NAIST-IS-DD1361014

Doctoral Dissertation

Microphone Array Processing based on Blind Source Separation for Robust Distant Speech Recognition System

Fine Dwinita Aprilyanti

March 14, 2016

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

Fine Dwinita Aprilyanti

Thesis Committee:

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

Microphone Array Processing based on Blind Source Separation for Robust Distant Speech Recognition System[*]

Fine Dwinita Aprilyanti

Abstract

Distant speech recognition system, in which a single or an array of microphone is utilized to capture the user's utterance as opposed to the user-attached microphone, is greatly affected by the presence of background noise and reverberation. Many researches have been conducted on developing an array processing method to improve the quality of the captured speech signals. Most of these methods are optimized to obtain the clean target speech, as measured by signal-to noise ratio or human perception. However, speech recognition system works as a statistical pattern classifier of features extracted from the speech waveform. Therefore, array front-end processing can only be expected to increase the recognition accuracy if it maximizes the likelihood of the correct hypothesis.

In this study, I propose array processing techniques based on blind source separation to suppress the background noise and late reverberation, optimized to maximize the likelihood to the acoustic model of speech recognizer. The first method utilizes frequency-domain blind signal extraction (BSE), which is an alternative to the conventional blind source separation specifically designed for the case of speech in the presence of diffuse noise, combined with two stages of multichannel Wiener filter. I extend this method by integrating information from the image sensor to achieve optimum performance regardless the interference level. In the second method, I combine BSE with multichannel generalized minimum mean-square error estimator of short time spectral amplitude (MMSE-STSA),

[*] Doctoral Dissertation, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1361014, March 14, 2016.

which can provide less distortion to the output signal owing to the use of speech spectral amplitude statistical model assumption and decision-directed signal-to-noise ratio estimation approach.

These methods, however, are able to perform optimally only if the background noise is diffused, due to the characteristics of BSE. The performance of BSE degrades in the presence of point-source noise, thus the first and second method are sub-optimal for such case. To further improve the capability of array front-end processing, in the third method, I develop a source-adaptive blind source separation, in which the activation function is parameterized according to the estimated statistical model of each source signal. Parameter for each activation function is derived from the parameter of the generalized MMSE-STSA postprocessing, which is optimized based on the acoustic model. Experimental evaluation shows that this proposed method is more robust to different types of interference than former array processing approaches.

Keywords:

Blind source separation, generalized MMSE-STSA, dereverberation, noise suppression

Contents

List of Figures

List of Tables

1. Introduction

1.1 Automatic Speech Recognition System in Distant-Talking Environments

The use of speech-based human-machine interface provides us the simple and natural way of communication with computers. Over the past decades, research in this field has progressed significantly such that the state-of-the-art automatic speech recognition (ASR) systems are able to achieve high recognition accuracy in a noise-free environment. In these systems, the speaker talks using a microphone that is put closely to mouth, so that the quality of the captured speech signal is quite high. However, there are many conditions where the implementation of this setting becomes difficult or even impossible for either safety or convenience reasons. For example, the act of wearing a microphone while operating a vehicle is distracting and may lead to danger. In a meeting room or during conferences' poster session, hand-held or head-mounted microphone is inconvenient because it restricts the movement of the participants.

This problem can be solved by placing a single or an array containing multiple microphones to captured speech signals at a certain distance. However, another problem arises, i.e, distortions from interferences that corrupt the target speech quality, as illustrated in Fig. 1. In general, these distortions can be distinguished into three main types [1]:

- Noise, or also known as background noise, which is any sound other than the target speech, e.g, sound from air conditioners, machines in a factory, or speech from other speakers.

- Reveberation, which is the reflection of the sound source that arrives some time after the direct sound. The severity of this distortion varies according to the distance between speaker and microphone, the geometry of the room and the material of the surface of the room.

- Other types of distortions that are introduced by environmental factors, e.g., room modes, the orientation of the speaker's head, or the Lombard effect.

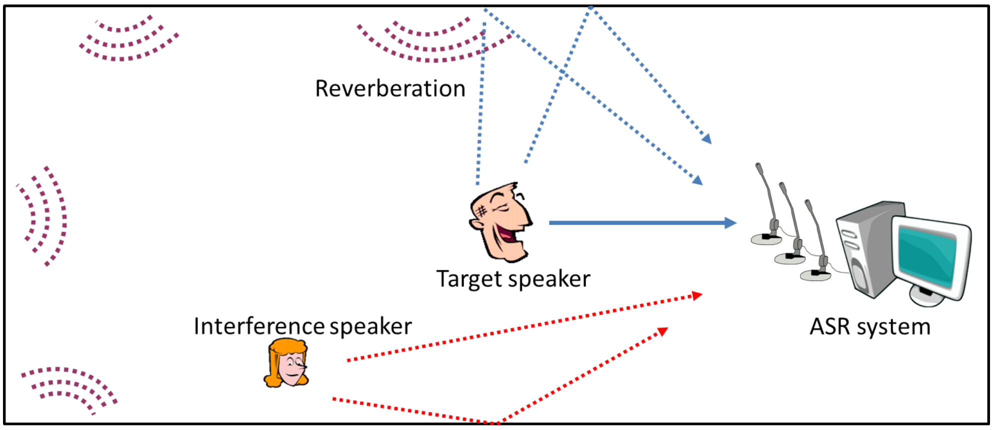

Figure 1. Illustration of the target speech and interference sounds captured at the microphone array in a distant speech recognition system.

Each type of distortion affects the captured speech signal in different manner. This dissertation is focused on the first two types of distortion.

1.1.1 Effect of Background Noise to Speech Recognition

Background noise is any additive sound other than the target speech that is captured by the microphone or microphone array. The term background noise covers a broad variety of additive sound, which can be classified as:

- *Stationarity noises*, which have statistics that relatively constant over long time spans. Some example sound from computer fans or air conditioning.
- *Nonstationary noises*, which have statistics that change significantly over short periods, such as music and people voice. This type of noise is usually produced by point source, unlike the stationary noise which is usually diffused and widespread.

Figure 2 illustrates the effect of the presence of noise on the quality of speech. It can be seen the noise components fill in the segment where the speech energy is low and mask the original clean speech. It will be difficult for ASR system to distinguish the speech pattern under such condition. Even though the recognition performance may be improved by using the noisy data for acoustic model training,

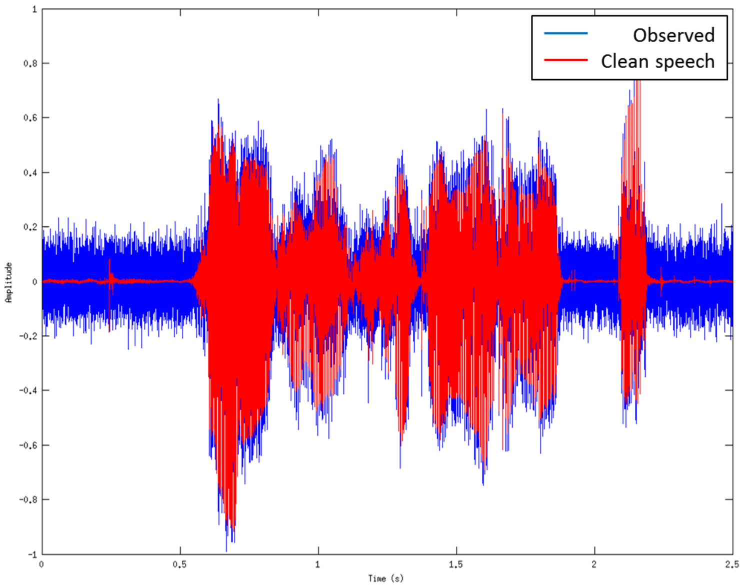

Figure 2. Illustration of speech distorted by background noise.

this option is not realistic because the noise conditions in training and testing process are seldom matched, particularly for nonstationary noise. Therefore, it is preferable to apply noise suppression algorithm as a front-end to the ASR system.

1.1.2 Effect of Reverberation to Speech Recognition

Reveberation occurs when the reflected sounds from the surfaces in an enclosure come simultaneously following the direct signal. In spectral domain, reverberation causes smearing of the original speech spectrum. For some cases, for example in a musical performance, the presence of reverberation is favored. However, reverberation also distorts both the envelope and fine structure of a speech signal. The reverberant speech becomes more difficult to be recognized by the ASR system [2].

Reverberation in one room depends on the room impulse response. An example of room impulse response is shown in Fig. 3. As shown in the figure, a room impulse response can be separated into three parts: direct-path response, early and late reverberations [3]. The direct-path response represents how the sound is received on the microphone without any reflection. The delay between the initial

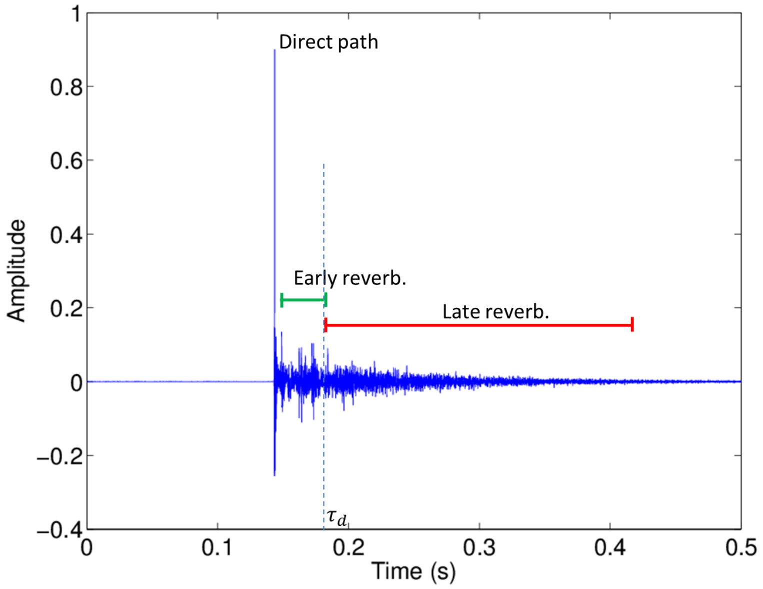

Figure 3. A sample of room impulse response.

excitation and its observation depends on the distance of the sound source from the speaker and also the velocity of the sound. Early reverberation arrives at the microphone a little later than the direct sounds. It is usually not perceived as a separate sound to the direct sound as long as the reflection delay does not exceed a certain limit which is often called *cutoff delay*. The cutoff delay τ_d is approximated between the range of 50-100 ms depending on the sound source. Due to its frequency response, early reverberation often causes the coloration of the original speech spectrum. However, early reverberation also tend to reinforce the direct sound owing to the precedence effect and therefore is considered useful to increase the speech intelligibility. The state-of-the-art ASR system also can handle the presence of early reverberation, for example by applying cepstral mean normalization.

Late reverberation is any component of reflected sound that arrives at the microphone after the cutoff delay. Because it is a combination of many reflected sounds, late reverberation loses its correlation to the direct sound. It is modeled in the time domain as an energy tail that decays according to the room reverberation time T_{60}. The late reverberation is harmful to the speech intelligibility as illustrated in Fig. 4. The strong reverberation smears the original speech

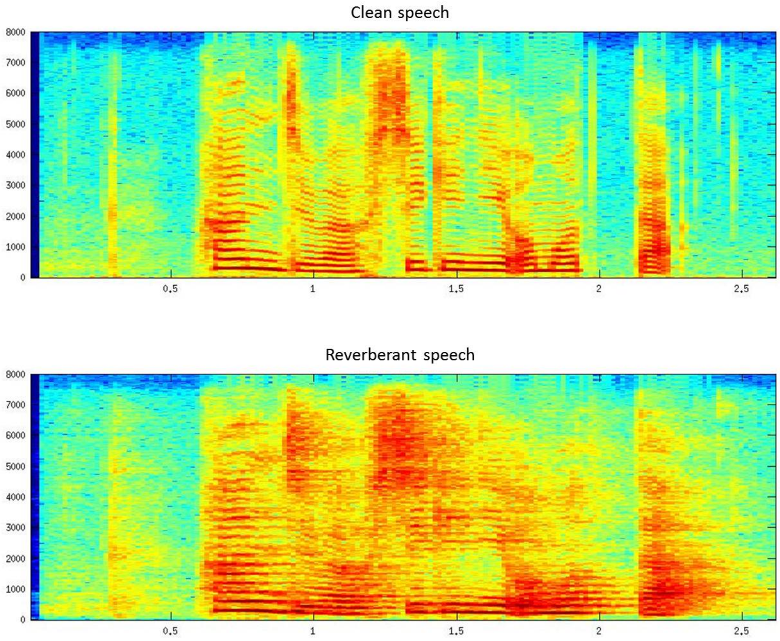

Figure 4. Distorted speech spectrum due to revereberation.

spectrum long enough that it causes the overlap-masking between the adjacent phoneme in an utterance. Therefore, researches on dereverberation algorithm put more focus in suppressing the late reverberation components.

1.2 Research Scope and Motivation

A vast amount of array processing techniques have been developed to reduce the effect of these distortions on speech. Many of these techniques only focus on one type of interference. For example, noise suppression algorithms are simulated under no reverberation condition or evaluated in rooms with short T_{60}, as in [4, 5]. On the other hands, the dereverberation algorithms usually developed under noise-free assumption, such as in [6, 7]. These approaches are unsuitable for the real world implementation.

Recently, more algorithms have been built to jointly suppress background noise and late reverberation, e.g., [8, 9, 10]. Often these algorithms are optimized to enhance the quality of the input speech waveform. The performance are evaluated objectively by the improvement of signal-to-noise ratio (SNR) or the reduced cepstral distortion, and subjectively by human listening test. This also applies to the ASR system array preprocessing, under assumption that the enhanced input speech waveform will lead to better accuracy in speech recognition. However, the ASR systems is basically a statistical pattern classifier that works on a sequence of features extracted from the speech waveform, not directly on the waveform itself. Therefore, the array preprocessing methods is expected to improve the recognition accuracy only if it can generate such feature sequence which increases the likelihood of the correct transcription, relative to other.

According to this principle, in this study I propose array prepocessing methods based on blind source separation (BSS) combined with nonlinear postprocessing, that is integrated with ASR system, as depicted in Fig. 5. Information from the acoustic model in ASR is used to tune the parameter of array processing method. This architecture will enable the output of the preprocessing methods to have the optimized likelihood to the best hypothesis, thus is expected to improve the recognition accuracy. The approach has been implemented for the beamforming-based array processing method in [11] and also spectral subtraction method [12].

The use of BSS-based approach in this study, including BSE as an alternated BSS, is based on several potential of this method. First, BSS and BSE requires almost no *a priori* knowledge, hence the term *blind.* Second, although both conventional BSS and BSE can only perform well on certain type of interference, we can develop new method that emphasize on their advantages, for example by combining with nonlinear postprocessing and modifying the activation function to be source-adaptive, as described in the third proposed method in this dissertation. Third, by incorporating the statistical model of the recognizer, i.e. acoustic model, into the array processing optimization scheme, we can focus on enhancing the signal components important for recognition accuracy, without putting too much emphasis on less important components.

All the proposed methods in this study requires the T_{60} information in the dereverberation stage. In real world implementation, the estimation of T_{60} can be done from the geometry of the room or by the impulse response test. Assuming the distant speech recognition system will be fixed in one place, the estimation process is only required once. For systems that can move to different room, such as robot implementation, we can integrate the information from other sensors,

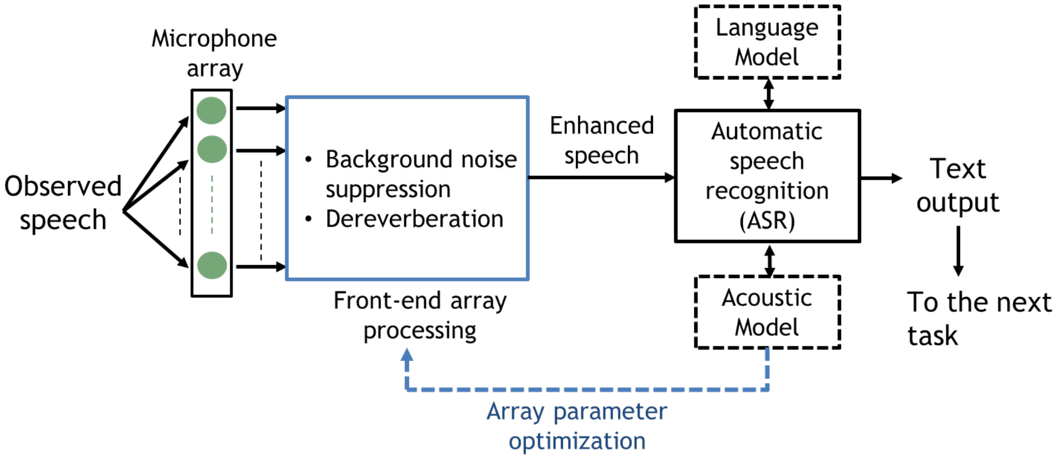
Microphone array
Observed speech
• Background noise suppression
• Dereverberation
Front-end array processing
Enhanced speech
Language Model
Automatic speech recognition (ASR)
Acoustic Model
Text output
To the next task
Array parameter optimization

Figure 5. The general architecture of array preprocessing optimized for the speech recognition performance, in which the information from statistical model of ASR is utilized to tune the array processing parameter.

such as image sensor, as described in the first proposed method. Furthermore, I also investigate the significance of acquiring the correct value of T_{60} for the performance of the source-adaptive proposed method as not to contradict the blind characteristics of the method.

1.3 Overview of Dissertation

The rest of dissertation is organized as follows.

First, I present related works on microphone array processing in Sect. 2. In this section, I will also explain about the data used for the experimental evaluation. The first proposed method combining FD-BSE with multichannel WF is described in Sect. 3. This semi-blind method assumes that the target user's distance information is accessible from the image sensor. The next method proposed in Sect. 4 works under no such assumption. In this section, FD-BSE is applied for noise suppression and the generalized MMSE-STSA is applied as postprocessing for dereverberation. These proposed method are developed only for the case of diffuse background noise, in which BSE excels in target speech enhancement and computational complexity.

