NAIST-IS-DD1261017

Doctoral Dissertation

Musical-Noise-Free Speech Enhancement Based on Higher-Order Statistics Pursuit

Ryoichi Miyazaki

February 6, 2014

Department of Information Science
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Ryoichi Miyazaki

Thesis Committee:

Professor Satoshi Nakamura	(Supervisor)
Professor Kenji Sugimoto	(Co-supervisor)
Associate Professor Hiroshi Saruwatari	(Co-supervisor)
Associate Professor Nobutaka Ono	(Co-supervisor)

Musical-Noise-Free Speech Enhancement Based on Higher-Order Statistics Pursuit*

Ryoichi Miyazaki

Abstract

In this dissertation, I propose a new speech enhancement theory for hearing aid and video conference systems, where the output speech quality of nonlinear signal processing is controlled using higher-order statistics. In these systems, since interference signals and noise deteriorate the quality of a users input speech, it is desirable to develop a digital signal processing technique to clean microphone signal before it is stored. In order to remove background noise, there have been many studies on noise reduction methods that have high noise reduction performance. However, the reduction of noise spectra often introduces an artificial distortion in the residual noise, which is the well-known phenomenon of so-called musical noise, leading to a serious deterioration of sound quality.

In this study, I first theoretically clarify that iterative spectral subtraction with a specific parameter generates almost no musical noise even with high noise reduction performance. On the basis of the fact, I propose a musical-noise-free theory for single-channel speech enhancement using iterative nonlinear signal processing. In the proposed theory, the fixed point in kurtosis yields the no-musical-noise state; we call this the "musical-noise-free condition." In addition, I mathematically derive the optimal internal parameter settings to satisfy the musical-noise-free condition based on higher-order statistics pursuit.

Next, I propose a new iterative blind signal extraction method that integrates blind noise estimation and iterative noise reduction to reduce nonstationary noise. This method includes a dynamic estimation of the noise power spectral density based on independent component analysis and multichannel Wiener filtering, which can provide

*Doctoral Dissertation, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1261017, February 6, 2014.

effective noise reduction even in the case that the noise has time-varying properties. From the experimental evaluation, it is asserted that the proposed methods are superior to conventional speech enhancement methods in terms of total sound quality.

Keywords:

Iterative spectral subtraction, musical-noise-free speech enhancement, blind speech extraction, higher-order statistics

Contents

List of Figures

List of Tables

1. Introduction

1.1 Background

Over the past few decades, many applications of speech communication systems, such as hands-free telecommunication systems,hearing aid systems, and video conference systems have been developed because speech is the most convenient medium for communication among human beings. However, since we live in an environment (noisy offices, crowded public spaces, and railway stations) where noise is inevitable and ubiquitous, speech signals are generally immersed in noise and can seldom be acquired and processed in a pure form. To make speech communication feasible, natural, and comfortable even in the presence of noise, it is desirable to develop a digital signal processing technique to clean a microphone signal before it is stored.

In order to remove background noise, there have been many studies on noise reduction. Noise reduction is concerned with improving some perceptual aspects of speech that has been degraded by additive noise. As single-channel noise reduction methods, subtraction-based methods [1, 2, 3], Wiener filtering approaches [4, 5], and statistical-model-based methods [6, 7, 8, 9, 10, 11] have been widely studied. These methods have high noise reduction performance with low computational complexity. However, in these methods, the enhancement of the noise spectra from the noisy spectra introduces an artificial distortion in the residual noise signal, which is known as *musical noise*, leading to a serious deterioration of sound quality.

To address the musical noise problem, there have been many studies on the analysis of musical noise generation in nonlinear signal processing, and methods aimed at its mitigation have been proposed [12, 13]. Such conventional musical noise mitigation methods are, unfortunately, designed to reduce musical noise generation at the cost of degrading the noise reduction performance. To achieve both high noise reduction performance and low musical noise generation, an *iterative spectral subtraction (SS)* method has recently been proposed [14, 15, 16, 17, 18, 19]. This method is performed through signal processing in which *weak* SS processes are recursively applied to the input signal. The methodology used in iterative SS is of great interest to researchers working on nonlinear signal processing and machine learning because it addresses the inherent question of whether or not recursive weak (nonlinear) signal processing can provide better performance. Although the effectiveness of the iterative SS method has

been reported experimentally, to the best of our knowledge, there have been no studies on the theoretical advantages of iterative SS. One reason for this is the difficulty of theoretical study due to the fact that no objective metric to measure how much musical noise is generated has been proposed.

Recently, it has been demonstrated that the amount of generated musical noise strongly correlates with the difference between higher-order statistics of the power spectra before and after nonlinear signal processing [20, 21, 22, 23]. This fact enables us to analyze the amount of musical noise generated through nonlinear signal processing. Furthermore, on the basis of higher-order statistics, a mathematical metric for musical noise generation that can be used as an objective measure has been established [20, 21]. Some researchers have theoretically clarified features of the musical noise generation in various speech enhancement methods based on this finding [24, 25, 26, 27, 28, 29]. However, no noise reduction method that generates no musical noise has been proposed.

On the other hand, in commonly used noise reduction methods, it is 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. In contrast, under real-world acoustical environments, such as a nonstationary noise field, it is necessary to dynamically estimate noise. As a well-known single-channel noise power spectral density (PSD) estimation method, Martin proposed an algorithm for noise estimation based on minimum statistics [3]. In this method, the noise is estimated from the minimum values of a smoothed power estimate of the noisy signal, which is multiplied by a factor to compensate for the bias. However, this noise estimate is sensitive to outliers and less stable when the noise is rapidly varying and speech is continuously present at a certain frequency. More recent spectral noise power estimators allow quicker tracking of noise power spectra, e.g., minimum-mean-square error (MMSE) based approaches [30, 31]. In the MMSE based estimator, a limited maximum likelihood estimate of the *a priori* SNR is used to estimate the periodogram of the noise signal. However, the accuracy of noise estimation is not sufficient.

It is well known that an approach using a microphone array is effective for improving the accuracy of noise estimation, and many methods of integrating microphone array signal processing for noise estimation and nonlinear signal processing

for noise reduction have been studied with the aim of achieving better noise reduction [32, 33, 34, 35, 36, 37, 38, 39]. These integrating methods can achieve higher noise reduction performance than that obtained using conventional adaptive microphone arrays. However, these methods always suffer from musical noise owing to nonlinear signal processing.

In conclusion, there is no effective method for achieving perfect "musical-noise-less" properties, even under stationary noise conditions. Also, in case of the nonstationary noise, the development of such a musical-noise-less method is strongly required. The above-mentioned problems require urgent attention.

1.2 Scope of thesis

1.2.1 Theory of musical-noise-free speech enhancement

To achieve high-quality speech enhancement with less musical noise, in this dissertation, I propose a new method for optimization of the performance in iterative SS. Although commonly used noise reduction methods have high noise reduction performance, musical noise arises, leading to a serious deterioration of sound quality (see Fig. 1). Therefore, it is desirable and very challenging to achieve both high noise reduction performance and less (or ideally no) musical noise generation.

Recently, a very interesting phenomenon has been found: the recursive use of very weak SS with appropriate parameters gives *equilibrium* behavior in the growth of higher-order statistics with increasing number of iterations. This means that almost no musical noise is generated even with high noise reduction (see Fig. 1), which is one of the most desirable properties of single-channel nonlinear noise reduction methods. Hereafter, I refer to this phenomenon as a *musical-noise-free* condition. In this study, I theoretically derive a closed-form solution of the internal parameters that satisfy the musical-noise-free condition based on the analysis of higher-order statistics.

1.2.2 Toward musical-noise-free blind speech extraction

In the previous proposed methods, however, it was assumed that the input noise signal is stationary, meaning that we can estimate the expectation of the noise power spectral density from a time-frequency period of a signal that contains only noise. In contrast,

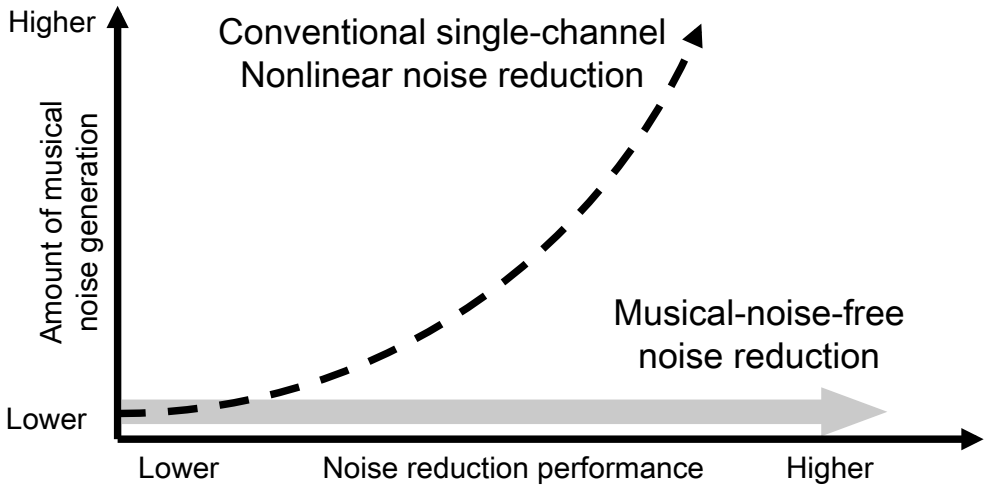

Fig. 1. Relation between conventional noise reduction methods and proposed method.

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 dissertation, I propose a new iterative signal extraction method using a microphone array that can be applied to nonstationary noise. This proposed method consists of iterative blind dynamic noise estimation by independent component analysis (ICA) [40] 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. This method can be applied to nonstationary noise, and almost no musical noise is generated because noisy speech is extracted by musical-noise-free speech extraction. Figure 2 shows the relationship between the conventional single-channel/multichannel noise reduction methods and the proposed method. As shown in Fig. 2, it is expected that the proposed method will generate less speech distortion than the conventional single-channel noise reduction methods and less musical noise than the conventional multichannel speech extraction methods.

Next, in relation to the proposed method, I discuss the justification of applying ICA to signals nonlinearly distorted by SS. I theoretically clarify that the degradation in ICA-based noise estimation obeys an amplitude variation in room transfer function between the target user and microphones. Furthermore, to reduce speech distortion, I introduce a channel selection strategy into ICA, where less varied inputs are automati-

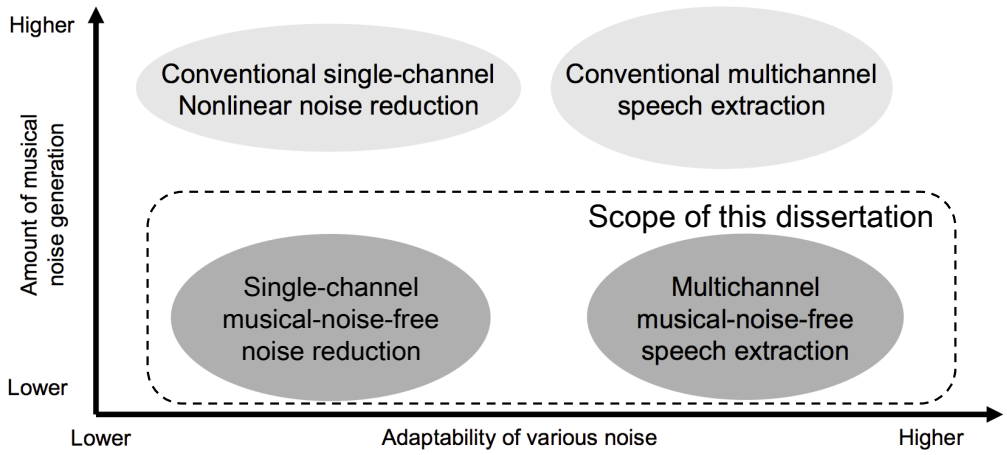

Fig. 2. Relationship between conventional single-channel/multichannel noise reduction method and proposed methods.

cally chosen to maintain the high accuracy of noise estimation. In addition, I introduce a time-variant noise PSD estimator [41] instead of ICA to improve the noise estimation accuracy.

1.3 Overview of dissertation

The dissertation is organized as follows.

First, I describe related works on non-iterative and iterative SSs in Sect. 2. In this section, the mathematical metric of musical noise generation is also explained.

In Sect. 3, a theoretical analysis of iterative SS is given. It is clarified by mathematical analysis that iterative SS with very weak processing can realize high-quality speech enhancement with a small amount of musical noise generated.

On the basis of this findings, I propose a new speech enhancement theory, i.e., musical-noise-free speech enhancement, in Sect. 4. In this section, I discuss a theorem of musical-noise-free conditions in iterative SS, and I mathematically derive the internal parameter settings to satisfy the musical-noise-free condition. It is clarified that the optimal parameters satisfying the musical-noise-free condition can generate almost no musical noise even with high noise reduction.

Next, I propose a musical-noise-free blind speech extraction method using a micro-

phone array that can be applied to nonstationary noise in Sect. 5. Also, in relation to the proposed method, I discuss the justification of applying ICA to signals nonlinearly distorted by SS. Moreover, to achieve higher accuracy of noise estimation, I propose the introduction of a channel selection strategy in ICA and a time-variant noise PSD estimator.

Finally, I summarize the contributions of this dissertation and provide suggestions for future work in Sect. 6.

2. Conventional Single-Channel Nonlinear Speech Enhancement Methods and Mathematical Metric of Musical Noise Generation

2.1 Introduction

In this section, I describe conventional speech enhancement methods and its problem. In recent years, many types of single-channel speech enhancement methods have been proposed and studied. Then, in this section, I review commonly used speech enhancement methods, *non-iterative SS* and *iterative SS*, respectively. These methods have high noise reduction performance with low computational complexity. However, these methods always suffer from artificial distortion, so-called *musical noise*, owing to nonlinear signal processing, leading to a serious deterioration of sound quality. Also, there is no general measure of the amount of musical noise. Moreover, it is well known that the degree of musical noise varies according to the noise environment; this leads to difficulty of parameter settings in SS.

Firstly, I review two types of single-channel nonlinear speech enhancement methods in Sect. 2.2. Next, I give a brief review of musical noise and its objective metric based on higher-order statistics in Sect. 2.3. Finally, Sect. 2.4 concludes this section.

2.2 Conventional single-channel nonlinear speech enhancement methods

2.2.1 Non-iterative SS

We apply short-time Fourier analysis 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:

$$y(f,\tau) = \begin{cases} \sqrt{|o(f,\tau)|^2 - \beta \cdot \mathrm{E}[|N|^2]}\, e^{j\arg(o(f,\tau))} \\ \quad \text{(where} \quad |o(f,\tau)|^2 - \beta \cdot \mathrm{E}[|N|^2] > 0\text{)}, \\ \eta o(f,\tau) \quad \text{(otherwise)}, \end{cases} \tag{1}$$

where $y(f,\tau)$ is the enhanced target speech signal, $o(f,\tau)$ is the observed signal, f denotes the frequency subband, τ is the frame index, β is the oversubtraction parameter,

and η is the flooring parameter. Here, $\mathrm{E}[|N|^2]$ is the expectation of the random variable $|N|^2$ corresponding to noise power spectra. Calculation of $\mathrm{E}[|N|^2]$ is a problem of noise PSD estimation. In practice, if we can use a voice activity detector (VAD), we can approximate $\mathrm{E}[|N|^2]$ by averaging the observed noise power spectra $|n(f,\tau)|^2$ in the specific K-sample frames, where we assume speech absence in this period;

$$\widehat{\mathrm{E}[|N|^2]} \approx \frac{1}{K}\sum_{\tau=k'}^{k'+K}|n(f,\tau)|^2. \tag{2}$$

In addition, many methods for dynamic estimation of the expectation of the noise PSD have been proposed [3].

Generally speaking, conventional SS 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.

2.2.2 Iterative SS

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, 15, 16, 17, 18, 19]. This method is performed through signal processing, in which the following *weak* SS processes are iteratively applied to the noise signal (see Fig. 3):

(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 η. Note that in this dissertation we still call such a large flooring case "weak" even when we employ a large oversubtraction parameter β because many subtracted components are floored.

(III) We then return to step (I) and substitute the resultant output (partially noise-reduced signal) for the input signal.

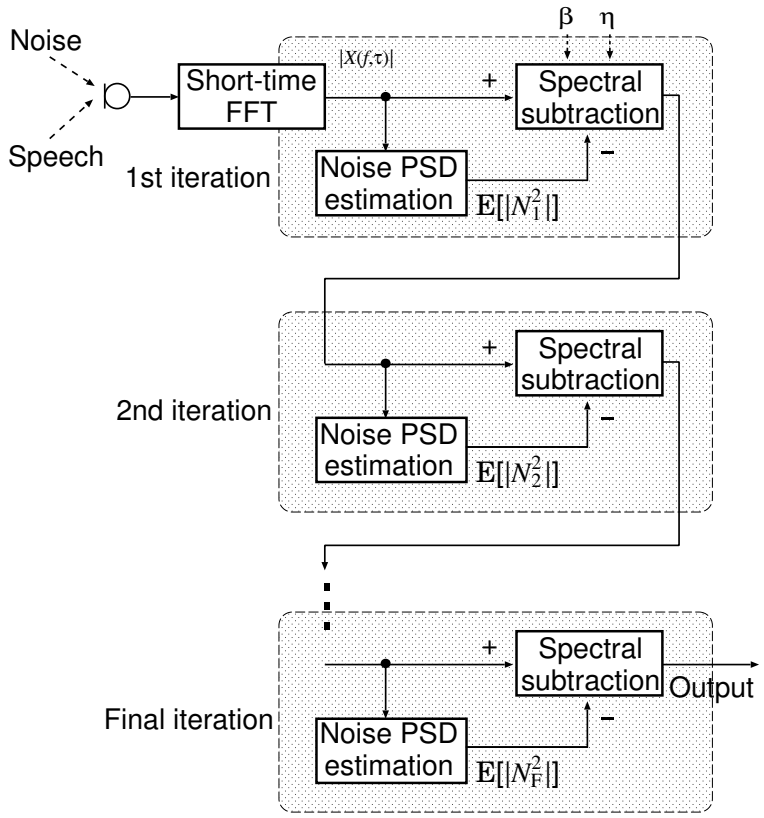

Fig. 3. Block diagram of iterative SS.

Although the efficacy of the iterative SS method has been reported experimentally, the theoretical or mathematical justification of its principles has not yet been presented. Intuitively, it appears that weak subtraction generates little musical noise in each iteration. However, if we require sufficient noise reduction, a huge number of iterations are needed, causing the amount of musical noise to accumulate. Moreover, it is not self-evident that the accumulated musical noise is smaller than that obtained by conventional non-iterative SS. Therefore, the lack of justification for iterative SS reduces its applicability to general noise reduction. Also the proof of the theoretical basis of the method remains as an open problem.

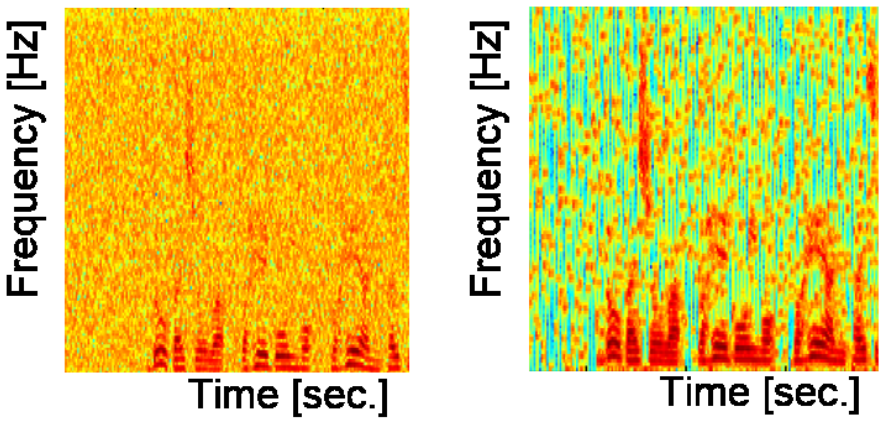

Fig. 4. (a) Observed spectrogram and (b) processed spectrogram.

2.3 Mathematical metric of musical noise generation via higher-order statistics [20]

2.3.1 Relation between kurtosis and musical noise generation

It is well known that the amount of musical noise is highly correlated with the number of isolated power spectral components and their level of isolation. In this dissertation, I call these isolated components *tonal components*. Figure 4 shows an example of a spectrogram of musical noise in which many tonal components can be observed. 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, Uemura, et al. have introduced kurtosis, one of the most commonly used higher-order statistics, to evaluate the percentage of tonal components among all components [20]. A larger kurtosis value indicates a signal with a heavy tail, meaning that the signal has many tonal components.

2.3.2 Kurtosis and kurtosis ratio

Kurtosis is one of the most commonly used higher-order statistics for the assessment of non-Gaussianity. Kurtosis is defined as

$$\text{kurt} = \frac{\mu_4}{\mu_2^2}, \tag{3}$$

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

$$\mu_m = \int_0^\infty x^m P(x) dx, \tag{4}$$

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, (3) is not kurtosis in the mathematically strict definition but a modified version; however, we still refer to (3) as kurtosis in this dissertation.

In this study, I apply such a kurtosis-based analysis to a *noise-only time-frequency period* of subject signals for the assessment of musical noise, even though these signals contain target-speech-dominant periods. Thus, this analysis should be conducted during, for example, speech absence periods. 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, the *kurtosis ratio* [20] has been proposed as a measure to assess musical noise:

$$\text{kurtosis ratio} = \frac{\text{kurt}_{\text{proc}}}{\text{kurt}_{\text{org}}}, \tag{5}$$

where $\text{kurt}_{\text{proc}}$ is the kurtosis of the processed signal and kurt_{org} is the kurtosis of the observed signal. This measure increases as the amount of generated musical noise increases. In Ref. [20], it was reported that the kurtosis ratio is strongly correlated with the human perception of musical noise.

2.4 Conclusion

In this section, conventional single-channel nonlinear speech enhancement methods were denoted. Next, mathematical metric of musical noise generation via higher-order statistics was reviewed.

3. Theoretical Analysis of Iterative SS

3.1 Introduction

In the previous section, I described the two types of conventional noise reduction methods and the objective measure for the musical noise generation on the basis of higher-order statistics. In this section, I conduct an analysis of the amounts of noise reduction performance and musical noise generation through iterative SS using higher-order statistics.

In this analysis, I first model a noise signal as a gamma distribution (see Sect. 3.2) and formulate the resultant p.d.f. after non-iterative SS (see Sect. 3.3). Then, the generalized form of the mth-order moment is derived (see Sect. 3.4). Next, on the basis of the above-mentioned analysis, I formulate the behavior of iteratively applied SS and compare the kurtosis values upon changing the parameter settings under the same amount of noise reduction (see Sect. 3.5).

Note that in Ref. [24] Inoue, et al. have partly formulated the mth-order moment only in the case that the flooring process is omitted. In contrast, the derivation of the mth-order moment in this study is a more general form taking into account the flooring effect that plays an important role for controlling the degree of weakness in SS.

3.2 Modeling of input signal

I assume that the input signal x in the power spectral domain can be modeled by the gamma distribution as [42, 43]

$$P(x) = \frac{x^{\alpha-1}\exp\{-x/\theta\}}{\theta^{\alpha}\Gamma(\alpha)}, \tag{6}$$

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

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}\exp(-t)dt. \tag{7}$$

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 $\alpha < 1$. I make assumption here that θ is assumed to be the deterministically known noise PSD and estimation artifacts of the noise PSD are not taken into account in this dissertation.

3.3 Process of deforming p.d.f. of noise via conventional non-iterative SS

In conventional non-iterative SS, 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 β is subtracted from the observed power spectrum. When a gamma distribution is used to model the noise signal, its mean is $\alpha\theta$. Thus, the amount of subtraction is $\beta\alpha\theta$. 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

$$\frac{1}{\theta^{\alpha}\Gamma(\alpha)}(z+\beta\alpha\theta)^{\alpha-1}\exp\left\{-\frac{z+\beta\alpha\theta}{\theta}\right\}, \tag{8}$$

where z is the random variable of the p.d.f. after SS.

