

# Chapter 13

## Musical-Noise-Free Blind Speech Extraction Based on Higher-Order Statistics Analysis

Hiroshi Saruwatari and Ryoichi Miyazaki

**Abstract** In this chapter, we introduce a musical-noise-free blind speech extraction method using a microphone array for application to nonstationary noise. In the recent noise reduction study, it was found that optimized iterative spectral subtraction (SS) results in speech enhancement with almost no musical noise generation, but this method is valid only for stationary noise. The method presented in this chapter consists of iterative blind dynamic noise estimation by, e.g., independent component analysis (ICA) or multichannel Wiener filtering, and musical-noise-free speech extraction by modified iterative SS, where multiple iterative SS is applied to each channel while maintaining the multichannel property reused for the dynamic noise estimators. Also, in relation to the method, we discuss the justification of applying ICA to signals nonlinearly distorted by SS. From objective and subjective evaluations simulating a real-world hands-free speech communication system, we reveal that the method outperforms the conventional speech enhancement methods.

### 13.1 Introduction

In the past few decades, many applications of speech communication systems have been investigated, but it is well known that these systems always suffer from the deterioration of speech quality under adverse noise conditions. In a study of speech enhancement, many types of statistical signal estimation methods have been proposed, e.g., the maximum likelihood estimator of short-time spectral amplitude (spectral subtraction (SS) [1–4]), the minimum mean-square error estimator of the complex-valued spectrum (Wiener filtering (WF) [5]), the Bayesian estimator of short-time spectral amplitude (the minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator [6] and the minimum mean-square error

---

H. Saruwatari (✉)  
The University of Tokyo, Tokyo 113-8656, Japan  
e-mail: [hiroshi\\_saruwatari@ipc.i.u-tokyo.ac.jp](mailto:hiroshi_saruwatari@ipc.i.u-tokyo.ac.jp)

R. Miyazaki  
Department of Computer Science and Electronic Engineering,  
National Institute of Technology, Tokuyama College, Gakuendai,  
Shunan, Yamaguchi 745-8585, Japan

log-spectral amplitude estimator (MMSE-LSA) [7]), and the MAP estimator [8]. SS is the commonly used noise reduction method that has high noise reduction performance with low computational complexity. However, in this method, artificial distortion, referred to as *musical noise*, arises owing to nonlinear signal processing, leading to a serious deterioration of sound quality [9, 10]. Therefore, to assess and control the generation of musical noise, several studies were conducted using higher-order statistics [11–13].

To achieve high-quality noise reduction with low musical noise, an iterative SS method has been proposed [14–16]. This method is performed through signal processing, in which weak SS processes are iteratively applied to the input signal. Also, Inoue et al. have reported the very interesting phenomenon that this method with appropriate parameters gives equilibrium behavior in the growth of higher-order statistics with increasing number of iterations [17]. This means that almost no musical noise is generated even with high noise reduction, which is one of the most desirable properties of single-channel nonlinear noise reduction methods. Following this finding, Miyazaki et al. have derived the optimal parameters satisfying the no-musical-noise-generation condition by analysis based on higher-order statistics [18]. We have defined this method as *musical-noise-free* speech enhancement, where no musical noise is generated even for a high signal-to-noise ratio (SNR) in iterative SS. In this chapter, firstly, we explain the overview of musical-noise-free iterative SS.

In conventional iterative SS, however, it is assumed that the input noise signal is stationary, meaning that we can estimate the expectation of noise power spectral density from a time-frequency period of a signal that contains only noise. In contrast, under real-world acoustical environments, such as a nonstationary noise field, although it is necessary to dynamically estimate noise, this is very difficult. Therefore, in this chapter, secondly, we describe an advanced iterative signal extraction method using a microphone array that can be applied to nonstationary noise [19]. This method consists of iterative blind dynamic noise estimation by independent component analysis (ICA) [20–23] and musical-noise-free speech extraction by modified iterative SS, where multiple iterative SS is applied to each channel while maintaining the multichannel property reused for ICA.

Thirdly, in relation to the above-mentioned method, we discuss the justification of applying ICA to signals nonlinearly distorted by SS. We theoretically clarify that the degradation in ICA-based noise estimation obeys an amplitude variation in room transfer functions between the target user and microphones. Next, to reduce speech distortion, we introduce a channel selection strategy into ICA, where we automatically choose less varied inputs to maintain the high accuracy of noise estimation. Furthermore, we introduce a time-variant noise power spectral density (PSD) estimator [24] instead of ICA to improve the noise estimation accuracy. From objective and subjective evaluations, it is revealed that the presented method outperforms various types of the conventional methods.

Note that there exist many investigations for musical noise assessment using higher-order statistics [25–29] and the study on musical-noise-free speech enhancement was carried out for several methods except for iterative SS, namely, iterative WF [30], the iterative MMSE-STSA estimator [31] and the iterative generalized

MMSE-STSA estimator [32]. In this chapter, however, only SS-based method is dealt with because of ease in the mathematical derivations and readers' understanding. Also, the theoretical analysis and results in ICA-based noise estimation are valid for other independent linear factor analysis algorithms, e.g., independent vector analysis [33–35] and independent low-rank matrix analysis [36–39]. However, we focus our attention on only ICA in this chapter owing to its simpleness.

## 13.2 Single-Channel Speech Enhancement with Musical-Noise-Free Properties

### 13.2.1 Conventional Non-iterative Spectral Subtraction

We apply a short-time discrete Fourier transform (DFT) to the observed signal, which is a mixture of target speech and noise, to obtain the time-frequency signal. We formulate conventional *non-iterative SS* [1] in the time-frequency domain as follows:

$$Y(f, \tau) = \begin{cases} \sqrt{|X(f, \tau)|^2 - \beta E[|N|^2]} \exp(j \arg(X(f, \tau))) \\ \text{(if } |X(f, \tau)|^2 > \beta E[|N|^2]), \\ \eta X(f, \tau) \quad \text{(otherwise),} \end{cases} \quad (13.1)$$

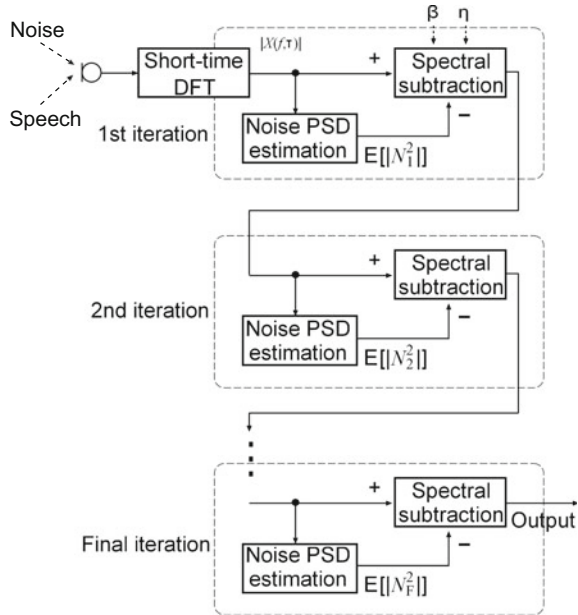
where  $Y(f, \tau)$  is the enhanced target speech signal,  $X(f, \tau)$  is the observed signal,  $f$  denotes the frequency subband,  $\tau$  is the frame index,  $\beta$  is the oversubtraction parameter, and  $\eta$  is the flooring parameter. Here,  $E[|N|^2]$  is the expectation of the random variable  $|N|^2$  corresponding to the noise power spectra. In practice, we can approximate  $E[|N|^2]$  by averaging the observed noise power spectra  $|N(f, \tau)|^2$  in the first  $K$ -sample frames, where we assume the absence of speech in this period and noise stationarity. However, this often requires high-accuracy voice activity detection. In addition, many methods for dynamic estimation of the expectation of the noise PSD have been proposed [4], but always suffered from difficulty in rapidly changing nonstationary noise.

Generally speaking, conventional spectral subtraction suffers from the inherent problem of musical noise generation. For example, a large oversubtraction parameter affords a large noise reduction but considerable musical noise is also generated. To reduce the amount of musical noise generated, we often increase the flooring parameter, but this decreases noise reduction; thus, there exists a trade-off between noise reduction and musical noise generation.

### 13.2.2 Iterative Spectral Subtraction

In an attempt to achieve high-quality noise reduction with low musical noise, an improved method based on iterative SS was proposed in previous studies [14–16].

**Fig. 13.1** Block diagram of iterative SS



This method is performed through signal processing, in which the following *weak SS* processes are recursively applied to the noise signal (see Fig. 13.1). (I) The average power spectrum of the input noise is estimated. (II) The estimated noise prototype is then subtracted from the input with the parameters specifically set for weak subtraction, e.g., a large flooring parameter  $\eta$  and a small subtraction parameter  $\beta$ . (III) We then return to step (I) and substitute the resultant output (partially noise reduced signal) for the input signal.

### 13.2.3 Modeling of Input Signal

In this chapter, we assume that the input signal  $X$  in the power spectral domain is modeled using the gamma distribution as

$$P(x) = \frac{x^{\alpha-1}}{\Gamma(\alpha)\theta^\alpha} \exp(-x/\theta), \tag{13.2}$$

where  $x \geq 0$ ,  $\alpha > 0$ , and  $\theta > 0$ . Here,  $\alpha$  is the shape parameter,  $\theta$  is the scale parameter, and  $\Gamma(\alpha)$  is the gamma function, defined as

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \exp(-t) dt. \tag{13.3}$$

If the input signal is Gaussian noise, 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 noise, the p.d.f. of its power spectra obeys the gamma distribution with  $\alpha < 1$ . We make assumption here that  $\theta$  is assumed to be the deterministically known noise PSD and estimation artifacts of the noise PSD are not taken into account in this chapter. Also, the estimation of  $\alpha$  for real-world (short-term) data is explained in, e.g., Ref. [40].

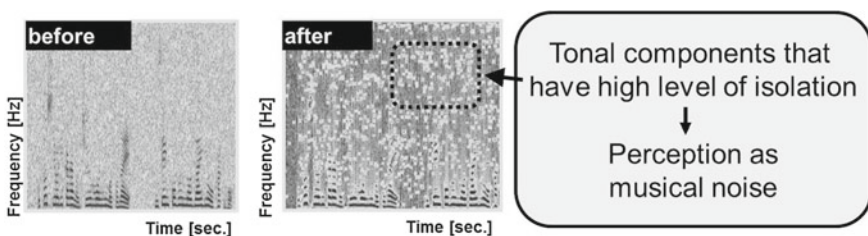
### 13.2.4 Metric of Musical Noise Generation: Kurtosis Ratio

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. 13.2). 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

$$\text{kurt} = \frac{\mu_4}{\mu_2^2}, \tag{13.4}$$

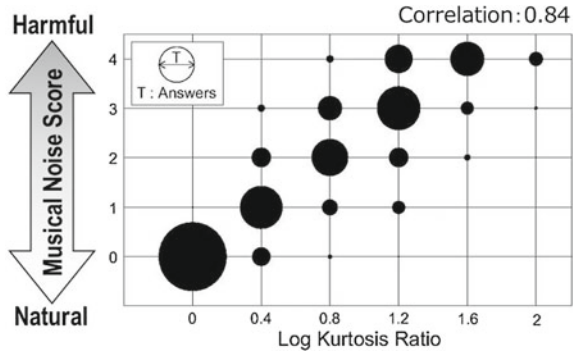
where “kurt” is the kurtosis and  $\mu_m$  is the  $m$ th-order moment, given by

$$\mu_m = \int_0^\infty x^m P(x) dx, \tag{13.5}$$



**Fig. 13.2** Example of generation of tonal component after signal processing, where input signal is speech with white Gaussian noise and output is processed signal by SS

**Fig. 13.3** Relation between kurtosis ratio (in log scale) and human-perceptual score of degree of musical noise generation [11]



where  $P(x)$  is the p.d.f. of the random variable  $X$ . Note that  $\mu_m$  is not a central moment but a raw moment. Thus, (13.4) is not kurtosis in the mathematically strict definition but a modified version; however, we still refer to (13.4) 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 [11–13]. This measure is defined as

$$\text{kurtosis ratio} = \frac{\text{kurt}_{\text{proc}}}{\text{kurt}_{\text{org}}}, \tag{13.6}$$

where  $\text{kurt}_{\text{proc}}$  is the kurtosis of the processed signal and  $\text{kurt}_{\text{org}}$  is the kurtosis of the original (unprocessed) signal. This measure increases as the amount of generated musical noise increases. In Ref. [11], it was reported that the kurtosis ratio is strongly correlated with the human perception of musical noise. Figure 13.3 shows an example of the relation between the kurtosis ratio (in log scale) and a human-perceptual score of degree of musical noise generation, where we can confirm the strong correlation.

