NAIST-IS-DD0761017

Doctoral Dissertation

Blind Speech Enhancement with Independent Component Analysis and Spectral Subtraction

Yu Takahashi

March 24, 2010

Department of Information Processing
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

Yu Takahashi

Thesis Committee:

Professor Kiyohiro Shikano	(Supervisor)
Professor Tsukasa Ogasawara	(Co-supervisor)
Associate Professor Kiyohiro Shikano	(Co-supervisor)

Blind Speech Enhancement with Independent Component Analysis and Spectral Subtraction[†]

Yu Takahashi

Abstract

A hands-free speech recognition system and a hands-free telecommunication system are essential for realizing an intuitive, unconstrained, and stress free human-machine interface. In an actual acoustic environment, however, not only user's speech but also interference source signals such as background noise and interference speech are existing. Such interferences disturb high-quality speech recognition or telecommunication. Therefore, a source extraction method is needed to realize high-quality hands-free systems. Particularly, blind source extraction methods are spotlighted. Since blind source extraction does not require any supervision, it can be applied to wide-area applications.

Independent component analysis (ICA) is a successful candidate of blind source extraction methods. There have been many studies on ICA, and they have provided strong evidences that ICA can extract blindly source signals from noisy observations. However, almost all studies on ICA only treat the limited case, i.e., all sound sources are point source like speech. Such an acoustic condition is very unrealistic; interferences are often widespread in an actual world.

In the thesis, I mainly deal with generalized noise that cannot be regarded as a point source. Then, first, I analyze ICA under a non-point-source noise condition, and theoretically point out that ICA is proficient in noise estimation rather than in speech estimation under such a noise condition. Namely, we cannot utilize ICA as a target speech estimator. However, we can still use ICA as an accurate noise

[†] Doctoral Dissertation, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0761017, March 24, 2010.

estimator. Based on the above-mentioned findings, I propose a new blind source extraction architecture, i.e., blind spatial subtraction array (BSSA). The proposed BSSA comprises an ICA-based noise estimator, and noise reduction is carried out by subtracting the power spectrum of the estimated noise via ICA from the power spectrum of the partly-speech-enhanced signal by microphone array technique. This "power-spectrum-domain subtraction" procedure accomplishes better noise reduction than the conventional ICA.

Furthermore, the proposed BSSA provides robustness against the permutation problem inherent in ICA. The frequency-domain ICA often causes source permutation ambiguity problem in each frequency bin, and the permutation problem markedly degrades the resultant signal quality. Therefore, it is indispensable for us to align the permutation problem so that each extracted signal contains frequency components from the same source. Indeed the proposed BSSA partially involves the permutation problem in the ICA-based noise estimator part. However, the proposed BSSA can efficiently reduce the negative affection of the permutation owing to the over-subtraction in the spectral subtraction and defocusing properties in the speech enhancement part. In addition, the proposed BSSA has a remarkable property that is the robustness against reverberation and microphone element errors. This fact is given by an alternative interpretation of the proposed BSSA.

These effectiveness of the proposed BSSA are shown several experiments. First, I gives an evidence of permutation robustness of the proposed BSSA in an artificial computer simulation. Next, I conduct experiments in a experimental room and an actual rail-way station. As a result of the experiments, it can be confirmed that the noise reduction and speech recognition performance of the proposed BSSA outperforms those of the conventional ICA. From these results, I conclude that the proposed BSSA is well applicable to the noise-robust hands-free system.

Next, I propose the real-time algorithm of the proposed BSSA. As for hand-free speech recognition system and telecommunication system, "real-time" property is a crucial factor. Indeed the proposed BSSA can reduce noises efficiently, BSSA is difficult to work in real-time because ICA-based noise estimation part consumes huge amount of computational complexities. Therefore, I take a strat-

egy in that the separation filter optimized by using the past time period data is applied to the current data. Although the separation filter update in the ICA part is not real-time processing but involves some latency, the entire system still seems to run in real-time because the other parts of BSSA can work in the current segment with no delay. Based on the real-time BSSA algorithm, I develop a hands-free spoken-oriented guidance system. The developed system can realize enough speech recognition performance, over 80% word correct, and low-latency, particularly about 50 ms, blind source extraction.

Next, I focus my attention to “musical-noise problem.” Musical noise is an artificially generated noise through nonlinear signal processing, and makes users uncomfortable. Unfortunately, the proposed BSSA suffers from the musical-noise problem because the proposed BSSA includes nonlinear spectral subtraction in its own structure. In the thesis, I analyze how much musical noise are generated through methods of integrating microphone array signal processing and spectral subtraction like the proposed BSSA on the basis of higher-order statistics. As a result of the analysis, I clarify that the specific integration structure can mitigate the musical-noise generation. The validity of the analysis is demonstrated via computer simulations and subjective listening tests.

Keywords:

Speech enhancement, blind source separation, independent component analysis, spectral subtraction, musical noise, higher-order statistics

Contents

Chapter 1
Prologue

Chapter 2
Data Model and Conventional BSS Methods with ICA

Chapter 3
Analysis of ICA under Non-Point-Source Noise Condition

Chapter 4
Blind Speech Extraction Method with ICA-based Noise Estimator

Chapter 5
Robustness against Reverberation and Microphone Element Errors in BSSA

Chapter 6

Real-time Implementation of Proposed BSSA for Hands-Free Spoken-Oriented Guidance System

Chapter 7

Musical Noise and Its Objective Measure

Chapter 8

Kurtosis-Based Musical-Noise Analysis for Microphone Array Signal Processing and SS

Chapter 9
Epilogue

Appendix

List of Figures

List of Tables

Chapter 1

Prologue

1.1. Background

These days, hands-free speech recognition systems (see Fig. 1), e.g., human-robot speech interaction system [1, 2, 3], and hands-free telecommunication systems [4] are in demand because such systems are essential for the realization of an intuitive, unconstrained, and stress-free human-machine interface. In such hands-free systems, however, not only the user's speech but also interference sounds such as background noise and interference speech are observed by the microphones in the systems. Thus, it is difficult to achieve high-quality speech recognition or telecommunication systems compared with the case of using a close-talking microphone (see Fig. 2) such as a headset microphone or a hand microphone. Therefore, interference sounds must be suppressed to realize a noise-robust hands-free system.

In order to remove interference sound sources, there have been many studies on source separation. Source separation for acoustic signals is the estimation of the original sound source signals from the mixed signals observed in each input channel. Various methods have been presented for acoustic source signal separation, which can be classified into two groups: methods based on single-channel input, e.g., spectral subtraction (SS) [5], and those based on multichannel input, e.g., microphone array signal processing [6]. There have been various studies on microphone array signal processing; in particular, the delay-and-sum (DS) [7, 8, 9] array and the adaptive beamformer (ABF) [10, 11, 12] are the most commonly used microphone arrays for source separation and noise reduction. The ABF can achieve higher performance than the DS array. However, the ABF requires *a priori* information, e.g., the look direction and speech break interval. These requirements are due to the fact that the conventional ABF is based on *supervised* adaptive filtering, which significantly limits its applicability of ABF to source separation in practical applications. Indeed, the ABF cannot work well when the interfering signal is nonstationary noise.

Recently, alternative approaches to acoustic source signal separation have been proposed. Blind source separation (BSS) is an approach to estimating original source signals using only the mixed signals observed in each input channel. In particular, BSS based on independent component analysis (ICA) [13], in which the independence among source signals is mainly used for the separation, has been studied actively [14, 15, 16, 17, 18, 19, 20, 21, 22]. Indeed, conventional ICA can work particularly in the case of speech-speech mixing, i.e., all sources can be regarded as point sources, but such a mixing condition is very rare and unrealistic; real noises are often *widespread* sources.

Furthermore, many methods of integrating microphone array signal processing and nonlinear signal processing such as SS have been studied with the aim of achieving better noise reduction [23, 24, 25, 26, 27, 28]. It has been well demonstrated that such integration methods can achieve higher noise reduction performance than that obtained using conventional adaptive microphone arrays [27], e.g., the Griffith-Jim array [11]. However, a serious problem exists in such methods: artificial distortion (so-called *musical noise* [29]) due to nonlinear signal processing. Since the artificial distortion causes discomfort to users, it is desirable that musical noise is controlled through signal processing. However, in almost all nonlinear noise reduction methods, the strength parameter to mitigate musical noise in nonlinear signal processing is determined heuristically. Although there have been some studies on reducing musical noise [29] and on nonlinear signal processing with less musical noise [30], evaluations have mainly depended on subjective tests by humans, and no objective evaluations have been performed to the best of my knowledge.

1.2. Scope of thesis

The aim of this study is to establish a blind source extraction method with the following three aspects.

- **Good source extraction performance in the real world:**
  The conventional ICA-based BSS cannot treat a realistic acoustical condi-

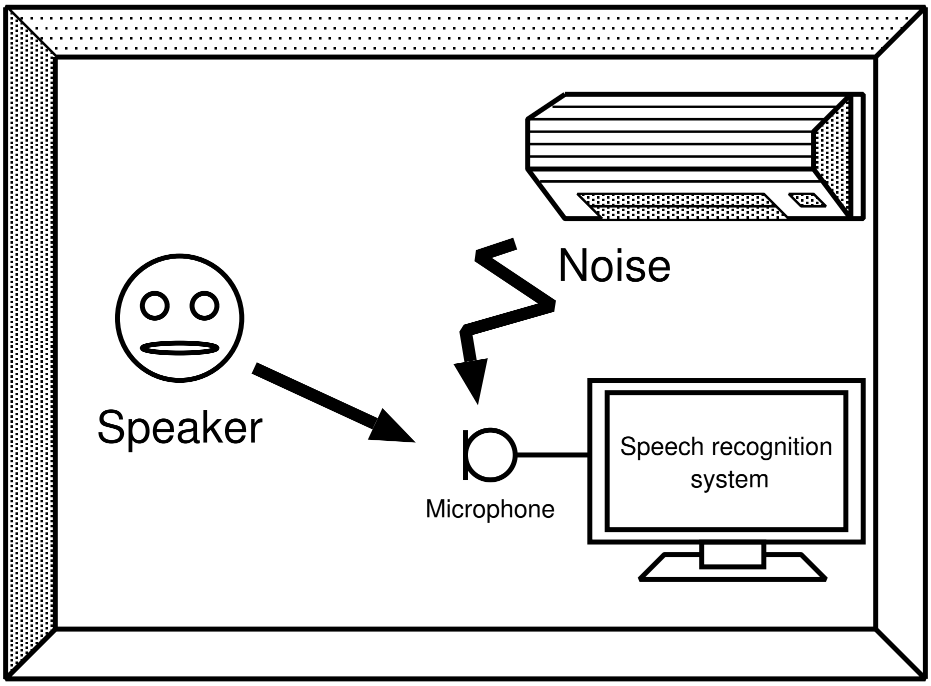


Figure 1. Configuration of hands-free speech recognition system.

tion that involves widespread noise. Thus, it is desirable that a blind source extraction method can deal with widespread noise. Hence, I propose a blind source extraction method that is a combination of conventional ICA and nonlinear SS in this study. The proposed blind source extraction method can handle widespread noise.

- **Real-time processing:**
  As for hand-free speech recognition and telecommunication systems, a real-time property is a crucial factor. In this study, I construct a real-time architecture for the proposed blind source extraction method that can be applied to real-world source extraction problem.

- **Good sound quality for human hearing:**
  In applications involving human hearing such as mobile phones, teleconference systems and a hearing-aid systems, the sound quality of the output is extremely important. In particular, an artificial distortion such as musical

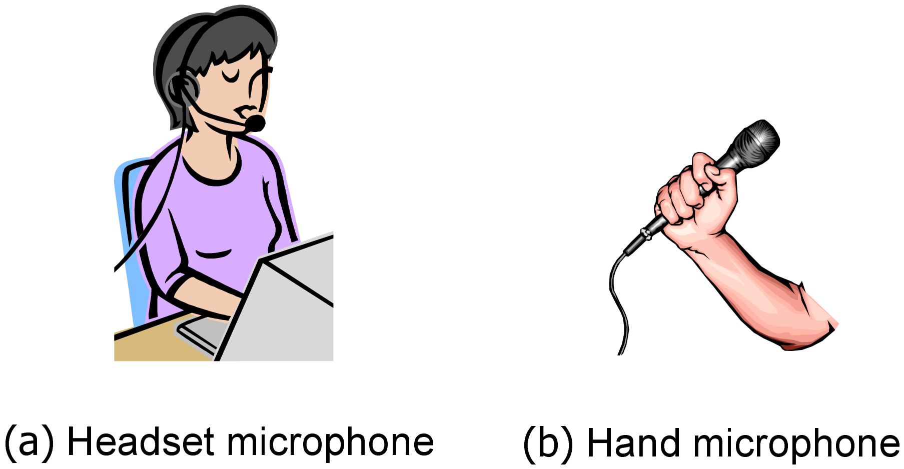


Figure 2. Example of close-talking microphones: (a) a headset microphone and (b) a hand microphone.

> noise originating from nonlinear signal processing is a critical weak point in such applications. Therefore, I analyze the generation of musical noise in methods of integrating microphone array signal processing and nonlinear signal processing, and clarify the type of integration structure that is suitable for human hearing.

In the following sections, I describe my approach to each aspect.

1.2.1 Approach to blind source extraction method in real world

Problems of ICA-based BSS in real world

Although conventional ICA-based BSS techniques can separate acoustic sound sources in the particular case that all sources can be approximated as point sources, such an acoustic condition is very rare and unrealistic. In actual environments, not only a point-source interference source signal but also non-point-source noise (widespread noise) often exists.

In this study, I mainly deal with generalized noise that cannot be regarded as a point source. Moreover, I assume this noise to be nonstationary noise that arises in many acoustical environments; however, ABF cannot treat this noise well.

Although ICA is not affected by nonstationarity of signals unlike ABF, the assumed noise environment is still a very challenging task that conventional ICA-based BSS cannot effectively address because ICA cannot separate widespread sources. In order to improve the performance of BSS, some techniques combining conventional ICA and beamforming have been proposed [31, 21]. However, these studies still deal with the separation of point sources, and the behavior of such methods under a non-point-source condition has not been explicitly analyzed to the best of my knowledge.

Approach

In this study, I first analyze ICA under a non-point-source noise condition and theoretically point out that ICA is effective for noise estimation rather than for speech estimation under such a noise condition. This analysis implies that we can still utilize ICA as an accurate noise estimator.

Next, I propose a new blind spatial subtraction array (BSSA). The proposed BSSA consists of an ICA-based noise estimator, and noise reduction in the proposed BSSA is achieved by subtracting the power spectrum of the noise estimated via ICA from the power spectrum of the noisy observations. This "power-spectrum-domain subtraction" procedure provides better noise reduction than conventional ICA with estimation-error robustness.

Another advantage of the proposed BSSA architecture is "*permutation robustness*." In frequency-domain ICA, a source permutation ambiguity arises in each frequency bin and markedly decreases the resultant quality. Therefore, it is indispensable to align the permutations so that each separated signal contains frequency components from the same source. Although various permutation solvers, e.g., spectral-continuity-based methods [16, 32], methods based on direction of arrival (DOA) [19, 33], and the method integrating spectral continuity and DOA [34], have been proposed, the permutation problem cannot be solved completely.

In addition, an increase in the permutation-salvaging accuracy requires an increase in computational cost. Permutation robustness indicates how little the BSS method is affected for a certain probability of a permutation arising, and such an important property has not yet been investigated in ICA studies. Note that permutation robustness in the BSSA does not conflict with any permutation solver. That is to say, all permutation solvers can be used in the ICA part of the BSSA. The BSSA reduces the number of remaining permuted components that cannot be solved by a permutation solver.

1.2.2 Approach to realizing real-time processing

Problems of real-time processing of proposed method

Although BSSA can reduce noises efficiently, it is difficult to operate in real-time because the ICA part of the BSSA requires a huge amount of computations. Thus, it is necessary to develop a real-time architecture for the BSSA.

Approach

In the proposed BSSA, it is toilsome to optimize the separation filter by ICA in real-time. In other words, the other parts of the BSSA operate in real-time. Therefore, I introduce a strategy in which the separation filter optimized using the data of the previous time period is applied to the current data. Although the update of the separation filter in the ICA part is not real-time processing and involves some latency, the entire system still appears to run in real-time because the other parts of the BSSA can operate in the current segment with no delay. In the system, the performance degradation due to the latency problem in ICA is mitigated by oversubtraction in the spectral subtraction.

1.2.3 Approach to obtaining good sound quality for human hearing

Problems of nonlinear signal processing

Although nonlinear signal processing such as by SS is a powerful noise reduction technique, it generates an artificial musical noise. It is desirable to reduce or control the amount of musical noise because it is unpleasant for users. Unfortunately, the proposed BSSA also suffers from the problem of musical noise because it involves SS in its structure. However, in almost all nonlinear signal processing methods, the strength parameter in the processing is determined heuristically to mitigate musical noise. This is because there are no objective criteria to measure the amount of musical noise generated via nonlinear processing.

Approach

Recently, it was reported that the amount of generated musical noise is strongly related to the difference between higher-order statistics (HOS) before and after nonlinear signal processing [35]. Moreover, an objective metric for the amount of musical noise generated has been established [35]. Furthermore, a detailed analysis of the amount of musical noise generated through SS has been given and the features of musical-noise generation in SS have been clarified [35].

In this study, I perform a musical-noise analysis on methods of integrating microphone array signal processing and SS on the basis of HOS, and I reveal that a specific integration structure can mitigate the amount of musical noise generated.

1.3. Overview of thesis

The thesis is organized as follows.

First, the sound-mixing model used in this study is described in Chapter 2. In this chapter, conventional ICA is also explained.

In Chapter 3, a theoretical investigation of ICA under non-point-source noise condition is presented. As a result of the investigation, I reveal that conventional ICA is proficient in noise estimation under a non-point-source noise condition. Moreover, a computer simulation result that supports this result is also demonstrated.

On the basis of the above-mentioned findings, I propose a novel blind source extraction method, i.e., the BSSA, in Chapter 4. In this chapter, I discuss signal processing on BSSA in detail and analyze its permutation robustness. Moreover, I provide strong evidence of the efficacy of the proposed BSSA via experimental results in not only an experimental room but also a real-world scenario.

Next, I give an alternative explanation of the proposed BSSA in Chapter 5. In this chapter, I first introduce the spatial subtraction array (SSA), which is a method of nonlinear microphone array signal processing and has a similar structure to the proposed BSSA. Next, I describe the problem of the SSA, and then perform the alternative analysis of the noise estimation part of the proposed BSSA by comparing it with the noise estimation part in the SSA. As a result of the analysis, I reveal that the proposed BSSA is robust against reverberation and microphone element errors.

In Chapter 6, I establish a real-time algorithm for the proposed BSSA, and construct a hands-free spoken-oriented guidance system using the real-time BSSA. Moreover, a result of the speech recognition test of the proposed real-time BSSA is also given.

In Chapter 7, the preliminaries to the musical-noise analysis in Chapter 8 are presented. First, I give formulations for two typical methods of integrating a microphone array and SS. Second, the objective metric for musical noise on the basis of HOS proposed in [35] is described.

In Chapter 8, HOS-based musical-noise analysis is carried out, and I reveal that a specific integration structure is preferable for applications involving human hearing. Moreover, several simulation results and the result of subjective listening test are illustrated in the chapter.

Finally, Chapter 9 concludes this thesis and clarifies remaining open problems.

Chapter 2

Data Model and Conventional BSS Methods with ICA

2.1. Introduction

In this chapter, I describe data model of speech enhancement problem in this study and conventional BSS methods with ICA applied to acoustical source separation problems. In recently years, many types of ICA-based BSS methods have been researched. Then, I review two typical ICA algorithms, second-order statistics-based ICA (SO-ICA) and higher-order statistics-based ICA (HO-ICA), in this chapter.

The chapter is organized as follows. Firstly, the sound mixing model to define the speech enhancement problem in Sect. 2.2. Next, I review the two types of ICA-based BSS methods in Sect. 2.3.

2.2. Sound mixing model

In this section, I represent the sound mixing model. I treat the convolutive sound mixing model which takes account into a time delay and a room reverberation.

In this study, a straight-line array is assumed. The coordinates of the elements are designated $d_j (j = 1, \ldots, J)$, and the DOAs of multiple sound sources are designated $\theta_k (k = 1, \ldots, K)$ (see Fig. 3). Then, I consider that only one target speech signal, some interference signals that can be regarded as point sources, and additive noise exist. This additive noise represents noises that cannot be regarded as point sources, e.g., spatially uncorrelated noises, background noises, and leakage of reverberation components outside the frame analysis. Multiple mixed signals are observed at microphone array elements, and a short-time analysis of the observed signals is conducted by frame-by-frame discrete Fourier transform (DFT).

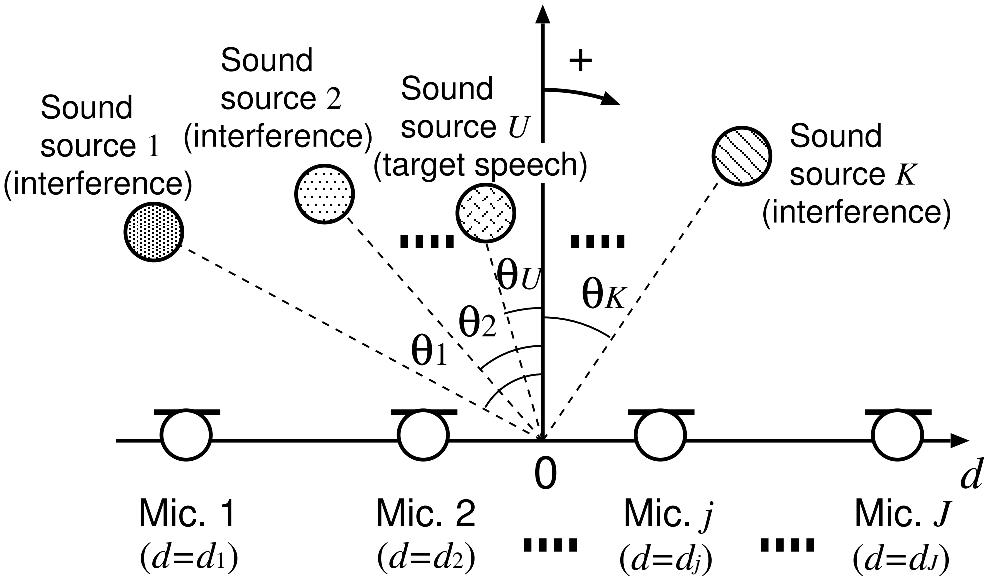

Sound source 1 (interference)
Sound source 2 (interference)
Sound source U (target speech)
+
Sound source K (interference)
θU
θK
θ2
θ1
0
d
Mic. 1 (d=d1)
Mic. 2 (d=d2)
Mic. j (d=dj)
Mic. J (d=dJ)

Figure 3. Configurations of microphone array and signals.

The observed signals are given by

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

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

These are defined as

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

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

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

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

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

2.3. Conventional BSS methods with frequency-domain ICA

In this section, I describe the BSS methods using ICA. In ICA algorithm, it is assumed that source signals are mutually independent, and an appropriate separation filter is optimized so that output signals are mutually independent. Indeed there exists many types of ICA, the filter is optimized by various iterative or non-iterative approaches. In this section, I review two typical types of ICA, SO-ICA and HO-ICA, on frequency-domain.

2.3.1 Demixing process

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

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

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

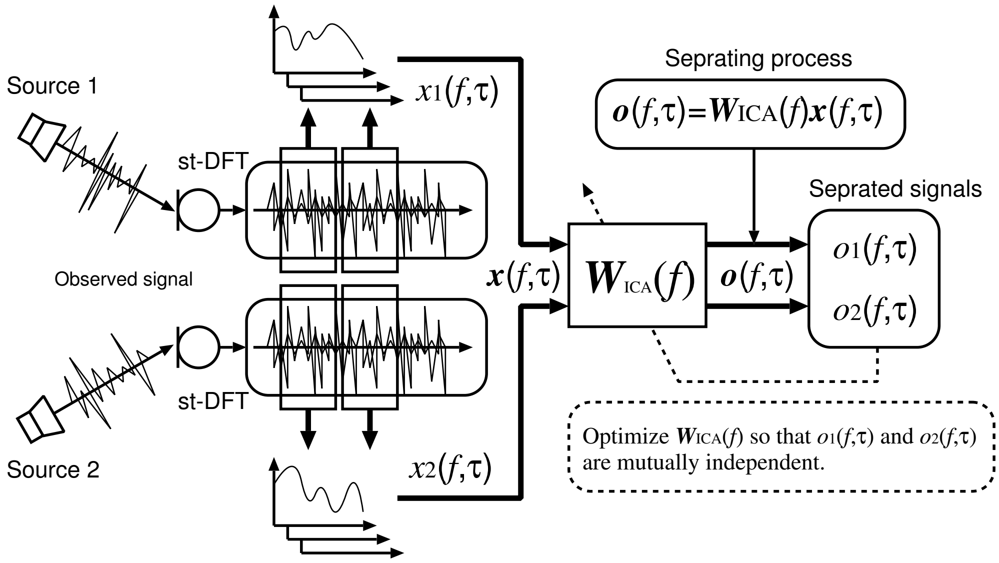

Source 1
st-DFT
$x_1(f,\tau)$
Seprating process
$\boldsymbol{o}(f,\tau)=\boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{x}(f,\tau)$
Seprated signals
$o_1(f,\tau)$
$o_2(f,\tau)$
Observed signal
$\boldsymbol{x}(f,\tau)$
$\boldsymbol{W}_{\mathrm{ICA}}(f)$
$\boldsymbol{o}(f,\tau)$
st-DFT
Optimize $\boldsymbol{W}_{\mathrm{ICA}}(f)$ so that $o_1(f,\tau)$ and $o_2(f,\tau)$ are mutually independent.
Source 2
$x_2(f,\tau)$

Figure 4. Blind source separation procedure in FDICA in case of $J = K = 2$.

where $\boldsymbol{o}(f,\tau)$ is the resultant output of the separation, and $\boldsymbol{W}_{\mathrm{ICA}}(f)$ is the complex-valued unmixing matrix (see Fig. 4).

2.3.2 Optimization of unmixing matrix

The unmixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$ is optimized by ICA so that the output entries of $\boldsymbol{o}(f,\tau)$ become mutually independent. Indeed, many kinds of ICA algorithms have been proposed. In SO-ICA [18, 20], the separation filter is optimized by joint diagonalization of co-spectra matrices using nonstationarity and coloration of the signal. For instance, the following iterative updating equation based on SO-ICA has proposed by Parra[18]:

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

where μ is the step-size parameter, $[p]$ is used to express the value of the pth step in iterations, off-diag$[\boldsymbol{X}]$ is the operation for setting every diagonal element of the matrix $\boldsymbol{X}$ to zero, and $\chi(f) = (\sum_{\tau_b}\|\boldsymbol{R}_{xx}(f,\tau_b)\|^2)^{-1}$ is a normalization factor ($\|\cdot\|$ represents the Frobenius norm). $\boldsymbol{R}_{xx}(f,\tau_b)$ and $\boldsymbol{R}_{oo}(f,\tau_b)$ are the cross-power spectra of the input $\boldsymbol{x}(f,\tau)$ and the output $\boldsymbol{o}(f,\tau)$, respectively, which are calculated around the multiple time blocks τ_b. Also, Pham et al. have proposed the following improved criterion for SO-ICA [20];

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

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

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

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

where $\boldsymbol{I}$ is the identity matrix, $\langle\cdot\rangle_\tau$ denotes the time-averaging operator, and $\boldsymbol{\varphi}(\cdot)$ is the nonlinear vector function. Many kinds of nonlinear function $\boldsymbol{\varphi}(f,\tau)$ have been proposed. Considering a batch algorithm of ICA, it is well-known that $\tanh(\cdot)$ or the sigmoid function is appropriate for super-Gaussian sources such as speech signals [36]. In this study, I define the nonlinear vector function $\boldsymbol{\varphi}(\cdot)$ as

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

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

where the superscripts (R) and (I) denote the real and imaginary parts, respectively. The nonlinear function given by (12) indicates that the nonlinearity is applied to the real and imaginary parts of the complex-valued signals separately. This type of complex-valued nonlinear function has been introduced by Smaragdis [17] for the FDICA, where it can be assumed in speech signals that the real (or imaginary) parts of the time-frequency representations of sources are mutually independent.

According to Refs. [22, 37], the source separation performance of HO-ICA is almost the same as or superior to that of SO-ICA. Thus, I utilize HO-ICA as basic ICA algorithm in simulations of this study.

2.3.3 Scaling and permutation problem

In FDICA, separation matrices are updated independently in each frequency bin. Therefore, source-gain ambiguity and source-order ambiguity arise in each frequency bin. The former problem is known as a *scaling problem*, and the latter problem is referred to as a *permutation problem*. The scaling problem can be solved by projection back (PB) method [16]. On the other hand, the permutation problem heavily decreases the resultant quality. Therefore, it is indispensable for us to align the permutation so that each separated signal contains frequency components from the same source. There have been several methods of solving permutation problem, e.g., a method based on correlations among neighbor frequency bins [16], a method based on DOA clustering [19, 33] and a integrated method of above mentioned methods [34]. However, the permutation problem cannot be solved completely. In addition, increase of the permutation-salvaging accuracy requires higher computational costs.

2.4. Conclusion

In this chapter, first, data model of speech enhancement system was denoted. Next two typical ICA algorithms for BSS were reviewed.

Chapter 3

Analysis of ICA under Non-Point-Source Noise Condition

3.1. Introduction

In this chapter, I investigate the proficiency of ICA under a non-point-source noise condition. In relation to the performance analysis of ICA, Araki et al. has mentioned that ICA-based BSS has equivalence to parallelly constructed ABFs [38]. However, this investigation was focused on separation with a non-singular mixing matrix, and thus was valid for only point sources.

First, I analyze beamformers that are optimized by ICA under a non-point-source condition in Sect. 3.2. In the analysis, I clarify that the beamformers optimized by ICA become specific beamformers that maximize the signal-to-noise ratio (SNR) in each output (so-called *SNR-maximize beamformers*). In particular, the beamformer for target speech estimation is optimized to be a DS beamformer, and the beamformer for noise estimation is likely to be a null beamformer (NBF) [19].

Next, a computer simulation is conducted in Sect. 3.3, and its result also indicates that ICA is proficient in noise estimation under a non-point-source noise condition. Then, I conclude that ICA is suitable for noise estimation such a condition.

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

3.2.1 Can ICA separate any source signals?

Many previous studies of BSS provided strong evidence in that conventional ICA could work in source separation, particularly in the special case of speech-speech mixing, i.e., all sound sources are point sources. However, such sound mixing is not realistic under common acoustic conditions; indeed the following scenario and problem are likely to arise (see Fig. 5):

- The target sound is the user's speech, which can be approximately regarded as a *point source*. In addition, the user themselves locates relatively *near the microphone array* (e.g., 1 m apart), and consequently the accompanying reflection and reverberation components are moderate.

- As for the noise, we are often confronted with interference sound(s) which is *not a point source* but a widespread source. Also the noise is usually far from the array and heavily reverberant.

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

3.2.2 SNR-Maximize beamformers optimized by ICA

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

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

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

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

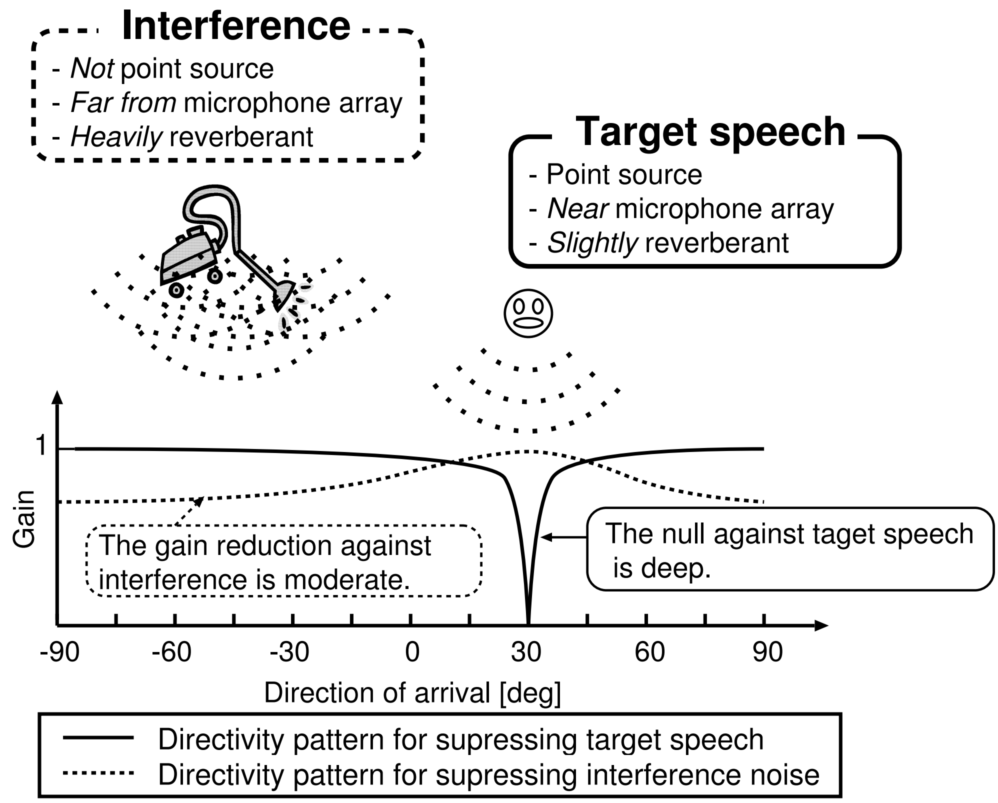

Interference
- *Not* point source
- *Far from* microphone array
- *Heavily* reverberant
Target speech
- Point source
- *Near* microphone array
- *Slightly* reverberant
1
Gain
The gain reduction against interference is moderate.
The null against taget speech is deep.
-90
-60
-30
0
30
60
90
Direction of arrival [deg]
Directivity pattern for supressing target speech
Directivity pattern for supressing interference noise

Figure 5. Expected directivity patterns that are shaped by ICA.

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

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

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

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

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

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

are diagonalized through the joint diagonalization.

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

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

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

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

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

$$C = \frac{\left|\langle y_{\mathrm{s}}(f,\tau) y_n^*(f,\tau)\rangle_\tau\right|}{\sqrt{\langle |y_{\mathrm{s}}(f,\tau)|^2\rangle_\tau}\sqrt{\langle |y_n(f,\tau)|^2\rangle_\tau}}, \tag{21}$$

$$y_{\mathrm{s}}(f,\tau) = \hat{s}(f,\tau) + r_{\mathrm{s}}\hat{n}(f,\tau), \tag{22}$$

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

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

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

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

where Σ_{s} is the SNR of $y_{\mathrm{s}}(f,\tau)$ and Σ_n is the SNR of $y_n(f,\tau)$. Using (22), (23), (24) and (25), we can rewrite (21) as

$$\begin{aligned} C &= \frac{\left|1/\sqrt{\Sigma_{\mathrm{s}}} \cdot e^{j\arg r_{\mathrm{s}}} + 1/\sqrt{\Sigma_n} \cdot e^{j\arg r_n^*}\right|}{\sqrt{1+1/\Sigma_{\mathrm{s}}}\sqrt{1+1/\Sigma_n}} \\ &= \frac{\left|1/\sqrt{\Sigma_{\mathrm{s}}} + 1/\sqrt{\Sigma_n} \cdot e^{j(\arg r_n^* - \arg r_{\mathrm{s}})}\right|}{\sqrt{1+1/\Sigma_{\mathrm{s}}}\sqrt{1+1/\Sigma_n}}, \end{aligned} \tag{26}$$

where $\arg r$ represents the argument of r. Thus, C is a function of only Σ_{s} and Σ_n. Therefore, the cross-correlation between $y_{\mathrm{s}}(f,\tau)$ and $y_n(f,\tau)$ only depends on the SNRs of beamformers $\boldsymbol{g}_{\mathrm{s}}(f)$ and $\boldsymbol{g}_n(f)$.

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

$$\begin{aligned} \frac{\partial C^2}{\partial \Sigma_{\mathrm{s}}} &= \frac{(1-\Sigma_{\mathrm{s}})}{(\Sigma_{\mathrm{s}}+1)^2(\Sigma_n+1)} \\ &\quad + \frac{\Sigma_{\mathrm{s}}\sqrt{\Sigma_{\mathrm{s}}\Sigma_n}(1-\Sigma_{\mathrm{s}})}{(\Sigma_{\mathrm{s}}+1)^2(\Sigma_n+1)} \cdot 2\mathrm{Re}\left[e^{j(\arg r_n^* - \arg r_{\mathrm{s}})}\right] < 0, \end{aligned} \tag{27}$$

where $\Sigma_{\mathrm{s}} > 1$ and $\Sigma_n > 1$. As for the partial differential of C^2 by Σ_n, I can also prove $\partial C^2/\partial\Sigma_n < 0$, where $\Sigma_{\mathrm{s}} > 1$ and $\Sigma_n > 1$ in the same manner. Therefore, C is a monotonically decreasing function of Σ_{s} and Σ_n. The above-mentioned fact indicates the following in ICA.

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

Thus, I conclude that ICA, in a parallel manner, optimizes multiple beamformers, i.e., $\boldsymbol{g}_{\mathrm{s}}(f)$ and $\boldsymbol{g}_n(f)$, so that the SNR of the output by each beamformer becomes maximum.

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

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

$$\min_{\boldsymbol{g}_s(f)} \boldsymbol{g}_{s}^{\mathrm{T}}(f)\boldsymbol{R}(f)\boldsymbol{g}_{s}(f) \quad \text{subject to } \boldsymbol{g}_{\mathrm{s}}^{\mathrm{T}}(f)\boldsymbol{a}(f,\theta_{\mathrm{s}}) = 1, \tag{28}$$

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

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

$$\boldsymbol{g}_{\rm s}(f)^{\rm T} = \frac{\boldsymbol{a}(f, \theta_{\rm s}(f))^{\rm H}\boldsymbol{R}^{-1}(f)}{\boldsymbol{a}(f, \theta_{\rm s}(f))^{\rm H}\boldsymbol{R}^{-1}(f)\boldsymbol{a}(f, \theta_{\rm s}(f))}. \tag{30}$$

This beamformer is called a minimum variance distortionless response (MVDR) beamformer [39]. Note that the MVDR beamformer requires the true DOA of the target speech and the noise-only time interval. However, we cannot determine the true DOA of the target source signal and noise-only interval because ICA is an *unsupervised* adaptive technique. Thus, the MVDR beamformer is expected to be the upper limit of ICA in the presence of non-point-source noises.

Although the correlation matrix is often not diagonalized in lower-frequency subbands [39], e.g., diffuse noise, I approximate that the correlation matrix is almost diagonalized in whole frequency subbands. Then, regarding the power of noise signal as approximately $\delta^2(f)$, the correlation matrix results in $\boldsymbol{R}(f) = \delta^2(f) \cdot \boldsymbol{I}$. Therefore, the inverse of correlation matrix $\boldsymbol{R}^{-1}(f) = \boldsymbol{I}/\delta^2(f)$ and (30) can be rewritten as

$$\boldsymbol{g}_{\rm s}(f)^{\rm T} = \frac{\boldsymbol{a}(f, \theta_{\rm s}(f))^{\rm H}}{\boldsymbol{a}(f, \theta_{\rm s}(f))^{\rm H}\boldsymbol{a}(f, \theta_{\rm s}(f))}. \tag{31}$$

Since $\boldsymbol{a}(f, \theta_{\rm s}(f))^{\rm H}\boldsymbol{a}(f, \theta_{\rm s}(f)) = J$, we finally obtain

$$\begin{aligned}\boldsymbol{g}_{\rm s}(f) = \frac{1}{J}[&\exp\left(-i2\pi(f/M)f_{\rm s}d_1 \sin\theta_{\rm s}(f)/c\right), \\ &\ldots, \exp\left(-i2\pi(f/M)f_{\rm s}d_J \sin\theta_{\rm s}(f)/c\right)]^{\rm T}.\end{aligned} \tag{32}$$

This filter $\boldsymbol{g}_{\rm s}(f)$ is approximately equal to a DS beamformer [7]. Note that the filter $\boldsymbol{g}_{\rm s}(f)$ is not a simple DS beamformer but a *reverberation-adapted DS beamformer* because it is optimized for distinct $\theta_{\rm s}(f)$ in each frequency bin. The resultant noise power is $\delta^2(f)/J$ when the noise is spatially uncorrelated and white Gaussian. Consequently the noise-reduction performance of the DS beamformer optimized by ICA under a non-point-source noise condition is proportional to $10\log_{10} J$ [dB]; this performance is not so good.

Next, I consider the other beamformer $\boldsymbol{g}_n(f)$ which picks up the noise source. As for the noise signal, the beamformer which removes the target signal arriving from $\theta_\mathrm{s}(f)$ is the SNR-maximize beamformer. Thus, the beamformer which steers the directional null to $\theta_\mathrm{s}(f)$ is the desired one for the noise signal. Such a beamformer is called NBF [19]. This beamformer compensates the phase of the signal arriving from $\theta_\mathrm{s}(f)$, and takes subtraction. Thus, the signal from arriving from $\theta_\mathrm{s}(f)$ is removed. For instance, NBF with two-element array is designed as

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

where $\sigma(f)$ is the gain compensate parameter. This beamformer surely satisfies $\boldsymbol{g}_n^\mathrm{T}(f)\cdot\boldsymbol{a}(f,\theta_\mathrm{s}(f)) = 0$. The steering vector $\boldsymbol{a}(f,\theta_\mathrm{s}(f))$ expresses the wavefront of the plane wave arriving from $\theta_\mathrm{s}(f)$. Thus, $\boldsymbol{g}_n(f)$ actually steers directional null to $\theta_\mathrm{s}(f)$. Note that this always holds regardless of the number of microphones (at least two microphones). Hence, this beamformer achieves quite high, ideally infinite, SNR for the noise signal.

Also, note that the filter $\boldsymbol{g}_n(f)$ is not a simple NBF but a *reverberation-adapted NBF* because it is optimized for distinct $\theta_\mathrm{s}(f)$ in each frequency bin. Overall, the performance of enhancing the target speech is very poor and that of estimating the noise source is good.