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 η (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

$$\frac{1}{(\eta^2\theta)^{\alpha}\Gamma(\alpha)}z^{\alpha-1}\exp\left\{-\frac{z}{\eta^2\theta}\right\}. \tag{9}$$

Note that the flooring parameter η is squared in the p.d.f. because the multiplication of η is conducted in the amplitude spectrum domain (see the second branch in (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 SS, $P_{\mathrm{SS}}(z)$, can be written as

$$P_{\mathrm{SS}}(z) = \begin{cases} \frac{1}{\theta^{\alpha}\Gamma(\alpha)}(z+\beta\alpha\theta)^{\alpha-1}\exp\left\{-\frac{z+\beta\alpha\theta}{\theta}\right\} \\ \quad + \frac{1}{(\eta^2\theta)^{\alpha}\Gamma(\alpha)}z^{\alpha-1}\exp\left\{-\frac{z}{\eta^2\theta}\right\} & (0 \le z < \beta\alpha\eta^2\theta), \\ \frac{1}{\theta^{\alpha}\Gamma(\alpha)}(z+\beta\alpha\theta)^{\alpha-1}\exp\left\{-\frac{z+\beta\alpha\theta}{\theta}\right\} & (\beta\alpha\eta^2\theta \le z). \end{cases} \tag{10}$$

3.4 The mth-order moment of $P_{\mathrm{SS}}(z)$ in conventional non-iterative SS

To characterize non-iterative SS, the mth-order moment of z is required. For $P_{\mathrm{SS}}(z)$, the mth-order moment is given by

$$\begin{aligned} \mu_m^{\mathrm{SS}} &= \int_0^{\infty} z^m \cdot P_{\mathrm{SS}}(z)dz \\ &= \int_0^{\infty} z^m \frac{1}{\theta^{\alpha}\Gamma(\alpha)}(z+\beta\alpha\theta)^{\alpha-1}\exp\left\{-\frac{z+\beta\alpha\theta}{\theta}\right\}dz \\ &\quad + \int_0^{\beta\alpha\eta^2\theta} z^m \frac{1}{(\eta^2\theta)^{\alpha}\Gamma(\alpha)}z^{\alpha-1}\exp\left\{-\frac{z}{\eta^2\theta}\right\}dz, \end{aligned} \tag{11}$$

where z is the random variable of the p.d.f. after SS. We now expand the first term of the right-hand side of (11). Here, let $t = (z+\beta\alpha\theta)/\theta$, then $\theta dt = dz$ and $z = \theta(t-\beta\alpha)$. Consequently,

$$\begin{aligned} &\int_0^{\infty} z^m \frac{1}{\theta^{\alpha}\Gamma(\alpha)}(z+\beta\alpha\theta)^{\alpha-1}\exp\left\{-\frac{z+\beta\alpha\theta}{\theta}\right\}dz \\ &= \int_{\beta\alpha}^{\infty} \theta^m(t-\beta\alpha)^m \frac{1}{\theta^{\alpha}\Gamma(\alpha)}(\theta t)^{\alpha-1}\exp\{-t\}\theta dt \\ &= \frac{\theta^m}{\Gamma(\alpha)}\int_{\beta\alpha}^{\infty}\sum_{l=0}^{m}(-\beta\alpha)^l\frac{\Gamma(m+1)}{\Gamma(l+1)\Gamma(m-l+1)}t^{m-l}t^{\alpha-1}\exp\{-t\}dt \\ &= \frac{\theta^m}{\Gamma(\alpha)}\sum_{l=0}^{m}(-\beta\alpha)^l\frac{\Gamma(m+1)}{\Gamma(l+1)\Gamma(m-l+1)}\Gamma(\alpha+m-l,\beta\alpha), \end{aligned} \tag{12}$$

where we use the binomial theorem given by

$$(t+a)^m = \sum_{l=0}^{m} a^l \frac{\Gamma(m+1)}{\Gamma(l+1)\Gamma(m-l+1)}t^{m-l}, \tag{13}$$

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

$$\Gamma(a, b) = \int_b^\infty t^{a-1} \exp\{-t\} dt. \tag{14}$$

Next we consider the second term of the right-hand side of (11). Here, let $t = z/(\eta^2\theta)$, then $\eta^2\theta dt = dz$. Thus,

$$\begin{aligned}\int_0^{\beta\alpha\eta^2\theta} z^m \frac{1}{(\eta^2\theta)^\alpha \Gamma(\alpha)} z^{\alpha-1} \exp\left\{-\frac{z}{\eta^2\theta}\right\} dz &= \int_0^{\beta\alpha} (\eta^2\theta t)^m \frac{1}{(\eta^2\theta)^\alpha \Gamma(\alpha)} (\eta^2\theta t)^{\alpha-1} \exp\{-t\} \eta^2\theta dt \\ &= \frac{\eta^{2m}\theta^m}{\Gamma(\alpha)} \int_0^{\beta\alpha} t^{\alpha-1+m} \exp\{-t\} dt \\ &= \frac{\eta^{2m}\theta^m}{\Gamma(\alpha)} \gamma(\alpha + m, \beta\alpha), \end{aligned} \tag{15}$$

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

$$\gamma(a, b) = \int_0^b t^{a-1} \exp\{-t\} dt. \tag{16}$$

As a result, the mth-order moment after SS, μ_m^{SS}, is a composite of (12) and (15), and is given as

$$\mu_m^{\mathrm{SS}} = \theta^m \mathcal{M}(\alpha, \beta, \eta, m), \tag{17}$$

where we refer to $\mathcal{M}(\alpha, \beta, \eta, m)$ as *normalized moment function* as

$$\begin{aligned}\mathcal{M}(\alpha, \beta, \eta, m) = &\frac{1}{\Gamma(\alpha)} \sum_{l=0}^{m} (-\beta\alpha)^l \frac{\Gamma(m+1)}{\Gamma(l+1)\Gamma(m-l+1)} \Gamma(\alpha + m - l, \beta\alpha) \\ &+ \frac{\eta^{2m}}{\Gamma(\alpha)} \gamma(\alpha + m, \beta\alpha). \end{aligned} \tag{18}$$

3.5 Analysis of behavior of iterative SS

3.5.1 Amount of musical noise generated

In this subsection, we formulate the amount of musical noise generated in the iterative SS method using the analytical result obtained in Sect. 3.4. Here we conduct a *recursively applied* kurtosis analysis in the following manner, where the subscript i represents the value in the ith iteration:

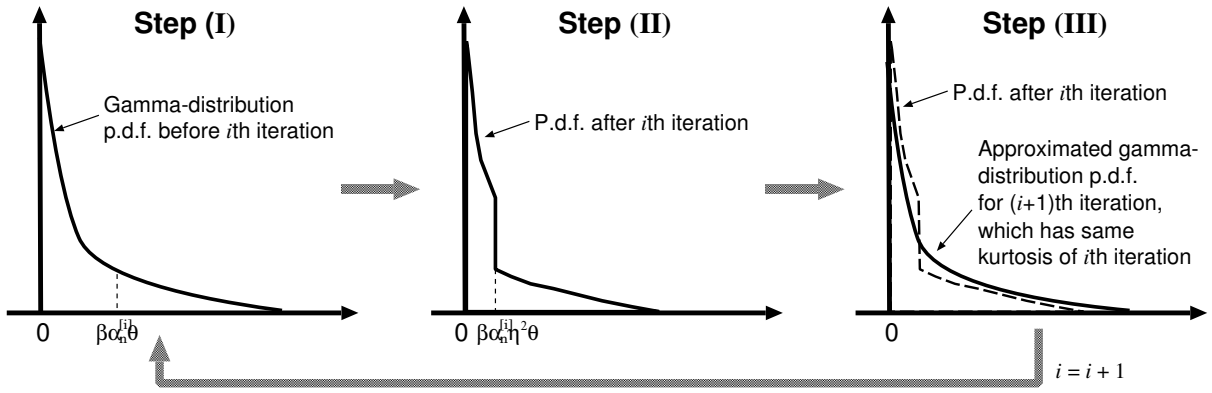

Fig. 5. P.d.f. deformation and approximated gamma-distribution p.d.f. for $(i+1)$th iteration, which has same kurtosis of p.d.f. after ith iteration.

(I) First, model the input noise p.d.f. as a gamma distribution with shape parameter α_i (initially $i = 0$).

(II) Next, apply SS to the signal using the oversubtraction parameter β and flooring parameter η. We calculate the kurtosis using (17); this is considered as the result of the ith iteration.

(III) Next, approximately remodel the resultant processed signal as a gamma distribution with the modified shape parameter α_{i+1} corresponding to the resultant kurtosis obtained in step (II) (see Fig. 5). Then return to step (I) with the updated value of α_{i+1}.

Note that this analysis includes an approximation of the p.d.f. modification in which the p.d.f. is always remodeled as a gamma distribution in each iteration. This is necessary because it is difficult to derive an exact analytical expression for the change in kurtosis of a non-gamma distribution. The proposed approximation is, however, still valid if the SS process in each step is weak and thus does not change the p.d.f. significantly (see appendix A).

In the following, full details of the iterative analysis are given. The kurtosis in the ith iteration is obtained via steps (I) and (II) using (17) with $\alpha = \alpha_i$ as

$$\begin{aligned}\mathrm{kurt}(\alpha_i, \beta, \eta) &= \frac{\mu_4^{\mathrm{SS}}}{(\mu_2^{\mathrm{SS}})^2} \\ &= \frac{\theta^4 \mathcal{M}(\alpha_i, \beta, \eta, 4)}{\{\theta^2 \mathcal{M}(\alpha_i, \beta, \eta, 2)\}^2} \\ &= \frac{\mathcal{M}(\alpha_i, \beta, \eta, 4)}{\mathcal{M}^2(\alpha_i, \beta, \eta, 2)}. \end{aligned} \tag{19}$$

In step (III), a new α_{i+1} can be calculated using the following relation between the kurtosis and the shape parameter. First, I obtain the 2nd-order moment of the gamma distribution with α_{i+1} as

$$\mu_2 = \int_0^\infty x^2 P(x) dx = \int_0^\infty x^2 \frac{1}{\theta^{\alpha_{i+1}} \Gamma(\alpha_{i+1})} \cdot x^{\alpha_{i+1}-1} \exp\{-x/\theta\} dx. \tag{20}$$

Here, let $X = x/\theta$, then this moment can be rewritten as

$$\begin{aligned}\mu_2 &= \frac{1}{\theta^{\alpha_{i+1}} \Gamma(\alpha_{i+1})} \cdot \theta^{\alpha_{i+1}+2} \int_0^\infty X^{(\alpha_{i+1}+2)-1} \exp\{-X\} dX \\ &= \frac{\theta^2}{\Gamma(\alpha_{i+1})} \Gamma(\alpha_{i+1} + 2) \\ &= \theta^2 (\alpha_{i+1} + 1) \alpha_{i+1}, \end{aligned} \tag{21}$$

where I use the following well-known functional equation of the gamma function:

$$\Gamma(\alpha + j) = (\alpha + j - 1)(\alpha + j - 2) \cdots (\alpha) \Gamma(\alpha). \tag{22}$$

Next, in the same manner, the 4th-order moment can be expressed as

$$\begin{aligned}\mu_4 &= \int_0^\infty x^4 \frac{1}{\theta^{\alpha_{i+1}} \Gamma(\alpha_{i+1})} \cdot x^{\alpha_{i+1}-1} \exp\{-x/\theta\} dx \\ &= \frac{\theta^4}{\Gamma(\alpha_{i+1})} \Gamma(\alpha_{i+1} + 4) \\ &= \theta^4 (\alpha_{i+1} + 3)(\alpha_{i+1} + 2)(\alpha_{i+1} + 1) \alpha_{i+1}. \end{aligned} \tag{23}$$

Using (21) and (23), I have

$$\frac{\mu_4}{\mu_2^2} = \frac{(\alpha_{i+1} + 3)(\alpha_{i+1} + 2)}{(\alpha_{i+1} + 1)\alpha_{i+1}} = \mathrm{kurt}(\alpha_i, \beta, \eta). \tag{24}$$

This results in the following quadratic equation in α_{i+1} to be solved:

$$(1 - \mathrm{kurt}(\alpha_i, \beta, \eta))\alpha_{i+1}^2 + (5 - \mathrm{kurt}(\alpha_i, \beta, \eta))\alpha_{i+1} + 6 = 0, \tag{25}$$

and I can derive a closed-form estimate of the shape parameter from the given kurtosis as

$$\begin{aligned}\alpha_{i+1} &= \frac{\mathrm{kurt}(\alpha_i, \beta, \eta) - 5 - \sqrt{\mathrm{kurt}(\alpha_i, \beta, \eta)^2 + 14\,\mathrm{kurt}(\alpha_i, \beta, \eta) + 1}}{2 - 2\,\mathrm{kurt}(\alpha_i, \beta, \eta)} \\ &= \mathcal{A}(\mathrm{kurt}(\alpha_i, \beta, \eta)),\end{aligned} \tag{26}$$

where I define

$$\mathcal{A}(k) = (k - 5 - \sqrt{k^2 + 14k + 1})(2 - 2k)^{-1}. \tag{27}$$

Note that in the derivation of (26), I chose the shape parameter α_{i+1} to be greater than 0, assuming that the p.d.f. of the noise power spectra is Gaussian or super-Gaussian and thus $\mathrm{kurt}(\alpha_i, \beta, \eta) \geq 6$. By applying the updated α_{i+1} to the new gamma distribution, I can obtain the following recursive equation for the kurtosis in the $(i + 1)$th iteration:

$$\begin{aligned}\mathrm{kurt}(\alpha_{i+1}, \beta, \eta) &= \frac{\theta^4 \mathcal{M}(\alpha_{i+1}, \beta, \eta, 4)}{\{\theta^2 \mathcal{M}(\alpha_{i+1}, \beta, \eta, 2)\}^2} \\ &= \frac{\mathcal{M}(\mathcal{A}(\mathrm{kurt}(\alpha_i, \beta, \eta)), \beta, \eta, 4)}{\mathcal{M}^2(\mathcal{A}(\mathrm{kurt}(\alpha_i, \beta, \eta)), \beta, \eta, 2)}.\end{aligned} \tag{28}$$

Thus, I can calculate the resultant kurtosis ratio as

$$\begin{aligned}\text{kurtosis ratio} &= \frac{\mathrm{kurt}(\alpha_{i+1}, \beta, \eta)}{\mathrm{kurt}(\alpha_0, 0, 0)} \\ &= \frac{\alpha_0(\alpha_0 + 1)}{(\alpha_0 + 3)(\alpha_0 + 2)} \frac{\mathcal{M}(\mathcal{A}(\mathrm{kurt}(\alpha_i, \beta, \eta)), \beta, \eta, 4)}{\mathcal{M}^2(\mathcal{A}(\mathrm{kurt}(\alpha_i, \beta, \eta)), \beta, \eta, 2)}.\end{aligned} \tag{29}$$

3.5.2 Amount of noise reduction

In this subsection, I analyze the amount of noise reduction by carrying out the same iterative analysis as that described in Sect. 3.5.1, including the approximation of the gamma distribution modeling. Hereafter, I define the *noise reduction rate* (NRR) as a measure of the noise reduction performance, which is defined as the output SNR in dB

minus the input SNR in dB [44]. The NRR is

$$\begin{aligned}\mathrm{NRR} &= 10\log_{10}\frac{\mathrm{E}[s_{\mathrm{out}}^2]/\mathrm{E}[n_{\mathrm{out}}^2]}{\mathrm{E}[s_{\mathrm{in}}^2]/\mathrm{E}[n_{\mathrm{in}}^2]} \\ &\simeq 10\log_{10}\frac{\mathrm{E}[n_{\mathrm{in}}^2]}{\mathrm{E}[n_{\mathrm{out}}^2]},\end{aligned} \tag{30}$$

where s_{in} and s_{out} are the input and output speech signals, respectively, and n_{in} and n_{out} are the input and output noise signals, respectively. In addition, I assume that the amount of noise reduction is much larger than that of speech distortion in , i.e., $\mathrm{E}[s_{\mathrm{out}}^2] \simeq \mathrm{E}[s_{\mathrm{in}}^2]$.

Here, the NRR achieved after the ith iteration is defined by $\mathrm{NRR}_i(\beta, \eta)$. It is obvious that the NRR additionally accumulates in each iteration because it is the logarithm of the power ratio between the input and processed noises. The relative improvement in the $(i+1)$th iteration can be given via the 1st-order moment with $\alpha = \alpha_{i+1}$, as

$$\begin{aligned}10\log_{10}\frac{\mathrm{E}[x]}{\mathrm{E}[z]} &= 10\log_{10}\frac{\theta\mathcal{M}(\alpha_{i+1},0,0,1)}{\theta\mathcal{M}(\alpha_{i+1},\beta,\eta,1)} \\ &= 10\log_{10}\frac{\alpha_{i+1}}{\mathcal{M}(\alpha_{i+1},\beta,\eta,1)}.\end{aligned} \tag{31}$$

Thus, using (26) and (31), the resultant NRR after the $(i+1)$th iteration is recursively expressed as

$$\begin{aligned}&\mathrm{NRR}_{i+1}(\beta,\eta) \\ &= 10\log_{10}\frac{\mathcal{A}(\mathrm{kurt}(\alpha_i,\beta,\eta))}{\mathcal{M}(\mathcal{A}(\mathrm{kurt}(\alpha_i,\beta,\eta),\beta,\eta,1))} + \mathrm{NRR}_i(\beta,\eta).\end{aligned} \tag{32}$$

In summary, I can derive theoretical estimates for the amount of musical noise generated and NRR using (29) and (32), respectively. This greatly simplifies the analysis because both equations are expressed analytically in a form that does not include any integrals.

3.6 Example of Theoretical Behavior

According to the previous analysis, I can trace the amount of musical noise generated in iterative SS along with the NRR. In addition, this can be compared with the amount

generated in the conventional non-iterative SS method under the same amount of noise reduction.

Figure 6(a) shows the theoretical behavior of the kurtosis ratio and NRR for several parameter settings, where the shape parameter α_0 is set to 1.0, i.e., the input noise signal is assumed to be Gaussian. In the iterative SS method, the oversubtraction parameter β is fixed to 2.4, and the flooring parameter η is set to 0.5, 0.7, and 0.9, corresponding to *normal*, *moderately weak*, and *very weak* processing in each iteration, respectively. In conventional non-iterative SS, the oversubtraction parameter β is manually adjusted (the flooring parameter η is fixed to 0.1) so that the NRR is varied as 0, 0.5, 1.0, ..., 12.0 dB. I plot black circles symbolic of the conventional non-iterative SS on the coordinates of the NRR and kurtosis ratio, which are given by

$$\begin{aligned}&(\text{NRR},\ \text{kurtosis ratio})\\&= \left(10\log_{10}\frac{\alpha_0}{\mathcal{M}(\alpha_0,\beta,\eta,1)},\ \frac{\alpha_0(\alpha_0+1)}{(\alpha_0+3)(\alpha_0+2)}\frac{\mathcal{M}(\alpha_0,\beta,\eta,4)}{\mathcal{M}^2(\alpha_0,\beta,\eta,2)}\right)\end{aligned} \tag{33}$$

for different β independently.

From Fig. 6(a), I can confirm the following interesting results:

- The iterative use of very weak SS (e.g., $\eta = 0.9$) can simultaneously achieve a large NRR and a small kurtosis ratio after a large number of iterations, meaning that I can realize high-quality speech enhancement with a small amount of musical noise generated. This is strong theoretical evidence of the advantageousness of iterative SS.

- Moreover, there exists an appropriate parameter setting ($\eta = 0.9$) that gives *equilibrium* behavior in the growth of the kurtosis ratio, i.e., a flat kurtosis ratio trajectory with a value of appropriately unity. This corresponds to the remarkable phenomenon that almost *no* musical noise is generated in iterative SS (i.e., musical-noise-free), unlike conventional single-channel nonlinear noise reduction, which always generates musical noise to some extent.

- In contrast, if we use strong subtraction with a small flooring parameter (e.g., $\eta = 0.5$), the above-mentioned equilibrium is violated, resulting in a large kurtosis ratio compared with that of conventional non-iterative SS under the same NRR.

This suggests that the iterative method is not always advantageous, and that the values of the parameters should be carefully set.

Next, Fig. 6(b) depicts the kurtosis ratio and NRR for another example, where the shape parameter α_0 is set to 0.2 with the assumption of super-Gaussian noise. The oversubtraction parameter β and flooring parameter η are set to 8.5 and 0.9, respectively, in iterative SS. From Fig. 6(b), I can also confirm that the iterative SS method generates less musical noise even for super-Gaussian noise, producing the equilibrium behavior of the kurtosis ratio.

3.7 Conclusion

In this section, I gave the theoretical analysis of iterative SS in terms of noise reduction performance and the amount of musical noise generation. My theoretical analysis indicates that the first-, second-, and fourth-order moments of the power spectra can be used to estimate the amount of noise reduction and musical noise generation. Also, I introduced a gamma-distribution approximation for simulating iteratively applied weak SS.

Next, I conducted a comparison of the amount of musical noise generated for different parameter settings under the same noise reduction performance. It was clarified from mathematical analysis that iterative SS with very weak processing can result in less musical noise being generated.

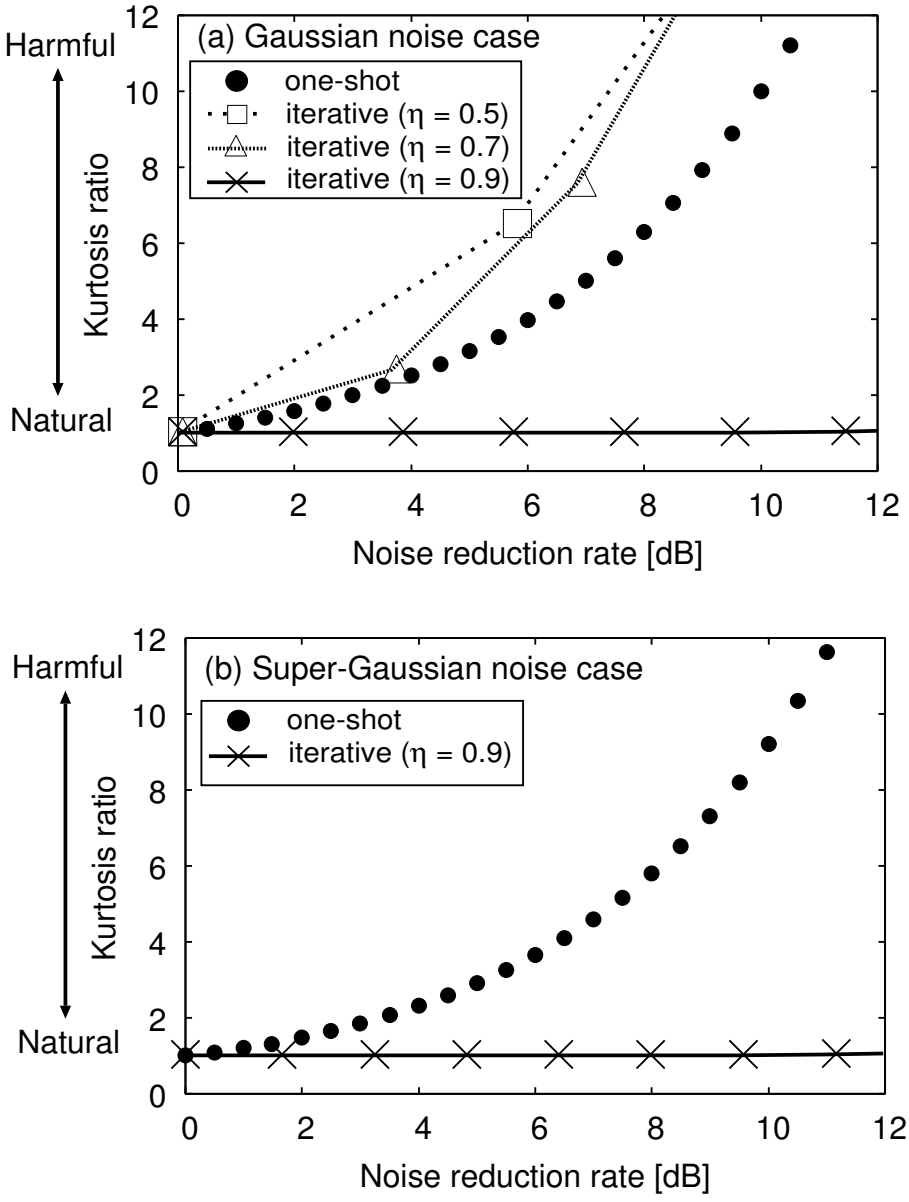

Fig. 6. Relation between NRR and kurtosis ratio obtained from theoretical analysis for (a) Gaussian noise case ($\alpha_0 = 1$) and (b) super-Gaussian noise case ($\alpha_0 = 0.2$).

4. Theory of Musical-Noise-Free Speech Enhancement

4.1 Introduction

In this section, I propose a new theory on musical-noise-free speech enhancement based on iterative SS. In the previous section, I analyzed the amounts of noise reduction performance and musical noise generation through iterative SS using higher-order statistics. It is of great interest now to know when the above-mentioned equilibrium behavior of the kurtosis ratio arises, i.e, which parameter settings give the output signal with the highest quality in iterative SS. However, the specific parameter settings was heuristically discovered. Therefore, in this section, I theoretically derive a closed-form solution of the internal parameters to satisfy the musical-noise-free condition.

I first describe an overview of the theory of the musical-noise-free condition (see Sect. 4.2). Next, I mathematically derive more general solution on the musical-noise-free condition (see Sect. 4.3). In Sect. 4.4, I show the example of the internal parameter settings that satisfy the musical-noise-free condition in iterative SS. In Sect. 4.5, I show the procedure of the musical-noise-free iterative SS. In Sect. 4.6, I conducted objective evaluation to confirm the validity of the musical-noise-free theory. Finally, I conducted objective and subjective evaluation experiments to compare the sound quality of the proposed method with those commonly used noise reduction methods (see Sect. 4.7).

4.2 Overview of musical-noise-free theory

As indicated by (28), iterative SS theory has an interesting *domino-toppling* phenomenon as follows. Given a specific parameter setting, if I am fortunate enough to obtain the same kurtosis as that of the input noise, i.e., $\mathrm{kurt}(\alpha_0, 0, 0)$, after the 1st iteration, i.e.,

$$\mathrm{kurt}(\alpha_0, \beta, \eta) = \mathrm{kurt}(\alpha_0, 0, 0) = \frac{(\alpha_0 + 3)(\alpha_0 + 2)}{(\alpha_0 + 1)\alpha_0}, \tag{34}$$

then from (26) I have $\alpha_1 = \alpha_0$. Obviously, this leads to the relation

$$\mathrm{kurt}(\alpha_1, \beta, \eta) = \mathrm{kurt}(\alpha_0, \beta, \eta) = \mathrm{kurt}(\alpha_0, 0, 0), \tag{35}$$

proving that the kurtosis in the 2nd iteration is also identical. The inductive result is that the kurtosis ratio never changes even at a large number of (ideally "infinite") iterations.

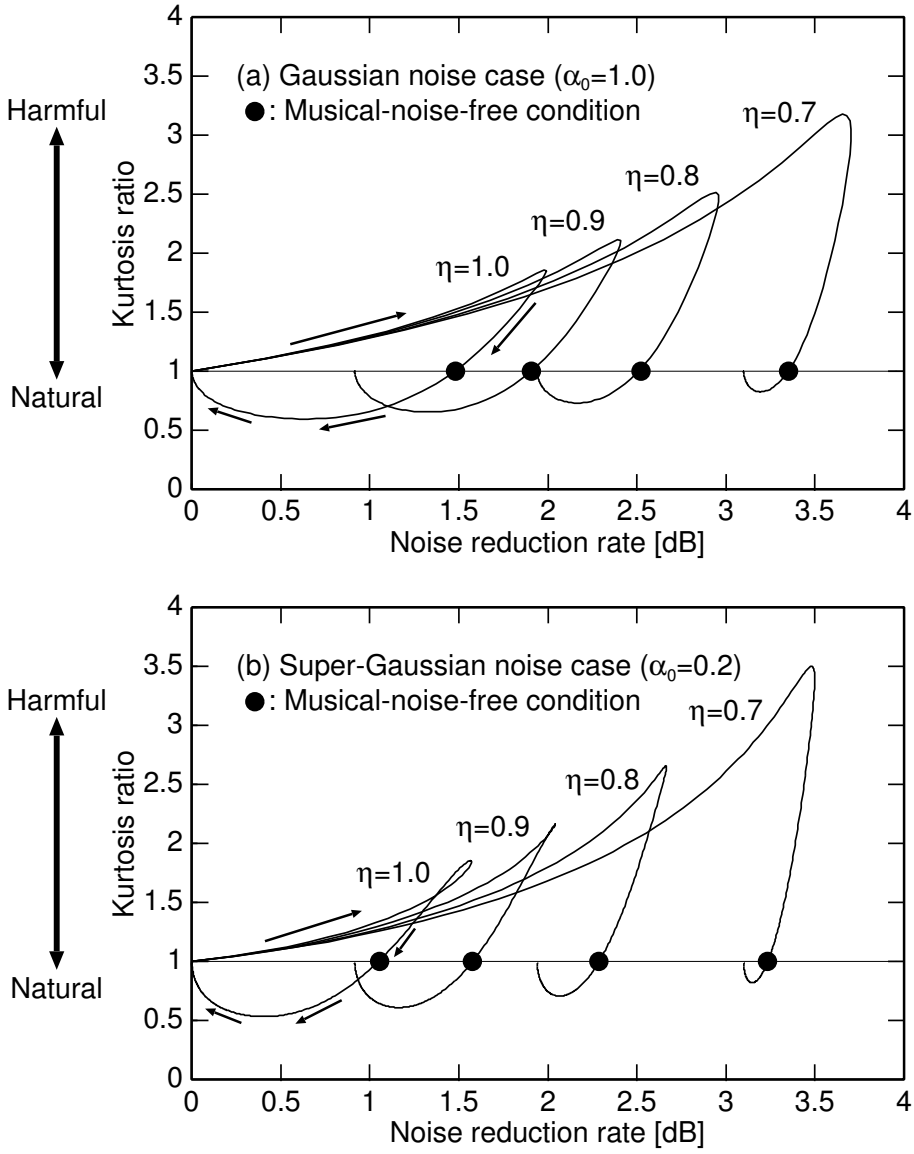

Fig. 7. Relation between NRR and kurtosis ratio from theoretical analysis with increasing β for (a) Gaussian noise case ($\alpha_0 = 1$) and (b) super-Gaussian noise case ($\alpha_0 = 0.2$).