Next, to achieve good performance not only in diffuse noise case but also in point-source noise case, I propose a novel source-adaptive BSS method in Sect. 5. The proposed method use the more appropriate parameterized activation function instead of fixed activation function in conventional BSS. The activation fuctions in this method correspond to the probability density function (PDF) of each source, which are modeled in generalized Gaussian distribution (GGD). I derive an approximated closed-form solution based on higher-order-statistics approach to estimate the target speech PDF from the parameter of the optimized generalized MMSE-STSA postprocessing. I also introduce the strategy to estimate the interference source's PDF using the optimized shape parameter of the target speech. Experiment is carried out using different type of interference to evaluate the performance of the proposed method.

As the parameter of the generalized MMSE-STSA is optimized according to the acoustic model, the resulting proposed method integrates the statistical modeling of speech in three different domains, i.e. complex spectral domain (as in source-adaptive BSS), amplitude domain (as in generalized MMSE-STSA), and mel-frequency cepstral domain (as in the acoustic model of ASR).

2. Review on Microphone Array Processing Method for Speech Recognition

2.1 Introduction

As pointed out in the previous section, the quality of speech signals captured by a microphone in a distance decreases significantly due to the distortion from the background noise and reverberation. It is also understandable that the use of multiple microphone can recover the target speech signal better than using single microphone owing to the spatial information it provides. In this section, first, the model of sound signal captured by a microphone array will be described. Next, two groups of well-known array processing method will be explained, i.e. BSS based method and spectral-based method. In the end of this section, I will also explain about the speech corpus and room impulse response data used in experimental evaluation.

2.2 Sound Mixture Model at Microphone Array

Signals captured by an M-channel microphone array are composed of a clean speech signal and interference sounds. This interference can be additive, that is background noise, and also convolutive, that is room reverberation. Mathematically, a mixture of signals at a microphone array can be modeled by

$$\boldsymbol{x}(t) = \boldsymbol{x}_{\mathrm{S}}(t) + \boldsymbol{x}_{\mathrm{N}}(t), \tag{1}$$

$$\begin{aligned} \boldsymbol{x}_{\mathrm{S}}(t) &= (\boldsymbol{h}_{\mathrm{E}}(\tau) + \boldsymbol{h}_{\mathrm{L}}(\tau)) * s(t) \\ &= \boldsymbol{x}_{\mathrm{E}}(t) + \boldsymbol{x}_{\mathrm{L}}(t), \\ \boldsymbol{x}_{\mathrm{N}}(t) &= (\boldsymbol{h}_{\mathrm{E}}(\tau) + \boldsymbol{h}_{\mathrm{L}}(\tau)) * s_{\mathrm{N}}(t), \end{aligned} \tag{2}$$

where $\boldsymbol{x}_{\mathrm{N}}(t)$ denotes the contribution from background noise and $\boldsymbol{x}_{\mathrm{S}}(t)$ denotes the contribution from speech and its reverberation. The signal $s(t)$, $\boldsymbol{h}_{\mathrm{E}}(\tau)$ and $\boldsymbol{h}_{\mathrm{L}}(\tau)$ are the clean speech source and the early and late parts of the room impulse response, respectively, with $*$ denoting the convolution operation. The term $s_{\mathrm{N}}(t)$ is the source of background noise, including interference speech if exists. The early and late impulse responses can be defined as

$$\boldsymbol{h}_{\mathrm{E}}(\tau) = \begin{cases} \boldsymbol{h}(\tau) & \text{for } \tau \leq \tau_{\mathrm{d}} \\ 0 & \text{for } \tau > \tau_{\mathrm{d}} \end{cases} \tag{3}$$

$$\boldsymbol{h}_{\mathrm{L}}(\tau) = \begin{cases} 0 & \text{for } \tau \leq \tau_{\mathrm{d}} \\ \boldsymbol{h}(\tau) & \text{for } \tau > \tau_{\mathrm{d}}. \end{cases} \tag{4}$$

The array processing method aims to suppress the background noise and late reverberation, and to obtain the estimated early reverberant speech signal $\hat{x}_{\mathrm{E}}(t)$.

In the time-frequency domain, the mixture model in each frequency bin can be simplified without explicitly separating the early and late reverberation components, as given by

$$\boldsymbol{X}(f,k) \approx \boldsymbol{H}_{\theta}(f)S(f,k) + \boldsymbol{X}_{\mathrm{N}}(f,k), \tag{5}$$

where $S(f,k)$ is the clean speech component, $\boldsymbol{X}_{\mathrm{N}}(f,k)$ contains components of background noise and

$$\boldsymbol{H}_{\theta}(f) = \{\exp(j2\pi(f/F)f_{\mathrm{s}}\frac{md}{c}\sin\theta(f)\}_{m\in[0,M-1]} \tag{6}$$

is an $M \times 1$ vector depending on the direction of arrival (DOA) $\theta(f)$ of the speech, with F the size of the discrete Fourier transform, d the microphone spacing, f_{s} the sampling frequency and c the sound velocity. Here, M denotes the number of microphone and f and k denote the frequency bin and time frame, respectively.

Without loss of generality, Eq. (5) can be reformulated as

$$\boldsymbol{X}(f,k) = \boldsymbol{A}(f)\boldsymbol{S}(f,k), \tag{7}$$

with $S_1(f,k) = S(f,k)$ and $[S_2(f,t), ..., S_N(f,k)] = \boldsymbol{X}_{\mathrm{N}}(f,k)$. It is also realistic to assume that, in each frequency bin, the speech component is statistically independent of the noise component. In this study, the number of sound source is limited to one target source and one interference, but the idea can be extended into any number of sound source.

2.3 Blind Source Separation

Blind source separation (BSS) is a class of array processing techniques that works on a principle that signal mixture captured at the multichannel microphone can be separated according to their source by applying a certain demixing matrix. As its name suggests, BSS was previously developed to separate multisource signals. But the separation function can also be utilized to separate the target speech from other unwanted sounds. The term blind is used because this class of method does not required *a priori* information about the relative positions of the sensors or the positions of the sources. Many BSS algorithms [13, 14, 15] are based on independent component analysis (ICA) [16, 17], in which the source signals are separated by utilizing their statistical independence. Also, it has been known that BSS in frequency domain performs better to separate convolutive mixture than the time-domain BSS. In this section, I discuss about the conventional ICA-based frequency domain BSS.

In this method, the estimated separated signals $\boldsymbol{Y}(f,k)$ in the fth frequency bin are obtained by applying an demixing matrix $\boldsymbol{W}(f)$ to the observed signals, as given by

$$\boldsymbol{Y}(f,k) = \boldsymbol{W}(f)\boldsymbol{X}(f,k) = \boldsymbol{W}(f)\boldsymbol{A}(f)\boldsymbol{S}(f,k). \tag{8}$$

$\boldsymbol{W}(f)$ is updated so that the output signals in $\boldsymbol{Y}(f,k)$ are mutually independent. Among the many proposed algorithms for ICA, an approach based on higher-order statistics exists, in which the optimization is based on the non-Gaussianity of the signal. The optimal demixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$ is then obtained from the iterative operation

$$\boldsymbol{W}^{[i+1]}_{\mathrm{ICA}}(f) = \mu[\boldsymbol{I} - \langle\phi(\boldsymbol{Y}(f,k))\boldsymbol{Y}^{\mathrm{H}}(f,k)\rangle_k]\boldsymbol{W}^{[i]}_{\mathrm{ICA}}(f) + \boldsymbol{W}^{[i]}_{\mathrm{ICA}}(f), \tag{9}$$

where $\boldsymbol{I}$ is the identity matrix, $\langle\cdot\rangle$ denotes the time-averaging operator and $\phi(\cdot)$ is a nonlinear activation function. Because of the super-Gaussian characteristics of the speech signal, the appropriate nonlinear function should not grow too fast with the signal amplitude, for example, tangent hyperbolic or the sigmoid function are suitable function [18].

Since the above calculations are carried out independently in each frequency bin, FD-ICA suffers from two problems, i.e., source permutation and scaling indeterminacy. This can be written as

$$\boldsymbol{Y}(f,k) = \boldsymbol{P}(f)\boldsymbol{\Lambda}(f)\boldsymbol{S}(f,k), \tag{10}$$

where $\boldsymbol{P}(f)$ is an $M \times M$ permutation matrix and $\boldsymbol{\Lambda}(f)$ is a diagonal $M \times M$ matrix.

The scaling indeterminacy filter $\boldsymbol{\Lambda}(f)$ can be solved by applying a *projection back* of the separated independent components to the microphone array input [14]. In the case of speech and diffuse background noise mixture, if $Y_m(f,t)$ is the estimated speech component, the projection back of the noise components can be defined by

$$\hat{\boldsymbol{X}}_N(f,k) = \boldsymbol{W}(f)^{-1}(\boldsymbol{I} - \boldsymbol{D}_m)\boldsymbol{Y}(f,k), \tag{11}$$

where $\boldsymbol{D}_m$ is a matrix having only one non-null entry $d_{mm} = 1$.

The remaining permutation problem $\boldsymbol{P}(f)$ requires the matching of the components belonging to the same signal across all the frequency bins. This is carried out by applying a *permutation resolution.* Methods of permutation resolution often utilize the DOA or temporal structure of signals. Assuming that the separation is perfect in each frequency bin, then

$$\boldsymbol{W}(f)\boldsymbol{A}(f) = \boldsymbol{P}(f)\boldsymbol{\Lambda}(f). \tag{12}$$

The authors of [19] showed that in the presence of nonpoint sources such as diffuse noise, the square matrix $\boldsymbol{W}_{\text{ICA}}(f)$ is such that the row corresponding to the estimated speech component is a delay-and-sum (DS) beamformer in the direction of the speech's apparent DOA at that frequency, while the other rows corresponding to the estimates of the noise components are null beamformers at that direction. Consequently, the quality of the noise estimate is superior to that of the speech estimate as the null beamformers efficiently suppress the speech (a point source) from the estimated noise components, whereas the DS beamformer does not suppress the noise from the estimated speech component. Therefore, performing frequency domain BSS alone may not be sufficient in the case of speech with diffuse noise mixture.

2.3.1 Blind Signal Exraction

For specific case of the presence of diffuse background noise, an alternative to BSS has been introduced in [20], namely frequency-domain blind signal extraction (BSE). The algorithm for BSS and BSE is depicted in Fig. 6. Unlike the

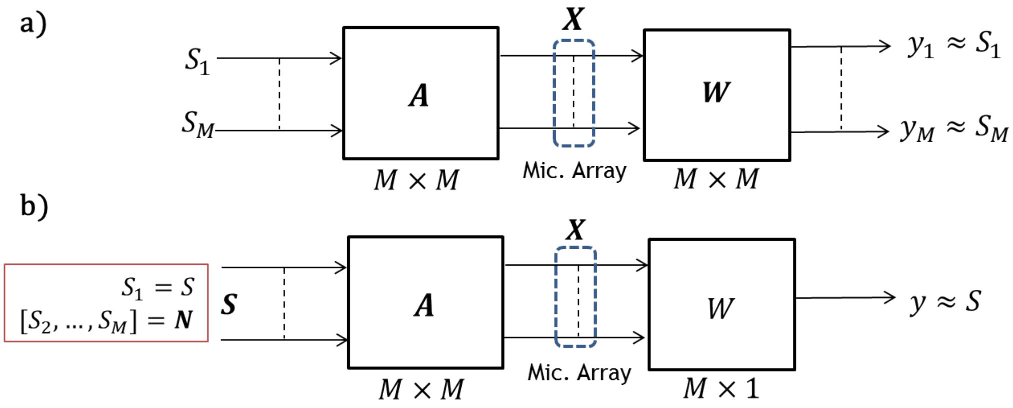

Figure 6. The block diagram of a) BSS and b) BSE.

conventional BSS, the BSE algorithm only extracts the desired speech components from the noise components.

Given the same observation signal $\boldsymbol{X}(f,k)$, BSE estimates only the components of $S_1(f,k)$ in each frequency bin by applying extracting vector $W(f)$, as given by

$$Y(f,k) = W(f)\boldsymbol{X}(f,k). \tag{13}$$

The vector $W(f)$ is updated using a gradient decent method to minimize the cost function $J(W(f))$ given by,

$$J(W(f)) = \frac{1}{2}\langle|Y(f,k)|\rangle^2, \tag{14}$$
$$\langle|Y(f,k)|^2\rangle = 1. \tag{15}$$

The cost function implies that the extracted component has a modulus with a small mean and a large variance, or in other word, the component is sparse that most of the values are close to zero and only a few are significantly large. In the case of a target speech within diffuse background noise, the speech modulus may be considered sparser than that of the noise components; thus the cost function is minimum when the target speech component is extracted. In this way, it is not required to confirm the selection of noise components, which means the permutation problem as in BSS can be avoided.

2.3.2 Performance Comparison Between Frequency-Domain BSE and BSS

BSE outperforms BSS in speech and diffuse noise mixture in terms of output signal quality and the computational complexity, owing to a simple nonparametric cost function. However, its performance is yet to be evaluated in the case of the presence of point-source noise. In order to investigate this, a preliminary experiment is conducted for both the mixture of speech with diffuse noise and speech with point-source noise. In this experiment, the clean speech was convoluted with recorded room impulse response with T_{60} of approximately 250 ms. For simulating the speech and diffuse noise, 10 female and 10 male utterances are each convoluted with room impulse response and mixed with recorded real noise at SNR of 10 dB.

For simulating the speech and point-source noise, 4 female and 4 male target speech are each mixed with interference speech convoluted with room impulse response. It is assumed that the target speaker always stays closer to the microphone than the interference speaker. Therefore, for near speaker distance, the interference speech comes from the far distance, and for far speaker distance, the interference speech comes from the same distance but has half the energy of the target speech signals. The performance of each method is evaluated using word recognition accuracy measure.

The experimental result for both cases is shown in Fig. 7. It is clearly shown that BSE outperforms BSS when the background noise is diffuse. However, its performance drops in the presence of speech as point-source noise because the sparseness assumption in the cost function does not hold when the interference has similar statistics to the target speech. From this preliminary result, it is of great interest to incorporate the benefit of BSE into BSS to built a new method that is more robust to various type of interference.

2.4 Spectral-based Speech Enhancement

The alternative approach in array processing is based on the spectral modification of the observed signal. Most of the methods in this field are the extension of single-channel processing algorithm. The methods work in frame-by-frame

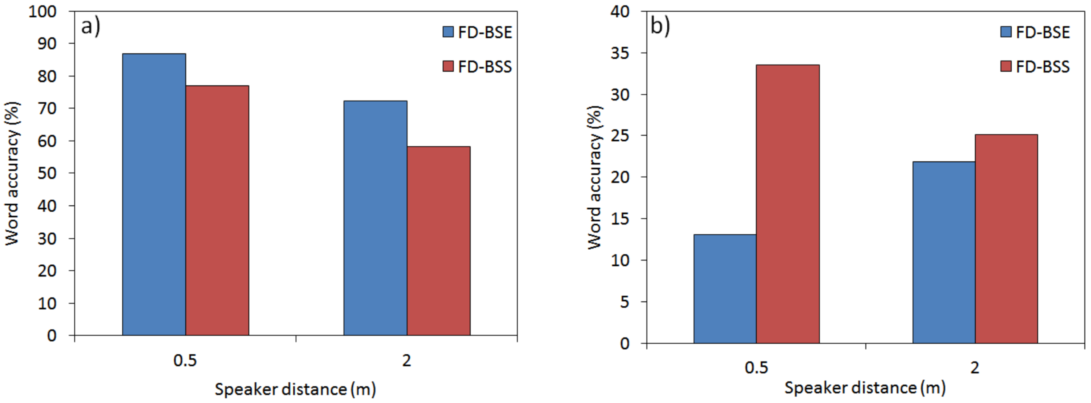

Figure 7. Performance of conventional frequency-domain BSE and BSS under different acoustic condition; a) speech mixed with diffuse background noise, and b) speech mixed with other interference speech.

manner, resulting in nonlinear processing that is more robust to nonstationary interferences but are prone to the artificial distortion of the residual noise, which is also known as the phenomenon of *musical noise*. One of the most well-known method is spectral subtraction [4], in which the estimated noise component is subtracted from the observed signal either in spectral or power spectral domain to obtain the clean target speech. There are also methods in which a filter is applied to the observed signal to suppress the interference, such as in Wiener filtering (WF) family [5].

Another spectral-based speech enhancement method is based on the assumption of the statistical model of the speech spectrum. The aim of this class of method is to estimate the spectrum of clean speech signal in noisy environment using statistical estimation framework [21]. Authors in [22] emphasizing on the estimation of speech spectral amplitude, acknowledging the importance of short-time spectral amplitude (STSA) on speech intelligibility and quality. In this method, speech STSA is obtained using a minimum mean-square error (MMSE) estimator. The optimal estimator is sought to minimized the mean-square error between the estimated and the true (clean) speech amplitudes. The clean speech prior is assumed to follow the Gaussian distribution, motivated by the central limit theorem. Furthermore, an estimator for the local *a priori* SNR is built using *decision-directed* approach. Then, the estimated clean speech is obtained

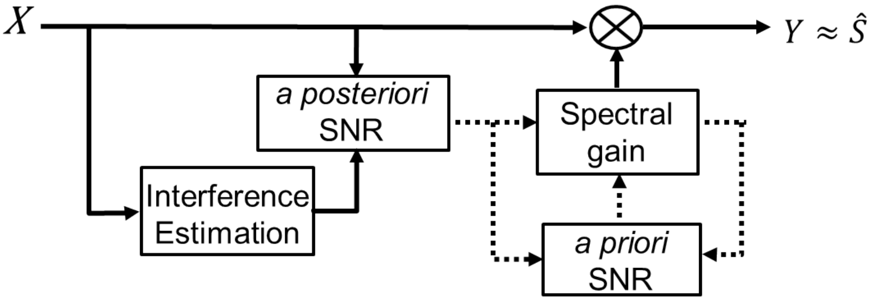

Figure 8. Block diagram of generalized MMSE-STSA estimator.

following the process flow as depicted in Fig. 8.