3.3. Computer simulation

I conduct computer simulations to confirm the performance of ICA under a non-point-source noise condition. Here, I used HO-ICA [17] as the ICA algorithm. I used the following 8 kHz-sampled signals as ICA's input; the original target speech (3 seconds) convoluted with impulse responses that were recorded in an actual environment, and to which three types of noise from 36 loudspeakers were added. The reverberation time (RT_{60}) is 200 ms; this corresponds to mixing fil-

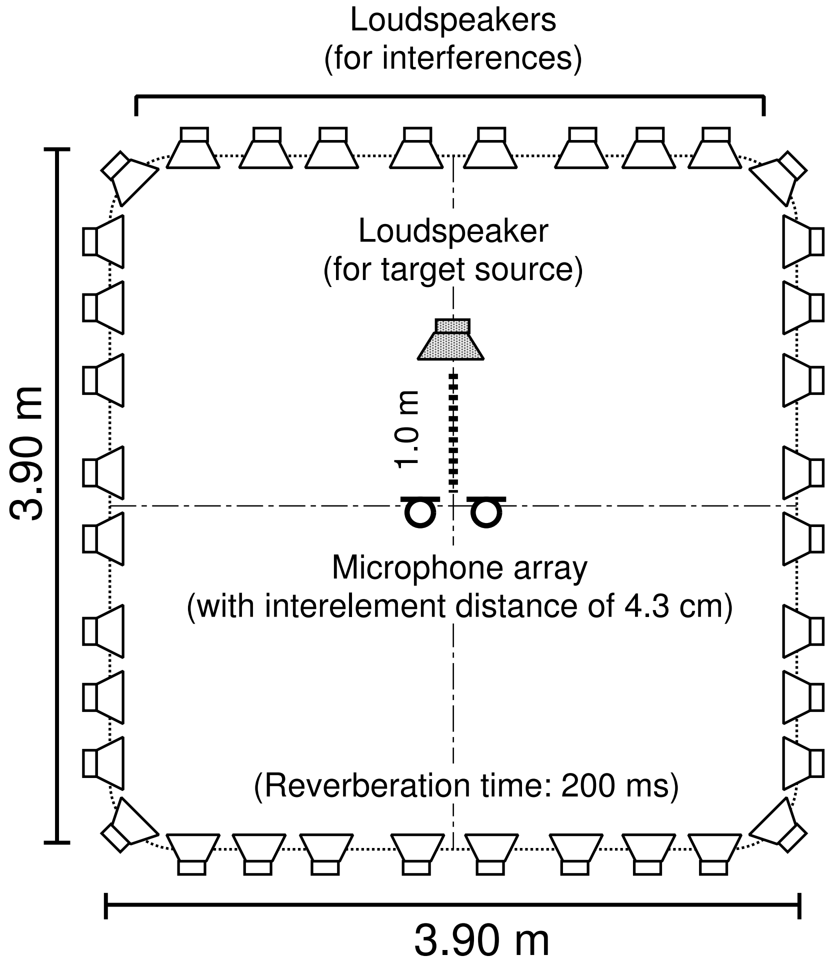


Figure 6. Layout of reverberant room in my simulation.

ters with 1600 taps in 8 kHz sampling. The three types of noise are an independent Gaussian noise, an actually recorded railway-station noise, and interference speech by 36 people. Figure 6 illustrates the reverberant room used in the simulation. I use 12 speakers (6 males and 6 females) as sources of the original target speech, and the input SNR of test data is set to 0 dB. I use a two-, three-, or four-element microphone array with an interelement spacing of 4.3 cm.

The simulation results are shown in Figs. 7 and 8. Figure 7 shows the result for the average noise reduction rate (NRR) [19] of all the target speakers. NRR is defined as the output SNR in dB minus the input SNR in dB. This measure

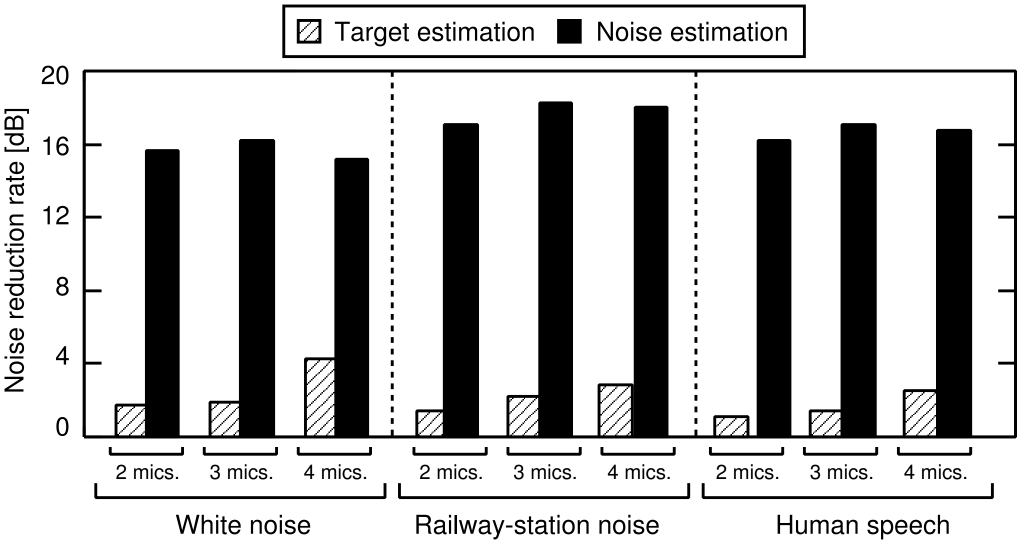


Figure 7. Simulation-based separation results under non-point-source noise condition.

indicates the objective performance of noise reduction. NRR is given by

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

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

From this result, we can see an imbalance between the target speech estimation and the noise estimation in every noise case; the performance of the target speech estimation is significantly poor, but that of noise estimation is very high. This result is consistent with the theory previously stated. Moreover, Fig. 8 shows directivity patterns shaped by the beamformers optimized by ICA in the simulation. It is clearly indicated that the beamformer $\boldsymbol{g}_{\mathrm{s}}(f)$ that picks up the target speech resembles the DS beamformer, and the beamformer $\boldsymbol{g}_n(f)$ that picks up the noise becomes NBF. From these results, we can confirm that the previously stated theory, i.e., the beamformers optimized by ICA under a non-point-source noise condition are DS and NBF, is valid.

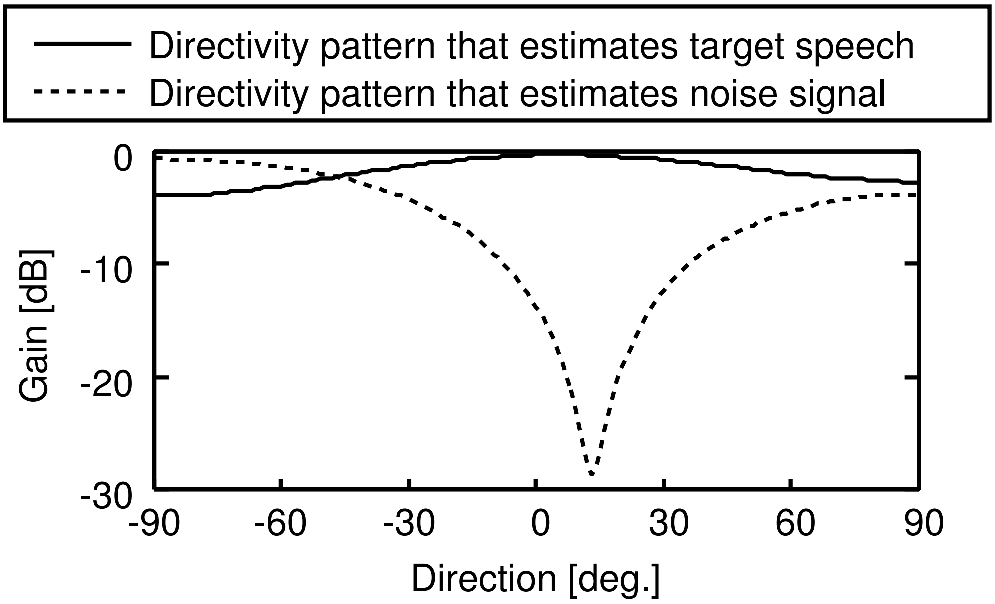


Figure 8. Typical directivity patterns under non-point-source noise condition shaped by ICA at 2 kHz and two-element array in white Gaussian noise case.

3.4. Conclusion

In this chapter, I gave the analysis of ICA under non-point-source noise condition. As a result of the analysis, I founded out that ICA optimizes SNR-maximize beamformers. Therefore, ICA generates NBF for target speech reduction, and DS for non-point-source noise reduction. That is to say, since the signal reduction performance of NBF is significantly high, ideally infinity, ICA is proficient in noise estimation that is equivalent to reduction of target speech. Also, the validity of this analysis was shown via a computer simulation. As a result of the simulation, it could be confirmed that ICA is proficient in estimation of non-point-source noise signal. Also, it could be shown that ICA generates DS for non-point-source signals and NBF for point source signals. For these reasons, I conclude that ICA is proficient in noise estimation under non-point-source noise condition.

Chapter 4

Blind Speech Extraction Method with ICA-based Noise Estimator

4.1. Introduction

In this chapter, I propose a new blind speech extraction method with using ICA-based noise estimator. In the previous chapter, I have clarified that ICA is proficient in noise estimation rather than in target-speech estimation under a non-point-source noise condition. This analytic result implies that ICA cannot be directly applied to the source separation problem which involves non-point-source noise signals. However, this analysis also insists that ICA can be still utilized as an accurate noise estimator. This fact motivates me to propose a new speech-enhancement strategy, i.e., BSSA. The proposed BSSA consists of a DS-based primary path and a reference path including ICA-based noise estimation (see Fig. 9). The estimated noise component in ICA is efficiently subtracted from the primary path in the power-spectrum domain without phase information. This procedure can yield better target-speech enhancement than the simple ICA, even with a benefit of estimation-error robustness in speech recognition applications.

Furthermore, the proposed BSSA has another advantage, i.e., *permutation robustness*. In frequency-domain ICA, source permutation ambiguity arises in each frequency bin, and markedly decreases the source separation quality. Therefore it is indispensable for us to align permutation so that each separated signals contains frequency components from the same source. Although various permutation solvers have been proposed, e.g., spectral-continuity-based methods [16, 32], DOA-based methods [19], and the integration method of spectral continuity and DOA [34], have been proposed, the permutation problem cannot be solved completely. In addition, an increase of the permutation-salvaging accuracy requires an increase in computational cost. Permutation robustness indicates how much the BSS method is not affected under a certain probability of arising permutation, and

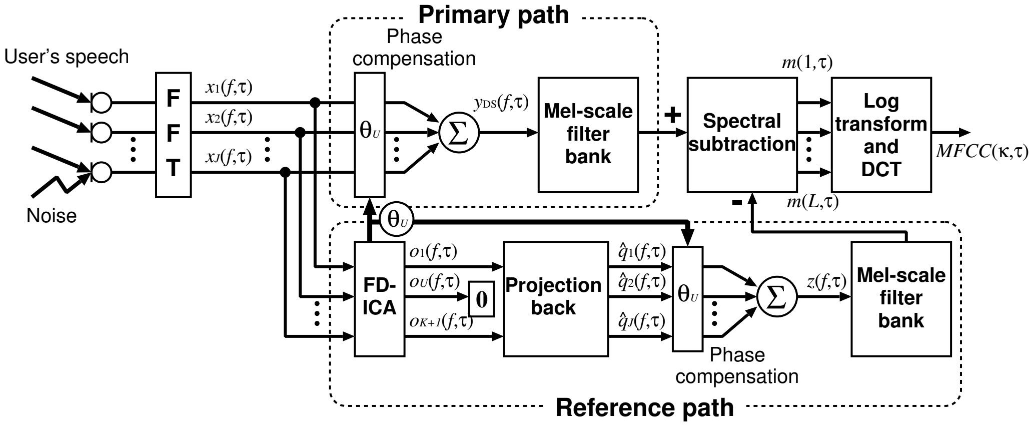


Figure 9. Block diagram of proposed blind spatial subtraction array.

such an important property has never been studied so far in previous ICA studies. Note that permutation robustness in BSSA does not conflict with any permutation solver. That is to say, any permutation solvers are available in ICA part of BSSA. BSSA reduces the remained permuted components which could not be solved by a permutation solver. An improvement in permutation robustness through small computations is a novel and efficient way of increasing BSS quality.

The chapter is organized as follows. In the following Sect. 4.2, I give a detailed signal processing in proposed BSSA. Next, the discussion about permutation robustness is described in Sect. 4.3. Next, the effectiveness of the proposed BSSA is shown via experimental results in Sect. 4.4. Finally, Sect. 4.5 concludes the chapter.

4.2. Algorithm

4.2.1 Partial speech enhancement in primary path

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

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

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

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

where $y_{\mathrm{DS}}(f,\tau)$ is the primary-path output that is a slightly enhanced target speech, $\boldsymbol{w}_{\mathrm{DS}}(f)$ is the filter coefficient vector of DS, and θ_U is the estimated DOA of the target speech given by the ICA part in Sect. 4.2.2. In (35), the second and third terms on the right-hand side express the remaining noise in the output of the primary path.

4.2.2 ICA-based noise estimation in reference path

The proposed BSSA provides ICA-based noise estimation. First, source separation by ICA is applied to the observed signal, and we obtain the separated signal vector $\boldsymbol{o}(f,\tau)$ is obtained as

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

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

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

where the unmixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$ is optimized by (11). Note that the number of ICA's outputs becomes $K + 1$ and thus the number of sensors, J, is equal to more than $K+1$ because it is supposed that the additive noise $\boldsymbol{n}_a(f,\tau)$ is not negligible. The additive noise cannot be perfectly estimated because the additive noise is deformed by the filter optimized by ICA. Moreover other components cannot also be estimated perfectly when the additive noise $\boldsymbol{n}_a(f,\tau)$ exists. However, it is possible to estimate at least that noises (including interference sounds that can be regarded as point sources, and the additive noise) that do not involve the target speech signal, as denoted in Sect. 3.2. Therefore, the estimated noise signal is still beneficial.

Next, DOAs are estimated from the unmixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$ [19, 34]. This procedure is represented by

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

where θ_u is the DOA of the uth sound source. Then, Uth source signal which is nearest the front of the microphone array are chosen, and the DOA of the chosen source signal is designated as θ_U in this paper. This is because almost all users will stand in front of the microphone array in a speech-oriented human-machine interface, e.g., a public guidance system, which is one of my target applications. Other strategies for choosing target speech signal can be considered (see Appendix B).

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

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

where $\boldsymbol{q}(f,\tau)$ is the "noise-only" signal vector that contains only noise components. Next, the projection back (PB) [16] method is performed to remove the ambiguity of amplitude. This procedure can be represented as

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

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

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

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

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

This equation means that $z(f,\tau)$ is a good candidate for noise terms of the primary path output $y_{\mathrm{DS}}(f,\tau)$ (see the 2nd and 3rd terms on the right-hand side of (35)). Of course this noise estimation is not perfect, but it is still possible to enhance the target speech signal via oversubtraction in the power-spectrum domain as described in Sect. 4.2.3. Note that $z(f,\tau)$ is a function of the frame index τ, unlike the constant noise prototype in the traditional SS method [5]. Therefore, the proposed BSSA can deal with *nonstationary* noise.

4.2.3 Noise reduction processing in BSSA

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

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

where $y_{\mathrm{BSSA}}(f,\tau)$ is the final output of BSSA, β is an oversubtraction parameter, and η is a flooring parameter. This is an extended formulation of SS [40]. The appropriate setting, e.g., $\beta > 1$ and $\eta \ll 1$, gives an efficient noise reduction. For example, too larger oversubtraction parameter ($\beta \gg 1$) leads the larger SNR improvement. However, the target signal would be distorted. On the other hand, the smaller oversubtraction parameter ($\beta \ll 1$) gives the low-distorted target signal. However the SNR improvement is decreased. In the end, the trade-off between SNR improvement and the distortion of the output signal exists with respect to the parameter β; $1 < \beta < 2$ is usually used.

The system switches in two equations depending on the conditions in (46). If the calculated noise components using ICA in (45) are underestimated, i.e., $|y_{\mathrm{DS}}(f,\tau)|^2 > \beta |z(f,\tau)|^2$, the resultant output $y_{\mathrm{BSSA}}(f,\tau)$ corresponds to the power-spectrum-domain subtraction among primary and reference paths with an over-subtraction rate of β. On the other hand, if the noise components are overestimated in ICA, i.e., $|y_{\mathrm{DS}}(f,\tau)|^2 < \beta |z(f,\tau)|^2$, the resultant output $y_{\mathrm{BSSA}}(f,\tau)$ is floored with a small positive value to avoid the negative-valued unrealistic spectrum. These *oversubtraction* and *flooring* procedures promise us an error-robust speech enhancement in the proposed BSSA rather than a simple linear subtraction. Although the nonlinear processing in (46) often generates an artificial distortion so called *musical noise*, it is still applicable in the speech recognition system because the speech decoder is not very sensitive to such a distortion.

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

The proposed BSSA involves a mel-scale filter bank analysis and directly outputs mel-frequency cepstrum coefficient (MFCC) [41] for speech recognition. Therefore, the proposed BSSA requires no transformation into the time-domain waveform for speech recognition. The detailed process is shown in Appendix C.

4.3. Permutation-robustness analysis in BSSA

In this section, I present a permutation-robustness analysis in BSSA architecture. In conventional FDICA, when the permutation arises, we directly suffer from a permuted noise component that is wrongly regarded as the target signal. Thus, the conventional FDICA has no robustness against permutation. For the permutation problem, FDICA requires special processing, i.e., permutation solvers [42]. On the other hand, in BSSA, the adverse effect of permutation is mitigated because the SS-based source extraction process in (46) reduces the power of permuted components (details will be shown in Sect. 4.3.1, and DS defocuses the component arriving from the out of look direction (details will be described in Sect. 4.3.2). These are performed without any special processing like permutation solvers. Therefore, it can be conclude that the BSSA architecture is a permutation-robust structure. Note that BSSA is not just a permutation solver but a mitigation of residual permutation effect. Indeed, BSSA can utilize any permutation solvers in ICA part. The BSSA structure can reduce remaining permuted components after permutation solver. The detailed analysis is shown below.

4.3.1 Permutation robustness by oversubtraction

Here, it is supposed that source separation was performed perfectly by FDICA except for the permutation that arises in the frequency bin f_p. Moreover, it is assumed that the additive noise $\boldsymbol{n}_a(f,\tau)$ can be made negligible to simplify discussion. Consequently, the observed signal in (1) can be rewritten as

$$\tilde{\boldsymbol{x}}(f,\tau) = \boldsymbol{A}(f)\{\boldsymbol{s}(f,\tau) + \boldsymbol{n}(f,\tau)\}. \tag{47}$$

Under this assumption, the estimated target speech signal in the frequency bin f_p by ICA (including PB processing) can be described as

$$\boldsymbol{y}_\mathrm{ICA}(f_\mathrm{p},\tau) = \boldsymbol{A}(f_\mathrm{p})\boldsymbol{n}_e(f_\mathrm{p},\tau), \tag{48}$$

$$\boldsymbol{n}_e(f_\mathrm{p},\tau) = [\underbrace{0,\ldots,0}_{V-1}, n_V(f_\mathrm{p},\tau), \underbrace{0,\ldots,0}_{K-V}]^\mathrm{T}, \tag{49}$$

where $\boldsymbol{y}_\mathrm{ICA}(f_\mathrm{p},\tau)$ is the output signal vector as a target speech signal by ICA, $\boldsymbol{n}_e(f_\mathrm{p},\tau)$ is the noise signal vector estimated as the target speech signal vector by mistake, $n_V(f_\mathrm{p},\tau)$ is the noise component estimated as the target speech component by mistake, and $V(\neq U)$ expresses the component number of noise. Moreover, since $\boldsymbol{n}_e(f_\mathrm{p},\tau)$ is composed of zero components except the specific noise component $n_V(f_\mathrm{p},\tau)$, $\boldsymbol{y}_\mathrm{ICA}(f_\mathrm{p},\tau)$ can be rewritten as

$$\boldsymbol{y}_\mathrm{ICA}(f_\mathrm{p},\tau) = \boldsymbol{h}(f_\mathrm{p})n_V(f_\mathrm{p},\tau), \tag{50}$$

$$\boldsymbol{h}(f_\mathrm{p}) = [A_{1V}(f_\mathrm{p}),\ldots,A_{JV}(f_\mathrm{p})]^\mathrm{T}, \tag{51}$$

where $\boldsymbol{h}(f_{\mathrm{p}})$ is a transfer function vector of the noise component $n_V(f_{\mathrm{p}},\tau)$, and $A_{ij}(f)$ expresses an element of the mixing matrix $\boldsymbol{A}(f)$.

On the other hand, the estimated noise signal in the reference path of BSSA, $\tilde{z}(f_{\mathrm{p}},\tau)$, can be represented by

$$\tilde{z}(f_{\mathrm{p}},\tau) = \boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{A}(f_{\mathrm{p}})\tilde{\boldsymbol{q}}(f_{\mathrm{p}},\tau), \tag{52}$$

$$\begin{aligned}\tilde{\boldsymbol{q}}(f_{\mathrm{p}},\tau) = [&n_1(f_{\mathrm{p}},\tau),\ldots,n_{V-1}(f_{\mathrm{p}},\tau),0,n_{V+1}(f_{\mathrm{p}},\tau),\ldots,\\ &n_{U-1}(f_{\mathrm{p}},\tau),s_U(f_{\mathrm{p}},\tau),n_{U+1}(f_{\mathrm{p}},\tau),\ldots,n_K(f_{\mathrm{p}},\tau)]^{\mathrm{T}},\end{aligned} \tag{53}$$

where $\tilde{\boldsymbol{q}}(f_{\mathrm{p}},\tau)$ is the estimated noise component vector including the target signal by mistake (here $V < U$ for simplification). Note that the observed signal $\tilde{\boldsymbol{x}}(f_{\mathrm{p}},\tau)$ can be rewritten as

$$\tilde{\boldsymbol{x}}(f_{\mathrm{p}},\tau) = \boldsymbol{A}(f_{\mathrm{p}})\{\tilde{\boldsymbol{q}}(f_{\mathrm{p}},\tau) + \boldsymbol{n}_e(f_{\mathrm{p}},\tau)\}. \tag{54}$$

Moreover, since the additive noise can be negligible in this section, the output of the primary path in BSSA (35) can be written as

$$\begin{aligned}\tilde{y}_{\mathrm{DS}}(f_{\mathrm{p}},\tau) &= \boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\tilde{\boldsymbol{x}}(f_{\mathrm{p}},\tau)\\ &= \boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{A}(f)\{\tilde{\boldsymbol{q}}(f_{\mathrm{p}},\tau) + \boldsymbol{n}_e(f_{\mathrm{p}},\tau)\}.\end{aligned} \tag{55}$$

When $|\tilde{y}_{\mathrm{DS}}(f_{\mathrm{p}},\tau)|^2 - \beta\cdot|\tilde{z}(f_{\mathrm{p}},\tau)|^2 \geq 0$, form (52) and (55), the expectation of the power spectrum of BSSA output $\tilde{y}_{\mathrm{BSSA}}(f_{\mathrm{p}},\tau)$ can be represented by

$$\begin{aligned}&E\left[|\tilde{y}_{\mathrm{BSSA}}(f_{\mathrm{p}},\tau)|^2\right]\\ &= E\left[|\tilde{y}_{\mathrm{DS}}(f_{\mathrm{p}},\tau)|^2 - \beta\cdot|\tilde{z}(f_{\mathrm{p}},\tau)|^2\right]\\ &= E\left[|\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{A}(f_{\mathrm{p}})\left\{\tilde{\boldsymbol{q}}(f_{\mathrm{p}},\tau) + \boldsymbol{n}_e(f_{\mathrm{p}},\tau)\right\}|^2\right]\\ &\qquad - E\left[\beta\cdot|\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{A}(f_{\mathrm{p}})\tilde{\boldsymbol{q}}(f_{\mathrm{p}},\tau)|^2\right]\\ &\simeq (1-\beta)\cdot E\left[|\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{A}(f_{\mathrm{p}})\tilde{\boldsymbol{q}}(f_{\mathrm{p}},\tau)|^2\right]\\ &\qquad + E\left[|\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{A}(f_{\mathrm{p}})\boldsymbol{n}_e(f_{\mathrm{p}},\tau)|^2\right],\end{aligned} \tag{56}$$

where $E[\cdot]$ denotes the expectation operator. I use the relation showing that the cross terms among the distinct noise components are negligible with taking expectation. Since the oversubtraction parameter is usually set to $\beta > 1$, it is evident that the first term on the right-hand side of (56) is a negative quantity and the following relation holds:

$$\begin{aligned} E\left[|\tilde{y}_{\mathrm{BSSA}}(f_{\mathrm{p}},\tau)|^2\right] &< E\left[|\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{A}(f_{\mathrm{p}})\boldsymbol{n}_e(f_{\mathrm{p}},\tau)|^2\right] \\ &= E\left[|\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{h}(f_{\mathrm{p}})n_V(f_{\mathrm{p}},\tau)|^2\right]. \end{aligned} \tag{57}$$

4.3.2 Permutation robustness by defocusing in DS

Under reverberant conditions, $\boldsymbol{h}(f_{\mathrm{p}})$ can be expressed by the superposition of all reflection components. Therefore $\boldsymbol{h}(f_{\mathrm{p}})$ can be rewritten as

$$\boldsymbol{h}(f_{\mathrm{p}}) = \sum_q r^{(q)}\boldsymbol{a}(f_{\mathrm{p}},\theta^{(q)}), \tag{58}$$

where (q) is used to express the index of the qth reflection component, $r^{(q)}$ is the reflection coefficient, $\theta^{(q)}$ is the DOA of the reflection component of the permuted noise $n_V(f_{\mathrm{p}},\tau)$, and $\boldsymbol{a}(f,\theta)$ is the steering vector described in (29). From (58), the resultant power of DS can be obtained by

$$\begin{aligned} &|\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{h}(f_{\mathrm{p}})n_V(f_{\mathrm{p}},\tau)|^2 \\ &= \Big|\sum_q r^{(q)}\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{a}(f_{\mathrm{p}},\theta^{(q)})n_V(f_{\mathrm{p}},\tau)\Big|^2 \\ &= \sum_q \Big|r^{(q)}\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{a}(f_{\mathrm{p}},\theta^{(q)})n_V(f_{\mathrm{p}},\tau)\Big|^2 + C_1, \end{aligned} \tag{59}$$

where C_1 is a term that contains all the cross terms among reflection components. Also, the power of the conventional ICA's output in the specific microphone j, $y_{\mathrm{ICA}}^{[j]}(f_{\mathrm{p}},\tau)$, can be written as

$$\begin{aligned} |y_{\mathrm{ICA}}^{[j]}(f_{\mathrm{p}},\tau)|^2 &= \Big|\sum_q r^{(q)}a_j(f_{\mathrm{p}},\theta^{(q)})n_V(f_{\mathrm{p}},\tau)\Big|^2 \\ &= \sum_q \Big|r^{(q)}a_j(f_{\mathrm{p}},\theta^{(q)})n_V(f_{\mathrm{p}},\tau)\Big|^2 + C_2, \end{aligned} \tag{60}$$

where $a_j(f,\theta)$ is the jth entry of $\boldsymbol{a}(f,\theta)$, and C_2 also expresses all the cross terms among reflection components. Here, the directivity gain of the DS filter $\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f)$ is unity only when θ equals the focus direction of DS, θ_U, and it is less than one (i.e., defocused) in the other directions. This is represented by

$$\left|\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{a}(f,\theta)\right| \leq 1. \tag{61}$$

Thus, the power of each reflection component satisfies

$$\begin{aligned}&|\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{a}(f_{\mathrm{p}},\theta^{(q)})|^2|r^{(q)}n_V(f_{\mathrm{p}},\tau)|^2\\ &\qquad\leq |a_j(f_{\mathrm{p}},\theta^{(q)})|^2|r^{(q)}n_V(f_{\mathrm{p}},\tau)|^2,\end{aligned} \tag{62}$$

because $|a_j(f,\theta)| = 1$. Here I assume that almost all the reflection components of $n_V(f_{\mathrm{p}},\tau)$ come from around the noise DOA and outside of θ_U. Hence, the following relation holds for almost all the reflection components except the specific reflection component arriving from θ_U, $\boldsymbol{a}(f_{\mathrm{p}},\theta^{(q')})n_V(f_{\mathrm{p}},\tau)$, where $\theta^{(q')} = \theta_U$;

$$\begin{aligned}&|r^{(q)}\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{a}(f_{\mathrm{p}},\theta^{(q)})n_V(f_{\mathrm{p}},\tau)|^2\\ &\qquad< |r^{(q)}a_j(f_{\mathrm{p}},\theta^{(q)})n_V(f_{\mathrm{p}},\tau)|^2.\end{aligned} \tag{63}$$

Moreover, if the interference with each reflection component arises statistically at random, it can be expected that C_1 in (59) and C_2 in (60) become statistically the same. Therefore, the following equation holds:

$$\begin{aligned}&\sum_q |r^{(q)}\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{a}(f_{\mathrm{p}},\theta^{(q)})n_V(f_{\mathrm{p}},\tau)|^2 + C_1\\ &\qquad< \sum_q |r^{(q)}a_j(f_{\mathrm{p}},\theta^{(q)})n_V(f_{\mathrm{p}},\tau)|^2 + C_2.\end{aligned} \tag{64}$$

This equation can be replaced by

$$|\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{h}(f_{\mathrm{p}})n_V(f_{\mathrm{p}},\tau)|^2 < |y_{\mathrm{ICA}}^{[j]}(f_{\mathrm{p}},\tau)|^2. \tag{65}$$

From (57) and (65), the following relation is valid:

$$\begin{aligned}E\left[|\tilde{y}_{\mathrm{BSSA}}(f_{\mathrm{p}},\tau)|^2\right] &< E\left[|\boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f_{\mathrm{p}})\boldsymbol{h}(f_{\mathrm{p}})n_V(f_{\mathrm{p}},\tau)|^2\right]\\ &< E\left[|y_{\mathrm{ICA}}^{[j]}(f_{\mathrm{p}},\tau)|^2\right].\end{aligned} \tag{66}$$

This relation indicates that the power of the BSSA output is less than that of the ICA output in the permutation-arising frequency bin f_{p}.

On the other hand, when $|\tilde{y}_{\mathrm{DS}}(f_{\mathrm{p}},\tau)|^2 - \beta \cdot |\tilde{z}(f_{\mathrm{p}},\tau)|^2 < 0$, the resultant power spectrum of BSSA is floored by the flooring parameter η. If η is sufficiently small, $\tilde{y}_{\mathrm{BSSA}}(f_{\mathrm{p}},\tau)$ becomes smaller than the error component of permutation.

From the above-mentioned fact, it can be concluded that BSSA is permutation-robust compared with ICA. However, we must pay attention to the setting of the oversubtraction parameter β. Although the oversized oversubtraction parameter β can suppress permutation perfectly, such a parameter reduces not only noise components but also the target component in other innocent (nonpermuted) frequency bins. Therefore, we should use an appropriate oversubtraction parameter, β, because such an oversized parameter causes artificial distortion.

4.4. Evaluation of proposed blind speech extraction method

In this section, I carry out the following experiments to show the effectiveness of the proposed BSSA.

- **Evaluation of permutation robustness in BSSA**
  In this experiment, I make a comparison between the conventional ICA and proposed BSSA from the viewpoint of permutation robustness.

- **Experiment in reverberant room**
  In this experiment, I represent a comparison result of ICA-based BSS, and the traditional SS cascaded with ICA, and the proposed BSSA under the reverberant room condition. The comparison is performed on the basis of NRR, cepstral distortion (CD), and speech recognition test.

- **Experiment in an actual world**
  The above evaluations are conducted in the experiment room. On the other hand, I evaluate the performance of the proposed BSSA in an actual railway-station in this experiment. The effectiveness of the proposed BSSA under an actual environment is revealed through this experiment.

4.4.1 Evaluation of permutation robustness in BSSA

In this experiment, I mainly evaluate permutation-robustness ability in BSSA. First, I compare ICA and BSSA on the basis of NRR. As well, HO-ICA algorithm is utilized as the conventional ICA [17]. Hereafter, the 'ICA' simply indicates HO-ICA. It is supposed that source separation is performed perfectly except for the permutation generated artificially in randomly selected frequency bins. I increase the percentage of permutation-arising frequency bins to assess the robustness against the permutation problem. Figure 10 shows a layout of the reverberant room used in this experiment, where the reverberation time is 200 ms; this corresponds to mixing filters of 3200 taps with 16 kHz sampling. I use 3-s speech signals (male and female) as an original speech, and input SNR is set to 0 dB at the array. The target signal is a male's speech, the noise is a female's speech, and the noise direction is 50 degrees. A four- or eight-element array with an interelement spacing of 2 cm is used, and DFT size is 512. The oversubtraction parameter β is 1.2 and the flooring coefficient γ is 0.0. Such parameters are experimentally determined. Figure 11 shows the resultant curve of the NRRs of ICA and BSSA with increasing the percentage of permutation-arising frequency bins. From these results, we can confirm that the NRR of BSSA outperforms that of ICA even if the percentage of permutation-arising frequency bins increases. These results evidently indicate that BSSA involves a permutation-robust structure.

Although the previous NRR results are positive for BSSA, one might speculate that sound distortion is enhanced; certainly, we can see musical noise in the resultant output of the proposed BSSA. Consequently, I show results of CD and speech recognition that is the final goal of BSSA, in which the separated sound quality is completely considered. I use an eight-element array, and I generate 5% or 10% permutations artificially. I use 46 speakers (200 sentences) as the original source and I use a male's speech (1 sentence) as an interference noise source. Noise direction is 50 or 80 degrees. The speech recognition task and conditions are shown in Table 1.

CD [43] is a measure of the degree of distortion via the cepstrum domain. CD indicates distortion among two signals, which is defined as

$$\mathrm{CD\ [dB]} \equiv \frac{1}{T}\sum_{\tau=1}^{T} D_b \sqrt{\sum_{\rho=1}^{B} 2(C_{\mathrm{out}}(\rho;\tau) - C_{\mathrm{ref}}(\rho;\tau))^2}, \tag{67}$$

$$D_b = \frac{20}{\log 10}, \tag{68}$$

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

$$\mathrm{WA\ [\%]} \equiv \frac{W_{\mathrm{WA}} - S_{\mathrm{WA}} - D_{\mathrm{WA}} - I_{\mathrm{WA}}}{W_{\mathrm{WA}}} \times 100, \tag{69}$$

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

Figures 12(c) and (d) illustrate the CD score under each condition. We can see that the proposed BSSA increases the degree of distortion slightly due to spectral oversubtraction. Figures 12(e) and (f) show the word accuracy under each condition. From these results, however, we can confirm that the word accuracy of the proposed BSSA is higher that of ICA under all conditions; this means that the marked improvement in NRR can dominantly contribute to word accuracy in BSSA.

4.4.2 Experiment in reverberant room

In this experiment, I present a comparison of typical blind noise reduction methods, namely, the conventional ICA [17] and the traditional SS [5] cascaded with ICA (ICA+SS). In ICA+SS, first, noise estimation is performed from the speech pause interval in the target speech estimation by ICA. The noise reduction is

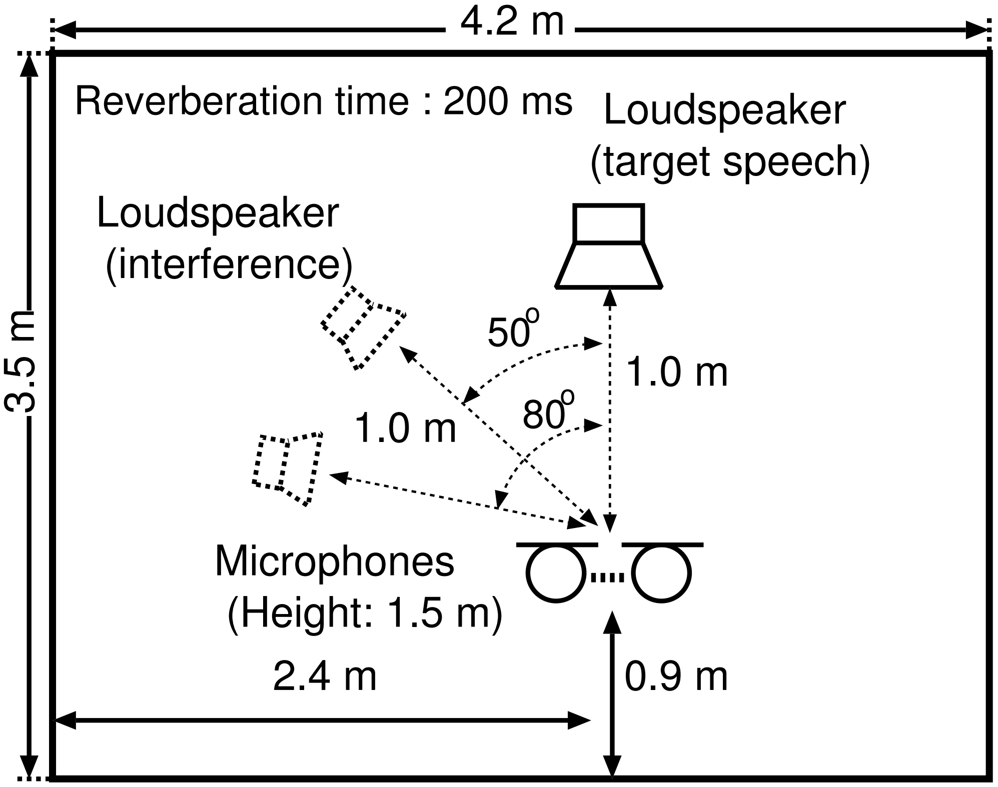


Figure 10. Layout of reverberant room used in experiment which simulates permutation problem.

Table 1. Conditions for Speech Recognition

Database	JNAS [44], 306 speakers (150 sentences/speaker)
Task	20 k words newspaper dictation
Acoustic model	phonetic tied mixture (PTM) [44], clean model
Number of training speakers for acoustic model	260 speakers (150 sentences/speaker)
Decoder	JULIUS [44] ver 3.5.1

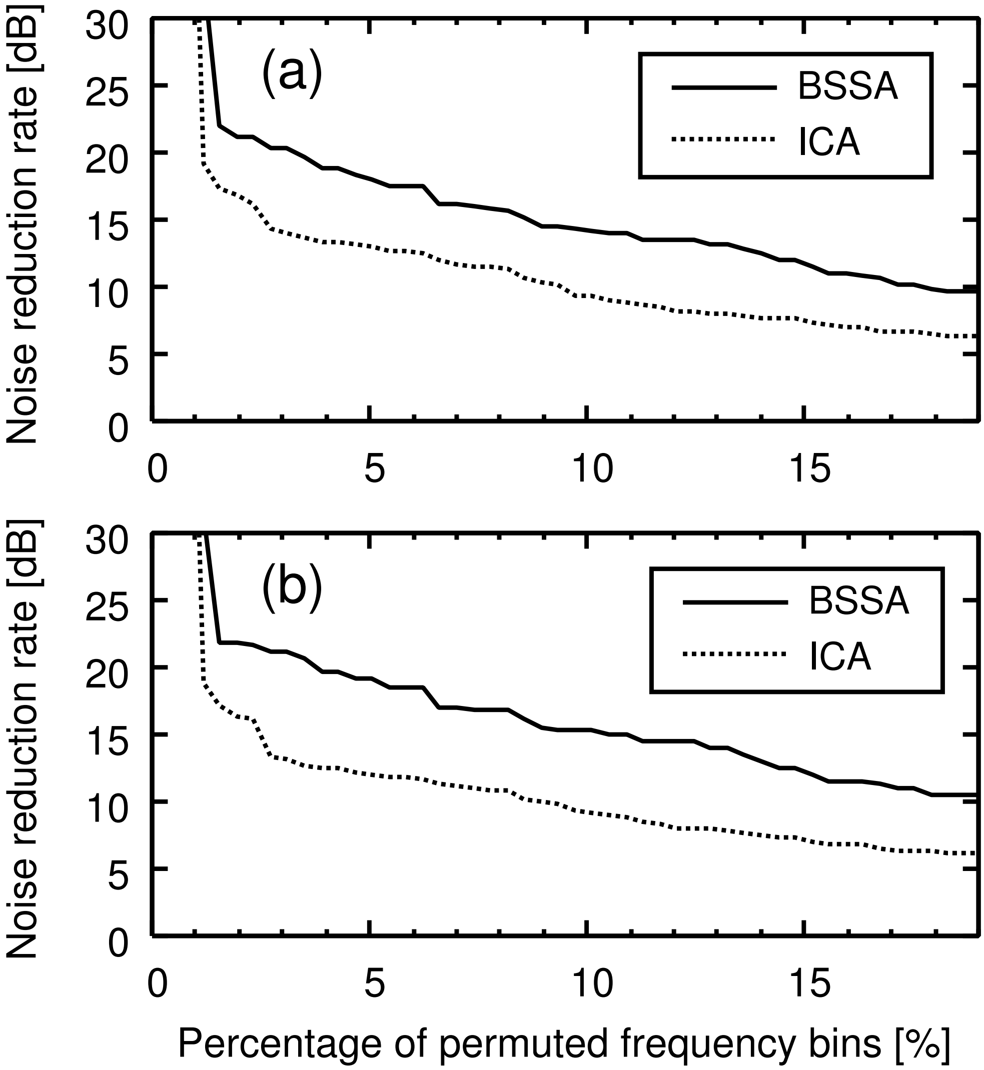


Figure 11. Curves of NRR with increasing the percentage of permutation-arising frequency bins by (a) four-element and (b) eight-element arrays.

achieved by SS as

$$y_{\text{ICA+SS}}(f,\tau) = \begin{cases} \left\{|o_U(f,\tau)|^2 - \beta|\hat{n}_{\text{remain}}(f)|^2\right\}^{\frac{1}{2}} \\ \qquad \text{(where } |o_U(f,\tau)|^2 - \beta|\hat{n}_{\text{remain}}(f,\tau)|^2 \geq 0\text{)}, \\ \gamma|o_U(f,\tau)| \quad \text{(otherwise)}, \end{cases} \tag{70}$$

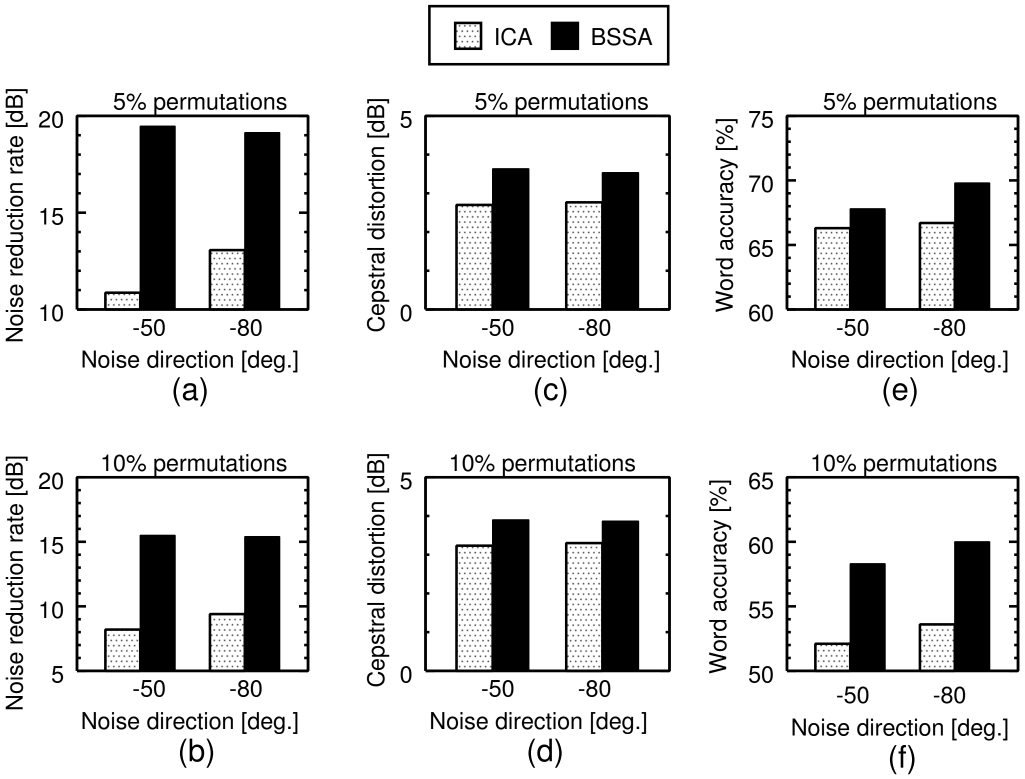


Figure 12. Experimental results of simulating permutation problem artificially. (a) and (b) are results of noise reduction rate, where 5% and 10% permutations arose, respectively. (c) and (d) indicate results of cepstral distortion, where 5% and 10% permutations arose, respectively. (e) and (f) show speech recognition results, where 5% and 10% permutations arose, respectively.

where $\hat{n}_{\text{remain}}(f)$ is the estimated noise signal from the speech pause in the target speech estimation by ICA. Moreover, DOA-based permutation solver[19] is used in the conventional ICA and ICA part in BSSA.