In this situation, sufficient noise reduction can be gained if the NRR improvement in each iteration is even small but positive. In this study, since first-, second-, or fourth-order statistics affects our sense of hearing [20, 21], I ignore the effect of a variation

of fifth- or more higher-order statistics. This corresponds to musical-noise-free noise reduction. Here, since the kurtosis (28) and the NRR (31) are not equations in terms of the scale parameter θ, we do not need to estimate the scale parameter *theta* in each iteration.

In summary, I can formulate a new theory on musical-noise-free conditions as follows.

(I) Fixed-point kurtosis condition: The kurtosis should be equal before and after SS 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.

In the previous section, I have discovered the limited number of examples in which the musical-noise-free conditions can hold, as shown in Fig. 6 (see the case of $\eta = 0.9$). Except for the parameter settings used in Fig. 6, I can also find the other cases that satisfy the musical-noise-free conditions. For example, Fig. 7 shows *hysteresis loops* in the relation between the NRR and kurtosis ratio of non-iterative SS (calculated by (29) and (31)) with various parameter settings. Note that each hysteresis loop corresponding to each η has its own intersection at the point of the kurtosis ratio of unity, showing the existence of multiple cases for realizing the fixed point in the kurtosis. In the following subsections, I mathematically derive more general solutions on the musical-noise-free conditions.

4.3 Musical-noise-free condition

4.3.1 Fixed-point kurtosis condition

Although the parameters to be optimized are η and β, I hereafter derive the optimal η given a fixed β for ease of closed-form analysis. First, I change (19) for

$$\mathrm{kurt}(\alpha_0, \beta, \eta) = \frac{\mathcal{S}(\alpha_0, \beta, 4) + \eta^8 \mathcal{F}(\alpha_0, \beta, 4)}{\left(\mathcal{S}(\alpha_0, \beta, 2) + \eta^4 \mathcal{F}(\alpha_0, \beta, 2)\right)^2}, \tag{36}$$

where

$$\mathcal{S}(\alpha_0,\beta,m)=\sum_{l=0}^{m}(-\beta\alpha_0)^l\frac{\Gamma(m+1)\Gamma(\alpha_0+m-l,\beta\alpha_0)}{\Gamma(\alpha_0)\Gamma(l+1)\Gamma(m-l+1)}, \tag{37}$$

$$\mathcal{F}(\alpha_0,\beta,m)=\frac{\gamma(\alpha_0+m,\beta\alpha_0)}{\Gamma(\alpha_0)}. \tag{38}$$

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

$$\frac{\mathcal{S}(\alpha_0,\beta,4)+\eta^8\mathcal{F}(\alpha_0,\beta,4)}{\left(\mathcal{S}(\alpha_0,\beta,2)+\eta^4\mathcal{F}(\alpha_0,\beta,2)\right)^2}=\frac{(\alpha_0+3)(\alpha_0+2)}{(\alpha_0+1)\alpha_0}. \tag{39}$$

Let $\mathcal{H}=\eta^4$, and (39) yields the following quadratic equation in $\mathcal{H}$.

$$\begin{aligned}&\left(\mathcal{F}(\alpha_0,\beta,4)(\alpha_0+1)\alpha_0-\mathcal{F}^2(\alpha_0,\beta,2)(\alpha_0+3)(\alpha_0+2)\right)\mathcal{H}^2\\&-2\mathcal{S}(\alpha_0,\beta,2)\mathcal{F}(\alpha_0,\beta,2)(\alpha_0+3)(\alpha_0+2)\,\mathcal{H}\\&+\mathcal{S}(\alpha_0,\beta,4)(\alpha_0+1)\alpha_0-\mathcal{S}^2(\alpha_0,\beta,2)(\alpha_0+3)(\alpha_0+2)=0.\end{aligned} \tag{40}$$

Thus, I can derive a closed-form estimate of $\mathcal{H}$ from the given oversubtraction parameter as

$$\begin{aligned}\mathcal{H}=&\{\mathcal{F}(\alpha_0,\beta,4)(\alpha_0+1)\alpha_0-\mathcal{F}^2(\alpha_0,\beta,2)(\alpha_0+3)(\alpha_0+2)\}^{-1}\\&\Bigg[\mathcal{S}(\alpha_0,\beta,2)\mathcal{F}(\alpha_0,\beta,2)(\alpha_0+3)(\alpha_0+2)\\&\pm\Big[\{\mathcal{S}(\alpha_0,\beta,2)\mathcal{F}(\alpha_0,\beta,2)(\alpha_0+3)(\alpha_0+2)\}^2\\&-\left\{\mathcal{F}(\alpha_0,\beta,4)(\alpha_0+1)\alpha_0-\mathcal{F}^2(\alpha_0,\beta,2)(\alpha_0+3)(\alpha_0+2)\right\}\\&\left\{\mathcal{S}(\alpha_0,\beta,4)(\alpha_0+1)\alpha_0-\mathcal{S}^2(\alpha_0,\beta,2)(\alpha_0+3)(\alpha_0+2)\right\}\Big]^{\frac{1}{2}}\Bigg].\end{aligned} \tag{41}$$

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

4.3.2 NRR growth condition

In this subsection, I reveal the range of the flooring parameter η that increases the NRR. From (31), the NRR growth condition is expressed as

$$\mathrm{NRR}=10\log_{10}\frac{\alpha_0}{\mathcal{S}(\alpha_0,\beta,1)+\eta^2\mathcal{F}(\alpha_0,\beta,1)}>0. \tag{42}$$

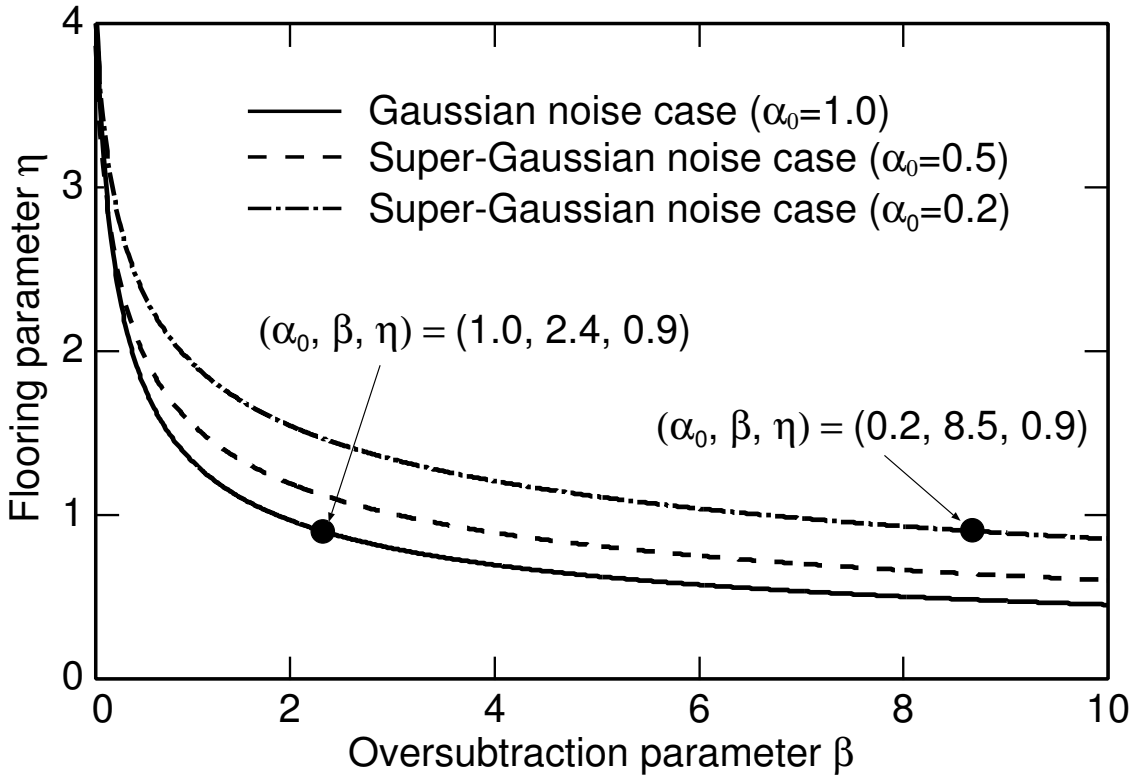

Fig. 8. Example of oversubtraction parameter β and flooring parameter η to satisfy musical-noise-free condition.

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

$$0 < \eta < \sqrt{\frac{\alpha_0 - \mathcal{S}(\alpha_0, \beta, 1)}{\mathcal{F}(\alpha_0, \beta, 1)}}. \tag{43}$$

In summary, I can choose the parameters simultaneously satisfying the fixed kurtosis point condition and NRR growth condition using (41) and (43).

4.4 Parameter example for musical-noise-free condition

According to the previous analysis, I can calculate combinations of the oversubtraction parameter β and the flooring parameter η that satisfy the musical-noise-free condition under the three types of shape parameter α_0, namely, 0.2, 0.5, and 1.0. Figure 8

shows examples of traces. It is worth mentioning that the specific setting $(\alpha, \beta, \eta) = (1.0, 2.0, 0.97)$ appears in Fig. 8, which was heuristically discovered in Sect. 3.6, but our theory can provide more wide-ranging solutions. Also, I show the typical example of the optimal parameter settings in Appendix B.

4.5 Procedure of musical-noise-free iterative SS

In this subsection, I conduct the procedure of the musical-noise-free iterative SS in the following manner.

(I) First, set the obersubtraction parameter β to arbitrary value.

(II) Next, estimate the shape parameter α_0 in the speech absence periods using the maximum likelihood estimation method as follows

$$\hat{\alpha} = \frac{3 - \gamma + \sqrt{(\gamma - 3)^2 + 24\gamma}}{12\gamma}, \tag{44}$$

where $\gamma = \log(\widehat{\mathrm{E}[|N|^2]}) - \mathrm{E}[\widehat{\log |N|^2}]$ (see Refs. [45, 46]).

(III) Next, calulate the flooring parameters η using (41) and choose is satisfying (43).

(IV) Finally, perform iterative SS with the oversubtraction parameter β and the flooring parameter η calculated by step (III).

4.6 Evaluation experiment for iterative SS with optimal parameter settings

4.6.1 Experimental conditions

I conducted objective evaluation to confirm the validity of the theoretical analysis described in the previous section. Noisy observation signals were generated by adding noise signals to target speech signals with an SNR of 0 dB. I conducted our experiments on white Gaussian noise and babble noise. The target speech signals were the utterances of two male and two female speakers in Japanese (4 sentences) from the JNAS database [47]. The noise signals were white Gaussian noise and babble noise, where the babble noise was recorded human speech emitted from 36 loudspeakers (this

simulates a crowded place). The estimated shape parameter of the power spectra of the white Gaussian noise was 0.97 and that of the babble noise was 0.21. The length of each signal was 7 s, and each signal was sampled at 16 kHz. The FFT size was 1024 and the frame shift length was 256. In these experiments, I calculated the noise PSD $\mathrm{E}[|N|^2]$ in the first 1 s frames, where I assume speech absence in this period in conventional non-iterative SS and iterative SS. In conventional non-iterative SS, the oversubtraction parameter β is manually adjusted (the flooring parameter η is fixed to 0.1) so that the NRR is varied as 0, 0.5, 1.0, ..., 12.0 dB. In iterative SS, the parameter settings of β and η are 2.4 and 0.9 for white Gaussian noise case and 8.5 and 0.9 for babble noise case. Those parameter settings satisfy the musical-noise-free condition.

4.6.2 Comparison between theoretical analysis and experiments

I conducted an objective evaluation experiment and evaluated the sound quality of processed signals on the basis of the kurtosis ratio and cepstral distortion [48]. Here, I calculated the kurtosis ratio from the noise-only period and the cepstral distortion from the target speech components. The cepstral distortion is a measure of the degree of distortion via the cepstrum domain. The cepstral distortion indicates the amount of distortion among two signals, which is defined as

$$\mathrm{CD}\ [\mathrm{dB}] \equiv \frac{20}{T \log 10} \sum_{\tau=1}^{T} \sqrt{\sum_{\rho=1}^{B} 2(C_{\mathrm{out}}(\rho,\tau) - C_{\mathrm{ref}}(\rho,\tau))^2}, \tag{45}$$

where T is the frame length, $C_{\mathrm{out}}(\rho,\tau)$ is the ρth cepstral coefficient of the output signal in frame τ, and $C_{\mathrm{ref}}(\rho,\tau)$ is the ρth cepstrum coefficient of the original speech signal. B is the number of dimensions of the cepstrum used in the evaluation; I set $B = 22$. A small value of cepstral distortion indicates that the sound quality of the target speech part is high.

The results of the experiment are depicted in Figs. 9 and 10, where the kurtosis ratio, cepstral distortion, and NRR were calculated from the observed and processed signals. All of the scores are the averages in terms of four target speakers. Figure 9(a) shows that the kurtosis ratio decreases as the flooring parameter increases, and from Figs. 9(a) and (b) I can confirm the efficacy of iterative SS for white Gaussian noise and babble noise if I use weak processing in each iteration. This tendency is in good agreement with the results of the theoretical analysis in Sect. 3.6. The discrepancy between

the kurtosis ratio obtained from the real processed data and the theoretical estimate is thought to be mainly due to the gamma-distribution approximation introduced in our analysis. For reference, the histogram of the noise power spectra in each iteration of iterative SS is shown in Appendix C and the result of the signal-to-distortion ratio (SDR) and the source-to-interference ratio (SIR) is shown in Appendix D.

In addition, from Figs. 10(a) and (b), I can see that cepstral distortion in the case of iterative SS is smaller than that for conventional non-iterative SS. This indicates that there are no side effects in the utilization of the iterative method because I confirmed the decrease in both kurtosis ratio and cepstral distortion in Figs. 9 and 10. Consequently, in all cases, I can achieve high sound quality upon setting appropriate parameters in iterative SS.

4.7 Comparison between proposed method and conventional noise reduction methods

4.7.1 Experimental Conditions

I conducted objective and subjective evaluation experiments to compare the sound quality of the proposed method with those of commonly used noise reduction methods. Noisy observation signals were generated by adding noise signals to target speech signals with SNRs of -5, 0, 5, and 10 dB. The target speech signals were the utterances of two male and two female speakers in Japanese (4 sentences) from the JNAS database [47]. The noise signals were white Gaussian noise, babble noise, railway station noise, museum noise and factory noise. The estimated shape parameter of the power spectra of the railway station noise was 0.33, that of the museum noise was 0.21, and that of the factory noise was 0.21. The length of each signal was 7 s, and each signal was sampled at 16 kHz. The FFT size was 1024 and the frame shift length was 256. I calculated the noise PSD $E[|N|^2]$ by using the following two methods. (A) The noise PSD is calculated in the first 1 s frames, where I assume speech absence in this period (*VAD-based*). (B) The noise PSD is dynamically estimated by using minimum statistics method [3] (*minimum-statistics-based*).

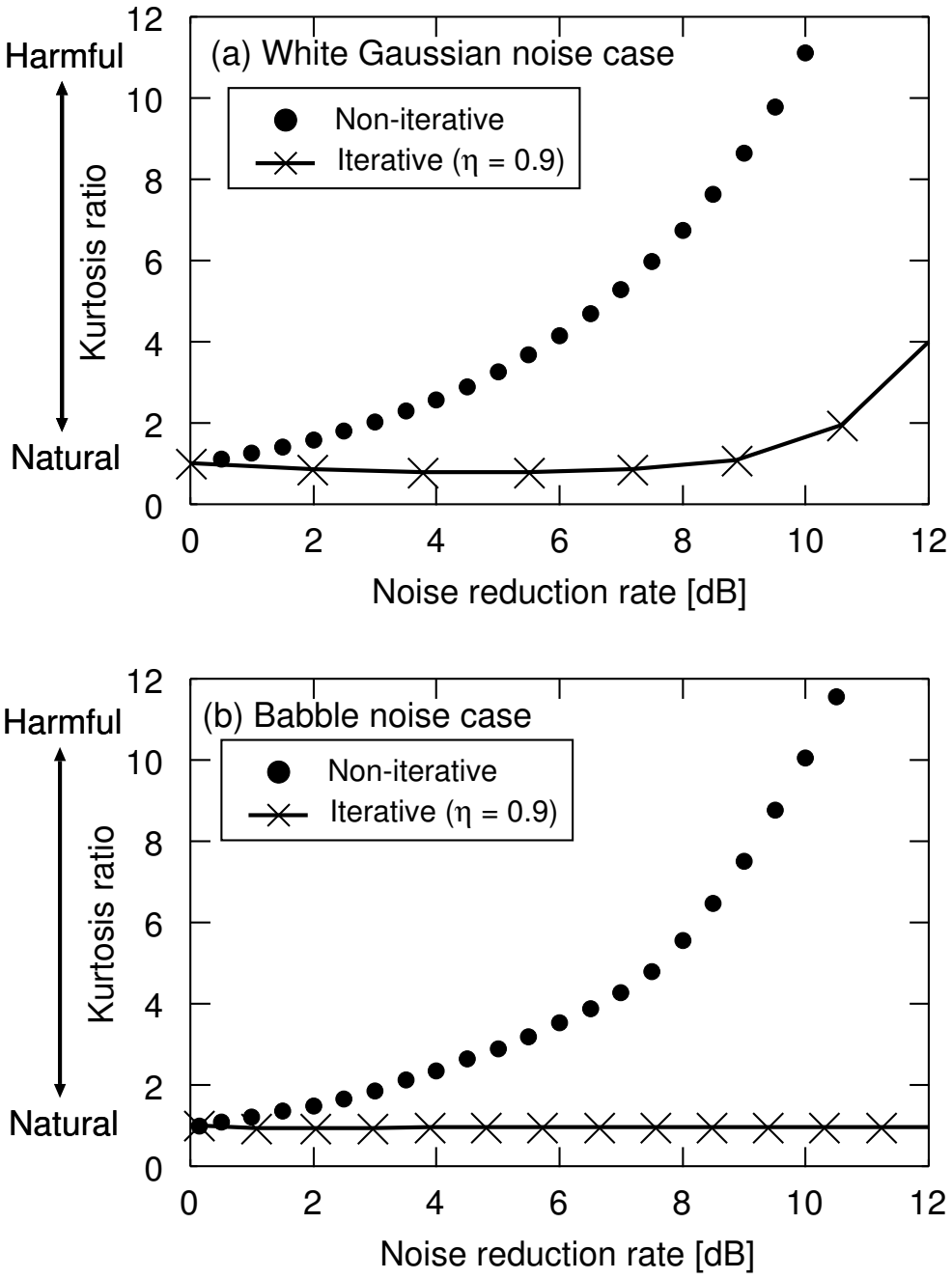

Fig. 9. Relation between NRR and kurtosis ratio obtained from experiment with real noisy speech data for (a) white Gaussian noise case ($\alpha_0 = 0.97$) and (b) babble noise case ($\alpha_0 = 0.21$).

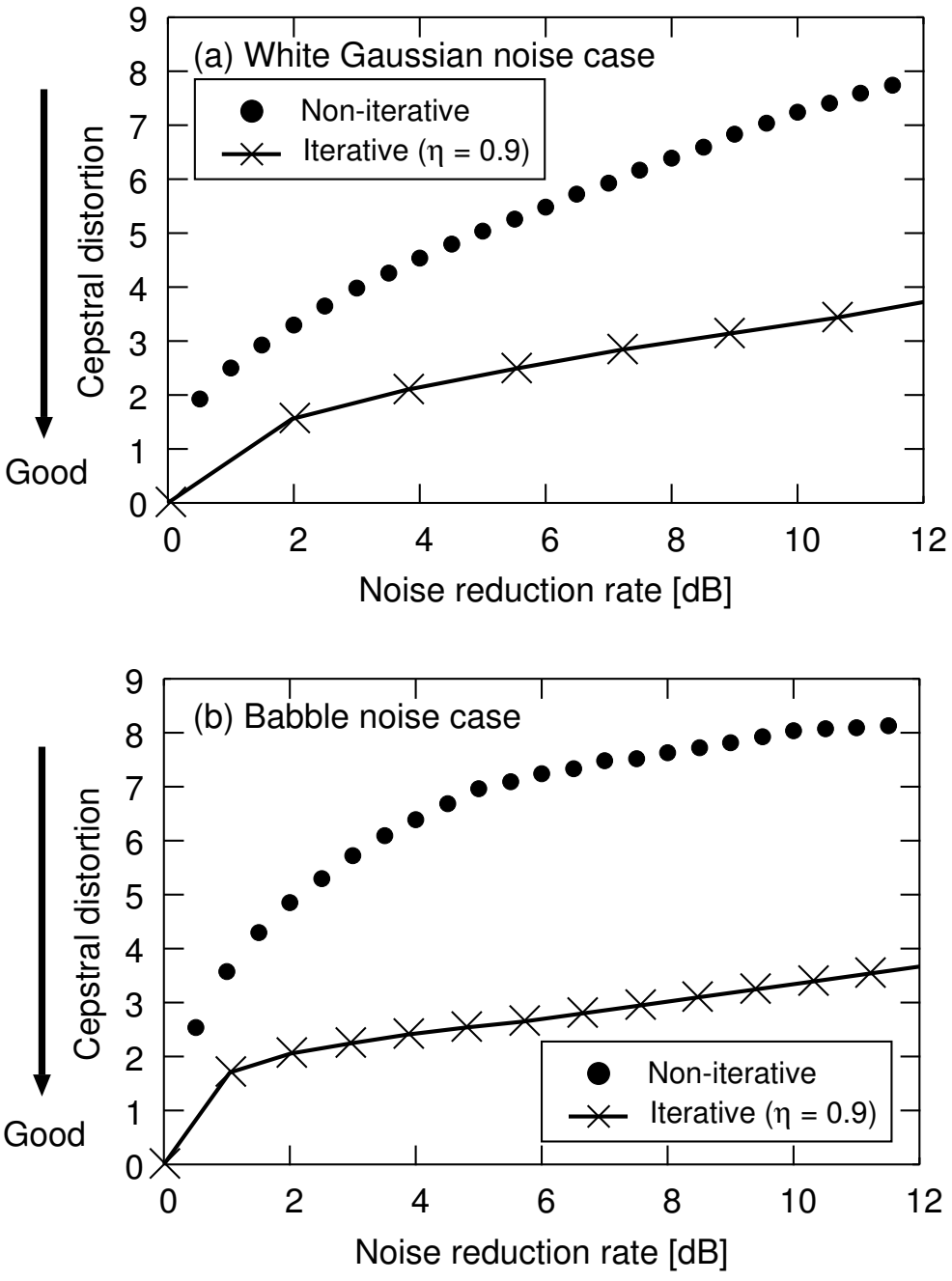

Fig. 10. Relation between NRR and cepstral distortion obtained from experiment with real noisy speech data for (a) white Gaussian noise case ($\alpha_0 = 0.97$) and (b) babble noise case ($\alpha_0 = 0.21$).

4.7.2 Objective Evaluation

I conducted an objective evaluation under the same NRR condition. Figures 11–14 show the kurtosis ratio and cepstral distortion obtained from the experiment with real noisy speech data for (a) white Gaussian noise, (b) babble noise, (c) railway station noise, (d) museum noise and (e) factory noise, where I evaluate 10-dB-NRR (i.e., output SNRs = 5, 10, 15, and 20 dB) signals processed by three conventional methods, namely, non-iterative SS, Wiener filtering [4], and the MMSE short-time spectral amplitude (STSA) estimator [7], and our proposed method, iterative SS with the optimal parameter settings (I apply two types of noise PSD estimators to each method). Here, I use the decision-directed approach for a priori SNR estimation in both of Wiener filtering and the MMSE STSA estimator. From Figs. 11–14, I can confirm that VAD-based and minimum-statistics-based iterative SS methods outperform other conventional methods in both the kurtosis ratio and cepstral distortion, except for the comparison with the MMSE STSA estimator in terms of the cepstral distortion for high input SNR cases. In particular, the kurtosis ratios of the proposed methods are mostly close to 1.0. Since Wiener filtering and the MMSE STSA estimator are often referred to as less musical noise method, this result greatly emphasizes the iterative SS's advantage, i.e., musical-noise-free property I have theoretically predicate.

4.7.3 Subjective Evaluation

First, I conducted subjective evaluation by comparing iterative SS with other commonly used noise reduction methods. I presented a pair of 10-dB-NRR signals processed by minimum-statistics-based non-iterative SS, minimum-statistics-based Wiener

filtering, minimum-statistics-based MMSE STSA estimator, and minimum-statistics-based iterative SS with the optimal parameter settings 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, etc.

The result of the experiment is shown in Fig. 15 for (a) white Gaussian noise, (b) babble noise, (c) railway station noise, (d) museum noise and (e) factory noise. It was found that the output signal of minimum-statistics-based iterative SS with the optimal parameters is preferred to those of conventional methods. This result is also consistent with our theoretical analysis, thus confirming the validity of the proposed method of theoretical analysis.

