Bayesian Microphone Array Processing

Takuma OTSUKA

Abstract

This dissertation presents Bayesian models of microphone array processing for computational auditory scene analysis in multisource environments. In such environments where multiple sounds are inevitably observed at a time, the decomposition function that extracts constituent sound source signals from the observed mixture of audio signals is essential for robust auditory processing because most audio decoding algorithms, such as speech recognition and sound source classification, assume clean and isolated audio signals as an input. We develop microphone array techniques that provide three fundamental functions: sound source separation, localization, and removal of reverberation (also known as dereverberation) to cope with multiple sound sources in practical environments.

Microphone array processing must overcome the following three auditory uncertainties for achieving a robust decomposition function in real environments: (1) uncertainty in the number of sound sources, (2) reverberation in indoor environments, and (3) dynamic environments such as moving sound sources. These uncertainty issues have been partly addressed: (1) Sound source separation methods assuming unknown number of sources ignore the reverberation, which results in degraded separation performance. (2) Methods coping with sound source separation and dereverberation simultaneously are limited to the case where the microphones outnumber the sound sources. (3) As for microphone array processing in dynamic environments, the main topic has been focused on only the localization function.

We overcome these uncertainties by using Bayesian nonparametrics so that our models can have an infinitely extensible flexibility to express the data and deal with observed mixture signals containing any number of sound sources. When the number of sound sources is uncertain, the selection of model complexity to handle the mixture signal is problematic because the model should be flexible enough to explain the observed mixture signal. Our Bayesian

nonparametrics-based models can bypass this elaborate model selection depending on environments. Thus, a wide applicability is realized in our models.

This dissertation is organized as follows. Chapter 2 provides a literature review on sound source separation methods. This review clarifies that the aforementioned three auditory uncertainties are critical but unsolved issues in the literature.

Chapter 3 first presents a method for sound source separation and localization based on Bayesian nonparametrics. This method addresses the source number uncertainty. Experiments using simulated and recorded audio mixtures show that this model achieves state-of-the-art sound source separation quality.

Chapter 4 extends the method presented in Chapter 3 such that the dereverberation is simultaneously carried out for a robust separation performance in reverberant environments. We also apply Bayesian nonparametric modeling to the dereverberation function so as to handle an arbitrary number of sound sources. Experiment with the sound source separation task using mixtures signals of various number of sound sources demonstrates that (1) our method is capable of separation and dereverberation of mixture signals where the number of sound sources is larger than that of microphones used to observe the mixture, and that (2) the source extraction performance is comparable to that of a state-of-the-art method suitable only for mixtures where the number of sources is less than the number of microphones.

Chapter 5 develops a Bayesian nonparametric infinite order autoregressive model for the dereverberation function. An appropriate order value of the autoregressive model should be determined depending on the reverberation time of the acoustic environment. While existing order determination methods choose an order value from a finite amount of candidate, our method infers a posterior belief over infinitely many order values, which can theoretically adapt to any reverberant environment. An experiment with the dereverberation task in three reverberant environments confirms the efficacy of our model compared to several order determination methods.

Chapters 6 and 7 tackle the third uncertainty of dynamic environments. Chapter 6 uses the method explained in Chapter 3 to separate sound sources in a dynamic environment, where the sound sources or the microphone array move over time. The method first simultaneously separates and splits the input mixture signal along the time axis, where each separated and split

segment can be considered a source signal arriving from a stationary direction. Then, the split segments of source signals are merged to reconstruct a certain moving sound source signal. The capability of this approach is demonstrated in two environments using a microphone array embedded in a mobile robot.

In Chapter 7, we develop a Bayesian sound source localization method robust against environment-dependent aspects, such as reverberation time and changes in the number of active sound sources over time. While existing localization method requires an environment-dependent tuning of the threshold for detecting sound sources, our method robustly estimates the threshold using the observed data in an unsupervised manner. Experimental results demonstrate our method robustly localizes multiple sound sources in a reverberant environment.

Chapter 8 discusses the contributions of this dissertation with some remarks on future work. We conclude this dissertation in Chapter 9.

Acknowledgments

Many people have supported me while I was engaged in this study. Without their help, this dissertation would not have been completed.

First of all, I wish to express my deepest gratitude to my advisor Professor Hiroshi G. Okuno for his supervision. His insightful and constructive comments for my research and dedicated commitment to his students helped me carry out this work. As a student of Okuno laboratory, I fortunately had many opportunities and experiences thanks to his generosity.

I am also grateful to the members of my dissertation committee, Professor Tatsuya Kawahara, Professor Marco Cuturi, and Dr. Kazuyoshi Yoshii. They gave me valuable comments and helpful suggestions for the improvement of this dissertation.

A large part of the work presented in this dissertation were conducted jointly with prominent researchers at NTT Communication Science Laboratories. In particular, I am indebted to the collaborators, Dr. Katsuhiko Ishiguro, Dr. Hiroshi Sawada (now with NTT Service Evolution Laboratories), and Dr. Takuya Yoshioka, for insightful advice and thorough comments for my research papers. With their knowledge about Bayesian models and multichannel or dereverberation techniques, the quality of the research is greatly advanced.