I used 16 kHz-sampled signals as test data; the original speech (6 s) convoluted with impulse responses recorded in an actual environment, and to which cleaner noise or a male's interfering speech that was recorded in an actual environment were added. Figure 13 shows the layout of the reverberant room used in the experiment. The reverberation time of the room is 200 ms; this corresponds to mixing filters of 3200 taps in 16 kHz sampling. The cleaner noise is not a simple point source signal but consists of several *nonstationary* noises emitted from a motor, air duct and nozzle. Also, a male's interfering speech is not a simple point source but is slightly moving. In addition, these interference noises involve background noise. The SNR of background noise (power ratio between target speech and background noise) is about 28 dB. I use 46 speakers (200 sentences) as the source of the target speech. The input SNR is set to 10 dB at the array. I use a four-element microphone array with an interelement spacing of 2 cm. The DFT size is 512. The oversubtraction parameter β is 1.4 and the flooring coefficient γ is 0.2. Such parameters are experimentally determined. The speech recognition task and conditions are the same as those in Sect. 4.4.1 as shown in Table 1.

First, I show actual separation results by ICA for the cleaner noise and interference speech cases in Fig. 14. We can confirm the imbalanced performance between target estimation and noise estimation similarly to the simulation-based results (see Sect. 3.3).

Next, I make a discussion of the NRR-based experimental result shown in Figs. 15(a) and 16(a). From the result, we can confirm that the NRRs of the proposed BSSA are greater than those of the conventional ICA and ICA+SS by more than 3 dB. However, we can see that the distortion of the proposed BSSA increases slightly from Figs. 15(b) and 16(b). This is due to the fact that the noise reduction of the proposed BSSA is performed based on SS. However, the amount of increase in the degree of distortion is expected to be negligible.

Finally, we can see the speech recognition result in Figs. 15(c) and 16(c). It is evident that the proposed BSSA is superior to the conventional ICA and ICA+SS.

4.4.3 Experiment in actual world

Finally, I conduct an experiment in an actual railway-station environment. Figure 17 shows a layout of the railway-station environment used in my experiment, where the reverberation time is about 1000 ms; this corresponds to mixing filters of 16000 taps in 16 kHz sampling. I used 16 kHz-sampled signals as test data; the original speech (6 s) convoluted with impulse responses that were recorded in an actual railway-station environment, and to which a real-recorded noise in the

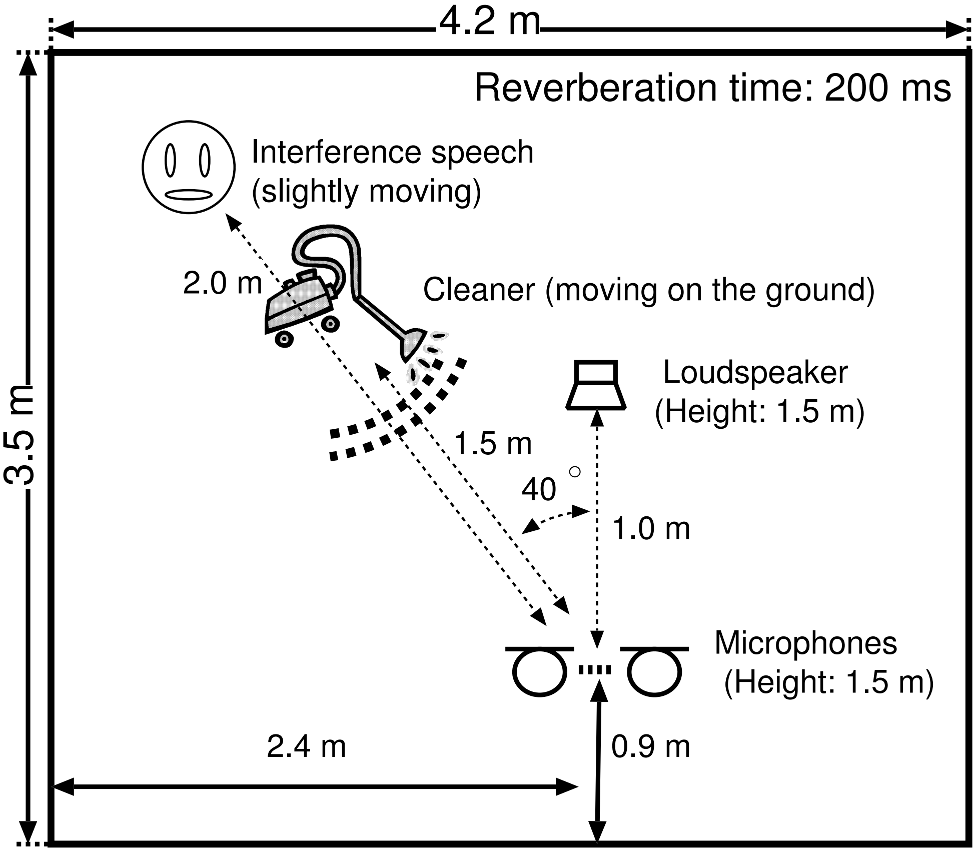


Figure 13. Layout of reverberant room used in my experiment.

environment was added. I use 46 speakers (200 sentences) as the original source of the target speech. The noise in the environment is nonstationary and is almost a non-point-source; consists of various kinds of interference noise, namely, background noise, and the sounds of trains, ticket-vending machines, automatic ticket wickets, foot steps, cars, and wind. Figure 18 shows two typical noises, noise 1 and noise 2, which are recorded in distinct time periods, and used in my experiment. A four-element array with an interelement spacing of 2 cm is used.

Figure 19 shows the real separation results by ICA in a railway-station environment. We can ascertain an imbalanced performance between target estimation and noise estimation similarly to the simulation-based results (see Sect. 3.3).

In the next experiment, I compare the conventional ICA, ICA+SS, and BSSA in terms of NRR, cepstral distortion, and speech recognition performance. Fig-

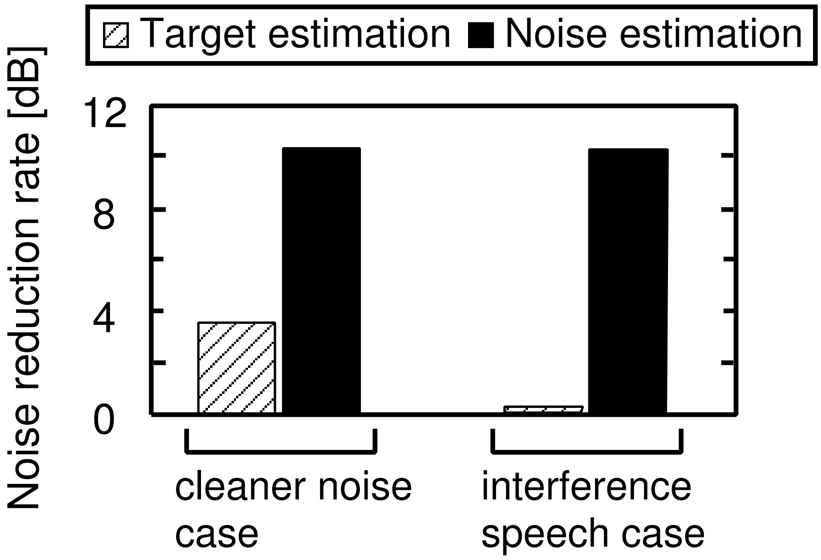


Figure 14. NRR-based separation performance of conventional ICA in environment shown in Fig. 13.

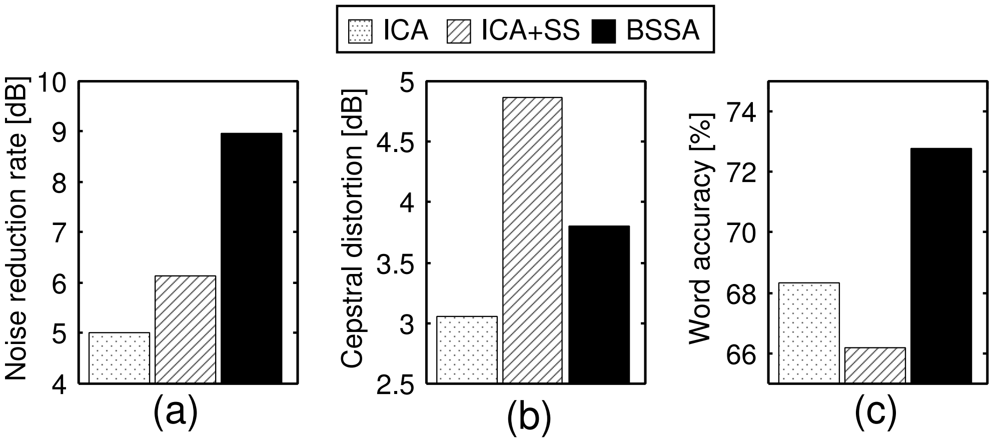


Figure 15. Results of (a) noise reduction rate, (b) cepstral distortion, and (c) speech recognition test for each method (cleaner noise case).

ure 20(a) shows the results of the average of NRR in whole sentences. From these results, we can see that the NRR of BSSA that utilizes ICA as a noise estimator is superior to those of the conventional methods. Figure 21 shows the waveform examples of each method. From this result, we can also see that the noise reduction performance of the proposed BSSA is better than those of the conventional

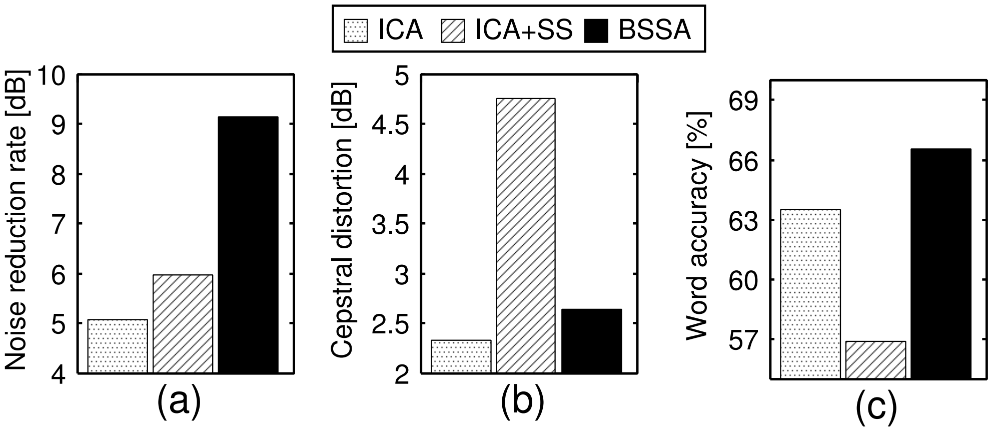


Figure 16. Results of (a) noise reduction rate, (b) cepstral distortion, and (c) speech recognition test using each method (interference speech case).

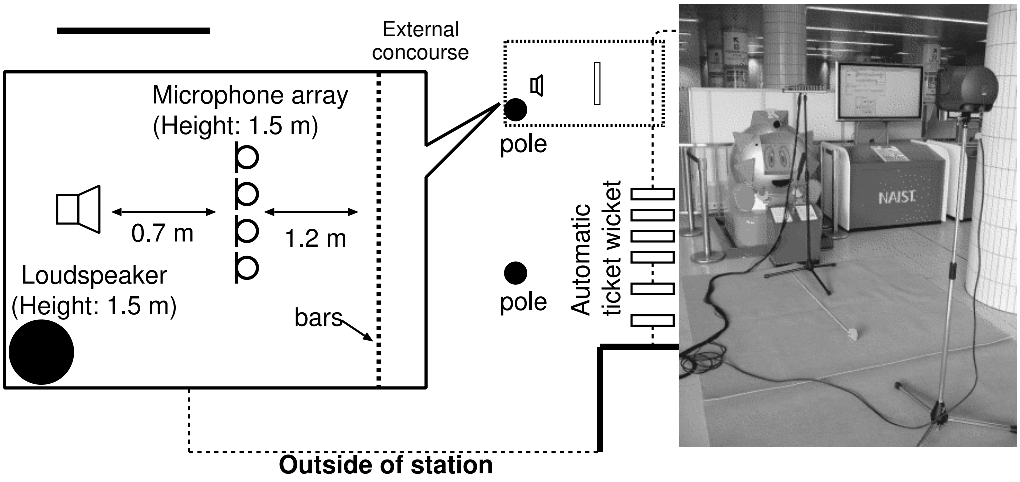


Figure 17. Layout of railway-station environment used in my experiment.

methods. However, we can find that the cepstral distortion in BSSA is increased compared with that in ICA from Fig. 20(b).

Finally, I show results of speech recognition, where the extracted sound quality is completely considered, in Fig. 20(c). The speech recognition task and conditions are the same as those in Sect. 4.4.1, as shown in Table 1. From this result, I conclude that the target-enhancement performance of BSSA, i.e., the method that uses ICA as a noise estimator, is evidently superior to the method that uses ICA

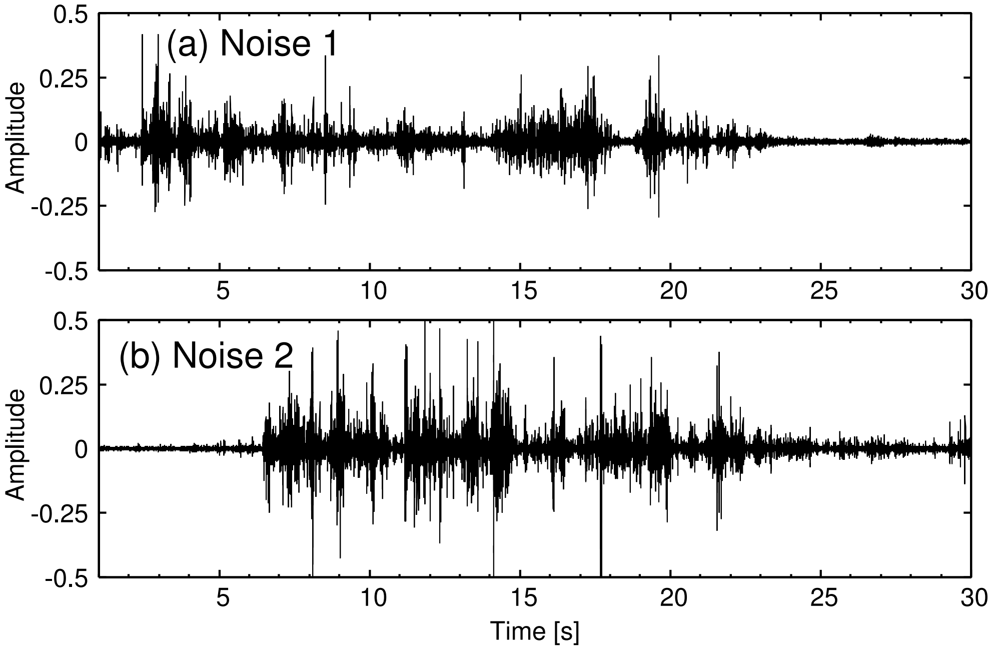


Figure 18. Two typical noises in railway-station environment.

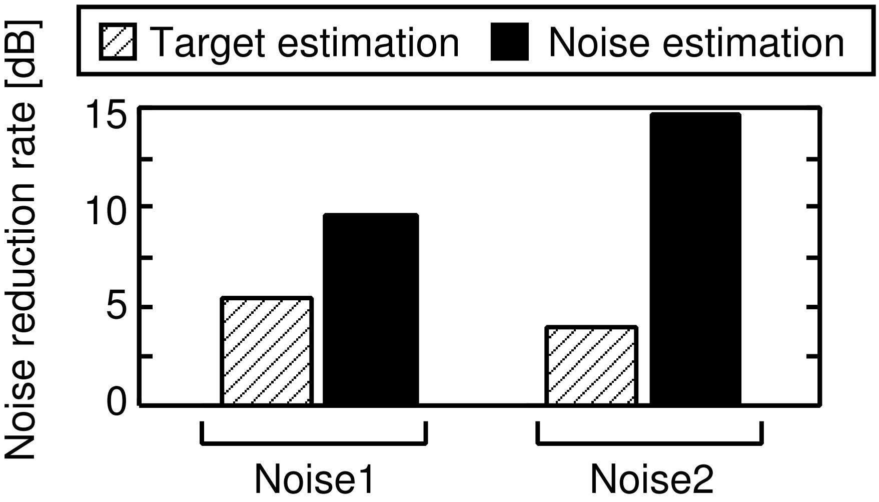


Figure 19. NRR-based noise reduction performance of conventional ICA in railway-station environment.

directly as well as ICA+SS.

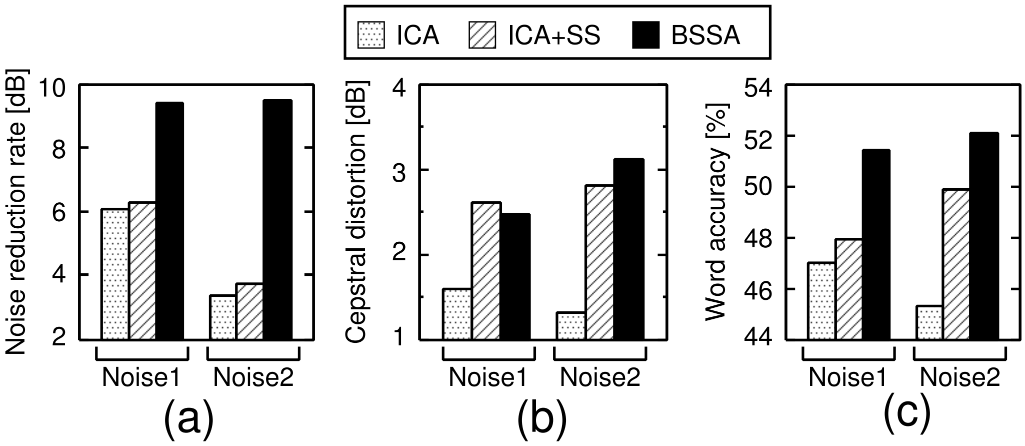


Figure 20. Experimental results of (a) noise reduction rate, (b) cepstral distortion, and (c) speech recognition test, in railway-station environment.

4.5. Conclusion

In this chapter, I proposed the new blind speech extraction method, i.e., BSSA. The proposed BSSA introduces the following two aspects:

BSSA can treat non-point-source noise The conventional ICA, which is the most popular blind source separation method, can work in the limited case such as a speech-speech mixing condition, i.e., all sound sources are point sources. However, an actual environment involves not only point-source interference signals but also non-point-source noise signals. On the other hand, the proposed BSSA architecture can work even under such the realistic acoustic condition. This is because that ICA is proficient in noise estimation under non-point-source noise condition and is utilized as a noise estimator in BSSA.

Permutation-robustness Moreover, the proposed BSSA has permutation-robust structure rather than the the conventional ICA. In the conventional ICA, a special processing is needed to solve the permutation problem. On the other hand, in the proposed BSSA, the structure itself has the robustness against the permutation problem. That is to say, the proposed BSSA can mitigate

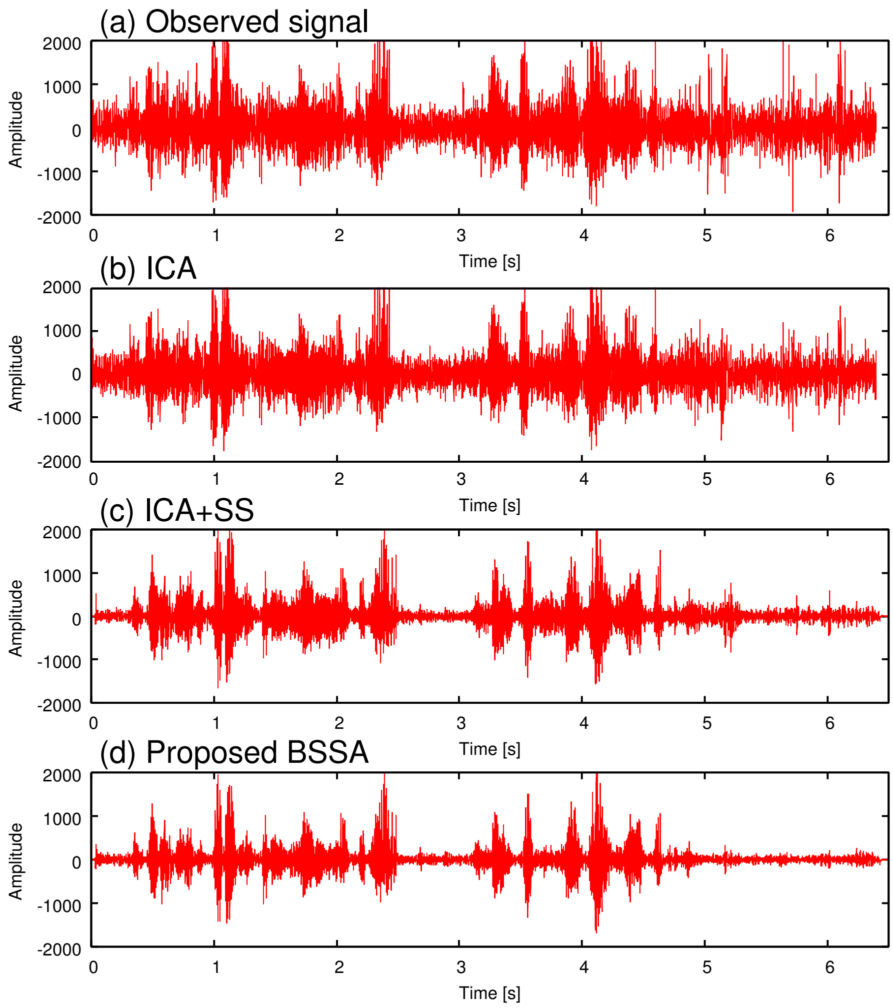


Figure 21. Examples of waveform; (a) observed signal, (b) output signal by ICA, (c) output signal by ICA+SS, and (d) output signal by proposed BSSA.

the negative effect of permutation problem without any special processing. Furthermore, this aspect does not conflict with any conventional permutation solvers. This means that any permutation solvers are available in the ICA part of the proposed BSSA. BSSA reduces the remained permuted components which could not be solved by a permutation solver.

In order to evaluate the efficacy of the proposed BSSA, three experiments were carried out. In the experiments including computer-simulation-based and realrecording-based data, the SNR improvement and speech recognition results of the proposed BSSA were superior to those of conventional methods. These facts evidently indicate that the proposed BSSA is beneficial to speech enhancement in adverse environments.

Chapter 5

Robustness against Reverberation and Microphone Element Errors in BSSA

5.1. Introduction

In the previous chapter, I proposed a new blind source extraction method, BSSA. In the previous chapter, I explained the proposition of BSSA is motivated by the fact that ICA is proficient in noise estimation under non-point-source noise condition. On the other hand, in this chapter, I give an alternative interpretation, i.e., the proposed BSSA is the extension of spatial subtraction array (SSA) [27].

The proposed BSSA involves an ICA-based noise estimation part, and nonlinear subtraction based on the estimated noise is applied to slightly-enhanced-speech signal by DS. Actually, this structure is similar to the conventional SSA[27]. The difference of SSA and BSSA is whether NBF or ICA is utilized for noise estimation part. In this chapter, I give an alternative interpretation of BSSA, i.e., BSSA is an extension of SSA, and I reveal that the proposed BSSA provides the robustness against reverberation and microphone element errors that is the important properties for the real world application.

The chapter is constructed as follows. In the following Sect.5.1, I review the conventional SSA. Next, I bring up the problem of the conventional SSA, and I theoretically analyze the behavior of the noise estimation filter by NBF in SSA and ICA in the proposed BSSA at Sect. 5.3. In the analysis, I give another interpretation of ICA-based noise estimation in the proposed BSSA. As a result of the analysis, I clarify that the proposed BSSA has the robustness against reverberation and microphone element errors. The validity of the analysis is shown via a computer simulation and a speech recognition test in Sect. 5.4. Finally, Sect. 5.5 concludes the chapter.

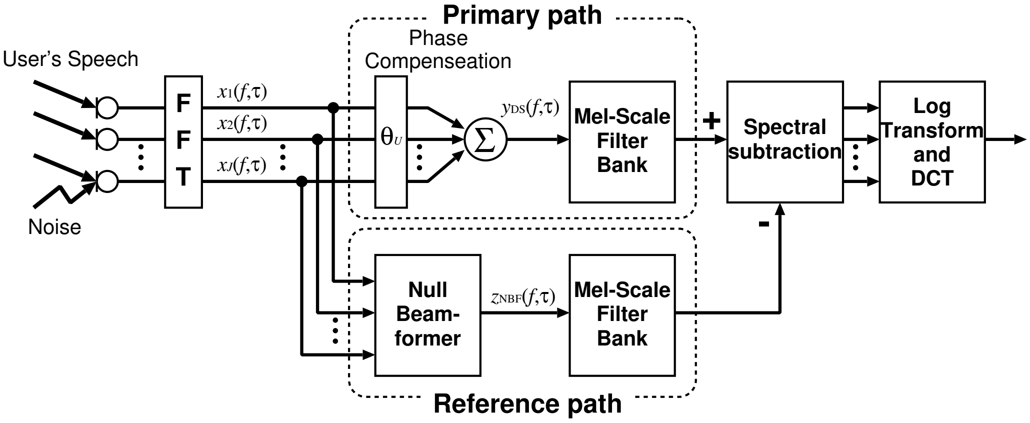


Figure 22. Block diagram of conventional SSA.

5.2. Spatial subtraction array

In this section, I present the conventional SSA [27]. SSA is a kind of non-linear microphone array processing and specifically designed for hands-free speech recognition. Firstly, I exposit the structure and the detailed signal processing of SSA. Next, I point out that the problems of the conventional SSA.

The conventional SSA consists of a DS-based primary path and a reference path via the NBF-based noise estimation (see Fig. 22). The estimated noise component by NBF is efficiently subtracted from the output of the primary path in the power spectrum domain without phase information. Like this, the structure of BSSA is similar to SSA. However, in SSA, the target-speech direction and speech break interval are needed to be known in advance. Thus, SSA is not a blind source extraction method. Detailed signal processing is shown below.

5.2.1 Partial speech enhancement in primary path

The target speech signal is partly enhanced in advance by DS. Thus, the output signal of the primary path is the same as that of BSSA. This procedure can be given as

$$y_{\mathrm{DS}}(f,\tau) = \boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f)\boldsymbol{x}(f,\tau). \tag{71}$$

Note that the look direction of $\boldsymbol{w}_{\mathrm{DS}}(f)$ is known in advance in SSA unlike BSSA.

5.2.2 Noise estimation in reference path

In the reference path, the estimated noise signal is derived by using NBF. This procedure is given as

$$z_{\mathrm{NBF}}(f,\tau) = \boldsymbol{w}_{\mathrm{NBF}}^{\mathrm{T}}(f)\boldsymbol{x}(f,\tau), \tag{72}$$

$$\boldsymbol{w}_{\mathrm{NBF}}(f) = \left\{[1,0]\cdot[\boldsymbol{a}(f,\theta_{\mathrm{O}}),\boldsymbol{a}(f,\theta_{\mathrm{U}})]^{+}\right\}^{\mathrm{T}}, \tag{73}$$

where $z_{\mathrm{NBF}}(f,\tau)$ is the estimated noise by NBF, $\boldsymbol{w}_{\mathrm{NBF}}(f)$ is a NBF-filter coefficient vector which steers the directional null against the DOA of the target speech, θ_{U}, and steers unit gain in the arbitrary direction $\theta_{\mathrm{O}}(\neq\theta_{\mathrm{U}})$. This processing can suppress the target speech arriving from θ_{U}, which is equal to an extraction of noises from sound mixtures if affections of sensor errors and reverberations are not considered. Thus, we can estimate the noise signals by NBF under ideal conditions. Note that $z_{\mathrm{NBF}}(f,\tau)$ is the function of the frame number τ, unlike the constant noise prototype estimated in the traditional SS. Therefore, SSA can also deal with a *non-stationary* noise such as the proposed BSSA.

5.2.3 Noise reduction part

In SSA, noise reduction is carried out by subtracting the estimated noise power spectrum from the partly enhanced target speech power spectrum in power domain. This can be designated as

$$y_{\mathrm{SSA}}(f,\tau) = \begin{cases} \left\{|y_{\mathrm{DS}}(f,\tau)|^2 - \cdot\beta\cdot|z_{\mathrm{NBF}}(f,\tau)|^2\right\}^{\frac{1}{2}} \\ \quad (\text{ if } |y_{\mathrm{DS}}(f,\tau)|^2 - \cdot\beta\cdot|z_{\mathrm{NBF}}(f,\tau)|^2 \geq 0) \\ \eta\cdot|y_{\mathrm{DS}}(f,\tau)| \quad (\text{otherwise}). \end{cases} \tag{74}$$

Like this, signal processing in SSA is almost the same as in BSSA except for the NBF-based noise estimation part.

5.3. Interpretation of BSSA as extended SSA

5.3.1 Problems of SSA

In this section, I denote the problem of the conventional SSA. The NBF-based noise estimator is used in the conventional SSA, but NBF suffers from the adverse effect of the microphone element error and the room reverberation. NBF is a technique to suppress an interference source signal by generating a null against the direction of the interference source signal [19]. If the interference source signal arrives from the same direction as the generated null, we can suppress the interference source signal perfectly. In a reverberant environment, however, the interference source signal arrives from not only the null's direction but also outside of the direction. Therefore, we cannot suppress the interference source signal sufficiently in the reverberant room. In addition, a generic microphone for products, which is not specialized for measurement, often comprises an element error. For some microphones, 3-dB-gain error is permissible in their design [45]. NBF is designed under the ideal assumption in that all elements have the same characteristics. In the real environment, however, the characteristics of each element are different. For these reasons, the directivity pattern shaped by NBF in the ideal environment is apart from that of in the real environment.

Figure 24 illustrates directivity patterns which are shaped by two-element NBF in the ideal (solid line) and the real (dotted line) environment (see Fig. 23) where the reverberation time is 200 ms. In this figure, the null direction is set to zero degree. Here, the ideal directivity pattern $G_{\mathrm{ideal}}(\theta)$ is derived by

$$G_{\mathrm{ideal}}(\theta) = 10 \log_{10} \frac{1}{M} \sum_{f} |\boldsymbol{w}_{\mathrm{NBF}}^{\mathrm{T}}(f) \cdot \boldsymbol{a}(f, \theta)|^2. \tag{75}$$

On the other hand, the directivity pattern in a real environment is given as

$$G_{\mathrm{real}}(\theta) = 10 \log_{10} \frac{1}{C_{\mathrm{adj}} \cdot M} \sum_{f} |\boldsymbol{w}_{\mathrm{NBF}}^{\mathrm{T}}(f) \cdot \boldsymbol{h}_{\mathrm{real}}(f, \theta)|^2, \tag{76}$$

where $\boldsymbol{h}_{\mathrm{real}}(f, \theta) = [h_{\mathrm{real}}^{(1)}(f, \theta), \ldots, h_{\mathrm{real}}^{(J)}(f, \theta)]^{\mathrm{T}}$ is the transfer function vector from a sound source signal to microphones observed in the experimental room illustrated in Fig 23. Besides, C_{adj} is the gain compensation parameter which makes

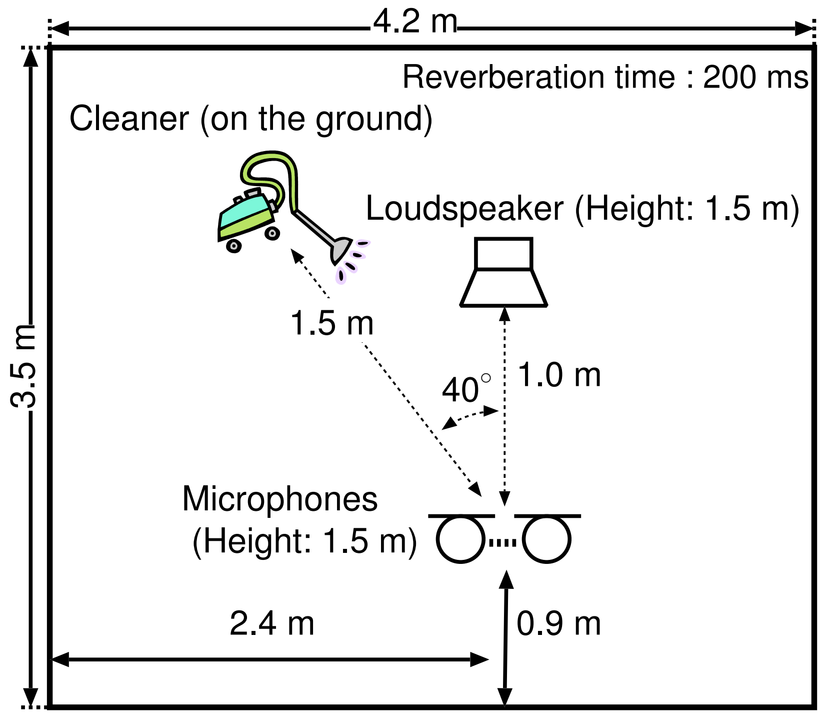


Figure 23. Acoustical environment used in my simulation.

the spatially gain at -90 degrees become zero. From Fig 24, we can see that the depth of the null in the real environment which contains the element error and the reverberation shallows. Therefore, we cannot suppress the interference source signal completely in the real environment by using NBF. Indeed, in SSA, we perform noise estimation via NBF which steers null against the target speech signal, but we cannot suppress the target speech signal sufficiently. In fact, NBF cannot estimate noise signal completely.

5.3.2 Alternative interpretation of ICA in BSSA

In this section, I make an alternative interpretation of ICA-based noise estimation in the proposed BSSA. In the proposed BSSA, ICA is utilizes as a noise estimator instead of NBF. In this section, I point out that ICA can adapt reverberation and microphone element errors which are the problems of NBF in SSA. In this section,

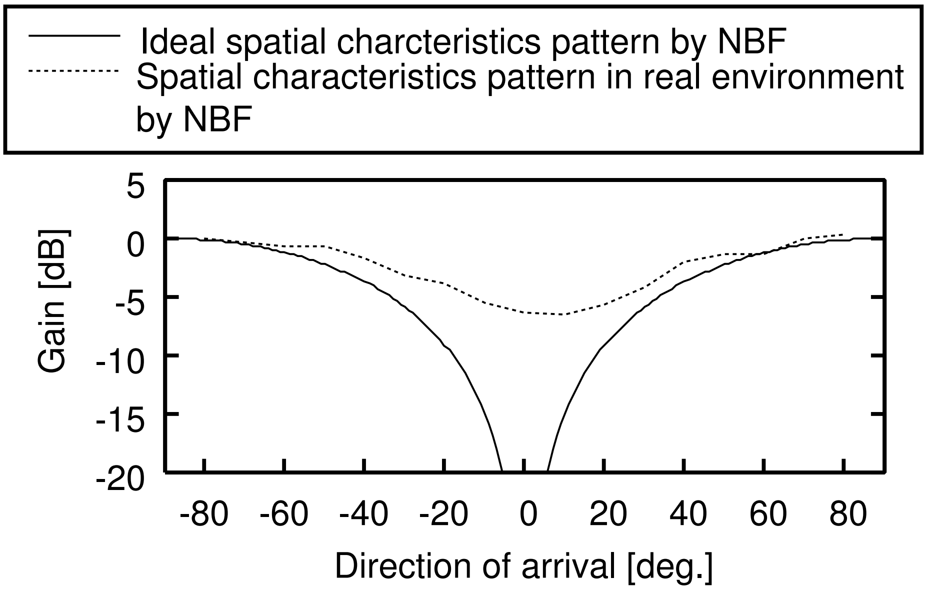


Figure 24. Directivity patterns shaped by NBF in ideal environment and real environment which contains element error and reverberation.

the following observation model as a matter of convenience:

$$\boldsymbol{x}(f,\tau) = \boldsymbol{h}_{\theta_{\mathrm{U}}}(f)s(f,\tau) + \boldsymbol{h}_{\theta_{\mathrm{O}}}(f)n(f,\tau), \tag{77}$$

where $s(f,\tau)$ is the target speech signal, $n(f,\tau)$ is the noise signal. Besides, $\boldsymbol{h}_{\theta_{\mathrm{U}}}(f) = [h_1^{(\theta_{\mathrm{U}})}(f),\ldots,h_J^{(\theta_{\mathrm{U}})}(f)]^{\mathrm{T}}$ is the transfer function vector from the target speech signal to microphones, and $\boldsymbol{h}_{\theta_{\mathrm{O}}}(f) = [h_1^{(\theta_{\mathrm{O}})}(f),\ldots,h_J^{(\theta_{\mathrm{O}})}(f)]^{\mathrm{T}}$ is the transfer function vector from the noise to microphones. These transfer function vectors $\boldsymbol{h}_{\theta_{\mathrm{U}}}(f)$ and $\boldsymbol{h}_{\theta_{\mathrm{O}}}(f)$ involve element errors and room reverberation.

In the proposed BSSA, the source separation is performed by optimized filter by ICA to estimate noise signal. For simplicity, I consider the two-output ICA, namely the mixed observations are separated into a target speech and a noise. Here, the demixing process of ICA defined in (38) and (40) can be rewritten as

$$\boldsymbol{o}(f,\tau) = [o_{\mathrm{n}}(f,\tau), o_{\mathrm{s}}(f,\tau)]^{\mathrm{T}} = \boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{x}(f,\tau), \tag{78}$$

$$\boldsymbol{W}_{\mathrm{ICA}}(f) = \begin{bmatrix} W_{11}^{(\mathrm{ICA})}(f) & \cdots & W_{1J}^{(\mathrm{ICA})}(f) \\ W_{21}^{(\mathrm{ICA})}(f) & \cdots & W_{2J}^{(\mathrm{ICA})}(f) \end{bmatrix}, \tag{79}$$

where $o_{\mathrm{n}}(f,\tau)$ is the estimated noise signal and $o_{\mathrm{s}}(f,\tau)$ is the estimated speech signal.

In the following, I compare the ICA-based noise estimation filter in BSSA and NBF-based noise estimation filter in SSA. Here, (78) can be modified as

$$\boldsymbol{o}(f) = \boldsymbol{D}(f)\tilde{\boldsymbol{W}}(f)[\boldsymbol{h}_{\theta_{\mathrm{O}}}(f), \boldsymbol{h}_{\theta_{\mathrm{U}}}(f)][n(f,\tau), s(f,\tau)]^{\mathrm{T}}, \tag{80}$$

where $\boldsymbol{D}(f)$ is a diagonal matrix that expresses the source-gain ambiguity, and $\tilde{\boldsymbol{W}}_{\mathrm{ICA}}(f)$ is the unmixing matrix removing the source-gain ambiguity from the unmixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$, namely $\boldsymbol{W}_{\mathrm{ICA}}(f) = \boldsymbol{D}(f)\tilde{\boldsymbol{W}}_{\mathrm{ICA}}(f)$. Here, if the source separation by ICA is completely accomplished, the following relation holds:

$$\boldsymbol{W}_{\mathrm{ICA}}(f) = \boldsymbol{D}(f)[\boldsymbol{h}_{\theta_{\mathrm{O}}}(f), \boldsymbol{h}_{\theta_{\mathrm{U}}}(f)]^{+}. \tag{81}$$

Consequently, the noise estimation filter by ICA, $\boldsymbol{w}_{\mathrm{ICA}}(f)$, corresponds to the first row of $\boldsymbol{W}_{\mathrm{ICA}}(f)$, and it can be given as

$$\begin{aligned} \boldsymbol{w}_{\mathrm{ICA}}(f) &= \{[1,0]\cdot\boldsymbol{W}_{\mathrm{ICA}}(f)\}^{\mathrm{T}} \\ &= \{[\sigma(f),0]\cdot[\boldsymbol{h}_{\theta_{\mathrm{O}}}(f), \boldsymbol{h}_{\theta_{\mathrm{U}}}(f)]^{+}\}^{\mathrm{T}} \end{aligned} \tag{82}$$

where $\sigma(f)$ is the entry of the first row in the diagonal matrix $\boldsymbol{D}(f)$. Note that the source-gain ambiguity $\boldsymbol{D}(f)$ is removed by PB method in the proposed BSSA. Here, I compare (82) and (73). The noise estimation filter $\boldsymbol{w}_{\mathrm{ICA}}(f)$ does not utilize the ideal steering vector $\boldsymbol{a}(f,\theta)$ unlike $\boldsymbol{w}_{\mathrm{NBF}}(f)$. Instead of $\boldsymbol{a}(f,\theta)$, the transfer function vectors $\boldsymbol{h}_{\theta_{\mathrm{O}}}(f)$ and $\boldsymbol{h}_{\theta_{\mathrm{U}}}(f)$ which involve reverberation and microphone element errors are used in $\boldsymbol{w}_{\mathrm{ICA}}(f)$. Therefore, ICA-based noise estimation filter can adapt reverberation and microphone element error. Furthermore, this ICA-based noise estimation filter is automatically optimized without pre-measured transfer functions. That is to say, such the ICA-based noise estimation filter is robust against reverberation and microphone element errors. Therefore, the proposed BSSA can be also explained as the extension of the conventional SSA.

5.4. Evaluation

In order to confirm the validity of the above-mentioned analysis, I conduct a computer simulation and a speech recognition test. In the computer simulation, I compare the accuracy of NBF-based noise estimation and ICA-based noise estimation. Moreover, in the speech recognition test, the comparison result of SSA and the proposed BSSA is shown and I reveal that the speech recognition performance of the proposed BSSA outperforms that of the conventional SSA.