From the previous subjective evaluation experiment, I reveal that our proposed method outperforms other conventional methods in total sound quality. However, I did not compare the processed signal using the proposed method with the unprocessed signal that has no musical noise, no speech distortion, but *no noise reduction*. Some people may guess that the unprocessed signal is more preferable from the viewpoint of the amounts of musical noise and speech distortion except for the amount of noise. Therefore, secondly, I conducted another subjective evaluation experiment for direct comparison between the unprocessed and processed signals. I presented a pair of unprocessed signal and 10-dB-NRR processed signal by minimum-statistics-based iterative SS with the optimal parameter settings in random order to 10 examinees, who selected which signal they preferred from the viewpoint of total sound quality.

The result of the experiment is shown in Fig. 16 for (a) white Gaussian noise, (b) babble noise, (c) railway station noise, (d) museum noise and (e) factory noise. It was found that the output signal of minimum-statistics-based iterative SS with the optimal parameters is preferred to the unprocessed signal. This result indicates that noise reduction performed by the proposed method is essentially valid in terms of human hearing.

4.8 Conclusion

In this section, I mathematically derived the internal parameter settings to satisfy the musical-noise-free condition. It was clarified that the optimal parameters satisfying

the fixed kurtosis point condition and NRR growth condition can generate almost no musical noise even with high noise reduction. This desirable property of iterative SS was well supported by the comparative experiment between iterative SS and commonly used noise reduction methods, e.g., conventional non-iterative SS, Wiener filtering, and the MMSE STSA estimator. In summary, proposed theory mathematically proves that iterative SS with the optimal parameters is advantageous for achieving high-quality noise reduction, which has only been experimentally shown in previous studies.

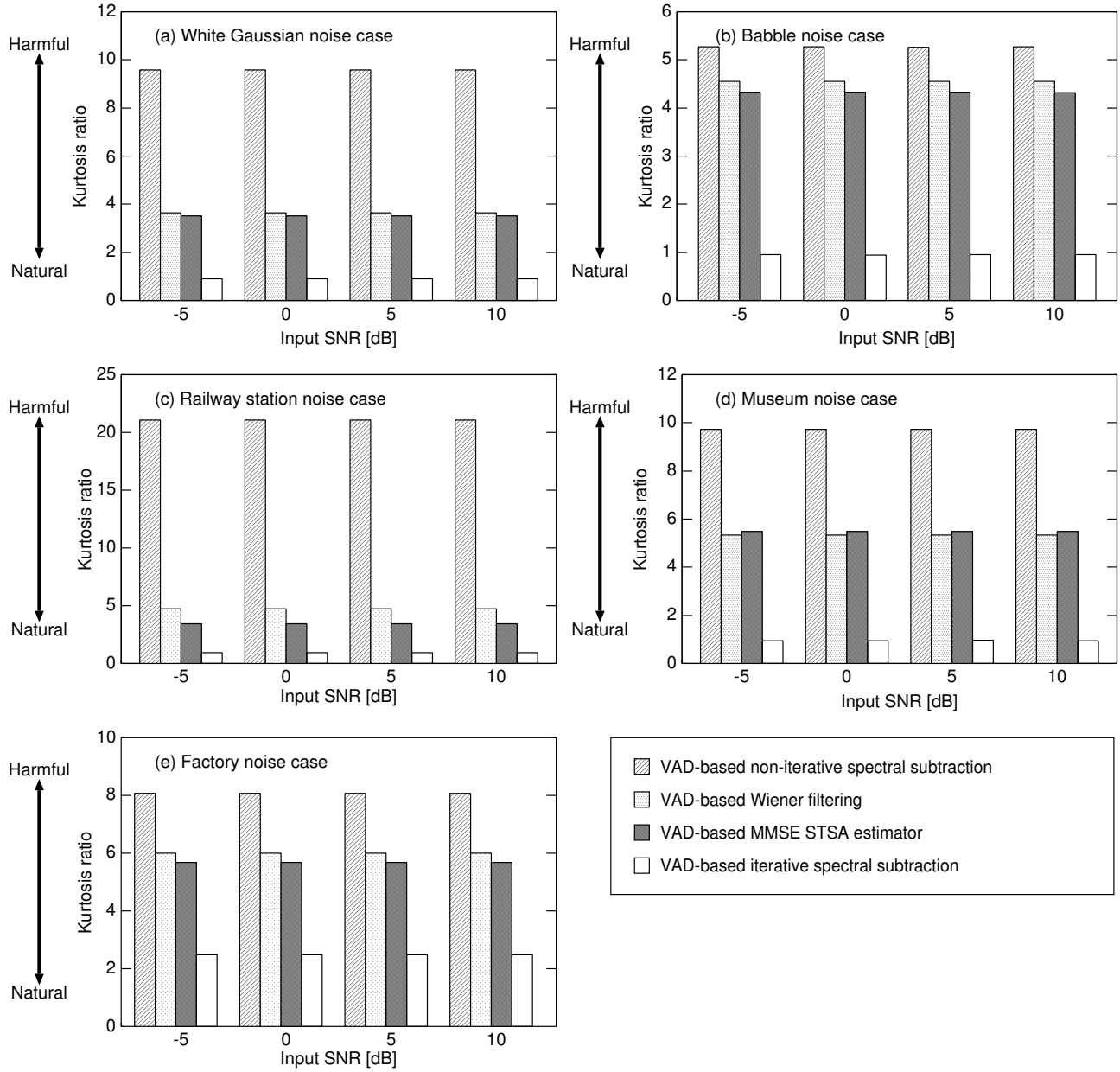

Fig. 11. Kurtosis ratio obtained from experiment with real noisy speech data for (a) white Gaussian noise, (b) babble noise, (c) railway station noise, (d) museum noise and (e) factory noise under 10-dB-NRR condition, where I compare VAD-based non-iterative SS, VAD-based Wiener filtering, VAD-based MMSE STSA estimator, and VAD-based iterative SS with the optimal parameter settings.

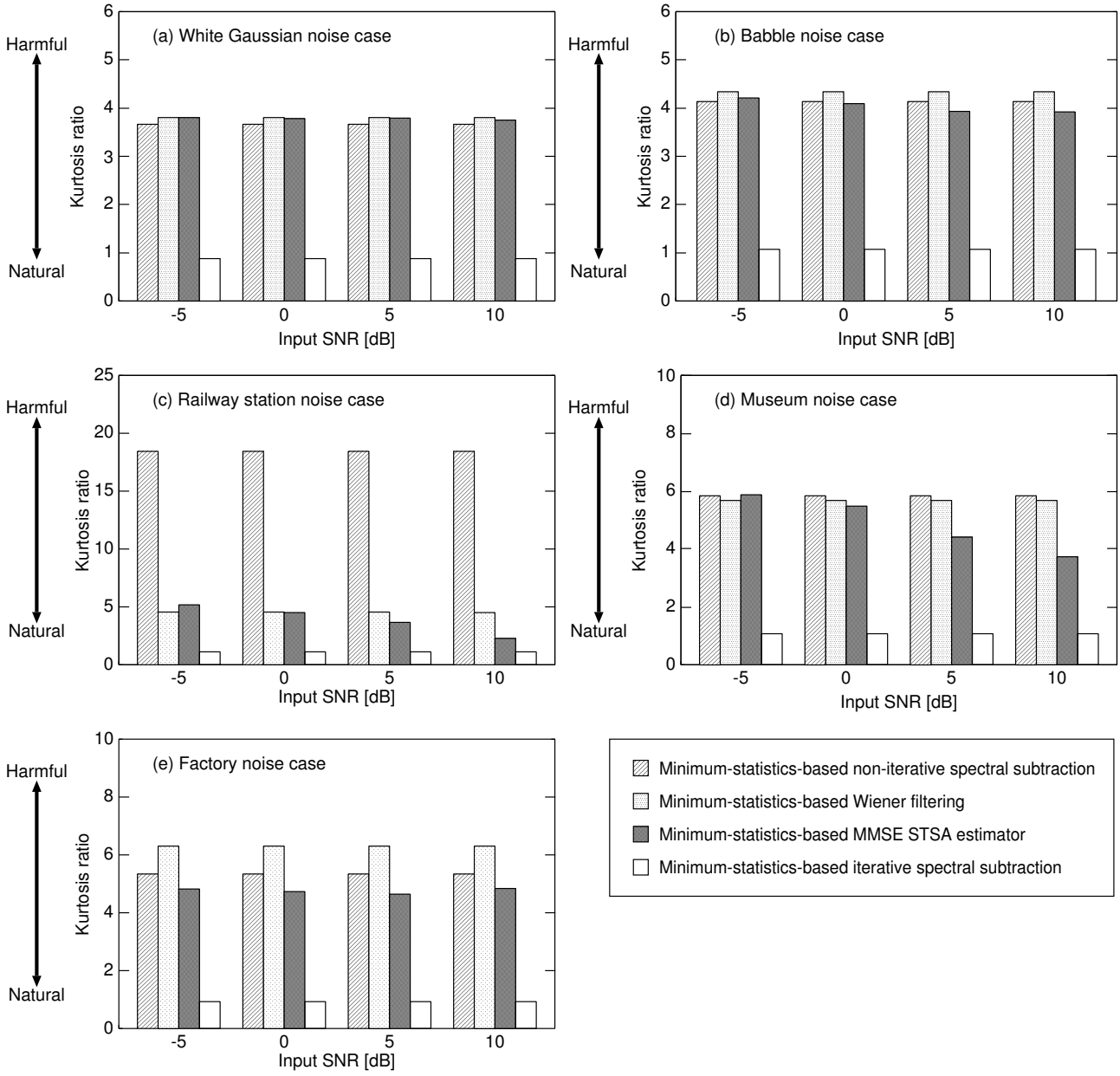

Fig. 12. Kurtosis ratio obtained from experiment with real noisy speech data for (a) white Gaussian noise, (b) babble noise, (c) railway station noise, (d) museum noise and (e) factory noise under 10-dB-NRR condition, where I compare minimum-statistics-based non-iterative SS, minimum-statistics-based Wiener filtering, minimum-statistics-based MMSE STSA estimator, and minimum-statistics-based iterative SS with the optimal parameter settings.

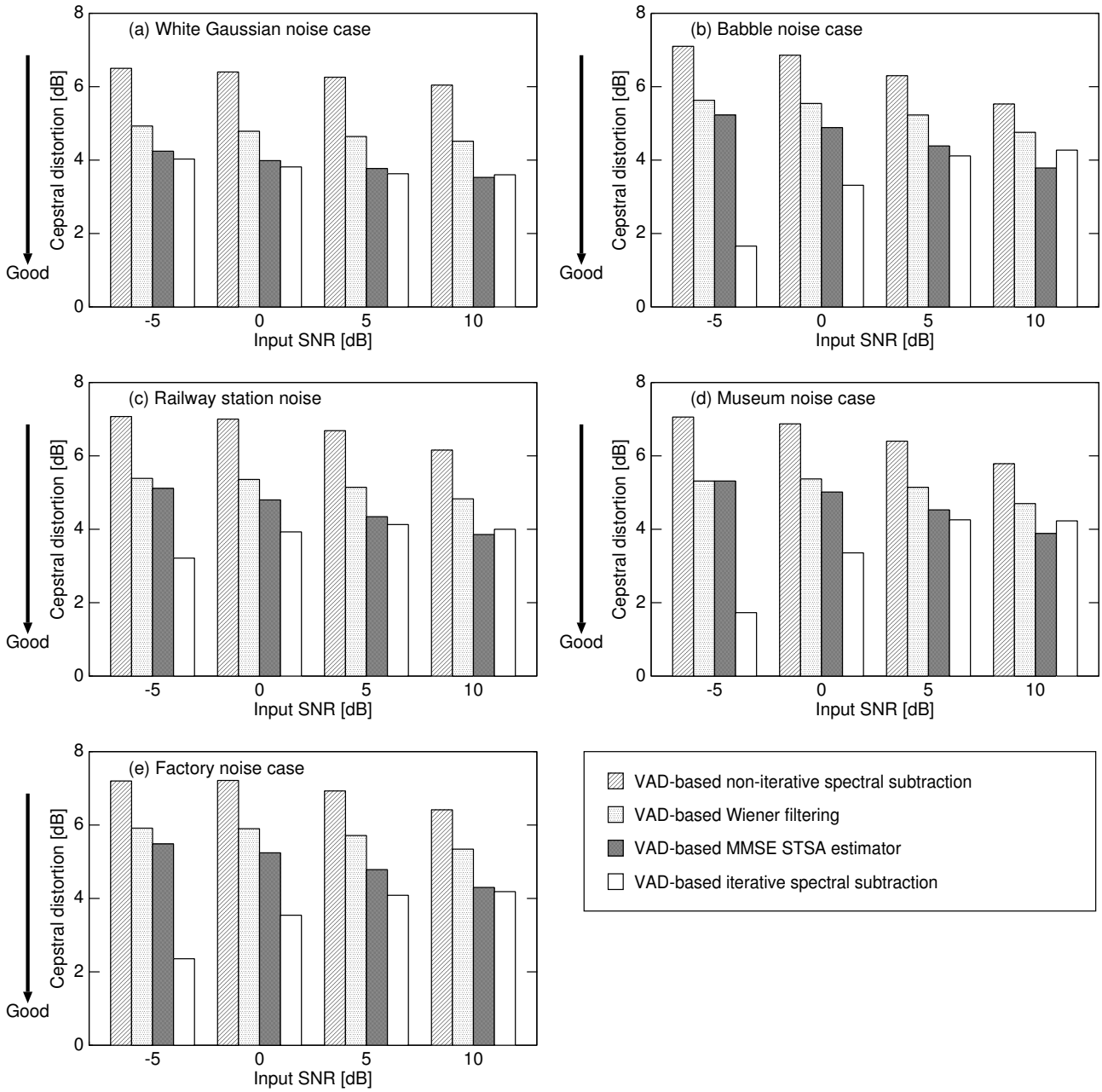

Fig. 13. Cepstral distortion obtained from experiment with real noisy speech data for (a) white Gaussian noise, (b) babble noise, (c) railway station noise, (d) museum noise and (e) factory noise under 10-dB-NRR condition, where I compare VAD-based non-iterative SS, VAD-based Wiener filtering, VAD-based MMSE STSA estimator, and VAD-based iterative SS with the optimal parameter settings.

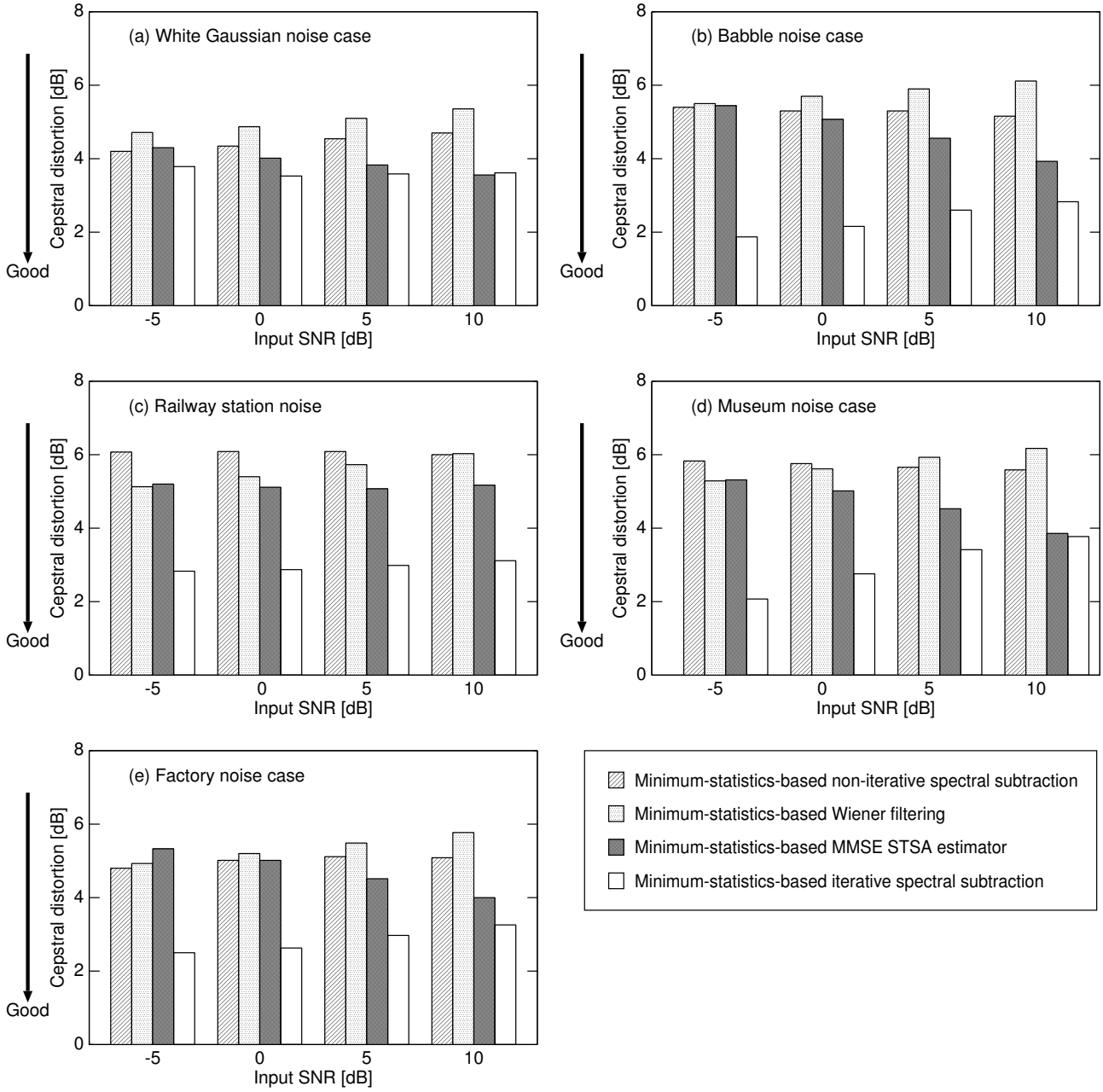

Fig. 14. Cepstral distortion obtained from experiment with real noisy speech data for (a) white Gaussian noise, (b) babble noise, (c) railway station noise, (d) museum noise and (e) factory noise under 10-dB-NRR condition, where I compare minimum-statistics-based non-iterative SS, minimum-statistics-based Wiener filtering, minimum-statistics-based MMSE STSA estimator, and minimum-statistics-based iterative SS with the optimal parameter settings.

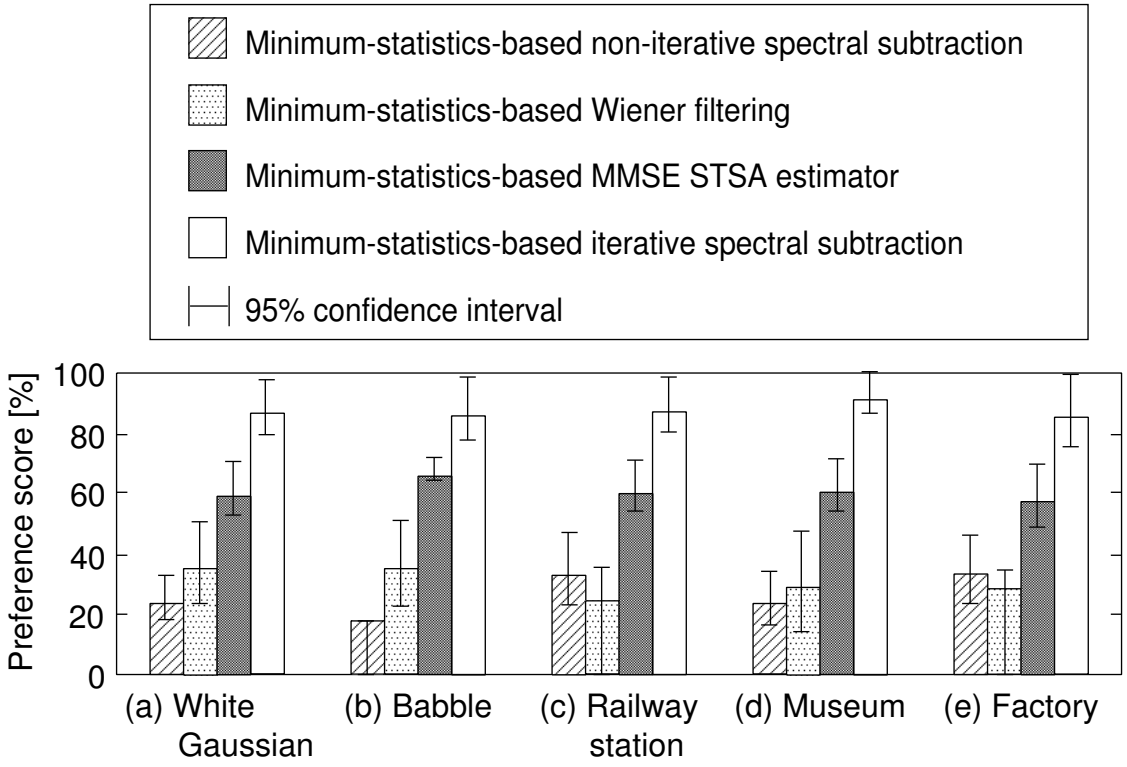

Fig. 15. Subjective evaluation results for (a) white Gaussian noise, (b) babble noise, (c) railway station noise, (d) museum noise and (e) factory noise. I presented a pair of 10-dB-NRR signals processed by minimum-statistics-based non-iterative SS, minimum-statistics-based Wiener filtering, minimum-statistics-based MMSE STSA estimator, and minimum-statistics-based iterative SS with the optimal parameter settings in random order to 10 examinees, who selected which signal they preferred from the viewpoint of total sound quality.

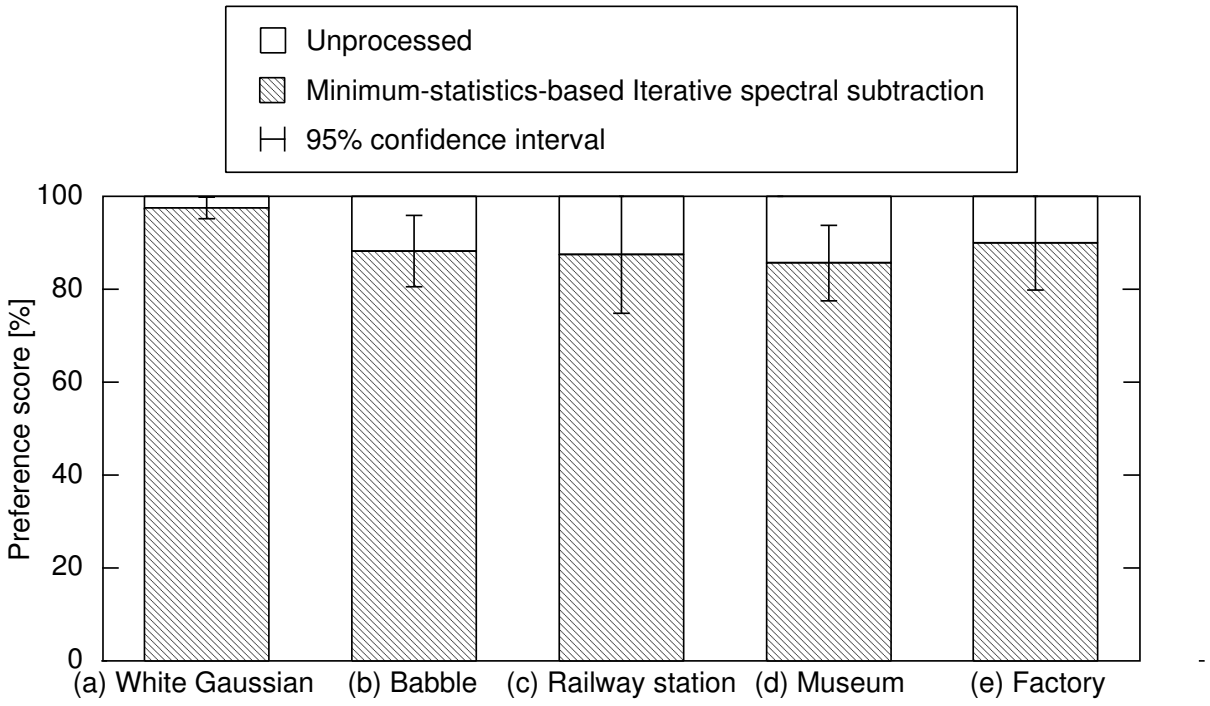

Fig. 16. Subjective evaluation results for (a) white Gaussian noise, (b) babble noise, (c) railway station noise, (d) museum noise and (e) factory noise. I presented a pair of unprocessed noisy speech signal and 10-dB-NRR signals processed by minimum-statistics-based iterative SS with the optimal parameter settings in random order to 10 examinees, who selected which signal they preferred from the viewpoint of total sound quality.

5. Extension to Microphone Array Signal Processing

5.1 Introduction

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, e.g., a nonstationary noise field, it is necessary to dynamically estimate the noise PSD.

To solve this problem, Takahashi previously proposed blind spatial subtraction array (BSSA) [36], which involves accurate noise estimation by ICA followed by a speech extraction procedure based on SS (see Fig. 17). 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.

In this section, I propose a new multi-iterative blind signal extraction method integrating iterative blind noise estimation by ICA and iterative noise reduction by SS, where multiple iterative SS is applied to each channel while maintaining the muli-channel property reused for ICA. Hereafter, I refer to this proposed method as *iterative BSSA*.

I first describe an overview of iterative BSSA (see Sect. 5.2). Next, I discuss the accuracy of the estimated noise signal in each iteration of iterative BSSA (see Sect. 5.3). In Sect. 5.4, I introduce the improvement scheme for poor noise estimation in iterative BSSA. Finally, in Sect. 5.5, I show the results of objective and subjective evaluation experiment.

5.2 Iterative blind spatial subtraction array

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, we propose a new BSSA structure that performs multiple independent SS in each channel before DS; we call this structure

channelwise SS [37, 38, 39]. This structure can reduce the amount of musical noise generation compared with conventional structure (see Appendix E). Using this structure, we can equalize the number of channels of the observed signal to that of the signals after channelwise SS. Therefore, we can iteratively apply noise estimation by ICA and speech extraction by SS (see Fig. 18). Also, the advantage of the proposed 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. 5.3.

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

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

$$\boldsymbol{x}^{[0]}(f,\tau) = \boldsymbol{h}(f)s(f,\tau) + \boldsymbol{n}(f,\tau), \tag{46}$$

where $\boldsymbol{h}(f) = [h_1(f), h_2(f) \ldots, h_K(f)]^{\mathrm{T}}$ 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 $\boldsymbol{n}(f,\tau)$ is a column vector of the additive noise.