A generalization of MMSE-STSA estimator is also referred to as *MMSE estimation with optimizable speech model and inhomogeneous error criterion* (MOSIE) estimator [23]. In this method, the PDF of the clean speech spectral amplitude is modeled by a chi distribution, as given by

$$p(s) = \frac{2}{\Gamma(\rho)}\left(\frac{\rho}{P_{\mathrm{S}}(f)}\right)^{\rho} s^{2\rho-1}\exp\left(-\frac{\rho}{P_{\mathrm{S}}(f)}s^2\right), \tag{16}$$

where ρ is the shape parameter of the speech model, $P_{\mathrm{S}}(f)$ is the mean of speech power spectrum, and $\Gamma(\cdot)$ is the complete Gamma function. The value of $\rho = 1$ corresponds to Gaussian-distributed speech prior as in the original MMSE-STSA estimator. As speech is known to be super-Gaussian-distributed, it is reasonable to set the value of ρ from the range of $0 < \rho \leq 1$ to model the speech spectral amplitude.

In similar manner to the original MMSE-STSA estimator, MOSIE is used for single-channel noise suppression by applying the gain function $G(f,k)$ to the observed signal, as given by

$$Y_{\mathrm{MOSIE}}(f,k) = G(f,k)X(f,k), \tag{17}$$

$$G(f,k) = \frac{\sqrt{\nu(f,k)}}{\hat{\gamma}(f,k)}\left[\frac{\Gamma(\rho+\beta/2)}{\Gamma(\rho)}.\frac{\Phi(1-\rho-\beta/2,1;-\nu(f,k))}{\Phi(1-\rho,1;-\nu(f,k)}\right]^{1/\beta}, \tag{18}$$

$$\nu(f,k) = \frac{\hat{\xi}(f,k)}{\rho+\hat{\xi}(f,k)}\hat{\gamma}(f,k), \tag{19}$$

where $\Phi(a, b, c)$ is the confluent hypergeometric function. $\hat{\xi}(f, k)$ and $\hat{\gamma}(f, k)$ are the estimated *a priori* and *a posteriori* SNRs, respectively, as given by

$$\hat{\xi}(f, k) = \alpha\hat{\gamma}(f, k-1)G^2(f, k-1) + (1-\alpha)\text{max}[\hat{\gamma}(f, k) - 1, 0], \tag{20}$$

$$\hat{\gamma}(f, k) = \frac{|X(f, k)|^2}{|X_{\mathrm{N}}(f, k)|^2}, \tag{21}$$

where α is the forgetting parameter in the decision-directed approach [22]. In addition, β is the compression parameter of the error function given by

$$e(S_{\mathrm{o}}(f, k), S_{\mathrm{p}}(f, k)) = |S_{\mathrm{o}}(f, k)|^{\beta} - |S_{\mathrm{p}}(f, k)|^{\beta}, \tag{22}$$

where $S_{\mathrm{o}}(f, k)$ and $S_{\mathrm{p}}(f, k)$ are the speech spectral amplitude before and after processing by MOSIE, respectively.

Several well-known estimators can be derived from the MOSIE estimator depending on the choice of ρ and β. For example, by applying $\rho = 1$ and $\beta = 1$, MOSIE will be equivalent to the conventional MMSE-STSA estimator. The log spectral amplitude estimator [24] can be approximated with $\rho = 1$ and $\beta = 0.001$.

2.5 Experimental Setup and Corpora

In this study, I use utterances from the Japanese newspaper article sample (JNAS) speech corpus [25] as target speech. The speech corpus contains speech recordings of 153 males and 153 females reading excerpts from newspaper's articles and phonetically balanced (PB) sentences, sampled at 16 kHz . It was developed to accommodate Japanese large vocabulary (20K) continuous speech recognition task. The training data for the acoustics model consists to about 100 sentences per speaker for over 100 speakers.

The room impulse responses used for experimental evaluations throughout this dissertation were recorded in real rooms using a microphone array. For the experiments in Sect. 3 and 4, the impulse response recorded in a large lecture room is used. The reverberation time T_{60} of the room is approximately 500 ms. The recording is carried out using a linear array of 8 omni-directional microphones, with the distance between each microphone is 2.5 cm. The target speech is assumed to be at the normal direction relative to the array. The background

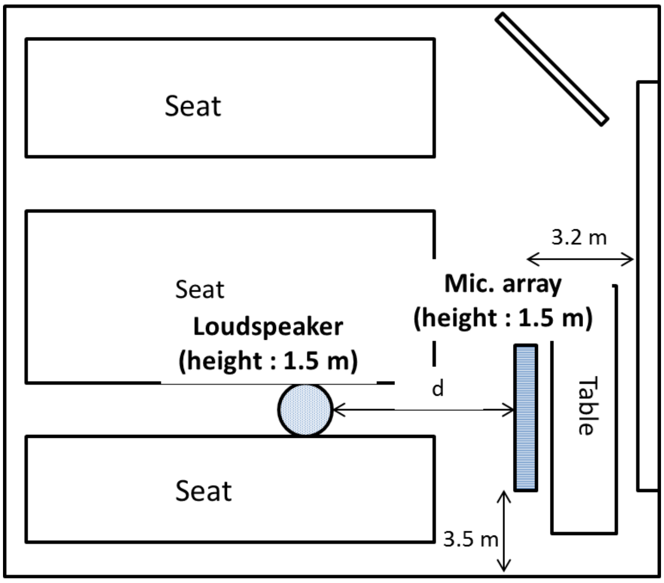

Figure 9. Recording setup for the room impulse response used in Sect. 3 and 4.

noise is also recorded separately using the same microphone array. The recording setup is illustrated in Fig. 9.

The preliminary experiment and all experiments in Sect. 5 are conducted using the room impulse response data from REVERB Challenge 2014 [26], which is an event to evaluate the dereverberation and ASR techniques based on a common database and evaluation metrics. There are impulse response data from 3 rooms provided in this challenge, i.e. a small, medium, and large-size meeting room with T_{60} of about 250 ms, 680 ms, and 730 ms, respectively. The recordings is carried out from two different angles using 8-channel circular microphone arrays with diameter of 20 cm. The recordings were taken for two speaker-microphone distance, the 'near' case being 0.5 m and the 'far' case being 2.0 m. The background noise is also recorded using the same microphone array.

In the first and third room (small and large-size meeting room), the position of the target speaker is set to 45° relative to the first microphone channel, while the interference speaker is positioned at 135° relative to the first microphone channel. For the second room (medium-size meeting room), the position of the target speaker is set to 45° relative to the seventh microphone channel, while the interference speaker is positioned at 135° relative to the seventh microphone

Table 1. System specification of ASR

Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCC, 12-order ΔMFCC, 1-order ΔE
Acoustic model	HMM phonetic tied mixture (PTM), 2000 states, GMM 64 mixtures
Language model	standard word trigram model
Training data	Adult JNAS database

channel.

To simplify the problem, in this study only the data from 2 channel are selected from the circular array, creating a linear array with inter-microphone space of 7.65 cm. This is done because the 2-channel linear configuration is more flexible to extend into any number of microphone. It is worth to note that the inter-microphone distance in this array is quite large, causing the spatial aliasing to occur [27]. This will disrupt the estimation of DOA of the target speech. To cope with this problem, only the low frequency components (up to about 2.2 kHz) is used for calculating the estimate DOA.

The performance of each method is evaluated by ASR system using the word recognition accuracy given by

$$\mathrm{WA} = 100 \times \frac{N - (I + S + D)}{N}, \tag{23}$$

where N is the number of words in the reference transcriptions, I is the number of insertions, S is the number of substitutions, and D is the number of deletions. Julius 4.2[28] is used as the decoder in ASR system. The system specification is shown in table 1.

2.6 Summary

In this section, the problem of distant speech recognition system has been modeled mathematically. Some important assumptions used in the model have also been described. Then, I provided short review on some well-known speech enhancement techniques, covering their advantages and weakness. Conventionally, these techniques have only been viewed as a means of improving the quality of the speech waveform. Throughout the next sections, I present a framework in which the microphone array processing are specifically optimized for improved speech recognition performance. The novel methods extend and combine the existing speech enhancement techniques and are optimized to maximize the likelihood of the acoustic model in ASR system. This integrated system is more appropriate for the implementation in distant speech recognition system.

3. Semi-Blind Noise Suppression and Dereverberation based on Frequency-Domain BSE

3.1 Introductions

The implementation of microphone array in a hands-free robot dialog system allows a more natural and stress-free interface for human-robot interaction. However, it is difficult to achieve accurate speech recognition, because the background noises always degrade the target speech quality. Furthermore, the distance between the speaker and the robot also causes the reverberation to be captured along with the target speech.

In this section, I propose a semi-blind method based on BSE that jointly suppress diffuse background noise and late reverberation, assuming only one speaker is active at a time. This method is the extension of the work in [29], with the optimization scheme that address the speech recognition accuracy improvement. The conventional BSE-based joint method has to be manually optimized, therefore it is unsuitable for the real environment implementation. Moreover, the performance of this method is not stable if the interference is not too severe.

In the proposed method, I make assumption that the robot has its own video camera, and thus, the position of the target speaker can be immediately detected. From this image information, I develop a semi-blind optimization scheme for the joint method. I conduct experiment to evaluate the performance of the proposed method.

3.2 Main Algorithm

3.2.1 Noise Suppression Stage

The block diagram of the optimized BSE-based joint method is depicted in Fig. 10. In this method, BSE is utilized to estimate the background noise component. This is done by subtracting the orthogonal projection of the extracted speech component $Y(f,k)$ obtained from Eq. (13), as given by

$$\widehat{\boldsymbol{X}}_{\mathrm{N}}(f,k) = (\boldsymbol{I}_{\mathrm{M}} - P_{\boldsymbol{X}}(f)\lambda^{\mathrm{H}}W^{\mathrm{H}}(f)\lambda W(f))\boldsymbol{X}(f,k), \tag{24}$$

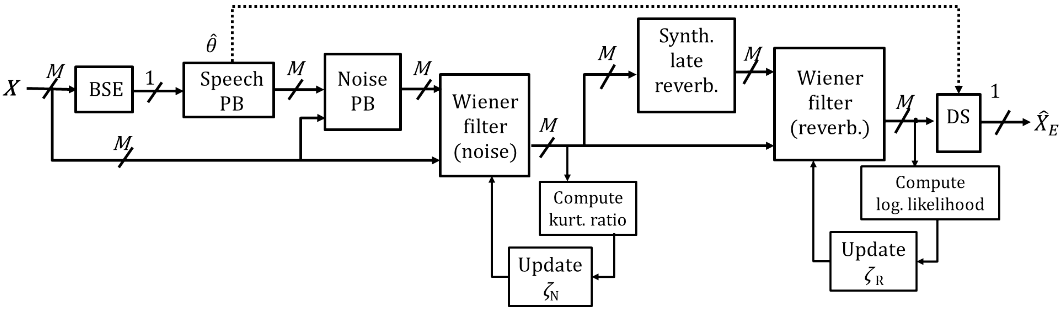

Figure 10. The block diagram of the optimized BSE-based joint method.

where λ is a scalar such that $Q(f,k) = \lambda W(f)\boldsymbol{X}(f,k)$ verifies $\langle |Q(f,k)|^2\rangle = 1$. $P_{\boldsymbol{X}}$ is given by

$$P_{\boldsymbol{X}} = \langle \boldsymbol{X}(f,k)\boldsymbol{X}^H(f,k)\rangle. \tag{25}$$

Then, the estimated noise is suppressed using multichannel WF, as given by

$$\begin{aligned} \widehat{\boldsymbol{X}}_{\mathrm{S}}(f,k) &= G|\boldsymbol{X}(f,k)|\mathrm{e}^{jarg(\boldsymbol{X}(f,k))}, && (26)\\ G &= \frac{|\boldsymbol{X}(f,k)|^2}{|\boldsymbol{X}(f,k)|^2 + \zeta_{\mathrm{N}}|\widehat{\boldsymbol{X}}_{\mathrm{N}}(f,k)|^2}, && (27)\end{aligned}$$

where β_{N} is a parameter used to control the strength of noise suppression.

The output of noise suppression stage will be used to synthesize the late reverberation component at the dereverberation stage. Therefore, it is important to optimize the parameter of WF in order to improve the quality of the output waveform of this stage. One problem that commonly arises in nonlinear signal processing methods such as WF and SS is the occurrence of musical noise. This artificial noise distorts the spectrum of the speech output of the noise suppression stage. Thus, the parameter of WF shall be optimized to avoid generating musical noise excessively.

I utilize kurtosis ratio (KR) in the optimization scheme, which measure a ratio of kurtosis of the residual noise in the processed signal to the kurtosis of unprocessed noise in the observed signal [30]. Kurtosis of noise spectrum is calculated frequency subband-wise in speech absence region, as given by

$$\mathrm{kurt}^{(i)} = \frac{(1/L)\sum_{f\in F_i}\sum_{k\in T}(|\boldsymbol{X}(f,k)|^2)^4}{\{(1/L)\sum_{f\in F_i}\sum_{k\in T}(|\boldsymbol{X}(f,k)|^2)^2\}^2}, \tag{28}$$

where $\mathrm{kurt}^{(i)}$ is the ith subband kurtosis of a signal x. F_i and T represent the evaluated subband time-frequency grid indexes, while L is the total number of grids in each subband. Here, a 250-Hz-width F_i and a T of 5 s are used, which are taken from a noise-only time-frequency region preceding a speech utterance. Then, ζ_{N} is updated in an iterative manner to achieve the optimum noise reduction ratio (NRR) under a KR constraint, as given by

$$\widehat{\zeta}_{\mathrm{N}} = \arg\max_{\zeta_{\mathrm{N}}} \mathrm{NRR}(\zeta_{\mathrm{N}}), \tag{29}$$

$$\frac{\mathrm{kurt}_{\mathrm{proc}}(\zeta_{\mathrm{N}})}{\mathrm{kurt}_{\mathrm{org}}} \leq \mathrm{KR}_{\mathrm{lim}}, \tag{30}$$

where $\mathrm{kurt}_{\mathrm{org}}$ and $\mathrm{kurt}_{\mathrm{proc}}(\zeta_{\mathrm{N}})$ are the kurtosis of the unprocessed noise and the output from noise suppression stage, respectively, and $\mathrm{KR}_{\mathrm{lim}}$ is the constraint value of KR. The procedure is described as follow.
Step 0: First, set initial ζ_{N}.
Step 1: Next, apply the WF for noise suppression using the value of ζ_{N}.
Step 2: Apply DS beamformer to the output signal, then calculate the kurtosis. Obtain kurtosis ratio by dividing noise kurtosis of the output signal by the noise kurtosis of observed signal.
Step 3: Increase the value of ζ_{N} by a certain amount $\Delta_{\zeta_{\mathrm{N}}}$. Return to Step 1 until the kurtosis ratio value reaches the given limit, or until the difference between updated value and previous value of kurtosis ratio is below certain threshold.

3.2.2 Dereverberation Stage

Provide that the noise suppression stage is effective, the estimated $\widehat{\boldsymbol{X}}_{\mathrm{S}}(f,k)$ contains only the early reverberant speech $\boldsymbol{X}_{\mathrm{E}}(f,k)$ and late reverberant speech $\boldsymbol{X}_{\mathrm{L}}(f,k)$. The task in dereverberation stage consists of the estimation and suppression of the late reverberation components.

The estimation of the late reverberation components can be separated into two tasks: estimating the late impulse response $\boldsymbol{h}_{\mathrm{L}}(\tau)$ and the clean speech signal $s(t)$. In this method, the estimated $\boldsymbol{h}_{\mathrm{L}}(\tau)$ is approximated by generating an exponentially decayed Gaussian random variable $\boldsymbol{u}(\tau)$ as given by [31]

$$\begin{aligned}\boldsymbol{h}_{\mathrm{L}}(\tau) &= au(\tau)\mathrm{e}^{-d(\tau-\tau_{\mathrm{d}})}, \qquad (31)\\ d &= \frac{\ln 10^6}{2(T_{60}-\tau_{\mathrm{d}})}, \qquad (32)\end{aligned}$$

where a is a scaling factor. The direct speech $s(t)$ is approximated by projecting back the output of the noise suppression stage $\widehat{\boldsymbol{X}}_{\mathrm{S}}(f,k)$ to the truncated FD-BSE filter $W_{\mathrm{trunc}}(f)$, as given by

$$\widehat{\boldsymbol{S}}(f,k) = P_{\widehat{\boldsymbol{X}}_{\mathrm{S}}(f)}\lambda^{\mathrm{H}}W^{\mathrm{H}}_{\mathrm{trunc}}(f)\lambda W_{\mathrm{trunc}}(f)\widehat{\boldsymbol{X}}_{\mathrm{S}}(f,k). \qquad (33)$$

The scaling factor a is obtained from the energy difference between $\widehat{\boldsymbol{X}}_{\mathrm{S}}(f,k)$ and $\widehat{\boldsymbol{S}}(f,k)$. Then, according to Eq. (2), $\widehat{\boldsymbol{X}}_{\mathrm{L}}(f,k)$ is obtained by applying a convolution in the time domain, given by

$$\widehat{\boldsymbol{x}}_{\mathrm{L}}(t) = \boldsymbol{h}_{\mathrm{L}}(\tau) * \widehat{\boldsymbol{s}}(t), \qquad (34)$$

then it transformed back to the time-frequency domain by applying an STFT. After that, the dereverberation process is carried out in the same manner as the noise suppression stage, using multichannel WF given by

$$\begin{aligned}\widehat{\boldsymbol{X}}_{\mathrm{E}}(f,k) &= G|\widehat{\boldsymbol{X}}_{\mathrm{S}}(f,k)|\mathrm{e}^{jarg(\widehat{\boldsymbol{X}}_{\mathrm{S}}(f,k))}, \qquad (35)\\ G &= \frac{|\widehat{\boldsymbol{X}}_{\mathrm{S}}(f,k)|^2}{|\widehat{\boldsymbol{X}}_{\mathrm{S}}(f,k)|^2+\zeta_{\mathrm{R}}|\widehat{\boldsymbol{X}}_{\mathrm{L}}(f,k)|^2}, \qquad (36)\end{aligned}$$

where ζ_{R} is a parameter to control the strength of dereverberation.