5.4.1 Setup

Figure 23 shows the experimental room used in the simulation and speech recognition test. In the experiments, I use the following 16 kHz-sampled signals as test data; the original speech convoluted with impulse responses that were recorded in the experimental room, and to which a real-recorded cleaner noise in the experimental room was added. The cleaner noise is nonstationary and not a point source but consists of several non-stationary noises emitted from, e.g., a motor, air duct and nozzle. Besides, the cleaner noise includes background noise because it is real recorded noise. The input SNR is set to be 5 dB, 10 dB, or 15 dB, and a four-element array with an interelement spacing of 2 cm is used. Moreover, DFT size is 512 points, window size is 256 points, and shift size is 128 points. As for parameters in SS, $\beta = 1.4$ and $\eta = 0.2$ are chosen. These parameters are determined so that the speech recognition performance is maximum.

5.4.2 Comparison of noise estimation accuracy

First, I compare directivity pattern shaped by ICA and NBF in the real environment. The broken line in Fig. 25 is the spatial directivity pattern shaped by ICA in the environment. From this result, we can confirm that the null shaped by ICA becomes deep compared with that of the NBF-based conventional SSA (dotted line in Fig. 25). Therefore, it can be expected that the noise estimation accuracy of ICA is better than that of NBF.

Next, I compare the accuracy of noise estimation by ICA and NBF. I use NRR and the signal-to-distortion ratio (SDR) to compare the noise estimation accuracy,

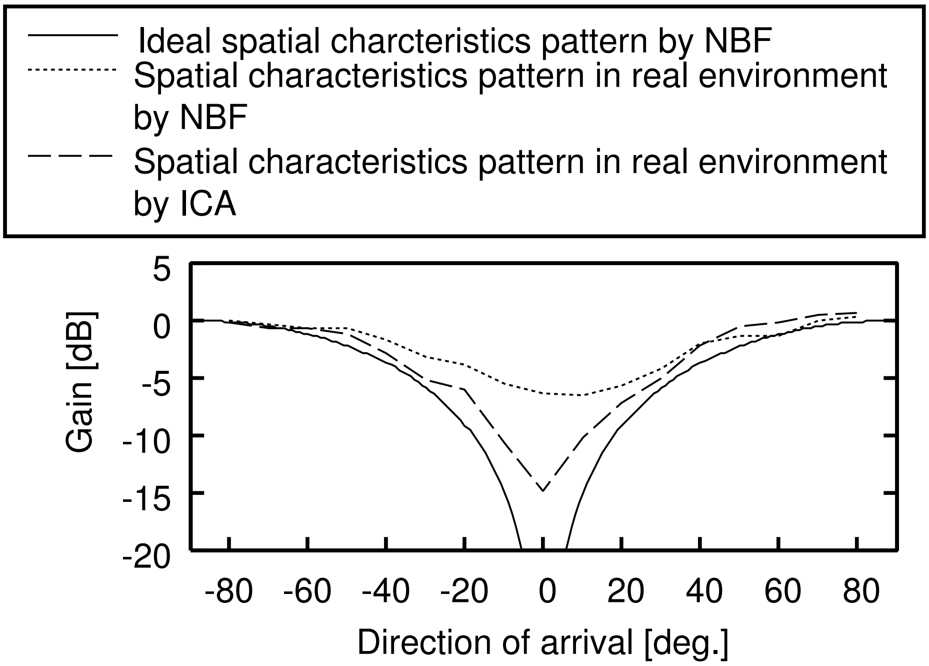


Figure 25. Directivity patterns shaped by ICA in real environment, NBF in real environment, and NBF in ideal environment.

which is defined as

$$\text{SDR [dB]} = 10 \log_{10} \frac{\sum_{\tau} \sum_{f} |n(f,\tau)|^2}{\sum_{\tau} \sum_{f} (|\hat{n}(f,\tau)| - |n(f,\tau)|)^2}, \tag{83}$$

where $n(f,\tau)$ is the true noise signal and $\hat{n}(f,\tau)$ is the estimated noise signal. Here, I consider the target speech signal as interference signal and derive NRR to compare the noise estimation performance in ICA and NBF.

The comparison result is shown in Table 2. The result is the averaged NRR and SDR over 200 utterances. From this result, both NRR and SDR of ICA overtake those of NBF. Therefore, it can be concluded that the noise estimation performance of ICA is better than that of NBF. Moreover, Fig. 26 illustrates an example of long-term-averaged power spectra of estimated noise by ICA and NBF. In the Fig. 26, the black solid line indicates the power spectrum of the noise signal in the primary path, and this power spectrum is needed to be estimated. The gray solid

Table 2. Comparison of noise estimation accuracy by NBF and ICA on the basis of NRR and SDR

	NRR [dB]	SDR [dB]
NBF	4.94	−10.6
ICA	12.4	4.09

line represents the power spectrum of the estimated noise signal by NBF, and the dotted line shows the power spectrum of the estimated noise signal by ICA. We can see that the power spectrum of the estimated noise signal by NBF is not accurate. This is due to that the target speech component still remains in the output of NBF because the null shaped by NBF is shallow. On the other hand, we can see that the power spectrum of the estimated noise signal by ICA is a good estimation because the depth of the null shaped by ICA is enough for suppressing the target speech. From these results, I conclude that ICA-based noise estimation is more accurate than NBF-based one and is robust against reverberation and microphone element errors.

5.4.3 Speech recognition test

Finally, I compare DS, the conventional SSA, and the proposed BSSA on the basis of word accuracy scores. Table 1 describes the conditions for speech recognition, and I use 46 speakers (200 sentences) as original speech. Figure 27 shows the word accuracy in each method. Here, “Unprocessed” refers to the result without any noise reduction processing. From this result, we can see that the word accuracy of the proposed method is obviously superior to those of the conventional methods. This is a promising evidence that the proposed BSSA has the robustness against reverberation and microphone element errors compared to the conventional SSA.

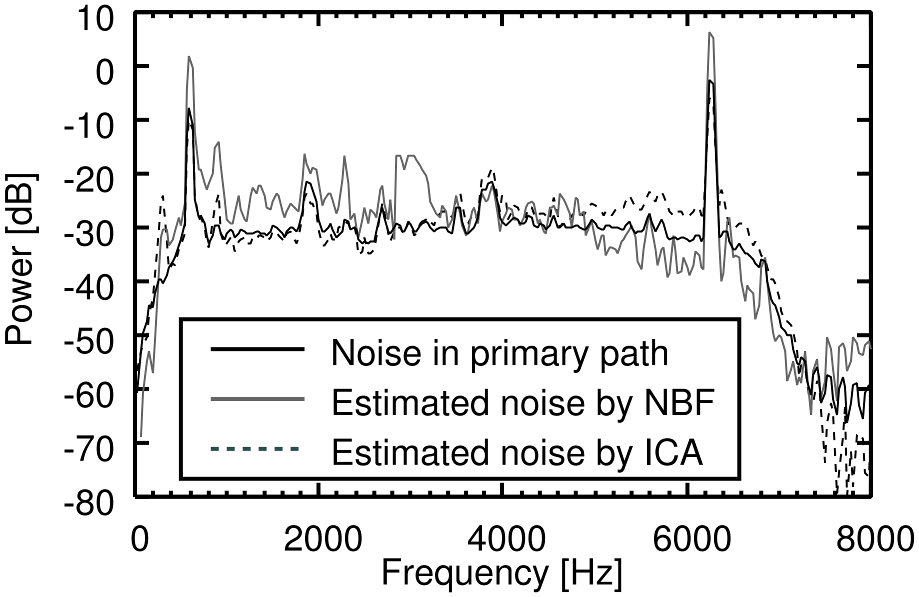


Figure 26. Comparison of long-term averaged power spectra estimated by NBF and ICA.

Table 3. Experimental conditions for speech recognition

Database	JNAS [44], 306 speakers (150 sentences / 1 speaker)
Speech recognition task	20 k words newspaper dictation
Acoustic model	phonetic tied mixture (PTM) [44], clean model
Number of training speakers for acoustic model	260 speakers (150 sentences / 1 speaker)
Decoder	Julius [44] ver 3.5.1

5.5. Conclusion

In this chapter, I gave an alternative interpretation of the proposed BSSA, i.e., the proposed BSSA is an extension of SSA. The difference of the proposed BSSA and the conventional SSA is whether ICA or NBF is utilized for noise estimation.

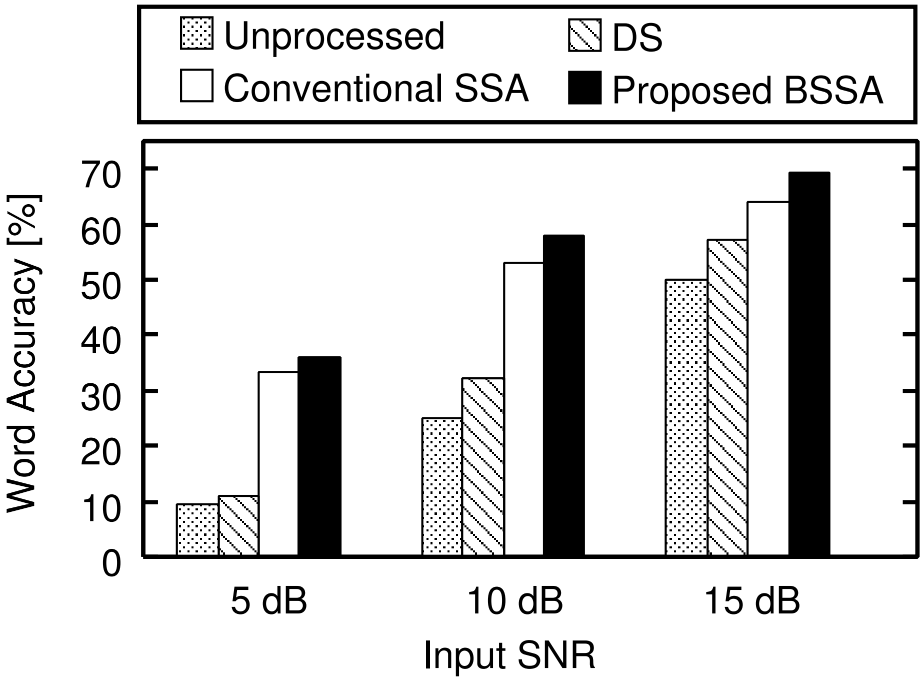


Figure 27. Results of word accuracy in various methods, unprocessed, DS, SSA, and robust SSA, used in simulation experiment.

Then, I analyzed the difference of the behavior of ICA-based noise estimation and NBF-based noise estimation. As a result of the analysis, I revealed that ICA-based noise estimation introduces the robustness against reverberation and microphone element errors, which is one important property for actual world applications. The propriety of the analysis was confirmed by a computer simulation and a speech recognition tests. As a result of experiments, ICA could estimate noise signals accurately compared with NBF. Moreover, the proposed BSSA could achieve better speech recognition performance than the speech recognition performance of the conventional SSA. Therefore, let me conclude that the proposed BSSA provides the robustness against reverberation and microphone element error because the proposed BSSA utilizes ICA as a noise estimator.

CHAPTER 6

REAL-TIME IMPLEMENTATION OF PROPOSED BSSA FOR HANDS-FREE SPOKEN-ORIENTED GUIDANCE SYSTEM

6.1. Introduction

In Chapter 4, I proposed a new blind speech extraction method, i.e. BSSA, and gave strong evidences of the efficacy of the proposed BSSA thorough experiments. However, the proposed algorithm described in Chapter 4 is basically a batch algorithm. That is to say, the proposed BSSA itself cannot be applied to applications require real-time processing, e.g., hands-free spoken-oriented guidance system. To work in real-time is one of the indispensable factors for a hands-free spoken-oriented system. Indeed BSSA can reduce noise efficiently, but BSSA is difficult to work in real-time because ICA part of BSSA consumes huge amount of computational complexities. Thus, it is required to develop a real-time architecture of BSSA.

In this chapter, I newly propose the real-time architecture of BSSA and implement the real-time BSSA. Moreover, I introduce the implemented real-time BSSA into the spoken-oriented guidance system “Kitarobo” which has already been installed at an actual railway station, and construct a hands-free spoken-oriented dialogue system. Although many real-time robot audition systems have been proposed [3], the behavior and performance are not explicitly analyzed under heavy widespread noise condition, e.g., an actual railway-station, as far as I know. Then, I evaluate the constructed hands-free spoken dialogue system with the real-time BSSA in an experimental room simulating actual railway-station environment based on the speech recognition test, and 6% improvement of the speech recognition result can be confirmed compared with the conventional speech enhancement methods.

This chapter is constructed as follows. In the following Sect. 6.2, I describe the strategy of the real-time architecture of BSSA. Next, detailed algorithm of the proposed real-time BSSA is explained in Sect. 6.3, and then the constructed hands-free spoken-oriented guidance system with the proposed real-time BSSA is illustrated in Sect 6.4. The following Sect. 6.5 gives evaluation results, and Sect. 6.6 concludes the chapter.

6.2. Strategy of real-time implementation of BSSA

The proposed BSSA can be decomposed to the following parts:

- **Partial speech enhancement part in primary path**
  In the primary path, DS is applied to the multichannel observation, and the partly-enhanced speech signal is obtained.

- **Noise estimation part in reference path**
  In the reference path, noise estimation is performed based on ICA. Then, the noise estimation part in reference path is additionally decomposed to the following parts:

  - **ICA optimization part**
    In the ICA optimization part, noise estimation filter is optimized by ICA.
  - **Noise estimation part by optimized filter**
    In this part, noise estimation is performed by the optimized noise estimation filter by ICA.

- **SS part for final output**
  In this part, the final output signal of BSSA is yielded by subtracting the power spectrum of estimated noise signal in reference path from the power spectrum of partly-speech-enhanced signal in primary path.

DS part in the primary math, filtering optimized noise estimation filter to observation in the reference path, and the SS part for the final output are possible to

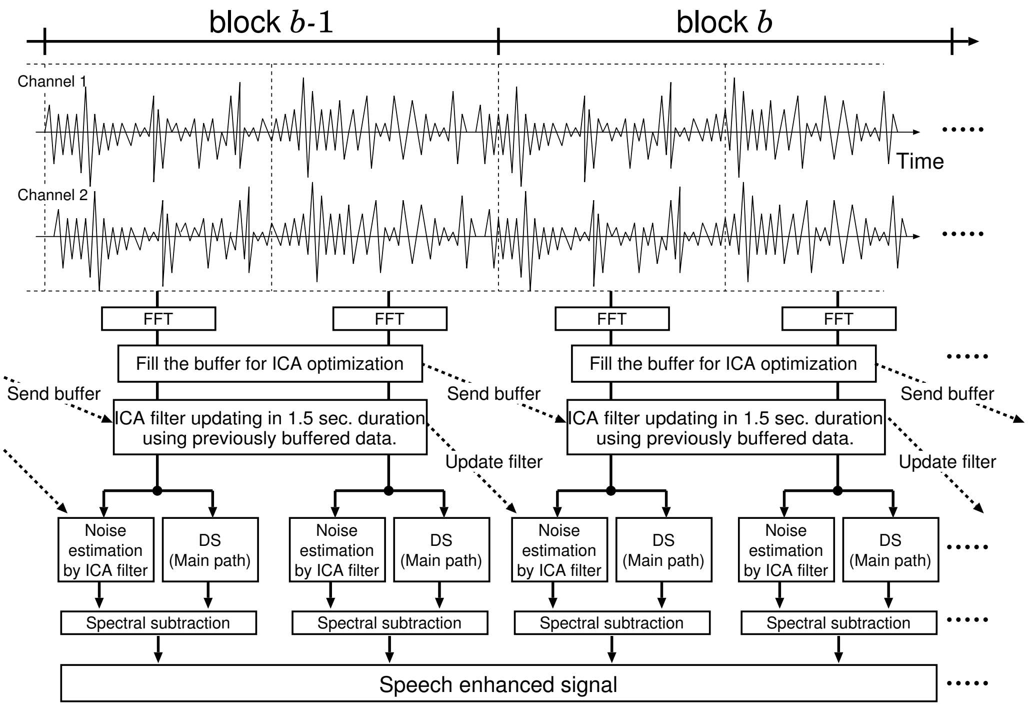

block b-1
block b
Channel 1
Channel 2
Time
FFT
Fill the buffer for ICA optimization
Send buffer
ICA filter updating in 1.5 sec. duration using previously buffered data.
Update filter
Noise estimation by ICA filter
DS (Main path)
Spectral subtraction
Speech enhanced signal

Figure 28. Signal flow in real-time implementation of proposed method.

work in real-time because these parts are enough simple and low-complexity signal processing. However, it is toilsome to optimize (update) the separation filter in real-time because the optimization of the unmixing matrix by ICA consumes huge amount of computational costs. Therefore, I introduce a strategy in that the separation filter optimized by using the past time period data is applied to the current data. Figure 28 illustrates a configuration of a real-time implementation for BSSA. Signal processing in this implementation is performed via the following manner.

S1. Inputted signals are converted into time-frequency domain series by using a frame-by-frame fast Fourier transform (FFT).

S2. ICA is conducted using the past 1.5-s-duration data for estimating separation filter while the current 1.5 s. The optimized separation filter is applied to the next (*not current*) 1.5 s samples. This staggered relation is due to the fact that the filter update in ICA requires substantial computational complexities and cannot provide the optimal separation filter for the current 1.5 s data.

S3. Inputted data is processed in two paths. In the primary path, target speech is partly enhanced by DS. In the reference path, ICA-based noise estimation is conducted. Again, note that the separation filter for ICA is optimized by using the past time period data.

S4. Finally, we obtain the target-speech-enhanced signal by subtracting the power spectrum of the estimated noise signal in the reference path from the power spectrum of the primary path's output.

Although the separation filter update in the ICA part is not real-time processing but involves totally a latency of 3.0 seconds, the entire system still seems to run in real-time because DS, SS and separation filtering can work in the current segment with no delay. In the system, the performance degradation due to the latency problem in ICA is mitigated by oversubtraction in SS. Detailed *real-time* signal processing is shown in the following sections.

6.3. Algorithm

In this section, I represent detailed signal processing of the real-time architecture of the proposed BSSA. Since the ICA part of the proposed BSSA needs huge amount of complexities, I divide the parts of BSSA into the part of ICA and the other parts. Consequently, the signal processing parts in the proposed BSSA are classified into the following two blocks:

- **Optimization of noise estimation filter by ICA**
  This is a part of reference path. In this part, the noise estimation filter is optimized by ICA.

- **Noise reduction part**
  This part includes the primary path, noise estimation with using optimized filter by ICA, and SS for final output.

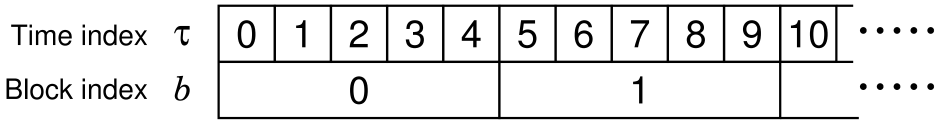


Figure 29. Relation between time index and block index ($\ell_{\text{sample}} = 5$ case).

These blocks are parallelly and independently executed in the proposed real-time architecture. In the following, I described the detailed signal processing of each block.

6.3.1 ICA part in real-time algorithm

In the ICA part of this algorithm, a sequential time-series input is divided into fixed-length blocks, and ICA is performed in each block. The number of samples in one block, ℓ_{sample}, is defined as

$$\ell_{\text{sample}} = \left\lfloor \frac{\ell_{\text{sec}} \cdot f_{\text{s}}}{T_{\text{shift}}} \right\rfloor, \tag{84}$$

where ℓ_{sec} is block length in seconds (I use 1.5 s in this study), T_{shift} is frame shift size for short-time Fourier transform, and $\lfloor \cdot \rfloor$ is the floor function. Thus, a set of time frame index belonging to a block b $(= 0, 1, 2, \ldots)$, T_b, can be given as

$$T_b = \{\, \tau \mid b \cdot \ell_{\text{sample}} \le \tau < (b+1) \cdot \ell_{\text{sample}} \}. \tag{85}$$

Figure 29 shows the relation between a time frame index and a block index, where, e.g., $\ell_{\text{sample}} = 5$.

The unmixing matrix for a block b, $\boldsymbol{W}^{\text{ICA}}_{(b)}(f)$, is optimized by the following iterative update equation:

$$\begin{aligned}\left[\boldsymbol{W}^{\text{ICA}}_{(b)}(f)\right]^{[p+1]} = {} & \mu[\boldsymbol{I} - \langle \boldsymbol{\varphi}(\hat{\boldsymbol{o}}(f,\tau))\hat{\boldsymbol{o}}^{\text{H}}(f,\tau)\rangle_{\tau \in T_b}] \left[\boldsymbol{W}^{\text{ICA}}_{(b)}(f)\right]^{[p]} \\ & + \left[\boldsymbol{W}^{\text{ICA}}_{(b)}(f)\right]^{[p]},\end{aligned} \tag{86}$$

where $\langle \cdot \rangle_{\tau \in T_b}$ is the time-averaging operator which is localized within block T_b, and $\hat{\boldsymbol{o}}(f,\tau) = [\hat{o}_1(f,\tau), \ldots, \hat{o}_K(f,\tau)]^{\text{T}}$ is the temporal separated signal vector given as

$$\hat{\boldsymbol{o}}(f,\tau) = \boldsymbol{W}^{\mathrm{ICA}}_{(b)}(f)\boldsymbol{x}(f,\tau) \quad (\tau \in T_b). \tag{87}$$

Here, if the average power of the specific block b is very small, the unmixing matrix should not be updated because the low-power block which does not contains any dominant signals leads to an unstable convergence of the unmixing matrix. Thus, the unmixing matrix is not updated in such a block b if the average power of the block b is very small. This can be represented by

$$\boldsymbol{W}^{\mathrm{ICA}}_{(b)}(f) = \boldsymbol{W}^{\mathrm{ICA}}_{(b-1)}(f) \quad (\text{If} \ \ \langle |x(f,\tau)|^2 \rangle_{\tau\in T_b} < th_{\mathrm{pow}}), \tag{88}$$

where th_{pow} is the threshold for the average power.

Moreover, the initial value of the unmixing matrix in the optimization at each block is represented by

$$\left[\boldsymbol{W}^{\mathrm{ICA}}_{(b)}(f)\right]^{[0]} = \begin{cases} \boldsymbol{W}_{\mathrm{initial}}(f) & (\text{if} \quad b \bmod b_{\mathrm{reset}} = 0), \\ \boldsymbol{W}^{\mathrm{ICA}}_{(b-1)}(f) & (\text{otherwise}), \end{cases} \tag{89}$$

where b_{reset} is the reset period of the unmixing matrix, and $\boldsymbol{W}_{\mathrm{initial}}(f)$ is the initial value of the unmixing matrix given in advance. This initial value is ordinarily generated using the observed signal via some methods, e.g., principle component analysis or beamforming. Thus, the optimized unmixing matrix is reset into the given initial value every b_{reset} blocks.

Furthermore, we can estimate DOAs from the unmixing matrix $\boldsymbol{W}^{\mathrm{ICA}}_{(b)}(f)$ as described in (41). This procedure is represented by

$$\theta_{u,b} = \sin^{-1}\left(\frac{\left[[\boldsymbol{W}^{\mathrm{ICA}}_{(b)}(f)]^{-1}\right]_{ju} / \left[[\boldsymbol{W}^{\mathrm{ICA}}_{(b)}(f)]^{-1}\right]_{j'u}}{2\pi f_s c^{-1}(d_j - d_{j'})}\right), \tag{90}$$

where $\theta_{u,b}$ is the DOA of the u-th sound source in the block b. Then, I choose the U-th source signal which is the nearest the front of the microphone array, and designate the DOA of the chosen source signal as $\theta_{U,b}$ in this study. This is because almost all users often stand in front of the microphone array in a spoken-oriented human-machine interface.

6.3.2 Noise reduction part in real-time algorithm

Noise reduction is carried out according to the following three steps;

1. First, DS is performed to enhance the target signal (primary path).
2. Next, we estimated noise signal based on ICA (reference path).
3. Finally, we obtain the target speech enhanced signal by subtracting the power spectrum of the estimated noise from the power spectrum of the primary path's output.

In the primary path, DS is performed to enhance the target speech signal. This procedure can be represented by

$$y_{(b)}^{\mathrm{DS}}(f,\tau) = \boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f,\theta_{U,b-2})\boldsymbol{x}(f,\tau) \ \ (\tau \in T_b), \tag{91}$$

where $y_{(b)}^{\mathrm{DS}}(f,\tau)$ is the primary path's output in a block b.

In the reference path, first, the signal separation is performed. This can be designated as

$$\boldsymbol{o}_{(b)}(f,\tau) = \boldsymbol{W}_{(b-2)}^{\mathrm{ICA}}(f)\boldsymbol{x}(f,\tau) \ \ (\tau \in T_b), \tag{92}$$

where $\boldsymbol{o}_{(b)}(f,\tau) = [o_{1,b}(f,\tau),\ldots,o_{K,b}(f,\tau)]^{\mathrm{T}}$ is the separated signal vector in a block b. Next, we obtain the estimated noise signal in a block b, $z_{(b)}(f,\tau)$, as

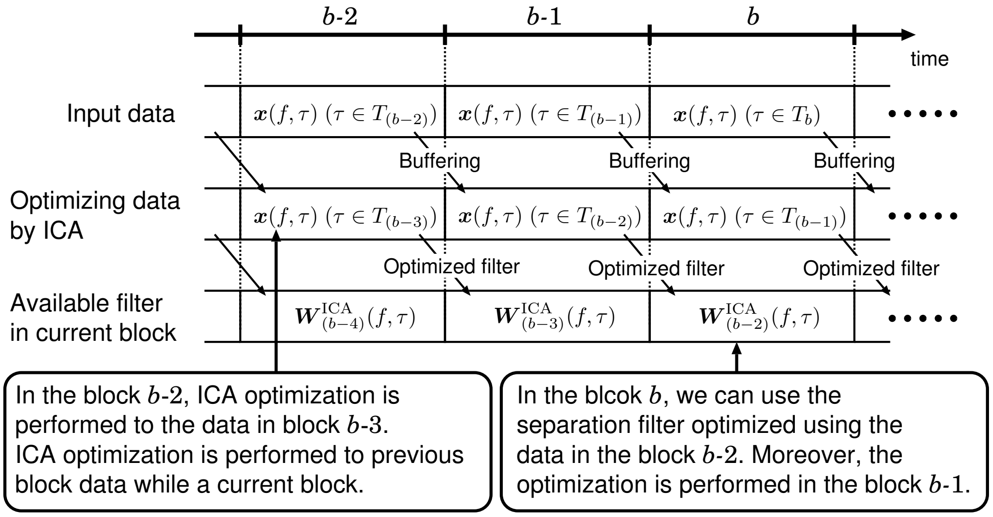
$$z_{(b)}(f,\tau) = \boldsymbol{g}_{\mathrm{DS}}^{\mathrm{T}}(f,\theta_{U,b-2})\left[\boldsymbol{W}_{(b-2)}^{\mathrm{ICA}}(f)\right]^{+}\boldsymbol{q}_{(b)}(f,\tau) \ (\tau \in T_b), \tag{93}$$

$$\begin{aligned}\boldsymbol{q}_{(b)}(f,\tau) = [&o_{1,b}(f,\tau),\ldots,o_{U-1,b}(f,\tau),0,\\ &o_{U+1,b}(f,\tau),\ldots,o_{K,b}(f,\tau)]^{\mathrm{T}},\end{aligned} \tag{94}$$

where $\boldsymbol{q}_{(b)}(f,\tau)$ is the vector in which the target speech component is removed.

Finally, we obtain the target speech enhanced signal $y_{(b)}^{\mathrm{BSSA}}(f,\tau)$ by SS. This can be given as

$$y_{(b)}^{\mathrm{BSSA}}(f,\tau) = \begin{cases} \left\{|y_{(b)}^{\mathrm{DS}}(f,\tau)|^2 - \beta\cdot|z_{(b)}(f,\tau)|^2\right\}^{\frac{1}{2}}, \\ \quad (\text{ if } \ |y_{(b)}^{\mathrm{DS}}(f,\tau)|^2 - \beta\cdot|z_{(b)}(f,\tau)|^2 \geq 0 \ ) \\ \gamma\cdot|y_{(b)}^{\mathrm{DS}}(f,\tau)| \quad \text{(otherwise).} \end{cases} \tag{95}$$


b-2
b-1
b
time
Input data
$\boldsymbol{x}(f,\tau)\ (\tau \in T_{(b-2)})$
$\boldsymbol{x}(f,\tau)\ (\tau \in T_{(b-1)})$
$\boldsymbol{x}(f,\tau)\ (\tau \in T_{b})$
Buffering
Buffering
Buffering
Optimizing data by ICA
$\boldsymbol{x}(f,\tau)\ (\tau \in T_{(b-3)})$
$\boldsymbol{x}(f,\tau)\ (\tau \in T_{(b-2)})$
$\boldsymbol{x}(f,\tau)\ (\tau \in T_{(b-1)})$
Optimized filter
Optimized filter
Optimized filter
Available filter in current block
$\boldsymbol{W}^{\mathrm{ICA}}_{(b-4)}(f,\tau)$
$\boldsymbol{W}^{\mathrm{ICA}}_{(b-3)}(f,\tau)$
$\boldsymbol{W}^{\mathrm{ICA}}_{(b-2)}(f,\tau)$

In the block b-2, ICA optimization is performed to the data in block b-3. ICA optimization is performed to previous block data while a current block.

In the blcok b, we can use the separation filter optimized using the data in the block b-2. Moreover, the optimization is performed in the block b-1.

Figure 30. Configuration of updating separation filter.

In (91) and (93), note that we have only to use the estimated DOA and the optimized unmixing matrix in the previous block $b-2$. This is due to data buffering and optimization process for ICA. ICA optimization requires a certain length of data, e.g., 1.5 s. data. Thus, we must buffer a certain length of input data for ICA optimization. Consequently, ICA optimization just starts after the buffering. Moreover, ICA optimization cannot finish in no time at all because ICA optimization consumes huge amount of computations. Thus ICA optimization is performed while one block. As a result, in a current block b, we are only admitted to utilize the separation filter optimized in the block $b-2$ (see Fig. 30). By the same manner, we can only apply the estimated DOA of the block $b-2$ to a current block b.

6.3.3 Algorithmic delay

For a real-time system, delay-time is a crucial factor. Hereinafter, I asses the algorithm delay of the proposed real-time BSSA.

The algorithmic delay of the proposed real-time BSSA only depends on the following; (a) DS filtering, (b) noise estimation by the separation filter, and (c) buffer size. Although ICA optimization is parallelly performed, the optimization result cannot be applied to current block. Thus, ICA optimization does not yield the algorithmic delay. In DS and the separation filter, for reducing the effect of the circular convolution, the main pulse of the filter is located at the center of the filter. Thus, the resultant signal of the filtering is delayed, and its delay is the half of the filter length. Note that the noise estimation is performed in parallel with DS. Therefore, the total delay of DS filtering and noise estimation is also the half of the filter length. Moreover, the buffer size for reading data from the hardware cannot be negligible. Consequently, the algorithm delay of the final output can be given by

$$\text{Delay [points]} = \text{Buffer Size} + \text{Filter Size}/2. \tag{96}$$

For instance, supposed that the buffer size is 512 points and the filter length is also 512 points, the algorithmic delay of the final output of the real-time BSSA is 768 points. This corresponds to 48 ms delay with 16 kHz sampling.

6.3.4 Robustness against acoustical environment change

Under an actual world, the change of acoustical environment, i.e., user's move, the change of noise environment, and so on, is considerable factor. However, it cannot be regarded that a user moves while the user talks to a hands-free spoken-oriented guidance system. Furthermore, we can easily classify target voice or interference voice by asking users to locate themselves in front of the system. Consequently, the gravest problem on the change of acoustical environment is the change of noise environment.

By the way, the proposed BSSA is subtracting the estimated noise by removing target speech signal, from the partly-enhanced-speech signal via DS. Thus, the proposed BSSA can reduce noises even if the noise environment is momentarily changing while an appropriate noise estimation filter is optimized. In the real-time BSSA, the noise estimation filter is learned with the past data block, e.g. 3.0 s. Since that filter removes the target speech signal, the filter can estimate noise accurately even if the noise environment of a block for optimizing noise estimation filter and a current block are different. For these reasons, it can be expected that the noise reduction performance of the real-time BSSA is not so degraded.

6.4. Hands-free robot spoken-oriented dialogue system with real-time BSSA

I introduce the real-time BSSA into the robot spoken-oriented dialogue system "Kitarobo" [46] which has already been installed in an actual railway station. In this study, I replace the input device of Kitarobo, i.e., a close-talking microphone, with the real-time BSSA to construct the hands-free robot spoken-oriented dialogue system. Figures 31 and 32 show an overview of the system and appearance of the hands-free robot spoken-oriented dialogue system with the real-time BSSA. Unlike the conventional Kitarobo, the input device is substituted with the real-time BSSA. Details of Kitarobo are described in the following subsection.

6.4.1 Brief review of spoken-oriented dialogue system "Kitarobo"

The spoken-oriented guidance robot "Kitarobo" is working in an actual railway station since end of March 2006. The system is installed besides the ticket gate and is adjacent to each other. Everybody can use the systems while the station is open. Since the station faces to a road, an automobile engine sound and sound of a bus horn are also inputted to the system. Kitarobo provides guidance information to visitors regarding issues on the station or around the station without resting. Kitarobo only can exchange one question and one answer, that is to say, any dialogue histories are not taken into account. The input device of the original Kitarobo is a close-talking microphone. Thus the original Kitarobo is not a hands-free system and is weak against the surrounding noises.

In original Kitarobo, firstly, an input signal is classified into valid speech or non-speech based on Gaussian mixture model [47]. If the input signal is regarded as non-speech, the input signal is dropped. Next, the voice activity detection (VAD) is applied to the input signal and the voice period of the input signal is

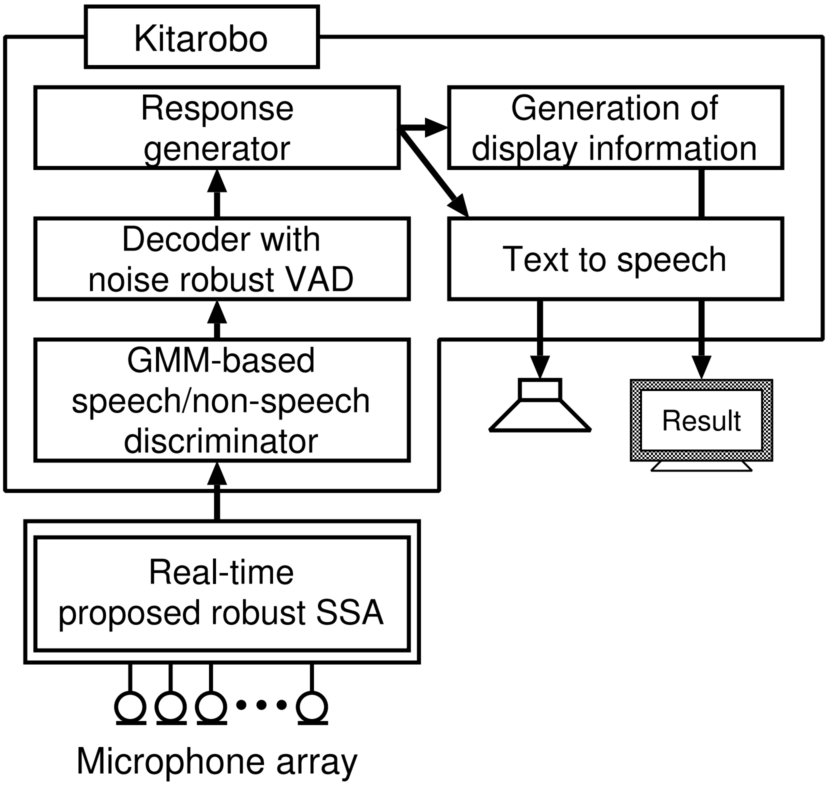

Kitarobo
Response generator
Generation of display information
Decoder with noise robust VAD
Text to speech
GMM-based speech/non-speech discriminator
Result
Real-time proposed robust SSA
Microphone array

Figure 31. Overview of hands-free robot spoken-oriented dialogue system with real-time BSSA.

clipped. In Kitarobo, speech-decoder-based VAD by Sakai et al. [48], that is robust against noise contaminated signals, is adopted. In the speech-decoder-based VAD, the speech recognition and VAD is performed at the same time. According to the result of speech recognition, responses are generated. Finally, based on the generated response, response sound is synthesized by text-to-speech synthesizer and information demand on the input speech is displayed. Reference [46] helps you to understand further details of Kitarobo.

6.4.2 Implementation of real-time BSSA

The proposed real-time BSSA is implemented by C/C++ on Debian/GNU Linux [49] 4.0 platform. Also, I utilize the computer with Intel Xeon X5355 processor 2.66 G

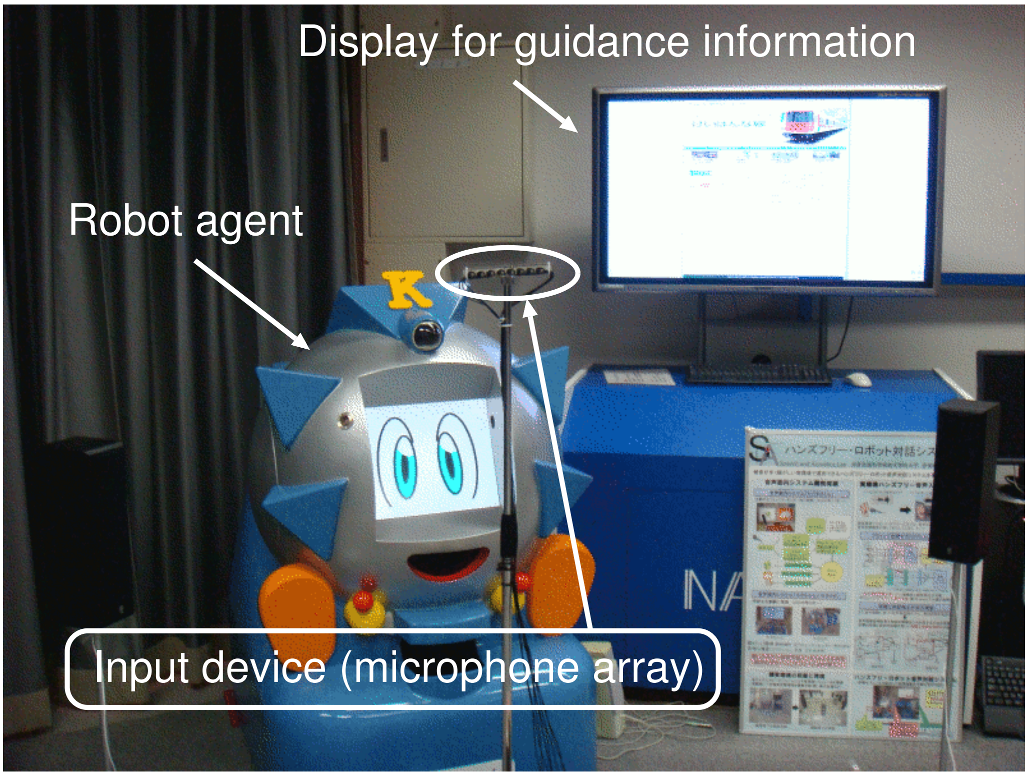

Display for guidance information
Robot agent
Input device (microphone array)

Figure 32. Appearance of my hands-free robot spoken-oriented dialogue system with real-time BSSA.

Hz for the implementation. The implemented real-time BSSA consumes about 52 M Bytes RAM. Moreover, I use RME Hammerfall DSP Multiface for AD/DA converter. In the implementation, the configuration of AD/DA is 16 kHz sampling frequency and 16 bits quantization. The parameters for frame-by-frame DFT analysis are the following; DFT size is 512 points, window size is 256 points, and shift size is 128 points. Although it seems to require high-spec hardware, Mori et al. have succeeded at the real-time implementation of ICA on a general purpose DSP [22]. The computational complexity of the proposed real-time BSSA is almost the same as the real-time ICA, i.e., DS and SS are only added compared with the real-time ICA. Thus, it can be expected that the real-time BSSA is also implemented on general purpose DSP.

Though the computational complexities of the proposed real-time BSSA depend on the number of filter update in ICA, the implemented real-time BSSA updates noise estimation filter to the maximum extent within the current 1.5 s. That is to say, the real-time factor always becomes 1.0 in the implementation. As a result of the implementation, 61 times filter update is performed on average with the utilized computer.

I adjust the buffer size and the filter size in the real-time BSSA so that the delay-time in the real-time BSSA becomes about 5% of that in the original Kitarobo. In the original Kitarobo, the averaged response time, i.e., the delay-time from a speech input to a response output, is about 994 ms. Thus, the buffer size is fixed to 512 points and filter size is also fixed to 512 points so that the algorithmic delay-time of the real-time BSSA becomes 48 ms. The actual delay-time of the implemented real-time BSSA is about 56.5 ms. There exists 8.5 ms difference between algorithmic delay and measured delay. This difference is caused by hardware latency.

6.4.3 Simulating railway-station noise

The main task of Kitarobo is station guidance, and always working in an actual railway-station. Thus, it is difficult to conduct various BSSA experiments in an arbitrary time. Therefore, I have a necessity to construct the noise environment simulator of railway-station for experiments. To solve the problem, our laboratory has constructed the experimental room for hands-free spoken dialogue system with the real-time BSSA. The experimental room contains Kitarobo with the real-time BSSA and railway-station noise simulator. The noise simulation is performed in the following;

1. Record noises in an actual railway station. In the experiment, eight-channel directional microphones are used to record the multi-channel railway-station noise.