### 13.2.5 Musical Noise Generation in Non-iterative Spectral Subtraction

In conventional non-iterative spectral subtraction, the long-term-averaged power spectrum of a noise signal is utilized as the estimated noise power spectrum. Then, the estimated noise power spectrum multiplied by the oversubtraction parameter  $\beta$  is subtracted from the observed power spectrum. When a gamma distribution is used to model the noise signal, its mean is  $\alpha_n \theta_n$ , where  $\alpha_n$  and  $\theta_n$  are the shape and scale parameters of noise, respectively (the subscript “n” indicates that the parameters belong to noise). Thus, the amount of subtraction is  $\beta \alpha_n \theta_n$ . The subtraction of the estimated noise power spectrum in each frequency band can be considered as a shift of the p.d.f. in the zero-power direction, given by

$$\frac{1}{\theta_n^{\alpha_n} \Gamma(\alpha_n)} (z + \beta \alpha_n \theta_n)^{\alpha_n - 1} \exp \left\{ -\frac{z + \beta \alpha_n \theta_n}{\theta_n} \right\}, \quad (13.7)$$

where  $z$  is the random variable of the p.d.f. after spectral subtraction.

As a result, negative-power components with nonzero probability arise. To avoid this, such negative components are replaced by observations that are multiplied by a positive value  $\eta$  (flooring parameter). This means that the region corresponding to the probability of the negative components, which forms a section cut from the original gamma distribution, is compressed by the effect of the flooring, resulting in

$$\frac{1}{(\eta^2 \theta_n)^{\alpha_n} \Gamma(\alpha_n)} z^{\alpha_n - 1} \exp \left\{ -\frac{z}{\eta^2 \theta_n} \right\}. \quad (13.8)$$

Note that the flooring parameter  $\eta$  is squared in the p.d.f. because the multiplication of  $\eta$  is conducted in the amplitude spectrum domain (see the second branch in (13.1)) but we now consider its effect in the power spectrum domain.

Finally, the floored components are superimposed on the laterally shifted p.d.f. Thus, the resultant p.d.f. after spectral subtraction,  $P_{SS}(z)$ , can be written as

$$P_{SS}(z) = \begin{cases} \frac{1}{\theta_n^{\alpha_n} \Gamma(\alpha_n)} (z + \beta \alpha_n \theta_n)^{\alpha_n - 1} \exp \left\{ -\frac{z + \beta \alpha_n \theta_n}{\theta_n} \right\} \\ + \frac{1}{(\eta^2 \theta_n)^{\alpha_n} \Gamma(\alpha_n)} z^{\alpha_n - 1} \exp \left\{ -\frac{z}{\eta^2 \theta_n} \right\} & (0 \leq z < \beta \alpha_n \eta^2 \theta_n), \\ \frac{1}{\theta_n^{\alpha_n} \Gamma(\alpha_n)} (z + \beta \alpha_n \theta_n)^{\alpha_n - 1} \exp \left\{ -\frac{z + \beta \alpha_n \theta_n}{\theta_n} \right\} & (\beta \alpha_n \eta^2 \theta_n \leq z), \end{cases} \quad (13.9)$$

To characterize non-iterative spectral subtraction, the  $m$ th-order moment of  $z$  is required. For  $P_{SS}(z)$ , the  $m$ th-order moment is given by

$$\begin{aligned} \mu_m^{SS} &= \int_0^\infty z^m \cdot P_{SS}(z) dz \\ &= \int_0^\infty z^m \frac{1}{\theta_n^{\alpha_n} \Gamma(\alpha_n)} (z + \beta\alpha_n\theta_n)^{\alpha_n-1} \exp\left\{-\frac{z + \beta\alpha_n\theta_n}{\theta_n}\right\} dz \\ &\quad + \int_0^{\beta\alpha_n\eta^2\theta_n} z^m \frac{1}{(\eta^2\theta_n)^{\alpha_n} \Gamma(\alpha_n)} z^{\alpha_n-1} \exp\left\{-\frac{z}{\eta^2\theta_n}\right\} dz, \end{aligned} \tag{13.10}$$

where  $z$  is the random variable of the p.d.f. after spectral subtraction. We now expand the first term of the right-hand side of (13.10). Here, let  $t = (z + \beta\alpha_n\theta_n)/\theta_n$ , then  $\theta_n dt = dz$  and  $z = \theta_n(t - \beta\alpha_n)$ . Consequently,

$$\begin{aligned} &\int_0^\infty z^m \frac{1}{\theta_n^{\alpha_n} \Gamma(\alpha_n)} (z + \beta\alpha_n\theta_n)^{\alpha_n-1} \exp\left\{-\frac{z + \beta\alpha_n\theta_n}{\theta_n}\right\} dz \\ &= \int_{\beta\alpha_n}^\infty \theta_n^m (t - \beta\alpha_n)^m \frac{1}{\theta_n^{\alpha_n} \Gamma(\alpha_n)} (\theta_n t)^{\alpha_n-1} \exp\{-t\} \theta_n dt \\ &= \frac{\theta_n^m}{\Gamma(\alpha_n)} \int_{\beta\alpha_n}^\infty \sum_{l=0}^m (-\beta\alpha_n)^l \frac{\Gamma(m+1)}{\Gamma(l+1)\Gamma(m-l+1)} t^{m-l} t^{\alpha_n-1} \exp\{-t\} dt \\ &= \frac{\theta_n^m}{\Gamma(\alpha_n)} \sum_{l=0}^m (-\beta\alpha_n)^l \frac{\Gamma(m+1)}{\Gamma(l+1)\Gamma(m-l+1)} \Gamma(\alpha_n + m - l, \beta\alpha_n), \end{aligned} \tag{13.11}$$

where we use the binomial theorem given by

$$(t + a)^m = \sum_{l=0}^m a^l \frac{\Gamma(m+1)}{\Gamma(l+1)\Gamma(m-l+1)} t^{m-l}, \tag{13.12}$$

and  $\Gamma(a, b)$  is the upper incomplete gamma function defined as

$$\Gamma(a, b) = \int_b^\infty t^{a-1} \exp\{-t\} dt. \tag{13.13}$$

Next we consider the second term of the right-hand side of (13.10). Here, let  $t = z/(\eta^2\theta_n)$ , then  $\eta^2\theta_n dt = dz$ . Thus,

$$\begin{aligned} &\int_0^{\beta\alpha_n\eta^2\theta_n} z^m \frac{1}{(\eta^2\theta_n)^{\alpha_n} \Gamma(\alpha_n)} z^{\alpha_n-1} \exp\left\{-\frac{z}{\eta^2\theta_n}\right\} dz \\ &= \int_0^{\beta\alpha_n} (\eta^2\theta_n t)^m \frac{1}{(\eta^2\theta_n)^{\alpha_n} \Gamma(\alpha_n)} (\eta^2\theta_n t)^{\alpha_n-1} \exp\{-t\} \eta^2\theta_n dt \end{aligned}$$



$$\begin{aligned}
&= \frac{\eta^{2m} \theta_n^m}{\Gamma(\alpha_n)} \int_0^{\beta \alpha_n} t^{\alpha_n - 1 + m} \exp\{-t\} dt \\
&= \frac{\eta^{2m} \theta_n^m}{\Gamma(\alpha_n)} \gamma(\alpha_n + m, \beta \alpha_n),
\end{aligned} \tag{13.14}$$

where  $\gamma(a, b)$  is the lower incomplete gamma function defined as

$$\gamma(a, b) = \int_0^b t^{a-1} \exp\{-t\} dt. \tag{13.15}$$

As a result, the  $m$ th-order moment after spectral subtraction,  $\mu_m^{\text{SS}}$ , is a composite of (13.11) and (13.14), and is given by [17]

$$\mu_m^{\text{SS}} = \theta_n^m \mathcal{M}(\alpha_n, \beta, \eta, m), \tag{13.16}$$

where

$$\mathcal{M}(\alpha_n, \beta, \eta, m) = \mathcal{S}(\alpha_n, \beta, \eta) + \eta^{2m} \mathcal{F}(\alpha_n, \beta, \eta), \tag{13.17}$$

$$\mathcal{S}(\alpha_n, \beta, m) = \sum_{l=0}^m (-\beta \alpha_n)^l \frac{\Gamma(m+1) \Gamma(\alpha_n + m - l, \beta \alpha_n)}{\Gamma(\alpha_n) \Gamma(l+1) \Gamma(m-l+1)}, \tag{13.18}$$

$$\mathcal{F}(\alpha_n, \beta, m) = \frac{\gamma(\alpha_n + m, \beta \alpha_n)}{\Gamma(\alpha_n)}. \tag{13.19}$$

From (13.4), (13.16), and (13.17), the kurtosis after SS can be expressed as

$$\text{kurt} = \frac{\mathcal{M}(\alpha_n, \beta, \eta, 4)}{\mathcal{M}^2(\alpha_n, \beta, \eta, 2)}. \tag{13.20}$$

Using (13.6) and (13.20), we also express the kurtosis ratio as

$$\text{kurtosis ratio} = \frac{\mathcal{M}(\alpha_n, \beta, \eta, 4) / \mathcal{M}^2(\alpha_n, \beta, \eta, 2)}{\mathcal{M}(\alpha_n, 0, 0, 4) / \mathcal{M}^2(\alpha_n, 0, 0, 2)}. \tag{13.21}$$

Also, as a measure of the noise reduction performance, the noise reduction rate (NRR) [41], the output SNR minus the input SNR in dB, can be given in terms of a 1st-order moment as [17]

$$\text{NRR} = 10 \log_{10} \frac{\alpha_n}{\mathcal{M}(\alpha_n, \beta, \eta, 1)}. \tag{13.22}$$

### 13.2.6 Musical-Noise-Free Speech Enhancement

In [18], Miyazaki et al. proposed musical-noise-free noise reduction, where no musical noise is generated even for a high SNR in iterative SS. In the study, some of the authors discovered an interesting phenomenon that the kurtosis ratio sometimes does not change even after SS via mathematical analysis based on (13.21) [17]. This indicates that the kurtosis ratio can be maintained at unity even after iteratively applying SS to improve the NRR, and thus no musical noise is generated owing to the *domino-toppling* phenomenon. Following this finding, the authors derived the optimal parameters satisfying the musical-noise-free condition [18] by finding a fixed-point status in the kurtosis ratio, i.e., by solving

$$\frac{\mathcal{M}(\alpha_n, 0, 0, 4)}{\mathcal{M}^2(\alpha_n, 0, 0, 2)} = \frac{\mathcal{M}(\alpha_n, \beta, \eta, 4)}{\mathcal{M}^2(\alpha_n, \beta, \eta, 2)}. \quad (13.23)$$

The inductive result is that the kurtosis ratio never changes even at a large number of (ideally “infinite”) iterations. In this situation, sufficient noise reduction can be gained if the NRR improvement in each iteration is even small but positive. This corresponds to musical-noise-free noise reduction. In summary, we can formulate a new theorem on musical-noise-free conditions as follows.

**(I) Fixed-point kurtosis condition:** The kurtosis should be equal before and after spectral subtraction in each iteration. This corresponds to a fixed point for the 2nd- and 4th-order moments.

**(II) NRR growth condition:** The amount of noise reduction should be larger than 0 dB in each iteration, relating to a change in the 1st-order moment.

