Evaluation of blind signal separation method using directivity pattern under reverberant conditions. In: Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2000), no. SAM-P2-5, pp. 3140–3143 (2000)
39. Sawada, H., Mukai, R., Araki, S., Makino, S.: A robust and precise method for solving the permutation problem of frequency-domain blind source separation. IEEE Trans. Speech Audio Process. **12**(5), 530–538 (2004)
40. Nishikawa, T., Saruwatari, H., Shikano, K.: Blind source separation of acoustic signals based on multistage ICA combining frequency-domain ICA and time-domain ICA. In: IEICE Trans. Fundam. **E86-A**(4), 846–858 (2003)
41. Nishikawa, T., Abe, H., Saruwatari, H., Shikano, K., Kaminuma, A.: Overdetermined blind separation for real convolutive mixtures of speech based on multistage ICA using subarray processing. IEICE Trans. Fundam. **E87-A**(8), 1924–1932 (2004)
42. Araki, S., Makino, S., Aichner, R., Nishikawa, T., Saruwatari, H.: Subband-based blind separation for convolutive mixtures of speech. IEICE Trans. Fundam. **E88-A**(12), 3593–3603 (2005)
43. Araki, S., Mukai, R., Makino, S., Nishikawa, T., Saruwatari, H.: The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. IEEE Trans. Speech Audio Process. **11**(2), 109–116 (2003)
44. Araki, S., Makino, S., Hinamoto, Y., Mukai, R., Nishikawa, T., Saruwatari, H.: Equivalence between frequency domain blind source separation and frequency domain adaptive beamforming for convolutive mixtures. EURASIP J. Appl. Signal Process. **2003**(11), 1157–1166 (2003)
45. Brandstein, M., Ward, D. (eds.): Microphone Arrays: Signal Processing Techniques and Applications. Springer, New York (2001)
46. Saruwatari, H., Hirata, N., Hatta, T., Wakisaka, R., Shikano, K., Takatani, T.: Semi-blind speech extraction for robot using visual information and noise statistics. In: Proceedings of 11th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT2011), pp. 238–243 (2011)
47. Lee, A., Nakamura, K., Nishimura, R., Saruwatari, H., Shikano, K.: Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs. In: Proceedings of 8th International Conference on Spoken Language Processing (ICSLP2004), vol. 1, pp. 173–176 (2004)

48. Sim, B.L., Tong, Y.C., Chang, J.S., Tan, C.T.: A parametric formulation of the generalized spectral subtraction method. IEEE Trans. Speech Audio Process. **6**(4), 328–337 (1998)
49. Stacy, E.W.: A generalization of the gamma distribution. Ann. Math. Stat. **33**(3), 1187–1192 (1962)
50. Shin, J.W., Chang, J.-H., Kim, N.S.: Statistical modeling of speech signal based on generalized gamma distribution. IEEE Signal Process. Lett. **12**(3), 258–261 (2005)
51. Saruwatari, H., Ishikawa, Y., Takahashi, Y., Inoue, T., Shikano, K., Kondo, K.: Musical noise controllable algorithm of channelwise spectral subtraction and adaptive beamforming based on higher-order statistics. IEEE Trans. Audio Speech Lang. Process. **19**(6), 1457–1466 (2011)
52. Inoue, T., Saruwatari, H., Takahashi, Y., Shikano, K., Kondo, K.: Theoretical analysis of musical noise in generalized spectral subtraction based on higher-order statistics. IEEE Trans. Audio Speech Lang. Process. **19**(6), 1770–1779 (2011)
53. Lee, A., Kawahara, T., Shikano, K.: Julius -An open source real-time large vocabulary recognition engine. In: Proceedings of Eurospeech, pp. 1691–1694 (2001)
54. Takahashi, Y., Osako, K., Saruwatari, H., Shikano, K.: Blind source extraction for hands-free speech recognition based on Wiener filtering and ICA-based noise estimation. In: Proceedings of 2008 Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA2008), pp. 164–167 (2008)
55. Even, J., Saruwatari, H., Shikano, K.: Enhanced Wiener post-processing based on partial projection back of the blind signal separation noise estimate. In: Proceedings of 17th European Signal Processing Conference (EUSIPCO2009), pp. 1442–1446 (2009)
56. Okamoto, R., Takahashi, Y., Saruwatari, H., Shikano, K.: MMSE STSA estimator with nonstationary noise estimation based on ICA for high-quality speech enhancement. In: Proceedings of 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2010), pp. 4778–4781 (2010)
57. Saruwatari, H., Go, M., Okamoto, R., Shikano, K.: Binaural hearing aid using sound-localization-preserved MMSE STSA estimator with ICA-based noise estimation. In: Proceedings of 2010 International Workshop on Acoustic Echo and Noise, Control (IWAENC2010) (2010)
58. Jan, T., Wang, W., Wang, D.L.: A multistage approach to blind separation of convolutive speech mixtures. Speech Commun. **53**, 524–539 (2011)
59. Inoue, T., Saruwatari, H., Shikano, K., Kondo, K.: Theoretical analysis of musical noise in Wiener filtering family via higher-order statistics. In: Proceedings of 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2011), pp. 5076–5079 (2011)
60. Yu, H., Fingscheidt, T.: A figure of merit for instrumental optimization of noise reduction algorithms. In: Proceedings of DSP in Vehicles (2011)
61. Kanehara, S., Saruwatari, H., Miyazaki, R., Shikano, K., Kondo, K.: Comparative study on various noise reduction methods with decision-directed a priori SNR estimator via higher-order statistics. In: Proceedings of 2012 APSIPA Annual Summit and Conference (APSIPA2012) (2012)
62. Yu, H., Fingscheidt, T.: Black box measurement of musical tones produced by noise reduction systems. In: Proceedings of ICASSP2012, pp. 4573–4576 (2012)
63. Saruwatari, H., Kanehara, S., Miyazaki, R., Shikano, K., Kondo, K.: Musical noise analysis for Bayesian minimum mean-square error speech amplitude estimators based on higher-order statistics. In: Proceedings of Interspeech 2013 (2013)
64. Miyazaki, R., Saruwatari, H., Inoue, T., Takahashi, Y., Shikano, K., Kondo, K.: Musical-noise-free speech enhancement based on optimized iterative spectral subtraction. IEEE Trans. Audio Speech Lang. Process. **20**(7), 2080–2094 (2012)
65. Miyazaki, R., Saruwatari, H., Shikano, K., Kondo, K.: Musical-noise-free blind speech extraction using ICA-based noise estimation and iterative spectral subtraction. In: Proceedings of 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA2012), pp. 322–327 (2012)

66. Miyazaki, R., Saruwatari, H., Shikano, K., Kondo, K.: Musical-noise-free blind speech extraction using ICA-based noise estimation with channel selection. In: Proceedings of 2012 International Workshop on Acoustic Signal Enhancement (IWAENC2012) (2012)
67. Buchner, H., Aichner, R., Kellermann, W.: A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. IEEE Trans. Speech Audio Process. **13**(1), 120–134 (2005)
68. Hiekata, T., Ikeda, Y., Yamashita, T., Morita, T., Zhang, R., Mori, Y., Saruwatari, H., Shikano, K.: Development and evaluation of pocket-size real-time blind source separation microphone. Acoust. Sci. Technol. **30**(4), 297–304 (2009)