(II) Next, we perform signal separation using ICA as [40]

$$\boldsymbol{o}^{[i]}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f)\boldsymbol{x}^{[i]}(f,\tau), \tag{47}$$

$$\boldsymbol{W}_{\mathrm{ICA}}^{[i][p+1]}(f) = \mu[\boldsymbol{I} - \langle \boldsymbol{\varphi}(\boldsymbol{o}^{[i]}(f,\tau))(\boldsymbol{o}^{[i]}(f,\tau))^{\mathrm{H}}\rangle_\tau] \cdot \boldsymbol{W}_{\mathrm{ICA}}^{[i][p]}(f) + \boldsymbol{W}_{\mathrm{ICA}}^{[i][p]}(f), \tag{48}$$

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

$$\boldsymbol{o}_{\mathrm{noise}}^{[i]}(f,\tau) = [o_1^{[i]}(f,\tau), \ldots, o_{U-1}^{[i]}, 0, o_{U+1}^{[i]}(f,\tau), \ldots, o_K^{[i]}(f,\tau)]^{\mathrm{T}}, \tag{49}$$

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, $\boldsymbol{z}^{[i]}(f,\tau)$, as

$$\boldsymbol{z}^{[i]}(f,\tau) = \boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f)^{-1}\boldsymbol{o}_{\mathrm{noise}}^{[i]}(f,\tau). \tag{50}$$

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

$$x_k^{[i+1]}(f,\tau) = \begin{cases} \sqrt{|x_k^{[i]}(f,\tau)|^2 - \beta|z_k^{[i]}(f,\tau)|^2}\exp(j\arg(x_k^{[i]}(f,\tau))) \\ (\text{if} \quad |x_k^{[i]}(f,\tau)|^2 > \beta|z_k^{[i]}(f,\tau)|^2), \\ \eta x_k^{[i]}(f,\tau) \quad (\text{otherwise}), \end{cases} \tag{51}$$

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 $\boldsymbol{x}^{[i+1]}(f,\tau)$. 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 $\boldsymbol{x}^{[*]}(f,\tau)$, where $*$ is the number of iterations after which sufficient noise reduction performance is obtained. This procedure can be expressed by

$$y(f,\tau) = \boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{x}^{[*]}(f,\tau), \tag{52}$$

$$\boldsymbol{w}_{\mathrm{DS}}(f) = [w_1^{(\mathrm{DS})}(f),\ldots,w_K^{(\mathrm{DS})}(f)], \tag{53}$$

$$w_k^{(\mathrm{DS})}(f) = \frac{1}{K}\exp(-2j(f/N)f_{\mathrm{s}}d_k\sin\theta_U/c), \tag{54}$$

$$\theta_U = \sin^{-1}\frac{\arg\left(\frac{\left[\boldsymbol{W}_{\mathrm{ICA}}^{[*]}(f)^{-1}\right]_{kU}}{\left[\boldsymbol{W}_{\mathrm{ICA}}^{[*]}(f)^{-1}\right]_{k'U}}\right)}{2\pi f_{\mathrm{s}}c^{-1}(d_k - d_{k'})}, \tag{55}$$

where $y(f,\tau)$ is the final output signal of iterative BSSA, $\boldsymbol{w}_{\mathrm{DS}}$ is the filter coefficient vector of DS, N is the DFT size, f_{s} is the sampling frequency, d_k is the microphone position, c is the sound velocity, and θ_U is the estimated direction of arrival of the target speech. Moreover, $[\boldsymbol{A}]_{lj}$ represents the entry of $\boldsymbol{A}$ in the lth row and jth column.

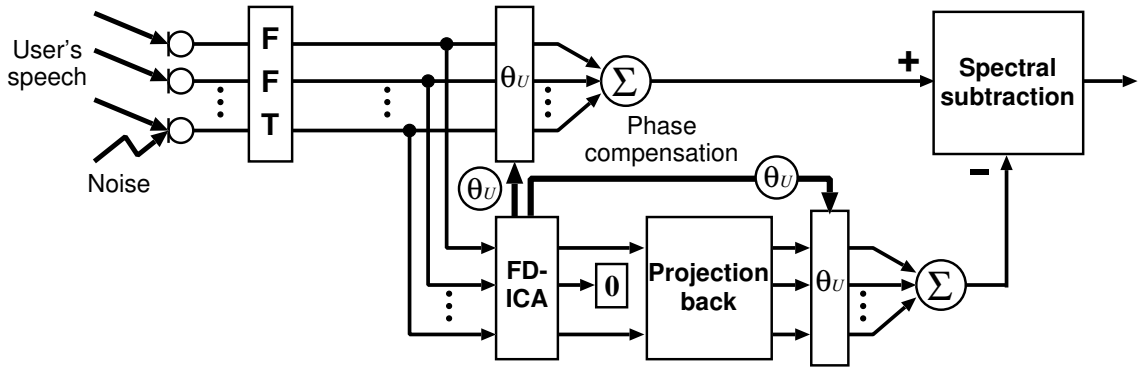

Fig. 17. Block diagram of conventional BSSA [36].

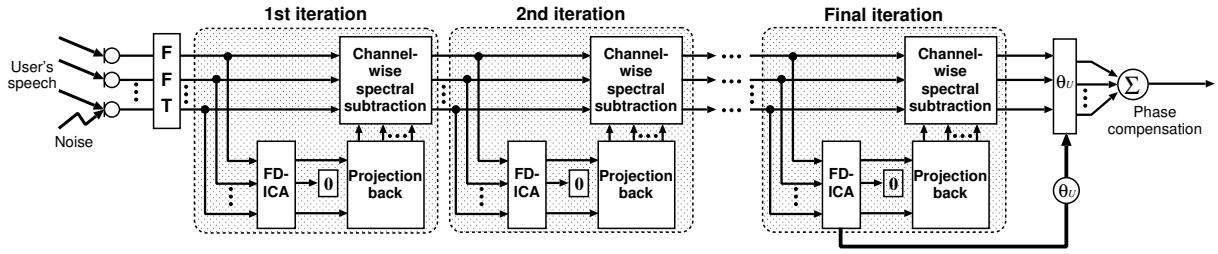

Fig. 18. Block diagram of proposed iterative BSSA.

5.3 Accuracy of wavefront estimated by ICA after SS

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 [36]. 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

$$\hat{s}_1(f,\tau) = h_1(f)s(f,\tau) + \Delta s_1(f,\tau), \tag{56}$$

$$\hat{s}_2(f,\tau) = h_2(f)s(f,\tau) + \Delta s_2(f,\tau), \tag{57}$$

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 kth 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

$$\begin{aligned}
&[\boldsymbol{W}_{\mathrm{ICA}}(f)]_{11}\hat{s}_1(f,\tau) + [\boldsymbol{W}_{\mathrm{ICA}}(f)]_{12}\hat{s}_2(f,\tau) \\
&= ([\boldsymbol{W}_{\mathrm{ICA}}(f)]_{11}h_1(f) + [\boldsymbol{W}_{\mathrm{ICA}}(f)]_{12}h_2(f))s(f,\tau) \\
&= 0,
\end{aligned} \tag{58}$$

where we assume $U = 1$ and, e.g., $[\boldsymbol{W}_{\mathrm{ICA}}(f)]_{11} = h_2(f)$ and $[\boldsymbol{W}_{\mathrm{ICA}}(f)]_{12} = -h_1(f)$. 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

$$\mathrm{COS}(f) = \frac{\sum_\tau |\hat{s}_1(f,\tau)|^2 |\hat{s}_2(f,\tau)|^2}{\sqrt{\sum_\tau |\hat{s}_1(f,\tau)|^4}\sqrt{\sum_\tau |\hat{s}_2(f,\tau)|^4}}. \tag{59}$$

From (59), 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 $\mathrm{COS}(f)$ 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 $\mathrm{COS}(f)$. We statistically estimate the distorted speech component of the enhanced signal in each iteration. Here, we assume that the original speech power

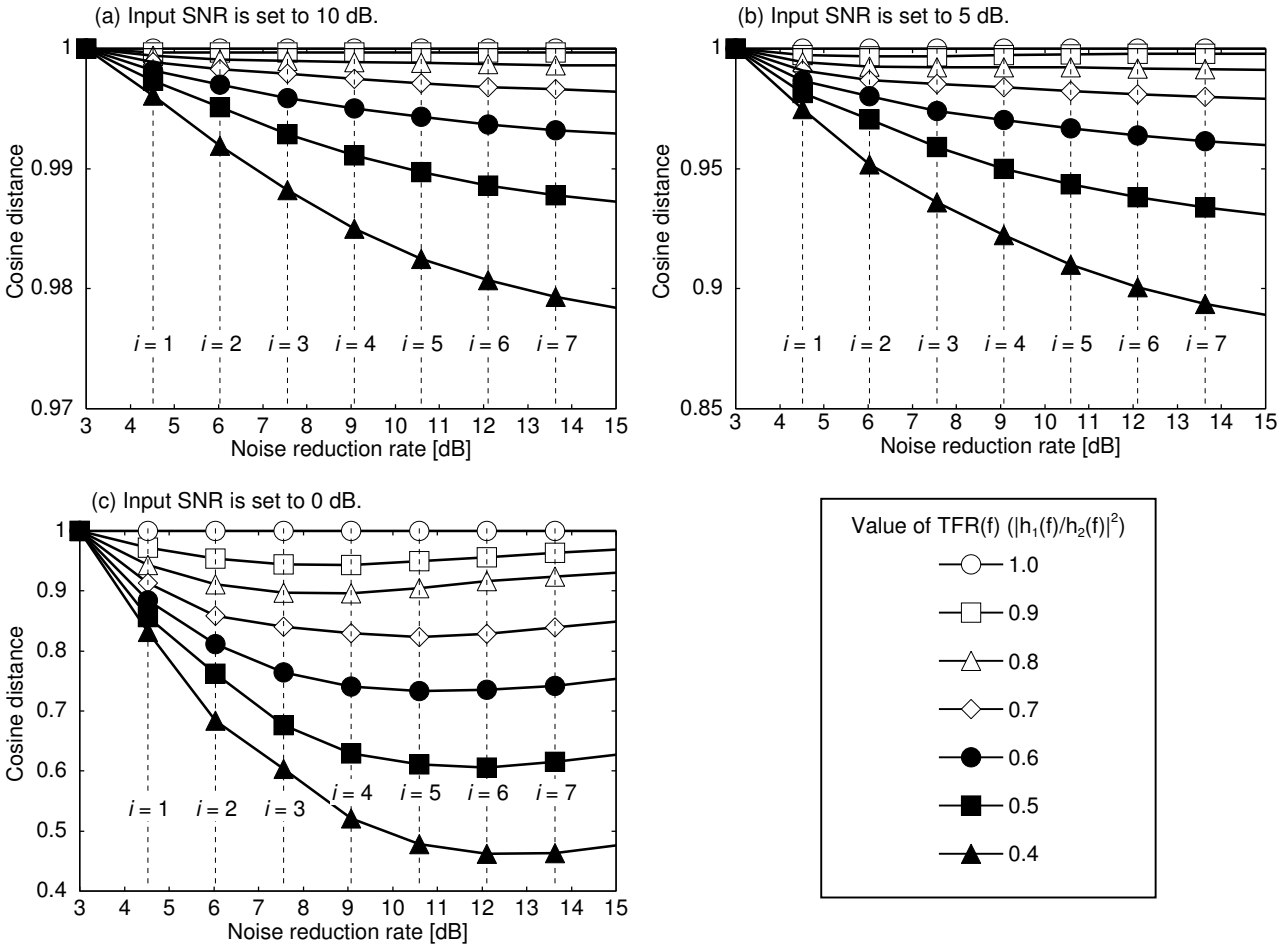

Fig. 19. Relation between number of iterations of iterative BSSA and cosine distance. Input SNR is (a) 10 dB, (b) 5 dB, and (c) 0 dB.

spectrum $|s(f,\tau)|^2$ obeys a gamma distribution with a shape parameter of 0.1 (this is a typical value for speech) as

$$|s(f,\tau)|^2 \sim \frac{x^{-0.9}}{\Gamma(0.1)\theta_{\mathrm{s}}^{0.1}} \exp(-x/\theta_{\mathrm{s}}), \tag{60}$$

where θ_{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 $\theta_{\mathrm{n}}\alpha_{\mathrm{n}}$ when the number of iterations, i, equals zero. Also, the value of α_{n} does not change in each iteration when we use the specific parameters β and η 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

$$\mathcal{M}(\alpha_{\mathrm{n}},\beta,\eta,1)/\alpha_{\mathrm{n}}, \tag{61}$$

and thus, the amount of residual noise after the ith iteration is given by

$$\begin{aligned}\mu_1^{[i]} &= \theta_{\mathrm{n}}\alpha_{\mathrm{n}}\left\{\mathcal{M}(\alpha_{\mathrm{n}},\beta,\eta,1)/\alpha_{\mathrm{n}}\right\}^i \\ &= \theta_{\mathrm{n}}\mathcal{M}^i(\alpha_{\mathrm{n}},\beta,\eta,1)\alpha_{\mathrm{n}}^{1-i}.\end{aligned} \tag{62}$$

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 kth channel after the ith iteration is represented by subtracting the amount of residual noise (62) as

$$\begin{aligned}&|\hat{s}_k^{[i+1]}(f,\tau)|^2 = \\ &\begin{cases} |\hat{s}_k^{[i]}(f,\tau)|^2 - \beta\theta_{\mathrm{n}}\mathcal{M}^i(\alpha_{\mathrm{n}},\beta,\eta,1)\alpha_{\mathrm{n}}^{1-i} \\ (\text{if}\quad |\hat{s}_k^{[i]}(f,\tau)|^2 > \beta\theta_{\mathrm{n}}\mathcal{M}^i(\alpha_{\mathrm{n}},\beta,\eta,1)\alpha_{\mathrm{n}}^{1-i}), \\ \eta^2|\hat{s}_k^{[i]}(f,\tau)|^2 \quad (\text{otherwise}). \end{cases}\end{aligned} \tag{63}$$

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

$$\begin{aligned}\mathrm{ISNR}(f) &= \frac{1}{2}\left(\frac{0.1|h_1(f)|^2\theta_{\mathrm{s}}}{\alpha_{\mathrm{n}}\theta_{\mathrm{n}}} + \frac{0.1|h_2(f)|^2\theta_{\mathrm{s}}}{\alpha_{\mathrm{n}}\theta_{\mathrm{n}}}\right) \\ &= \frac{0.1\theta_{\mathrm{s}}}{2\alpha_{\mathrm{n}}\theta_{\mathrm{n}}}(|h_1(f)|^2 + |h_2(f)|^2).\end{aligned} \tag{64}$$

If we normalize the speech scale parameter θ_{s} to unity, from (64), the noise scale parameter θ_{n} is given by

$$\theta_{\mathrm{n}} = \frac{0.1(|h_1(f)|^2 + |h_2(f)|^2)}{2\alpha_{\mathrm{n}}\mathrm{ISNR}(f)}, \tag{65}$$

and using (65), we can reformulate (63) as

$$\begin{aligned}&|\hat{s}_k^{[i+1]}(f,\tau)|^2 = \\ &\begin{cases} |\hat{s}_k^{[i]}(f,\tau)|^2 - \beta\frac{0.1(|h_1(f)|^2+|h_2(f)|^2)}{2\mathrm{ISNR}(f)}\mathcal{M}^i(\alpha_{\mathrm{n}},\beta,\eta,1)\alpha_{\mathrm{n}}^{-i} \\ (\text{if } |\hat{s}_k^{[i]}(f,\tau)|^2 > \beta\frac{0.1(|h_1(f)|^2+|h_2(f)|^2)}{2\mathrm{ISNR}(f)}\mathcal{M}^i(\alpha_{\mathrm{n}},\beta,\eta,1)\alpha_{\mathrm{n}}^{-i}), \\ \eta^2|\hat{s}_k^{[i]}(f,\tau)|^2 \quad (\text{otherwise}). \end{cases}\end{aligned} \tag{66}$$

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

$$\mathrm{TFR}(f) = |h_1(f)/h_2(f)|^2, \tag{67}$$

and if we normalize $|h_1(f)|^2$ to unity in each frequency subband, $|h_1(f)|^2 + |h_2(f)|^2$ becomes $1 + 1/\mathrm{TFR}(f)$. Finally, we express (66) in terms of the input SNR $\mathrm{ISNR}(f)$ and the transfer function ratio $\mathrm{TFR}(f)$ as

$$|\hat{s}_k^{[i+1]}(f,\tau)|^2 = \begin{cases} |\hat{s}_k^{[i]}(f,\tau)|^2 - \beta \frac{0.1(1+1/\mathrm{TFR}(f))}{2\mathrm{ISNR}(f)} \mathcal{M}^i(\alpha_\mathrm{n}, \beta, \eta, 1)\alpha_\mathrm{n}^{-i} \\ (\text{if} \quad |\hat{s}_k^{[i]}(f,\tau)|^2 > \beta \frac{0.1(1+1/\mathrm{TFR}(f))}{2\mathrm{ISNR}(f)} \mathcal{M}^i(\alpha_\mathrm{n}, \beta, \eta, 1)\alpha_\mathrm{n}^{-i}), \\ \eta^2 |\hat{s}_k^{[i]}(f,\tau)|^2 \quad (\text{otherwise}). \end{cases} \tag{68}$$

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

Figure 19 shows the relation between the TFR and the corresponding value of $\mathrm{COS}(f)$ calculated by (59) and (68). In Fig. 19, we plot the average of $\mathrm{COS}(f)$ over whole frequency subbands. The noise shape parameter α_n is set to 0.2 with the assumption of super-Gaussian noise the input SNR is set to 10 dB, 5 dB, or 0 dB, and the noise scale parameter θ_n is uniquely determined by (65) 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 (see the example of the TFR in Fig. 20). For the internal parameters used in iterative BSSA in this simulation, β and η 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. 19 because DS is still performed even when $i = 0$.

From Figs. 19(a) and (b), which correspond to relatively high input SNRs, we can confirm that the degree of point-source-likeness, i.e., $\mathrm{COS}(f)$, 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 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. 19(c), which shows the case of a low

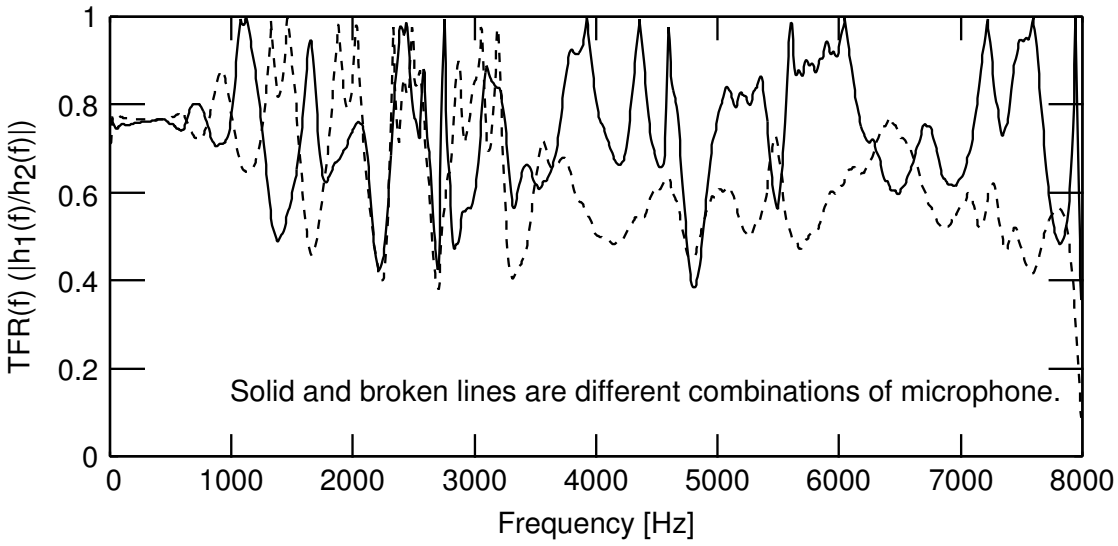

Fig. 20. Typical examples of TFR(f) ($|h_1(f)/h_2(f)|^2$) in each frequency subband.

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.

5.4 Improvement scheme for poor noise estimation

5.4.1 Channel selection in ICA

In this subsection, we propose a channel selection strategy in ICA for achieving high accuracy of noise estimation. As mentioned previously, speech distortion is subjected to ISNR(f) and TFR(f), and the accuracy of noise estimation is degraded along with speech distortion. Figure 20 shows a typical example of the TFR. From Fig. 20, we can confirm that the TFRs in different combinations of microphones are not the same in each frequency subband; in a specific frequency, one microphone pair has higher TFR(f) than another pair, and vice verse in another frequency. Thus, we are able to select the 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 kth channel as

$$\mathrm{E}_\tau[|x_k(f,\tau)|^2] = \mathrm{E}_\tau[|s(f,\tau)|^2]|h_k(f)|^2 + \mathrm{E}_\tau[|n_k(f,\tau)|^2]. \tag{69}$$

Here, $\mathrm{E}_\tau[|s(f,\tau)|^2]$ is a constant, and if we assume a diffuse noise field, $\mathrm{E}_\tau[|n_k(f,\tau)|^2]$ is also a constant. Thus, we can estimate the relative order of $|h_k(f)|^2$ by comparing (69) for every k.

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

$$\max_k \mathrm{E}_\tau[|x_k(f,\tau)|^2] \cdot \xi \leq \mathrm{E}_\tau[|x_k(f,\tau)|^2], \tag{70}$$

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.

5.4.2 Time-variant noise PSD 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 [41] 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 (51), we can perform the speech extraction. The procedure of this noise PSD estimator is described in Appendix F.

5.5 Experiment in real world

5.5.1 Experimental conditions

We conducted objective and subjective evaluation experiments to confirm the validity of the proposed methods under the diffuse and nonstationary noise condition. The

size of the experimental room was $4.2 \times 3.5 \times 3.0$ m^3 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 FFT size was 1024, and the frame shift length was 256. We used 10 speakers (5 males and 5 females) as sources of the original target speech signal. The input SNR was -5, 0, 5, and 10 dB.

5.6 Objective evaluation

We conducted an objective experimental evaluation under the same NRR condition. Figures. 21, 22, 23, and 24 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 minimum mean-square error (MMSE) short-time spectral amplitude (STSA) estimator [7], the Log MMSE estimator incorporating speech-presence uncertainty [9], 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 our proposed methods of iterative BSSA (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 needs at least two-channel signals. Also, we applied a minimum statistics noise PSD estimator [3] to the MMSE STSA estimator, the Log MMSE estimator and musical-noise-free iterative SS, and we use the decision-directed approach for a priori SNR estimation in the MMSE STSA estimator. From Figs. 21 and 23, we can confirm that iterative BSSA methods outperform the MMSE STSA estimator, the Log MMSE estimator and the conventional BSSA in terms of kurtosis ratio. In particular, the kurtosis ratios of the proposed methods are mostly close to 1.0. This means that the proposed 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. 22 and 24). Therefore, a trade-off exists between the amount of musical noise generation and speech distortion in the conventional BSSA and iterative BSSA methods.

5.7 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 our proposed 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. 25 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 proposed method in terms of human perception. This result is plausible because humans are often more sensitive to musical noise than to speech distortion as indicated in past studies, e.g., [21].

5.8 Conclusion

In this section, I addressed a musical-noise-free blind speech extraction method using a microphone array that can be applied to nonstationary noise. First, I proposed iterative BSSA using a new BSSA structure, which generates almost no musical noise even with increasing noise reduction performance. The proposed method consists of iterative blind dynamic noise estimation by ICA, 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.

Secondly, in relation to the proposed method, I discussed the justification of applying ICA to signals nonlinearly distorted by SS. I theoretically clarify that the degradation in ICA-based noise estimation obeys an amplitude variation in room transfer functions between the target users and microphones. Therefore, I proposed the introduction of a channel selection strategy in ICA to achieve higher accuracy of noise estimation. Furthermore I introduced a time-variant noise PSD estimator instead of ICA to improve the noise estimation accuracy.

Finally, from the objective evaluation experiments, it was shown that there is a trade-off between the amount of musical noise generation and speech distortion in

both the conventional BSSA 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 total sound quality.

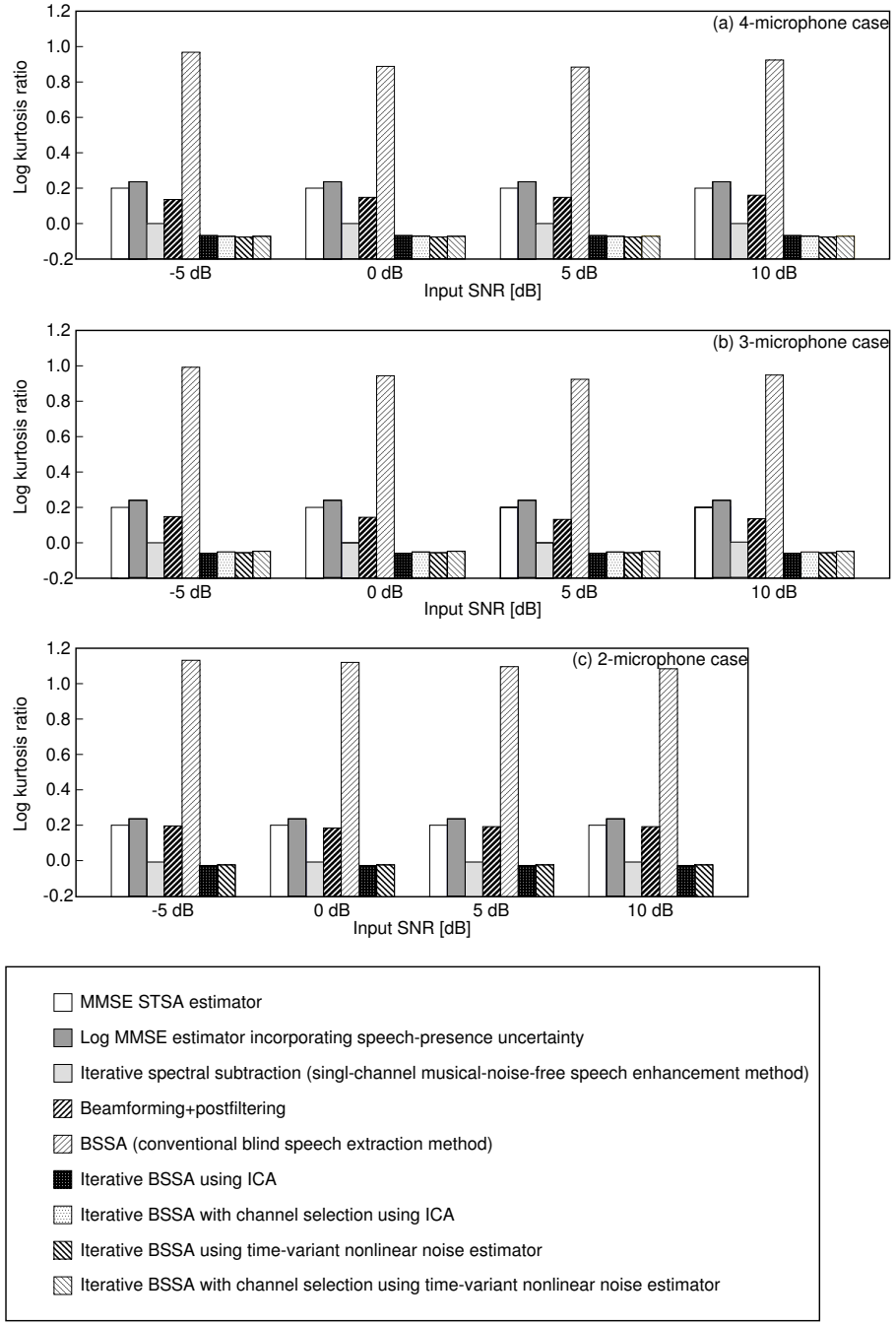

Fig. 21. Kurtosis ratio obtained from experiment for traffic noise under 10-dB NRR condition.