Although the parameters to be optimized are  $\eta$  and  $\beta$ , we hereafter derive the optimal  $\eta$  given a fixed  $\beta$  for ease of closed-form analysis. First, we change (13.20) for

$$\text{kurt}(\alpha_n, \beta, \eta) = \frac{\mathcal{S}(\alpha_n, \beta, 4) + \eta^8 \mathcal{F}(\alpha_n, \beta, 4)}{(\mathcal{S}(\alpha_n, \beta, 2) + \eta^4 \mathcal{F}(\alpha_n, \beta, 2))^2}. \quad (13.24)$$

Next, the fixed-point kurtosis condition corresponds to the kurtosis being equal before and after spectral subtraction, thus

$$\frac{\mathcal{S}(\alpha_n, \beta, 4) + \eta^8 \mathcal{F}(\alpha_n, \beta, 4)}{(\mathcal{S}(\alpha_n, \beta, 2) + \eta^4 \mathcal{F}(\alpha_n, \beta, 2))^2} = \frac{(\alpha_n + 3)(\alpha_n + 2)}{(\alpha_n + 1)\alpha_n}. \quad (13.25)$$

Let  $\mathcal{H} = \eta^4$ , and (13.25) yields the following quadratic equation in  $\mathcal{H}$ .

$$\begin{aligned} & (\mathcal{F}(\alpha_n, \beta, 4)(\alpha_n + 1)\alpha_n - \mathcal{F}^2(\alpha_n, \beta, 2)(\alpha_n + 3)(\alpha_n + 2)) \mathcal{H}^2 \\ & - 2\mathcal{S}(\alpha_n, \beta, 2)\mathcal{F}(\alpha_n, \beta, 2)(\alpha_n + 3)(\alpha_n + 2)\mathcal{H} \\ & + \mathcal{S}(\alpha_n, \beta, 4)(\alpha_n + 1)\alpha_n - \mathcal{S}^2(\alpha_n, \beta, 2)(\alpha_n + 3)(\alpha_n + 2) = 0. \end{aligned} \quad (13.26)$$

Thus, we can derive a closed-form estimate of  $\mathcal{H}$  from the given noise shape parameter  $\alpha_n$  and oversubtraction parameter  $\beta$  as

$$\begin{aligned} \mathcal{H} = & \{ \mathcal{F}(\alpha_n, \beta, 4)(\alpha_n+1)\alpha_n - \mathcal{F}^2(\alpha_n, \beta, 2)(\alpha_n+3)(\alpha_n+2) \}^{-1} \\ & \left[ \mathcal{S}(\alpha_n, \beta, 2)\mathcal{F}(\alpha_n, \beta, 2)(\alpha_n+3)(\alpha_n+2) \right. \\ & \pm \left[ \{ \mathcal{S}(\alpha_n, \beta, 2)\mathcal{F}(\alpha_n, \beta, 2)(\alpha_n+3)(\alpha_n+2) \}^2 \right. \\ & \left. - \{ \mathcal{F}(\alpha_n, \beta, 4)(\alpha_n+1)\alpha_n - \mathcal{F}^2(\alpha_n, \beta, 2)(\alpha_n+3)(\alpha_n+2) \} \right. \\ & \left. \left. \{ \mathcal{S}(\alpha_n, \beta, 4)(\alpha_n+1)\alpha_n - \mathcal{S}^2(\alpha_n, \beta, 2)(\alpha_n+3)(\alpha_n+2) \} \right]^{\frac{1}{2}} \right]. \end{aligned} \quad (13.27)$$

Finally,  $\eta = \mathcal{H}^{1/4}$  is the resultant flooring parameter that satisfies the fixed-point kurtosis condition.

From (13.22), the NRR growth condition is expressed as

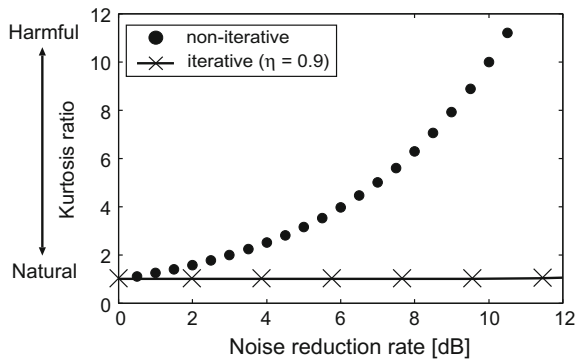
$$\text{NRR} = 10 \log_{10} \frac{\alpha_n}{\mathcal{S}(\alpha_n, \beta, 1) + \eta^2 \mathcal{F}(\alpha_n, \beta, 1)} > 0. \quad (13.28)$$

Here, since  $\eta > 0$ , we can solve the inequality as

$$0 < \eta < \sqrt{\frac{\alpha_n - \mathcal{S}(\alpha_n, \beta, 1)}{\mathcal{F}(\alpha_n, \beta, 1)}}. \quad (13.29)$$

In summary, we can choose the parameters simultaneously satisfying the fixed kurtosis point condition and NRR growth condition using (13.27) and (13.29). Figure 13.4 shows an example of the kurtosis ratio in optimized iterative SS, where Gaussian noise is assumed. We can confirm the flat trace of the kurtosis, indicating no musical noise generation.

**Fig. 13.4** Relation between NRR and kurtosis ratio obtained from theoretical analysis for case of Gaussian noise



### 13.3 Extension to Multichannel Blind Signal Processing

#### 13.3.1 Blind Spatial Subtraction Array

In the previous section, we assumed that the input noise signal is stationary, meaning that we can estimate the expectation of a noise signal from a time-frequency period of a signal that contains only noise, i.e., speech absence. However, in actual environments, such as a nonstationary noise field, it is necessary to dynamically estimate the noise PSD.

To solve this problem, Takahashi et al. previously proposed blind spatial subtraction array (BSSA) [42], which involves accurate noise estimation by ICA followed by a speech extraction procedure based on SS (see Fig. 13.5). BSSA improves the noise reduction performance, particularly in the presence of both diffuse and nonstationary noises; thus, almost all the environmental noise can be dealt with. However, BSSA always suffers from musical noise owing to SS. In addition, the output signal of BSSA degenerates to a *monaural* (not multichannel) signal, meaning that ICA cannot be reapplied; thus, we cannot iteratively estimate the noise power spectra. Therefore, it is impossible to directly apply iterative SS to the conventional BSSA.

#### 13.3.2 Iterative Blind Spatial Subtraction Array

In this section, we introduce a multi-iterative blind signal extraction method integrating iterative blind noise estimation by ICA and iterative noise reduction by SS. As mentioned previously, the conventional BSSA cannot iteratively and accurately estimate noise by ICA because the conventional BSSA performs a delay and sum (DS) operation before SS. To solve this problem, Takahashi et al. have proposed an improved BSSA structure that performs multiple independent SS in each channel before DS; we call this structure *channelwise SS* [43–45]. Using this structure, we can equalize the number of channels of the observed signal to that of the signals after

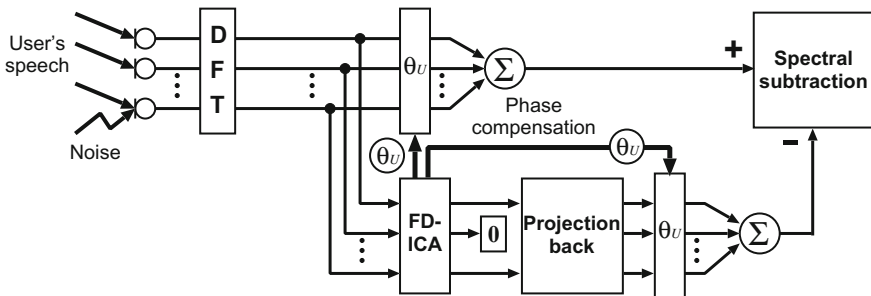


Fig. 13.5 Block diagram of BSSA [42]

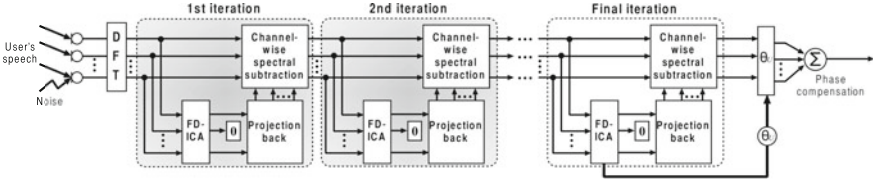


Fig. 13.6 Block diagram of iterative BSSA [19]

channelwise SS. Therefore, we can iteratively apply noise estimation by ICA and speech extraction by SS (see Fig. 13.6). Also, the advantage of the structure is that ICA has the possibility of adaptively estimating the *distorted wavefront* of a speech signal to some extent even after SS, because ICA is a blind signal identification method that does not require knowledge of the target signal direction. Details of this issue will be discussed in Sect. 13.3.3. Hereafter, we refer to this type of BSSA as *iterative BSSA*.

We conduct iterative BSSA in the following manner, where the superscript  $[i]$  represents the value in the  $i$ th iteration of SS (initially  $i = 0$ ).

- (I) The observed signal vector of the  $K$ -channel array in the time-frequency domain,  $\mathbf{X}^{[0]}(f, \tau)$ , is given by

$$\mathbf{X}^{[0]}(f, \tau) = \mathbf{H}(f)S(f, \tau) + \mathbf{N}(f, \tau), \quad (13.30)$$

where  $\mathbf{H}(f) = [H_1(f), H_2(f), \dots, H_K(f)]^T$  is a column vector of the transfer functions from the target signal position to each microphone,  $S(f, \tau)$  is the target speech signal, and  $\mathbf{N}(f, \tau)$  is a column vector of the additive noise.

- (II) Next, we perform signal separation using ICA as [20]

$$\mathbf{O}^{[i]}(f, \tau) = \mathbf{W}_{\text{ICA}}^{[i]}(f)\mathbf{X}^{[i]}(f, \tau), \quad (13.31)$$

$$\begin{aligned} \mathbf{W}_{\text{ICA}}^{[i][p+1]}(f) = & \mu[\mathbf{I} - \langle \varphi(\mathbf{O}^{[i]}(f, \tau))(\mathbf{O}^{[i]}(f, \tau))^H \rangle_{\tau}] \\ & \cdot \mathbf{W}_{\text{ICA}}^{[i][p]}(f) + \mathbf{W}_{\text{ICA}}^{[i][p]}(f), \end{aligned} \quad (13.32)$$

where  $\mathbf{W}_{\text{ICA}}^{[i][p]}(f)$  is a demixing matrix,  $\mu$  is the step-size parameter,  $[p]$  is used to express the value of the  $p$ th step in the ICA iterations,  $\mathbf{I}$  is the identity matrix,  $\langle \cdot \rangle_{\tau}$  denotes a time-averaging operator, and  $\varphi(\cdot)$  is an appropriate nonlinear vector function. Then, we construct a *noise-only vector*,

$$\begin{aligned} \mathbf{O}_{\text{noise}}^{[i]}(f, \tau) = & [O_1^{[i]}(f, \tau), \dots, O_{U-1}^{[i]}, 0, \\ & O_{U+1}^{[i]}(f, \tau), \dots, O_K^{[i]}(f, \tau)]^T, \end{aligned} \quad (13.33)$$

where  $U$  is the signal number for speech, and we apply the projection back operation to remove the ambiguity of the amplitude and construct the estimated

noise signal,  $\mathbf{Z}^{[i]}(f, \tau)$ , as

$$\mathbf{Z}^{[i]}(f, \tau) = \mathbf{W}_{\text{ICA}}^{[i]}(f)^{-1} \mathbf{O}_{\text{noise}}^{[i]}(f, \tau). \quad (13.34)$$

(III) Next, we perform SS independently in each input channel and derive the multiple target-speech-enhanced signals. This procedure can be given by

$$X_k^{[i+1]}(f, \tau) = \begin{cases} \sqrt{|X_k^{[i]}(f, \tau)|^2 - \beta |Z_k^{[i]}(f, \tau)|^2} \exp(j \arg(X_k^{[i]}(f, \tau))) \\ \text{(if } |X_k^{[i]}(f, \tau)|^2 > \beta |Z_k^{[i]}(f, \tau)|^2), \\ \eta X_k^{[i]}(f, \tau) \quad \text{(otherwise),} \end{cases} \quad (13.35)$$

where  $X_k^{[i+1]}(f, \tau)$  is the target-speech-enhanced signal obtained by SS at a specific channel  $k$ . Then we return to step (II) with  $\mathbf{X}^{[i+1]}(f, \tau)$ . When we obtain sufficient noise reduction performance, we proceed to step (IV).