DS beamformer is applied in the direction of the target speech to merge the output of the dereverberation stage $\widehat{\boldsymbol{X}}_{\mathrm{E}}(f,k)$ as given by

$$\begin{aligned}\widehat{X}_{\mathrm{E}}(f,k) &= \boldsymbol{w}^{\mathrm{T}}_{\mathrm{DS}}(f)\widehat{\boldsymbol{X}}_{\mathrm{E}}(f,k), \qquad (37)\\ \boldsymbol{w}^{\mathrm{T}}_{\mathrm{DS}}(f) &= \left[w^{(\mathrm{DS})}_1(f), w^{(\mathrm{DS})}_2(f), ...w^{(\mathrm{DS})}_{\mathrm{M}}(f)\right]^{\mathrm{T}}, \qquad (38)\\ w^{(\mathrm{DS})}_{\mathrm{m}}(f) &= \frac{1}{\mathrm{M}}\exp(-i2\pi(f/F)f_s d\ \sin\widehat{\theta}(f)/c). \qquad (39)\end{aligned}$$

The estimated DOA $\widehat{\theta}$ is calculated from the projection back of the FD-BSE filter as given by

$$\begin{aligned}\boldsymbol{K}(f) &= P_{\boldsymbol{X}}(f)\lambda^{\mathrm{H}}\boldsymbol{W}^{\mathrm{H}}(f)\lambda\boldsymbol{W}(f), \qquad (40)\\ \widehat{\theta}(f) &= \arg\sin\left(\frac{cF}{2\pi f f_{\mathrm{s}} d}\mathrm{angle}\left(\frac{\{\boldsymbol{K}(f)\}_{i+1,j}}{\{\boldsymbol{K}(f)\}_{i,j}}\right)\right), \qquad (41)\end{aligned}$$

Finally, the desired output $\widehat{x}_{\mathrm{E}}(t)$ is obtained by applying an inverse STFT to the output of DS beamformer.

Because this method is designed to be implemented in hands-free robot dialogue system, the final stage must be optimized to improve the speech recognition accuracy. Therefore, the WF parameter at dereverberation stage is optimized to maximizes the likelihood of the acoustic model of the ASR system.

The ASR system works on an extracted feature vector of the speech waveform. It hypothesizes the correct transcription of an utterance by finding the sequence that has the maximum likelihood of generating the extracted feature vector given the statistical acoustic models of the recognizers. First, a series of vectors $\boldsymbol{o} = [o_1, ..., o_T]$ containing Mel-frequency cepstral coefficient (MFCC) [32] is extracted from the speech waveform. Then, during decoding, the ASR system attempts to find the word sequence $\boldsymbol{Z} = [z_1, ..., z_K]$ that is most likely to generate the sequence $\boldsymbol{o}$, as expressed by

$$\widehat{\boldsymbol{Z}} = \arg\max_{\boldsymbol{Z}} P(\boldsymbol{Z}|\boldsymbol{o}). \tag{42}$$

By applying Bayes' theorem, the above expression can be written as

$$\widehat{\boldsymbol{Z}} = \arg\max_{\boldsymbol{Z}} \frac{P(\boldsymbol{o}|\boldsymbol{Z})P(\boldsymbol{Z})}{P(\boldsymbol{o})}, \tag{43}$$

where $P(\boldsymbol{o}|\boldsymbol{Z})$ is the acoustic likelihood or acoustic score, representing the probability that acoustic feature sequence $\boldsymbol{o}$ is observed given that word sequence $\boldsymbol{Z}$ was spoken, and $P(\boldsymbol{Z})$ is the language score, i.e., the *a priori* probability of a particular word sequence $\boldsymbol{Z}$, which is calculated using a language model. Taking into account only the part related to the acoustic feature of the signal, the WF parameter in the dereverberation stage is optimized to maximize

$$\widehat{\zeta}_{\mathrm{R}} = \arg\max_{\zeta_{\mathrm{R}}} P(\boldsymbol{o}(\zeta_{\mathrm{R}})|\boldsymbol{Z}). \tag{44}$$

For an HMM-based speech recognition system, the solution of this problem can be computed using the Viterbi algorithm [33].

In practice, the calculation of (44) requires the correct transcription $\boldsymbol{Z}$, but if it is already known, the speech recognizer is not required anymore. Therefore, the optimization procedure is carried out as follows.

Step 0: First, set initial ζ_{R}.
Step 1: Next, apply the WF for dereverberation using the value of ζ_{R}.
Step 2: Apply DS beamformer to the output signal, then calculate the log likelihood $P(\boldsymbol{o}(\zeta_{\mathrm{R}})|\boldsymbol{Z})$ using Viterbi algorithm.
Step 3: Increase the value of ζ_{R} by a certain amount $\Delta_{\zeta_{\mathrm{R}}}$. Return to Step 1 and use $\boldsymbol{Z}$ obtained from initial ζ_{R} for Viterbi alignment. Repeat the process until maximum score achieved.

3.3 Semi-blind Implementation of Joint Method

Although the proposed optimization scheme performs well under heavily reverberant conditions, it is still difficult to achieve the optimum performance when the level of interference is not severe. This may be caused by the distortion from the long processing. Therefore, we need a system that can adjust the signal processing method to be applied according to the environmental conditions. Here I incorporate the user position information provided by the robot's camera and develop a multimodal switching scheme based on distance information. An RGB camera with depth sensor is utilized to obtain the speaker distance information.

The block diagram of the proposed semi-blind scheme is shown in Fig. 11. It is assumed that the speaker distance from the robot corresponds to the severity of the interference. First, an offline training stage is conducted to estimate the distance at which the method should be switched. Both frequency domain BSE and the optimized joint method are applied to signals at various speaker-to-microphone distances and the results are compared. Next, after the switching distance has been decided, the system applies two different schemes according to the user position:

- For a short user distance, the reverberation is not severe so only frequency-domain BSE is applied to the input signal to suppress the background noise. The extracted speech from FD-BSE becomes the output signal.
- For a longer user distance, the complete optimized joint method is applied instead of only FD-BSE.

The switching point also depends on the SNR condition of the signal. The SNR can be easily approximated by utilizing the noise estimation from BSE. The

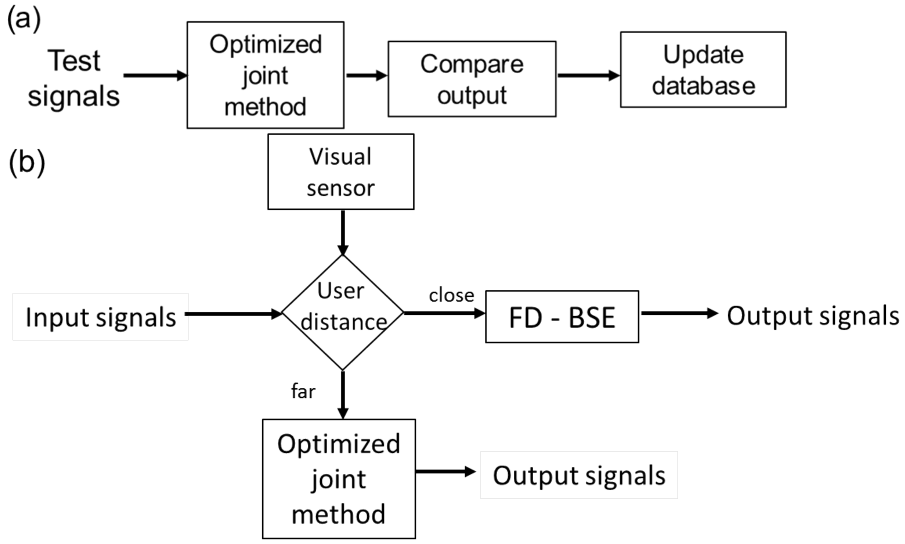

Figure 11. The block diagram of the semi-blind joint method.

calculation is carried out channel-wise, as given by

$$\mathrm{SNR}_{est} = 10 \log_{10} \frac{E[x(t)]^2 - E[\widehat{x}_{\mathrm{N}}(t)]^2}{E[\widehat{x}_{\mathrm{N}}(t)]^2}. \tag{45}$$

The lower average SNR indicates not only more severe background noise but also late reverberation, because the noise estimation from FD-BSE may also contain the late reverberation components since they have similar characteristics. In this case, the optimized joint method is more preferable, thus the switching distance is shorter than in the case of higher SNR.

3.4 Experimental Result and Discussion

For the evaluation, 100 utterances from the female JNAS corpus as the clean speech, each was convoluted with the room impulse response ($T_{60} = 500$ ms) and mixed with noise at SNRs of 0, 10, and 20 dB. The τ_d value was set to 75 ms, which corresponds to the effect of a room impulse response that can still be handled by the speech recognizer. The time-frequency domain processing was done by implementing the short-time Fourier transform (STFT) with a 1024 point FFT size, a Hanning window and 50% overlap. The BSE algorithm was performed

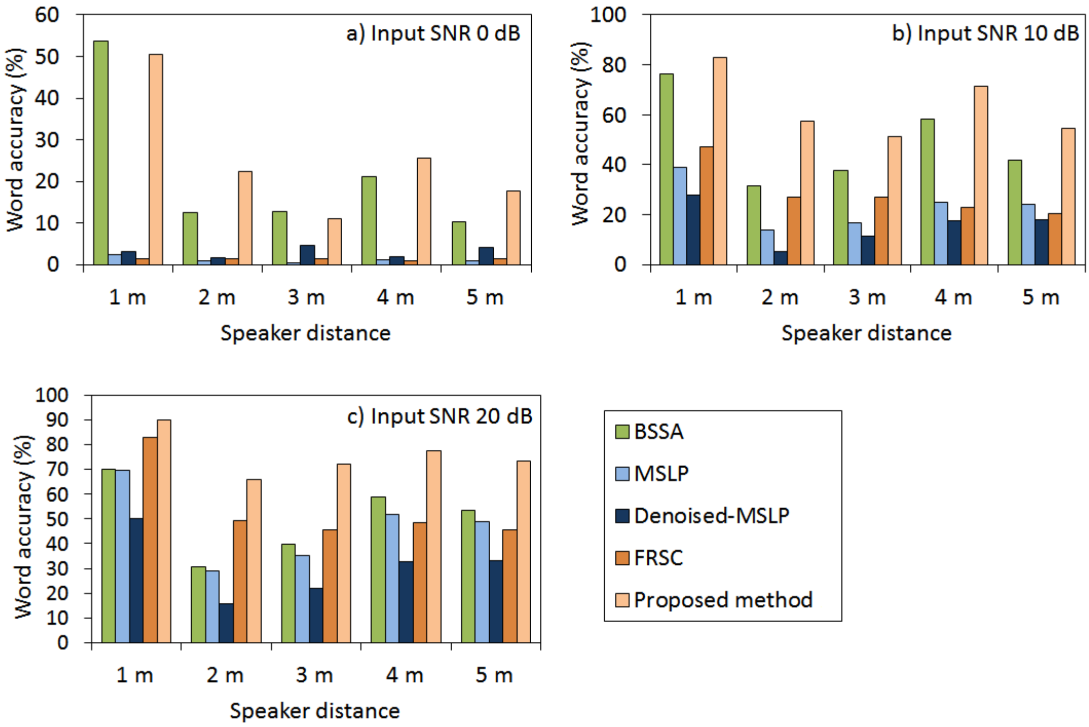

Figure 12. Word accuracy comparison among conventional methods and the proposed method in various input SNR condition: (a) 0 dB, (b) 10 dB and (c) 20 dB.

for 600 iterations with an adaptation step of 0.3, which was halved every 200 iterations.

The proposed method is compared to the currently known method, namely, blind spatial subtraction array (BSSA) [19], multi-step linear prediction based dereverberation method (MSLP) [34], and the adaptation from full-rank spatial covariance model (FRSC) [35]. Originally, MSLP works under noise-free assumption, so I also investigate the performance of MSLP combined with multichannel WF for suppressing the background noise (Denoised-MSLP). The background noise for this method is estimated with *a priori* SNR [21].

The word recognition accuracy result is shown in Fig. 12. It can seen that BSSA performs well under severe SNR conditions and short user distance, since this method only suppress background noise. On the other hand, MSLP fails to perform under severe SNR conditions. This is mainly because the method relies on noise-free assumption. Combination with WF for noise suppression do not improve the recognition accuracy, probably because the noise suppression causes more distortion to the target speech. The poor performance of FRSC may be caused by failure in initialize the parameters, as this method is sensitive to initialization process. Overall, the proposed method outperforms the other methods under most of conditions.

3.5 Summary

In this section, I present a semi-blind method to suppress diffuse background noise and late reverberation for a distant-talking robot system. This method combines BSE with two stages of multichannel WF, and utilizes the image information from the robot's camera to know the position (distance) of the speaker. Then, the information is used to select the optimum method between the optimized joint method and BSE to be implemented to the observed signal.

This semi-blind configuration can maintain the stable performance regardless the severity of interferences. However, there are many situations where the information of speaker position is not provided. Under such conditions, the proposed semi-blind method is impossible to be implemented. Therefore, in the next section I present the modified blind method based on BSE. I also introduce the extension of generalized MMSE-STSA as a nonlinear postprocessing for dereverberation in the next section.

4. Joint Noise Suppression and Dereverberation Combining FD-BSE and Generalized MMSE-STSA

4.1 Introduction

The semi-blind method proposed in Sect. 3 is effective to suppress the diffuse background noise and late reverberation. However, its implementation is limited to the situation where the image information revealing the user's position is accessible. Therefore, it is preferable to develop a blind joint method that can maintain a stable performance without the aid of another sensor.

In this section, I propose the modified blind joint noise suppression and dereverberation method based on BSE. Under assumption that only one speaker is active at a time, BSE is utilized as speech extractor, in contrast to its role as noise estimator in the previous method. Simultaneously, BSE also suppress the background noise as it extracts the target speech component. Thus, only one stage of postprocessing is required, that is the dereverberation stage.

I also introduce the extension of generalized MMSE-STSA (MOSIE) as a postprocessing for dereverberation stage. As described in Sect. 2, MOSIE works based on the prior assumption of the statistical model of speech STSA. An experiment is carried out to compare the performance of the proposed blind method with different postprocessing choices.

4.2 Main Algorithm

In the proposed method, the FD-BSE is used to extract the speech component from the observed signal, based on the cost function in Eq. (14). The scaling problem is solved by applying projection back (PB), as given by

$$\hat{\boldsymbol{X}}_{\mathrm{S}}(f,k) = P_{\mathrm{X}}(f)W^{\mathrm{H}}(f)Y_{\mathrm{BSE}}(f,k), \tag{46}$$

where $P_{\mathrm{X}}(f)$ indicates the covariance of the observed signals. Assuming that the extraction is effective, $\hat{\boldsymbol{X}}_{\mathrm{S}}(f,k)$ will only consist of clean speech and its reverberation. First, we synthesize the late reverberation in the time domain using (32),

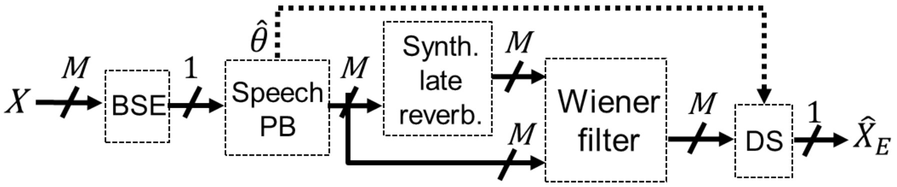

Figure 13. Block diagram of FD-BSE combined with multichannel WF postprocessing.

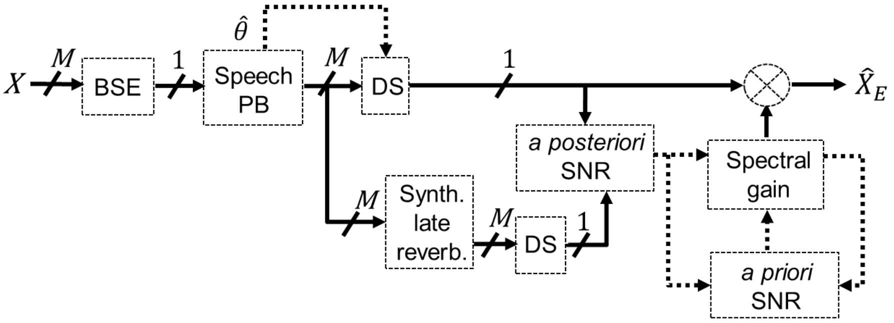

Figure 14. Block diagram of FD-BSE combined with single-channel MOSIE estimator postprocessing.

(33), and (34). Next, we compare three different postprocessing for dereverberation. The first method utilizes a multichannel WF as shown in Fig. 13, given by

$$\hat{\boldsymbol{X}}_{\mathrm{E}}(f,k) = G|\hat{\boldsymbol{X}}_{\mathrm{S}}(f,k)|\mathrm{e}^{jarg(\hat{\boldsymbol{X}}_{\mathrm{S}}(f,k))}, \quad (47)$$

$$G = \frac{|\hat{\boldsymbol{X}}_{\mathrm{S}}(f,k)|^2}{|\hat{\boldsymbol{X}}_{\mathrm{S}}(f,k)|^2 + \zeta_{\mathrm{R}}|\hat{\boldsymbol{X}}_{\mathrm{L}}(f,k)|^2}, \quad (48)$$

where ζ_{R} is a parameter for controlling the strength of dereverberation.

The second and third methods utilize single-channel and multichannel MOSIE estimator postprocessing, as shown in Fig. 14 and Fig. 15, respectively. The estimated speech is obtained by applying a gain function to the observed signal, as given by Eq. (19). Then, DS beamformer is applied in the end of multichannel processing to obtain single-channel output, with the direction-of-arrival calculated from the PB of the FD-BSE filter using Eq. (41).

Parametric postprocessing allows flexible control of the level of dereverberation. However, it is important to set the parameter to obtain the best speech

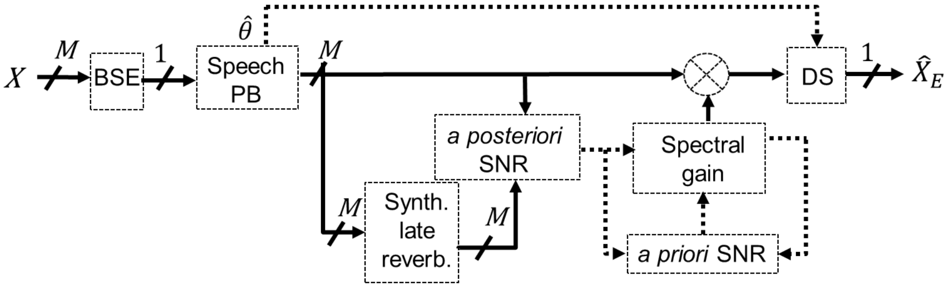

Figure 15. Block diagram of FD-BSE combined with multichannel MOSIE estimator postprocessing.