As a member of HARK developing team, I experienced various outreach activities of the outcome of the research. I am convinced this experience would be substantial to further carry on research projects in the future. I would like to thank Professor Kazuhiro Nakadai at Honda Research Institute Japan for inviting me to the HARK team and giving me meaningful opportunities such as a visit to other research institutes and an experience as a lecturer of HARK tutorials.

The joint work on the field research of frog behaviors provided a wider perspective on various research areas. I am obliged to Dr. Ikkyu Aihara at RIKEN Brain Science Institute for

his invitation to the field work.

I wish to deeply thank the past and present members of Okuno Laboratory, especially Professor Tetsuya Ogata (now with Waseda University), Professor Kazunori Komatani (now with Nagoya University), Dr. Ryu Takeda (now with Central Research Laboratory, Hitachi, Ltd.), and Dr. Takeshi Mizumoto (now with Honda Research Institute Japan). I received a great deal of inspiration from discussions with these members.

I sincerely thank my parents for their support and encouragement for my long student life.

Finally, I would like to thank my wife, Kyoko, for her devoted support and patience with our postponed honeymoon trip.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Auditory sense is essential for understanding a scene. For example, some species of bats and dolphins use audio signals to detect surrounding objects, which is known as echolocation (Griffin 1958, Evans 1973). Humans also exert the auditory sense for various purposes such as verbal communications and an awareness of surrounding events by sometimes using artificial auditory signs like a phone ring and a siren. Thanks to the capability of perception of the scene including the auditory sense, we can take an adaptive action in various environments.

Over the last decade, much progress has been made for computational scene understanding such as surveillance systems (Hu et al. 2004) and automated robots (Thrun 2002). These systems are well-appreciated since computers and robots can deal with tasks that are almost impossible or costly for humans. For example, an automated surveillance system can monitor a street all day and night, an automatic-driving car may achieve a continuous transportation (Thrun 2010), or a robot may probe hazardous places that humans cannot reach (Weisbin and Rodriguez 2000). A key to successful computational scene understanding using various sensor inputs is to specify a certain task and to develop an algorithm for the task. This is because what computers do is after all algorithmic numerical manipulations of data. For example, one of the tasks for an automated car is to find a path that minimizes the distance between the current position and the goal while avoiding obstacles in its way using the sensors embedded on the car. In the development of computational algorithms, probabilistic models, optimization problems, automatic acquisition of numerical representation have widely used (Murphy 2012, Bengio 2009). These approaches are useful for developing feasible algorithms thanks to the abstraction of the task and sensor data into mathematical procedures.

1.1 Computational Auditory Scene Analysis

Computational auditory scene analysis (CASA) (Rosenthal and Okuno 1998, Wang and Brown 2006), recently also known as machine listening (Vincent et al. 2013), tackles an auditory aspect of scene understanding for computers. This dissertation aims at a development of CASA system that can work in real environments. The motivations are twofold:

1. realization of auditory functions for computers and robots, and
2. enhancement of auditory intelligibility of humans.

Figure 1.1 outlines our envisioned system. The primary challenge is that real environments often contain multiple sound sources. Therefore, an observation of a certain sound source is interfered by other sound sources. To realize practical CASA systems, the decomposition of the multisource observation audio signal into the constituent sound source signals is essential. This is because most high-level auditory processing algorithms, such as speech recognition and sound source classification, typically assume clean audio signals. If the observed signal contains the interference of other sound source signals, the performance of these auditory decodings is degraded.

The decomposition function is beneficial to the enhancement of human auditory intelligibility as well. Humans cope with multisource environments by the selective attention to a certain sound source, which is known as the cocktail party effect (Cherry 1953, Bronkhorst 2000). In the selective attention mechanism, humans are capable of understanding a part of sounds of interest, typically a single source, but the rest are ignored (Wood 1990). The computational decomposition function in multisource environments can facilitate our understanding of respective sound sources.

We pursue the machine listening function that satisfies the following two requirements:

1. less a priori knowledge about auditory environments and
2. ability for spatial interpretation of sound sources.

First, auditory environments have a wide variety of sound sources and acoustic characteristics: multiple humans may talk together with some music in an indoor situation or wild animals call and song in an outdoor place. An ideal decomposition function should cope with

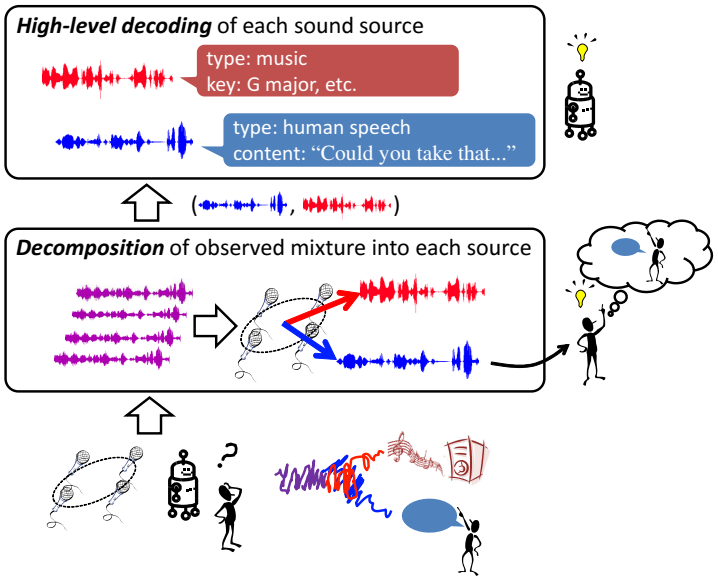


Figure 1.1: Machine listening for enhancing human auditory intelligibility and auditory functions on computers. Observed mixture in a multisource environment is first decomposed into constituent sound source signals. This decomposition contributes to auditory intelligibility of humans. Then, for each decomposed source, high-level audio decoding for information extraction and decision making of system behavior.