(IV) Finally, we obtain the resultant target-speech-enhanced signal by applying DS to  $\mathbf{X}^{[*]}(f, \tau)$ , where  $*$  is the number of iterations after which sufficient noise reduction performance is obtained. This procedure can be expressed by

$$Y(f, \tau) = \mathbf{W}_{\text{DS}}^T(f) \mathbf{X}^{[*]}(f, \tau), \quad (13.36)$$

$$\mathbf{W}_{\text{DS}}(f) = [W_1^{(\text{DS})}(f), \dots, W_K^{(\text{DS})}(f)], \quad (13.37)$$

$$W_k^{(\text{DS})}(f) = \frac{1}{K} \exp(-2\pi j(f/M) f_s d_k \sin \theta_U / c), \quad (13.38)$$

$$\theta_U = \sin^{-1} \frac{\arg \left( \frac{[\mathbf{w}_{\text{ICA}}^{[*]}(f)^{-1}]_{kU}}{[\mathbf{w}_{\text{ICA}}^{[*]}(f)^{-1}]_{k'U}} \right)}{2\pi f_s c^{-1} (d_k - d_{k'})}, \quad (13.39)$$

where  $Y(f, \tau)$  is the final output signal of iterative BSSA,  $\mathbf{w}_{\text{DS}}$  is the filter coefficient vector of DS,  $M$  is the DFT size,  $f_s$  is the sampling frequency,  $d_k$  is the microphone position,  $c$  is the sound velocity, and  $\theta_U$  is the estimated direction of arrival of the target speech obtained by ICA's demixing matrix [46]. Moreover,  $[\mathbf{A}]_{lj}$  represents the entry in the  $l$ th row and  $j$ th column of  $\mathbf{A}$ .

### 13.3.3 Accuracy of Wavefront Estimated by Independent Component Analysis After Spectral Subtraction

In this subsection, we discuss the accuracy of the estimated noise signal in each iteration of iterative BSSA. In actual environments, not only point-source noise but also non-point-source (e.g., diffuse) noise often exists. It is known that ICA is proficient in noise estimation rather than speech estimation under such a noise

condition [42, 47, 48]. This is because the target speech can be regarded as a point-source signal (thus, the wavefront is static in each subband) and ICA acts as an effective blocking filter of the speech wavefront even in a time-invariant manner, resulting in good noise estimation. However, in iterative BSSA, we should address the inherent question of whether the distorted speech wavefront after nonlinear noise reduction such as SS can be blocked by ICA or not; thus, we determine whether the speech component after channelwise SS can become a point source again.

Hereafter, we quantify the degree of point-source-likeness for SS-applied speech signals. For convenience of discussion, a simple two-channel array model is assumed. First, we define the speech component in each channel after channelwise SS as

$$\hat{S}_1(f, \tau) = H_1(f)S(f, \tau) + \Delta S_1(f, \tau), \quad (13.40)$$

$$\hat{S}_2(f, \tau) = H_2(f)S(f, \tau) + \Delta S_2(f, \tau), \quad (13.41)$$

where  $S(f, \tau)$  is the original point-source speech signal,  $\hat{S}_k(f, \tau)$  is the speech component after channelwise SS at the  $k$ th channel, and  $\Delta S_k(f, \tau)$  is the speech component distorted by channelwise SS. Also, we assume that  $S(f, \tau)$ ,  $\Delta S_1(f, \tau)$ , and  $\Delta S_2(f, \tau)$  are uncorrelated with each other. Obviously,  $\hat{S}_1(f, \tau)$  and  $\hat{S}_2(f, \tau)$  can be regarded as being generated by a point source if  $\Delta S_1(f, \tau)$  and  $\Delta S_2(f, \tau)$  are zero, i.e., a valid static blocking filter can be obtained by ICA as

$$\begin{aligned} & [\mathbf{W}_{\text{ICA}}(f)]_{11} \hat{S}_1(f, \tau) + [\mathbf{W}_{\text{ICA}}(f)]_{12} \hat{S}_2(f, \tau) \\ &= ([\mathbf{W}_{\text{ICA}}(f)]_{11} H_1(f) + [\mathbf{W}_{\text{ICA}}(f)]_{12} H_2(f)) S(f, \tau) \\ &= 0, \end{aligned} \quad (13.42)$$

where we assume  $U = 1$  and, e.g.,  $[\mathbf{W}_{\text{ICA}}(f)]_{11} = H_2(f)$  and  $[\mathbf{W}_{\text{ICA}}(f)]_{12} = -H_1(f)$ . However, if  $\Delta S_1(f, \tau)$  and  $\Delta S_2(f, \tau)$  become nonzero as a result of SS, ICA does not have a valid speech-blocking filter with a static (time-invariant) form.

Second, the cosine distance between speech power spectra  $|\hat{S}_1(f, \tau)|^2$  and  $|\hat{S}_2(f, \tau)|^2$  is introduced in each frequency subband to indicate the degree of point-source-likeness as

$$\text{COS}(f) = \frac{\sum_{\tau} |\hat{S}_1(f, \tau)|^2 |\hat{S}_2(f, \tau)|^2}{\sqrt{\sum_{\tau} |\hat{S}_1(f, \tau)|^4} \sqrt{\sum_{\tau} |\hat{S}_2(f, \tau)|^4}}. \quad (13.43)$$

From (13.43), the cosine distance reaches its maximum value of unity if and only if  $\Delta S_1(f, \tau) = \Delta S_2(f, \tau) = 0$ , regardless of the values of  $H_1(f)$  and  $H_2(f)$ , meaning that the SS-applied speech signals  $\hat{S}_1(f, \tau)$  and  $\hat{S}_2(f, \tau)$  can be assumed to be produced by the point source. The value of  $\text{COS}(f)$  decreases with increasing magnitudes of  $\Delta S_1(f, \tau)$  and  $\Delta S_2(f, \tau)$  as well as with increasing difference between  $H_1(f)$  and  $H_2(f)$ ; this indicates the non-point-source state.

Third, we evaluate the degree of point-source-likeness in each iteration of iterative BSSA by using  $\text{COS}(f)$ . We statistically estimate the distorted speech component

of the enhanced signal in each iteration. Here, we assume that the original speech power spectrum  $|S(f, \tau)|^2$  obeys a gamma distribution with a shape parameter of 0.1 (this is a typical value for speech [49–54]) as

$$|S(f, \tau)|^2 \sim \frac{x^{-0.9}}{\Gamma(0.1)\theta_s^{0.1}} \exp(-x/\theta_s), \quad (13.44)$$

where  $\theta_s$  is the speech scale parameter. Regarding the amount of noise to be subtracted, the 1st-order moment of the noise power spectra is equal to  $\theta_n \alpha_n$  when the number of iterations,  $i$ , equals zero. Also, the value of  $\alpha_n$  does not change in each iteration when we use the specific parameters  $\beta$  and  $\eta$  that satisfy the musical-noise-free condition because the kurtosis ratio does not change in each iteration. If we perform SS only once, the rate of noise decrease is given by

$$\mathcal{M}(\alpha_n, \beta, \eta, 1)/\alpha_n, \quad (13.45)$$

and thus, the amount of residual noise after the  $i$ th iteration is given by

$$\begin{aligned} \mu_1^{[i]} &= \theta_n \alpha_n \{\mathcal{M}(\alpha_n, \beta, \eta, 1)/\alpha_n\}^i \\ &= \theta_n \mathcal{M}^i(\alpha_n, \beta, \eta, 1) \alpha_n^{1-i}. \end{aligned} \quad (13.46)$$

Next, we assume that the speech and noise are disjoint, i.e., there are no overlaps in the time-frequency domain, and that speech distortion is caused by subtracting the average noise from the pure speech component. Thus, the speech component  $|\hat{S}_k^{[i+1]}(f, \tau)|^2$  at the  $k$ th channel after the  $i$ th iteration is represented by subtracting the amount of residual noise (13.46) as

$$\begin{aligned} |\hat{S}_k^{[i+1]}(f, \tau)|^2 &= \\ &\begin{cases} |\hat{S}_k^{[i]}(f, \tau)|^2 - \beta \theta_n \mathcal{M}^i(\alpha_n, \beta, \eta, 1) \alpha_n^{1-i} \\ \text{(if } |\hat{S}_k^{[i]}(f, \tau)|^2 > \beta \theta_n \mathcal{M}^i(\alpha_n, \beta, \eta, 1) \alpha_n^{1-i}), \\ \eta^2 |\hat{S}_k^{[i]}(f, \tau)|^2 \quad \text{(otherwise)}. \end{cases} \end{aligned} \quad (13.47)$$

Here, we define the input SNR as the average of both channel SNRs,

$$\begin{aligned} \text{ISNR}(f) &= \frac{1}{2} \left( \frac{0.1 |H_1(f)|^2 \theta_s}{\alpha_n \theta_n} + \frac{0.1 |H_2(f)|^2 \theta_s}{\alpha_n \theta_n} \right) \\ &= \frac{0.1 \theta_s}{2 \alpha_n \theta_n} (|H_1(f)|^2 + |H_2(f)|^2). \end{aligned} \quad (13.48)$$

If we normalize the speech scale parameter  $\theta_s$  to unity, from (13.48), the noise scale parameter  $\theta_n$  is given by



$$\theta_n = \frac{0.1(|H_1(f)|^2 + |H_2(f)|^2)}{2\alpha_n \text{ISNR}(f)}, \quad (13.49)$$

and using (13.49), we can reformulate (13.47) as

$$|\hat{S}_k^{[i+1]}(f, \tau)|^2 = \begin{cases} |\hat{S}_k^{[i]}(f, \tau)|^2 - \beta \frac{0.1(|H_1(f)|^2 + |H_2(f)|^2)}{2\text{ISNR}(f)} \mathcal{M}^i(\alpha_n, \beta, \eta, 1) \alpha_n^{-i} \\ \text{(if } |\hat{S}_k^{[i]}(f, \tau)|^2 > \beta \frac{0.1(|H_1(f)|^2 + |H_2(f)|^2)}{2\text{ISNR}(f)} \mathcal{M}^i(\alpha_n, \beta, \eta, 1) \alpha_n^{-i}), \\ \eta^2 |\hat{S}_k^{[i]}(f, \tau)|^2 \quad \text{(otherwise)}. \end{cases} \quad (13.50)$$

Furthermore, we define the transfer function ratio (TFR) as

$$\text{TFR}(f) = |H_1(f)/H_2(f)|^2, \quad (13.51)$$

and if we normalize  $|H_1(f)|^2$  to unity in each frequency subband,  $|H_1(f)|^2 + |H_2(f)|^2$  becomes  $1 + 1/\text{TFR}(f)$ . Finally, we express (13.50) in terms of the input SNR  $\text{ISNR}(f)$  and the transfer function ratio  $\text{TFR}(f)$  as

$$|\hat{S}_k^{[i+1]}(f, \tau)|^2 = \begin{cases} |\hat{S}_k^{[i]}(f, \tau)|^2 - \beta \frac{0.1(1+1/\text{TFR}(f))}{2\text{ISNR}(f)} \mathcal{M}^i(\alpha_n, \beta, \eta, 1) \alpha_n^{-i} \\ \text{(if } |\hat{S}_k^{[i]}(f, \tau)|^2 > \beta \frac{0.1(1+1/\text{TFR}(f))}{2\text{ISNR}(f)} \mathcal{M}^i(\alpha_n, \beta, \eta, 1) \alpha_n^{-i}), \\ \eta^2 |\hat{S}_k^{[i]}(f, \tau)|^2 \quad \text{(otherwise)}. \end{cases} \quad (13.52)$$

As can be seen, the speech component is subjected to greater subtraction and distortion as  $\text{ISNR}(f)$  and/or  $\text{TFR}(f)$  decrease.