2. Playback the multi-channel recorded railway-station noise by eight surrounded loudspeaker (see Fig. 33).

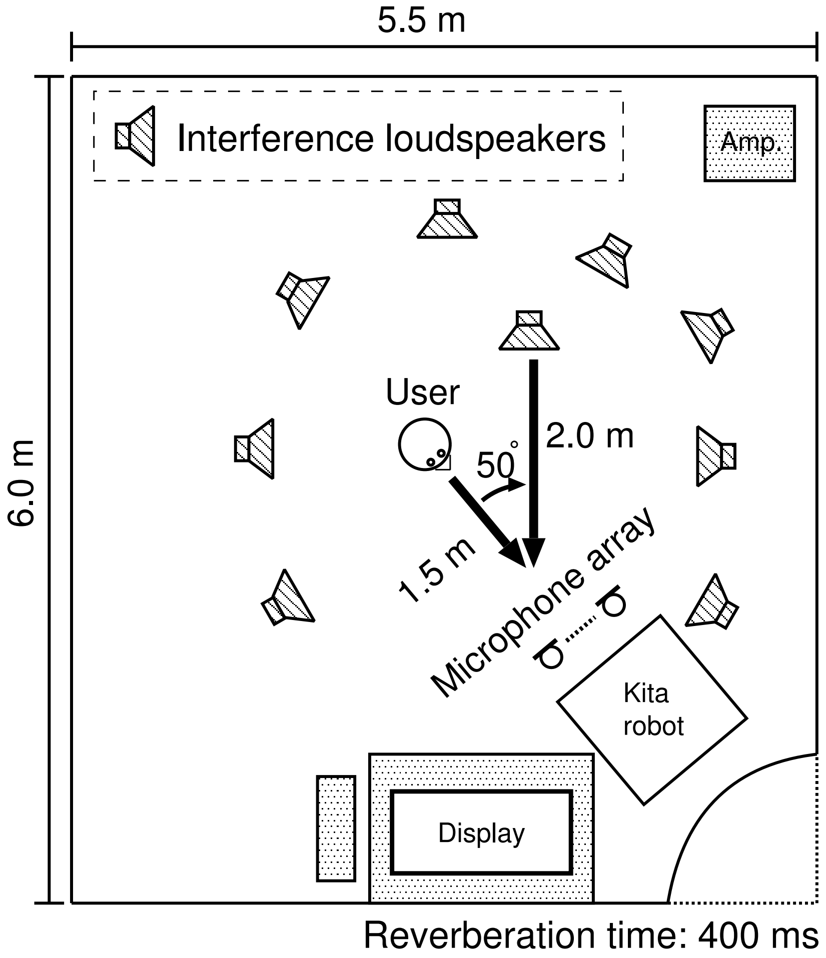


Figure 33. Layout of reverberant room in my experiment.

This noise consists of various kinds of interference noises, namely, background noise, sounds of trains, ticket-vending machines, automatic ticket wickets, foot steps, cars, and wind. In addition, this noise is highly nonstationary.

6.5. Evaluation of implemented system

6.5.1 Configurations of evaluation

To evaluate the hands-free spoken dialogue system with the real-time BSSA, the speech recognition test is conducted. In the experiment, in order to evaluate only the speech recognition performance of the real-time BSSA, the dialogue part in the system is stopped. Since Kitarobo exchanges one question and one answer, each response is independently-generated of each input speech. Thus, we can evaluate the hands-free Kitarobo by speech recognition test. I compare the speech recognition performance of the proposed real-time BSSA, the off-line BSSA, the real-time ICA, and DS. In the off-line BSSA, the number of filter update in ICA part is aligned to that of real-time BSSA, i.e., 61 times.

Figure 33 depicts a layout of a reverberant room in our experiment where the reverberation time is about 400 ms. The following real-recorded 16 kHz-sampled signals were used in the experiments. The target signal is real-recorded user's speech which is talked in front of a microphone array and 1.5 m apart from the array. The contents of the target utterances are all related to Kitarobo task, i.e., questions about transfer, station's facilities, sights around the station, and so on. As for noise, two noises were added simultaneously. First noise is the real-recorded noise in an actual railway-station noise (it simulates railway-station noise) emitted from surrounded 8 loudspeakers. Second noise is an interference speech located at 50 degrees in the right direction of the microphone array, and its distance is 2.0 m. I use 5 speakers (250 words) as target user, and Julius [44] ver. 4.0 RC2 as speech decoder. An eight-element array with the interelement spacing of 2.15 cm is used. The array consists of directional microphone SHURE MX-184.

6.5.2 Experimental result

Figure 34 shows speech recognition result. In the result, I describe the word correct (WC) score in addition to WA score. The WC score is defined as

$$\mathrm{WC}\ [\%] \equiv \frac{W_{\mathrm{WC}} - S_{\mathrm{WC}} - D_{\mathrm{WC}}}{W_{\mathrm{WC}}} \times 100, \tag{97}$$

where W_{WC} is the number of words, S_{WC} is the number of substitution errors, and D_{WC} is the number of dropout errors. In the WC score, the number of insertion errors is neglected unlike WA score defined in (69). In a spoken-oriented dialogue system, whether 'words' are properly decoded or not is crucially important.

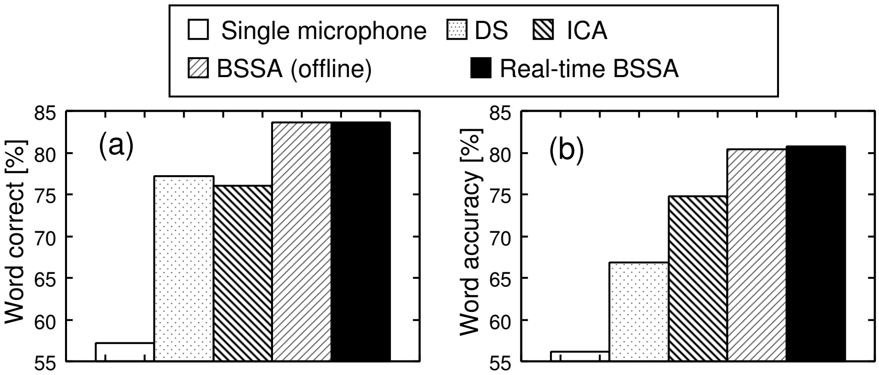


Figure 34. Result of speech recognition test in (a) word correct, and (b) word accuracy.

Hence, I show the WC score along with the WA score.

From this result, we can see that both the WC score and the WA score of the proposed BSSA are obviously superior to those of DS and the conventional ICA. In particular, 8% (in WC) or 6% (in WA) improvement of the speech recognition result can be confirmed. Besides, the WC performance is over 80% that is sufficient speech recognition performance to construct spoken-oriented guidance system. Furthermore, we can confirm that the speech recognition performance of the proposed real-time BSSA and the off-line BSSA is almost the same. This results implies robustness of the proposed real-time BSSA against the change of noise environment described in Sect. 6.3.4. From the result, it can be concluded the proposed real-time BSSA can achieve sufficient speech recognition performance for hands-free spoken-oriented guidance system.

6.6. Conclusion

In this chapter, I proposed a real-time architecture of the proposed BSSA. Based on the proposed real-time BSSA, I constructed hands-free spoken-oriented guidance system. As a speech recognition test, the speech recognition performance of the proposed real-time BSSA outperformed those of conventional methods.

Also, the speech recognition performance of the proposed real-time BSSA was almost the same as that of the off-line BSSA. Furthermore, the proposed real-time BSSA could realize sufficient short delay of its algorithm with keeping enough speech recognition performance, e.g., about 50 ms in the implementation. For these reasons, I conclude that the proposed BSSA can accomplish hands-free spoken-oriented guidance system.

Chapter 7

Musical Noise and Its Objective Measure

7.1. Introduction

In the previous chapter, it was demonstrated that the proposed BSSA integrating ICA and SS can achieve good noise reduction performance. However, a serious problem still exists in BSSA; artificial distortion (the so-called *musical noise*) [29] due to nonlinear SS. Since the artificial distortion causes discomfort to users, it is desirable that we control musical noise through signal processing. However, in almost all nonlinear noise reduction methods, the strength parameter to mitigate the musical noise in nonlinear signal processing is determined heuristically. Although there have been some studies on reducing musical noise [29] and on nonlinear signal processing with less musical noise [30], evaluations mainly depended on subjective tests by humans, and no objective evaluations have been performed to the best of my knowledge.

In recent years, it was reported that the amount of generated musical noise is strongly related to the difference between higher-order statistics (HOS) before and after nonlinear signal processing [35]. This fact enables us to analyze the amount of musical noise arising through nonlinear signal processing. Furthermore, on the basis of HOS, a mathematical metric for musical-noise generation in an objective manner has been established [35]. Uemura et al. have analyzed single-channel nonlinear signal processing based on the objective metric and clarified features of the amount of musical noise. Hereafter, I give an analysis of the amount of musical noise generated via methods of integrating microphone array signal processing and SS on the basis of HOS.

Methods of integrating microphone array signal processing and nonlinear signal processing such as the proposed BSSA can be basically classified into two types. Figure 35 shows a typical architecture used for the integration of microphone array signal processing and SS, where SS is performed after beamforming. Thus, I call this type of architecture *BF+SS*. The proposed BSSA can be classi-

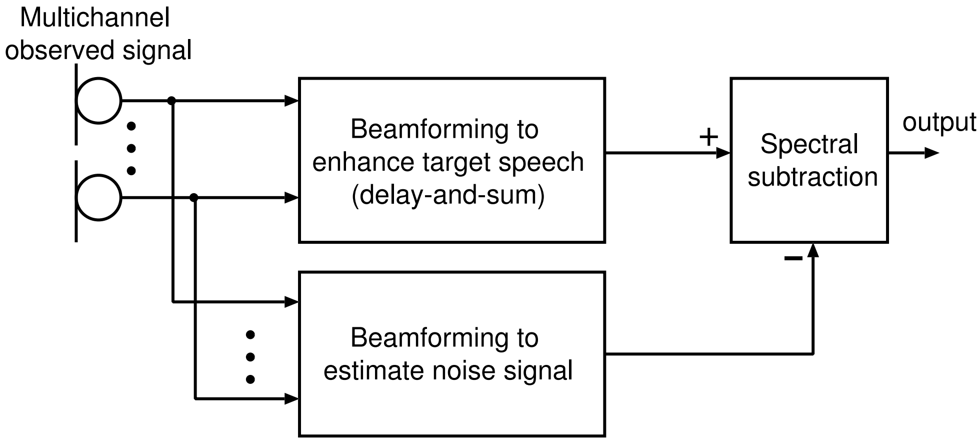


Figure 35. Block diagram of architecture for SS after beamforming (BF+SS).

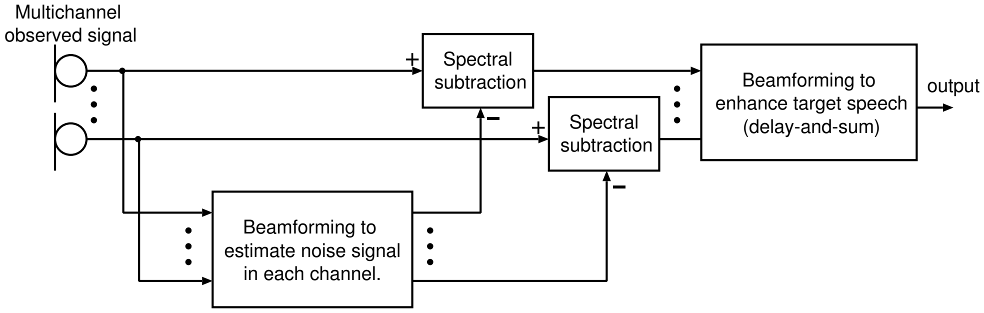


Figure 36. Block diagram of architecture for channelwise SS before beamforming (chSS+BF).

fied into this BF+SS. Such a structure is adopted in many integration methods, e.g., [25, 27]. On the other hand, the integration architecture illustrated in Fig. 36 is an alternative architecture used when SS is performed before beamforming. Such a structure is less commonly used, but some integration methods use this structure [26, 28]. In this architecture, channelwise SS is performed before beamforming, and I call this type of architecture *chSS+BF*.

In the following chapters, I would analyze these two architectures on the basis of HOS and obtain the following results:

- The amount of musical noise generated strongly depends on not only the oversubtraction parameter of SS but also *the statistical characteristics of* *the input signal*.

- Except for the specific condition that the input signal is Gaussian, the noise reduction performances of the two methods are not equivalent even if we set the same SS parameters.

- As a result of analysis under equivalent noise reduction performance conditions, chSS+BF generates less musical noise than BF+SS for almost all practical cases.

The most important contribution is that these findings are mathematically proved. In particular, the amount of musical noise generated and the noise reduction performance resulting from the integration of microphone array signal processing and SS are analytically formulated on the basis of HOS. Although there have been many studies on optimization methods based on HOS, this is the first time they have been used for musical-noise assessment. The validity of the analysis based on HOS, is demonstrated via a computer simulation and a subjective evaluation by humans. Moreover, this analysis can be applied to BSSA as well as other methods of integrating microphone array signal processing and SS.

In this chapter, first, I describe the two methods of integrating microphone array and SS in Sect. 7.2. Next I give a brief review of musical noise and its objective metric based on HOS in Sect 7.3. Finally, Sect. 7.4 concludes this chapter. The musical-noise analysis of SS, microphone array signal processing, and their integration methods are discussed in the next chapter.

7.2. Methods of integrating microphone array signal processing and SS

In this section, the formulations of the two methods of integrating microphone array signal processing and SS are described. First, BF+SS, which is a typical method of integration, is formulated. Next, an alternative method of integration, chSS+BF, is introduced.

7.2.1 Sound mixing model

In this and the next chapters, I consider one target speech signal and an additive interference signal. Hence, the sound mixing model defined in (1) can be rewritten as

$$\boldsymbol{x}(f,\tau) = \boldsymbol{h}(f)s(f,\tau) + \boldsymbol{n}_a(f,\tau), \tag{98}$$

where $\boldsymbol{h}(f) = [h_1(f), \ldots, h_J(f)]^{\mathrm{T}}$ is the transfer function vector of target speech signal $s(f,\tau)$.

7.2.2 SS after beamforming

In BF+SS, the single-channel target-speech-enhanced signal is first obtained by beamforming, e.g., DS. Next, single-channel noise estimation is performed by a beamforming technique, e.g., null beamformer [19] or adaptive beamforming [39]. Finally, we extract the resultant target-speech-enhanced signal via SS. The full details of signal processing are given below.

To enhance the target speech, DS is applied to the observed signal. This can be represented by

$$y_{\mathrm{DS}}(f,\tau) = \boldsymbol{w}_{\mathrm{DS}}(f,\theta_{\mathrm{U}})^{\mathrm{T}}\boldsymbol{x}(f,\tau). \tag{99}$$

Finally, we obtain the target-speech-enhanced spectral amplitude based on SS. This procedure can be expressed as

$$|y_{\mathrm{SS}}(f,\tau)| = \begin{cases} \sqrt{|y_{\mathrm{DS}}(f,\tau)|^2 - \beta \cdot \mathrm{E}_\tau[|\hat{n}(f,\tau)|^2]} \\ \qquad\qquad (\text{where } |y_{\mathrm{DS}}(f,\tau)|^2 - \beta \cdot \mathrm{E}_\tau[|\hat{n}(f,\tau)|^2] \geq 0), \\ \eta \cdot |y_{\mathrm{DS}}(f,\tau)| \qquad\qquad (\text{otherwise}), \end{cases} \tag{100}$$

where β is the oversubtraction parameter, η is the flooring parameter, and $\hat{n}(f,\tau)$ is the estimated noise signal, which can be generally obtained by some beamforming techniques, e.g., fixed or adaptive beamforming. In BSSA, noise estimation is performed through ICA. $\mathrm{E}_\tau[\cdot]$ denotes the expectation operator with respect to the time-frame index.

7.2.3 Channelwise SS before beamforming

In chSS+BF, we first perform SS independently in each input channel then we derive a multichannel target-speech-enhanced signal by channelwise SS. This can be expressed as

$$|y_j^{(\mathrm{chSS})}(f,\tau)| = \begin{cases} \sqrt{|x_j(f,\tau)|^2 - \beta \cdot \mathrm{E}_\tau[|\tilde{n}_j(f,\tau)|^2]} \\ \qquad\qquad (\text{where } |x_j(f,\tau)|^2 - \beta \cdot \mathrm{E}_\tau[|\tilde{n}_j(f,\tau)|^2] \geq 0), \\ \eta \cdot |x_j(f,\tau)| \qquad\qquad (\text{otherwise}), \end{cases} \tag{101}$$

where $y_j^{(\mathrm{chSS})}(f,\tau)$ is the target-speech-enhanced signal obtained by SS at a specific channel j and $\tilde{n}_j(f,\tau)$ is the estimated noise signal in the jth channel. For instance, the multichannel noise can be estimated by single-input multiple-output ICA (SIMO-ICA) [50] or a combination of ICA and the projection back method [16]. These techniques can provide the multichannel estimated noise signal, unlike traditional ICA. SIMO-ICA can separate mixed signals not into monaural source signals but into SIMO-model signals at the microphone. Here “SIMO” represents the specific transmission system in which the input signal is a single source signal and the outputs are its transmitted signals observed at multiple microphones. Thus, the output signals of SIMO-ICA maintain the rich spatial qualities of the sound sources. Also the projection back method provides SIMO-model-separated signals using inverse of an optimized ICA filter.

Finally, the target-speech-enhanced signal can be extracted by applying DS to $\boldsymbol{y}_{\mathrm{chSS}}(f,\tau) = [y_1^{(\mathrm{chSS})}(f,\tau), \ldots, y_J^{(\mathrm{chSS})}(f,\tau)]^{\mathrm{T}}$. This procedure can be expressed by

$$y(f,\tau) = \boldsymbol{w}_{\mathrm{DS}}^{\mathrm{T}}(f,\theta_{\mathrm{U}})\boldsymbol{y}_{\mathrm{chSS}}(f,\tau), \tag{102}$$

where $y(f,\tau)$ is the final output of chSS+BF.

Such a chSS+BF structure performs DS after (multichannel) SS. Since DS is basically signal processing in which the summation of the multichannel signal is taken, it can be considered that interchannel smoothing is applied to the multichannel spectral-subtracted signal. On the other hand, the resultant output signal of BF+SS remains as it is after SS. That is to say, it can be expected that the output signal of chSS+BF is more natural (contains less musical noise) than that of BF+SS.

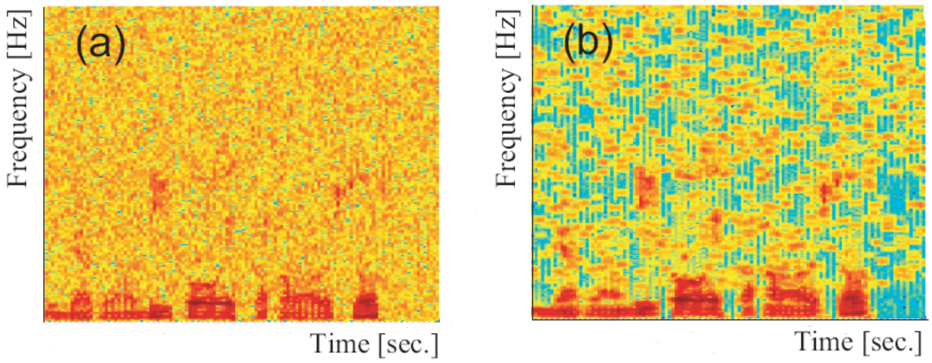


Figure 37. (a) Observed spectrogram and (b) processed spectrogram.

7.3. Kurtosis-based musical-noise generation metric

7.3.1 Introduction

Uemura et al. have been reported that the amount of musical noise generated is strongly related to the difference between the kurtosis of a signal before and after signal processing [35]. Thus, I can analyze the amount of musical noise generated through BF+SS and chSS+BF on the basis of the change in the measured kurtosis. Hereinafter, I give details of the kurtosis-based musical-noise metric.

7.3.2 Relation between musical-noise generation and kurtosis

Generally, musical noise can be considered as the audible isolated spectral components generated through signal processing. Figure 37(b) shows an example of a spectrogram of musical noise in which many isolated components can be observed. Then, it can be speculated that the amount of musical noise is strongly related to the number of such isolated components and their level of isolation.

Hence, Uemura et al. have introduced kurtosis to quantify the isolated spectral components, and they focus their attention on the changes in kurtosis. Since isolated spectral components are dominant, they are heard as tonal sounds, which results in our perception of musical noise. Therefore, it is expected that obtaining the number of tonal components will enable us to quantify the amount of musical noise. However, such a measurement is extremely complicated, so instead they have introduced a simple statistical estimate, i.e., kurtosis.

This strategy allows us to obtain the characteristics of tonal components. The adopted kurtosis can be used to evaluate the width of the probability density function (p.d.f.) and the weight of its tails, i.e., kurtosis can be used to evaluate the percentage of tonal components among the total components. A larger value indicates a signal with a heavy tail in its p.d.f., meaning that it has a large number of tonal components. Also, kurtosis has the advantageous property that it can be easily calculated in a concise algebraic form.

7.3.3 Kurtosis

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

$$\mathrm{kurt}_x = \frac{\mu_4}{\mu_2^2}, \tag{103}$$

where x is a random variable, kurt_x is the kurtosis of x, and μ_n is the nth-order moment of x. Here μ_n is defined as

$$\mu_n = \int_{-\infty}^{+\infty} x^n P(x) \mathrm{d}x, \tag{104}$$

where $P(x)$ denotes the p.d.f. of x. Note that this μ_n is not a central moment *but a raw moment*. Thus, (103) is not kurtosis according to the mathematically strict definition, but a modified version; however, I refer to (103) as kurtosis in this study.

7.3.4 Kurtosis ratio

Although we can measure the number of tonal components by kurtosis, it is worth mentioning that kurtosis itself is not sufficient to measure musical noise. This is because that the kurtosis of some unprocessed signals such as speech signals is also high, but we do not perceive speech as musical noise. Since we aim to count only the musical-noise components, we should not consider genuine tonal components. To achieve this aim, we should focus on the fact that musical noise is generated only in artificial signal processing. Hence, we should consider the change in kurtosis during signal processing. Consequently, the following *kurtosis ratio* [35] has been proposed to measure the kurtosis change:

$$\text{kurtosis ratio} = \frac{\text{kurt}_{\text{proc}}}{\text{kurt}_{\text{input}}}, \tag{105}$$

where $\text{kurt}_{\text{proc}}$ is the kurtosis of the processed signal and $\text{kurt}_{\text{input}}$ is the kurtosis of the input signal. A larger kurtosis ratio ($\gg 1$) indicates a marked increase in kurtosis as a result of processing, implying that a larger amount of musical noise is generated. On the other hand, a smaller kurtosis ratio ($\simeq 1$) implies that less musical noise is generated. It has been confirmed that this kurtosis ratio closely matches the amount of musical noise in a subjective evaluation based on human hearing [35].

7.4. Conclusion

In this chapter, I pointed out the problem of the methods of integrating microphone array signal processing and SS such as the proposed BSSA, i.e., musical noise problem. Next, I mentioned the typical methods of integrating microphone array and SS. Finally, I gave a brief explanation of objective measure of musical noise on the basis of HOS.

Chapter 8

Kurtosis-Based Musical-Noise Analysis for Microphone Array Signal Processing and SS

8.1. Introduction

In the previous chapter, the objective measure for the amount of musical noise generated on the basis of HOS, which is kurtosis ratio, was described. In this chapter, I perform an analysis on musical-noise generation in BF+SS and chSS+BF on the basis of kurtosis.

The analysis is composed of the following three parts:

- First, an analysis on musical-noise generation in BF+SS and chSS+BF based on kurtosis that does not take noise reduction performance into account is performed in Sect. 8.3.
- The noise reduction performance is analyzed in Sect. 8.4, and I reveal that the noise reduction performances of BF+SS and chSS+BF are not equivalent. Moreover, a flooring parameter design to align the noise reduction performances of BF+SS and chSS+BF is also derived for the fair comparison of BF+SS and chSS+BF.
- The kurtosis-based comparison between BF+SS and chSS+BF under the same noise reduction performance condition is carried out in Sect. 8.5.

Note that my analysis has no limitations regarding assumptions on the statistical characteristics of noise, thus, all noises including Gaussian and super-Gaussian noise can be considered.

8.2. Signal model used for analysis

Musical-noise components generated from the noise-only period are dominant in spectrograms (see Fig. 37); hence, I mainly focus my attention on musical-noise components originating from input noise signals.

Moreover, to evaluate the resultant kurtosis of SS, we introduce a gamma distribution to model the noise in the power domain [51, 52, 53]. The p.d.f. of the gamma distribution for random variable x is defined as

$$P_{\mathrm{GM}}(x) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} \cdot x^{\alpha-1} \exp\left\{-\frac{x}{\theta}\right\}, \tag{106}$$

where $x \geq 0, \alpha > 0$, and $\theta > 0$. Here, α denotes the shape parameter, θ is the scale parameter, and $\Gamma(\cdot)$ is the gamma function. The gamma distribution with $\alpha = 1$ corresponds to the chi-square distribution with two degrees of freedom. Moreover, it is well known that the mean of x for a gamma distribution is $\mathrm{E}[x] = \alpha\theta$, where $\mathrm{E}[\cdot]$ is the expectation operator. Furthermore, the kurtosis of a gamma distribution, $\mathrm{kurt}_{\mathrm{GM}}$, can be expressed as [35]

$$\mathrm{kurt}_{\mathrm{GM}} = \frac{(\alpha+2)(\alpha+3)}{\alpha(\alpha+1)}. \tag{107}$$

Moreover, let me consider the power-domain noise signal, x_{p}, in the frequency domain, which is defined as

$$x_{\mathrm{p}} = |x_{\mathrm{re}} + i \cdot x_{\mathrm{im}}|^2 = (x_{\mathrm{re}} + i \cdot x_{\mathrm{im}})(x_{\mathrm{re}} + i \cdot x_{\mathrm{im}})^* = x_{\mathrm{re}}^2 + x_{\mathrm{im}}^2, \tag{108}$$

where x_{re} is the real part of the complex-valued signal and x_{im} is the imaginary part of the complex-valued signal. They are independent and identically distributed (i.i.d.) with each other, and the superscript $*$ expresses complex conjugation. Thus, the power-domain signal is the sum of two squares of random variables with the same distribution.

Hereinafter, let x_{re} and x_{im} be the signals after DFT analysis of x_{j} ($j = 1, \ldots, J$), and we suppose that the statistical properties of x_{j} equal to x_{re} and x_{im}. Moreover, we assume the following; x_j is i.i.d. in each channel, the p.d.f. of x_j is symmetrical, and its mean is zero. These assumptions mean that the odd-order cumulants and moments are zero except for the first order.

Although $\mathrm{kurt}_x = 3$ if x is a Gaussian signal, note that the kurtosis of a Gaussian signal in the power spectral domain is 6. This is because a Gaussian signal in the time domain obeys the chi-square distribution with two degrees of freedom in the power spectral domain; for such a chi-square distribution, $\mu_4/\mu_2^2 = 6$.

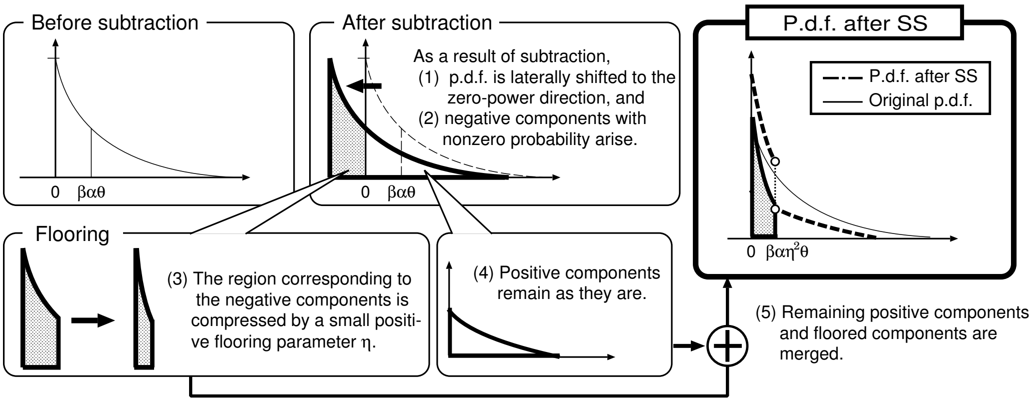

Before subtraction
0
βαθ
After subtraction
As a result of subtraction,
(1) p.d.f. is laterally shifted to the zero-power direction, and
(2) negative components with nonzero probability arise.
0
βαθ
P.d.f. after SS
P.d.f. after SS
Original p.d.f.
0
βαη²θ
Flooring
(3) The region corresponding to the negative components is compressed by a small positive flooring parameter η.
(4) Positive components remain as they are.
(5) Remaining positive components and floored components are merged.

Figure 38. Deformation of original p.d.f. of power-domain signal via SS.

8.3. Kurtosis analysis on BF+SS and chSS+BF

8.3.1 Resultant kurtosis after SS

In this section, I analyze the kurtosis after SS. In traditional SS, the long-term-averaged power spectrum of a noise signal is utilized as the estimated noise power spectrum. Then, the estimated noise spectrum multiplied by the oversubtraction parameter β is subtracted from the observed power spectrum. When a gamma distribution modeling 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. to the zero-power direction (see Fig. 38). 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 small positive value η (the so-called flooring technique). This means that the region corresponding to the probability of the negative components, which form a section cut from the original gamma distribution, is compressed by the effect of the flooring. Finally, the floored components are superimposed on the laterally shifted p.d.f. (see Fig. 38). Thus, the resultant p.d.f. after SS, $P_{\mathrm{SS}}(z)$, can be written as

$$P_{\mathrm{SS}}(z) = \begin{cases} \dfrac{1}{\theta^{\alpha}\Gamma(\alpha)}(z+\beta\alpha\theta)^{\alpha-1}\exp\left\{-\dfrac{z+\beta\alpha\theta}{\theta}\right\} & (z \geq \beta\alpha\eta^2\theta), \\ \\ \dfrac{1}{\theta^{\alpha}\Gamma(\alpha)}(z+\beta\alpha\theta)^{\alpha-1}\exp\left\{-\dfrac{z+\beta\alpha\theta}{\theta}\right\} + \dfrac{1}{(\eta^2\theta)^{\alpha}\Gamma(\alpha)}z^{\alpha-1}\exp\left\{-\dfrac{z}{\eta^2\theta}\right\} & \\ & (0 < z < \beta\alpha\eta^2\theta), \end{cases} \tag{109}$$

where z is the random variable of the p.d.f. after SS. The derivation of $P_{\mathrm{SS}}(z)$ is described in Appendix D.

From (109), the kurtosis after SS can be expressed as

$$\mathrm{kurt}_{\mathrm{SS}} = \Gamma(\alpha)\frac{\mathcal{F}(\alpha,\beta,\eta)}{\mathcal{G}^2(\alpha,\beta,\eta)}, \tag{110}$$

where

$$\mathcal{G}(\alpha,\beta,\eta) = \Gamma(\alpha)\Gamma(\beta\alpha,\alpha+2) - 2\beta\alpha\Gamma(\beta\alpha,\alpha+1) + \beta^2\alpha^2\Gamma(\beta\alpha,\alpha) + \eta^4\gamma(\beta\alpha,\alpha+2), \tag{111}$$

$$\begin{aligned}\mathcal{F}(\alpha,\beta,\eta) = {} & \Gamma(\beta\alpha,\alpha+4) - 4\beta\alpha\Gamma(\beta\alpha,\alpha+3) + 6\beta^2\alpha^2\Gamma(\beta\alpha,\alpha+2) \\ & - 4\beta^3\alpha^3\Gamma(\beta\alpha,\alpha+1) + \beta^4\alpha^4\Gamma(\beta\alpha,\alpha) + \eta^8\gamma(\beta\alpha,\alpha+4).\end{aligned} \tag{112}$$

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

$$\Gamma(b,a) = \int_b^{\infty} t^{a-1}\exp\{-t\}\mathrm{d}t, \tag{113}$$

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

$$\gamma(b,a) = \int_0^{b} t^{a-1}\exp\{-t\}\mathrm{d}t. \tag{114}$$

The detailed derivation of (110) is given in Appendix E. Although Uemura et al. have given an approximated form (lower bound) of the kurtosis after SS in Ref. [35], (110) involves no approximation throughout its derivation. Furthermore, (110) takes into account *the effect of the flooring technique* unlike Ref. [35].

Figure 39(a) depicts the theoretical output kurtosis ratio after SS, $\mathrm{kurt}_{\mathrm{SS}}$ / $\mathrm{kurt}_{\mathrm{GM}}$, for various values of oversubtraction parameter β and flooring parameter η. In the figure, the kurtosis of the input signal is fixed to 6.0, which corresponds to a Gaussian signal. From this figure, it is confirmed that the output kurtosis ratio is basically proportional to the oversubtraction parameter β. However, kurtosis does not monotonically increase when the flooring parameter is nonzero. For instance, the output kurtosis ratio is smaller than the peak value when $\beta = 4$ and $\eta = 0.4$. This phenomenon can be explained as follows. For a large oversubtraction parameter, almost all the spectral components become negative due to the larger lateral shift of the p.d.f. by SS. Since flooring is applied to avoid such negative components, almost all the components are reconstructed by flooring. Therefore, the statistical characteristics of the signal do not change except for its amplitude if $\eta \neq 0$. Generally, kurtosis does not depend on the change in amplitude; consequently, it can be considered that kurtosis does not markedly increase when a larger oversubtraction parameter and a larger flooring parameter are set.

The relation between the theoretical output kurtosis ratio and the kurtosis of the original input signal is shown in Fig. 39(b). In the figure, η is fixed to 0.0. It is revealed that the output kurtosis ratio after SS rapidly decreases as the input kurtosis increases, even with the same oversubtraction parameter β. Therefore, the output kurtosis ratio after SS, which is related to the amount of musical noise, strongly depends on the statistical characteristics of the input signal. That is to say, SS generates a larger amount of musical noise for a Gaussian input signal than for a super-Gaussian input signal. This fact has been reported in Ref. [35].

8.3.2 Resultant kurtosis after DS

In this section, I analyze of the kurtosis after DS, and I reveal that DS can reduce the kurtosis of input signals. Since I assume that the p.d.f. of x_{re} or x_{im} corresponds to the time-domain signal x_j, the effect of DS on the change in kurtosis can be derived from the cumulants and moments of x_j.

For cumulants, when X and Y are independent random variables it is well known that the following relation holds:

$$\mathrm{cum}_n(aX + bY) = a^n \,\mathrm{cum}_n(X) + b^n \,\mathrm{cum}_n(Y), \tag{115}$$

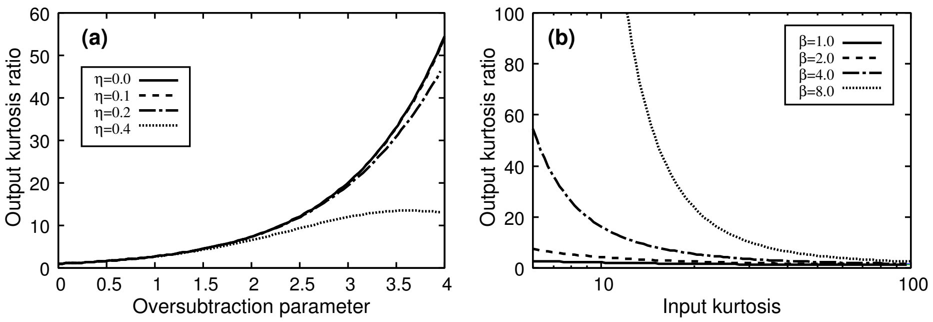


Figure 39. (a) Theoretical output kurtosis ratio after SS for various values of oversubtraction parameter β and flooring parameter η. In this figure, kurtosis of input signal is fixed to 6.0. (b) Theoretical output kurtosis ratio after SS for various values of input kurtosis. In this figure, flooring parameter η is fixed to 0.0.

where $\mathrm{cum}_n(\cdot)$ denotes the nth-order cumulant. The cumulants of the random variable X, $\mathrm{cum}_n(X)$, are defined by a cumulant-generating function, which is the logarithm of the moment-generating function. The cumulant-generating function $C(\zeta)$ is defined as

$$C(\zeta) = \log(\mathrm{E}[\exp\{\zeta X\}]) = \sum_{n=1}^{\infty} \mathrm{cum}_n(X)\frac{\zeta^n}{n!}, \tag{116}$$

where ζ is an auxiliary variable and $\mathrm{E}[\exp\{\zeta X\}]$ is the moment-generating function. Thus, the nth-order cumulant $\mathrm{cum}_n(X)$ is represented by

$$\mathrm{cum}_n(X) = C^{(n)}(0), \tag{117}$$

where $C^{(n)}(\zeta)$ is the nth-order derivative of $C(\zeta)$.

Now I consider the DS beamformer, which is steered to $\theta_{\mathrm{U}} = 0$ and whose array weights are $1/J$. Using (115), the resultant nth-order cumulant after DS, $\mathcal{K}_n$, can be expressed by

$$\mathcal{K}_n = \frac{1}{J^{n-1}} K_n, \tag{118}$$

where K_n is the nth order cumulant of x_j. Therefore, using (118) and the well-known mathematical relation between cumulants and moments, the power-spectral-

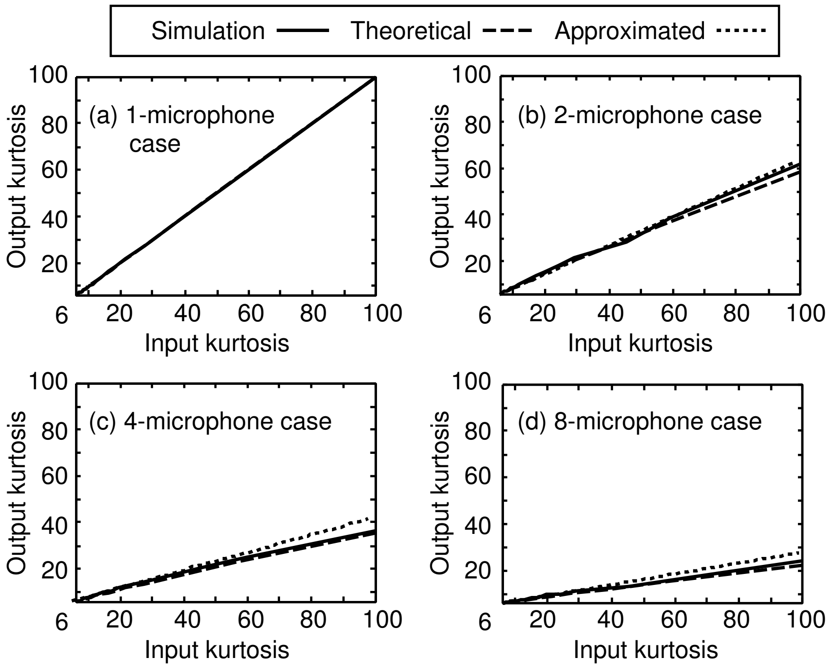


Figure 40. Relation between input kurtosis and output kurtosis after DS. Solid lines indicate simulation results, broken lines express theoretical plots obtained by (119), and dotted lines show approximate results obtained by (120).

domain kurtosis after DS, $\mathrm{kurt}_{\mathrm{DS}}$ can be expressed by

$$\mathrm{kurt}_{\mathrm{DS}} = \frac{\mathcal{K}_8 + 38\mathcal{K}_4^2 + 32\mathcal{K}_2\mathcal{K}_6 + 288\mathcal{K}_2^2\mathcal{K}_4 + 192\mathcal{K}_2^4}{2\mathcal{K}_4^2 + 16\mathcal{K}_2^2\mathcal{K}_4 + 32\mathcal{K}_2^4}. \tag{119}$$

The detailed derivation of (119) is described in Appendix F.

Regarding the power-spectral components obtained from a gamma distribution, the relation between input kurtosis and output kurtosis after DS is illustrated in Fig. 40. In the figure, solid lines indicate simulation results and broken lines show theoretical relations given by (119). The simulation results are derived as follows. First, multichannel signals with various values of kurtosis are generated artificially from a gamma distribution. Next, DS is applied to the generated signals. Finally, kurtosis after DS is estimated from the signal resulting from DS. From this figure, it is confirmed that the theoretical plots closely fit the simulation results. The relation between input/output kurtosis behaves as follows: (I) The output kurtosis is very close to a linear function of the input kurtosis, and (II)

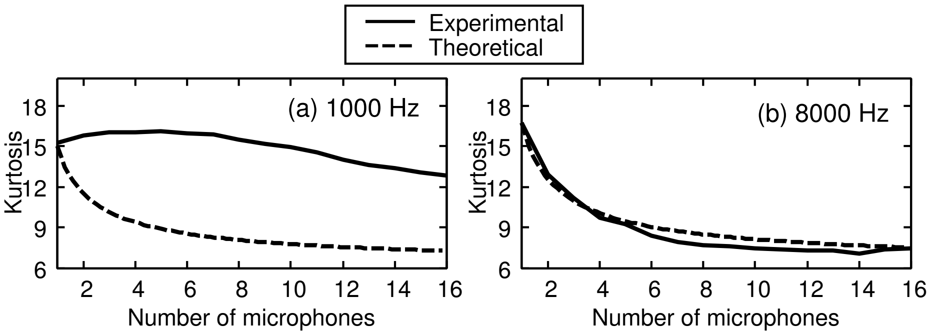


Figure 41. Simulation result for noise with interchannel correlation (solid line) and theoretical effect of DS assuming no interchannel correlation (broken line) in each frequency subband.

the output kurtosis is almost inversely proportional to the number of microphones. These behaviors result in the following simplified (but useful) approximation with an explicit function form:

$$\mathrm{kurt}_{\mathrm{DS}} \simeq J^{-0.7} \cdot (\mathrm{kurt}_{\mathrm{in}} - 6) + 6, \tag{120}$$

where $\mathrm{kurt}_{\mathrm{in}}$ is the input kurtosis. The approximated plots also match the simulation results in Fig. 40.

When input signals involve interchannel correlation, the relation between input kurtosis and output kurtosis after DS approaches that for only one microphone. If all input signals are identical signals, i.e., the signals are completely correlated, the output after DS also becomes the same as the input signal. In such a case, the effect of DS on the change in kurtosis corresponds to that for only one microphone. However, the interchannel correlation is not completely unit within all frequency subbands for a diffuse noise field that is a typically considered noise field. It is well known that the intensity of the interchannel correlation is strong in lower-frequency subbands and weak in higher-frequency subbands for a diffuse noise field [39]. Therefore, in lower-frequency subbands, it can be expected that DS does not significantly reduce the kurtosis of the signal.

As it is well known that the interchannel correlation for the diffuse noise field between two measurement locations can be expressed by the sinc function [39],

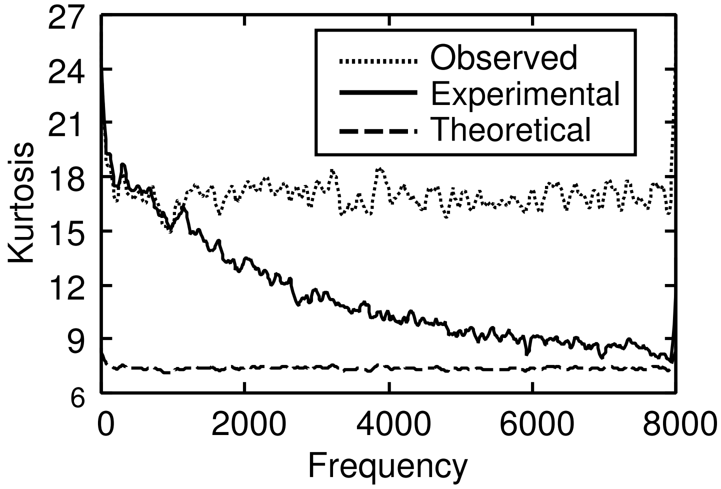


Figure 42. Simulation result for noise with interchannel correlation (solid line), and theoretical effect of DS assuming no interchannel correlation (broken line), and observed kurtosis (dotted line), in eight-microphone case.

we can state how array signal processing is affected by the interchannel correlation. However, we cannot know exactly how cumulants are changed by the interchannel correlation because (115) only holds when signals are mutually independent. Therefore, we cannot formulate how kurtosis is changed via DS for signals with interchannel correlation. For this reason, I experimentally investigate the effect of interchannel correlation in the following.