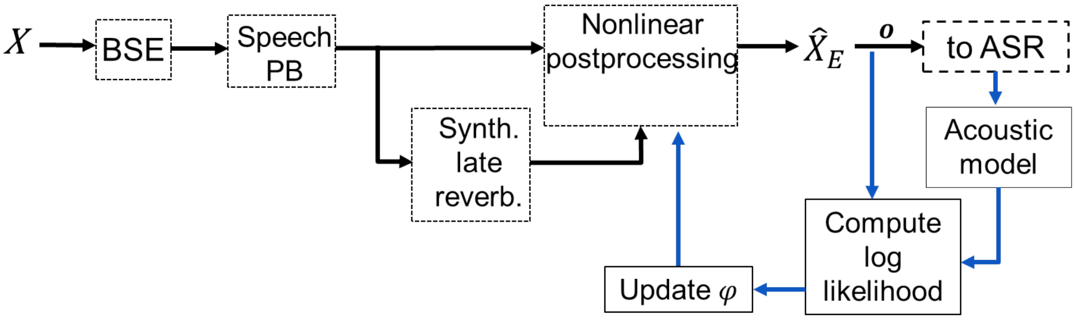

Figure 16. Block diagram of the optimized one stage blind noise suppression and dereverberation based on BSE.

recognition accuracy. In this study, only the internal parameter in nonlinear postprocessing, i.e., ζ_{R} in the multichannel WF and ρ in the MOSIE estimator, will be optimized. Other parameters, such as T_{60}, are assumed to be known. The parameters of nonlinear postprocessing are optimized to maximized the likelihood of the acoustic model of the speech recognizer. The optimization scheme is carried out in iterative manner, with the hypothesized transcription from the first iteration is used in Viterbi alignment for the rest of iteration. The general flow of the optimization scheme is depicted in Fig. 16.

4.3 Experimental Result and Discussion

Two experiments have been carried out for evaluation purposes. The observed signal was created by convolution of the clean speech with the impulse response

Table 2. Word accuracy results (%) of Experiment 1, input SNR = 10 dB

Distance	1 m	2 m	3 m	4 m	5 m
Reference	97.20	90.22	90.11	92.91	92.85
FD-BSE	89.65	61.55	56.32	70.59	41.24
FD-BSE + MC-WF	91.79	**79.28**	**73.13**	82.09	55.86
FD-BSE + SC-MOSIE $\alpha = 0.98$	74.44	56.36	55.97	61.94	36.04
FD-BSE + MC-MOSIE $\alpha = 0.96$	91.79	75.00	**73.13**	**85.07**	**58.56**
FD-BSE + MC-MOSIE $\alpha = 0.98$	**95.50**	65.91	72.07	82.84	52.25
FD-BSE + MOSIE-LSA $\alpha = 0.98$	71.64	60.23	61.26	73.87	48.86

($T_{60} = 500$ ms), and recorded real noise was added at SNR of 10 dB. The frequency domain processing was carried out with a 512-point Hamming window and 50% overlap of the STFT. FD-BSE was performed in 600 iterations with an adaptation step of 0.3. The parameter τ_{d} was set to 75 ms, corresponding to the delay that can still be handled by the speech recognizer.

4.3.1 Manual Tuning of the Parameter of the MOSIE Estimator

The first experiment was carried out using 5 male and 5 female utterances. The purpose of this experiment is to find the optimum parameter sets for the MOSIE estimator other than the shape parameter of chi-distributed speech prior ρ. The first parameter is α, i.e. the forgetting factor of decision-directed *a priori* SNR estimator. It is well known that the optimum value of α for hearing purpose is 0.98, as any value below 0.98 generates a noticeable amount of musical noise. In the experiment, I compare the performance of MOSIE postprocessing with α set to 0.98 and 0.96.

The second parameter to be tuned is β, i.e. the compression parameter of the error function. As described in Sect. 2, β value of 0.001 to represent the MOSIE-LSA estimator and 1 to represent the MOSIE-STSA estimator [23]. The best result among these combinations are manually selected. The results are shown in Table 2. It is shown that nonlinear postprocessing improves the recognition accuracy compared with the FD-BSE method, except for the case of the single-channel MOSIE-STSA estimator. This is understandable as single-channel processing tends to result in higher speech output distortion due to the lost of spatial information. It can be observed that the MOSIE-STSA estimator performs better than the MOSIE-LSA estimator for dereverberation in terms of word recognition accuracy.

It is also shown that the $\alpha = 0.96$ results in better word accuracy than $\alpha = 0.98$. This is an interesting finding because $\alpha = 0.98$ is a preferred setting for speech enhancement such as that for hearing aid system. This is possibly because a high quality output signal waveform is less important for speech recognition purposes. The performance of the MOSIE estimator with $\alpha = 0.96$ is similar to multichannel WF postprocessing in the case of small user-to-microphone distances, except for a distance of 2 m. For the long user distances, postprocessing with the MOSIE estimator achieves slightly better word accuracy.

4.3.2 Evaluation of Optimized Blind Joint Method

The optimization scheme of the proposed blind joint method is evaluated in the second experiment. The utterances from 50 male and 50 female speakers is used as the target speech. In this experiment, I compare the performance of combination of BSE and multichannel WF postprocessing to the combination of BSE and multichannel MOSIE postprocessing. For the multichannel MOSIE estimator, α and β were set to 0.96 and 1, respectively, following to the result in the first experiment.

The results of the second experiment is shown in Table 3. We can see that both optimized methods outperform frequency domain BSE, with an average improvement of 11.4% for the multichannel WF and 12.9% for MOSIE estimator postprocessing, and the highest improvement of 17.85% achieved by the multichannel WF at a distance of 3 m. This implies that by shortening the processing path, a more stable performance of blind noise suppression and dereverberation method can be achieved, hence the better recognition accuracy. It is also shown that the parameter optimization based on acoustic likelihood is effective for the

Table 3. Word accuracy results (%) of Experiment 2, input SNR = 10 dB

Distance	1 m	2 m	3 m	4 m	5 m
Reference	90.86	78.20	80.30	83.96	84.97
FD-BSE	77.50	52.99	47.59	60.63	33.80
FD-BSE+ MC-WF	82.00	**68.33**	**65.44**	68.15	45.68
FD-BSE+ MC-MOSIE	**83.82**	65.75	64.02	**73.38**	**50.05**

proposed blind method.

Despite its advantages, the current proposed method leaves potential and unsolved problems. The nonparametric characteristic of BSE causes the noise suppression stage to be non-optimizable. Thus, the performance of this method is greatly dependent to the capability of BSE to extract the clean speech from noisy mixture. From the preliminary experiment result in Sect. 2, we can only expect the current proposed method to perform effectively in the case of speech and diffuse noise mixture.

4.4 Summary

In this section, I propose a modified blind noise suppression and dereverberation method based on frequency-domain BSE. In this method, BSE is used to extract speech component and suppress the diffuse background noise. Then, a nonlinear postprocessing is applied to suppress the late reverberation, with the parameter optimized according to the acoustic likelihood. Experimental result confirms the effectiveness and stability of the proposed method, particularly for the BSE combined with multichannel MOSIE.

The BSE-based proposed methods perform well in the mixture of speech and diffuse background noise, under assumption that only one speaker, i.e. target speaker, is active at a time. However, there are many situations in which such assumption does not hold. As pointed out in Sect. 2, BSE performance is prone to the presence of point-source interference, for example the overlapping speech from other speaker. The poor performance of BSE in noise suppression stage will lead to the poor overall performance of the proposed method, as conventional BSE is non-optimizable. These problems will be addressed in the next section.

5. Robust Noise Suppression and Dereverberation using Source-Adaptive BSS and Generalized MMSE-STSA

5.1 Introduction

In general, the conventional ICA-based BSS suffers from poor and slow convergence due to the fact that the simultaneous identification of statistical model of the sound source and the estimation of demixing matrix for source separation is a difficult task in an unsupervised optimization viewpoint. The utilization of fixed activation function, such as described in Sect. 2.3 aids in improving the performance. The use of BSE instead of BSS also reduces the computational complexity by only considering the difference of sparseness of the signal spectral modulus in the case of one target speech in the presence of diffuse background noise.

However, the above-mentioned approaches result in lost applicability to varying type of sound signals. Thus, it is preferable to develop a strategy for efficient estimation of the statistical model of each component and then building the corresponding activation function for the demixing matrix update. Authors in [36] has proposed a time-domain ICA method with such flexible activation function, however it can only perform well for additive mixture of signals.

In this section, I propose the new method of BSS based on source-adaptive ICA in the frequency domain. The activation functions for updating the separation filter correspond to the statistical model of each sound source spectrum. First, the main algorithm of the proposed method is explained. Then, I describe the estimation strategy for each parameter of source-adaptive BSS. Experiments are then carried out to evaluate the performance of the proposed method and highlight some remarks on the implementation.

5.2 Main Algorithm

Many signal processing methods apply different kinds of statistical models assumption to different variables in each operational domain. For example, conventional ICA-based BSS introduces the tangent hyperbolic or sigmoid activation function that corresponds to the super-Gaussian-distributed sound source PDF in the time-frequency (spectral) domain. MMSE-STSA estimator assumes that the speech spectral amplitude follows Gaussian distribution, and MOSIE utilizes chi-distributed speech spectral amplitude prior model.

The proposed method in Sect. 4 has been shown to perform optimally for the case of speech in the presence of diffuse background noise, owing to the utilization of BSE to extract the speech components. The simple cost function within frequency domain BSE exploits the difference of statistical properties of each sound source, i.e. the sparseness of its modulus. However, the lacks of the statistical model assumption causes the performance of BSE to drop in the presence of interference speech, as the statistical properties of each sound source is similar. Therefore, it is of great interest to develop a BSS method that has flexible statistical model assumption according to the estimate of each sound source's PDF.

Figure 17 shows the main idea of the proposed method, which is the extension of the method proposed in Sect. 4. Here, we take advantage from the MOSIE postprocessing that works based on a statistical model assumption of speech, i.e. chi-distributed speech spectral amplitude prior. The shape parameter ρ of chi distribution, which is optimized to maximize the acoustic likelihood, can be utilized to obtain the optimum internal parameter of source-adaptive BSS.

Although the idea is promising, there is an inherent problem arised because BSS and MOSIE works in different domain. There is no known explicit relationship between the statistical model in amplitude domain (such as in MOSIE) and the statistical model in the complex spectral domain (such as in BSS). Thus development of the proposed method includes the following tasks:

- Approximation of the relationship between the parameter of speech statistical model in source-adaptive BSS, which is in complex spectral domain, and the shape parameter of chi-distributed speech amplitude prior in MOSIE.

- Estimation of the parameter corresponds to the remaining sound source, i.e. interference signal, in source-adaptive BSS.

- Integrating the source-adaptive BSS into the proposed method and building

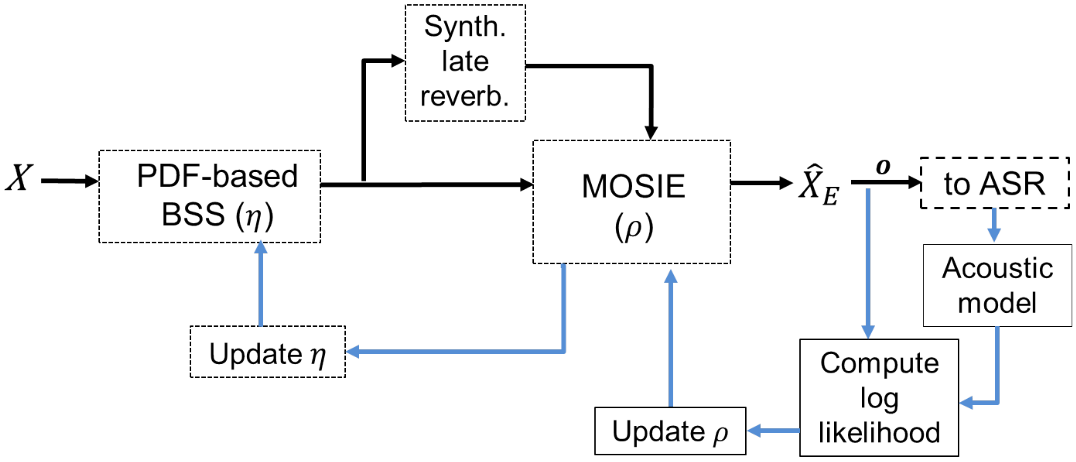

Figure 17. Block diagram of the optimized joint noise suppression and dereverberation combining source-adaptive BSS and MOSIE postprocessing.

optimization strategy.

If such tasks are well-performed, then the proposed method will bridge the difference of speech component modelling in three different domain, i.e. spectral domain in BSS, spectral amplitude domain in MOSIE, and mel-cepstral domain in ASR acoustic model. This relation is shown in Fig. 18.

5.3 Estimating the Parameter of Source-Adaptive BSS

In this method, the generalized Gaussian distribution is used to model the sound source PDF in the complex spectral domain. This distribution is selected amongst other distribution functions because it can represent various shape of distribution according the value of its parameter. The PDF for an arbitrary random variable z in the form of GGD function is given by

$$p_{GGD}(z;\vartheta,\eta) = \frac{\eta}{2\vartheta\Gamma(\frac{1}{\eta})}\exp\left(-\left[\frac{z-\bar{z}}{\vartheta}\right]^{\eta}\right), \tag{49}$$

where $\bar{z}$ is the mean of z, ϑ and η is the scaling parameter and shape parameter of GGD function, respectively. In this study, all the sound source are assumed to be zero-mean.

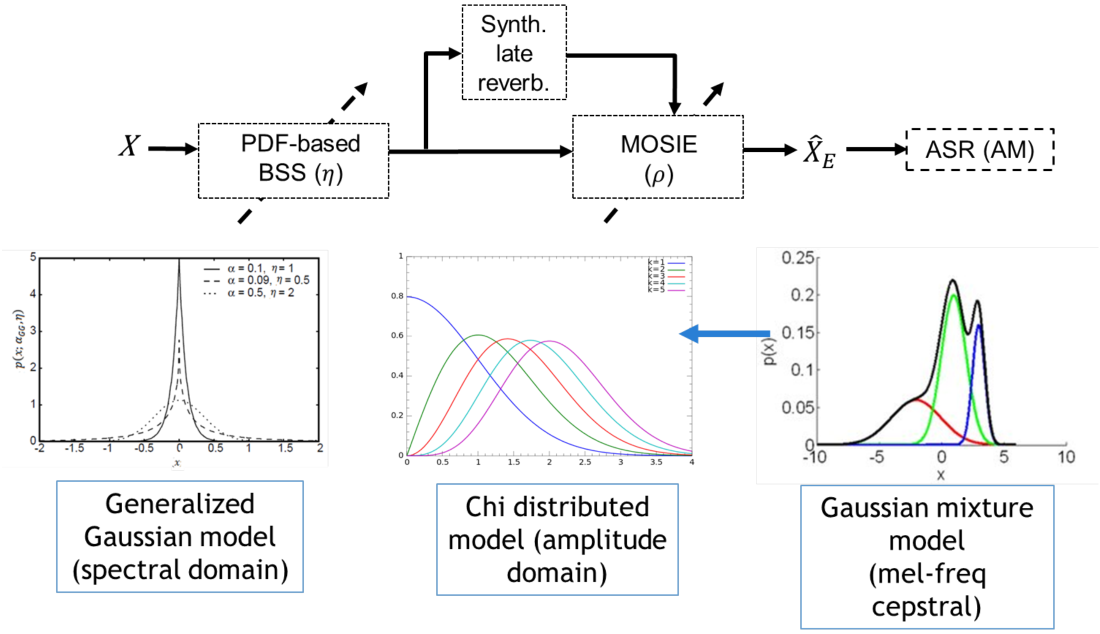
Synth. late reverb.
X
PDF-based BSS (η)
MOSIE (ρ)
$\hat{X}_E$
ASR (AM)
Generalized Gaussian model (spectral domain)
Chi distributed model (amplitude domain)
Gaussian mixture model (mel-freq cepstral)
$\alpha = 0.1,\ \eta = 1$
$\alpha = 0.09,\ \eta = 0.5$
$\alpha = 0.5,\ \eta = 2$
$p(x;\alpha_{GG},\eta)$
x
k=1
k=2
k=3
k=4
k=5
p(x)
x

Figure 18. Block diagram showing how the proposed joint method will connect different modelling of speech component in different domain.

Figure 19 depicts the various shape of GGD function according to different values of its shape parameter. It is shown that for $\eta = 2$, GGD function represents Gaussian distribution, which is commonly used to model the diffuse noise PDF. Also, for $\eta = 1$, GGD function represents Laplacian distribution, which is commonly used to model the speech PDF [37]. In general $\eta < 2$ indicates super-Gaussian distributed random variable, while $\eta > 2$ indicates sub-Gaussian distributed random variable.

5.3.1 Estimation of Speech PDF Shape Parameter

The relation between the statistical model of the target speech component in source-adaptive BSS and MOSIE is derived in this section. Recall that in MOSIE, the target signal amplitude spectrum is modeled as chi distribution by Eq. 16. It is known that chi distribution have a useful relation between the shape parameter

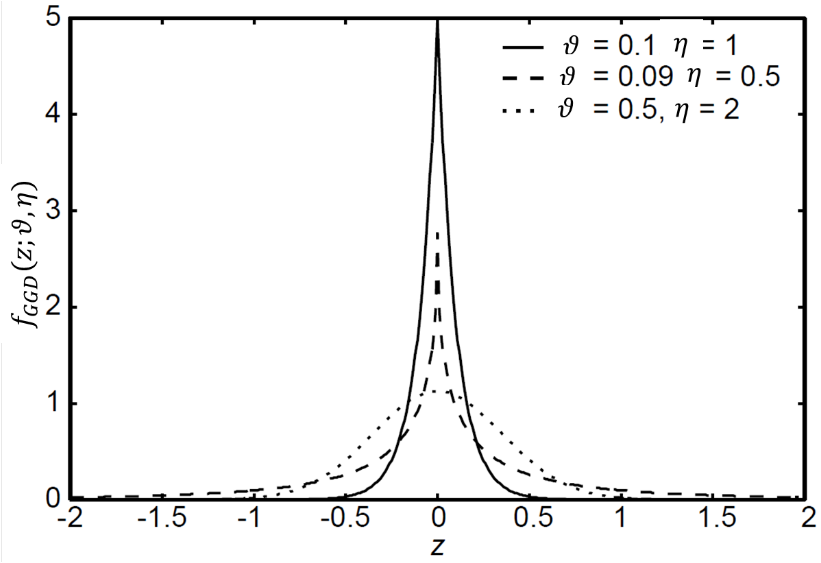

Figure 19. Sample of GGD functions for different values of ϑ and η.