Figure 13.7 shows the relation between the TFR and the corresponding value of  $\text{COS}(f)$  calculated by (13.43) and (13.52). In Fig. 13.7, we plot the average of  $\text{COS}(f)$  over whole frequency subbands. The noise shape parameter  $\alpha_n$  is set to 0.2 with the assumption of super-Gaussian noise (this corresponds to the real noises used in Sect. 13.5), the input SNR is set to 10, 5, or 0 dB, and the noise scale parameter  $\theta_n$  is uniquely determined by (13.49) and the previous parameter settings. The TFR is set from 0.4 to 1.0 ( $|h_1(f)|$  is fixed to 1.0). Note that the TFR is highly correlated to the room reverberation and the interelement spacing of the microphone array; we determined the range of the TFR by simulating a typical moderately reverberant room and the array with 2.15 cm interelement spacing used in Sect. 13.5 (see the example of the TFR in Fig. 13.8). For the internal parameters used in iterative BSSA in this simulation,  $\beta$  and  $\eta$  are 8.5 and 0.9, respectively, which satisfy the musical-noise-free condition. In addition, the smallest value on the horizontal axis is 3 dB in Fig. 13.7 because DS is still performed even when  $i = 0$ .

From Figs. 13.7a and b, which correspond to relatively high input SNRs, we can confirm that the degree of point-source-likeness, i.e.,  $\text{COS}(f)$ , is almost maintained when the TFR is close to 1 even if the speech components are distorted by iterative BSSA. Also, it is worth mentioning that the degree of point-source-likeness is still

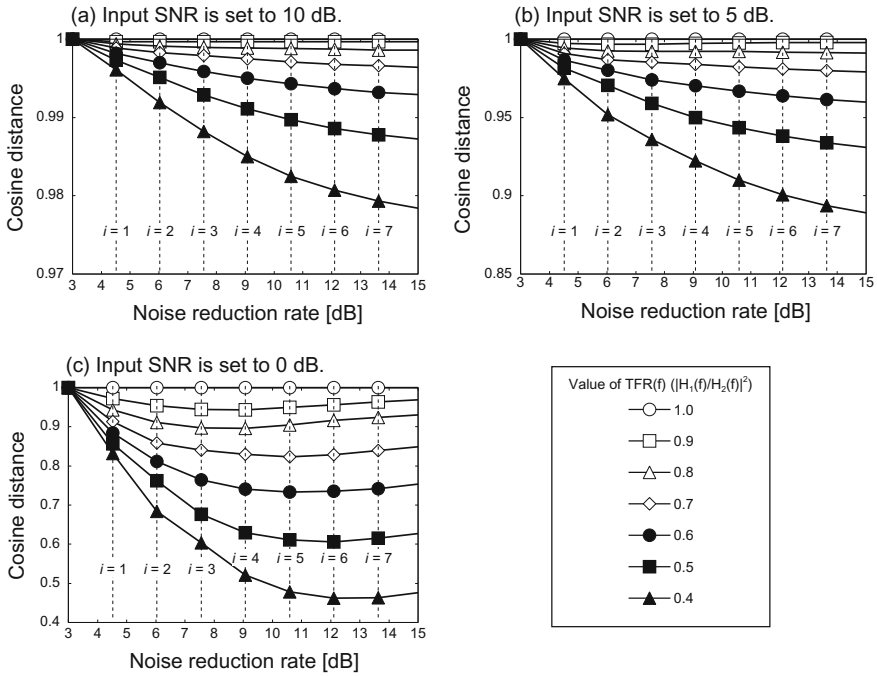


Fig. 13.7 Relation between number of iterations of iterative BSSA and cosine distance. Input SNR is a 10 dB, b 5 dB, and c 0 dB

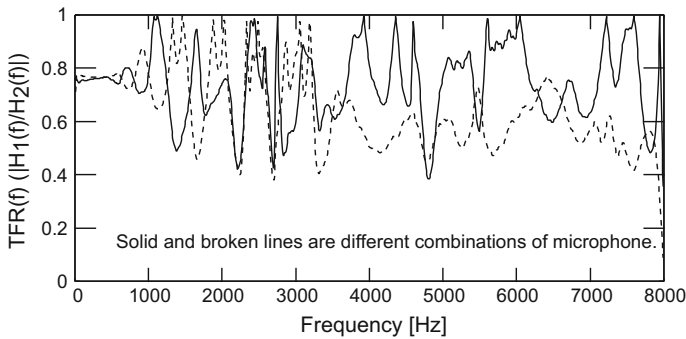


Fig. 13.8 Typical examples of  $TFR(f) (|H_1(f)/H_2(f)|^2)$  in each frequency subband

above 0.9 even when the TFR is decreased to 0.4 and  $i$  is increased to 6. This means that almost 90% of the speech components can be regarded as a point source and thus can be blocked by ICA. In contrast, from Fig. 13.7c, which shows the case of a low input SNR, when the TFR is dropped to 0.4 and  $i$  is more than 3, the degree of point-source-likeness is lower than 0.6. Thus, less than 60% of the speech components can be regarded as a point source, and this leads to poor noise estimation.

## 13.4 Improvement Scheme for Poor Noise Estimation

### 13.4.1 Channel Selection in Independent Component Analysis

In this subsection, we introduce a channel selection strategy in ICA for achieving high accuracy of noise estimation. As mentioned previously, speech distortion is subjected to  $\text{ISNR}(f)$  and  $\text{TFR}(f)$ , and the accuracy of noise estimation is degraded along with speech distortion. Figure 13.8 shows typical examples of the TFR. From Fig. 13.8, we can confirm that the TFRs in different combinations of microphones are not the same in each frequency subband; at a specific frequency, one microphone pair has higher  $\text{TFR}(f)$  than another pair, and vice versa at another frequency. Thus, we are able to select an appropriate combination of microphones to obtain a higher TFR.

Therefore, we introduce the channel selection method into ICA in each frequency subband, where we automatically choose less varied inputs to maintain high accuracy of noise estimation. Hereafter, we describe the detail of the channel selection method. First, we calculate the average power of the observed signal  $X_k(f, \tau)$  at the  $k$ th channel as

$$E_{\tau}[|X_k(f, \tau)|^2] = E_{\tau}[|S(f, \tau)|^2]|H_k(f)|^2 + E_{\tau}[|N_k(f, \tau)|^2]. \quad (13.53)$$

Here,  $E_{\tau}[|S(f, \tau)|^2]$  is a constant, and if we assume a diffuse noise field,  $E_{\tau}[|N_k(f, \tau)|^2]$  is also a constant. Thus, we can estimate the relative order of  $|H_k(f)|^2$  by comparing (13.53) for every  $k$ .

Next, we sort  $E_{\tau}[|X_k(f, \tau)|^2]$  in descending order and select the channels corresponding to a high amplitude of  $|H_k(f)|^2$  satisfying the following condition:

$$\max_k E_{\tau}[|X_k(f, \tau)|^2] \cdot \xi \leq E_{\tau}[|X_k(f, \tau)|^2], \quad (13.54)$$

where  $\xi (< 1)$  is the threshold for the selection.

Finally, we perform noise estimation based on ICA using the selected channels in each frequency subband, and we apply the projection back operation to remove the ambiguity of the amplitude and construct the estimated noise signal.

### 13.4.2 Time-Variant Noise Power Spectral Density Estimator

In the previous section, we revealed that the speech components cannot be regarded as a point source, and this leads to poor noise estimation in iterative BSSA. To solve this problem, we introduce a time-variant noise PSD estimator [24] instead of ICA to improve the noise estimation accuracy. This method has been developed for future

high-end binaural hearing aids and performs a prediction of the left noisy signal from the right noisy signal via the Wiener filter, followed by an auto-PSD of the difference between the left noisy signal and the prediction. By applying the noise PSD estimated from this estimator to (13.35), we can perform speech extraction. The procedure of this noise PSD estimator is described in Appendix.

## 13.5 Experiments in Real World

### 13.5.1 *Experimental Conditions*

We conducted objective and subjective evaluation experiments to confirm the validity of iterative BSSA under the diffuse and nonstationary noise condition. The size of the experimental room was  $4.2 \times 3.5 \times 3.0$  m<sup>3</sup> and the reverberation time was approximately 200 ms. We used a two-, three-, or four-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. All the signals used in this experiment were sampled at 16 kHz with 16-bit accuracy. The DFT size was 1024, and the frame shift length was 256. We used 5 male and 5 female speakers (one utterance per speaker) as sources of the original target speech signal. The input SNR was -5, 0, 5, and 10 dB.

### 13.5.2 *Objective Evaluation*

We conducted an objective experimental evaluation under the same NRR condition. First, Figs. 13.9, 13.10, 13.11, and 13.12 show the kurtosis ratio and cepstral distortion obtained from the experiments with real traffic noise and railway station noise, where we evaluate 10-dB NRR (i.e., output SNRs = 5, 10, 15, and 20 dB) signals processed by five conventional methods, namely, the MMSE-STSA estimator, the Log MMSE estimator incorporating speech-presence uncertainty [55], single-channel musical-noise-free iterative spectral subtraction, the multichannel speech enhancement method integrating the minimum variance beamformer and the Log MMSE estimator for postfiltering, and BSSA, in addition to several types of iterative BSSAs (using ICA or a time-variant noise estimator with/without channel selection). Here, we did not apply the channel selection method to the two-microphone case because ICA or time-variant noise estimation requires at least two-channel signals. Also, we applied a minimum statistics noise PSD estimator [4] to the MMSE STSA estimator and musical-noise-free iterative spectral subtraction, and we use the decision-directed approach for a priori SNR estimation in the MMSE STSA estimator and the log MMSE estimator. From Figs. 13.9 and 13.11, we can confirm that the iterative BSSA methods outperform the MMSE STSA estimator, the Log MMSE estimator, and the conventional BSSA in terms of kurtosis ratio. In particular, the