this acoustic variety without a priori knowledge about the environment. The second requirement refers to the necessity of the localization of sound sources. Some applications, such as probing robots (Sasaki et al. 2009a) or telepresence robots driven by a remote human operator (Mizumoto et al. 2011), require an ability of spatial understanding of the sound sources in the environment.

As shown in the bottom of Figure 1.1, we use a microphone array that consists of multiple microphones to achieve the auditory decomposition function that satisfies the requirements (Benesty et al. 2008). The above-mentioned two requirements motivate us to use microphone array processing, also referred to as a multichannel method, for two reasons. First, multichannel methods are applicable to various types of sound sources. Therefore, multichannel methods can widely be used when the types of sound sources in the environment are unknown. Second, the localization of sound sources can be carried out with multiple microphones to realize the spatial interpretation of the auditory scene. Microphone array processing provides three functions for CASA to cope with multisource environments:

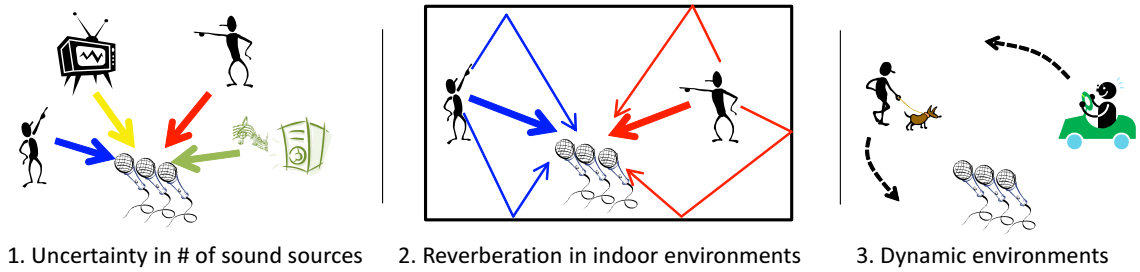


Figure 1.2: Three auditory uncertainties tackled by this dissertation.

(i) sound source separation,
(ii) sound source localization, and
(iii) removal of reverberation, also referred to as dereverberation.

The sound source separation function is the core of the auditory decomposition: from the observed mixture of sound signals, this function extracts the constituent sound signals. The sound source localization enables the spatial interpretation by estimating the direction of arrival for each sound source. This function satisfies the second requirement above. Reverberation is typically observed in an indoor environment and caused by the reflections of audio signals on the walls and floor. Reverberation impairs the auditory intelligibility as well as degrade the separation performance using a microphone array processing. Therefore, dereverberation is also an important function provided by microphone array techniques.

1.2 Technical Issues: Auditory Uncertainties

Although various progress has been made in the literature, microphone array processing confronts a challenging problem with regard to the first requirement–the auditory uncertainties in the environment. These uncertainties include

1. uncertainty in the number of sound sources,
2. reverberation, and
3. dynamic environments such as moving sound sources or microphones,

as depicted in Figure 1.2.

The uncertainty of the number of sources has been a critical parameter in the literature because most sound source separation algorithms assume that the number of sources in the observed mixture signal is known. Furthermore, different types of algorithms have been developed depending on the relationship between the number of sources and that of microphones. Under the uncertainty of the number of sound sources, we need a consistent algorithm that is applicable to any number of sources.

To alleviate the degradation of separation performance due to reverberation, dereverberation is often incorporated in microphone array techniques. Here, the uncertainty of the number of sound sources is again a problem for the dereverberation function because existing approaches assume that the source number is less than the number of microphones. However, this assumption may be violated because the sound sources can possibly outnumber the microphones in the source number uncertainty. Another uncertainty about reverberation is the amount of reverberation. The dereverberation model should adapt to the actual amount of reverberation for the improved dereverberation performance.

A dynamic environment in the sense that the sound sources are moving is problematic for microphone array-based sound source separation methods because most source separation methods assume that the source positions as well as microphone positions are static. Taking into account that moving sound sources can be observed in a natural scene, we need a separation method that is applicable to dynamic environments.

1.3 Benefits from Bayesian Modeling

This dissertation make the best use of Bayesian models for microphone array processing to overcome auditory uncertainties. The following two properties of Bayesian modeling are suitable points for our problem.

First, some sorts of prior distributions in Bayesian models can be regarded as regularization terms from the viewpoint of optimization problems (Bishop 2003, Fraley and Raftery 2007), when point estimates are computed from posterior distributions. The prior distributions often prevents latent parameters from overfitting the observed data, which stabilizes the parameter inference and results in better performance. For example, our Bayesian formulation of dereverberation uses a prior distribution corresponding to an anechoic situation. Given a

reverberant observation signal, the reverberation signal component is estimated through the posterior inference. In this way, our Bayesian dereverberation prevents the direct signal components from being regarded as reverberation.