Figures 41 and 42 show preliminary simulation results of DS. In this simulation, SS is first applied to a multichannel Gaussian signal with interchannel correlation. Next, DS is applied to the signal after SS. In the preliminary simulation, the interelement distance between microphones is 2.15 cm each. From the results shown in Fig. 41(a) 42, we can confirm that the effect of DS on kurtosis is weak in lower-frequency subbands, although it should be noted that the effect does not completely disappear in lower-frequency subbands. Also, the theoretical kurtosis curve is in good agreement with the actual results in higher-frequency subbands (see Figs. 41(b) 42). This is because interchannel correlation is weak in higher-frequency subbands. Consequently, for the diffuse noise field, DS can reduce the kurtosis of the input signal even if interchannel correlation exists.

If input noise signals contain no interchannel correlation, the distance between microphones does not affect the results. That is to say, the kurtosis change via DS can be well fit to (120). Otherwise, in lower-frequency subbands, it is expected that the mitigation effect of kurtosis by DS degrades with decreasing of the microphone distance. This is because the interchannel correlation in lower-frequency subbands increases with decreasing distance between microphones. In higher-frequency subbands, the effect of distance between microphones is thought to be small.

8.3.3 Resultant kurtosis: BF+SS vs. chSS+BF

In the previous subsections, I discussed the resultant kurtosis after SS and DS. In this subsection, I analyze the resultant kurtosis for two types of composite systems, i.e., BF+SS and chSS+BF, and compare their effect on musical-noise generation. As described in Sect. 7.3, it can be expected that a smaller increase in kurtosis leads to a smaller amount of musical noise generated.

In BF+SS, DS is first applied to a multichannel input signal. At this point, the resultant kurtosis in the power spectral domain, $\mathrm{kurt}_{\mathrm{DS}}$, can be represented by (120). Using (107), we can derive a shape parameter for the gamma distribution corresponding to $\mathrm{kurt}_{\mathrm{DS}}$, $\hat{\alpha}$, as

$$\hat{\alpha} = \frac{\sqrt{\mathrm{kurt}_{\mathrm{DS}}^2 + 14\,\mathrm{kurt}_{\mathrm{DS}} + 1} - \mathrm{kurt}_{\mathrm{DS}} + 5}{2\,\mathrm{kurt}_{\mathrm{DS}} - 2}. \tag{121}$$

The derivation of (121) is shown in Appendix G. Consequently, using (110) and (121), the resultant kurtosis after BF+SS, $\mathrm{kurt}_{\mathrm{BF+SS}}$, can be written as

$$\mathrm{kurt}_{\mathrm{BF+SS}} = \Gamma(\hat{\alpha}) \frac{\mathcal{F}(\hat{\alpha}, \beta, \eta)}{\mathcal{G}^2(\hat{\alpha}, \beta, \eta)}. \tag{122}$$

In chSS+BF, SS is first applied to each input channel. Thus, the output kurtosis after channelwise SS, $\mathrm{kurt}_{\mathrm{chSS}}$, can be given by

$$\mathrm{kurt}_{\mathrm{chSS}} = \Gamma(\alpha) \frac{\mathcal{F}(\alpha, \beta, \eta)}{\mathcal{G}^2(\alpha, \beta, \eta)}. \tag{123}$$

Finally, DS is performed and the resultant kurtosis after chSS+BF, $\mathrm{kurt}_{\mathrm{chSS+BF}}$, can be written as

$$\mathrm{kurt}_{\mathrm{chSS+BF}} = J^{-0.7}\left[\Gamma(\alpha)\frac{\mathcal{F}(\alpha,\beta,\eta)}{\mathcal{G}^{2}(\alpha,\beta,\eta)} - 6\right] + 6, \tag{124}$$

where I use (120).

I should compare $\mathrm{kurt}_{\mathrm{BF+SS}}$ and $\mathrm{kurt}_{\mathrm{chSS+BF}}$ here. However, one problem still remains: comparison under equivalent noise reduction performance; the noise reduction performances of BF+SS and chSS+BF are not equivalent as described in the next section. Moreover, the design of a flooring parameter so that the noise reduction performances of both methods become equivalent will be discussed in the next section. Therefore, $\mathrm{kurt}_{\mathrm{BF+SS}}$ and $\mathrm{kurt}_{\mathrm{chSS+BF}}$ will be compared in Sect. 8.5 under equivalent noise reduction performance conditions.

8.4. Noise reduction performance analysis

In the previous section, I did not discuss the noise reduction performances of BF+SS and chSS+BF. In this section, a mathematical analysis of the noise reduction performances of BF+SS and chSS+BF is given. As a result of this analysis, it is revealed that the noise reduction performances of BF+SS and chSS+BF are not equivalent even if the same parameters are set in the SS part. I then derive a flooring-parameter design strategy for aligning the noise reduction performances of BF+SS and chSS+BF.

8.4.1 Noise reduction performance of SS

I utilize the following index to measure the noise reduction performance (NRP),

$$\mathrm{NRP} = 10\log_{10}\frac{E[n_{\mathrm{in}}]}{E[n_{\mathrm{out}}]}, \tag{125}$$

where n_{in} is the power-domain (noise) signal of the input and n_{out} is the power-domain (noise) signal of the output after processing.

First, I derive the average power of the input signal. I assume that the input signal in the power domain can be modeled by a gamma distribution. Then, the average power of the input signal can be given as

$$\begin{aligned}
E[n_{\mathrm{in}}] &= E[x] \\
&= \int_0^\infty x P_{\mathrm{GM}}(x)\mathrm{d}x \\
&= \int_0^\infty x \cdot \frac{1}{\theta^\alpha \Gamma(\alpha)} x^{\alpha-1} \exp\left\{-\frac{x}{\theta}\right\} \mathrm{d}x \\
&= \frac{1}{\theta^\alpha \Gamma(\alpha)} \int_0^\infty x^\alpha \exp\left\{-\frac{x}{\theta}\right\} \mathrm{d}x. \qquad (126)
\end{aligned}$$

Here, let $t = x/\theta$, then $\theta\mathrm{d}t = \mathrm{d}x$. Thus,

$$\begin{aligned}
E[n_{\mathrm{in}}] &= \frac{1}{\theta^\alpha \Gamma(\alpha)} \int_0^\infty (\theta t)^\alpha \exp\{-t\}\, \theta \mathrm{d}t \\
&= \frac{\theta^{\alpha+1}}{\theta^\alpha \Gamma(\alpha)} \int_0^\infty t^\alpha \exp\{-t\}\, \mathrm{d}t \\
&= \frac{\theta \Gamma(\alpha+1)}{\Gamma(\alpha)} \\
&= \theta\alpha. \qquad (127)
\end{aligned}$$

This corresponds to the mean of a random variable with a gamma distribution.

Next, the average power of the signal after SS is calculated. Here, let z have the p.d.f. of the signal after SS, $P_{\mathrm{SS}}(z)$, defined by (109), then the average power of the signal after SS can be expressed as

$$\begin{aligned}
E[n_{\mathrm{out}}] &= E[z] \\
&= \int_0^\infty z P_{\mathrm{SS}}(z)\mathrm{d}z \\
&= \int_0^\infty \frac{z}{\theta^\alpha \Gamma(\alpha)} (z + \beta\alpha\theta)^{\alpha-1} \exp\left\{-\frac{z + \beta\alpha\theta}{\theta}\right\} \mathrm{d}z \\
&\qquad + \int_0^{\beta\alpha\eta^2\theta} \frac{z}{(\eta^2\theta)^\alpha \Gamma(\alpha)} z^{\alpha-1} \exp\left\{-\frac{z}{\eta^2\theta}\right\} \mathrm{d}z. \qquad (128)
\end{aligned}$$

I now consider the first term of the right-hand side in (128). Here let $t = z + \beta\alpha\theta$, then d$t$ = dz. As a result,

$$\begin{aligned}
&\int_0^\infty \frac{z}{\theta^\alpha \Gamma(\alpha)} (z+\beta\alpha\theta)^{\alpha-1} \exp\left\{-\frac{z+\beta\alpha\theta}{\theta}\right\} \mathrm{d}z \\
&\quad = \int_{\beta\alpha\theta}^\infty (t-\beta\alpha\theta) \cdot \frac{1}{\theta^\alpha\Gamma(\alpha)} \cdot t^{\alpha-1} \, \exp\left\{-\frac{t}{\theta}\right\} \mathrm{d}t \\
&\quad = \int_{\beta\alpha\theta}^\infty \frac{1}{\theta^\alpha\Gamma(\alpha)} \cdot t^{\alpha} \, \exp\left\{-\frac{t}{\theta}\right\} \mathrm{d}t - \int_{\beta\alpha\theta}^\infty \frac{\beta\alpha\theta}{\theta^\alpha\Gamma(\alpha)} \cdot t^{\alpha-1} \, \exp\left\{-\frac{t}{\theta}\right\} \mathrm{d}t \\
&\quad = \frac{\theta \cdot \Gamma(\beta\alpha, \alpha+1)}{\Gamma(\alpha)} - \beta\alpha\theta \cdot \frac{\Gamma(\beta\alpha,\alpha)}{\Gamma(\alpha)}. \qquad (129)
\end{aligned}$$

Also, I deal with the second term of the right-hand side in (128). Let $t = z/(\eta^2\theta)$ then $\eta^2\theta$dt = dz, resulting in

$$\begin{aligned}
&\int_0^{\beta\alpha\eta^2\theta} \frac{z}{(\eta^2\theta)^\alpha\Gamma(\alpha)} z^{\alpha-1} \exp\left\{-\frac{z}{\eta^2\theta}\right\} \mathrm{d}z \\
&\quad = \frac{1}{(\eta^2\theta)^\alpha\Gamma(\alpha)} \int_0^{\beta\alpha} (\eta^2\theta t)^\alpha \cdot \exp\{-t\}\, \eta^2\theta \mathrm{d}t = \frac{\eta^2\theta}{\Gamma(\alpha)} \gamma(\beta\alpha, \alpha+1). \qquad (130)
\end{aligned}$$

Using (127), (129), and (130), the noise reduction performance of SS, NRP$_{\mathrm{SS}}$, can be expressed by

$$\begin{aligned}
\mathrm{NRP}_{\mathrm{SS}} &= 10 \log 10 \left(\frac{E[z]}{E[x]}\right) \\
&= -10 \log_{10} \left[\frac{\Gamma(\beta\alpha, \alpha+1)}{\Gamma(\alpha+1)} - \beta \cdot \frac{\Gamma(\beta\alpha,\alpha)}{\Gamma(\alpha)} + \eta^2 \frac{\gamma(\beta\alpha,\alpha+1)}{\Gamma(\alpha+1)}\right]. \qquad (131)
\end{aligned}$$

Figure 43(a) shows the theoretical value of NRP$_{\mathrm{SS}}$ for various values of oversubtraction parameter β and flooring parameter η, where the kurtosis of the input signal is fixed to 6.0, corresponding to a Gaussian signal. From this figure, it is confirmed that NRP$_{\mathrm{SS}}$ is proportional to β. However, NRP$_{\mathrm{SS}}$ hits a peak when η is nonzero even for large values of β. The relation between the theoretical NRP$_{\mathrm{SS}}$ and the kurtosis of the input signal is illustrated in Fig. 43(b). In this figure, η is fixed to 0.0. It is revealed NRP$_{\mathrm{SS}}$ decreases as the input kurtosis increases. This is because the mean of a high-kurtosis signal tends to be small. Since the shape parameter α of a high-kurtosis signal is small, the mean $\alpha\theta$ corresponding to the amount of subtraction, also becomes small. As a result, NRP$_{\mathrm{SS}}$ decreases as the input kurtosis increases. That is to say, NRP$_{\mathrm{SS}}$ strongly depends on the statistical characteristics of the input signal as well as the values of the oversubtraction and flooring parameters.

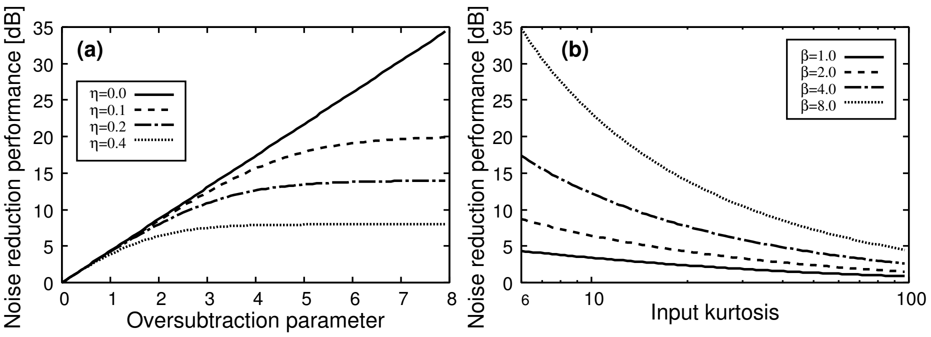


Figure 43. (a) Theoretical noise reduction performance of SS with various oversubtraction parameters β and flooring parameters η. In this figure, kurtosis of input signal is fixed to 6.0. (b) Theoretical noise reduction performance of SS with various values of input kurtosis. In this figure, flooring parameter η is fixed to 0.0.

8.4.2 Noise reduction performance of DS

It is well known that the noise reduction performance of DS ($\mathrm{NRP_{DS}}$) is proportional to the number of microphones. In particular, for spatially uncorrelated multichannel signals, $\mathrm{NRP_{DS}}$ is given as [39]

$$\mathrm{NRP_{DS}} = 10 \log_{10} J. \tag{132}$$

8.4.3 Resultant noise reduction performance: BF+SS vs. chSS+BF

In the previous subsections, the noise reduction performances of SS and DS were discussed. In this subsection, I derive the resultant noise reduction performances of the composite systems of SS and DS, i.e., BF+SS and chSS+BF.

The noise reduction performance of BF+SS is analyzed as follows. In BF+SS, DS is first applied to a multichannel input signal. If this input signal is spatially uncorrelated, its noise reduction performance can be represented by $10 \log_{10} J$. After DS, SS is applied to the signal after DS. Note that DS affects the kurtosis of the input signal. As described in Sect. 8.3.2, the resultant kurtosis after DS can be approximated as $J^{-0.7} \cdot (\mathrm{kurt_{in}} - 6) + 6$. Thus, SS is applied to the kurtosis-modified signal. Consequently, using (121), (131), and (132), the noise reduction performance of BF+SS, $\mathrm{NRP}_{\mathrm{BF+SS}}$, is given as

$$\begin{aligned}\mathrm{NRP}_{\mathrm{BF+SS}} &= 10\log_{10} J - 10\log_{10}\left[\frac{\Gamma(\beta\hat{\alpha},\hat{\alpha}+1)}{\Gamma(\hat{\alpha}+1)} - \beta\cdot\frac{\Gamma(\beta\hat{\alpha},\hat{\alpha})}{\Gamma(\hat{\alpha})} + \eta^2\frac{\gamma(\beta\hat{\alpha},\hat{\alpha}+1)}{\Gamma(\hat{\alpha}+1)}\right] \\ &= -10\log_{10}\frac{1}{J\cdot\Gamma(\hat{\alpha})}\left[\frac{\Gamma(\beta\hat{\alpha},\hat{\alpha}+1)}{\hat{\alpha}} - \beta\cdot\Gamma(\beta\hat{\alpha},\hat{\alpha}) + \eta^2\frac{\gamma(\beta\hat{\alpha},\hat{\alpha}+1)}{\hat{\alpha}}\right],\end{aligned} \tag{133}$$

where $\hat{\alpha}$ is defined by (121).

In chSS+BF, SS is first applied to a multichannel input signal, then DS is applied to the resulting signal. Thus, using (131) and (132), the noise reduction performance of chSS+BF, $\mathrm{NRP}_{\mathrm{chSS+BF}}$, can be represented by

$$\mathrm{NRP}_{\mathrm{chSS+BF}} = -10\log_{10}\frac{1}{J\cdot\Gamma(\alpha)}\left[\frac{\Gamma(\beta\alpha,\alpha+1)}{\alpha} - \beta\cdot\Gamma(\beta\alpha,\alpha) + \eta^2\frac{\gamma(\beta\alpha,\alpha+1)}{\alpha}\right]. \tag{134}$$

Figure 44 depicts the theoretical values of $\mathrm{NRP}_{\mathrm{BF+SS}}$ and $\mathrm{NRP}_{\mathrm{chSS+BF}}$. From this result, we can see that the noise reduction performances of both methods are equivalent when the input signal is Gaussian. However, if the input signal is super-Gaussian, $\mathrm{NRP}_{\mathrm{BF+SS}}$ exceeds $\mathrm{NRP}_{\mathrm{chSS+BF}}$. This is due to the fact that DS is first applied to the input signal in BF+SS; thus, DS reduces the kurtosis of the signal. Since $\mathrm{NRP}_{\mathrm{SS}}$ for a low-kurtosis signal is greater than that for a high-kurtosis signal (see Fig. 43(b)), the noise reduction performance of BF+SS is superior to that of chSS+BF.

This discussion implies that the $\mathrm{NRP}_{\mathrm{BF+SS}}$ and $\mathrm{NRP}_{\mathrm{chSS+BF}}$ are not equivalent under some conditions. Thus the kurtosis-based analysis described in Sect. 8.3 is biased and requires some adjustment. In the following subsection, I will discuss how to align the noise reduction performances of BF+SS and chSS+BF.

8.4.4 Flooring-parameter design in BF+SS for equivalent noise reduction performance

In this section, we describe the flooring-parameter design in BF+SS so that $\mathrm{NRP}_{\mathrm{BF+SS}}$ and $\mathrm{NRP}_{\mathrm{chSS+BF}}$ become equivalent.

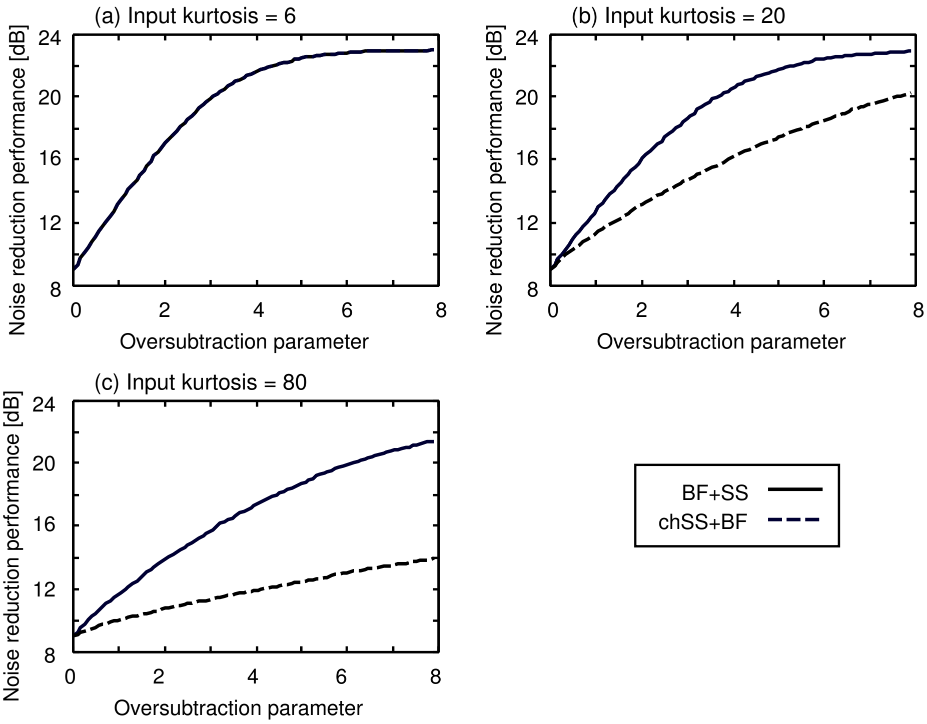


Figure 44. Comparison of noise reduction performances of chSS+BF with BF+SS. In this figure, flooring parameter is fixed to 0.2 and number of microphones is 8.

Using (133) and (134), the flooring parameter $\hat{\eta}$ that makes $\mathrm{NRP}_{\mathrm{BF+SS}}$ equal to $\mathrm{NRP}_{\mathrm{chSS+BF}}$, is

$$\hat{\eta} = \sqrt{\frac{\hat{\alpha}}{\gamma(\beta\hat{\alpha}, \hat{\alpha}+1)} \cdot \left[\frac{\Gamma(\hat{\alpha})}{\Gamma(\alpha)}\mathcal{H}(\alpha,\beta,\eta) - \mathcal{I}(\hat{\alpha},\beta)\right]}, \tag{135}$$

where

$$\mathcal{H}(\alpha,\beta,\eta) = \frac{\Gamma(\beta\alpha, \alpha+1)}{\alpha} - \beta\cdot\Gamma(\beta\alpha,\alpha) + \eta^2\frac{\gamma(\beta\alpha,\alpha+1)}{\alpha}, \tag{136}$$

$$\mathcal{I}(\hat{\alpha},\beta) = \frac{\Gamma(\beta\hat{\alpha}, \hat{\alpha}+1)}{\hat{\alpha}} - \beta\cdot\Gamma(\beta\hat{\alpha},\hat{\alpha}). \tag{137}$$

The detailed derivation of (135) is given in Appendix H. By replacing η in (100) with this new flooring parameter $\hat{\eta}$, we can align $\mathrm{NRP}_{\mathrm{BF+SS}}$ and $\mathrm{NRP}_{\mathrm{chSS+BF}}$ to ensure a fair comparison.

8.5. Output kurtosis comparison under equivalent NRP condition

In this section, using the new flooring parameter for BF+SS, $\hat{\eta}$, I compare the output kurtosis of BF+SS and chSS+BF.

Setting $\hat{\eta}$ to (122), the output kurtosis of BF+SS is modified to

$$\mathrm{kurt}_{\mathrm{BF+SS}} = \Gamma(\hat{\alpha})\frac{\mathcal{F}(\hat{\alpha},\beta,\hat{\eta})}{\mathcal{G}^2(\hat{\alpha},\beta,\hat{\eta})}. \tag{138}$$

Here, I adopt the following index to compare the resultant kurtosis after BF+SS and chSS+BF:

$$R = \ln\frac{\mathrm{kurt}_{\mathrm{BF+SS}}}{\mathrm{kurt}_{\mathrm{chSS+BF}}}, \tag{139}$$

where R expresses the resultant kurtosis ratio between BF+SS and chSS+BF. Note that a positive R indicates that chSS+BF reduces the kurtosis more than BF+SS, implying that less musical noise is generated in chSS+BF. The behavior of R is depicted in Figs. 45 and 46. Figure 45 illustrates theoretical values of R for various values of input kurtosis. In this figure, β is fixed to 2.0 and the flooring parameter in chSS+BF is set to $\eta = 0.0$, 0.1, 0.2, and 0.4. The flooring parameter for BF+SS is automatically determined by (135). From this figure, we can confirm that chSS+BF reduces the kurtosis more than BF+SS for almost all input signals with various values of input kurtosis. Theoretical values of R for various oversubtraction parameters are depicted in Fig. 46. Figure 46(a) shows that the output kurtosis after chSS+BF is always less than that after BF+SS for a Gaussian signal, even if η is nonzero. On the other hand, Fig. 46(b) implies that the output kurtosis after BF+SS becomes less than that after chSS+BF for some parameter settings. However, such phenomena only occur for a large oversubtraction parameter, e.g., $\beta \geq 7$, which is not often applied in practical use. Therefore, it can be considered that chSS+BF reduces the kurtosis and musical noise more than BF+SS in almost all cases.

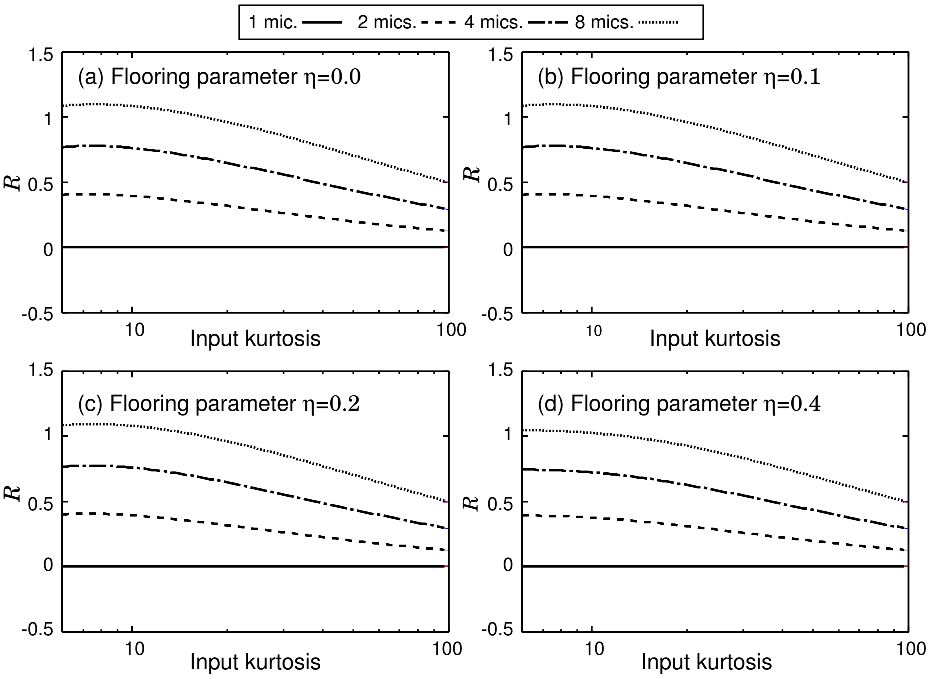


Figure 45. Theoretical kurtosis ratio between BF+SS and chSS+BF for various values of input kurtosis. In this figure, oversubtraction parameter is $\beta = 2.0$ and flooring parameter in chSS+BF is (a) $\eta = 0.0$, (b) $\eta = 0.1$, (c) $\eta = 0.2$, and (d) $\eta = 0.4$.

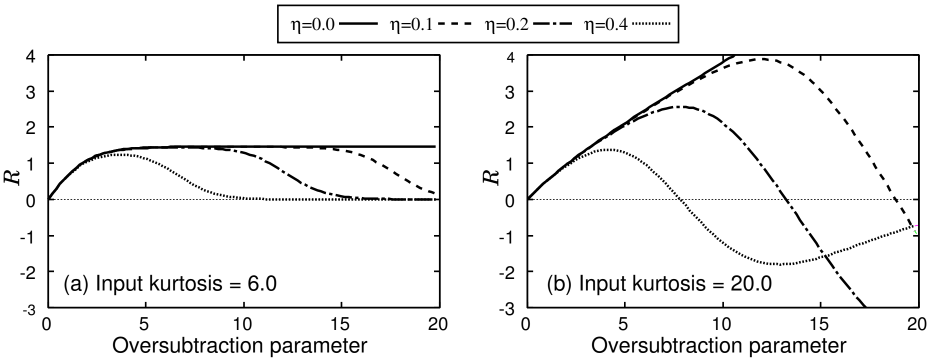


Figure 46. Theoretical kurtosis ratio between BF+SS and chSS+BF for various oversubtraction parameters. In this figure, number of microphones is fixed to 8, and input kurtosis is (a) 6.0 (Gaussian) and (b) 20.0 (super-Gaussian).

8.6. Evaluation

8.6.1 Computer simulations

First, I compare BF+SS and chSS+BF in terms of kurtosis ratio and noise reduction performance. I use 16-kHz-sampled signals as test data, in which the target speech is the original speech convoluted with impulse responses recorded in a room with 200 ms reverberation (see Fig. 47), and to which an artificially generated spatially uncorrelated white Gaussian or super-Gaussian signal is added. I use six speakers (six sentences) as sources of the original clean speech. The number of microphone elements in the simulation is varied from 2 to 16, and their interelement distance is 2.15 cm each. The oversubtraction parameter β is set to 2.0 and the flooring parameter for BF+SS, η, is set to 0.0, 0.2, 0.4, or 0.8. Note that the flooring parameter in chSS+BF is set to 0.0. In the simulation, I assume that the long-term-averaged power spectrum of noise is estimated perfectly in advance.

Here, I utilize the kurtosis ratio defined in Sect. 7.3.4 to measure the kurtosis difference, which is related to the amount of musical noise generated. The kurtosis ratio is given by

$$\text{Kurtosis ratio} = \frac{\text{kurt}(n_{\text{proc}}(f,\tau))}{\text{kurt}(n_{\text{org}}(f,\tau))}, \tag{140}$$

where $n_{\text{proc}}(f,\tau)$ is the power spectra of the residual noise signal after processing, and $n_{\text{org}}(f,\tau)$ is the power spectra of the original noise signal before processing. This kurtosis ratio indicates the extent to which kurtosis is increased with processing. Thus, a smaller kurtosis ratio is desirable. Moreover, the noise reduction performance is measured using (125).

Figures 48–50 show the simulation results for a Gaussian input signal. From Fig. 48(a), we can see that the kurtosis ratio of chSS+BF is decreases almost monotonically with increasing number of microphones. On the other hand, the kurtosis ratio of BF+SS does not exhibit such a tendency regardless of the flooring parameter. Also, the kurtosis ratio of chSS+BF is lower than that of BF+SS for all cases except for $\eta = 0.8$. Moreover, we can confirm from Fig. 48(b) that the values of noise reduction performance for BF+SS with flooring parameter $\eta = 0.0$ and chSS+BF are almost the same. When the flooring parameter for BF+SS is nonzero, the kurtosis ratio of BF+SS becomes smaller but the noise reduction performance degrades. On the other hand, for Gaussian signals, chSS+BF can reduce the kurtosis ratio, i.e., reduce the amount of musical noise generated, without degrading the noise reduction performance. Indeed BF+SS with $\eta = 0.8$ reduces the kurtosis ratio more than chSS+BF, but the noise reduction performance of BF+SS is extremely degraded. Furthermore, we can confirm from Figs. 49 and 50 that the theoretical kurtosis ratio and noise reduction performance closely fit the experimental results. These findings also support the validity of the analysis in Sects. 8.3, 8.4, and 8.5.

Figures 51–53 illustrate the simulation results for a super-Gaussian input signal. It can be confirmed from Fig. 51(a) that the kurtosis ratio of chSS+BF also decreases monotonically with increasing the number of microphones. Unlike the case of the Gaussian input signal, the kurtosis ratio of BF+SS with $\eta = 0.8$ also decreases with increasing number of microphones. However, for a lower value of the flooring parameter, the kurtosis ratio of BF+SS is not degraded. Moreover, the kurtosis ratio of chSS+BF is lower than that of BF+SS for almost all cases. For the super-Gaussian input signal, in contrast to the case of the Gaussian input signal, the noise reduction performance of BF+SS with $\eta = 0.0$ is greater than that of chSS+BF (see Fig. 51(b)). That is to say, the noise reduction performance of BF+SS is superior to that of chSS+BF for the same flooring parameter. This result is consistent with the analysis in Sect. 8.4. The noise reduction performance of BF+SS with $\eta = 0.4$ is comparable to that of chSS+BF. However, the kurtosis ratio of chSS+BF is still lower than that of BF+SS with $\eta = 0.4$. This result also coincides with the analysis in Sect. 8.5. On the other hand, the kurtosis ratio of BF+SS with $\eta = 0.8$ is almost the same as that of chSS+BF. However, the noise reduction performance of BF+SS with $\eta = 0.8$ is lower than that of chSS+BF. Thus, it is confirmed that chSS+BF reduces the kurtosis ratio more than BF+SS for a super-Gaussian signal under the same noise reduction performance. Furthermore, the theoretical kurtosis ratio and noise reduction performance closely fit the experimental results in Figs. 52 and 53.

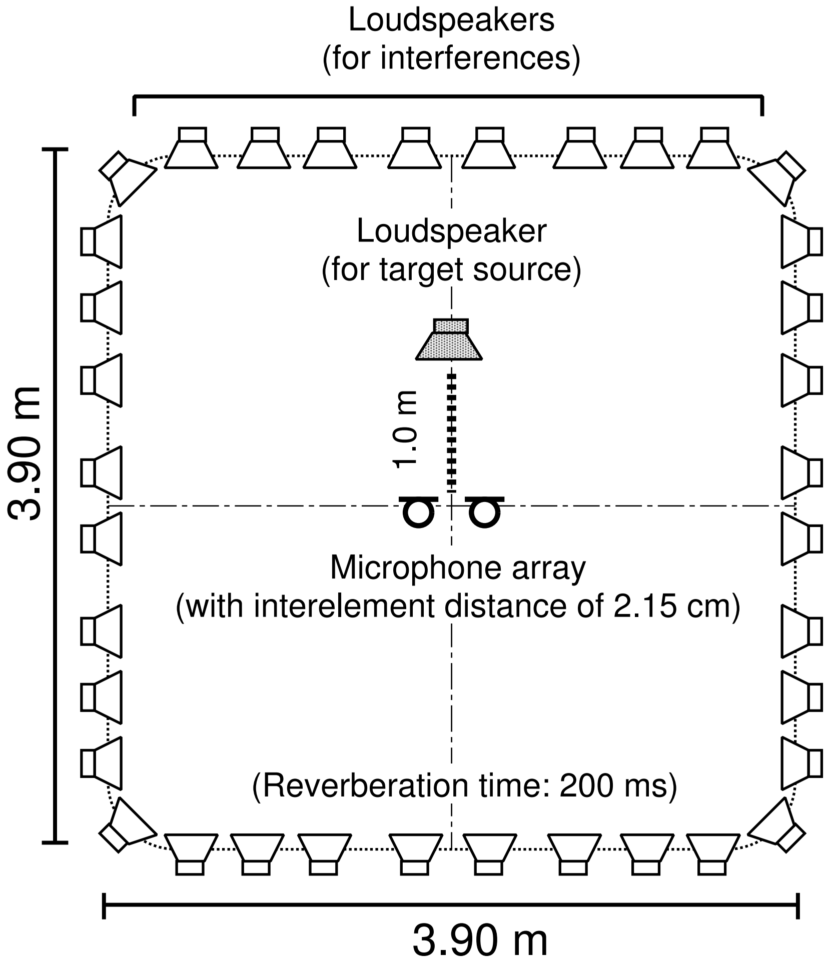


Figure 47. Reverberant room used in my simulations.

I also compare speech distortion originating from BF+SS and chSS+BF on the basis of cepstral distortion (CD) [43] for the four-microphone case. The comparison is made under the condition that the noise reduction performances of both methods are almost the same. For the Gaussian input signal, the same parameters $\beta = 2.0$ and $\eta = 0.0$ are utilized for BF+SS and chSS+BF. On the other hand, $\beta = 2.0$ and $\eta = 0.0$ are utilized for chSS+BF and $\beta = 2.0$ and $\eta = 0.4$ are utilized for BF+SS for the super-Gaussian input signal. Table 4 shows the result of the comparison, from which we can see that the amount of speech distortion originating from BF+SS and chSS+BF is almost the same for the Gaussian in-

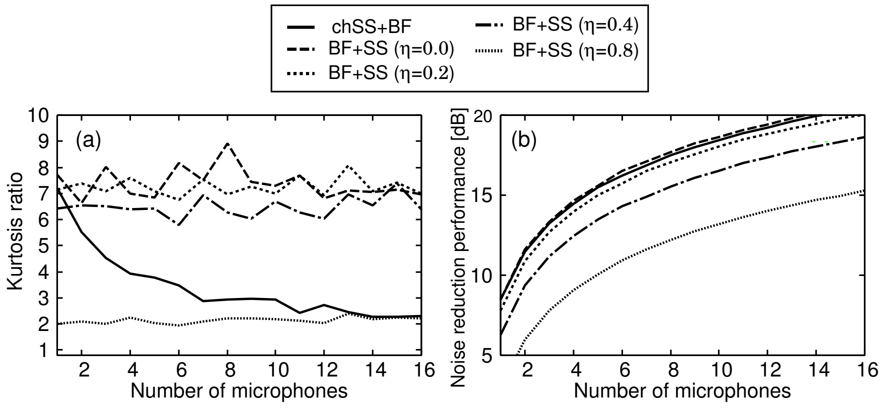


Figure 48. Results for Gaussian input signal. (a) Kurtosis ratio and (b) noise reduction performance for BF+SS with various flooring parameters.

Table 4. Speech distortion comparison of BF+SS and chSS+BF on the basis of CD for four-microphone case

Input noise type	chSS+BF	BF+SS
Gaussian	6.15 dB	6.45 dB
Super-Gaussian	6.17 dB	5.12 dB

put signal. For the super-Gaussian input signal, the speech distortion originating from BF+SS is less than that from chSS+BF. This is owing to the difference in the flooring parameter for each method.

In conclusion, all of these results are strong evidence for the validity of the analysis in Sects. 8.3, 8.4, and 8.5. These results suggest the following:

- Although BF+SS can reduce the amount of musical noise by employing a larger flooring parameter, it leads to a deterioration of the noise reduction performance.

- In contrast, chSS+BF can reduce the kurtosis ratio, which corresponds to the amount of musical noise generated, without degradation of the noise reduction performance.

- Under the same level of noise reduction performance, the amount of musical

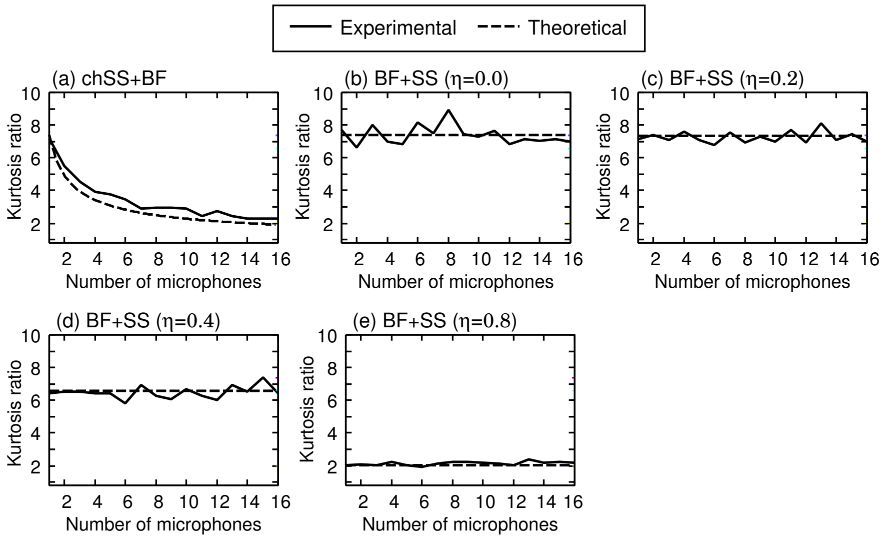


Figure 49. Comparison between experimental and theoretical kurtosis ratios for Gaussian input signal.

noise generated via chSS+BF is less than that generated via BF+SS.

- Thus, the chSS+BF structure is preferable from the viewpoint of musical-noise generation.
- However, the noise reduction performance of BF+SS is superior to that of chSS+BF for a super-Gaussian signal when the same parameters are set in the SS part for both methods.
- These results imply a trade-off between the amount of musical noise generated and the noise reduction performance. Thus, we should use an appropriate structure depending on the application.

These results should be applicable under different SNR conditions because our analysis is independent of the noise level. In the case of more reverberation, the observed signal tends to become Gaussian because many reverberant components are mixed. Therefore, the behavior of both methods under more reverberant conditions should be similar to that in the case of a Gaussian signal.

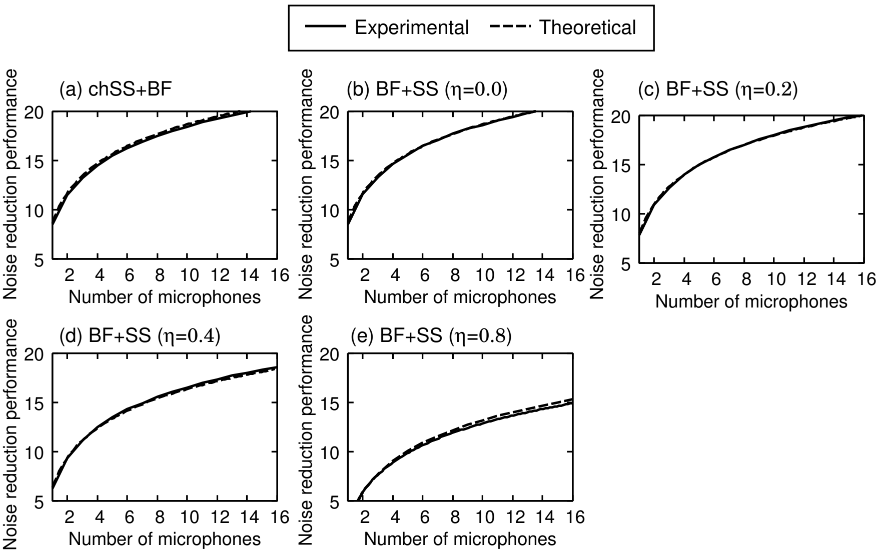


Figure 50. Comparison between experimental and theoretical noise reduction performances for Gaussian input signal.

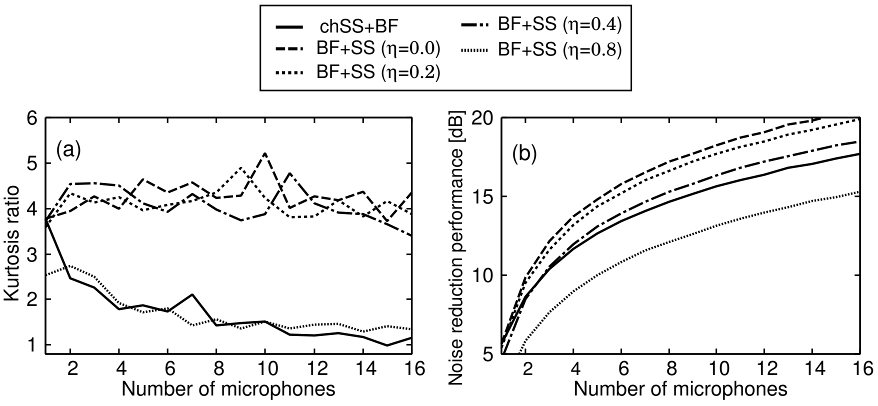


Figure 51. Results for super-Gaussian input signal. (a) Kurtosis ratio and (b) noise reduction performance for BF+SS with various flooring parameters.