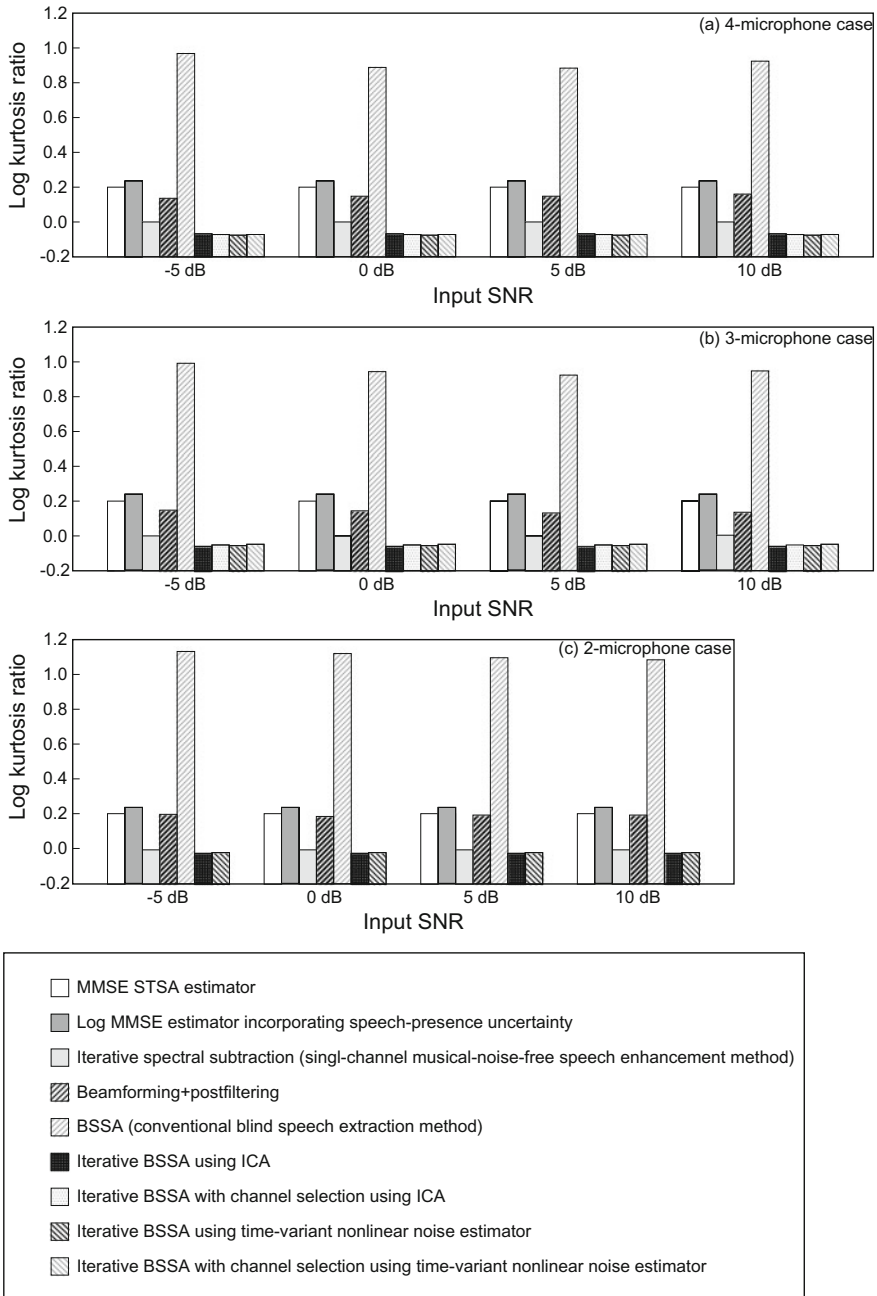
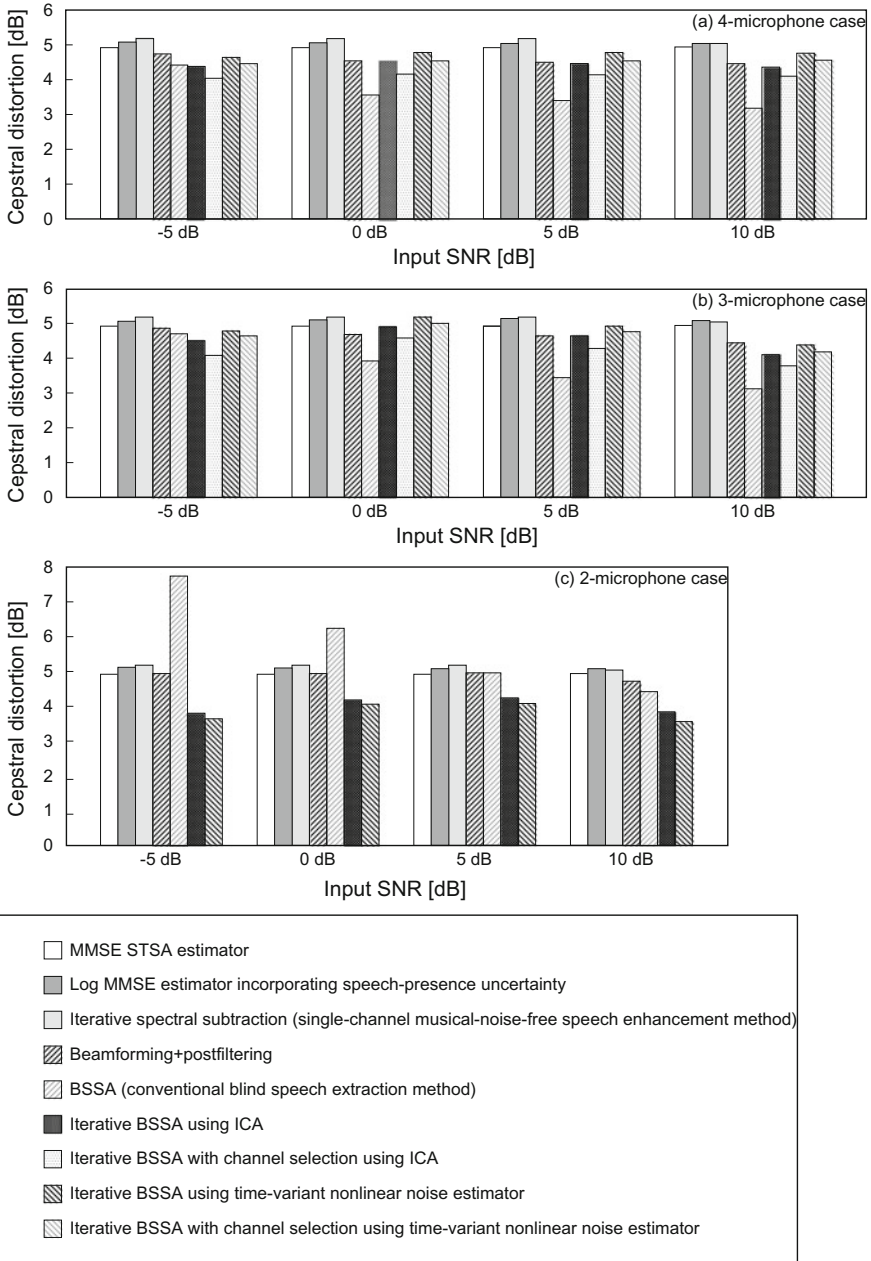
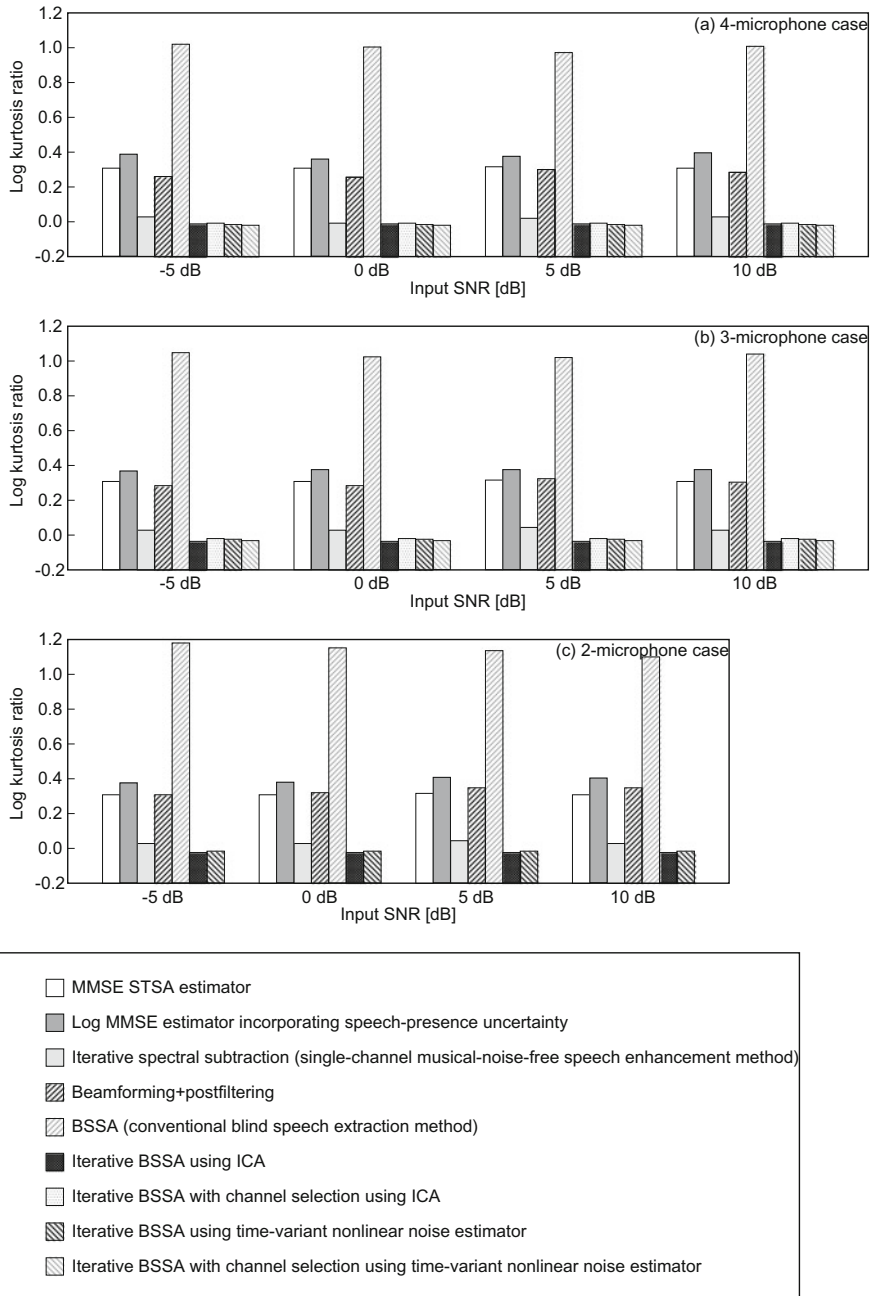


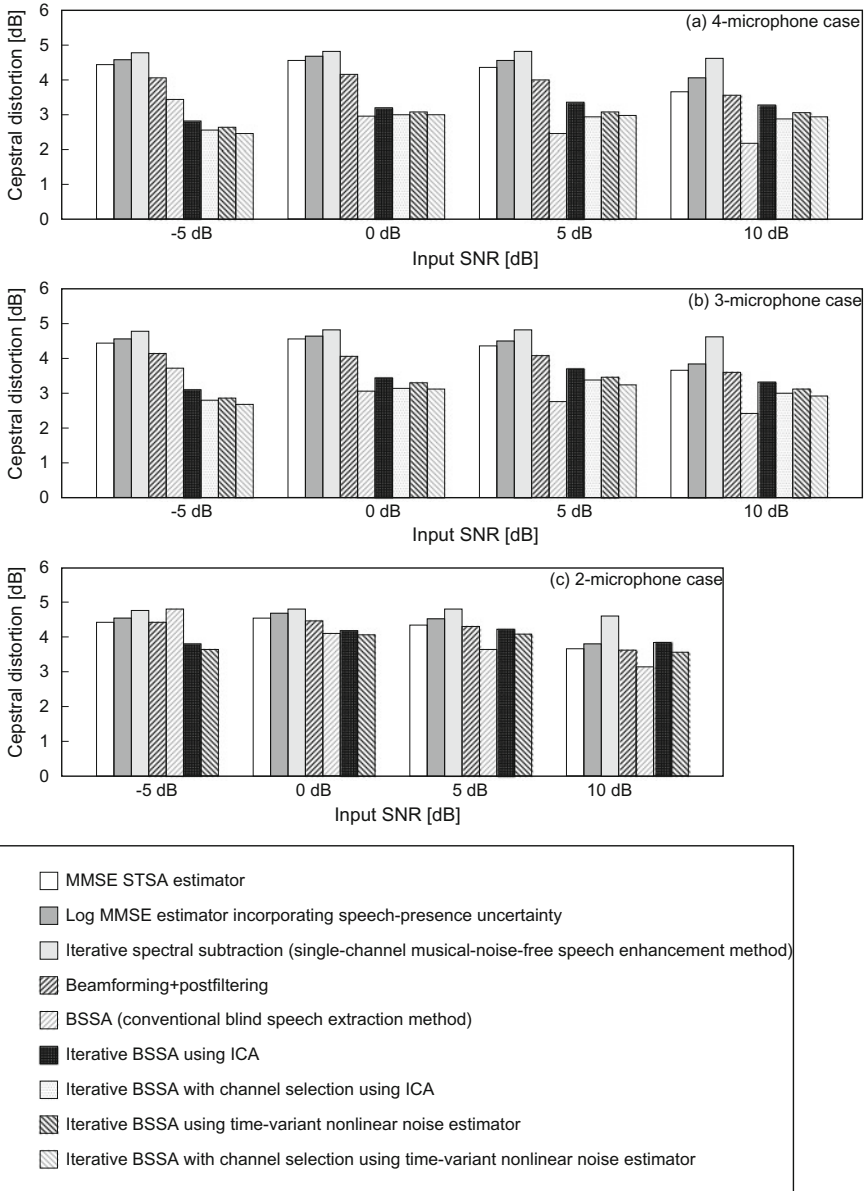
Fig. 13.9 Kurtosis ratio obtained from experiment for traffic noise under 10-dB NRR condition