ρ and the higher-order statistics, i.e. kurtosis, as given by

$$\rho = \frac{1}{\text{kurtosis}(s) - 1}. \tag{50}$$

The complex-valued speech variable in spectral domain is defined by $(s_R + js_I)$, where s_R and s_I are real and imaginary parts of the speech signal spectrum, respectively. Thus, Eq. (50) can be reformulated as

$$\frac{\mu_4(\sqrt{s_R^2 + s_I^2})}{\mu_2^2(\sqrt{s_R^2 + s_I^2})} = \rho^{-1} + 1, \tag{51}$$

where $\mu_l(s)$ is the lth-order moment of s, as defined by $\mu_l(s) = \mathrm{E}[s^l]$. The statistics of squared variable of s_R and s_I, respectively, are given by:

$$\mu_l(s_R^2) = \mu_{2l}(s_R), \tag{52}$$

$$\mu_l(s_I^2) = \mu_{2l}(s_I). \tag{53}$$

However, moments of summed random variables generally do not equal the sum of each random variable's moments. Therefore, cumulants, which hold the additivity property of additive variables, are introduced in the derivation using *moment-cumulant transformation* [38].

Given Eq. (52) and (53), the lth-order cumulant of power spectrum $s_R^2 + s_I^2$ can be calculated by

$$\begin{aligned}\kappa_l(s_R^2 + s_I^2) =&\kappa_l(s_R^2) + \kappa_l(s_I^2)\\ =&\sum_{\pi(l)}(-1)^{(|\pi(l)|-1)}(|\pi(l)|-1)! \prod_{B\in pi(l)} \mu_{|B|}(s_R^2)+\\ &\sum_{\pi(l)}(-1)^{(|\pi(l)|-1)}(|\pi(l)|-1)! \prod_{B\in pi(l)} \mu_{|B|}(s_I^2),\end{aligned} \tag{54}$$

and the lth-order moment of the power spectrum is given by

$$\mu_l(s_R^2 + s_I^2) = \sum_{\pi(l)} \prod_{B\in\pi(l)} \kappa_{|B|}(s_R^2 + s_I^2). \tag{55}$$

Furthermore, the lth-order moment of the amplitude of spectrum $\sqrt{s_R^2 + s_I^2}$ can be written as

$$\mu_l\left(\sqrt{s_R^2 + s_I^2}\right) = \mu_{\frac{l}{2}}(s_R^2 + s_I^2). \tag{56}$$

Finally, using Eq. (56), the resultant 2nd- and 4th-order moments of the speech amplitude can be estimated as

$$\mu_2\left(\sqrt{s_R^2 + s_I^2}\right) = 2\mu_2(s_R), \tag{57}$$

$$\mu_4\left(\sqrt{s_R^2 + s_I^2}\right) = 2\mu_4(s_R) + 2\mu_2^2(s_R), \tag{58}$$

under assumption that s_R and s_I are i.i.d to each other. Since both are modeled by GGD, the lth-order moment of each component corresponds to the shape parameter η_S through the following relationship:

$$\mu_l(s_R) = \vartheta^l \Gamma\left(\frac{l+1}{\eta_S}\right)\Gamma\left(\frac{1}{\eta_S}\right)^{-1}. \tag{59}$$

Thus, by incorporating Eq. (57), (58), and (59) into (50), we obtain the relation between shape parameter of chi-distributed spectral amplitude ρ and shape parameter of GGD-distributed spectrum η_S as given by

$$\Gamma\left(\frac{5}{\eta_S}\right)\Gamma\left(\frac{1}{\eta_S}\right)\Gamma\left(\frac{3}{\eta_S}\right)^{-2} = 2\rho^{-1} + 1. \tag{60}$$

The value of η_S from given rho can be obtained by calculating the inverse function of (60). However, such calculation is a difficult task and there is no exact closed-form solution w.r.t to η_S. Therefore, the closed-form derivation of η_S is approximated using modified Stirling's formula on gamma function [39],

$$\Gamma(z) \approx \sqrt{2\pi} \cdot \exp(-z) \cdot z^{z-0.5} \cdot \exp\left(\frac{1}{12z}\right). \tag{61}$$

In logarithmic scale, Eq. (61) is equal to

$$\log \Gamma(z) \approx \frac{1}{2}\log(2\pi) - z + (z-0.5)\log z + \frac{1}{12z}. \tag{62}$$

Then the relation in (60) can be approximated by

$$\begin{aligned} \log\Gamma\left(\frac{5}{\eta_S}\right) + \log\Gamma\left(\frac{1}{\eta_S}\right) - 2\log\Gamma\left(\frac{3}{\eta_S}\right) &\approx \frac{1}{\eta_S}\log\left(\frac{5^5}{3^6}\right) + \log\left(\frac{3}{\sqrt{5}}\right) + \eta_S \cdot \frac{2}{45} \\ &= \log(2\rho^{-1}+1). \end{aligned} \tag{63}$$

This results in quadratic equation

$$\eta_S^2 + \eta_S \frac{45}{2}\log\left(\frac{3}{\sqrt{5}(2\rho^{-1}+1)}\right) + \frac{45}{2}\log\left(\frac{5^5}{3^6}\right) = 0, \tag{64}$$

and closed-form estimate of η_S from given ρ can be obtained from

$$\begin{aligned} \eta_S = &-\frac{45}{4}\log\left(\frac{3}{\sqrt{5}(2\rho^{-1}+1)}\right) \\ &- \frac{1}{2}\sqrt{\frac{2025}{4}\log\left(\frac{3}{\sqrt{5}(2\rho^{-1}+1)}\right)^2 - 90\left(\frac{5^5}{3^6}\right)}. \end{aligned} \tag{65}$$

The comparison between the theoretical relation in Eq. (60) and the estimated value in Eq. (65) is depicted in Fig. 20. From the figure it is shown that the estimation is very good in consistency, particularly for $\eta < 2$ which is the range of shape parameter for super-Gaussian signal.

Next, an experiment is conducted to evaluate Eq. (65) when dealing with real data. I use 20 clean utterances (10 males and 10 females) in this experiment, with each utterance' length varies between 1.8 - 13 s. The estimated η_S from the proposed approach is compared to its true value which is estimated using the kurtosis of the real component of each speech spectrum.

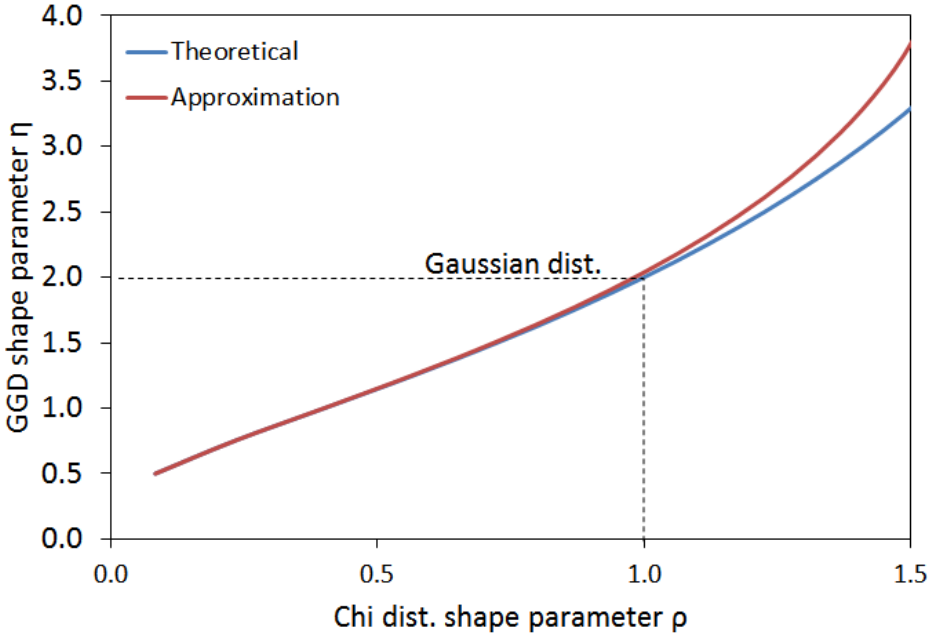

Figure 20. Relation between the shape parameter of Chi distribution ρ and the shape parameter of GGD η. The blue line shows the theoretical relation (60) and the red line is the proposed approximation (65).

The experiment result is shown in Fig. 21. We can see that the shape parameter of clean speech indicates the super-Gaussian nature of the signal. The estimated η_S is found to be always less than the true value, but the average estimation error does not exceed 5%. Thus, the proposed approach for the estimation of η_S is valid.

5.3.2 Estimation of Noise PDF shape Parameter

The approximated relationship between ρ and η only holds for the speech component, as MOSIE do not have statistical assumption on noise component. Therefore, the solution (65) cannot be applied to estimate η_N. Instead, the shape parameter of speech η_S is utilized to the estimation of η_N. This is done by estimating the PDF of the output of the source-adaptive BSS in which η_S has been optimized. The optimized η_S will provide the appropriate activation function corresponds to the speech component, thus improve the separation of speech signal

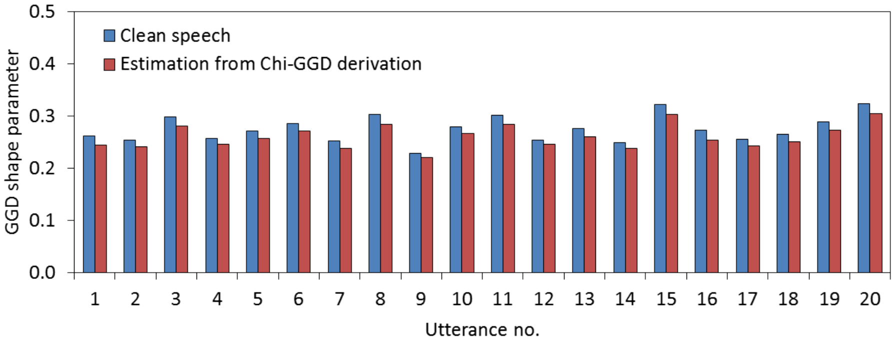

Figure 21. Estimated speech shape parameter corresponding to 20 utterances.

from the interferences. This is also mean that the residual output will contain less speech component, and the PDF of the residual output will be close to the real PDF of noise source. In practice, the shape parameter η_N can be estimated from the kurtosis of the output signal that corresponds to the noise component.

5.4 Source-Adaptive BSS

The previous PDF estimation enables BSS to use more appropriate activation function and mitigate the drawbacks on the poor convergence. Recall that the separation filter matrix is update as given in Eq. (9). Instead of using fixed tangent hyperbolic or sigmoid nonlinear function, the update rule is modified into

$$\boldsymbol{W}_{\mathrm{ICA}}^{[i+1]}(f) = \mu[\boldsymbol{I} - \langle \phi(\boldsymbol{Y}(f,k);\vartheta,\eta)\boldsymbol{Y}^{\mathrm{H}}(f,k)\rangle_k]\boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f) + \boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f), \tag{66}$$

where

$$\begin{aligned}\phi(\boldsymbol{Y}(f,k);\vartheta,\eta) =&[\phi(Y_1^{(R)}(f,k);\vartheta_1,\eta_1) + j\phi(Y_1^{(I)}(f,k);\vartheta_1,\eta_1),\\ &\cdots, Y_M^{(R)}(f,k);\vartheta_M,\eta_M) + j\phi(Y_M^{(I)}(f,k);\vartheta_M,\eta_M)]^T\end{aligned} \tag{67}$$

is the source PDF-adaptive activation function for each real and imaginary parts of the separated signal $[Y_1(f,k),\cdots,Y_M(f,k)]$ modeled in GGD function with parameters ϑ and η. The appropriate activation function is derived from (49) as given by

$$\phi(z;\vartheta\eta) = -\frac{\partial}{\partial z}\log(p_{GGD}(z;\vartheta,\eta)) = \begin{cases} \frac{\eta}{\vartheta^\eta}|z|^{\eta-1}, (z \geq 0), \\ -\frac{\eta}{\vartheta^\eta}|z|^{\eta-1}, (z > 0). \end{cases} \tag{68}$$

Assuming that the output of BSS is normalized to have unit variance, the scale parameter ϑ can be calculated as a function of η as given by

$$\vartheta = \sqrt{\frac{\Gamma(1/\eta)}{\Gamma(3/\eta)}}. \tag{69}$$

Figure 22 provides examples of activation functions for GGD with various shape parameter η. It is shown that although $\eta \leq 1$ closely resembles the distribution shape of speech component which is super-Gaussian distributed, the corresponding activation function is unstable, particularly for near-zero region. Therefore, there is a trade-off between correctly modeling the source PDF and maintaining good separation performance. In the proposed method, we will view this trade-off problem from the recognition performance.

5.5 Experimental Evaluation

Three experiments are conducted to evaluate the performance of the proposed method. The room impulse response from REVERB Challenge data are used in the experiment, with the specific information as described in Sect. 2. In the first experiment, I compare the source-adaptive BSS to conventional BSS with fixed tanh activation function and BSE. For the speech and diffuse noise condition, 20 clean utterances from JNAS corpus are convoluted with room impulse response and mixed with pre-recorded noise at input SNR of 10 dB.

For simulating the speech and point-source noise, 4 female and 4 male target speech are each mixed with interference speech convoluted with room impulse response. It is assumed that the target speaker always stays closer to the microphone than the interference speaker. Therefore, for near speaker distance, the interference speech comes from the far distance, and for far speaker distance, the interference speech comes from the same distance but has half the energy of the target speech signals. The performance of each method is evaluated using word recognition accuracy measure.

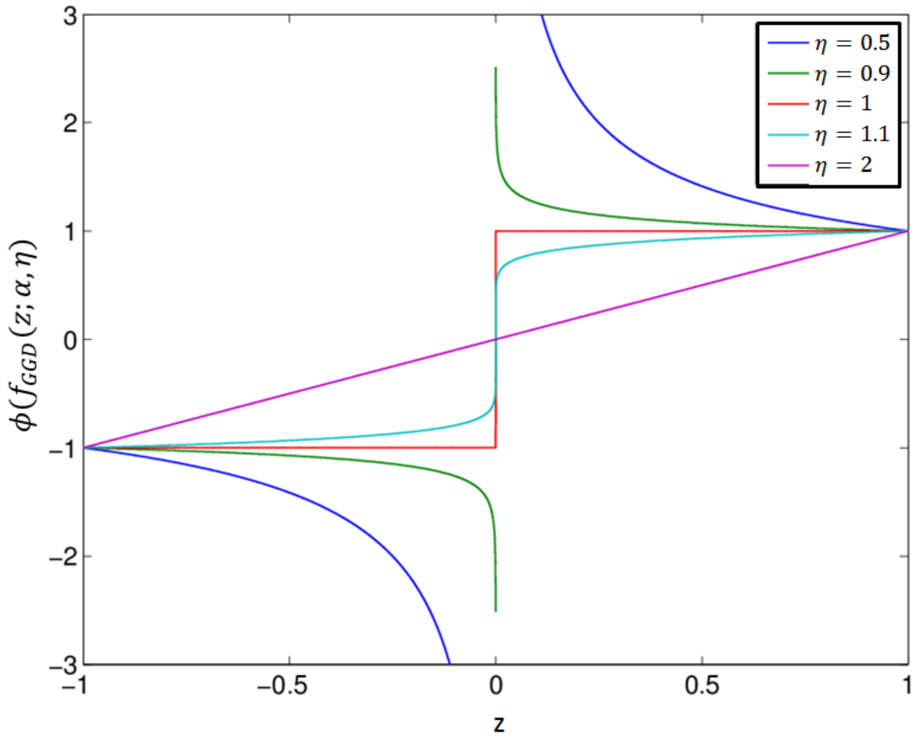

Figure 22. Samples of activation function corresponds to the various shape of GGD function.

In the second experiment, I evaluate the performance of the proposed joint method based on source-adaptive BSS in comparison to other BSS-based joint method. In this experiment, 20 clean utterances from JNAS corpus are convoluted with room impulse response and mixed with pre-recorded noise at input SNR of 10 dB. Using the same experimental conditions, additional experiment is conducted to investigate the effect of the incorrect information of T_{60} to the performance of the proposed method. This experiment is to ensure whether the proposed method can perform well blindly, with no proper *a priori* knowledge.

5.5.1 Evaluation of Source-Adaptive BSS

In this experiment, the performance of source-adaptive BSS (PDFBSS) is compared with conventional BSS with fixed tanh activation function (tanh-BSS) and frequency domain BSE (FD-BSE). To compensate the poor performance of conventional BSS in the case of diffuse background noise mixture, I also try to apply several iteration of BSE to provide initial unmixing matrix for BSS (BSE-BSS). The shape parameter η_S and η_N is estimated directly from the reference speech and reference noise, so only the algorithm of source-adaptive BSS is evaluated in this experiment.The estimation is carried out using kurtosis matching method in each frequency bin, and the mean of the obtained η is used throughout all frequency bins.

In Sect. 5.4 it has been shown that $\eta < 1$ yields the corresponding activation function of source-adaptive BSS that is unstable near the origin. Therefore, in practice, a certain limit η_{lim} is applied to the source-adaptive BSS, so that if the estimated η falls below the limit, the value will be replaced by η_{lim}. This process can be rationalized as the source-adaptive BSS is intended to act as a noise suppressor, hence the extracted speech should include the reverberation components, which causes the output to have more Gaussian-like distribution (larger η_S) according to central limit theorem.

The experiment results is depicted in Fig. 23. As expected, it is shown that in the case of speech and diffuse background noise, BSE outperforms the conventional BSS. On the other hands, BSE performance drops in the case of speech with interference speech. The performance of BSS also decreases in the significant presence of reverberation in room 2 and room 3. The BSE-based initialization helps improving the performance of BSS in the presence of diffuse background noise, but it deteriorates the performance in the case of interference speech.