Second, Bayesian nonparametric models (Ferguson 1973, Antoniak 1974, Gershman and Blei 2012) provide an infinitely extensible flexibility of the model complexity to explain the observed data. As we explain in the subsequent chapters, microphone array-based methods, such as sound source separation and dereverberation, are related to the selection of model complexity depending on the number of sound sources in the observed mixture signal. For example, our sound source separation method is formulated as a clustering of observed data points where each cluster corresponds to each sound source. Note that the model complexity, the number of clusters in this case, should suit the observation: a sufficiently large number of clusters should be prepared to cope with the sound source separation of unknown number of sound sources in the observed signal. Some model selection procedure may be necessary to adapt the model to various observed signals with different number of sound sources. The strength of Bayesian nonparametric models is explained as follows: they provides an infinitely many number of clusters so that an elaborate model selection can be bypassed thanks to infinitely many clusters capable of handling any number of sound sources. These benefits motivates us to develop a Bayesian microphone array processing.

1.4 Organization

The organization of this dissertation is outlined in Figure 1.3. In the following chapters, we address each of the three auditory uncertainties described in Section 1.2.

Chapter 2 provides a literature review to clarify that the three auditory uncertainties are still open problems. The review includes the comparison of separation methods with a single-microphone observation and microphone array-based methods, multichannel dereverberation methods, and microphone array techniques for dynamic environments.

Chapter 3 presents a Bayesian nonparametric microphone array processing for sound source separation and localization. This method tackles the first uncertainty–the unknown number of sources.

Chapter 4 presents a multichannel source separation and dereverberation method based on

Bayesian nonparametrics. This method is an extension of the method presented in Chapter 3 so that the separation performance can be robust against the reverberation. This chapter deals with the uncertainties about the number of sources and reverberation at the same time by presenting a method that is consistently applicable to any number of sound sources.

In Chapter 5, we discuss an adaptive fitting of dereverberation model depending on the amount of reverberation, where the amount of reverberation is posed as an uncertainty. We develop a Bayesian nonparametric model called infinite order autoregressive process to realize an adaptive dereverberation method.

Chapters 6 and 7 tackle the third uncertainty of dynamic environments. In Chapter 6, we apply the Bayesian nonparametric sound source separation method presented in Chapter 3 so as to separate the audio signal of moving sound sources. Chapter 7 presents a sound source localization method for moving sound sources. We construct a Bayesian model for a robust localization against environment-dependent aspects.

Chapter 8 discusses the contributions of this dissertation with some remarks on future work. We conclude this dissertation in Chapter 9.

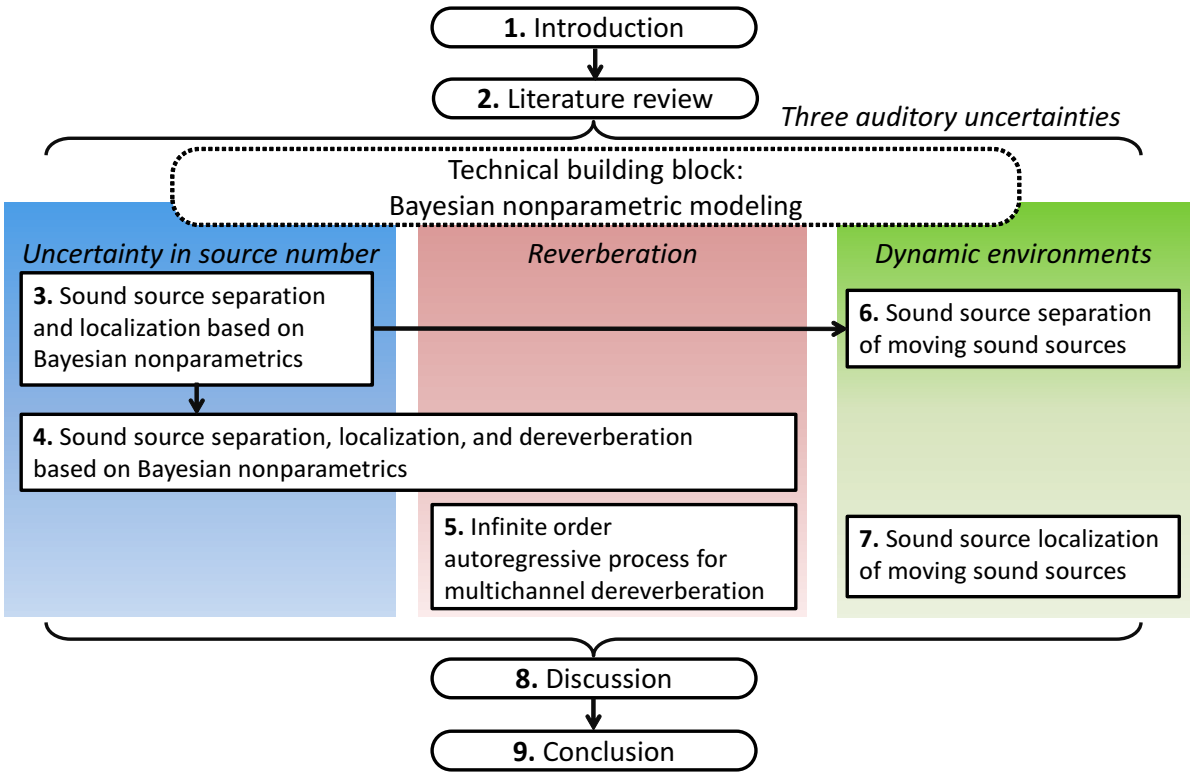


Figure 1.3: Organization of this dissertation.