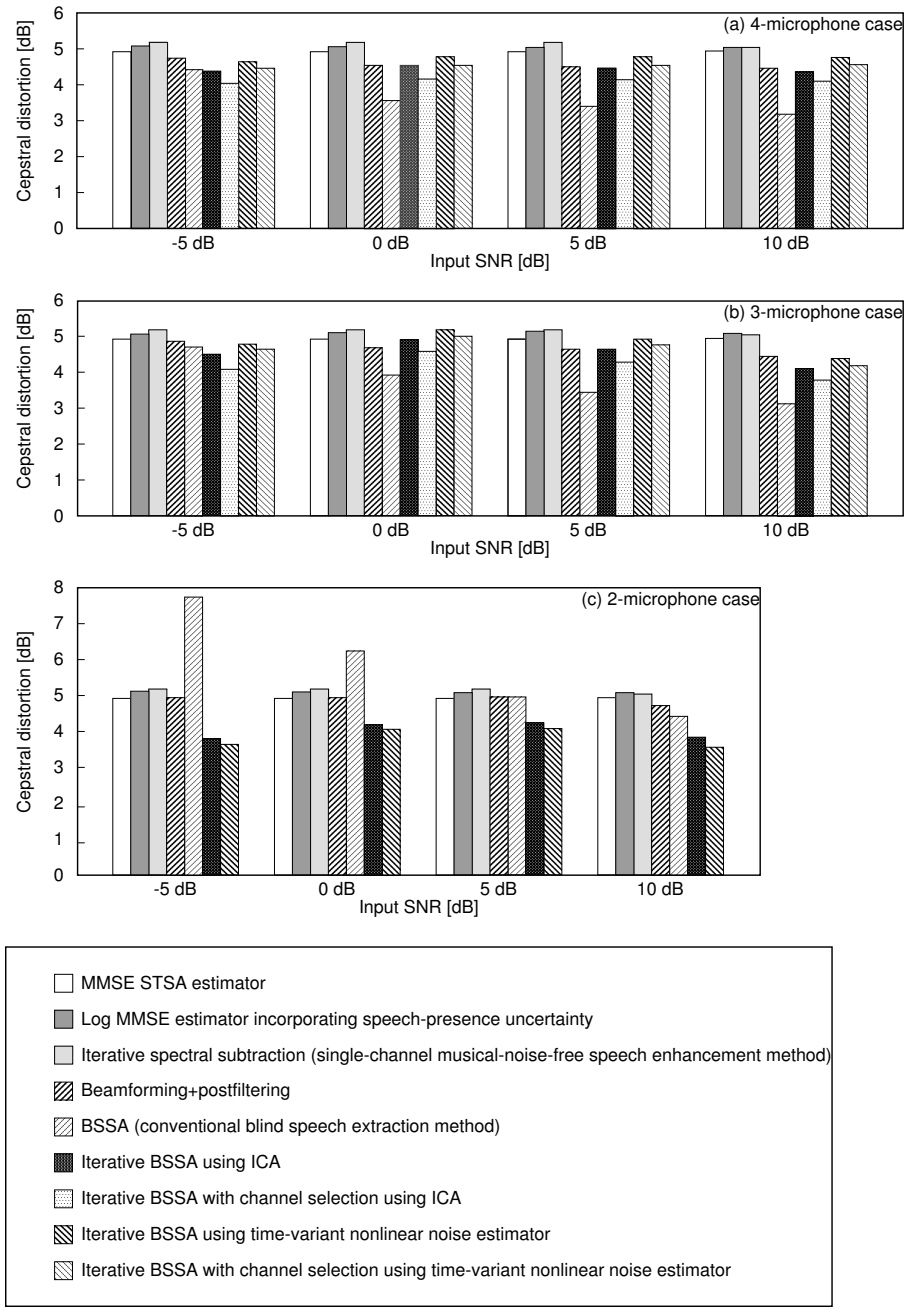

Fig. 22. Cepstral distortion obtained from experiment for traffic noise under 10-dB NRR condition.

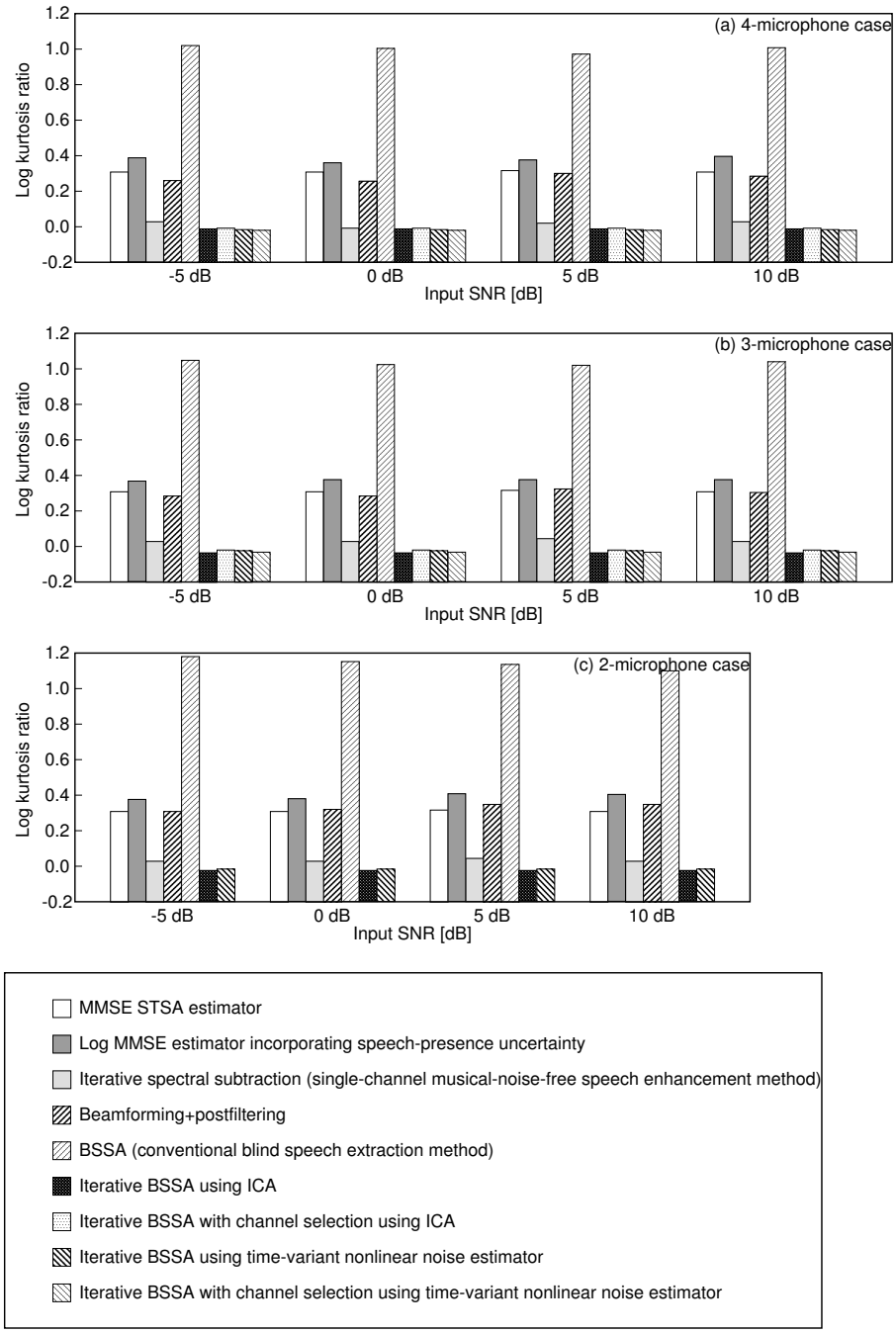

Fig. 23. Kurtosis ratio obtained from experiment for railway station noise under 10-dB NRR condition.

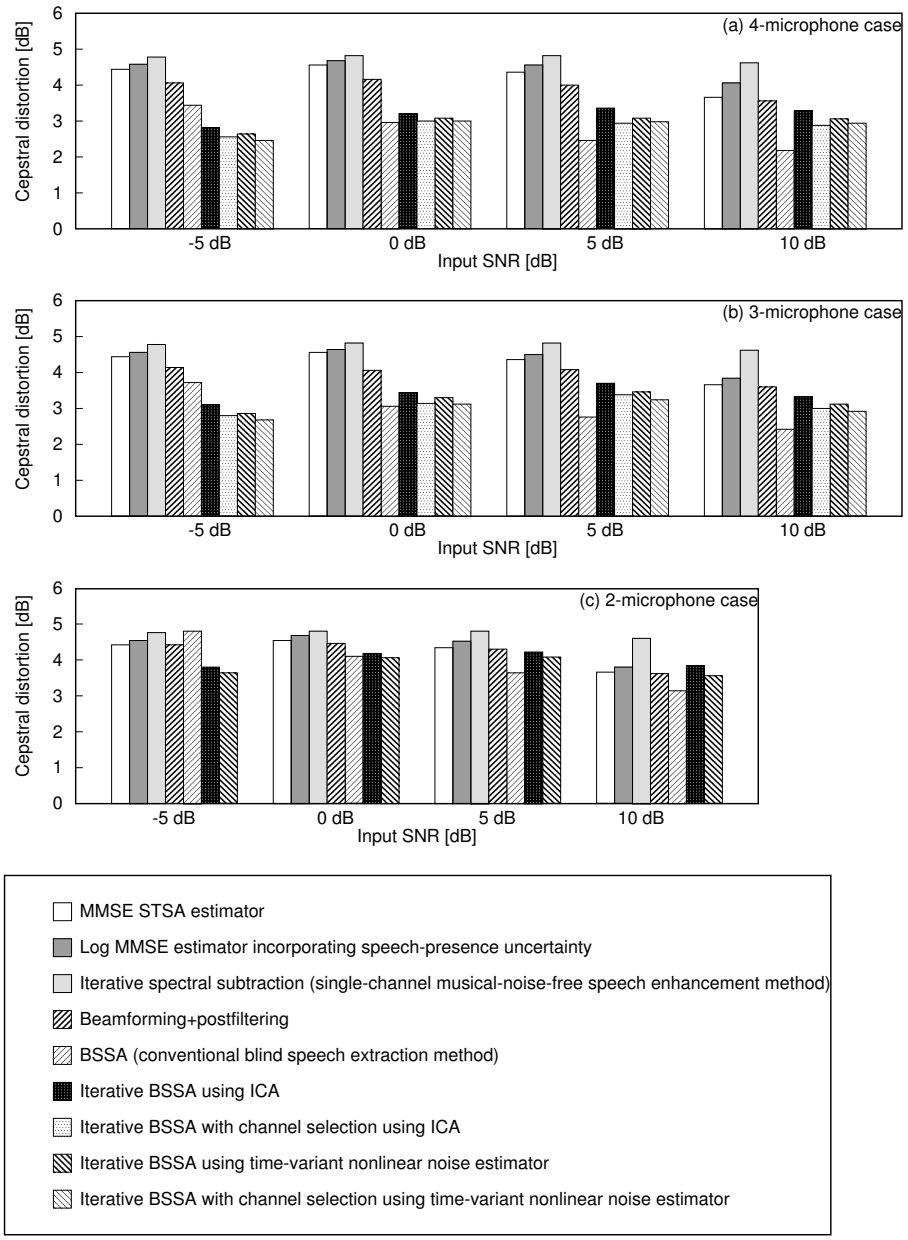

Fig. 24. Cepstral distortion obtained from experiment for railway station noise under 10-dB NRR condition.

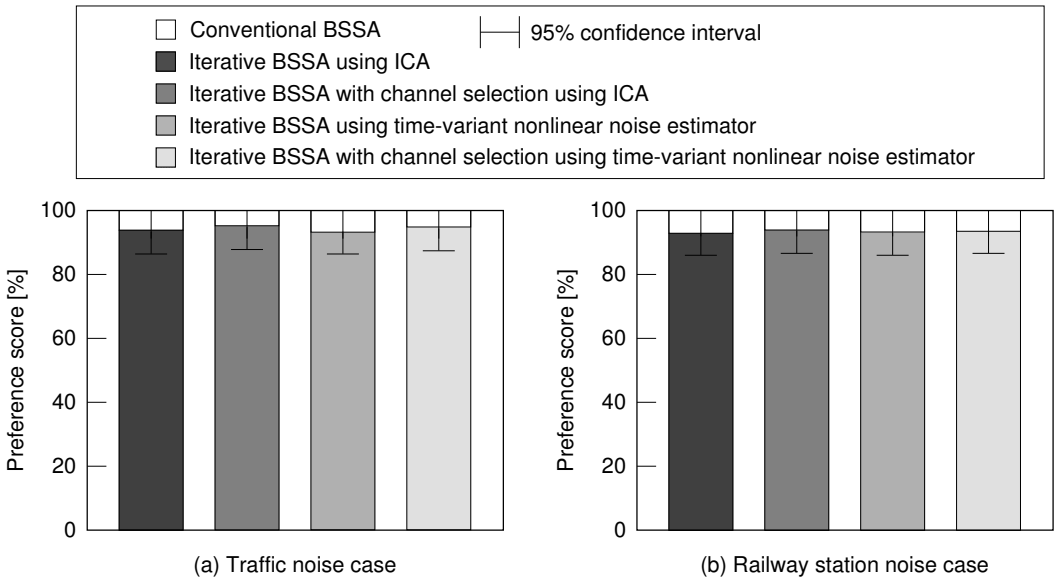

Fig. 25. Subjective evaluation results for (a) traffic noise and (b) railway station noise.

6. Conclusion

6.1 Summary of dissertation

In this dissertation, I proposed a new speech enhancement theory, i.e., musical-noise-free speech enhancement, which can be applied to the musical noise mitigation problem. From a theoretical behavior and objective experiments, it was revealed that the proposed method outperforms conventional methods in both the amount of musical noise generation and speech distortion. Furthermore, I proposed a new iterative signal extraction method that can be applied to nonstationary noise. Also, I discussed the justification of applying ICA to signals nonlinearly distorted by SS. From the theoretical analysis, it was clarified that the degradation in ICA-based noise estimation obeys an amplitude variation in room transfer function between the target user and microphones. According to the results of objective and subjective evaluations, the a proposed method is superior to the conventional BSSA in terms of total sound quality.

In Sect. 3, a theoretical analysis of iterative SS was given. The theoretical analysis indicates that the first-, second-, and fourth-order moments of the power spectra can be used to estimate the amount of noise reduction and musical noise generation, and I introduced a gamma-distribution approximation to simulate iteratively applied weak SS. Next, I conducted a comparison of the amount of musical noise generated for different parameter settings under the same noise reduction performance. It was clarified from mathematical analysis that iterative SS with very weak processing can realize high-quality speech enhancement with a small amount of musical noise generated.

On the basis of the above-mentioned findings, I proposed a new speech enhancement theory, i.e., musical-noise-free speech enhancement, in Sect. 4. In this section, I discussed a theorem of musical-noise-free conditions in iterative SS, and I mathematically derived the internal parameter settings to satisfy the musical-noise-free condition. It was clarified that the optimal parameters satisfying the musical-noise-free condition can generate almost no musical noise even with high noise reduction. This desirable property of iterative SS was well supported by the results of a comparative experiment between iterative SS and commonly used noise reduction methods, including conventional non-iterative SS, Wiener filtering, and the MMSE STSA estimator. In summary, the proposed theory mathematically proves that iterative SS with the optimal parameters is advantageous for achieving high-quality noise reduction, which has only been

experimentally shown in previous studies.

Next, I proposed a musical-noise-free blind speech extraction method using a microphone array that can be applied to nonstationary noise in Sect. 5. Also, in relation to the proposed method, I discussed the justification of applying ICA to signals nonlinearly distorted by SS. From the theoretical analysis, I showed that the degradation in ICA-based noise estimation obeys an amplitude variation in room transfer function between the target user and microphones. Moreover, to achieve higher accuracy of noise estimation, I proposed the introduction of a channel selection strategy in ICA and a time-variant noise PSD estimator. From objective evaluation experiments, it was shown that there is a trade-off between the amount of musical noise generation and speech distortion in both the conventional BSSA and iterative BSSA. However, in a subjective preference test, the iterative BSSA obtained a higher preference score than the conventional BSSA. Thus, the iterative BSSA is superior to the conventional BSSA in terms of total sound quality.

6.2 Future work

In this dissertation, I have reported a method for improving the sound quality for human-hearing applications. However, the following problems still remain to be solved.

Although I have mathematically optimized the internal parameter settings based on higher-order statistics in iterative SS, this is merely one example of a nonlinear speech enhancement technique. There is a strong possibility that a musical-noise-free theory can be applied to other nonlinear speech enhancement methods (e.g., Wiener filtering and Bayesian MMSE estimator). These speech enhancement methods are often considered to generate as less musical noise and less speech distortion than SS. Therefore, it is expected that the development of a musical-noise-free theory for these speech enhancement methods will greatly improve the sound quality.

Moreover, in the musical-noise-free theory, I pursued higher-order statistics in only the speech absence periods, and optimized the internal parameters in terms of musical noise generation. However, I have not taken into account of the distortion of speech component by speech enhancement. The calculation of the distorted speech component by speech enhancement, e.g., cepstral distortion, requires a reference (clean) speech signal. However, in actual situations, the speech component is always overlapped with noise, and we cannot obtain a clean speech signal. To overcome this problem, as an

unsupervised measure of speech distortion estimated in a reference-free manner, *the kurtosis of the speech power spectrum* has recently been proposed, which is effective for optimizing parameters in the speech enhancement method and the speech recognition performance [39, 49, 50, 51]. This method is based on a moment-cumulant transformation technique with respect to the statistical estimates of observable noise and noisy speech signals. In the future, internal parameters can be optimized in terms of both speech distortion and musical noise generation on the basis of this finding.

Acknowledgements

This dissertation is a summary of four years of study carried out at Graduate School of Information Science, Nara Institute of Science and Technology, Japan.

I would like to express my sincere thanks to Emeritus Professor Kiyohiro Shikano and Professor Satoshi Nakamura of Nara Institute of Science and Technology, my dissertation advisers, for their valuable guidance and constant encouragement.

I would also like to express my appreciation to Professor Kenji Sugimoto of Nara Institute of Science and Technology, a member of the dissertation committee, for his valuable comments on the dissertation.

I would especially like to express my deep gratitude to Associate Professor Hiroshi Saruwatari of Nara Institute of Science and Technology for his continuous teaching and essential advice on both technical and non technical issues. This work could not have been accomplished without his well-directed advice, helpful suggestions, and fruitful discussions with him. I have learned many valuable aspects of being a researcher from his attitude toward study and have always enjoyed conducting research with him.

I would also like to express my appreciation to Associate Professor Nobutaka Ono of National Institute of Informatics, a member of the dissertation committee, for his valuable comments on the dissertation.

I would like to express my gratitude to Professor Martin Bouchard for his valuable suggestions and shrewd advice on the time-variant noise PSD estimator, as well as Associate Professor Tomoki Toda and Assistant Professors Hiromichi Kawanami, Sakriani Sakti, and Graham Neubig of Nara Institute of Science and Technology, and Assistant Professor Sunao Hara of Okayama University for their instructive comments.

The former part of this work could not have been completed without the collaboration of many researchers. I would like to express my gratitude to Dr. Tomoya Takatani for his valuable suggestions and shrewd advice on ICA-based techniques. Also, I would like to acknowledge Dr. Noriyoshi Kamado, currently a researcher at NTT laboratory, and Dr. Hironori Doi, currently an engineer at Dwango Corporation, for valuable discussions and suggestions on this work.

The latter part of this work also could not have been achieved without the collaboration of many researchers. I especially thank Mr. Takayuki Inoue, Dr. Yu Takahashi, and Dr. Kazunobu Kondo, researchers at Yamaha Corporation, for their beneficial and valuable comments on musical-noise-free speech enhancement and its applications.

Many staff and members of my research group have supported me in carrying out experiments and writing this dissertation at Nara Institute of Science and Technology; I would especially like to express my appreciation of the valuable discussions with them on technical issues of speech signal processing and digital signal processing, and their provision of a comfortable computing environment in our laboratory. I also wish to express my deep gratitude to Mrs. Toshie Nobori and Mrs. Manami Matsuda, secretaries at our laboratory, for their kind help and support in all aspects of my research.

I appreciate the opportunity to have studied with all the students in our research group at Nara Institute of Science and Technology. I thank Mr. Keigo Kubo and Mrs. Fine Dwinita April, who are Ph.D. candidates at Nara Institute of Science and Technology, for useful discussions on this work.

Finally, I wish to thank all members of my family for their support over many years.

References

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

[2] M. Berouti, R. Schwartz, J. Makhoul, "Enhancement of speech corrupted by acoustic noise," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP1979)*, pp.208–211, 1979.

[3] R. Martin, "Spectral subtraction based on minimum statistics," *Proceedings of European Signal Processing Conference (EUSIPCO1994)*, pp.1182–1185, 1994.

[4] J. Chen, J. Benesty, Y. Huang, S. Doclo, "New insights into the noise reduction Wiener filter," *IEEE Transactions of Audio, Speech, and Language Processing*, vol.14, no.4, pp.1218–1234, 2006.

[5] J. Lim, A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol.67, no.12, pp.1586–1604, 1979.

[6] R. McAulay, M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol.28, no.2, pp.137–145, 1980.

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

[8] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol.23, no.2, pp.1109–1121, 1985.

[9] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectra amplitude estimator," *IEEE Signal Processing Letters*, vol.9, no.4, pp.113–116, 2002.

[10] T. Lotter, P. Vary, "Speech enhancement by maximum a posteriori spectral amplitude estimation using a supergaussian speech model," *EURASIP Journal on Applied Signal Processing*, no.7, pp.1110–1126, 2005.

[11] C. Breithaupt, R. Martin, "Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.19, no.2, pp.277–289, 2011.

[12] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Transactions on Speech and Audio Processing*, vol.2, no.2, pp.345–349, 1994.

[13] Z. Goh, K.-C. Tan, and B. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Transactions on Speech and Audio Processing*, vol.6, no.3, pp.287–292, 1998.

[14] K. Yamashita, S. Ogata, T. Shimamura, "Spectral subtraction iterated with weighting factors," *Proceedings of IEEE Speech Coding Workshop*, pp.138–140, 2002.

[15] K. Yamashita, S. Ogata, T. Shimamura, "Improved spectral subtraction utilizing iterative processing," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol.J88-A, no.11, pp.1246–1257, 2005 (in Japanese).

[16] M. R. Khan, T. Hasan, "Iterative noise power subtraction technique for improved speech quality," *Proceedings of International Conference on Electrical and Computer Electrical Engineering (ICECE2008)*, pp.391–394, 2008.

[17] X. Li, G. Li, X. Li, "Improved voice activity detection based on iterative spectral subtraction and double thresholds for CVR" *Proceedings of 2008 Workshop on Power Electronics and Intelligent Transportation System*, pp.153–156, 2008.

[18] T. Fukumori, M. Morise, T. Nishiura, H. Nanjo, "Musical tone reduction on iterative spectral subtraction based on optimum flooring parameters," *IEICE Technical Report*, vol.110, no.54, pp.43–48, 2010 (in Japanese).

[19] S. Li, J.-Q. Wang, M. Niu, X.-J. Jing, T. Liu, "Iterative spectral subtraction method for millimeter-wave conducted speech enhancement," *Journal of Biomedical Science and Engineering*, vol.2010, no.3, pp.187–192, 2010.

[20] 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," *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC2008)*, 2008.

[21] 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," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2009)*, pp.4433–4436, 2009.

[22] H. Yu, T. Fingscheidt, "Black box measurement of musical toens produced by noise reduction systems," *Proceedings of IEEE International Conference in Acoustics, Speech and Signal Processing (ICASSP2012)*, pp.4573–4576, 2012.

[23] N. Derakhshan, M. Rahmani, A. Akbari, A. Ayotollahi, "An objective measure for musical noise assessment in noise reduction systems," *Proceedings of IEEE International Conference in Acoustics, Speech and Signal Processing (ICASSP2009)*, pp.4429–4432, 2009.

[24] 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 Transactions on Audio, Speech, and Language Processing*, vol.19, No.6, pp.1770–1779, 2011.

[25] T. Inoue, H. Saruwatari, K. Shikano, K. Kondo, "Theoretical analysis of musical noise in Wiener filtering family via on higher-order statistics," *Proceedings of IEEE International Conference in Acoustics, Speech and Signal Processing (ICASSP2011)*, pp.5076–5079, 2011.

[26] S. Kanehara, H. Saruwatari, R. Miyazaki, K. Shikano, K. Kondo, "Theoretical analysis of musical noise generation in noise reduction method with decision-directed a priori SNR estimator," *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC2012)*, 2012.

[27] 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," *Proceedings of APSIPA annual Summit and Conference (APSIPA2012)*, 2012.

[28] H. Saruwatari, S. Kanehara, R. Miyazaki, K. Shikano, K. Kondo, "Musical noise analysis for Bayesian minimum mean-square error speech amplitude estimator based on higher-order statistics," *Proceedings of INTERSPEECH2013*, pp.441–445, 2013.