In contrast, the proposed source-adaptive BSS is able to maintain the stable performance regardless the severity and the type of interference. In average, the proposed source-adaptive BSS outperforms conventional methods. We can also see that applying different η_{lim} results in varying performance, but in average, choosing the $\eta_{lim} = 1.05$ provides the best performance. The corresponding activation function of $\eta_{lim} = 1.05$ is similar to the tangent hyperbolic function in conventional BSS, but it provides more flexibility in the activity function corresponds to the noise component, thus the better performance. From the ASR point of view, this finding supports the hypothesis that recognition accuracy performance does not directly correlated to the fine modeling of speech waveform.

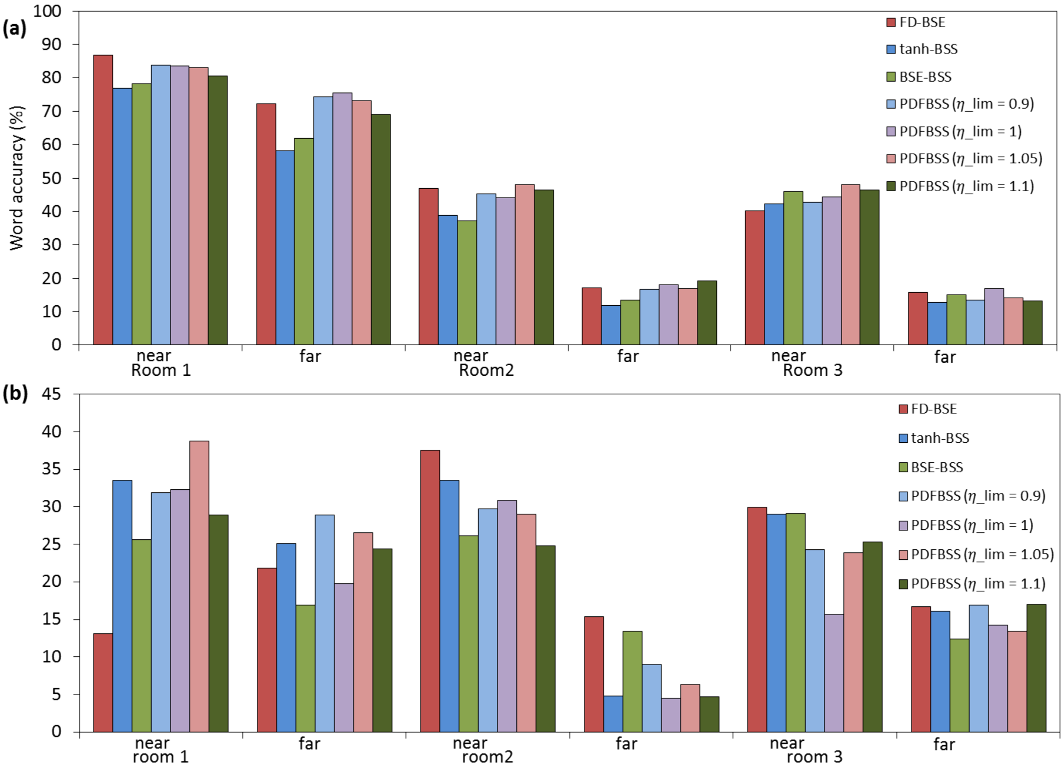

Figure 23. Performance comparison of source-adaptive BSS to conventional BSS and BSE, in (a) speech with diffuse background noise, and (b) speech with interference speech.

5.5.2 Evaluation of the Proposed Robust Joint Method

The first experiment confirms the robustness of the proposed source-adaptive BSS *provided* the true η value of each sound source. However, in the real environment implementation, such information is not accessible. Therefore, the source-adaptive BSS is combined with MOSIE postprocessing to build a robust joint method. The optimum shape parameter η_S, η_N and ρ are obtained following these procedures:

Step 0: First, set initial η_S and η_N to $\eta_{lim} = 1.05$ according to the previous experiment.

Step 1: Next, run several iteration of initial source-adaptive BSS using initial η.

Step 2: Estimate the η_S and η_N from the output of initial source-adaptive BSS

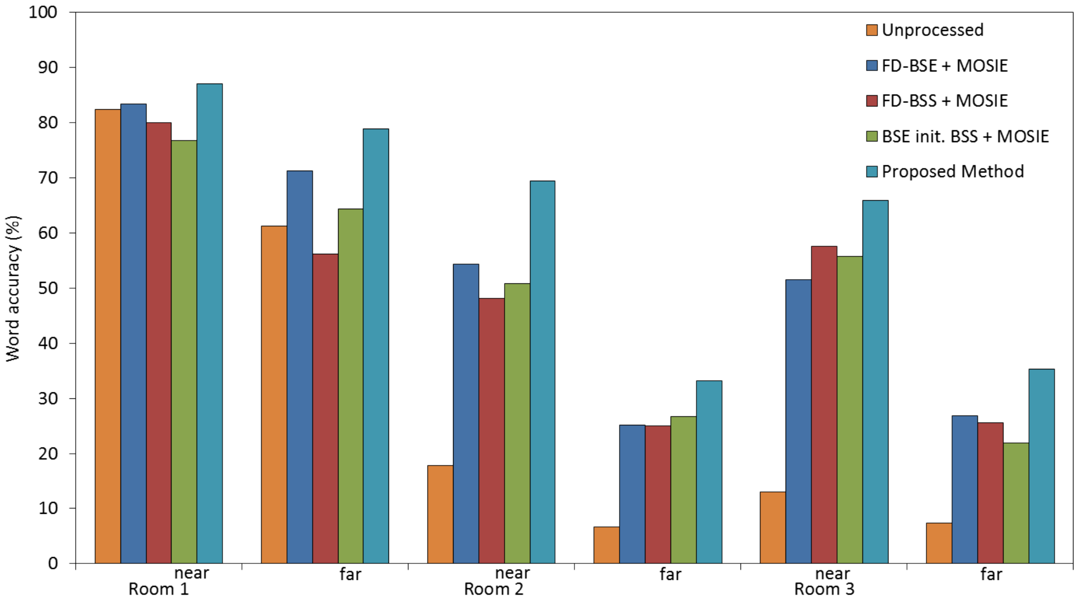

Figure 24. Performance comparison of proposed joint method based on source-adaptive BSS and other joint method.

according to the estimation strategy in Sect. 5.3.2, by applying $\eta_{lim} = 1.05$. Run the rest of source-adaptive BSS iteration.

Step 3: Apply MOSIE postprocessing to the output of source-adaptive BSS. Optimize the shape parameter ρ according to the acoustic likelihood as described in Sect. 3.

Step 4: Derive the optimum η_S from ρ using (65). Then run several iteration of source-adaptive BSS to re-estimate η_N.

Step 5: Apply the source-adaptive BSS with optimized η value to the observed signal, then apply MOSIE to obtained the desired output. Optimize ρ according to the acoustic likelihood.

In this experiment, single channel MOSIE is applied as postprocessing. I compare the performance of the proposed method with the BSE+ MOSIE joint method and BSS+MOSIE joint method. The experiment results is depicted in Fig. 24. From the result, it is shown that the proposed joint method outperforms other method in all conditions. The previously proposed BSE+MOSIE joint method fails to perform effectively particularly where the reverberation is significant (case 'room 2' and 'room 3'), most likely because the sparseness difference between speech and noise component is not as distinct as in the case of low reverberation. The proposed source-adaptive BSS method also benefits from the joint optimization of both source-adaptive BSS in noise suppression stage and MOSIE in dereverberation stage, in contrast to the BSE-based joint method that is only optimized in dereverberation stage.

5.5.3 Evaluation of the Importance of T_{60} information

Throughout this dissertation, it is assumed that T_{60} used in dereverberation stage is accessible. This seems contradictory to the name blind in the proposed method. Therefore, I conducted additional experiment to investigate the effect of incorrect T_{60} value to the performance of the proposed method. In this experiment, T_{60} is set to 500 ms for all rooms. It means that the value is overestimated for 'room 1' case, and underestimated for 'room 2' and 'room 3'. The performance of the proposed method is compared to the BSE+MOSIE joint method and the proposed method with correct T_{60}.

The experimental result is depicted in Fig. 25. It is shown that the use of incorrect value caused the performance of the proposed method to drop. However, it still outperforms the BSE+MOSIE joint method in general. The optimization of source-adaptive BSS parameters may have provided a compensation for the mismatched T_{60}. From the result, it is safe to assume that the proposed method can cope with the lack of T_{60} information to some extent, although further research may be required to improve its robustness.

5.6 Summary

In this section I propose a novel frequency-domain BSS algorithm that is adaptive to the PDF of each sound source. This source-adaptive BSS is combined with MOSIE postprocessing, and the parameters in each part are jointly optimized to maximize the acoustic likelihood to the acoustic model in speech recognizer. I also present the approximated relation of the shape parameter of different statistical model in BSS and MOSIE.

The resulting algorithm connects the statistical model of speech in three differ-

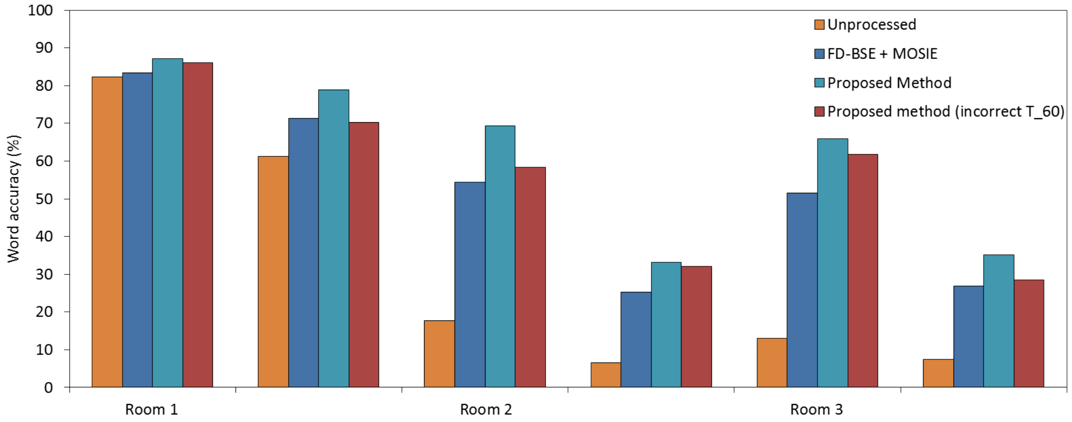

Figure 25. Effect of the value of T_{60} to the recognition performance of the proposed method.

ent domain, i.e. spectral domain, amplitude domain, and mel-frequency cepstral domain, and utilizes these models to improve the speech recognition accuracy. The experimental result confirms the robustness of the proposed method for any type of interference. Further improvement may be achieved by implementing multichannel MOSIE instead of single channel MOSIE, under consideration that the computational complexity will also increase.

6. Conclusion

6.1 Summary of Dissertation

This study aims to develop a new microphone array processing to suppress the interferences, i.e. background noise and late reverberation, and to improve the speech recognition accuracy for the speech recognition system in distant-talking environment. Conventional microphone array processing attempt to do so by manipulating the captured signal according to waveform-level objective criterion. These approaches assume that enhancement of the target speech waveform quality will lead to increased recognition accuracy.

The work presented in this dissertation take a different approach by considering that the microphone array processing and the ASR system as single system. The proposed methods, also named as joint method, are based on BSS combined with nonlinear postprocessings. The algorithms in microphone array front-end are optimized with the objective of improved speech recognition accuracy. This is done by optimizing the parameter of the proposed joint method to maximize the likelihood of the output speech to the acoustic model in the ASR system.

A preliminary experiment in Sect. 2 shows the potentials and disadvantages of array processing method belongs to BSS family. While conventional ICA-based BSS performs optimally only in the case of target speech mixed with point-source interference, the BSE algorithms achieves best performance in the case of target speech mixed with diffuse background noise. Based on this findings, one of the main goal in this work is to develop a novel adaptive BSS algorithms that is robust to various type of interference.

In Sect. 3, I proposed a semi-blind joint method for the implementation in a robot dialogue system. The main algorithm of this method consists of BSE, which is the alternative of BSS specifically designed for the mixture of speech and diffuse background noise, as noise estimator, combined with two stages of multichannel parametric WF as postprocessings. The proposed method utilizes the image information from robot's camera to decide whether to apply the postprocessings or not, assuming the position of the speaker correspond to the severity of interferences. Experimental evaluation confirm the robustness of the proposed method under varying level of background noise (diffuse) and reverberation, in comparison to some well-known competing method. However, this method can only be implemented along with the presence of image sensor.

To cope with the above-mentioned limitation, I proposed the modified BSE-based joint method which can perform without the aid of image information in Sect. 4. In contrast to the previous method, in this method BSE is utilized as speech extractor, and also simultaneously suppress the background noise. Thus, only one additional stage of postprocessing is required for dereverberation. In this section, I discussed the performance of the joint method with various nonlinear postprocessing, with the parameter of postprocessing is optimized to maximize the acoustic likelihood in the speech recognizer. Experimental evaluation shows that the joint method combining BSE and generalized MMSE-STSA estimator (MOSIE) outperforms other methods in terms of recognition accuracy. Another interesting finding is that the preferable value of forgetting factor in decision-directed-based SNR estimator in MOSIE is different from what is commonly used for the hearing purpose. This supports the hypothesis that the optimization criterion for speech recognition is different to that for human hearing purpose.

While these BSE-based joint methods have achieved quite successful improvement in recognition accuracy, both are limited in the implementation. The reason is the underlying assumption that the background noise contains only diffuse noise, or in other words, only one speaker is active at a time frame. Although this assumption holds for several acoustic conditions, there are many situations where more than one speaker exists and the overlapping speech is unavoidable. Moreover, the poor performance of BSE in noise suppression stage will lead to the poor overall performance of the proposed method, as conventional BSE is non-optimizable.

To extend the flexibility of joint method in the mixture of speech and various type of interference, I proposed a source-adaptive BSS and its implementation in joint method in Sect. 5. I take advantage from the provided statistical prior of clean speech amplitude in MOSIE postprocessing and derive an approximation of closed-form solution of the relationship between shape parameter of two different statistical model in different domain. The MOSIE shape parameter ρ that is optimized according to the acoustic likelihood is utilized to estimate the shape parameter of speech component η_S of GGD function used in source-adaptive BSS.

Furthermore, the shape parameter of noise component η_N is obtained from the estimated η_S through kurtosis comparison.

In this source-adaptive approach, I managed to connect the statistical model of speech in three different domain, i.e. spectral domain as in source-adaptive BSS, amplitude domain as in MOSIE, and mel-frequency cepstral domain as in acoustic model of speech recognizer. This configuration results in a robust joint method that maintain a stable performance regardless the type and degree of interferences. The proposed source-adaptive joint method can also handle the mismatched T_{60} information, though further works remain to be done to improve its robustness.

6.2 Direction for Future Works

The algorithms developed in this study have been successful at improving the recognition accuracy using microphone arrays. Nevertheless, the proposed method still leave plenty of room for improvement. In this section, I would like to give some remarks regarding the current work, in the hope it will give hints for further research.

One of the drawbacks of the proposed method is the inability to adapt to a moving source. The current algorithm performs well in utterance level, but cannot cope with a moving user during an utterance. Ideally, extending the proposed method in an online manner should solve the problem. An example of online algorithm of BSS has been developed in [40].

The parameter optimization in this dissertation are carried out using a simple line search. This is because there is no closed-form solution can be derived for the acoustic likelihood-based optimization. The future work may seek for the solution with stronger mathematical basis, such as by implementing gradient-based parameter update.

It is also worth to note that the improvement in speech recognition accuracy throughout this dissertation is solely the contribution from the proposed array processing techniques. Therefore, it is not surprising that under some conditions the recognition accuracy are very low. Higher accuracy rate shall be expected by combining the proposed array processing method with improvement in the speech recognizer part of the system, for example by incorporating the acoustic model adaptation and speech recognition compensation methods.

One of the well-known acoustic model adaptation method is maximum likelihood linear regression (MLLR) [41]. This method assumes that Gaussian means of the state distribution of HMM acoustic model representing noisy speech are related to that of clean speech by a linear regression function.The adaptation can be carried out in unsupervised manner on the test data itself. This method has been observed to perform well in many situations where interference condition is relatively stationary. However, as this method really depends on the amount of available adaptation data, a modification may be required to integrate the technique to the proposed joint method.

Acknowledgements

All praises belong to Allah, for all the guidance and blessings throughout the completion of this task. This research became possible because of the following institutions: the Japanese Government via the MEXT scholarship, and the Augmented Human Communication Laboratory of Graduate School of Information Science, Nara Institute of Science and Technology.

I would like to express my sincere gratitude to Professor Satoshi Nakamura of Nara Institute of Science and Technology as the advisor of my dissertation, for the valuable advices and constant encouragement through the period of this research. I have been very fortunate to be accepted as a doctoral student in his laboratory.

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

This work would not have been possible without the continuous support from Professor Hiroshi Saruwatari of The University of Tokyo. I am profoundly grateful for the countless guidances, encouragements, and advices on both technical and nontechnical issues he has given me during my study in Japan. It is a great experience to be able to conduct research under his supervision and learned many valuable aspects of being a researcher from him.

I would also like to express my appreciation to Professor Tomoki Toda of Nagoya University for the valuable suggestions on the source-adaptive BSS method, as well as Assistant Professor Sakriani Sakti, Assistant Professor Graham Neubig, Assistant Professor Koichiro Yoshino of Nara Institute of Science and Technology for their useful insights and comments on my research.

I would like to especially thank Dr. Tomoya Takatani from Toyota Motor Corp., for the fruitful discussion on the semi-blind method, and Dr. Jani Even from ATR, who taught me the concept of frequency-domain BSE in the clearest way possible.

Great thanks to Emeritus Professor Kiyohiro Shikano and Assistant Professor Hiromichi Kawanami for the support and the guidance since I was a research student at Nara Institute of Science and Technology.

I wish to express my deep gratitude for Dr. Ryoichi Miyazaki, for the fruitful

discussions and valuable supports related to this work. I wish him the best for his future career and life. Also my thanks to Mr. Kodai Okamoto, Mr. Yuji Onuma and Mr. Toshihide Aketo, for kindly helping me recording the impulse response database used in this research. I wish them the best of luck in their professional career.