Chapter 2

Literature Review

This chapter reviews existing methods for handling multisource environments in the literature. The review includes methods for noise reduction, sound source separation, reduction of reverberation (referred to as dereverberation), and tracking of moving sound sources. Through this review, we reveal that the following issues are critical:

1. uncertainty in the number of sources,
2. performance degradation of source separation by reverberation, and
3. dynamic environment such as movements of sound sources or microphones.

Sound source separation methods are divided into two categories as to whether the mixture of sound sources is observed by a single microphone (single channel methods) or by a microphone array (multichannel methods). In the following, single channel methods are outlined with some remarks on their limitations. Then, we review existing multichannel methods for sound source separation, dereverberation, and dynamic environments.

2.1 Single Channel Sound Source Separation Methods

Single channel methods are advantageous since they can be implemented with a simpler hardware architecture with only one microphone in comparison with multichannel methods. Early work on single channel noise reduction includes the Wiener filter (Wiener 1949), and spectral subtractions (Boll 1979, Martin 2001). These methods assume that the noise to be suppressed is generated from a stationary distribution. In other words, the statistical characteristics of the

interfering noise are time-invariant, or slowly change over time. Therefore, these methods are inapplicable to the separation of two or more dynamic signals such as speech signals.

Recent single channel source separation methods belong to two categories. Methods in the first category carry out a pre-training using isolated source signals so as to generate dictionaries or to train statistical models for the target sounds (Roweis 2000, Davies and James 2001, Jang and Lee 2003, Benaroya and Bimbot 2003, Pearlmutter and Zador 2004, Smaragdis et al. 2009). Some methods decompose the source signals for pre-training into fragment signals, the dictionary of target signals, using independent component analysis (ICA) (Davies and James 2001, Jang and Lee 2003), and reconstruct the constituent sources from the observed mixture using the fragments. Some methods train the parameters of statistical models such as hidden Markov models (HMMs) so that the constituent sources in the observed mixture signal can be restored by the statistical models (Roweis 2000, Benaroya and Bimbot 2003, Pearlmutter and Zador 2004). Smaragdis et al. develop a single channel source separation method that extracts the target signals using the pre-training audio signals themselves where a plenty amount of pre-training dataset is assumed to be available (Smaragdis et al. 2009). The limitation of these methods is that the coverage of the dataset in the pre-training phase is critical to the separability of a variety of sound sources. For example, if the pre-training data contains only speech signals, the resulting algorithm is only applicable to mixtures of speech signals.

Methods in the second category assume a stationary form or slow transitions of spectra over time in the frequency domain. Nonnegative matrix factorization (NMF) is often used to handle the spectral stationarity (Schmidt and Olsson 2006, Virtanen 2007, Ozerov et al. 2007, Hoffman et al. 2010a). While NMF is suitable for audio signals with little fluctuation in the spectral form such as instrumental sounds in music audio signals Kameoka et al. (2012a), it is difficult to apply NMF to the separation of multiple source signals that dynamically change the spectral forms over time such as speech signals.

2.1.1 Limitations of single channel methods

Though single channel methods are easily deployable with a simple observation device, they have two limitations:

1. limited applicability to various types of source signals, and

2. lack of spatial interpretation.

The first limitation has been explained above. The pre-training methods require an appropriate coverage of training dataset for specific applications whereas NMF-based methods assume the spectral stationarity. Regarding the second limitation, the lack of the spatial interpretation means that single channel methods give no idea about the location of the sound source in the environment. Localization of sound sources is an essential function for certain applications such as mobile probing robots (Sasaki et al. 2010).

2.2 Multichannel Sound Source Separation Methods

Multichannel methods that use a microphone array can overcome the aforementioned limitations: multichannel source separation methods make much milder assumptions on the sound source signals to extract and are capable of estimating the direction of arrivals of the constituent sound sources as a sound source localization function. This is because multichannel methods in principle analyzes the modulation of audio signals during the propagation of them instead of the characteristics of audio signals themselves. To be specific, multichannel methods use the difference in time of arrivals and in sound pressure levels of the signals observed at each microphone for the separation and localization. Therefore, multichannel methods are suitable to cope with multisource environment where the type of sound sources are unknown a priori, or the estimation of source location is necessary. One limitation about multichannel methods is the difficulty in separating source signals that come from the same direction because the time difference of arrivals between the microphones is identical when the sources arrive from the same direction. While we may need to combine a single channel separation method with a multichannel method to cope with this source overlap problem, the multichannel approach provides a general framework to cope with multisource situations with less assumptions on the type of sound sources.