**Fig. 13.10** Cepstral distortion obtained from experiment for traffic noise under 10-dB NRR condition

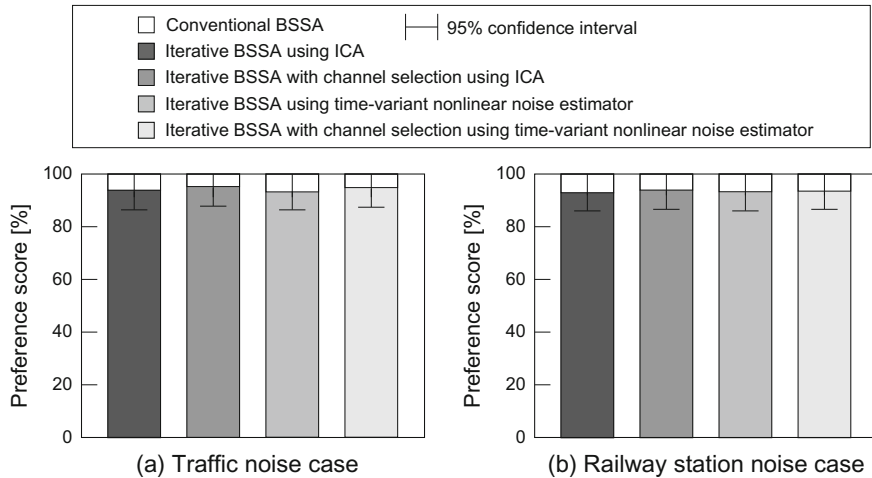


**Fig. 13.11** Kurtosis ratio obtained from experiment for railway station noise under 10-dB NRR condition



**Fig. 13.12** Cepstral distortion obtained from experiment for railway station noise under 10-dB NRR condition





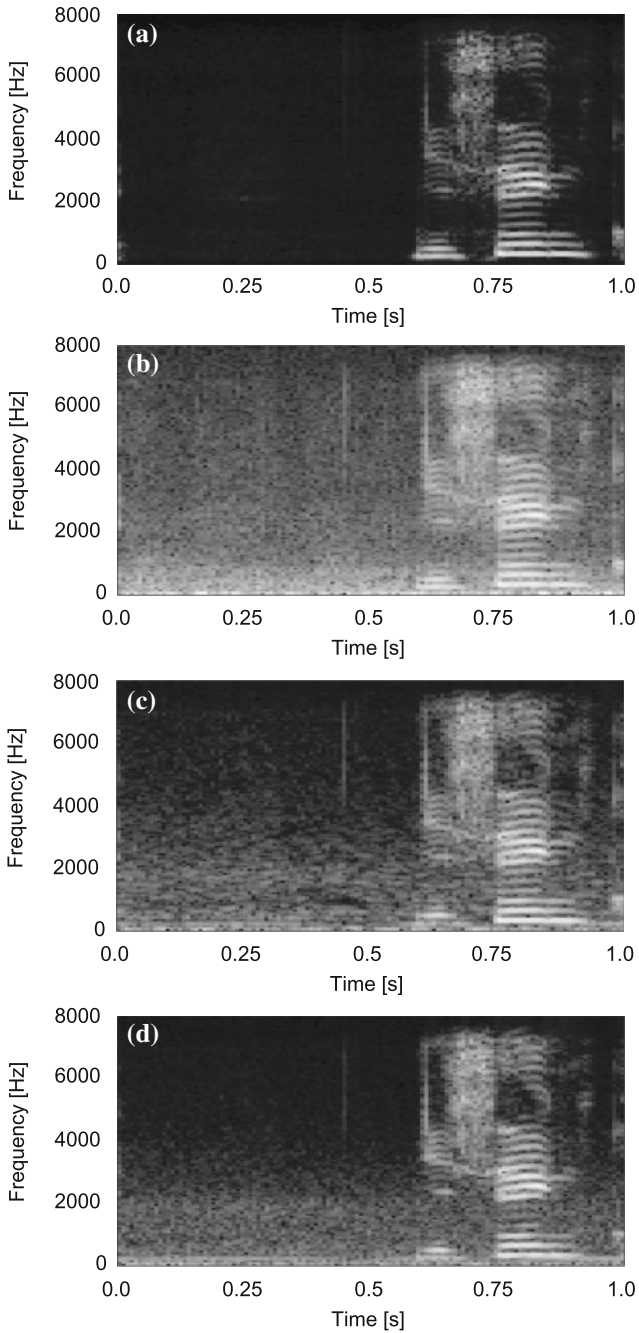
**Fig. 13.13** Subjective evaluation results for **a** traffic noise and **b** railway station noise

kurtosis ratios of the iterative BSSA methods are mostly close to 1.0. This means that the iterative methods did not generate any musical noise. However, the iterative BSSA methods lead to greater speech distortion compared with the conventional BSSA (see Figs. 13.10 and 13.12). Therefore, a trade-off exists between the amount of musical noise generation and speech distortion in the conventional BSSA and iterative BSSA methods. This result implies the disadvantage of iterative BSSA, i.e., large speech distortion, which has been theoretically predicted in Sect. 13.3.3. However, since the speech distortion of the proposed iterative BSSA with channel selection is lower than that of the original iterative BSSA, we can confirm the validity of the channel selection method.

### 13.5.3 Subjective Evaluation

Since we found the above-mentioned trade-off, we next conducted a subjective evaluation for setting the performance competition. In the evaluation, we presented a pair of 10-dB NRR signals processed by the conventional BSSA and four of iterative BSSAs (using ICA or a time-variant noise estimator with/without channel selection) in random order to 10 examinees, who selected which signal they preferred from the viewpoint of total sound quality, e.g., less musical noise, less speech distortion, and so forth.

The result of this experiment is shown in Fig. 13.13 for (a) traffic noise and (b) railway station noise. It is found that the output signals of some iterative BSSAs are preferred to that of the conventional BSSA, indicating the higher sound quality of the iterative methods in terms of human perception. This result is plausible because



**Fig. 13.14** Spectrogram for **a** clean signal, **b** observed signal, **c** signal extracted by BSSA, and **d** signal extracted by iterative BSSA

humans are often more sensitive to musical noise than to speech distortion as indicated in past studies, e.g., [12].

To visually understand the above-mentioned result, Fig. 13.14 shows part of the spectrograms of (a) a clean signal, (b) an observed signal, (c) a signal extracted by BSSA, and (d) a signal extracted by iterative BSSA, where the input SNR is set to 5 dB with real traffic noise and the NRR is 10 dB. From Fig. 13.14, it is confirmed that iterative BSSA reduces the number of isolated components in time-frequency domain sequences, which is a factor contributing to musical noise, compared with BSSA. Also, there are no major differences in the speech components of the clean signal, the signal processed by BSSA, and the signal processed by iterative BSSA; thus, we can conclude that the intelligibility of iterative BSSA is no less than that of BSSA.

## 13.6 Conclusions and Remarks

In this chapter, we addressed a musical-noise-free blind speech extraction method using a microphone array that can be applied to nonstationary noise. Firstly, we introduced iterative BSSA using a new BSSA structure, which generates almost no musical noise even with increasing noise reduction performance.

Secondly, in relation to the method, we discussed the justification of applying ICA to signals nonlinearly distorted by SS. We theoretically clarified that the degradation in ICA-based noise estimation obeys an amplitude variation in room transfer functions between the target user and microphones. Therefore, we gave the introduction of a channel selection strategy in ICA and a time-variant noise PSD estimator to improve the noise estimation accuracy.

Finally, from the objective evaluation experiments, we confirmed a trade-off between the amount of musical noise generation and speech distortion in the conventional and iterative BSSA. However, in a subjective preference test, iterative BSSA obtained a higher preference score than the conventional BSSA. Thus, iterative BSSA is advantageous to the conventional BSSA in terms of sound quality.

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 [56–60] have dealt with the issue of real-time implementation of ICA and BSSA, which would be helpful for the readers.

**Acknowledgements** This work was partially supported by SECOM Science and Technology Foundation.

## Appendix

This appendix provides a brief review of the time-variant nonlinear noise estimator. For more detailed information, Ref. [24] is available.

Let  $X_1(f, \tau)$  and  $X_2(f, \tau)$  be noisy signals received at the microphones in the time-frequency domain, defined as

$$X_1(f, \tau) = H_1(f)S(f, \tau) + N_1(f, \tau), \quad (13.55)$$

$$X_2(f, \tau) = H_2(f)S(f, \tau) + N_2(f, \tau), \quad (13.56)$$

where  $H_1(f)$  and  $H_2(f)$  are the transfer functions from the target signal position to each microphone. Next, the auto-power PSDs in each microphone,  $\Gamma_{11}(f)$  and  $\Gamma_{22}(f)$ , can be expressed as follows:

$$\Gamma_{11}(f, \tau) = |H_1(f)|^2 \Gamma_{SS}(f, \tau) + \Gamma_{NN}(f, \tau), \quad (13.57)$$

$$\Gamma_{22}(f, \tau) = |H_2(f)|^2 \Gamma_{SS}(f, \tau) + \Gamma_{NN}(f, \tau), \quad (13.58)$$

where  $\Gamma_{SS}(f, \tau)$  is the PSD of the target speech signal and  $\Gamma_{NN}(f, \tau)$  is the PSD of the noise signal. In this chapter, we assume that the left and right noise PSDs are approximately the same, i.e.,  $\Gamma_{N_1N_1}(f, \tau) \simeq \Gamma_{N_2N_2}(f, \tau) \simeq \Gamma_{NN}(f, \tau)$ .

Next, we consider the Wiener solution between the left and right transfer functions, which is defined as

$$H_W(f, \tau) = \frac{\Gamma_{12}(f, \tau)}{\Gamma_{22}(f, \tau)}, \quad (13.59)$$

where  $\Gamma_{12}(f, \tau)$  is the cross-PSD between the left and right noisy signals. The cross-PSD expression then becomes

$$\Gamma_{12}(f, \tau) = \Gamma_{SS}(f, \tau)H_1(f)H_2^*(f). \quad (13.60)$$

Therefore, substituting (13.60) into (13.59) yields

$$H_W(f, \tau) = \frac{\Gamma_{SS}(f, \tau)H_1(f)H_2^*(f)}{\Gamma_{22}(f, \tau)}. \quad (13.61)$$

Furthermore, using (13.57) and (13.58), the squared magnitude response of the Wiener solution in (13.61) can also be expressed as

$$|H_W(f, \tau)|^2 = \frac{(\Gamma_{11}(f, \tau) - \Gamma_{NN}(f, \tau))(\Gamma_{22}(f, \tau) - \Gamma_{NN}(f, \tau))}{\Gamma_{22}^2(f, \tau)}. \quad (13.62)$$

Equation (13.62) is rearranged into the following quadratic equation:

$$\begin{aligned} \Gamma_{\text{NN}}^2(f, \tau) - \Gamma_{\text{NN}}(f, \tau) (\Gamma_{11}(f, \tau) + \Gamma_{22}(f, \tau)) \\ + \Gamma_{\text{EE}}(f, \tau) \Gamma_{22}(f, \tau) = 0, \end{aligned} \quad (13.63)$$

where

$$\Gamma_{\text{EE}}(f, \tau) = \Gamma_{11}(f, \tau) - \Gamma_{22}(f, \tau) |H_{\text{W}}(f)|^2. \quad (13.64)$$

Consequently, the noise PSD  $\Gamma_{\text{NN}}(f)$  can be estimated by solving the quadratic equation in (13.63) as follows:

$$\Gamma_{\text{NN}}(f, \tau) = \frac{1}{2} (\Gamma_{11}(f, \tau) + \Gamma_{22}(f, \tau)) - \Gamma_{\text{avg}}(f, \tau), \quad (13.65)$$

$$\begin{aligned} \Gamma_{\text{avg}}(f, \tau) = \frac{1}{2} \{ (\Gamma_{11}(f, \tau) + \Gamma_{22}(f, \tau))^2 \\ - 4\Gamma_{\text{EE}}(f, \tau) \Gamma_{22}(f, \tau) \}^{0.5}. \end{aligned} \quad (13.66)$$