Great thanks to all of the staff and member of the Augmented Human Communication Laboratory and the former Speech and Acoustics Processing Laboratory of NAIST, for the friendship, countless supports in daily life, and great working environment where I can share valuable discussions and talks on speech signal processing and speech recognition. I would especially like to express my gratitude to Mrs. Manami Matsuda and Mrs. Toshie Nobori for their kind help and support in all aspects of my research.

I also would like to especially thank all the past and current staff members of NAIST International Student Affair Section for the continuous assistance in both academic and daily life.

To all my friends in NAIST and Osaka, I am greatly thankful for abundant favors that made my whole experience living in Japan very exciting. My gratitude also to all my Indonesian friends who are always ready to support no matter the distance. My parents, brother, and sisters, who never forget me in their prayers.

Last but not the least, to my dear Adhitya Gandaryus Saputro, Hikari Dafira Nusayba, and Mamoru Sabar Ibrahim. I owe them beyond deepest gratitude for all the cheers, inspirations, and understanding, for keeping me going when things become difficult. I am eager to step into the next journey with them.

References

[1] M. Wolfel and J. McDonough. *Distant Speech Recognition.* John Wiley & Sons, West Sussex, UK, 2009.

[2] B. E. D. Kingsbury and N. Morgan. Recognizing reverberant speech with rasta-plp. In *Proc. ICASSP-97*, pages 1259–1262. IEEE, 1997.

[3] J. Benesty Y. Huang and J. Chen. Dereverberation. In J Benesty, M M Sondhi, and Y Huang, editors, *Handbook of Speech Processing*, pages 929–943. Springer, London, UK, 2008.

[4] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust. Speech, Signal Process.*, ASSP-27(2):113–120, Apr. 1979.

[5] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proc. IEEE*, 67(12):1586–1604, Dec. 1979.

[6] M. Miyoshi and Y. Kaneda. Inverse filtering of room acoustics. *IEEE Trans. Acoust. Speech Signal Process.*, 36:145–152, 1988.

[7] A. Schwarz, K. Reindl, and W. Kellermann. A two-channel reverberation suppression scheme based on blind signal separation and wiener filtering. In *Proc. ICASSP*, pages 113–116. IEEE, 2012.

[8] M. Delcroix, T. Hikichi, and M. Miyoshi. Dereverberation and denoising using multichannel linear prediction. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(6):1791–1801, aug 2007.

[9] J. S. Erkelens and R. Heusdens. Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments. *IEEE Trans. Audio, Speech, Lang. Process.*, 18(7):1746–1765, Sep. 2010.

[10] E. A. P. Habets and J. Benesty. Joint dereverberation and noise reduction using a two-stage beamforming approach. *Proc. HSCMA*, pages 192–195, 2011.

[11] M. L. Seltzer and R. M. Stern. Subband likelihood-maximizing beamforming for speech recognition in reverberant environments. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(6):2109–2121, 2006.

[12] R. Gomez and T. Kawahara. Robust speech recognition based on dereverberation parameter optimization using acoustic model likelihood. *IEEE Trans. Audio, Speech, Lang. Process.*, 18(7):1708–1716, 2010.

[13] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[14] S. Ikeda and N. Murata. A method of ICA in time-frequency domain. *Proc. ICA*, pages 365–371, 1999.

[15] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Neurocomputing*, 22(1–3):21–34, 1998.

[16] P Comon. Independent component analysis, a new concept? *Signal Process.*, 36:287–314, 1994.

[17] A. Hyvrinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13:411–430, 2000.

[18] A. Hyvrinen. One-unit contrast functions for independent component analysis: A statistical analysis. *Proc. IEEE Neural Networks for Signal Processing Workshop*, pages 388–397, 1997.

[19] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano. Blind spatial subtraction array for speech enhancement in noisy environment. *IEEE Trans. Audio, Speech, Lang. Process.*, 17(4):650–664, 2009.

[20] J. Even, H. Saruwatari, and K. Shikano. Blind signal extraction based speech enhancement in presence of diffuse background noise. *Proc. IEEE Workshop on Statistical Signal Processing*, pages 513–516, 2009.

[21] P. C. Loizou. *Speech Enhancement: Theory and Practice.* CRC Press, Inc., Boca Raton, FL, USA, 2nd edition, 2013.

[22] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.*, 32(6):1109–1121, 1984.

[23] C. Breithaupt and R. Martin. Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient condition. *IEEE Trans. Audio Speech Lang. Process.*, 19(2):277–289, 2011.

[24] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.*, 33(2):443–445, 1985.

[25] K Ito, M Yamamoto, K Takeda, T Takezawa, T Matsuoka, T Kobayashi, K Shikano, and S Itahashi. JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research. *J. Acoust. Soc. Jpn*, 20:196–206, 1999.

[26] The REVERB challenge - evaluating dereverberation and ASR techniques in reverberant environments.

[27] J Benesty, J Chen, and Y Huang. *Microphone Array Signal Processing.* Springer-Verlag, Berlin, Germany, 2008.

[28] A. Lee, T. Kawahara, and K. Shikano. Julius -an open source realtime large vocabulary recognition engine. *Proc. European Conference on Speech Communication Technology*, pages 1691–1694, 2001.

[29] J. Even, H. Saruwatari, K. Shikano, and T. Takatani. Blind signal extraction based joint suppression of diffuse background noise and late reverberation. *Proc. EUSIPCO*, pages 1534–1538, 2010.

[30] H Saruwatari, Y Ishikawa, Y Takahashi, T Inoue, K Shikano, and K Kondo. Musical noise controllable algorithm of channelwise spectral subtraction and adaptive beamforming based on higher order statistics. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(6):1457–1466, 2011.

[31] K Lebart and J M Boucher. A new method based on spectral subtraction for speech dereverberation. *Acta Acoustica*, 87:359–356, 2001.

[32] S. Davis and P. Mermelstein. Comparison of parametric representation for monosyllablic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.*, 28:357–366, 1980.

[33] L Rabiner and B Juang. *Fundamentals of speech recognition.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[34] K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi. Suppression of late reverberation effect on speech signal using long-term multiplestep linear prediction. *IEEE Trans. Speech Audio Process.*, 17(4):534–545, 2009.

[35] N. Q. K. Duong, E. Vincent, and R. Gribonval. Under-determined reverberant audio source separation using a full-rank spatial covariance model. *IEEE Trans. Speech Audio Process.*, 18(7):1830–1840, 2010.

[36] S. Choi, A. Cichocki, and S. Amari. Flexible independent component analysis. *Journal of VLSI Signal Processing System for Signal, Image, and Video Technology*, 26(1-2):25–38, 2000.

[37] S. Gazor and W. Zhang. Speech probability distribution. *IEEE Signal Process. Letters*, 10(7):204–207, 2003.

[38] R. Wakisaka, H. Saruwatari, K. Shikano, and T. Takatani. Speech prior estimation for generalized minimum mean-square error short-time spectral amplitude estimator. *IEICE Trans. Fundamentals*, E95-A(2):591–595, 2012.

[39] P.J. Davis. Gamma function and related functions. In M. Abramowitz and I. Stegun, editors, *Handbook of Mathematical Functions*. Dover Publications, 1965.

[40] Hua Yang and Shun-ichi Amari. Adaptive online learning algorithms for blind separation: maximum entropy and minimum mutual information. *Neural computation*, 9(7):1457–1482, 1997.

[41] Christopher J Leggetter and Philip C Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language*, 9(2):171–185, 1995.

Appendix

A. Statistical Distribution of Diffuse Background Noise

In this appendix, I show the goodness-of-fit test for the noise spectrum component. I investigate the statistical distribution of real recorded background noise signal with each length is approximately 20 s. Using Kolmogorov-Smirnov (KS) test, I test the null hypothesis that the histogram of the noise spectrum and the approximated pdf of a Gaussian distribution come from populations with similar distribution.

The experiment result confirms that the KS test do not reject the null hypothesis at the default 5% significance level. Figure 26 shows an example of the histogram of the noise spectrum and the pdf of a Gaussian distribution corresponding to it.

B. Statistical Distribution of Speech

I conduct the goodness-of-fit test for the speech spectrum component to validate the use of generalized Gaussian distribution to model the speech spectral pdf in source-adaptive BSS. I use 20 clean utterances as observed data with each length is approximately 30 s. Using Kolmogorov-Smirnov (KS) test, I test the null hypothesis that the histogram of the speech spectrum and the approximated pdf of a generalized Gaussian distribution come from populations with similar distribution.

The experiment result confirms that the KS test do not reject the null hypothesis at the default 5% significance level. Figure 27 shows an example of the histogram of the speech spectrum and the pdf of a generalized Gaussian distribution corresponding to it. It is also confirmed that noise and speech have different statistical characteristics, thus this difference can be utilized for BSS activation function.

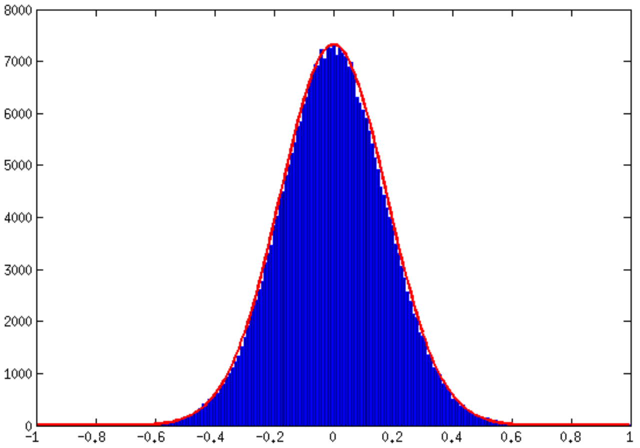

Figure 26. Histogram of the noise spectrum and the pdf of a Gaussian distribution corresponding to it.

C. Moment-Cumulant Transformation

Moments and cumulants of a statistical distribution function have unique relations in each order. These relations are very useful, for example for estimating the kurtosis of the speech amplitude from its complex-valued spectrum. Some formula regarding moment-cumulant transformation are derived in this section.

The probability distribution of random variable x is defined by the characteristic function $\phi_x(it)$, as given by

$$\phi_x(it) = \int_{-\infty}^{\infty} e^{itx} P(x) dx. \tag{70}$$

From this function, the lth-order moment $\mu_l(x)$ and cumulant $\kappa_l(x)$ of x, respectively, can be defined as

$$\mu_l(x) = \left. \frac{\partial^{(l)} \phi_x(it)}{\partial it^{(l)}} \right|_{t=0}, \tag{71}$$

$$\kappa_l(x) = \left. \frac{\partial^{(l)} \log \phi_x(it)}{\partial it^{(l)}} \right|_{t=0}. \tag{72}$$

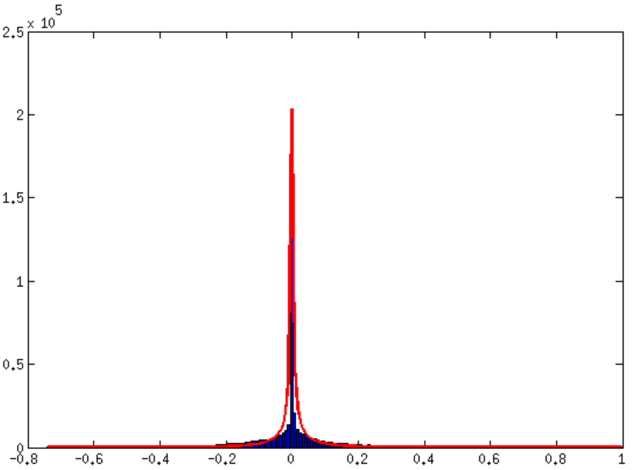

Figure 27. Histogram of the speech spectrum and the pdf of a generalized Gaussian distribution corresponding to it.

From Eq. 71, the lth-order moment $\mu_l(x)$ can be rewritten as

$$\begin{aligned}\mu_l(x) &= \left.\frac{\partial^{(l)} \exp(\log \phi_x(it))}{\partial it^{(l)}}\right|_{t=0} \\ &= \sum_{\pi(l)} \exp^{(|\pi(l)|)}(\log \phi_x(it)) \prod_{B\in\pi(l)} [\log \phi_x(it)]^{(|B|)}\Bigg|_{t=0} \\ &= \sum_{\pi(l)} \prod_{B\in\pi(l)} \kappa_{|B|}(x),\end{aligned} \tag{73}$$

using a combinatorial form of Faà di Bruno's formula,

$$\frac{\partial^{(l)} f(g(x))}{\partial x^{(l)}} = \sum_{\pi(l)} f^{(|\pi(l)|)}(g(x)) \prod_{B\in\pi(l)} [g(x)]^{(|B|)}. \tag{74}$$

The sum is over all partitions π of the set $\{1, \cdots, l\}$ and the product is over all of the blocks B of the partition π, and the number of members of B is denoted by $|B|$.

In the same manner, the lth-order cumulant $\kappa_l(x)$ in Eq. 72 can be rewritten

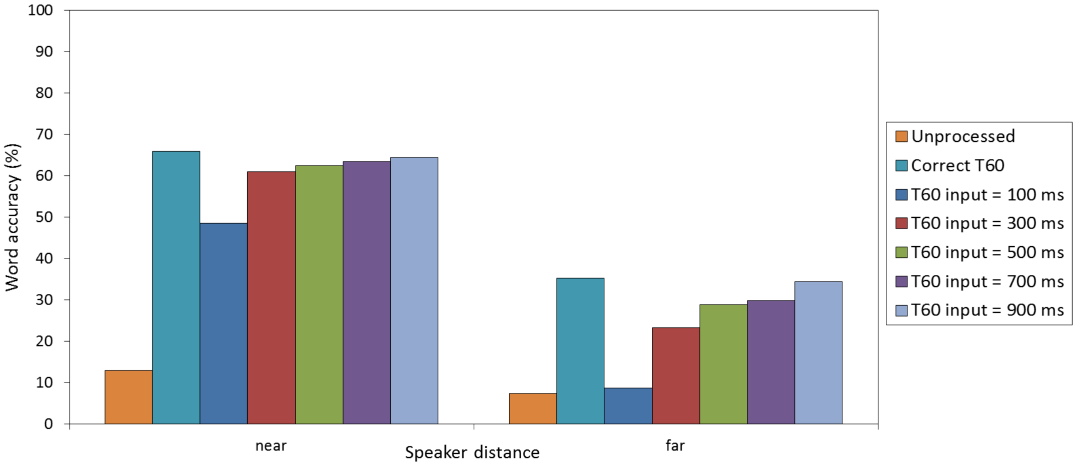

Figure 28. Effect of the value of T_{60} to the recognition performance of the proposed source-adaptive BSS joint method.

as

$$\begin{aligned}\kappa_l(x) &= \sum_{\pi(l)} \log^{(|\pi(l)|)}(\phi_x(it)) \prod_{B\in\pi(l)} [\phi_x(it)]^{(|B|)}\Big|_{t=0} \\ &= \sum_{\pi(l)} (-1)^{(|\pi(l)|-1)}(|\pi(l)|-1)! \prod_{B\in\pi(l)} \mu_{|B|}(x).\end{aligned} \tag{75}$$

The moment-cumulant transformations up to the 4th-order are given by

$$\begin{aligned}\mu_1 &= \kappa_1, \\ \mu_2 &= \kappa_1^2 + \kappa_2, \\ \mu_3 &= \kappa_1^3 + 3\kappa_1\kappa_2 + \kappa_3, \\ \mu_4 &= \kappa_1^4 + 6\kappa_1^2\kappa_2 + 3\kappa_2^2 + 4\kappa_1\kappa_3 + \kappa_4, \\ \\ \kappa_1 &= \mu_1, \\ \kappa_2 &= \mu_2 - \mu_1^2, \\ \kappa_3 &= 2\mu_1^3 - 3\mu_1\mu_2 + \mu_3, \\ \kappa_4 &= -6\mu_1^4 + 12\mu_1^2\mu_2 - 3\mu_2^2 - 4\mu_1\mu_3 + \mu_4.\end{aligned} \tag{76}$$

D. Effect of T_{60} Input Value to the Robustness of Joint Method Based on Source-Adaptive BSS

Additional experiment was conducted to further investigate the robustness of source-adaptive BSS based joint method against mismatched T_{60} input. I use impulse response from 'room 3' to simulate the observed data. For the experiment, I test the proposed method using several incorrect T_{60} values. The experiment result is depicted in Fig. 28.

From the result it can be seen that the proposed joint method is relatively robust to the mismatched T_{60} input, except when it is far underestimated as in the case of $T_{60} = 100$ ms. It is also shown that, for room with significant level of reverberation, it is better to use overestimated T_{60} input than the underestimated one.

List of Publications

Journal

1. Fine D. Aprilyanti, Jani Even, Hiroshi Saruwatari, Kiyohiro Shikano, Satoshi Nakamura, and Tomoya Takatani, "Suppression of noise and late reverberation based on blind signal extraction and Wiener filtering," *Acoustical Science and Technology*, vol. 36, no. 4, pp. 302–313, July 2015.

International Conferences (Peer-Reviewed)

1. Fine D. Aprilyanti, Hiroshi Saruwatari, Kiyohiro Shikano, Satoshi Nakamura, and Tomoya Takatani, "Semi-blind algorithm for joint noise suppression and dereverberation based on higher-order statistics and acoustic model likelihood," *Proceeding of APSIPA Annual Summit and Conference 2013*, October 2013.

2. Fine D. Aprilyanti, Hiroshi Saruwatari, Satoshi Nakamura, and Tomoya Takatani, "Optimized joint noise suppression and dereverberation based on blind signal extraction for hands-free speech recognition system," *Proceeding of Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA 2014)*, May 2014.

Domestic Meetings

1. Fine D. Aprilyanti, Hiroshi Saruwatari, Satoshi Nakamura, and Tomoya Takatani, "Joint noise suppression and dereverberation combining frequency-domain blind signal extraction and multichannel Wiener filter for hands-free spoken dialogue system," *The Meeting of ASJ*, September 2013.

2. Fine D. Aprilyanti, Hiroshi Saruwatari, and Satoshi Nakamura, "Joint suppression of background noise and late reverberation combining blind speech extraction and generalized MMSE-STSA estimator," *The Meeting of ASJ*, March 2014.

Technical Report

1. Fine D. Aprilyanti, Hiroshi Saruwatari, Kiyohiro Shikano, Satoshi Nakamura, and Tomoya Takatani, “Semi-blind optimization scheme of joint suppression of background noise and late reverberation,” *IEICE Technical Report*, EA2013-52, pp. 111–116, July 2013.