In the following, we outline the existing multichannel methods keeping in mind the three uncertainties to overcome. Specifically, Section 2.2.1 reviews existing multichannel separation methods from the viewpoints of the relationship between the number of sources and the number of microphones, and the uncertainty in the source number. Section 2.2.2 presents dere-

verberation methods with a discussion on the combination with other multichannel processing such as a separation of sound sources. Section 2.2.3 outlines the microphone array-based methods for dynamic environments where the position of the sound sources or the microphone array may change over time.

2.2.1 Source separation in source number uncertainty

ICA is one of the most widely used models for multichannel sound source separation (Lee et al. 1999, Hyvärinen et al. 2001, Common and Jutten 2010, Sawada et al. 2007). This is because ICA is applicable to the separation of a wide range of sound sources and works well even when the propagation of the sound source signals to each microphone is unknown, which is highly dependent on the acoustic environment. Variants of ICA, called independent vector analysis (IVA), have also been developed (Lee et al. 2007, Ono 2011) for a unified modeling for the separation in the frequency domain to cope with the reverberation to a certain degree. These ICA and IVA-based multichannel source separation methods have two problems when the number of sources is uncertain. First, most methods assume that the number of sources is known in advance. Second, ICA and IVA are intractable when the sources outnumber the microphones, i.e., the mixture is underdetermined. This is problematic because underdetermined conditions may possibly occur in practical situations when the number of sources is uncertain. While the first problem is addressed in (Knowles and Ghahramani 2007, Nagira et al. 2013), the second problem remains as a limitation.

For the separation of more sources than microphones, time-frequency (TF) masking methods are helpful (Yilmaz and Rickard 2004, Mandel et al. 2010, Araki et al. 2010, Sawada et al. 2011). These methods extract the constituent source signals based on the assumption that at most one sound source is dominant at a time frame in a certain frequency bin since the power of each source signal in the TF domain is sparsely distributed. TF masking methods provide a consistent way to separate any number of sources because TF masking-based separation is applicable even when the source number is equal to or less than that of microphones.

Here, the uncertainty in the number of sources arises as a problem again. In order to separate the sound sources, at least the same number of TF masks as that of the sound sources are necessary. To cope with the source number uncertainty issue, some methods use a sufficiently

large number of TF masks (Araki et al. 2009, Loesch and Yang 2010, Taghia et al. 2012). In order to avoid the degradation of the separation performance cause by unused TF masks, these methods impose some regularization on the TF mask estimation algorithms to suppress unnecessary TF masks.

Indeed, the source number uncertainty is closely related to the selection of model complexity in that an appropriate choice of the number of TF masks is crucial for the separation performance. Since redundant TF masks may degrade the separation performance at the risk of splitting a single source signal into multiple TF masks, the TF mask number is preferably matched with the actual number of sources. While source number estimation methods are developed in the literature (Yamamoto et al. 2003, Loesch and Yang 2008), these methods lacks the general applicability since these source number estimation requires an environment-dependent training.

In the subsequent chapter, a Bayesian nonparametric model is used for the capability of handling an arbitrary number of sound sources as well as avoiding the performance degradation due to unnecessary TF masks. Concretely, the prior distribution of Bayesian nonparametric modeling allows for infinitely many TF masks for a theoretical flexibility, while also penalizes the emergence of unnecessary TF masks so as to stabilize the parameter estimation for the source separation.

2.2.2 Dereverberation methods

When we observe a sound in an indoor environment, the observed sound signal inevitably contains reverberation caused by the reflections of the sound signals on the walls and floor. The reverberation is known to degrade the sound source separation performance because the sound reflections virtually act as ghost sounds, which complicates the separation process. Dereverberation methods are categorized into two classes: spectral subtraction methods and linear transform methods.

Spectral subtraction methods reduce reverberation from the reverberant observation in the power spectrum domain, where the phase information is discarded (Lebart et al. 2001, Habets et al. 2008, Löllmann and Vary 2009, Wang et al. 2012, Erkelens and Heusdens 2010). Spectral subtraction methods are often designed for single channel observations (Lebart et al.

2001, Habets et al. 2008, Erkelens and Heusdens 2010). The limitation of these methods is the difficulty of the incorporation of these dereverberation techniques with multichannel separation algorithms because the dereverberation process discards the phase information that convey the time difference of arrivals information, which is essential to the separation process. Furthermore, the lack of the phase information may produce distorted resulting signals.

Dereverberation methods on the basis of linear transform of the multichannel observation overcome these limitations. The inverse filtering using a microphone array explicitly uses the phase information for a better dereverberation quality (Miyoshi and Kaneda 1988). Many multichannel dereverberation methods formulate the linear transform as an autoregressive process, or also called linear prediction, of the sequential observations (Triki and Slock 2006, Delcroix et al. 2007, Nakatani et al. 2010). A hybrid dereverberation method combining an autoregressive process and spectral subtraction is also developed for improved dereverberation performance (Kinoshita et al. 2009). The problem with these methods is that these methods assumes a single sound source in the reverberant observation, which results in a limited dereverberation performance when applied to a mixture of sound sources in a reverberant environment. The autoregression-based dereverberation model has been extended to multiple sound sources (Yoshioka and Nakatani 2012).