## References

1. S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech Signal Process.* **27**(2), 113–120 (1979)
2. M. Berouti, R. Schwartz, J. Makhoul, Enhancement of speech corrupted by acoustic noise, in *Proceeding of ICASSP* (1979), pp. 208–211
3. R. McAulay, M. Malpass, Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust. Speech Signal Process.* **28**(2), 137–145 (1980)
4. R. Martin, Spectral subtraction based on minimum statistics, in *Proceeding of EUSIPCO* (1994), pp. 1182–1185
5. P.C. Loizou, *Speech Enhancement Theory and Practice* (CRC Press, Taylor & Francis Group FL, 2007)
6. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **32**(6), 1109–1121 (1984)
7. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean square error log-spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **33**(2), 443–445 (1985)
8. T. Lotter, P. Vary, Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model. *EURASIP J. Appl. Signal Process.* **2005**, 1110–1126 (2005)
9. O. Cappe, Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Trans. Speech Audio Process.* **2**(2), 345–349 (1994)
10. Z. Goh, K.-C. Tan, B. Tan, Postprocessing method for suppressing musical noise generated by spectral subtraction. *IEEE Trans. Speech Audio Process.* **6**(3), 287–292 (1998)
11. Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, K. Kondo, Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics, in *Proceeding of IWAENC* (2008)
12. Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, K. Kondo, Musical noise generation analysis for noise reduction methods based on spectral subtraction and MMSE STSA estimation, in *Proceeding of ICASSP* (2009), pp. 4433–4436
13. Y. Takahashi, R. Miyazaki, H. Saruwatari, K. Kondo, Theoretical analysis of musical noise in nonlinear noise reduction based on higher-order statistics, in *Proceeding of APSIPA Annual Summit and Conference* (2012)

14. K. Yamashita, S. Ogata, T. Shimamura, Spectral subtraction iterated with weighting factors, in *Proceeding of IEEE Speech Coding Workshop* (2002), pp. 138–140
15. M.R. Khan, T. Hansen, Iterative noise power subtraction technique for improved speech quality, in *Proceeding of ICECE* (2008), pp. 391–394
16. S. Li, J.-Q. Wang, M. Niu, X.-J. Jing, T. Liu, Iterative spectral subtraction method for millimeter-wave conducted speech enhancement. *J. Biomed. Sci. Eng.* **2010**(3), 187–192 (2010)
17. T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano, K. Kondo, Theoretical analysis of iterative weak spectral subtraction via higher-order statistics, in *Proceeding of IEEE International Workshop on Machine Learning for Signal Processing* (2010), pp. 220–225
18. R. Miyazaki, H. Saruwatari, T. Inoue, Y. Takahashi, K. Shikano, K. Kondo, Musical-noise-free speech enhancement based on optimized iterative spectral subtraction. *IEEE Trans. Audio Speech Lang. Process.* **20**(7), 2080–2094 (2012)
19. R. Miyazaki, H. Saruwatari, S. Nakamura, K. Shikano, K. Kondo, J. Blanchette, M. Bouchard, Musical-noise-free blind speech extraction integrating microphone array and iterative spectral subtraction. *Signal Process.* (Elsevier) **102**, 226–239 (2014)
20. P. Comon, Independent component analysis, a new concept? *Signal Process.* (Elsevier) **36**, 287–314 (1994)
21. S. Araki, R. Mukai, S. Makino, T. Nishikawa, H. Saruwatari, The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech. *IEEE Trans. Speech Audio Process.* **11**(2), 109–116 (2003)
22. H. Sawada, R. Mukai, S. Araki, S. Makino, 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)
23. H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, K. Shikano, Blind source separation based on a fast-convergence algorithm combining ICA and beamforming. *IEEE Trans. Audio Speech Lang. Process.* **14**(2), 666–678 (2006)
24. A. Homayoun, M. Bouchard, Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment. *IEEE Trans. Audio Speech Lang. Process.* **17**(4), 521–533 (2009)
25. T. Inoue, H. Saruwatari, Y. Takahashi, K. Shikano, K. Kondo, 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)
26. H. Yu, T. Fingscheidt, A figure of merit for instrumental optimization of noise reduction algorithms, in *Proceeding of DSP in Vehicles* (2011)
27. H. Yu, T. Fingscheidt, Black box measurement of musical tones produced by noise reduction systems, in *Proceeding of ICASSP* (2012), pp. 4573–4576
28. S. Kanehara, H. Saruwatari, R. Miyazaki, K. Shikano, K. Kondo, Theoretical analysis of musical noise generation in noise reduction methods with decision-directed a priori SNR estimator, in *Proceeding of IWAENC* (2012)
29. S. Kanehara, H. Saruwatari, R. Miyazaki, K. Shikano, K. Kondo, Comparative study on various noise reduction methods with decision-directed a priori SNR estimator via higher-order statistics, in *Proceeding of APSIPA Annual Summit and Conference* (2012)
30. R. Miyazaki, H. Saruwatari, K. Shikano, K. Kondo, Musical-noise-free speech enhancement based on iterative Wiener filtering, in *Proceeding of IEEE International Symposium on Signal Processing and Information Technology* (2012)
31. S. Nakai, H. Saruwatari, R. Miyazaki, S. Nakamura, K. Kondo, Theoretical analysis of biased MMSE short-time spectral amplitude estimator and its extension to musical-noise-free speech enhancement, in *Proceeding of Hands-Free Speech Communication and Microphone Arrays* (2014)
32. H. Saruwatari, Statistical-model-based speech enhancement with musical-noise-free properties, in *Proceeding of IEEE International Conference on Digital Signal Processing* (2015), pp. 1201–1205
33. A. Hiroe, Solution of permutation problem in frequency domain ICA using multivariate probability density functions, in *Proceeding of ICA* (2006), pp. 601–608

34. T. Kim, H.T. Attias, S.-Y. Lee, T.-W. Lee, Blind source separation exploiting higher-order frequency dependencies. *IEEE Trans. Audio Speech Lang. Process.* **15**(1), 70–79 (2007)
35. N. Ono, Stable and fast update rules for independent vector analysis based on auxiliary function technique, in *Proceeding of WASPAA* (2011), pp. 189–192
36. D. Kitamura, N. Ono, H. Sawada, H. Kameoka, H. Saruwatari, Relaxation of rank-1 spatial constraint in overdetermined blind source separation, in *Proceeding of EUSIPCO* (2015), pp. 1271–1275
37. D. Kitamura, N. Ono, H. Sawada, H. Kameoka, H. Saruwatari, Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization. *IEEE/ACM Trans. Audio Speech Lang. Process.* **24**(9), 1626–1641 (2016)
38. Y. Mitsui, D. Kitamura, S. Takamichi, N. Ono, H. Saruwatari, Blind source separation based on independent low-rank matrix analysis with sparse regularization for time-series activity, in *Proceeding of ICASSP* (2017), pp. 21–25
39. S. Mogami, D. Kitamura, Y. Mitsui, N. Takamune, H. Saruwatari, N. Ono, Independent low-rank matrix analysis based on complex Student's  $t$ -distribution for blind audio source separation, in *Proceeding of IEEE International Workshop on Machine Learning for Signal Processing* (2017)
40. F.D. Aprilyanti, J. Even, H. Saruwatari, K. Shikano, S. Nakamura, T. Takatani, Suppression of noise and late reverberation based on blind signal extraction and Wiener filtering. *Acoust. Sci. Technol.* **36**(4), 302–313 (2015)
41. H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, Blind source separation combining independent component analysis and beamforming. *EURASIP J. Appl. Signal Process.* **2003**, 1135–1146 (2003)
42. Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, K. Shikano, Blind spatial subtraction array for speech enhancement in noisy environment. *IEEE Trans. Audio Speech Lang. Process.* **17**(4), 650–664 (2009)
43. Y. Takahashi, H. Saruwatari, K. Shikano, K. Kondo, Musical-noise analysis in methods of integrating microphone array and spectral subtraction based on higher-order statistics. *EURASIP J. Adv. Signal Process.* **2010**(431347), 25 pages (2010)
44. H. Saruwatari, Y. Ishikawa, Y. Takahashi, T. Inoue, K. Shikano, K. Kondo, 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)
45. R. Miyazaki, H. Saruwatari, K. Shikano, Theoretical analysis of amounts of musical noise and speech distortion in structure-generalized parametric spatial subtraction array. *IEICE Trans. Fundam.* **95-A**(2), 586–590 (2012)
46. S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, F. Itakura, Evaluation of blind signal separation method using directivity pattern under reverberant conditions, in *Proceeding of ICASSP*, vol. 5 (2000), pp. 3140–3143
47. J. Even, H. Saruwatari, K. Shikano, T. Takatani, Speech enhancement in presence of diffuse background noise: Why using blind signal extraction? in *Proceeding of ICASSP* (2010), pp. 4770–4773
48. J. Even, C. Ishi, H. Saruwatari, N. Hagita, Close speaker cancellation for suppression of non-stationary background noise for hands-free speech interface, in *Proceeding of INTERSPEECH* (2010), pp. 977–980
49. R. Prasad, H. Saruwatari, K. Shikano, Probability distribution of time-series of speech spectral components, *IEICE Trans. Fundam.* **E87-A**(3), 584–597 (2004)
50. R. Prasad, H. Saruwatari, K. Shikano, Estimation of shape parameter of GGD function by negentropy matching. *Neural Process. Lett.* **22**, 377–389 (2005)
51. T.H. Dat, K. Takeda, F. Itakura, Generalized gamma modeling of speech and its online estimation for speech enhancement, in *Proceeding of ICASSP*, vol. 4 (2005), pp. 181–184
52. I. Andrianakis, P.R. White, MMSE speech spectral amplitude estimators with chi and gamma speech priors, in *Proceeding of ICASSP* (2006), pp. III-1068–III-1071
53. R. Wakisaka, H. Saruwatari, K. Shikano, T. Takatani, Speech prior estimation for generalized minimum mean-square error short-time spectral amplitude estimator. *IEICE Trans. Fundam.* **95-A**(2), 591–595 (2012)

54. R. Wakisaka, H. Saruwatari, K. Shikano, T. Takatani, Speech kurtosis estimation from observed noisy signal based on generalized Gaussian distribution prior and additivity of cumulants, in *Proceeding of ICASSP* (2012), pp. 4049–4052
55. I. Cohen, Optimal speech enhancement under signal presence uncertainty using log-spectra amplitude estimator. *IEEE Signal Process. Lett.* **9**(4), 113–116 (2002)
56. H. Buchner, R. Aichner, W. Kellermann, 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)
57. Y. Mori, H. Saruwatari, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, Y. Ikeda, H. Hashimoto, T. Morita, Blind separation of acoustic signals combining SIMO-model-based independent component analysis and binary masking. *EURASIP J. Appl. Signal Process.* **2006**(34970), 17 pages (2006)
58. T. Hiekata, Y. Ikeda, T. Yamashita, T. Morita, R. Zhang, Y. Mori, H. Saruwatari, K. Shikano, Development and evaluation of pocket-size real-time blind source separation microphone. *Acoust. Sci. Technol.* **30**(4), 297–304 (2009)
59. Y. Omura, H. Kamado, H. Saruwatari, K. Shikano, Real-time semi-blind speech extraction with speaker direction tracking on Kinect, in *Proceeding of APSIPA Annual Summit and Conference* (2012)
60. Y. Bando, H. Saruwatari, N. Ono, S. Makino, K. Itoyama, D. Kitamura, M. Ishimura, M. Takakusaki, N. Mae, K. Yamaoka, Y. Matsui, Y. Ambe, M. Konyo, S. Tadokoro, K. Yoshii, H.G. Okuno, Low-latency and high-quality two-stage human-voice-enhancement system for a hose-shaped rescue robot. *J. Robot. Mechatron.* **29**(1), 198–212 (2017)