[29] H. Yu, T. Fingscheidt, "A figure of metric for instrumental optimization of noise reduction algorithms," *Proceedings of 5th Biennial Workshop on DSP for In-Vehicle Systems*, pp.1—-8, 2011.

[30] R. C. Hendriks, R. Heusdens, J. Jensen, "MMSE based noise PSD tracking with low complexity," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2010)*, pp.4266–4269, 2010.

[31] T. Gerkmannm R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.20, no.4m pp.1383–1393, 2012.

[32] S. Fischer, K. D. Kammeyer, "Broadband beamforming with adaptive post filtering for speech acquisition in noisy environment," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP1997)*, pp.359–362, 1997.

[33] J. Cho, A. Krishnamurthy, "Speech enhancement using microphone array in moving vehicle environment," *Proceedings of Intelligent Vehicles Symposium*, pp.366–371, 2003.

[34] Y. Ohashi, T. Nishikawa, H. Saruwatari, A. Lee, K. Shikano, "Noise robust speech recognition based on spatial subtraction array," *International Workshop on Nonlinear Signal and Image Processing (NSIP2005)*, pp.324–327, 2005.

[35] J. Even, H. Saruwatari, K. Shikano, "New architecture combining blind signal extraction and modified spectral subtraction for suppression of background

noise," *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC2008)*, 2008.

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

[37] 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 Jouranl on Advances in Signal Processing*, vol.2010, Article ID 431347, 25 pages, 2010.

[38] 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 Transactions on Audio, Speech and Language Processing*, vol.19, no.6, pp.1457–1466, 2011.

[39] R. Miyazaki, H. Saruwatari, K. Shikano, "Theoretical analysis of amounts of musical noise and speech distortion in structure-generalized parametric spatial subtraction array," *IEICE Transactions Fundamentals*, vol.95-A, no.2, pp.586–590, 2012.

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

[41] A. Homayoun, M. bouchard, "Improved noise power spectrum density estimation for binaural hearing aids operating in a diffuse noise field environment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.17, no.4, pp.521–533, 2009.

[42] E. W. Stacy, "A generalization of the gamma distribution," *The Annals of Mathematical Statistics*, vol.33, no.3, pp.1187–1192, 1962.

[43] J. W. Shin, J-H. Chang, N. S. Kim, "Statistical modeling of speech signal based on generalized gamma distribution," *IEEE Signal Processing Letters*, vol.12, no.3, pp.258–261, 2005.

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

[45] M. Evans, N. A. J. Hastings, B. Peacock, Eds., *Statistical Distributions, 2nd ed,* Wiley-Interscience, 1993.

[46] W. Q. Meeker, L. A. Escobar, *Statistical methods for reliability data,* Wiley-Interscience, 1998.

[47] A. Lee, T. Kawahara, K. Shikano, "Julius – An open source real-time large vocabulary recognition engine," *Proceedings of European Conference on Speech Communication and Technology (INTERSPEECH2001)*, pp.1691–1694, 2001.

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

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

[50] R. Wakisaka H. Saruwatari, K. Shikano, T. Takatani, "Blind speech prior estimation for generalized minimum mean-square error short-time spectral amplitude estimator," *Proceedings of INTERSPEECH*, pp.361–364, 2011.

[51] 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," *Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP2012)*, pp.4049–4052, 2012.

[52] E. Vincent, R. Gribonval, C. Fevotte, "Performance measurement in blind audio source separation," *Transactions of Audio, Speech, and Language Pricessing*, vol.12, no.4, pp.1462–1469, 2006.

Appendix

A. Approximate Accuracy of Noise Power Spectra after Weak SS

In this appendix, I conduct the goodness-of-fit test for the noise power spectra to measure the goodness of fit an approximated gamma distribution p.d.f. (see Fig. 5). I test the null hypothesis that the histogram of the noise power spectra in 1st iteration of SS and the approximated p.d.f. of a gamma distribution come from populations with the same distribution, using the Kolmogorov-Smirnov test.

Noisy observation signal was generated by adding noise signals to target speech signals with an SNR of 0 dB. I conducted the experiment on white Gaussian noise. The target speech signal was two male and two female speakers in Japanese. In iterative SS, the parameter settings of β and η are 2.4 and 0.9. This parameter setting satisfies the musical-noise-free condition.

From this test, I can confirm that the Kolmogorov-Smirnov test dose not reject the null hypothesis at the default 5% significance level. Figure 26 shows an example of the histogram of the noise power spectra and p.d.f. of a gamma distribution corresponding to its histogram. Therefore, it can be said that the approximation of the noise power spectra after weak SS using a gamma distribution is valid.

B. Typical Example of Optimal Parameter Settings Satisfying Musical-Noise-Free Condition

This appendix shows a typical example of the optimal parameter settings satisfying the musical-noise-free condition. I calculate combinations of the oversubtraction parameter β and the flooring parameter η under three types of shape parameter α_0, namely 0.2, 0.5, and 1.0. Tables 1, 2, and 3 show the typical value of the optimal parameter settings in each shape parameter α_0.

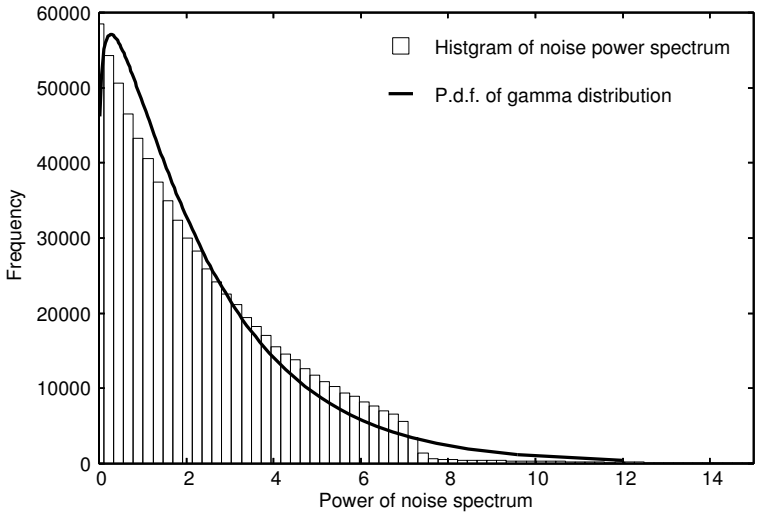

Fig. 26. Histogram of noise power spectra in 1st iteration of iterative SS and p.d.f. of gamma distribution corresponding to its histogram.

C. Histogram of Noise Power Spectra in Each Iteration of Iterative SS

This appendix provides a typical example of the histogram of the noise power spectra in each iteration of iterative SS. Noisy observation signal was generated by adding noise signals to target speech signals with an SNR of 0 dB. I conducted the experiment on white Gaussian noise. The target speech signal was two male and two female speakers in Japanese. In iterative SS, the parameter settings of β and η are 2.4 and 0.9. This parameter setting satisfies the musical-noise-free condition.

Figure 27 shows an example of the histogram of noise power spectra in each iteration. From Fig. 27, since the SS process in each step is weak, the histograms are not changed significantly. Also, it is confirmed that the average of noise power spectrum is reduced without change in kurtosis.

Table 1. Example of oversubtraction parameter β and flooring parameter η satisfying musical-noise-free condition for $\alpha_0 = 0.2$

Oversubtraction parameter β	Flooring parameter η
8.0	0.9314
8.1	0.9269
8.2	0.9226
8.3	0.9182
8.4	0.9140
8.5	0.9098
8.6	0.9057
8.7	0.9017
8.8	0.8977
8.9	0.8938
9.0	0.8899

D. Evaluation of Total Sound Quality for iterative SS

I conduct an objective evaluation to assess the total sound quality of the signal enhanced by iterative SS in this appendix. The SDR and SIR are defined in [52] as the evaluation scores. Here, the estimated signal $\hat{s}(t)$ is defined as

$$\hat{s}(t) = s_{\text{target}}(t) + s_{\text{interf}}(t) + s_{\text{artif}}(t), \tag{71}$$

where $s_{\text{target}}(t)$ is the allowable deformation of the target source, $s_{\text{interf}}(t)$ is the allowable deformation of the sources that account for the interferences of the undesired sources, and $s_{\text{artif}}(t)$ is an *artifact* term that may correspond to the artifacts of the separation algorithm, such as musical noise, or simply undesirable deformation induced by the nonlinear property of the separation algorithm.

The formulas for SDR and SIR are defined as

$$\text{SDR} = 10\log_{10}\frac{\sum_t s_{\text{target}}(t)^2}{\sum_t \{e_{\text{interf}}(t) + e_{\text{artif}}(t)\}^2}, \tag{72}$$

$$\text{SIR} = 10\log_{10}\frac{\sum_t s_{\text{target}}(t)^2}{\sum_t e_{\text{interf}}(t)^2}. \tag{73}$$

Table 2. Example of oversubtraction parameter β and flooring parameter η satisfying musical-noise-free condition for $\alpha_0 = 0.5$

Oversubtraction parameter β	Flooring parameter η
4.0	0.9637
4.1	0.9540
4.2	0.9446
4.3	0.9356
4.4	0.9268
4.5	0.9183
4.6	0.9101
4.7	0.9021
4.8	0.8943
4.9	0.8867
5.0	0.8794

SDR indicates the quality of the separated target sound and SIR indicates the degree of separation between the target and other sounds. Therefore, SDR indicates the total evaluation score that involves SIR.

Noisy observation signal was generated by adding noise signals to target speech signals with an SNR of 0 dB. I conducted the experiment on white Gaussian noise and babble noise. The target speech signal was two male and two female speakers in Japanese. This parameter settings satisfy the musical-noise-free condition.

Figures 28 and 29 show the relation between the number of iterations and SDR or SIR in each noise case. From Figs. 28 and 29, the SDR and SIR increase with increasing the number of iterations: this indicates that the total sound quality improves while reducing noise in each iteration.

Table 3. Example of oversubtraction parameter β and flooring parameter η satisfying musical-noise-free condition for $\alpha_0 = 1.0$

Oversubtraction parameter β	Flooring parameter η
2.0	0.9683
2.1	0.9458
2.2	0.9248
2.3	0.9052
2.4	0.8869
2.5	0.8697
2.6	0.8535
2.7	0.8382
2.8	0.8238
2.9	0.8101
3.0	0.7971

E. Theoretical Analysis of Amount of Musical Noise Generation and Speech Distortion

This appendix provides a brief review of the amount of musical noise generation and speech distortion in parametric BSSA.

E-I Analysis of amount of musical noise

E-I-I Analysis in the case of parametric BSSA

In this section, I 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, $\mathrm{kurt}_{\mathrm{in}}^{(\mathrm{n})}$, as [20]

$$\mathrm{kurt}_{\mathrm{in}}^{(\mathrm{n})} = \frac{(\alpha_{\mathrm{n}} + 2)(\alpha_{\mathrm{n}} + 3)}{\alpha_{\mathrm{n}}(\alpha_{\mathrm{n}} + 1)}. \tag{74}$$

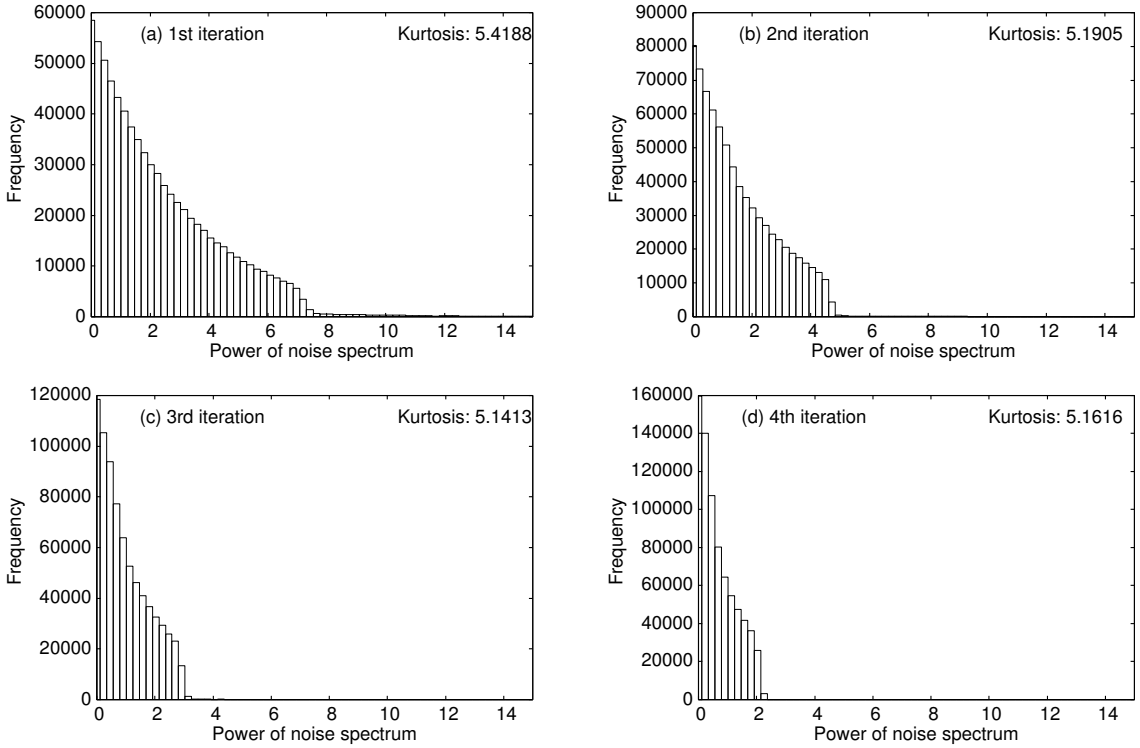

Fig. 27. Histograms of noise power spectra in each iteration of iterative SS. (a) 1st iteration, (b) 2nd iteration, (c) 3rd iteration, and (d) 4th iteration.

The kurtosis in the power spectral domain after DS is given by [37]

$$\mathrm{kurt}_{\mathrm{DS}}^{(\mathrm{n})} \simeq J^{-0.7} \cdot (\mathrm{kurt}_{\mathrm{in}}^{(\mathrm{n})} - 6) + 6. \tag{75}$$

Similarly to Eq. (74), 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}:

$$\mathrm{kurt}_{\mathrm{DS}}^{(\mathrm{n})} = \frac{(\alpha_{\mathrm{DS}} + 2)(\alpha_{\mathrm{DS}} + 3)}{\alpha_{\mathrm{DS}}(\alpha_{\mathrm{DS}} + 1)}. \tag{76}$$

This can be expanded as

$$\alpha_{\mathrm{DS}}^2(\mathrm{kurt}_{\mathrm{DS}}^{(\mathrm{n})} - 1) + \alpha_{\mathrm{DS}}(\mathrm{kurt}_{\mathrm{DS}}^{(\mathrm{n})} - 5) - 6 = 0, \tag{77}$$

and we have

$$\alpha_{\mathrm{DS}} = \frac{-\mathrm{kurt}_{\mathrm{DS}} + 5 + \sqrt{\mathrm{kurt}_{\mathrm{DS}}^2 + 14\,\mathrm{kurt}_{\mathrm{DS}} + 1}}{2\,\mathrm{kurt}_{\mathrm{DS}} - 2}. \tag{78}$$

Then, using Eqs. (74) and (75), α_{DS} can be expressed in terms of α_{n} as

$$\begin{aligned}\alpha_{\mathrm{DS}} = &\left[2J^{-0.7}\cdot\left\{\frac{(\alpha_{\mathrm{n}}+2)(\alpha_{\mathrm{n}}+3)}{\alpha_{\mathrm{n}}(\alpha_{\mathrm{n}}+1)}-6\right\}+10\right]^{-1}\\ &\cdot\left[\left\{\left(J^{-0.7}\cdot\left\{\frac{(\alpha_{\mathrm{n}}+2)(\alpha_{\mathrm{n}}+3)}{\alpha_{\mathrm{n}}(\alpha_{\mathrm{n}}+1)}-6\right\}+6\right)^{2}\right.\right.\\ &\left.+14J^{-0.7}\cdot\left\{\frac{(\alpha_{\mathrm{n}}+2)(\alpha_{\mathrm{n}}+3)}{\alpha_{\mathrm{n}}(\alpha_{\mathrm{n}}+1)}-6\right\}+85\right\}^{0.5}\\ &\left.-\left(J^{-0.7}\cdot\left\{\frac{(\alpha_{\mathrm{n}}+2)(\alpha_{\mathrm{n}}+3)}{\alpha_{\mathrm{n}}(\alpha_{\mathrm{n}}+1)}-6\right\}\right)-1\right].\end{aligned} \tag{79}$$

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

$$\mathrm{kurt}^{(\mathrm{n})}_{\mathrm{BSSA}} = \mathcal{M}(\alpha_{\mathrm{DS}},\beta,4,n)/\mathcal{M}^{2}(\alpha_{\mathrm{DS}},\beta,2,n), \tag{80}$$

where $\mathcal{M}(\alpha,\beta,m,n)$ can be expressed as [24]

$$\begin{aligned}\mathcal{M}(\alpha,\beta,m,n) = &\sum_{l=0}^{m/n}\frac{(-\beta)^{l}\Gamma^{l}(\alpha+n)\Gamma(m/n+1)}{\Gamma^{l+1}(\alpha)\Gamma(l+1)\Gamma(m/n-l+1)}\\ &\Gamma\left(\alpha+m-nl,\left(\beta\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}\right)^{\frac{1}{n}}\right),\end{aligned} \tag{81}$$

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

$$\Gamma(\alpha,z) = \int_{z}^{\infty} t^{\alpha-1}\exp(-t)dt. \tag{82}$$

Finally, using Eqs. (5), (74), and (80), we can determine the resultant kurtosis ratio through a parametric BSSA as

$$\text{kurtosis ratio}^{(\mathrm{n})}_{\mathrm{BSSA}} = \mathrm{kurt}^{(\mathrm{n})}_{\mathrm{BSSA}} \,/\, \mathrm{kurt}^{(\mathrm{n})}_{\mathrm{in}}\,. \tag{83}$$

E-I-II 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 Eq. (80) with the shape

parameter of input noise α_{n}, we can express the resultant kurtosis after channelwise GSS as

$$\mathrm{kurt}^{(\mathrm{n})}_{\mathrm{chGSS}} = \mathcal{M}(\alpha_{\mathrm{n}}, \beta, 4, n)/\mathcal{M}^2(\alpha_{\mathrm{n}}, \beta, 2, n). \tag{84}$$

Next, using Eqs. (75) and (84), we can derive the change in kurtosis after a parametric chBSSA as

$$\mathrm{kurt}^{(\mathrm{n})}_{\mathrm{chBSSA}} \simeq J^{-0.7} \cdot (\mathrm{kurt}^{(\mathrm{n})}_{\mathrm{chGSS}} - 6) + 6. \tag{85}$$

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

$$\text{kurtosis ratio}^{(\mathrm{n})}_{\mathrm{chBSSA}} = \mathrm{kurt}^{(\mathrm{n})}_{\mathrm{chBSSA}} \,/\, \mathrm{kurt}^{(\mathrm{n})}_{\mathrm{in}}\,. \tag{86}$$

E-II Analysis of amount of speech distortion

E-II-I 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, respectability. 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

$$\begin{aligned} |s_{\mathrm{BSSA}}(f,\tau)| &= \sqrt[2n]{|s(f,\tau)|^{2n} - \beta\overline{|z_{\mathrm{DS}}(f,\tau)|^{2n}}} \\ &= \sqrt[2n]{|s(f,\tau)|^{2n} - \beta C_{\mathrm{BSSA}}\overline{|s(f,\tau)|^{2n}}}, \end{aligned} \tag{87}$$

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

$$\begin{aligned} C_{\mathrm{BSSA}} &= \overline{|z_{\mathrm{DS}}(f,\tau)|^{2n}}/\overline{|s(f,\tau)|^{2n}} \\ &= J^{-n}\overline{|n(f,\tau)|^{2n}}/\overline{|s(f,\tau)|^{2n}} \\ &= J^{-n}\left(\frac{\alpha_{\mathrm{s}}}{\alpha_{\mathrm{n}}}\right)^n \frac{\Gamma(\alpha_{\mathrm{n}}+n)/\Gamma(\alpha_{\mathrm{n}})}{\Gamma(\alpha_{\mathrm{s}}+n)/\Gamma(\alpha_{\mathrm{s}})}\left(\frac{\overline{|n(f,\tau)|^2}}{\overline{|s(f,\tau)|^2}}\right)^n, \end{aligned} \tag{88}$$

where α_{s} is the shape parameter of the input speech. Equation (88) 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 Eqs. (81) and (88) with the input speech shape parameter α_{s}, we can obtain the speech kurtosis ratio through BSSA as

$$\begin{aligned}&\text{kurtosis ratio}^{(\mathrm{s})}_{\mathrm{BSSA}}\\&= \frac{\mathcal{M}(\alpha_{\mathrm{s}},\beta C_{\mathrm{BSSA}},4,n)}{\mathcal{M}^2(\alpha_{\mathrm{s}},\beta C_{\mathrm{BSSA}},2,n)}\frac{\alpha_{\mathrm{s}}(\alpha_{\mathrm{s}}+1)}{(\alpha_{\mathrm{s}}+2)(\alpha_{\mathrm{s}}+3)}.\end{aligned} \tag{89}$$

E-II-II Analysis in the case of chBSSA

In chBSSA, since channelwise GSS is performed before DS, C_{BSSA} is therefore replaced with

$$\begin{aligned}C_{\mathrm{chBSSA}} &= (\overline{|n(f,\tau)|^{2n}}/\overline{|s(f,\tau)|^{2n}})\\&= \left(\frac{\alpha_{\mathrm{s}}}{\alpha_{\mathrm{n}}}\right)^n \frac{\Gamma(\alpha_n+n)/\Gamma(\alpha_{\mathrm{n}})}{\Gamma(\alpha_{\mathrm{s}}+n)/\Gamma(\alpha_{\mathrm{s}})}\left(\frac{\overline{|n(f,\tau)|^2}}{\overline{|s(f,\tau)|^2}}\right)^n.\end{aligned} \tag{90}$$

Equation (90) 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 Eqs. (81) and (90) with α_{s}, we can obtain the speech kurtosis ratio through chBSSA as

$$\begin{aligned}&\text{kurtosis ratio}^{(\mathrm{s})}_{\mathrm{chBSSA}}\\&= \frac{\mathcal{M}(\alpha_{\mathrm{s}},\beta C_{\mathrm{chBSSA}},4,n)}{\mathcal{M}^2(\alpha_{\mathrm{s}},\beta C_{\mathrm{chBSSA}},2,n)}\frac{\alpha_{\mathrm{s}}(\alpha_{\mathrm{s}}+1)}{(\alpha_{\mathrm{s}}+2)(\alpha_{\mathrm{s}}+3)}.\end{aligned} \tag{91}$$

E-III 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 noise reduction rate (NRR) [36] (output SNR - input SNR in dB). Figure 30 shows the theoretical behaviors of the noise kurtosis ratio and speech kurtosis ratio. In Figs. 30(a) and 30(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 Figs. 30(c) and 30(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 BSSA and parametric chBSSA, the signal exponent parameter $2n$ is set to 2.0, 1.0, and 0.5.

Figures 30(a) and 30(b) indicate 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, Figs. 30(c) and 30(d) show 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.

F. Time-Variant Nonlinear Noise Estimator

This appendix provides a brief review of the time-variant nonlinear noise estimator. For more detailed information, Ref. [41] can be available.

Let $x_L(f,\tau)$ and $x_R(f,\tau)$ be noisy signals received at the left and right microphones in the time-frequency domain, defined as

$$x_L(f,\tau) = h_L(f)s(f,\tau) + n_L(f,\tau), \tag{92}$$

$$x_R(f,\tau) = h_R(f)s(f,\tau) + n_R(f,\tau), \tag{93}$$

where $h_L(f)$ and $h_R(f)$ are the left and right transfer functions, respectively. Next, the left and right auto-power spectral densities, $\Gamma_{LL}(f)$ and $\Gamma_{RR}(f)$, can be expressed as follows:

$$\Gamma_{LL}(f,\tau) = |H_L(f)|^2\Gamma_{SS}(f,\tau) + \Gamma_{NV}(f,\tau), \tag{94}$$

$$\Gamma_{RR}(f,\tau) = |H_R(f)|^2\Gamma_{SS}(f,\tau) + \Gamma_{CNN}(f,\tau), \tag{95}$$

where $\Gamma_{SS}(f,\tau)$ is the power spectral density of the target speech signal, and $\Gamma_{NT}(f,\tau)$ is the power spectral density of the noise signal. In this paper, we assume that the left and right noise power spectral densities are approximately the same, i.e., $\Gamma_{N_L N_L}(f,\tau) \simeq \Gamma_{N_R N_R}(f,\tau) \simeq \Gamma_{ANN}(f,\tau)$.