Multichannel dereverberation methods based on an autoregressive process can be combined with source separation methods to attain the robust extraction of sound sources in reverberant environments (Huang et al. 2005, Buchner and Kellermann 2010, Yoshioka et al. 2011, Takeda et al. 2012, Togami et al. 2013). While these joint models for sound source separation and dereverberation achieve a robust separation of reverberant sound source mixtures, these models assumes the number of sources is known in advance. In this sense, existing simultaneous separation and dereverberation methods fail to overcome the issue of the uncertainty in the source number. In addition, the liner multichannel dereverberation methods using an autoregressive process has a limitation that the reverberation is accurately estimated when the mixture is overdetermined (Gorokhov and Loubaton 1997). In summary, the linear multichannel dereverberation is suitable for combination with multichannel separation methods, though the state-of-the-art dereverberation methods are limited to overdetermined conditions, which is possibly undermined in the source number uncertainty.

2.2.3 Moving sound sources in dynamic acoustic environment

Most multichannel sound source separation methods are designed under the assumption that the propagation of the source signals to each microphone is time-invariant, that is, the relative position between the sources and microphones are fixed. The difficulty of sound source separation using only observation mixtures is that both the propagation of each sound source and the source signals themselves are unknown. To resolve this problem, most separation methods assume that only source signals are time-varying whereas the propagation of each source signal is fixed over time. If the movements of sound sources are allowed, the separation is regarded as a severely ill-posed problem because the temporal changes in the observed mixture may result from either the sound propagation or the changes in the source signals.

The application of microphone arrays to dynamic environments have been mostly focused on the tracking of sound sources (Affes and Grenier 1995, Brandstein et al. 1997, Valin et al. 2003, Murase et al. 2005, Ma et al. 2006), i.e., the temporal smoothing of sound source localization. The source number uncertainty problem is addressed in (Ma et al. 2006) in the sound source tracking task.

Some methods tackle the separation of moving sound sources by segmenting the observed mixture signals along the time axis, such that the source position can be regarded as stationary within each segment, to apply an ordinary separation method such as ICA (Koutvas et al. 2000, Prieto and Jinachitra 2005, Addison and Roberts 2006). Here, the determination of the window length for the segmentation is a critical problem. We prefer a long segment containing more data samples so as to stabilize the parameter estimation for the separation, although the stationarity of the sound sources may be unsatisfied in a too long segment.

A multimodal method has been proposed to address this problem, where the visual information is used to track the sound sources (Naqvi et al. 2010). If the sources are not moving, ICA is applied to separate the sources; whereas a beamforming is applied to separate moving sound sources. Though beamforming-based separation methods dispense with the assumption of source position stationarity, the separation quality is usually worse than separation methods that adaptively estimates the parameters such as ICA.

2.3 Summary

This chapter has reviewed various auditory processing methods to handle multisource environments. Multichannel methods have advantages in comparison with single channel methods due to their general applicability to various kinds of sound sources as well as the capability of source localization. Nevertheless, multichannel processing still faces three challenges regarding the auditory uncertainties: the uncertainty in the source number, reverberation, and dynamic environments.

In the subsequent chapters, Bayesian nonparametric models play the key role to cope with these three auditory uncertainties. In particular, the number of sources is a critical parameter to determine the model complexity of multichannel sound source separation and dereverberation. Bayesian nonparametric models are beneficial in that they are capable of bypassing the selection of model complexity thanks to the infinitely extensible modeling capability.

Chapter 3

Bayesian Nonparametric Multichannel Sound Source Separation and Localization

This chapter presents a basic multichannel model for sound source separation and localization on the basis of Bayesian nonparametrics. The Bayesian nonparametric model is designed to copes with the uncertainty in the number of sound sources.

3.1 Introduction

Computational auditory scene analysis (CASA) aims at a machine listening that can extract and analyze useful information and/or meaningful auditory events such as speech content and sound source type from audio recordings (Rosenthal and Okuno 1998, Wang and Brown 2006). The decomposition of these constituent sound sources is essential for CASA systems because a mixture of audio signals containing multiple sound sources is common in our daily environment (Common and Jutten 2010).

Many CASA systems use multiple sensors, e.g., a microphone array, to decompose the observed mixture into the individual sound sources (Benesty et al. 2008). Microphone arrays spatially filter the sound sources to act as a decomposition function. That is, they retrieve audio signals from different directions, which is referred to as sound source separation (Common and Jutten 2010, Sawada et al. 2011). If the alignment of the microphone array is available, the direction of arrival of each sound source can be realized as a sound source localization

function (Mandel et al. 2007). While these two problems of separation and localization are mutually dependent, most existing methods deal with a specific part of these problems, and combined in a cascade manner to handle both problems. The overall quality of this cascade approach is prone to be determined by the worst component. For example, the HARK sound source localization and separation system first localizes the sound sources, and then separates the mixture signal using the estimated direction of each source (Nakadai et al. 2010). Therefore, a failure in the localization step affects the separation. Thus, a unified method is necessary to optimize the mutually dependent problems.

Designing a unified framework for sound source localization and separation involves two challenges: how to model a unified microphone array processing and how to overcome the auditory uncertainties such as reverberation and an unknown source number. Though the localization and separation have been unified by a Bayesian topic model (Otsuka et al. 2012), this method assumes that the source number is available a priori, which is not always the case in practice. On the other hand, the estimation of the source number has also been tackled separately from the separation (Yamamoto et al. 2003, Loesch and Yang 2008). The drawbacks of these approaches are the necessity of parameter learning in advance or elaborate configuration depending on the auditory environments. An overall framework that unifies the localization and separation under the uncertainty of the source number will contribute to a more flexible CASA system than that by cascade approaches.