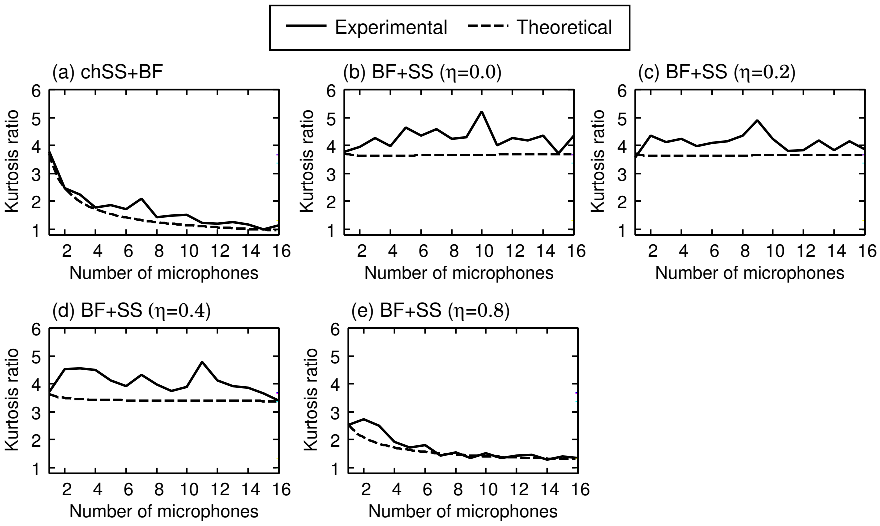


Figure 52. Comparison between experimental and theoretical kurtosis ratios for super-Gaussian input signal.

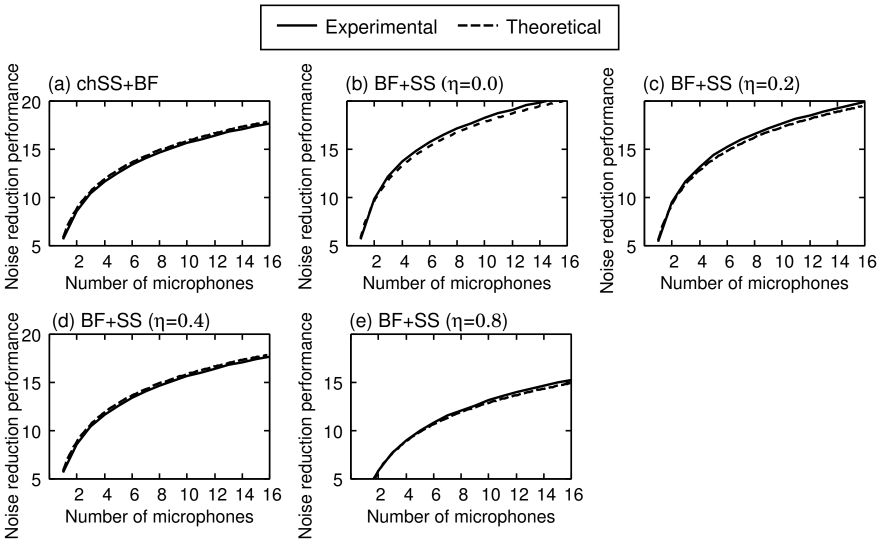


Figure 53. Comparison between experimental and theoretical noise reduction performances for super-Gaussian input signal.

8.6.2 Subjective evaluation

Next, I conduct a subjective evaluation to confirm that chSS+BF can mitigate musical noise. In the evaluation, I presented two signals processed by BF+SS and by chSS+BF to seven male examinees in random order, who were asked to select which signal they considered to contain less musical noise (the so-called AB method). Moreover, I instructed examinees to evaluate only the musical noise and not to consider the amplitude of the remaining noise. Here, the flooring parameter in BF+SS was automatically determined so that the output SNR of BF+SS and chSS+BF was equivalent. I used the preference score as the index of evaluation, which is the frequency of the selected signal.

In the experiment, three types of noise, (a) artificial spatially uncorrelated white Gaussian noise, (b) recorded railway-station noise emitted from 36 loudspeakers, and (c) recorded human speech emitted from 36 loudspeakers, were used. Note that noises (b) and (c) were recorded in the room shown in Fig. 47, and therefore include interchannel correlation because they were recordings of actual noise signals.

Each test sample is 16-kHz-sampled signal, and the target speech is the original speech convoluted with impulse responses recorded in a room with 200 ms reverberation (see Fig. 47) and to which the above-mentioned recorded noise signal is added. Ten pairs of signals per type of noise, i.e., a total of 30 pairs of processed signals, were presented to each examinee.

Figure 54 shows the subjective evaluation results, which confirm that the output of chSS+BF is preferred to that of BF+SS, even for actual acoustic noises including non-Gaussianity and interchannel correlation properties.

8.6.3 Subjective evaluation in BSSA architecture

Finally, I compare the amount of musical noise generated via two methods, i.e., the original BSSA and BSSA with channel-wise SS (chBSSA), on the basis of the informal listening test.

Figure 55 depicts the block diagram of chBSSA. In chBSSA, channel-wise spectral subtraction is performed before DS unlike the original BSSA.

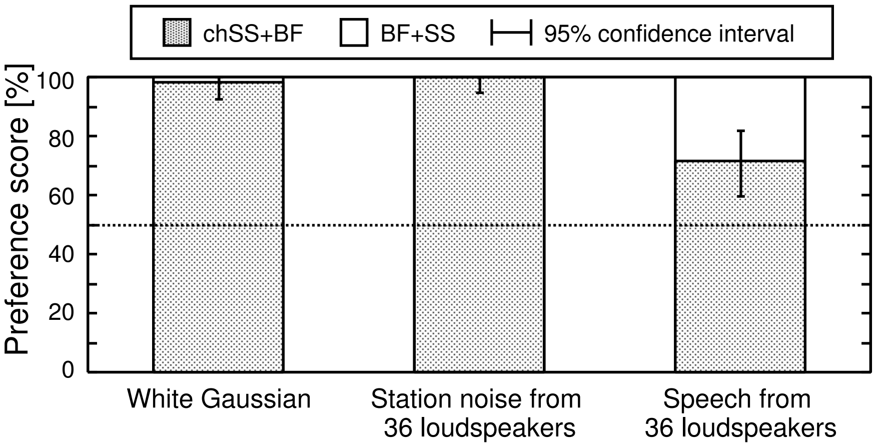


Figure 54. Subjective evaluation results: BF+SS vs. chSS+BF.

In the evaluation, I also presented two signals processed by BSSA and by chBSSA to seven male examinees in random order, who were asked to choose which signal they considered to contain less musical noise. The experimental configurations are the same as the configurations of Sect. 8.6.2 except for the number of displayed signals to examinees. In this evaluation, 20 pairs of signals per type of noise, i.e., a total of 60 pairs of processed signals, were presented to each examinee.

Figure 56 illustrates the subjective evaluation result, and Fig. 57 shows example spectrograms of signals processed by BSSA and by chBSSA. From Fig 56, we can confirm that the output signal of chBSSA is preferred to that of the original BSSA. Actually, it is confirmed that chBSSA reduces isolated components in time-frequency domain sequences, which is a factor of musical noise, rather than BSSA from Fig. 57. Therefore I conclude that the chSS+BF structure is applicable to less-musical noise methods integrating microphone array and spectral subtraction.

8.7. Conclusion

In this chapter, I carried out the analysis on the amount of musical noise generated via BF+SS and chSS+BF on the basis of kurtosis. First, I conducted an analysis on

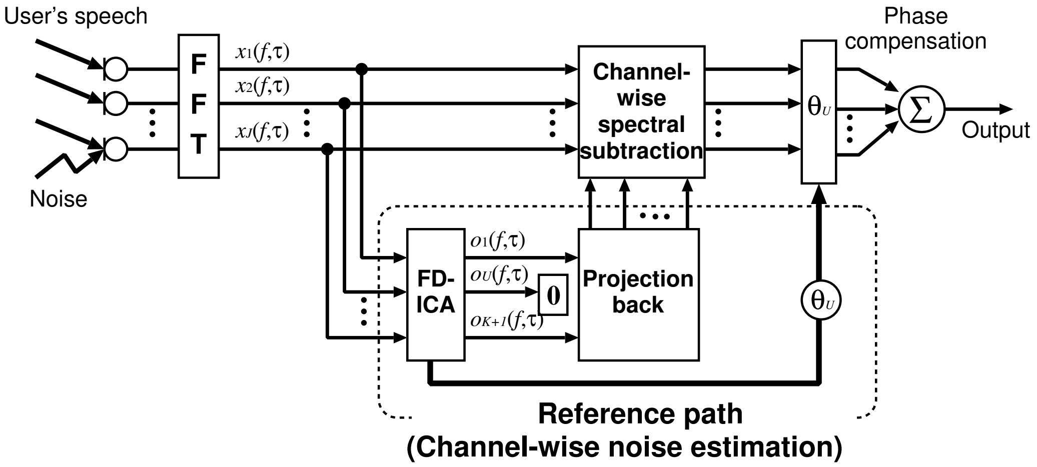


Figure 55. Block diagram of chBSSA.

the amount of musical noise generation thorough BF+SS and chSS+BF without consideration of noise reduction performance. However, I also revealed that the noise reduction performance of both methods are not equivalent even if the same parameters are set in SS part. Therefore, I introduced the new flooring parameter so that the noise reduction performance of both methods become equivalent. As a result of the analysis under the same noise reduction performance condition, it could be concluded that chSS+BF reduces the kurtosis and musical noise more than BF+SS for almost all cases. Moreover, the analysis validity is supported by computer simulations and subjective evaluations.

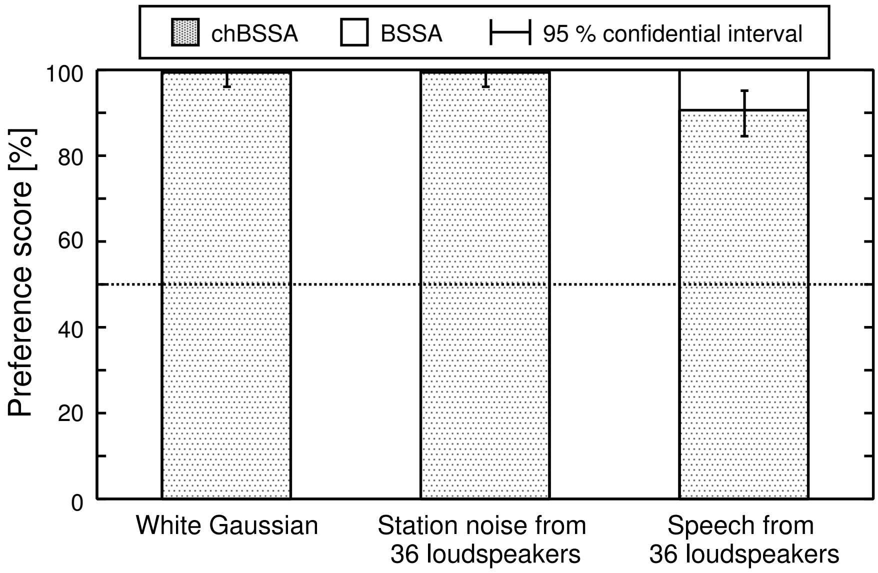


Figure 56. Subjective evaluation results: BSSA vs. chBSSA.

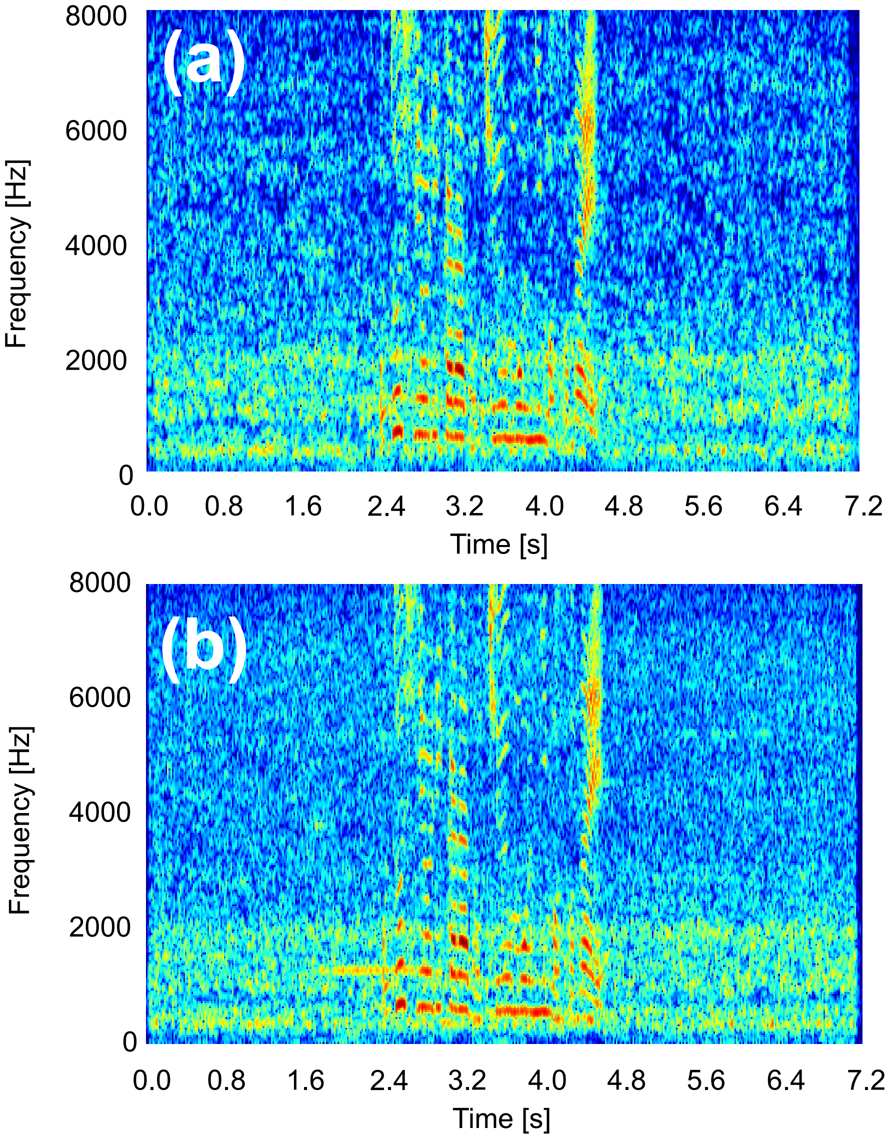


Figure 57. Example spectrograms of signals processed by (a) BSSA and by (b) chBSSA.

CHAPTER 9

EPILOGUE

9.1. Thesis summary

In this thesis, I proposed a novel blind speech extraction method, i.e., BSSA, that can be applied to actual world problem. Moreover, I constructed the real-time algorithm of the proposed BSSA, and built the hands-free spoken-oriented guidance system with the proposed real-time BSSA. As a result of computer simulations and real-world experiments, it was revealed that the proposed BSSA and real-time BSSA improve the speech recognition performance. Furthermore, I performed an analysis on the amount of musical-noise generated via methods of integrating microphone array and SS like BSSA on the basis of HOS. The analytic result clarified that the specific integration structure, i.e., chSS+BF, is proper to applications for human hearing.

In Chapter 3, the theoretical analysis of ICA under non-point-source noise condition was given. As a result of the analysis, I founded out that the conventional ICA is proficient in noise estimation under non-point-source noise condition. Besides, a computer simulation result that supports the analysis result was also demonstrated.

Based on the above-mentioned findings, I proposed a novel blind source extraction method, i.e., BSSA, in Chapter 4. In the chapter, I provided detailed signal processing of BSSA and the analysis of the permutation robustness in BSSA. Moreover, I showed strong evidences of the efficacy of the proposed BSSA via experimental results in not only an experimental room but also an actual world scenario.

In Chapter 5, I presented an alternative analysis of the proposed BSSA with comparing to the conventional SSA. As a result of the alternative analysis, it is clarified that the proposed BSSA has the robustness against room reverberation and microphone element errors.

In Chapter 6, I established the real-time algorithm of the proposed BSSA, and developed a hands-free spoken-oriented guidance system with the real-time BSSA. Furthermore, the result of the speech recognition test of the proposed real-time BSSA was also provided. The proposed real-time BSSA achieved enough speech recognition performance, particularly over 80% word correct. Also the delay of the proposed real-time BSSA was about 50 ms. For these reasons, I concluded that the proposed real-time BSSA satisfies requirements of a real-time hand-free spoken-oriented guidance system, which are both speech recognition performance and real-time properties.

In Chapter 7, a preliminary preparation for musical-noise analysis was expounded. Firstly, I described formulae for two typical methods of integrating microphone array and SS. Secondly, the objective metric for musical noise on the basis of HOS was briefly reviewed.

In the following Chapter 8, HOS-based musical-noise analysis in methods of integrating microphone array signal processing and SS were conducted. In that analysis, first, the amount of musical noise generated via DS and SS including the effect of the flooring technique was firstly investigated. Next, the musical-noise generation in two methods of integrating DS and SS, i.e., BF+SS and chSS+BF, were analyzed based on the above-mentioned investigation under the same noise reduction performance condition. The analytic result suggested that chSS+BF outputs less-musical noise signals. Also, the informal listening test advocated the analytic result. These results let me conclude that the chSS+BF structure is proper to applications for human hearing.

In summary, the acquired conclusions of the study are outlined in the following points:

- It is theoretically clarified that the conventional ICA is proficient in noise estimation rather than in target speech estimation under non-point-source noise condition.

- The proposed BSSA that utilizes ICA as an accurate noise estimator achieves better noise reduction performance than that by the conventional ICA.

- Also the proposed BSSA has a remarkable property that is the robustness against reverberation and microphone element errors.

- The proposed real-time BSSA accomplishes enough speech recognition performance and low-latency blind source extraction for hand-free systems.
- The chSS+BF structure is preferable for human-hearing application.

9.2. Future work

In the thesis, I have improved the source extraction performance for hand-free systems, and the proposed algorithm has realized enough performance for developing spoken-oriented speech guidance systems. However, the following problems are still opened.

For human-hearing applications, the output sound quality including not only noise reduction performance but also listenability is the most important factor. However, musical noise always deteriorates listenability of the output signal. This problem cannot be avoided in methods utilizing nonlinear signal processing like SS. Indeed, I have provided the less musical noise structure for methods of integrating microphone array signal processing and SS, but it cannot control the amount of musical noise generated. Then, a method can take control of the musical-noise generation is needed to develop. Fortunately, I have gained the objective metric for musical noise, i.e. kurtosis-based musical noise metric. On the basis of this objective metric, we would establish the optimization techniques from the viewpoint of not only noise reduction performance but also the amount musical noise generated.

Moreover, in this dissertation, I have analyzed the amount of musical noise generated through only methods of integrating microphone array and spectral subtraction. However, there exists various kinds of method using another nonlinear signal processing. For instance, Okamoto et al. have proposed the methods of integrating ICA and MMSE STSA estimator, it has been reported that the integration method is proper to human hearing [54]. In the future, it is needed that the analysis in terms of another nonlinear signal processing and its integration methods.

Appendix

A. When $|\arg r_n^* - \arg r_s| > \pi/2$, higher-order cross correlation is not minimized

In this section, I clarify that higher-order cross correlation between $y_{\mathrm{s}}(f,\tau)$ and $y_n(f,\tau)$ is not minimized when $y_{\mathrm{s}}(f,\tau)$ and $y_n(f,\tau)$ are orthogonalized. However, it is difficult to give the generalized proof, I give the analysis for the specific case below.

Here, I consider the case where the 2nd-order cross correlation coefficient of the ICA's output $y_s(f,\tau)$ and $y_n(f,\tau)$ is completely zero. The following r_s and r_n make the 2nd-order cross correlation coefficient zero:

$$r_n = -\frac{\langle \hat{n}(f,\tau)\hat{n}^*(f,\tau)\rangle_\tau}{\langle \hat{s}(f,\tau)\hat{s}^*(f,\tau)\rangle_\tau}, \tag{141}$$

$$r_s = 1, \tag{142}$$

where $|\arg r_n^* - \arg r_s| = \pi(> \pi/2)$. Actually, using these r_s and r_n,

$$\begin{aligned}\langle y_s(f,\tau)y_n^*(f,\tau)\rangle_\tau &= \langle(\hat{s}(f,\tau) + r_s\hat{n}(f,\tau))(\hat{n}(f,\tau) + r_n\hat{s}(f,\tau))\rangle_\tau \\ &= r_s\langle \hat{n}(f,\tau)\hat{n}^*(f,\tau)\rangle_\tau + r_n\langle \hat{s}(f,\tau)\hat{s}^*(f,\tau)\rangle_\tau \\ &= \langle \hat{n}(f,\tau)\hat{n}^*(f,\tau)\rangle_\tau - \frac{\langle \hat{n}(f,\tau)\hat{n}^*(f,\tau)\rangle_\tau}{\langle \hat{s}(f,\tau)\hat{s}^*(f,\tau)\rangle_\tau}\langle \hat{s}(f,\tau)\hat{s}^*(f,\tau)\rangle_\tau \\ &= 0. \end{aligned} \tag{143}$$

Anyway, the nonlinear function $\varphi(x) = \tanh(x^{(\mathrm{R})}) + i\tanh(x^{(\mathrm{I})})$ $(x \in \mathbb{C})$ can be expanded by Tailor expansion as,

$$\varphi(x) = x - \frac{1}{3}(\mathrm{Re}\,[x]^3 + i\cdot\mathrm{Im}\,[x]^3) + \cdots. \tag{144}$$

Therefore, the nonlinear cross correlation matrix is

$$\begin{aligned}
&\left\langle \varphi\left(\begin{bmatrix} y_s(f,\tau) \\ y_n(f,\tau) \end{bmatrix}\right) [y_s^*(f,\tau) y_n^*(f,\tau)] \right\rangle_\tau \\
&\quad = \left\langle \begin{bmatrix} y_s(f,\tau) - \frac{1}{3}(\mathrm{Re}\,[y_s(f,\tau)]^3 + i\cdot \mathrm{Im}\,[y_s(f,\tau)]^3) + \cdots \\ y_n(f,\tau) - \frac{1}{3}(\mathrm{Re}\,[y_n(f,\tau)]^3 + i\cdot \mathrm{Im}\,[y_n(f,\tau)]^3) + \cdots \end{bmatrix} [y_s^*(f,\tau), y_n^*(f,\tau)] \right\rangle_\tau
\end{aligned} \tag{145}$$

To analyze the higher-order cross correlation coefficient, I focus my attention on the 1st row and 2nd column factor of (145), C_{12}. C_{12} is represented by

$$C_{12} = \langle y_s(f,\tau) y_s^*(f,\tau)\rangle_\tau - \left\langle \frac{1}{3}\left(\mathrm{Re}\,[y_s(f,\tau)]^3 + i\cdot \mathrm{Im}\,[y_s(f,\tau)]^3\right) y_n^*(f,\tau) \right\rangle_\tau + \cdots . \tag{146}$$

The 4th-order cross correlation coefficient is

$$\begin{aligned}
&-\frac{1}{3}\left\langle \left(\mathrm{Re}\,[y_s(f,\tau)]^3 + i\cdot \mathrm{Im}\,[y_s(f,\tau)]^3\right) y_n^*(f,\tau) \right\rangle_\tau \\
&\quad = -\frac{1}{3}\left\langle (\mathrm{Re}\,[y_s(f,\tau)]^3 + i\cdot \mathrm{Im}\,[y_s(f,\tau)]^3)(\hat{n}^* + r_s \hat{s}^*(f,\tau)) \right\rangle_\tau \\
&\quad = -\frac{1}{3}\Bigg(\left\langle \mathrm{Re}\,[y_s(f,\tau)]^3\, \hat{n}^*(f,\tau) \right\rangle_\tau + i\cdot \left\langle \mathrm{Im}\,[y_s(f,\tau)]^3\, \hat{n}^*(f,\tau) \right\rangle_\tau \\
&\qquad + r_s \left\langle \mathrm{Re}\,[y_s(f,\tau)]^3\, \hat{s}^*(f,\tau) \right\rangle_\tau + i\cdot r_s \left\langle \mathrm{Im}\,[y_s(f,\tau)]^3\, \hat{s}^*(f,\tau) \right\rangle_\tau \Bigg) \\
&\quad = \frac{\langle \hat{n}(f,\tau)\hat{n}^*(f,\tau)\rangle_\tau}{3\,\langle \hat{s}(f,\tau)\hat{s}^*(f,\tau)\rangle_\tau} \left(\left\langle \mathrm{Re}\,[y_s(f,\tau)]^3\, \hat{s}^*(f,\tau) \right\rangle_\tau + i\cdot \left\langle \mathrm{Im}\,[y_s(f,\tau)]^3\, \hat{s}^*(f,\tau) \right\rangle_\tau \right) \\
&\quad \neq 0 \quad (\text{where } \langle \hat{n}(f,\tau)\hat{n}^*(f,\tau)\rangle_\tau \neq 0).
\end{aligned} \tag{147}$$

Consequently, indeed r_s and r_n which satisfy $|\arg r_n^* - \arg r_s| > \pi/2$ let $y_s(f,\tau)$ and $y_n(f,\tau)$ be orthogonalized but those let higher-order cross correlation not be zero.

B. Strategy of Selecting Target Speech Signal

For noise estimation in BSSA, the target speech must be removed from the separation results of ICA. Therefore, a method of choosing the target speech from ICA outputs is required in BSSA. Some methods of choosing the target speech signal from ICA outputs are considered as follows:

- If an approximate location of a target speaker is known in advance, we can utilize the location of a target speaker. For instance, we can know the approximate location of the target speaker at a hands-free speech recognition system in a car or of a public guidance system in advance. Then, the DOA of the target speech signal is approximately known. For such systems, we can choose the target speech signal, selecting the specific component in which the estimated DOA by ICA is nearest the known target-speech DOA. the basis of the estimated DOA by ICA.

- For an interaction robot system, we can utilize image information from a camera mounted on a robot. Therefore, we can estimate DOA from this information, and we can choose the target speech signal on the basis of this estimated DOA.

- If the only target signal is speech, i.e., all noises are not speech, we can choose the target speech signal on the basis of the Gaussian mixture model (GMM) that can classify sound signals into voices and nonvoices [47].

C. Mel-Scale Filter Bank Analysis

The proposed BSSA involves mel-scale filter bank analysis, and directly outputs MFCC. The triangular window $W_{\mathrm{mel}}(f;l)$ $(l = 1, \cdots, L)$ for performing mel-scale filter bank analysis is designated as

$$W_{\mathrm{mel}}(f;l) = \begin{cases} \dfrac{f - f_{\mathrm{lo}}(l)}{f_{\mathrm{c}}(l) - f_{\mathrm{lo}}(l)} & (f_{\mathrm{lo}}(l) \leq f \leq f_{\mathrm{c}}(l)), \\ \dfrac{f_{\mathrm{hi}}(l) - f}{f_{\mathrm{hi}}(l) - f_{\mathrm{c}}(l)} & (f_{\mathrm{c}}(l) \leq f \leq f_{\mathrm{hi}}(l)), \end{cases} \tag{148}$$

where $f_{\mathrm{lo}}(l)$, $f_{\mathrm{c}}(l)$, and $f_{\mathrm{hi}}(l)$ are the lower, center, and higher frequency bins of each triangle window, respectively. Furthermore, L is the dimension of the mel-scale filter bank. They satisfy the relation among adjacent windows as

$$f_{\mathrm{c}}(l) = f_{\mathrm{hi}}(l-1) = f_{\mathrm{lo}}(l+1). \tag{149}$$

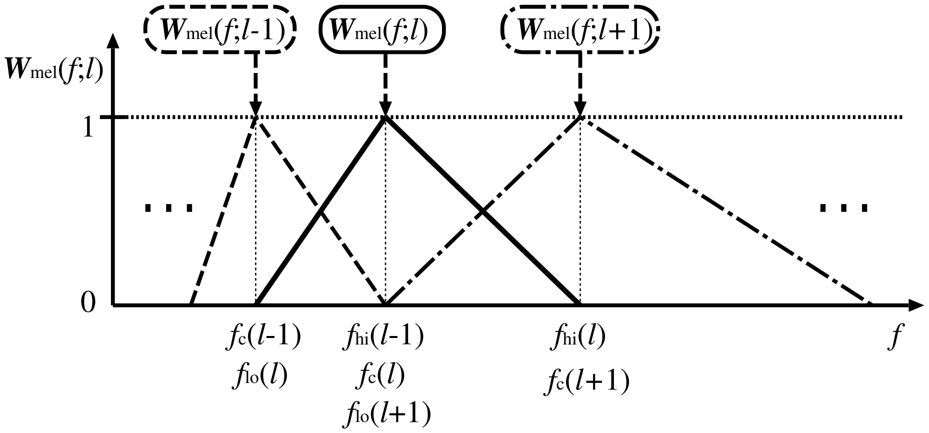


Figure 58. Configuration of mel-scale filter bank.

Moreover, $f_{\mathrm{c}}(l)$ is arranged at regular intervals on a mel-frequency domain. The mel-scale frequency $Mel_{f_{\mathrm{c}}(l)}$ for $f_{\mathrm{c}}(l)$ is calculated using

$$Mel_{f_{\mathrm{c}}(l)} = 2595 \log_{10}\{1 + f_{\mathrm{c}}(l) f_{\mathrm{s}}/(700 \cdot M)\}. \tag{150}$$

The mel-scale filter bank analysis is given by

$$m(l, \tau) = \sum_{f=f_{\mathrm{lo}}(l)}^{f_{\mathrm{hi}}(l)} W_{\mathrm{mel}}(f; l) y_{\mathrm{BSSA}}(f, \tau), \tag{151}$$

where $m(l, \tau)$ is the output of the mel-scale filter bank. Moreover, the logarithm transform and discrete cosine transform are performed in the mel-scale filter bank domain to obtain the MFCC for the speech recognizer; this processing can be written as

$$MFCC(\kappa,\tau) = \sqrt{\frac{2}{L}} \sum_{l=1}^{L} \log\{m(l,\tau)\} \cos\left\{\left(l - \frac{1}{2}\right)\frac{\kappa\pi}{L}\right\}, \tag{152}$$

where κ denotes the dimension of MFCC. The proposed BSSA requires no transformation into the time-domain waveform.

D. Derivation of (109)

When we assume that the input signal of the power domain can be modeled by a gamma distribution, the amount of subtraction is $\beta\alpha\theta$. The subtraction of the estimated noise power spectrum in each frequency subband can be considered as a lateral shift of the p.d.f. to the zero-power direction (see Fig. 38). As a result of this subtraction, the random variable x is replaced with $x + \beta\alpha\theta$ and the gamma distribution becomes

$$\hat{P}_{\mathrm{GM}}(x) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} \cdot (x + \beta\alpha\theta)^{\alpha-1} \exp\left\{-\frac{x + \beta\alpha\theta}{\theta}\right\} \quad (x \geq -\beta\alpha\theta). \tag{153}$$

Since the domain of the original gamma distribution is $x \geq 0$, the domain of the resultant p.d.f. is $x \geq -\beta\alpha\theta$. Thus, negative-power components with nonzero probability arise, which can be represented by

$$\hat{P}_{\mathrm{negative}}(x) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} \cdot (x + \beta\alpha\theta)^{\alpha-1} \exp\left\{-\frac{x + \beta\alpha\theta}{\theta}\right\} \quad (-\beta\alpha\theta \leq x \leq 0), \tag{154}$$

where $\hat{P}_{\mathrm{negative}}(x)$ is part of $\hat{P}_{\mathrm{GM}}(x)$. To remove the negative-power components, the signals corresponding to $\hat{P}_{\mathrm{negative}}(x)$ are replaced by observations multiplied by a small positive value η. The observations corresponding to (154), $\hat{P}_{\mathrm{obs}}(x)$, are given by

$$\hat{P}_{\mathrm{obs}}(x) = \frac{1}{\Gamma(\alpha)\theta^{\alpha}} \cdot (x)^{\alpha-1} \exp\left\{-\frac{x}{\theta}\right\} \quad (0 \leq x \leq \beta\alpha\theta). \tag{155}$$

Since a small positive flooring parameter η is applied to (155), the scale parameter θ becomes $\eta^2\theta$ and the range is changed from $0 \leq x \leq \beta\alpha\theta$ to $0 \leq x \leq \beta\alpha\eta^2\theta$. Then, (155) is modified to

$$\hat{P}_{\mathrm{floor}}(x) = \frac{1}{\Gamma(\alpha)(\eta^2\theta)^{\alpha}} \cdot (x)^{\alpha-1} \exp\left\{-\frac{x}{\eta^2\theta}\right\} \quad (0 \leq x \leq \beta\alpha\eta^2\theta), \tag{156}$$

where $\hat{P}_{\mathrm{floor}}(x)$ is the probability of the floored components. This $\hat{P}_{\mathrm{floor}}(x)$ is superimposed on the p.d.f. given by (153) within the range $0 \leq x \leq \beta\alpha\eta^2\theta$. By considering the positive range of (153) and $\hat{P}_{\mathrm{floor}}(x)$, the resultant p.d.f. of SS can be formulated as

$$P_{\mathrm{SS}}(z) = \begin{cases} \dfrac{1}{\theta^{\alpha}\Gamma(\alpha)}(z + \beta\alpha\theta)^{\alpha-1} \exp\left\{-\dfrac{z + \beta\alpha\theta}{\theta}\right\} & (z \geq \beta\alpha\eta^2\theta), \\ \dfrac{1}{\theta^{\alpha}\Gamma(\alpha)}(z + \beta\alpha\theta)^{\alpha-1} \exp\left\{-\dfrac{z + \beta\alpha\theta}{\theta}\right\} + \dfrac{1}{(\eta^2\theta)^{\alpha}\Gamma(\alpha)} z^{\alpha-1} \exp\left\{-\dfrac{z}{\eta^2\theta}\right\} & \\ \qquad (0 < z < \beta\alpha\eta^2\theta), & \end{cases} \tag{157}$$

where the variable x is replaced with z for convenience.

E. Derivation of (110)

To derive the kurtosis after SS, the 2nd- and 4th-order moments of z are required. For $P_{\mathrm{SS}}(z)$, the 2nd-order moment can be given by

$$\begin{aligned}\mu_2 &= \int_0^\infty z^2 \cdot P_{\mathrm{SS}}(z)\mathrm{d}z \\ &= \int_0^\infty z^2 \frac{1}{\theta^\alpha\Gamma(\alpha)}(z+\beta\alpha\theta)^{\alpha-1}\exp\left\{-\frac{z+\beta\alpha\theta}{\theta}\right\}\mathrm{d}z \\ &\qquad + \int_0^{\beta\alpha\eta^2\theta} z^2 \frac{1}{(\eta^2\theta)^\alpha\Gamma(\alpha)}z^{\alpha-1}\exp\left\{-\frac{z}{\eta^2\theta}\right\}\mathrm{d}z. \qquad (158)\end{aligned}$$

We now expand the first term of the right-hand side of (158). Here, let $t = (z + \beta\alpha\theta)/\theta$, then $\theta\mathrm{d}t = \mathrm{d}z$ and $z = \theta(t - \beta\alpha)$. Consequently,

$$\begin{aligned}&\int_0^\infty z^2 \frac{1}{\theta^\alpha\Gamma(\alpha)}(z+\beta\alpha\theta)^{\alpha-1}\exp\left\{-\frac{z+\beta\alpha\theta}{\theta}\right\}\mathrm{d}z \\ &\quad= \int_{\beta\alpha}^\infty \theta^2(t-\beta\alpha)^2\frac{1}{\theta^\alpha\Gamma(\alpha)}(\theta t)^{\alpha-1}\exp\{-t\}\theta\mathrm{d}t \\ &\quad= \frac{\theta^2}{\Gamma(\alpha)}\int_{\beta\alpha}^\infty (t^2 - 2\beta\alpha t + \beta^2\alpha^2)t^{\alpha-1}\exp\{-t\}\mathrm{d}t \\ &\quad= \frac{\theta^2}{\Gamma(\alpha)}\left[\Gamma(\beta\alpha,\alpha+2) - 2\beta\alpha\Gamma(\beta\alpha,\alpha+1) + \beta^2\alpha^2\Gamma(\beta\alpha,\alpha)\right]. \qquad (159)\end{aligned}$$

Next we consider the second term of the right-hand side of (158). Here, let $t = z/(\eta^2\theta)$ then $\eta^2\theta\mathrm{d}t = \mathrm{d}z$. Thus,

$$\begin{aligned}&\int_0^{\beta\alpha\eta^2\theta} z^2 \frac{1}{(\eta^2\theta)^\alpha\Gamma(\alpha)}z^{\alpha-1}\exp\left\{-\frac{z}{\eta^2\theta}\right\}\mathrm{d}z \\ &\quad= \int_0^{\beta\alpha}(\eta^2\theta t)^2\frac{1}{(\eta^2\theta)^\alpha\Gamma(\alpha)}(\eta^2\theta t)^{\alpha-1}\exp\{-t\}\eta^2\theta\mathrm{d}t \\ &\quad= \frac{\eta^4\theta^2}{\Gamma(\alpha)}\int_0^{\beta\alpha} t^{\alpha+1}\exp\{-t\}\mathrm{d}t \\ &\quad= \eta^4\theta^2\frac{\gamma(\beta\alpha,\alpha+2)}{\Gamma(\alpha)}. \qquad (160)\end{aligned}$$

As a result, the 2nd-order moment after SS, $\mu_2^{(\mathrm{SS})}$, is a composite of (159) and (160), and can be given as

$$\mu_2^{(\mathrm{SS})} = \frac{\theta^2}{\Gamma(\alpha)}\left[\Gamma(\beta\alpha,\alpha+2) - 2\beta\alpha\Gamma(\beta\alpha,\alpha+1) + \beta^2\alpha^2\Gamma(\beta\alpha,\alpha) + \eta^4\gamma(\beta\alpha,\alpha+2)\right]. \qquad (161)$$

In the same manner, the 4th-order moment after SS, $\mu_4^{(\mathrm{SS})}$, can be represented by

$$\mu_4^{(\mathrm{SS})} = \frac{\theta^4}{\Gamma(\alpha)}\Big[\Gamma(\beta\alpha, \alpha+4) - 4\beta\alpha\Gamma(\beta\alpha, \alpha+3) + 6\beta^2\alpha^2\Gamma(\beta\alpha, \alpha+2) \\ - 4\beta^3\alpha^3\Gamma(\beta\alpha, \alpha+1) + \beta^4\alpha^4\Gamma(\beta\alpha, \alpha) + \eta^8\gamma(\beta\alpha, \alpha+4)\Big]. \quad (162)$$

Consequently, using (161) and (162), the kurtosis after SS can be given as

$$\mathrm{kurt}_{\mathrm{SS}} = \Gamma(\alpha)\frac{\mathcal{F}(\alpha,\beta,\eta)}{\mathcal{G}^2(\alpha,\beta,\eta)}, \quad (163)$$

where

$$\mathcal{G}(\alpha,\beta,\eta) = \Gamma(\alpha)\Gamma(\beta\alpha, \alpha+2) - 2\beta\alpha\Gamma(\beta\alpha, \alpha+1) + \beta^2\alpha^2\Gamma(\beta\alpha, \alpha) + \eta^4\gamma(\beta\alpha, \alpha+2), \quad (164)$$

$$\mathcal{F}(\alpha,\beta,\eta) = \Gamma(\beta\alpha, \alpha+4) - 4\beta\alpha\Gamma(\beta\alpha, \alpha+3) + 6\beta^2\alpha^2\Gamma(\beta\alpha, \alpha+2) \\ - 4\beta^3\alpha^3\Gamma(\beta\alpha, \alpha+1) + \beta^4\alpha^4\Gamma(\beta\alpha, \alpha) + \eta^8\gamma(\beta\alpha, \alpha+4). \quad (165)$$

F. Derivation of (119)

As described in (108), the power-domain signal is the sum of two squares of random variables with the same distribution. Using (115), the power-domain cumulants $K_n^{(\mathrm{p})}$ can be written as

$$\text{power-domain cumulants}\begin{cases} K_1^{(\mathrm{p})} = 2K_1^{(2)}, \\ K_2^{(\mathrm{p})} = 2K_2^{(2)}, \\ K_3^{(\mathrm{p})} = 2K_3^{(2)}, \\ K_4^{(\mathrm{p})} = 2K_4^{(2)}, \end{cases} \quad (166)$$

where $K_n^{(2)}$ is the nth square-domain moment. Here, the p.d.f. of such a square-domain signal is not symmetrical and its mean is not zero. Thus, we utilize the following relations between the moments and cumulants around the origin:

$$\text{moments}\begin{cases} \mu_1 = \kappa_1, \\ \mu_2 = \kappa_2 + \kappa_1^2, \\ \mu_4 = \kappa_4 + 4\kappa_3\kappa_1 + 3\kappa_2^2 + 6\kappa_2\kappa_1^2 + \kappa_1^4, \end{cases} \quad (167)$$

where μ_n is the nth-order raw moment and κ_n is the nth-order cumulant. Moreover, the square-domain moments $\mu_n^{(2)}$ can be expressed by

$$\text{squared-domain moments}\begin{cases}\mu_1^{(2)} = \mu_2,\\ \mu_2^{(2)} = \mu_4,\\ \mu_4^{(2)} = \mu_8.\end{cases} \tag{168}$$

Using (166)–(168), the power-domain moments can be expressed in terms of the 4th- and 8th-order moments in the time domain. Therefore, to obtain the kurtosis after DS in the power domain, the moments and cumulants after DS up to the 8th-order are needed.

The 3rd-, 5th-, and 7th-order cumulants are zero because we assume that the p.d.f. of x_j is symmetrical and that its mean is zero. If these conditions are satisfied, the following relations between moments and cumulants hold

$$\text{moments}\begin{cases}\mu_1 = 0,\\ \mu_2 = \kappa_2,\\ \mu_4 = \kappa_4 + 3\kappa_2^2,\\ \mu_6 = \kappa_6 + 15\kappa_4\kappa_2 + 15\kappa_2^3,\\ \mu_8 = \kappa_8 + 35\kappa_4^2 + 28\kappa_2\kappa_6 + 210\kappa_2^2\kappa_4 + 105\kappa_2^4.\end{cases} \tag{169}$$

Using (118) and (169), the time-domain moments after DS are designated as

$$\text{moments after DS}\begin{cases}\mu_2^{(\text{DS})} = \mathcal{K}_2,\\ \mu_4^{(\text{DS})} = \mathcal{K}_4 + 3\mathcal{K}_2^2,\\ \mu_6^{(\text{DS})} = \mathcal{K}_6 + 15\mathcal{K}_2\mathcal{K}_4 + 15\mathcal{K}_2^3,\\ \mu_8^{(\text{DS})} = \mathcal{K}_8 + 35\mathcal{K}_4^2 + 28\mathcal{K}_2\mathcal{K}_6 + 210\mathcal{K}_2^2\mathcal{K}_4 + 105\mathcal{K}_2^4,\end{cases} \tag{170}$$

where $\mu_n^{(\text{DS})}$ is the nth-order raw moment after DS in the time domain.

Using (167), (168) and (170), the square-domain cumulants can be written as

$$\text{square-domain cumulants}\begin{cases}\mathcal{K}_1^{(2)} = \mathcal{K}_2,\\ \mathcal{K}_2^{(2)} = \mathcal{K}_4 + 2\mathcal{K}_2^2,\\ \mathcal{K}_3^{(2)} = \mathcal{K}_6 + 12\mathcal{K}_4\mathcal{K}_2 + 8\mathcal{K}_2^3,\\ \mathcal{K}_4^{(2)} = \mathcal{K}_8 + 32\mathcal{K}_4^2 + 24\mathcal{K}_2\mathcal{K}_6 + 144\mathcal{K}_2^2\mathcal{K}_4 + 48\mathcal{K}_2^4,\end{cases} \tag{171}$$

where $\mathcal{K}_n^{(2)}$ is the nth-order cumulant in the square domain.