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

$$H_W(f,\tau) = \frac{\Gamma_{LR}(f,\tau)}{\Gamma_{RR}(f,\tau)}, \tag{96}$$

where $\Gamma_{LR}(f)$ is the cross-power spectral density between the left and the right noisy signals. The cross-power spectral density expression then becomes

$$\Gamma_{LR}(f,\tau) = \Gamma_{SS}(f,\tau) H_L(f) H_R^*(f). \tag{97}$$

Therefore, substituting (97) into (96) yields

$$H_W(f,\tau) = \frac{\Gamma_{SS}(f,\tau) H_L(f) H_R^*(f)}{\Gamma_{RR}(f,\tau)}. \tag{98}$$

Furthermore, using (94) and (95), the squared magnitude response of the Wiener solution in (98) can be also expressed as

$$|H_W(f,\tau)|^2 = \frac{(\Gamma_{LL}(f,\tau) - \Gamma_{NS}(f,\tau))(\Gamma_{RR}(f,\tau) - \Gamma_{NO}(f,\tau))}{\Gamma_{RR}^2(f,\tau)}. \tag{99}$$

Equation (99) is rearranged into a quadratic equation as in the following:

$$\begin{aligned} &\Gamma_{NR}^2(f,\tau) - \Gamma_{NI}(f,\tau)\left(\Gamma_{LL}(f,\tau) + \Gamma_{RR}(f,\tau)\right) \\ &+ \Gamma_{SEE}(f,\tau)\Gamma_{RR}(f,\tau) = 0, \end{aligned} \tag{100}$$

where

$$\Gamma_{WE}(f,\tau) = \Gamma_{LL}(f,\tau) - \Gamma_{RR}(f,\tau)|H_W(f)|^2. \tag{101}$$

Consequently, the noise power spectral density $\Gamma_{NP}(f)$ can be estimated by solving the quadratic equation in (100) as follows:

$$\Gamma_{NY}(f,\tau) = \frac{1}{2}\left(\Gamma_{LL}(f,\tau) + \Gamma_{RR}(f,\tau)\right) - \Gamma_{LRavg}(f,\tau), \tag{102}$$

$$\begin{aligned} \Gamma_{LRavg}(f,\tau) = \frac{1}{2}\{&(\Gamma_{LL}(f,\tau) + \Gamma_{RR}(f,\tau))^2 \\ &- 4\Gamma_{PEE}(f,\tau)\Gamma_{RR}(f,\tau)\}^{0.5}. \end{aligned} \tag{103}$$

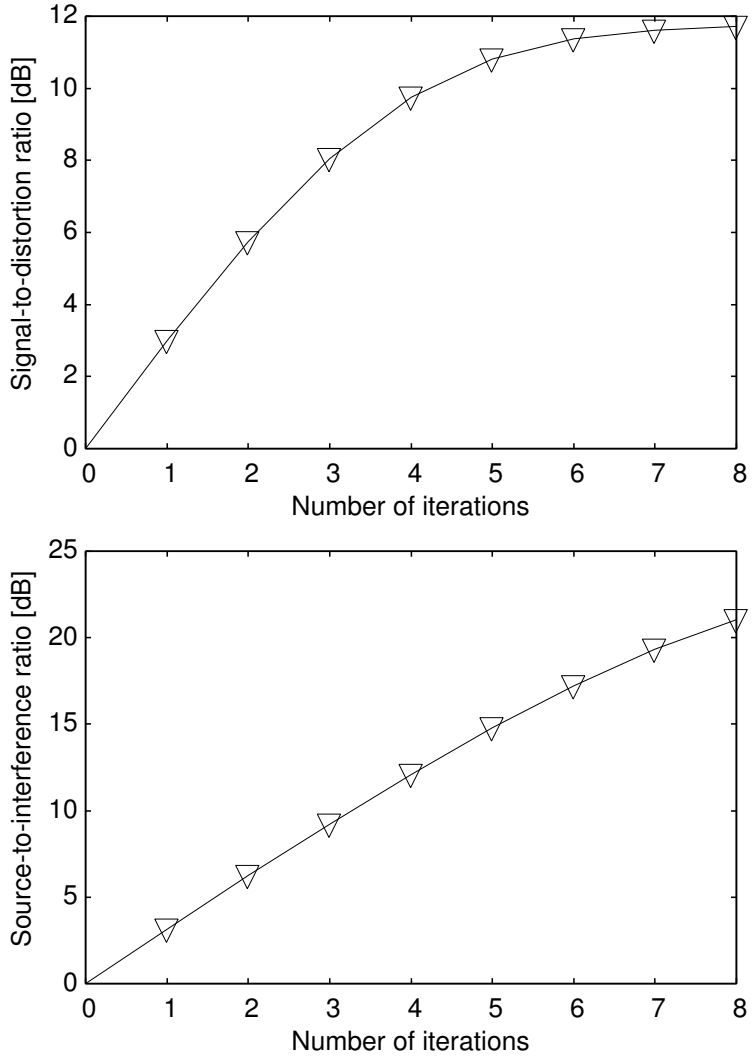

Fig. 28. (a) Relation between number of iterations and SDR and (b) relation between number of iterations and SIR for white Gaussian noise.

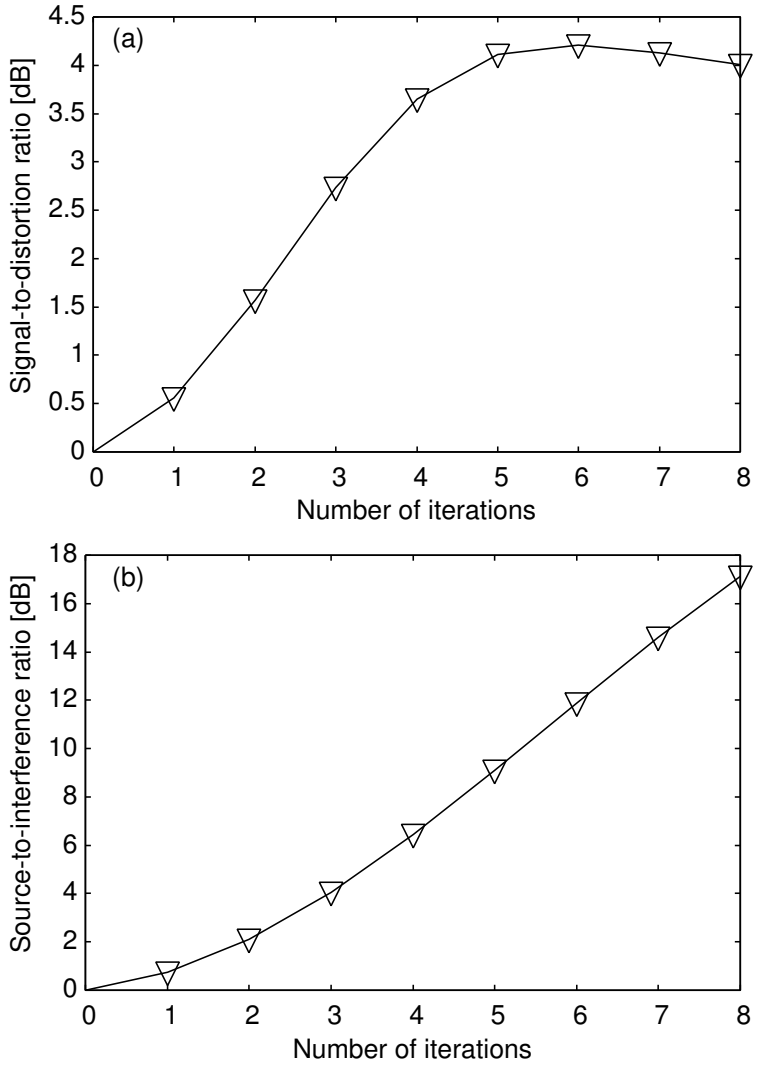

Fig. 29. (a) Relation between number of iterations and SDR and (b) relation between number of iterations and SIR for babble noise.

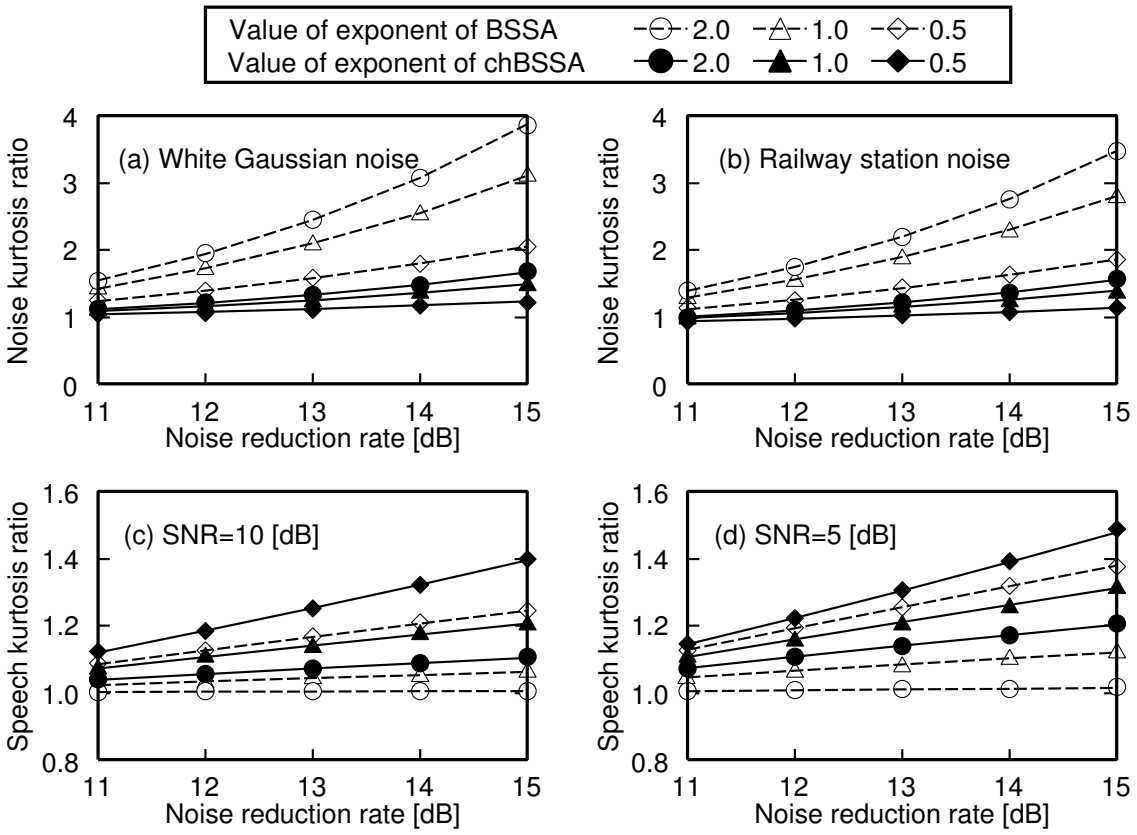

Fig. 30. (a) and (b) are theoretical behaviors of noise kurtosis ratio in structure-generalized parametric BSSA. (a) is for white Gaussian noise and (b) is for railway station noise. (c) and (d) are theoretical behaviors of speech kurtosis ratio in structure-generalized parametric BSSA, where the input SNR is set to 10 and 5 dB, respectively.

List of Publications

Journal Papers

1. Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, "Theoretical analysis of amounts of musical noise and speech distortion in structure-generalized parametric blind spatial subtraction array," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol.95-A, no.2, pp.586–590, February 2012.

2. Ryoichi Miyazaki, Hiroshi Saruwatari, Takayuki Inoue, Yu Takahashi, Kiyohiro Shikano, Kazunobu Kondo, "Musical-noise-free speech enhancement based on optimized iterative spectral subtraction," *IEEE Transactions on Audio, Speech and Language Processing,* vol.20, no.7, pp.2080–2094, September 2012.

3. Ryoichi Miyazaki, Hiroshi Saruwatari, Satoshi Nakamura, Kiyohiro Shikano, Kazunobu Kondo, Jonathan Blanchette, Martin Bouchard, "Musical-noise-free blind speech extraction integrating microphone array and iterative spectral subtraction," *Signal Processing,* (conditionally accepted).

Review Paper

1. Yu Takahashi, Ryoichi Miyazaki, Hiroshi Saruwatari, "An analysis of nonlinear noise reduction techniques based on higher-order statistics and its application," *Journal of Acoustical Society of Japan*, vol.68, no.11, 2012 (in Japanese).

Peer Reviewed International Conference Proceedings

1. Ryoichi Miyazaki, Hiroshi Saruwatari, Ryo Wakisaka, Kiyohiro Shikano, Tomoya Takatani, "Theoretical analysis of parametric blind spatial subtraction array and its application to speech recognition performance prediction," *Proceedings of Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA2011)*, pp.81–86, May 2011.

2. Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, "Theoretical analysis of musical noise and speech distortion in structure-generalized parametric

blind spatial subtraction array," *Proceedings of INTERSPEECH2011*, pp.341–344, August 2011.

3. Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Musical-noise-free speech enhancement: theory and evaluation," *Proceedings of IEEE International Conference in Acoustics, Speech, and Signal Processing (ICASSP2012)*, pp.4565–4568, March 2012.

4. Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Musical-noise-free blind speech extraction using ICA-based noise estimation and iterative spectral subtraction," *Proceedings of 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA2012),* pp.322–327, July 2012.

5. Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Musical-noise-free blind speech extraction using ICA-based noise estimation with channel selection," *Proceedings of International Workshop on Acoustic Signal Enhancement (IWAENC2012),* September 2012.

6. Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Musical-noise-free speech enhancement based on iterative Wiener filtering," *Proceedings of IEEE International Symposium on Signal Processing and Information Technology (ISSPIT2012),* December 2012.

7. Ryoichi Miyazaki, Hiroshi Saruwatari, Satoshi Nakamura, Kiyohiro Shikano, Kazunobu Kondo "Toward musical-noise-free blind speech extraction: concept and its applications," *Proceedings of APSIPA Annual Summit and Conference 2013,* October 2013 (APSIPA ASC 2013 The Best Paper Award).

8. Suzumi Kanehara, Hiroshi Saruwatari, Ryoichi Miyazaki, Kiyohiro Shikano, Kazunobu Kondo, "Theoretical analysis of musical noise generation in noise reduction methods with decision-directed a priori SNR estimator," *International Workshop on Acoustic Signal Enhancement (IWAENC2012)*, September 2012.

9. Yuji Onuma, Noriyoshi Kamado, Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, "Real-time semi-blind speech extraction with motion tracking on

Kinect," *International Workshop on Acoustic Signal Enhancement (IWAENC2012)*, September 2012.

10. Yu Takahashi, Ryoichi Miyazaki, Hiroshi Saruwatari, Kazunobu Kondo, "Theoretical analysis of musical noise in nonlinear noise reduction based on higher-order statistics," *Proceedings of APSIPA Annual Summit and Conference 2012*, December 2012.

11. Suzumi Kanehara, Hiroshi Saruwatari, Ryoichi Miyazaki, Kiyohiro Shikano, Kazunobu Kondo, "Comparative study on various noise reduction methods with decision-directed a priori SNR estimator via higher-order statistics," *Proceedings of APSIPA Annual Summit and Conference 2012*, December 2012.

12. Miyuki Itoi, Ryoichi Miyazaki, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Blind speech extraction for non-audible murmur speech with speaker's movement noise," *Proceedings of IEEE International Symposium on Signal Processing and Information Technology (ISSPIT2012)*, December 2012.

13. Hiroshi Saruwatari, Suzumi Kanehara, Ryoichi Miyazaki, Kiyohiro Shikano, Kazunobu Kondo, "Musical noise analysis for Bayesian minimum mean-square error speech amplitude estimators based on higher-order statistics," *Proceedings of INTERSPEECH2013*, pp.441–445, August 2013.

14. Hiroshi Saruwatari, Ryoichi Miyazaki, "Information-geometric optimization for nonlinear noise reduction systems," *Proceedings of Intelligent Signal Processing and Communication Systems*, November 2013.

15. Shunsuke Nakai, Ryoichi Miyazaki, Hiroshi Saruwatari, Satoshi Nakamura, "Theoretical analysis of musical noise generation for blind speech extraction with generalized MMSE short-time spectral amplitude estimator," *Proceedings of Intelligent Signal Processing Conference*, December 2013.

16. Shunsuke Nakai, Ryoichi Miyazaki, Hiromichi Kawanami, Hiroshi Saruwatari, Satoshi Nakamura, Kazunobu Kondo, "Semi-blind multiple speech extraction under diffuse noise condition for smart posterboard," *RISP 2014 International Workshop on Nonlinear Circuits, Communication and Signal Processing*, February 2014 (accepted).

17. Yuka Hirano, Ryoichi Miyazaki, Hiromichi Kawanami, Hiroshi Saruwatari, Satoshi Nakamura, Tomoya Takatani, "Control method of the number of iterations based on speech kurtosis ratio in musical-noise-free blind speech extraction," *RISP 2014 International Workshop on Nonlinear Circuits, Communication and Signal Processing*, February 2014.

Technical Reports

1. Ryoichi Miyazaki, Takayuki Inoue, Nobuhisa Hirata, Hiroshi Saruwatari, Kiyohiro Shikano, Tomoya Takatani, "Relation between musical noise generation in nonlinear signal processing and speech recognition performance," *IEICE Technical Report*, SP2010-106, pp.19–24, January 2011 (in Japanese).

2. Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, "Mathematical metric of speech distortion in various types of BSSA," *IEICE Technical Report*, SP2011-9 (EA2011-9, SIP2011-9), pp.49–54, May 2011.

3. Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Generalized theory of musical-noise-free noise reduction and its application to blind signal extraction," *26th SIP Symposium*, pp.436–441, November 2012 (in Japanese).

4. Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Evaluation of musical-noise-free noise reduction under real acoustic environments," *IEICE Technical Report*, EA2011-108, pp.25–30, January 2012.

5. Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Iterative blind spatial subtraction array for musical-noise-free speech enhancement in diffuse noise," *IEICE Technical Report*, EA2011-125, pp.31–36, March 2012.

6. Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Evaluation of musical-noise-free noise reduction methods based on various dynamic noise estimators," *27th SIP Symposium*, pp.436–441, November 2012 (in Japanese).

7. Ryoichi Miyazaki, Hiroshi Saruwatari, Satoshi Nakamura, Kiyohiro Shikano, Kazunobu Kondo, "Evaluation of musical-noise-free blind speech extraction under nonstationary noise environment," *IEICE Technical Report*, EA2013-51, pp.105–110, July 2013.

8. Kodai Okamoto, Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, "Evaluation of blind speech extraction for speech archiving of poster session," *IEICE Technical Report*, EA2011-108, pp.19–24, January 2012 (in Japanese).

9. Hiroshi Saruwatari, Ryoichi Miyazaki, Kiyohiro Shikano, "Application of higher-order statistics in speech enhancement," *IEICE Technical Report*, SP2012-21 (EA2011-21, SIP2011-21), pp.121–126, May 2012.

10. Miyuki Itoi, Ryoichi Miyazaki, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano, "Blind speech extraction for non-audible murmur speech with speaker's movement noise," *IEICE Technical Report*, EA2012-40, pp.43–48, June 2012 (in Japanese).

11. Suzumi Kanehara, Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Mathematical metric of musical noise for various nonlinear speech enhancement algorithms," *IEICE Technical Report*, EA2012-44, pp.67–72, June 2012.

12. Suzumi Kanehara, Hiroshi Saruwatari, Ryoichi Miyazaki, Kiyohiro Shikano, Kazunobu Kondo, "Analysis of musical noise generation for various nonlinear speech enhancement algorithms based on approximated model" *27th SIP Symposium*, pp.430–435, November 2012 (in Japanese).

13. Miyuki Itoi, Ryoichi Miyazaki, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano, "Blind speech extraction based on multichannel diverse sensing for non-audible murmur speech with speaker's movement noise," *IEICE Technical Report*, EA2012-119, pp.1–6, January 2013 (in Japanese).

14. Shunsuke Nakai, Suzumi Kanehara, Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, "Analysis on musical noise generation for blind speech extraction based on generalized MMSE short-time spectral amplitude estimator," *IEICE*

Technical Report, SP2013-13 (EA2013-13, SIP2013-13), pp.73–78, May 2013 (in Japanese).

15. Shunsuke Nakai, Ryoichi Miyazaki, Hiroshi Saruwatari, Satoshi Nakamura, Kazunobu Kondo, "Musical-noise-free speech enhancement based on generalized MMSE short-time spectral amplitude estimator with biased a priori SNR estimation," *28th SIP Symposium*, pp.342–347, November 2013 (in Japanese).

16. Shunsuke Nakai, Ryoichi Miyazaki, Hiroshi Saruwatari, Satoshi Nakamura, Kazunobu Kondo, "Study on biased a priori SNR estimation for musical-noise-free MMSE short-time spectral amplitude estimator," *IEICE Technical Report*, EA2012-118, pp.87–92, January 2013 (in Japanese).

17. Yuka Hirano, Ryoichi Miyazaki, Hiroshi Saruwatari, Satoshi Nakamura, Tomoya Takatani, "Study on speech recognition performance prediction using unsupervised measurement for speech quality evaluation based on higher-order statistics," *IEICE Technical Report*, EA2013-119, pp.93–98, January 2013 (in Japanese).

Domestic Meetings

1. Ryoichi Miyazaki, Takayuki Inoue, Nobuhisa Hirata, Hiroshi Saruwatari, Kiyohiro Shikano, Tomoya Takatani, "Relation between musical noise generation and speech recognition performance in generalized SS," *The young researchers meetings of ASJ Kansai-section*, B-11, December 2010 (in Japanese).

2. Ryoichi Miyazaki, Takayuki Inoue, Nobuhisa Hirata, Hiroshi Saruwatari, Kiyohiro Shikano, Tomoya Takatani, "Relation between musical noise generation and speech recognition performance in nonlinear noise reduction based on noise estimation," *The Meeting of ASJ*, 1-9-14, pp.547–650, March 2011 (in Japanese).

3. Ryoichi Miyazaki, Hiroshi Saruwatari, Takayuki Inoue, Kiyohiro Shikano, Kazunobu Kondo, "Musical-noise-free speech enhancement: its theory and evaluation," *The Meeting of ASJ*, 1-4-2, pp.601–604, September 2011 (in Japanese).

4. Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Evaluation of musical-noise-free noise reduction in actual environment," *The*

young researchers meetings of ASJ Kansai-section, p.29, December 2011 (in Japanese).

5. Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Evaluation of speech distortion in musical-noise-free noise reduction method," *The Meeting of ASJ*, 3-1-22, pp.805–808, March 2012 (in Japanese).

6. Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Musical-noise-free blind speech extraction using ICA-based noise estimation with channel selection," *The Meeting of ASJ*, 3-9-9, pp.691–694, September 2012 (in Japanese).

7. Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Investigation of noise estimation accuracy of signal after nonlinear signal processing," *The Meeting of ASJ*, 1-P-45, pp.927–930, March 2013 (in Japanese).

8. Ryoichi Miyazaki, Hiroshi Saruwatari, Satoshi Nakamura, Kazunobu Kondo, "Evaluation of various musical-noise-free speech enhancement methods," *The Meeting of ASJ*, 2-1-2, pp.619–622, September 2013 (in Japanese).

9. Kodai Okamoto, Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, "Blind Source Separation for Speech Archiving of Poster Session," *The Meeting of ASJ*, 2-6-11, pp.669–672, September 2011 (in Japanese).

10. Kodai Okamoto, Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, "Blind source separation for speech archiving of poster session," *The young researchers meetings of ASJ Kansai-section*, p.19, December 2011 (in Japanese).

11. Kodai Okamoto, Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, "Blind source separation and diarization for speech archiving of poster session," *The Meeting of ASJ*, 1-1-17, pp.707–710, March 2012 (in Japanese).

12. Suzumi Kanehara, Hiroshi Saruwatari, Ryoichi Miyazaki, Kiyohiro Shikano, Kazunobu Kondo, "Mathematical analysis of musical noise for speech enhancement algorithms with decision-directed a priori SNR estimator,' *The Meeting of ASJ*, 2-9-3, pp.633–636, September 2012 (in Japanese).

13. Miyuki Itoi, Ryoichi Miyazaki, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano, "Blind speech extraction for non-audible murmur speech recorded by various microphones," *The Meeting of ASJ*, 3-9-10, pp.695–698, September 2012 (in Japanese).

14. Yuji Onuma, Shunsuke Nakai, Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, "Voice activity detection for the construction of multi-modal archive system of a poster session presentation," *The Meeting of ASJ*, 1-P-35, pp.901–904, March 2013 (in Japanese).

15. Miyuki Itoi, Ryoichi Miyazaki, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano, "Evaluation of blind speech extraction based on multichannel diverse sensing for non-audible murmur speech with speakerfs movement noise,," *The Meeting of ASJ*, 2-10-2, pp.725–728, March 2013 (in Japanese).

16. Shunsuke Nakai, Yuji Onuma, Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, "Improvement of blind source separation performance for speech archiving of poster session," *The Meeting of ASJ*, 2-10-3, pp.729–732, March 2013 (in Japanese).

17. Suzumi Kanehara, Hiroshi Saruwatari, Ryoichi Miyazaki, Kiyohiro Shikano, Kazunobu Kondo, "Study on validity of mathematical analysis for amount of musical noise generation in various speech enhancement methods," *The Meeting of ASJ*, 3-10-15, pp.789–792, March 2013 (in Japanese).

18. Shunsuke Nakai, Ryoichi Miyazaki, Hiroshi Saruwatari, Satoshi Nakamura, Kazunobu Kondo, "Theoretical analysis on biased MMSE short-time spectral amplitude and its extension to musical-noise-free speech enhancement," *The Meeting of ASJ*, 2-1-1, pp.615–618, September 2013 (in Japanese).

19. Yuka Hirano, Ryoichi Miyazaki, Hiroshi Saruwatari, Satoshi Nakamura, Tomoya Takatani, "Speech kurtosis estimation in blind speech extraction for speech recognition performance prediction," *The Meeting of ASJ*, 2-1-6, pp.635–638, September 2013 (in Japanese).

20. Yuko Wakabayashi, Koji Inoue, Tatsuya Kawahara, Shunsuke Nakai, Ryoichi Miyazaki, Hiroshi Saruwatari, Satoshi Nakamura, "Speaker diarization based on audio-

visual integration for smart posterboard," *The Meeting of ASJ*, 2-Q4-7, March 2014 (in Japanese).

21. Shunsuke Nakai, Ryoichi Miyazaki, Hiroshi Saruwatari, Satoshi Nakamura, Koji Inoue, Yuko Wakabayashi, Tatsuya Kawahara, ""Multiple source separation for smart posterboard in real environment," *The Meeting of ASJ*, 2-Q4-8, March 2014 (in Japanese).

22. Yuka Hirano, Ryoichi Miyazaki, Hiroshi Saruwatari, Satoshi Nakamura, Tomoya Takatani, "Unsupervised control of speech quality based on higher-order statistics in musical-noise-free blind speech extraction," *The Meeting of ASJ*, 2-1-3, September 2013 (in Japanese).

Awards

1. Student Award of Acoustical Society of Japan, Ryoichi Miyazaki, Hiroshi Saruwatari, Takayuki Inoue, Kiyohiro Shikano, Kazunobu Kondo, September, 2011.

2. SIP Student Award of Signal Processing Symposium, Ryoichi Miyazaki, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, November, 2011.

3. Ericsson Best Student Award 2012, Ryoichi Miyazaki, Hiroshi Saruwatari, Takayuki Inoue, Yu Takahashi, Kiyohiro Shikano, Kazunobu Kondo, November 2012.

4. IEEE Signal Processing Society Japan Chapter Student Paper Award, Ryoichi Miyazaki, Hiroshi Saruwatari, Takayuki Inoue, Yu Takahashi, Kiyohiro Shikano, Kazunobu Kondo, November 2012.

5. TELECOM System Technology Student Award of the Telecommunications Advancement Foundation, Ryoichi Miyazaki, Hiroshi Saruwatari, Takayuki Inoue, Yu Takahashi, Kiyohiro Shikano, Kazunobu Kondo, March 2013.

6. APSIPA ASC 2013 Best Paper Award, Ryoichi Miyazaki, Hiroshi Saruwatari, Satoshi Nakamura, Kiyohiro Shikano, Kazunobu Kondo, Jonathan Blanchette, Martin Bouchard, October 2013.