Moreover, using (166), (167), and (171), the 2nd- and 4th-order power-domain moments can be written as

$$\mu_2^{(\mathrm{p})} = 2\left(\mathcal{K}_4 + 4\mathcal{K}_2^2\right), \tag{172}$$

$$\mu_4^{(\mathrm{p})} = 2\left(\mathcal{K}_8 + 38\mathcal{K}_4^2 + 32\mathcal{K}_6\mathcal{K}_2 + 288\mathcal{K}_4\mathcal{K}_2^2 + 192\mathcal{K}_2^4\right). \tag{173}$$

As a result, the power-domain kurtosis after DS, $\mathrm{kurt}_{\mathrm{DS}}$, can be given as

$$\mathrm{kurt}_{\mathrm{BF}} = \frac{\mathcal{K}_8 + 38\mathcal{K}_4^2 + 32\mathcal{K}_2\mathcal{K}_6 + 288\mathcal{K}_2^2\mathcal{K}_4 + 192\mathcal{K}_2^4}{2\mathcal{K}_4^2 + 16\mathcal{K}_2^2\mathcal{K}_4 + 32\mathcal{K}_2^4}. \tag{174}$$

G. Derivation of (121)

According to (107), the shape parameter $\hat{\alpha}$ corresponding to the kurtosis after DS, $\mathrm{kurt}_{\mathrm{DS}}$, is given by the solution of the quadratic equation

$$\mathrm{kurt}_{\mathrm{DS}} = \frac{(\hat{\alpha}+2)(\hat{\alpha}+3)}{\hat{\alpha}(\hat{\alpha}+1)}. \tag{175}$$

This can be expanded as

$$\hat{\alpha}^2(\mathrm{kurt}_{\mathrm{DS}} - 1) + \hat{\alpha}(\mathrm{kurt}_{\mathrm{DS}} - 5) + 6 = 0. \tag{176}$$

Using the quadratic formula,

$$\hat{\alpha} = \frac{-\mathrm{kurt}_{\mathrm{DS}} + 1 \pm \sqrt{\mathrm{kurt}_{\mathrm{DS}}^2 + 14\,\mathrm{kurt}_{\mathrm{DS}} + 1}}{2\,\mathrm{kurt}_{\mathrm{DS}} - 2}, \tag{177}$$

whose denominator is larger than zero because $\mathrm{kurt}_{\mathrm{DS}} > 1$. Here, since $\hat{\alpha} > 0$, we must select the appropriate numerator of (177). First, suppose that

$$-\mathrm{kurt}_{\mathrm{DS}} + 1 + \sqrt{\mathrm{kurt}_{\mathrm{DS}}^2 + 14\,\mathrm{kurt}_{\mathrm{DS}} + 1} > 0. \tag{178}$$

This inequality clearly holds when $1 < \mathrm{kurt}_{\mathrm{DS}} < 5$ because $-\mathrm{kurt}_{\mathrm{DS}} + 5 > 0$ and $\sqrt{\mathrm{kurt}_{\mathrm{DS}}^2 + 14\,\mathrm{kurt}_{\mathrm{DS}} + 1} > 0$. Thus,

$$-\mathrm{kurt}_{\mathrm{DS}} + 5 > -\sqrt{\mathrm{kurt}_{\mathrm{DS}}^2 + 14\,\mathrm{kurt}_{\mathrm{DS}} + 1}. \tag{179}$$

When $\text{kurt}_{\text{DS}} \geq 5$, then the following relation also holds:

$$\begin{aligned}&(-\text{kurt}_{\text{DS}}+5)^2 < \text{kurt}_{\text{DS}}^2 + 14\,\text{kurt}_{\text{DS}} + 1,\\ \Longleftrightarrow \quad & 24\,\text{kurt}_{\text{DS}} > 24. \end{aligned} \tag{180}$$

Since (180) is true when $\text{kurt}_{\text{DS}} \geq 5$, (178) holds. In summary, (178) always holds for $1 < \text{kurt}_{\text{DS}} < 5$ and $5 \leq \text{kurt}_{\text{DS}}$. Thus,

$$-\text{kurt}_{\text{DS}}+5+\sqrt{\text{kurt}_{\text{DS}}^2 + 14\,\text{kurt}_{\text{DS}} + 1} > 0 \quad \text{for} \quad \text{kurt}_{\text{DS}} > 1. \tag{181}$$

Overall,

$$\frac{-\text{kurt}_{\text{DS}}+5+\sqrt{\text{kurt}_{\text{DS}}^2 + 14\,\text{kurt}_{\text{DS}} + 1}}{2\,\text{kurt}_{\text{DS}} - 2} > 0. \tag{182}$$

On the other hand, let

$$-\text{kurt}_{\text{DS}}+5-\sqrt{\text{kurt}_{\text{DS}}^2 + 14\,\text{kurt}_{\text{DS}} + 1} > 0, \tag{183}$$

then this inequality is *not* satisfied when $\text{kurt}_{\text{DS}} > 5$ because $-\text{kurt}_{\text{DS}}+5 < 0$ and $\sqrt{\text{kurt}_{\text{DS}}^2 + 14\,\text{kurt}_{\text{DS}} + 1} > 0$. Now (183) can be modified as

$$-\text{kurt}_{\text{DS}}+5 > \sqrt{\text{kurt}_{\text{DS}}^2 + 14\,\text{kurt}_{\text{DS}} + 1}, \tag{184}$$

then the following relation also holds for $1 < \text{kurt}_{\text{DS}} \leq 5$;

$$\begin{aligned}&(-\text{kurt}_{\text{DS}}+5)^2 > \text{kurt}_{\text{DS}}^2 + 14\,\text{kurt}_{\text{DS}} + 1,\\ \Longleftrightarrow \quad & 24\,\text{kurt}_{\text{DS}} < 24. \end{aligned} \tag{185}$$

This is *not* true for $1 < \text{kurt}_{\text{DS}} \leq 5$. Thus, (183) is not appropriate for $\text{kurt}_{\text{DS}} > 1$. Therefore, $\hat{\alpha}$ corresponding to kurt_{DS} can be given by

$$\hat{\alpha} = \frac{-\text{kurt}_{\text{DS}}+5+\sqrt{\text{kurt}_{\text{DS}}^2 + 14\,\text{kurt}_{\text{DS}} + 1}}{2\,\text{kurt}_{\text{DS}} - 2}. \tag{186}$$

H. Derivation of (135)

For $0 < \alpha \leq 1$, which corresponds to a Gaussian or super-Gaussian input signal, it is revealed that noise reduction performance of BF+SS is superior to that of chSS+BF from the numerical simulation in Sect. 8.4. Thus, the following relation holds:

$$\begin{aligned} &-10\log_{10}\frac{1}{J\cdot\Gamma(\hat{\alpha})}\left[\frac{\Gamma(\beta\hat{\alpha},\hat{\alpha}+1)}{\hat{\alpha}}-\beta\cdot\Gamma(\beta\hat{\alpha},\hat{\alpha})+\eta^2\frac{\gamma(\beta\hat{\alpha},\hat{\alpha}+1)}{\hat{\alpha}}\right] \\ &\qquad\geq -10\log_{10}\frac{1}{J\cdot\Gamma(\alpha)}\left[\frac{\Gamma(\beta\alpha,\alpha+1)}{\alpha}-\beta\cdot\Gamma(\beta\alpha,\alpha)+\eta^2\frac{\gamma(\beta\alpha,\alpha+1)}{\alpha}\right]. \end{aligned} \tag{187}$$

This inequality corresponds to

$$\begin{aligned} &\frac{1}{\Gamma(\hat{\alpha})}\left[\frac{\Gamma(\beta\hat{\alpha},\hat{\alpha}+1)}{\hat{\alpha}}-\beta\cdot\Gamma(\beta\hat{\alpha},\hat{\alpha})+\eta^2\frac{\gamma(\beta\hat{\alpha},\hat{\alpha}+1)}{\hat{\alpha}}\right] \\ &\qquad\leq \frac{1}{\Gamma(\alpha)}\left[\frac{\Gamma(\beta\alpha,\alpha+1)}{\alpha}-\beta\cdot\Gamma(\beta\alpha,\alpha)+\eta^2\frac{\gamma(\beta\alpha,\alpha+1)}{\alpha}\right]. \end{aligned} \tag{188}$$

Then, the new flooring parameter $\hat{\eta}$ in BF+SS, which makes the noise reduction performance of BF+SS equal to that of chSS+BF, satisfies $\hat{\eta} \geq \eta$ (≥ 0) because

$$\frac{\gamma(\beta\hat{\alpha},\hat{\alpha}+1)}{\hat{\alpha}} \geq 0. \tag{189}$$

Moreover, the following relation for $\hat{\eta}$ also holds:

$$\begin{aligned} &\frac{1}{\Gamma(\hat{\alpha})}\left[\frac{\Gamma(\beta\hat{\alpha},\hat{\alpha}+1)}{\hat{\alpha}}-\beta\cdot\Gamma(\beta\hat{\alpha},\hat{\alpha})+\hat{\eta}^2\frac{\gamma(\beta\hat{\alpha},\hat{\alpha}+1)}{\hat{\alpha}}\right] \\ &\qquad= \frac{1}{\Gamma(\alpha)}\left[\frac{\Gamma(\beta\alpha,\alpha+1)}{\alpha}-\beta\cdot\Gamma(\beta\alpha,\alpha)+\eta^2\frac{\gamma(\beta\alpha,\alpha+1)}{\alpha}\right]. \end{aligned} \tag{190}$$

This can be rewritten as

$$\begin{aligned} &\hat{\eta}^2\frac{\Gamma(\alpha)}{\Gamma(\hat{\alpha})}\frac{\gamma(\beta\hat{\alpha},\hat{\alpha}+1)}{\hat{\alpha}} \\ &\qquad= \left[\frac{\Gamma(\beta\alpha,\alpha+1)}{\alpha}-\beta\cdot\Gamma(\beta\alpha,\alpha)+\eta^2\frac{\gamma(\beta\alpha,\alpha+1)}{\alpha}\right] \\ &\qquad\qquad -\frac{\Gamma(\alpha)}{\Gamma(\hat{\alpha})}\left[\frac{\Gamma(\beta\hat{\alpha},\hat{\alpha}+1)}{\hat{\alpha}}-\beta\cdot\Gamma(\beta\hat{\alpha},\hat{\alpha})\right], \end{aligned} \tag{191}$$

and consequently

$$\hat{\eta}^2 = \frac{\hat{\alpha}}{\gamma(\beta\hat{\alpha}, \hat{\alpha} + 1)} \cdot \left[\frac{\Gamma(\hat{\alpha})}{\Gamma(\alpha)}\mathcal{H}(\alpha,\beta,\eta) - \mathcal{I}(\hat{\alpha},\beta)\right], \tag{192}$$

where $\mathcal{H}(\alpha,\beta,\eta)$ is defined by (136) and $\mathcal{I}(\alpha,\beta)$ is given by (137). Using (189) and (190), the right-hand side of (191) is clearly greater than or equal to zero. Moreover, since $\Gamma(\alpha) > 0$, $\Gamma(\hat{\alpha}) > 0$, $\hat{\alpha} > 0$, and $\gamma(\beta\hat{\alpha}, \hat{\alpha} + 1) > 0$, the right-hand side of (192) is also greater than or equal to zero. Therefore,

$$\hat{\eta} = \sqrt{\frac{\hat{\alpha}}{\gamma(\beta\hat{\alpha}, \hat{\alpha} + 1)} \cdot \left[\frac{\Gamma(\hat{\alpha})}{\Gamma(\alpha)}\mathcal{H}(\alpha,\beta,\eta) - \mathcal{I}(\hat{\alpha},\beta)\right]}. \tag{193}$$

References

[1] K. Nakadai, D. Matsuura, H. Okuno, and H. Kitano, "Applying scattering theory to robot audition system: robust sound source localization and extraction," *Proceedings of International Conference on Intelligent Robots and Systems*, pp. 1147–1152, 2003.

[2] R. Prasad, H. Saruwatari, and K. Shikano, "Robots that can hear, understand and talk," *Advanced Robotics*, vol. 18, pp. 533–564, 2004.

[3] H. G. Okuno and S. Yamamoto, "Computing for computational auditory scene analysis," *Journal of The Japanese Society for Artificial Intelligence*, vol. 22, no. 6, pp. 846–854, 2001.

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

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

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

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

[8] M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, "Microphone array based speech recognition with different talker-array positions," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 227–230, 1997.

[9] H. F. Silverman and W. R. Pattterson, "Visualizing the performance of large-aperture microphone arrays," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 962–972, 1999.

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

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

[12] Y. Kaneda and J. Ohga, “Adaptive microphone-array system for noise reduction,” *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. 34, no. 6, pp. 1391–1400, 1986.

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

[14] J. F. Cardoso, “Eigenstructure of the 4th-order cumulant tensor with application to the blind source separation problem,” *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2109–2112, 1989.

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

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

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

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

[19] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, and T. Nishikawa, “Blind source separation combining independent component analysis and beam-

forming," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1135–1146, 2003.

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

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

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

[23] J. Meyer and K. Simmer, "Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1167–1170, 1997.

[24] S. Fischer and 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*, pp. 359–362, 1997.

[25] R. Mukai, S. Araki, H. Sawada, and S. Makino, "Removal of residual cross-talk components in blind source separation using time-delayed spectral subtraction," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1789–1792, 2002.

[26] J. Cho and A. Krishnamurthy, “Speech enhancement using microphone array in moving vehicle environment,” *Proc. Intelligent Vehicles Symposium*, pp. 366–371, 2003.

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

[28] J. Even, H. Saruwatari, and K. Shikano, “New architecture combining blind signal extraction and modified spectral subtraction for suppression of background noise,” *Proc. IWAENC2008*, 2008.

[29] S. B. Jebara, “A perceptual approach to reduce musical noise phenomenon with Wiener denoising technique,” *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. III, pp. 49–52, 2006.

[30] Y. Ephrain and D. Malah, “Speech enhancement using minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol. ASSP-32, no. 6, pp. 1109–1121, 1984.

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

[32] C. Serviere and D.-T. Pham, “Permutation correction in the frequnecy-domain in blind separation of speech mixtures,” *EURASIP Journal on Applied Signal Processing*, 2006.

[33] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, “Evaluation of blind signal separation method using directivity pattern under reverberant conditions,” *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 3140–3143, 2000.

[34] H. Sawada, R. Mukai, S. Araki, and S. Makino, “A robust and precise method for solving the permutation problem of frequency-domain blind

source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 530–538, 2004.

[35] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics," *Proc. IWAENC2008*, 2008.

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

[37] S. Ukai, T. Takatani, T. Nishikawa, and H. Saruwatari, "Blind source separation combining SIMO-model-based ICA and adaptive beamforming," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. III, pp. 85–88, 2005.

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

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

[40] M. Mizumachi and M. Akagi, "Noise reduction by paired-microphone using spectral subtraction," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 1001–1004, 1998.

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

[42] S. Makino and T.-W. Lee, Eds., *Blind Speech Separation*. Springer-Verlag, 2007.

[43] L. Rabiner and B. Juang, *Fundamentals of speech recognition*. Upper Saddle River, NJ: Prentice Hall PTR, 1993.

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

[45] PRIMO, Co. Ltd, “Primo electret condenser microphone EM160,” http://www.primocorp.co.jp/product/PDF/EM160.pdf.

[46] H. Kawanami, “Development and operational result of real environment speech-oriented guidance systems kita-robo and kita-chan,” *Oriental COCOSDA 2007*, pp. 132–136, 2007.

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

[48] H. Sakai, T. Cincarek, H. Kawanami, H. Saruwatari, K. Shikano, and A. Lee, “Voice activity detection applied to hands-free spoken dialogue robot based on decoding using acoustic and language model,” *2007 International Conference on Robot Communication and Coordination*, 2007.

[49] Debian Project, “Debian/GNU Linux,” http://www.debian.org/.

[50] T. Takatani, T. Nishikawa, H. Saruwatari, and K. Shikano, “High-fidelity blind separation of acoustic signals using SIMO-model-based independent component analysis,” *IEICE Transactions on fundamentals of electronics, communications and computer sciences*, vol. E87-A, no. 8, pp. 2063–2072, 2004.

[51] E. W. Stacy, “A generalization of the gamma distribution,” *Ann. Math. Statist.*, pp. 1187–1192, 1962.

[52] K. Kokkinakis and A. K. Nandi, “Generalized gamma density-based score functions for fast and flexible ICA,” *Signal Process.*, pp. 1156–1162, 2007.

[53] J. W. Shin, J. Chang, and N. Kim, “Statistical modeling of speech signals based on generalized gamma distribution,” *IEEE Signal Processing Letters*, vol. 12, no. 3, pp. 258–261, 2005.

[54] R. Okamoto, Y. Takahashi, H. Saruwatari, and K. Shikano, “MMSE STSA estimator with nonstationary noise estimation based on ICA for high-quality speech enhancement,” *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. –, 2010.

Acknowledgements

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

I would like to express my sincere thanks to Professor Kiyohiro Shikano in Nara Institute of Science and Technology, my dissertation adviser, for his valuable guidance and constant encouragement.

I also would like to express my appreciation Professor Tsukasa Ogasawara of Nara Institute of Science and Technology, a member of the thesis committee, for his valuable comments to the thesis.

I would especially like to express my deeply grateful and sincere acknowledgement to Associate Professor Hiroshi Saruwatari in 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 his well-directed advice, beneficial suggestion, and fruitful discussion with him. I could learn a lot of things as a researcher from his attitude toward a study. I have always been happy to conduct researches with him.

I would also like to thank to Assistant Professor Tomoki Toda and Assistant Professor Hiromichi Kawanami of Nara Institute of Science and Technology for their instructive comments.

The former part of this work could never have been completed without the collaboration of many researchers. I would like to express my gratitude to Dr. Tomoya Takatani, who is currently a researcher in Toyota Motor Corporation, and Mr. Keiichi Osako, who is currently a researcher in Sony Corporation, for their valuable suggestions and shrewd advice on ICA-based techniques. Also, I would like to acknowledge to Dr. Takashi Hiekata, who is currently a researcher in Kobe Steel Co., Ltd., for his instructive advice on real-time implementation of my proposed algorithm. Moreover, thanks also go to Dr. Jani Even, who is currently a Post-Doctorate Fellow in Nara Institute of Science and Technology, for his fruitful discussion and helpful advice on blind source extraction and statistical signal processing.

The later part of this work also could not have been achieved without collaborations of many researchers. I especially thank Mr. Yoshihisa Uemura, who is currently a researcher in Softbank Corporation, for his fruitful discussion on HOS-based musical noise analysis. I also wish to express my acknowledgement to Mr. Hideki Kenmochi, Mr. Kazunobu Kondo and Mr. Makoto Yamada, who are researchers in Yamaha Corporation, for their beneficial and valuable comments on musical noise analysis.

A lot of staff and my research group members had supported me to carry out experiments and write this dissertation at Nara Institute of Science and Technology; I especially would like to express my appreciation to Dr. Shigeki Miyabe, who is currently a Post-Doctorate Fellow in the University of Tokyo, and Mr. Yoshimitsu Mori for their valuable discussion on technical issues of speech signal processing and digital signal processing, and their arrangement of the comfortable computer environment in our laboratory. I also wish to represent my deeply gratitude to Mrs. Toshie Nobori, who is the secretary in our laboratory, for her kind help and support in all aspects of my research.

I appreciate to study together with all of students who had been in our research group at Nara Institute of Science and Technology. I thank Mr. Keigo Nakamura, Mr. Yamato Ohtani, Mr. Shota Takeuchi, and Mr. Noriyoshi Kamado, who are Ph.D. candidates in Nara Institute of Science and Technology, for their useful discussions on this work.

Finally, I am deeply grateful to all members of my family for their support they gave me during all these years.

List of Publications

Journal Papers

1. Keiichi Osako, Yoshimitsu Mori, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, “Fast Convergence Blind Source Separation Using Frequency Subband Interpolation by Null Beamforming,” *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol.E91-A, no.6, pp.1329–1336, 2008.

2. Kentaro Tachibana, Yu Takahashi, Yoshimitsu Mori, Hiroshi Saruwatari,Kiyohiro Shikano, Akira Tanaka, “Fast Nonclosed-Form Higher-Order ICA with Frequency-Subband Selection Using Closed-Form Second-Order ICA,” *Journal of Signal Processing*, vol.12, no.4, pp.327–330, 2008.

3. Yu Takahashi, Tomoya Takatani, Keiichi Osako, Hiroshi Saruwatari, Kiyohiro 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, May. 2009.

4. Yuki Fujihara, Yu Takahashi, Kentaro Tachibana, Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, Akira Tanaka, “Real-time blind source extraction with learning period detection based on closed-form second ICA and kurtosis,” *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences (Japanese edition)*, vol. J92-A, no. 5, pp.314–326, May 2009.

5. Takashi Hiekata, Youhei Ikeda, Hiroshi Hashimoto, Ruoyu Zhang, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, “Development of real-time blind source extraction microphone based on noise estimation by parallel ICAs,” *IEICE Transcations on Information and Systems (Japanese edition)*, vol. J92-D, no. 10, pp. 1772–1783, 2009. (in printing)

6. Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, “Development of hands-free speech recognition system based on spatial sbutraction array with independent component analysis,” *IEICE Transcations on Information and Systems (Japanese edition)*, vol.J93-D, no.3,pp.–,Mar. 2010. (in printing).

Peer Reviewed Proceedings

1. Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, "Blind spatial subtraction array with independent component analysis for hands-free speech recognition," *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, September 2006.

2. Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, "Blind spatial subtraction array based on independent component analysis for speech enhancement and recognition," *The Journal of the Acoustic Society of America*, Vol.120, No.5, Pt. 2 of 2 (4th Joint Meeting), pp.3047–3048, November 2006.

3. Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, "Robust Spatial Subtraction Array with Independent Component Analysis for Speech Enhancement," *International Symposium on Signal Processing and its Applications (ISSPA)*, Digital Object Identifier: 10.1109/ISSPA.2007.4555589, February 2007.

4. Naoya Tanaka, Yu Takahashi, Yoshimitsu Mori, Hiroshi Saruwatari, Kiyohiro Shikano, "Internal robot noise reduction by using NAM microphone for hands-free speech recognition," *RISP International Workshop on Nonlinear Circuits and Signal Processing*, pp.473–476, March 2007.

5. Ayase Takagi, Yoshimitsu Mori, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, "Improvement of Acoustic Model for Hands-Free Speech Recognition Using Spatial Subtraction Array," *RISP International Workshop on Nonlinear Circuits and Signal Processing*, pp.579–582, March 2007.

6. Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, "Permutation-robust Structure for ICA-based Blind Source Extraction," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2007.

7. Yoshimitsu Mori, Keiichi Osako, Shigeki Miyabe, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, "MLSP 2007 Data Analysis Competition: Two-Stage Blind Source Separation Combining SIMO-Model-Based ICA and Binary Masking," *2007 IEEE International workshops on Machine Learning for Signal Processing (MLSP 2007)*, August 2007. (invited talk)

8. Keiichi Osako, Yoshimitsu Mori, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, “Fast Convergence Blind Source Separation Based on Frequency Subband Interpolation by Null Beamforming,” *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2007)*, pp.42–45, Oct. 2007.

9. Kentaro Tachibana, Yu Takahashi, Yoshimitsu Mori, Hiroshi Saruwatari, Kiyohiro Shikano, Akira Tanaka, “Fast Nonclosed-Form Higher-Order ICA with Frequency-subband Selection Using Closed-Form Second-Order ICA, ” *2008 RISP International Workshop on Nonlinear Circuits and Signal Processing (NCSP 2008)*, pp.188–191, March 2008.

10. Yu Takahashi, Keiichi Osako, Hiroshi Saruwatari and Kiyohiro Shikano, “Blind Source Extraction For Hands-Free Speech Recognition based on Wiener Filtering and ICA-based Noise Estimation,” *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSMCA2008)*, pp. 164–167, May 2008.

11. Yu Takahashi, Hiroshi Saruwatari Kiyohiro Shikano, “Real-Time Implementation of Blind Spatial Subtraction Array For Hands-Free Robot Spoken Dialogue System,” *IEEE/RSJ International Conference on Intelligent Robotics and Systems (IROS2008)*, pp.1687–pp.1692, September 2008.

12. Yoshihisa Uemura, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, “Automatic Optimization Scheme of Spectral Subtraction based on musical noise assessment via higher-order statistics,” *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, September 2008.

13. Yuuki Fujihara, Yu Takahashi, Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, Akira Tanaka “Performance improvement of higher-order ICA using learning period detection based on closed-form second-order ICA and kurtosis,” *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, September 2008.

14. Hiroshi Saruwatari, Yu Takahashi, Hiroyuki Sakai, Shota Takeuchi, Tobias Cincarek, Hiromichi Kawanami, Kiyohiro Shikano, “Development and Evaluation of Hands-Free Spoken Dialogue System for Railway Station Guidance,” *INTERSPEECH2008*, pp.455–pp.458, September 2008.

15. Keisuke Masatoki, Shigeki Miyabe, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, Toshiyuki Nomura, “Timing Optimization of Filter Replacement in Com-

pressive Coding for Stereo Audio Signals Using Independent Component Analysis," *International Conference on Signal Processing (ICSP'08)*, Beijing, China, pp. 510–513, October 2008.

16. Yu Takahashi, Hiroshi Saruwatari, Yuki Fujihara, Kentaro Tachibana, Yoshimitsu Mori, Shigeki Miyabe, Kiyohiro Shikano, Akira Tanaka, "Source adaptive blind signal extraction using closed-form ICA for hands-free robot spoken dialogue system," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP2009)*, pp.3361–3364, April 2009 (invited).

17. Yu Takahashi, Yoshihisa Uemura, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Musical noise analysis based on higher order statistics for microphone array and nonlinear signal processing," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP2009)*, pp.229–232, April 2009.

18. Hiroshi Saruwatari, Hiromichi Kawanami, Shota Takeuchi, Yu Takahashi, Tobias Cincarek, Kiyohiro Shikano "Hands-free speech recognition challenge for real-world speech dialogue systems," International Conference on Acoustics, Speech, and Signal Processing (ICASSP2009), pp.3729–3782, April 2009.

19. Yoshihisa Uemura, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, and Kazunobu Kondo "Musical noise generation analysis for noise reduction methods based on spectral subtraction and MMSE STSA estimation," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP2009)*, pp.4433–4436, April 2009.

20. Takashi Hiekata, Takashi Morita, Youhei Ikeda, Hiroshi Hashimoto, Ruoyu Zhan, Yu Takahashi, Hiroshi Saruwatari, and Kiyohiro Shikano, "Multiple ICA-based real-time blind source extraction applied to handy size microphone," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP2009)*, pp.121–124, April 2009.

21. Yu Takahashi, Yoshihisa Uemura, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, 'Structure selection algorithm for less musical-noise generation in integration systems of beamforming and spectral subtraction," *2009 International Workshop on Statistical Signal Processing (SSP2009)*, pp.701–704, September 2009.

22. Hiroshi Saruwatari, Yu Takahashi, Kentaro Tachibana, Akira Tanaka, “Fast and Versatile Blind Separation of Diverse Sounds Using Closed-Form Estimation of Probability Density Functions of Sources,” *The Third International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP09)*, pp.249–252, December 2009.

23. Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, and Kazunobu Kondo, ‘Theoretical musical-noise analysis and its generalization for methods of integrating beamforming and spectral subtraction based on higher-order statistics.” *International Conference on Acoustics, Speech, and Signal Processing (ICASSP2010)*, pp. 93–96, 2010.

24. Ryoi Okamoto, Yu Takahashi, Hiroshi Saruwatari, and Kiyohiro Shikano “MMSE STSA estimator with nonstationary noise estimation based on ICA for high-quality speech enhancement,” *International Conference on Acoustics, Speech, and Signal Processing (ICASSP2010)*, pp. 4778–4781, 2010.

Technical Reports

1. Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, “Noise reduction using spatial subtraction array based on independent component analysis,” *IEICE Technical Report EA 2006-22*, vol. 166, no.125, pp. 13–18, June 2006.

2. Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, “Improvement of Accuracy of Noise Estimation Based on Independent Component Analysis in Spatial Subtraction Array,” *Proceedings of the 24th Meeting of Special Interest Group on AI Challenges*, pp.17–22, November 2006.

3. Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, “Blind source extraction method with permutation-robust structure,” *IEICE Technical Report EA 2006-95*, vol.166, no.125, pp.37–42, December 2006.

4. Keiichi Osako, Yu Takahashi, Yoshimitsu Mori, Hiroshi Saruwatari, Kiyorhiro Shikano, “Fast Algorithm for Cancellation of Near Point Source Based on Independent Component Analysis,” *IEICE Technical Report EA2007-57*, vol. 107, no. 240, pp. 13–18, September 2007. (in Japanese)

5. Yu Takahashi, Keiichi Osako, Hiroshi Saruwatari, Kiyohiro Shikano, “Development of Hands-Free Spoken Dialogue System with Real-Time Blind Spatial Subtraction Array,” *IEICE Technical Report EA2008-16*, vol. 108, no.68 , pp.59–64, May 2008.

6. Yuki Fujihara, Yu Takahashi, Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, Akira Tanaka, “Acceleration of Higher-Order ICA Using Lerning Period Detection Based on Closed-Form Second-Order ICA and Kurtosis,” *IEICE Technical Report EA2008-15*, vol.108, no. 68, pp. 53–58, May 2008. (in Japanese)

7. Yoshihisa Uemura, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kando, “Relationship between logarithmic kurtosis ratio and degree of musical noise generation on spectral subtraction,” *IEICE Technical Report EA2008-44, vol. 108, no.143*, pp.43–48 , July 2008.

8. Yu Takahashi, Yoshihisa Uemura, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, “Objective sound quality evaluation for combination method of beamforming and spectral subtraction,” *IEICE Technical Report EA2008-128*, vol. 108, no. 411, pp. 73–78, January 2009.

9. Yoshihisa Uemura, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, “Objective sound quality comparison based on higher-order statistics for nonlinear noise reduction methods,” *IEICE Technical Report EA2008-127*, vol. 108, no. 411, pp. 67–72, January 2009.

10. Takashi Hiekata, Youhei Ikeda, Hiroshi Hashimoto, Takashi Morita, Ruoyu Zhang, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, “Development of blind signal extraction device based on noise estimation by parallel ICAs,” *IEICE Technical Report EA2008-135*, vol. 108, no. 411, pp. 115–120, January 2009.

11. Yohei Ishikawa, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, “Musical-noise regulation for combination method of channel-domain nonlinear speech enhancement and adaptive array signal processing,” *IEICE Technical Report EA2009-3*, vol. 109, no. 55 , pp.11–16 , May 2009. (in Japanese)

12. Ryoi Okamoto, Yu Takahashi, Hiroshi Saruwatari,Kiyohiro Shikano “Blind signal extraction based on minimum mean-square error short-time spectral amplitude es-

timator," *IEICE Technical Report EA2009-39*, vol.109, no. 136, pp.13–18, July 2009. (in Japanese)

Domestic Meetings

1. Yu Takahashi, Tomoya Takatani, Chie Kiuchi, Hiroshi Saruwatari, Kiyohiro Shikano, "Noise reduction using spatial subtraction array with independent component analysis," *The Meeting of ASJ*, 3-5-8, pp.619–pp.620, March 2006 (in Japanese).

2. Yu Takahashi, Tomoya Takatani, Naoya Tanaka, Hiroshi Saruwatari, Kiyohiro Shikano, "Hands-free speech recognition using blind spatial subtraction array," *The Meeting of ASJ*, 3-Q-11, pp.619–pp.620, September 2006 (in Japanese).

3. Ayase Takagi, Yoshimitsu Mori, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, "A Study on Acoustic Model Reflecting Actual Environment for Hands-free Speech Recognition Using Spatial Subtraction Array," *The Meeting of ASJ*, 1-2-14, pp.27–28, September 2006 (in Japanese).

4. Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, "Blind spatial subtraction array providing accurate noise estimator based on independent component analysis," *Kansai-section Joint Convention of Institute of Electrical Engineering*, G16-17, pp.G380, November 2006 (in Japanese).

5. Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, "Blind spatial subtraction array with independent component analysis for hands-free speech recognition," *The young researchers meetings of ASJ Kansai-section*, 24-B, December 2006 (in Japanese).

6. Ayase Takagi, Yoshimitsu Mori, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, "A Study of Acoustic Model for Spatial Subtraction Array in Actual Environment," *The young researchers meetings of ASJ Kansai-section*, 6-B, December 2006 (in Japanese).

7. Ayase Takagi, Yoshimitsu Mori, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, "A Study of Acoustic Model for Spatial Subtraction Array," *Kansai-section Joint Convention of Institute of Electrical Engineering*, G16-16, pp.G379, November 2006 (in Japanese).

8. Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, “Analysis of permutation robustness in blind spatial subtraction array,” *The Meeting of ASJ*, 3-P-9, March 2007 (in Japanese).

9. Ayase Takagi, Yoshimitsu Mori, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, “Actual Environment Evaluation of Hands-Free Speech Recognition System Based on Spatial Subtraction Array,” *The Meeting of ASJ*, 3-9-18, March 2007 (in Japanese).

10. Naoya Tanaka, Yu Takahashi, Yoshimitsu Mori, Hiroshi Saruwatari, Kiyohiro Shikano, “Internal Moving-Robot-Noise Reduction by Using NAM Microphone for Spatial Subtraction Array,” *The Meeting of ASJ*, 2-1-16, March 2007 (in Japanese).

11. Yu Takahashi, Keiichi Osako, Hiroshi Saruwatari, Kiyohiro Shikano, “Speech Recognition in Railway-Station Environment by Using Blind Spatial Subtraction Array,” *The Meeting of ASJ*, 1-7-18, September 2007 (in Japanese).

12. Keiichi Osako, Yu Takahashi, Yoshimitsu Mori, Hiroshi Saruwatari, Kiyohiro Shikano, “A Study on Near-Point-Source Cancellation Based on Independent Component Analysis,” *The Meetings of ASJ*, 1-7-16, September 2007 (in Japanese).

13. Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, “Blind source extraction method combining ICA-based noise estimation and Wiener filterng,” *Kansai-seciont Joint Convention of Institute of Electrical Engineering*, G15-4, pp. G345, November 2007.

14. Keiichi Osako, Jani Even, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, “A permutation solving method on independent component analysis for near-point-source reduction,” *Kansai-seciont Joint Convention of Institute of Electrical Engineering*, G15-8, pp. G349, November 2007 (in Japanese).

15. Kentaro Tachibana, Yu Takahashi, Jani Even, Yoshimitsu Mori, Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, Akira Tanaka, “Fast nonclosed-form higher-order ICA with freqency-subband selection and probability density function estimation based on closed-form second-order ICA,” *24th SIP Symposium*, P1-1, pp.560–565, November 2007 (in Japanese).

16. Yu Takahashi, Shigeki Miyabe, Keiichi Osako, Cincarek Tobias, Shota Takeuchi, Hiroyuki Sakai, Hiromichi Kawanami, Hiroshi Saruwatari, Kiyohiro Shikano, “De-

velopment of Hands-Free Spoken Dialogue System with Real-Time Blind Spatial Subtraction Array," *The Meetings of ASJ*, 2-6-17, March 2008 (in Japanese).

17. Hiroshi Saruwatari, Makoto Shozakai, Katsumasa Nagahama, Masashi Yamada, Takanobu Nishiura, Yuuki Denda, Yu Takahashi, Kiyohiro Shikano "Development of hands-free speech recognition system based on spatial subtraction array," *The Meetings of IPSJ*, 4L-2, pp.5-351–5-352, March 2008 (in Japanese).

18. Hiroshi Saruwatari, Yu Takahashi, Cincarek Tobias, Hiroyuki Sakai, Shota Takeuchi, Keiichi Osako, Shigeki Miyabe, Yoshimitsu Mori, Hiromichi Kawanami, Lee Akinobu, Kiyohiro Shikano, "Development of hands-free robot spoken dialogue system," *The Meetings of IPSJ*, 4L-3, pp.5-353–5-354, March 2008 (in Japanese).

19. Keiichi Osako, Yu Takahashi, Yoshimitsu Mori, Jani Even, Hiroshi Saruwatari, Kiyohiro Shikano "Blind Spatial Subtraction Array with Fast Near-Point-Source Cancellation Algorithm," *The Meetings of ASJ*, 2-6-16, March 2008 (in Japanese).

20. Yuuki Fujihara, Yu Takahashi, Shigeki Miyabey, Hiroshi Saruwatari, Kiyohiro Shikano, Akira Tanaka "Voice Activity Ditection Using Closed-Form Second-Order ICA and Subband Kurtosis for Lerning of Nonclosed-Form Higher-Order ICA," *The Meetings of ASJ*, 2-6-12, March 2008 (in Japanese).

21. Yu Takahashi, Yoshihisa Uemura, Hiroshi Saruwatari, Kiyohiro Shikano "Musical noise reduction with channel-wise spectral subtraction in microphone-array post processing," *The Meetings of ASJ*, 2-8-17, pp.671–674, September 2008 (in Japanese).

22. Keisuke Masatoki, Shigeki Miyabe, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, Toshiyuki Nomura, "Compressive coding of stereo audio signal using cosine-distance weighted k-means," *The Meetings of ASJ*, 1-8-1, pp.581–584, September 2008 (in Japanese).

23. Yoshihisa Uemura, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, "Automatic Optimization Scheme in Spectral Subtraction with Higher-Order Statistics-Based Musical Noise Assessment," *The Meetings of ASJ*, 3-8-4, pp.691–694, September 2008 (in Japanese).

24. Yuuki Fujihara, Yu Takahashi, Shigeki Miyabey, Hiroshi Saruwatari, Kiyohiro Shikano, Akira Tanaka “Selection of correlation matrix for learning point detection of source separation filter using closed-form ICA and kurtosis,” *The Meetings of ASJ*, 3-8-6, pp.697–698, September 2008 (in Japanese).

25. Yoshihisa Uemura, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, “Short-time automatic parameter optimization of spectral subtraction based on musical noise metric via higher-order statistics,” *23rd SIP Symposium*, pp.241–246, November 2008 (in Japanese).

26. Yu Takahashi, Yoshihisa Uemura, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, “Higher-order statistics-based analysis on integration method of beamforming and spectral subtraction,” *The Meetings of ASJ*, 3-9-9, pp.727–730, March 2009 (in Japanese).

27. Shota Suzuki, Keisuke Masatoki, Shigeki Miyabe, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, Toshiyuki Nomura, “Sound-localization control method based on spatial representation vector transformation for wide area sound reproduction,” *The Meetings of ASJ* 1-9-17, pp.1497–1500, March 2009 (in Japanese).

28. Ryoi Okamoto, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano “MMSE STSA with Noise Estimation Based on Independent Component Analysis,” *The Meetings of ASJ* 2-9-6, pp.667–670, March 2009 (in Japanese).

29. Yuki Fujihara, Yu Takahashi, Shigeki Miyabe, Hiroshi Saruwatari, Kiyohiro Shikano, Akira Tanaka, “Theoretical study on selection of correlation matrix for learning period detection of source separation filter using closed-form ICA and kurtosis,” 2-9-7, pp.671–674, March 2009 (in Japanese).

30. Takashi Hiekata, Yohei Ikeda, Hiroshi Hashimoto, Takashi Morita, Ruoyu Zhang, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, “Development of blind source extraction microphone based on noise estimation using multiple ICAs,” *The Meetings of ASJ* 3-P-29, pp.819–820, March 2009 (in Japanese).

31. Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, “Automatic structure selection method in integration of beamforming and spectral subtraction based on metric for the amount of musical-noise generation,,, *The Meetings of ASJ*, 2-4-12, pp.635–638, September 2009 (in Japanese).

32. Yohei Ishikawa, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, “Subjective Evaluation of Musical Noise Controlable Array Signal Processing,” *The Meetings of ASJ*, 2-4-13, pp.639–642, September 2009 (in Japanese).

33. Ryoi Okamoto, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano “Subjective Evaluation of MMSE STSA Estimator with Noise Estimation Based on Independent Component Analysis,” *The Meetings of ASJ*, 2-4-19, pp.659–662, September 2009 (in Japanese).

34. Makoto Yamada, Yu Takahashi, Kazunobu Kondo, Hiroshi Saruwatari, “Bootstrap Aggregating Spectral Subtraction for Musical Noise Reduction,” *The Meetings of ASJ*, 2-P-26, pp.851–854, March 2010 (in Japanese).

35. Ryoi Okamoto, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, “Blind Adaptation of Target Signals Statistical Model in MMSE STSA Estimator with Independent Component Analysis,” *The Meetings of ASJ*, 2-5-11, pp.745–748, March 2010 (in Japanese).

36. Takayuki Inoue, Yu Takahashi, Yohei Ishikawa, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, “Mathematical Analysis of Musical Noise for Generalized Spectral Subtraction Method,” *The Meetings of ASJ*, 3-5-4, pp.759–762, March 2010 (in Japanese).

37. Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, “Generalization of theoretical musical-noise analysis in methods of integrating beamforming and spectral subtraction based on higher-order statistics,” *The Meetings of ASJ*, 3-5-5, pp.763–766, March 2010 (in Japanese)

38. Yohei Ishikawa, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, Kazunobu Kondo, “Evaluation of Musical Noise Controllable Array Signal Processing in Real Environment,” *The Meetings of ASJ*, 3-5-6, pp.767–768, March 2010 (in Japanese).

Awards

1. An Encouraging Prize on the young researchers meetings of the Acoustical Society of Japan Kansai section, Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, December 2006.

2. 2006 Technical Report Award on the Japanese Society for Artificial Intelligence, Yu Takahashi, Tomoya Takatani, Hiroshi Saruwatari, Kiyohiro Shikano, July 2007.

3. 2007 IEEE workshops on Machine Learning for Signal Processing (MLSP2007) Data Analysis Competition Winner on Nonlinear Separation, Yoshimitsu Mori, Shigeki Miyabe, Keiichi Osako, Yu Takahashi, Hiroshi Saruwatari, Kiyohiro Shikano, 2007.

4. Student Paper Award on 2008 RISP International Workshop on Nonlinear Circuit and Signal Processing (NCSP08), Kentaro Tachibana, Yu Takahashi, Yoshimitsu Mori, Hiroshi Saruwatari, Kiyohiro Shikano, Akira Tanaka, 2008.

5. 2009 IEEE Signal Processing Society Japan Chapter Student Paper Award, Yu Takahashi, Tomoya Takatani, Keiichi Osako, Hiroshi Saruwatari, Kiyohiro Shikano, November 2009.

Open Source Software

1. Open ICA
   An implementation of blind source separation for acoustic sound sources based on frequency-domain independent component analysis
   http://openica.sourceforge.jp/