This chapter presents a model based on Bayesian nonparametrics for sound source separation and localization with source number estimation using a microphone array. We formulate this as a unified twofold clustering problem in which the sound source separation is formulated as a clustering of time-frequency points in the time-frequency domain of the observed spectrogram and sound source localization is formulated as an assignment of each cluster to a certain direction. The clusters corresponding to the different sound sources are generated using a hierarchical Dirichlet process to cope with the source number uncertainty. To infer the latent variables, we derive a collapsed Gibbs sampler that drastically improves the source number estimation accuracy.

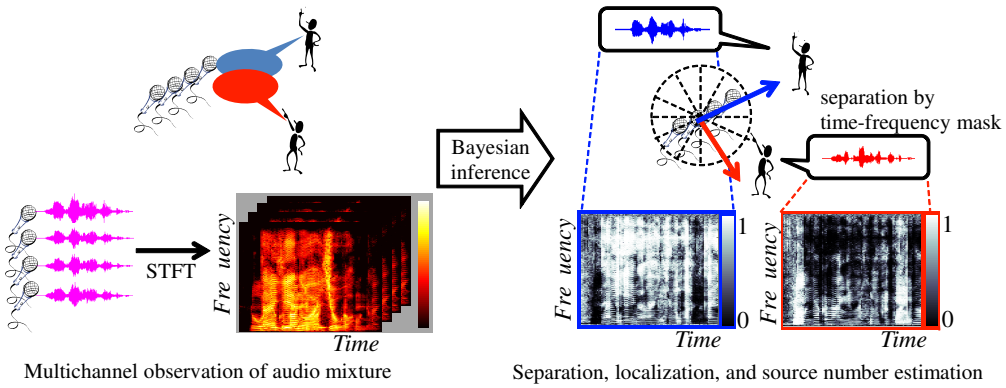


Figure 3.1: Illustration of our problem; sound source localization and separation with source number estimation. The process is carried out in the time-frequency domain to generate TF masks. Our Bayesian nonparametrics-based model dispenses with environment-dependent model configurations such as a priori source number information.

3.2 Problem and Related Work

Figure 3.1 outlines our problem. The inputs are a multichannel mixture audio signal and steering vectors that carry information about the alignment of microphones. The outputs are the respective audio signals comprising the observed mixture, the arrival directions of the sound sources, and the number of sources. Steering vectors are necessary for sound source localization. A steering vector conveys the time difference of sound arrivals at each microphone given a certain direction and a frequency bin. We use the steering vectors measured in an anechoic chamber so that the vectors can be used independently of the reverberation or environment. We can also synthesize steering vectors when we use a microphone array in a simple shape such as a linear array in Figure 3.1.

We make three assumptions on our auditory setup: (1) spatial sparsity of sound sources, (2) power sparsity of the audio in the time-frequency domain, and (3) non-moving sound sources. The first assumption means that all sound sources are located in different directions because a microphone array extracts audio signals coming from distinct directions. The second assumption is often satisfied by most audio signals due to their non-stationarity: non-stationary audio signals, such as speech, music, and so on, have their power on certain frequency bands that may vary over time. This leads to a sparse power distribution in the time-frequency domain. By the second assumption, we can assume that only one source signal is likely to be dominant

at each time-frequency point even for a mixture of sound sources. This supports the use of a clustering-based approach for sound source separation (Sawada et al. 2011, Otsuka et al. 2012, Bofill and Zibulevsky 2001). The third assumption means that the sound sources do not change their directions over time and is made for simplicity.

Sound source separation and localization in practical situations have two inherent problems: reverberation and source number uncertainty. When we observe a sound in a room, the observation contains reverberation that can be modeled as a convolutive process (Pedersen et al. 2007). Though methods in the time-frequency domain through a short-time Fourier transform (STFT) are often used to cope with the reverberation, this causes a permutation problem (Sawada et al. 2004). The permutation problem occurs when the separation is carried out independently of frequency bins in an unsupervised manner, e.g., using independent component analysis (ICA) (Common and Jutten 2010). To aggregate the spectrogram of a certain source, we must identify the signals of the same sound source from all frequency bins. Independent vector analysis (IVA) (Lee et al. 2007, Ono 2011) avoids the permutation problem by maximizing the independence of the constituent sound sources across all frequency bins simultaneously.

Due to the uncertainty of the number of sources, we have to deal with a model complexity problem and a possibly underdetermined situation. With ICA and IVA, the number of sources N is assumed not to exceed the number of microphones M. However, in practice, N is not always guaranteed to be capped at M, especially when we are unaware of the source number. The case in which $N > M$ is called an underdetermined problem. An approach to this condition is the clustering formulation that generates a time-frequency (TF) mask for each sound source (Yilmaz and Rickard 2004, Mandel et al. 2010, Sawada et al. 2011, Taghia et al. 2012, Mandel et al. 2007).

3.2.1 Problem formulation

We outline a general convolutive mixture model for observation of multiple sound sources with a microphone array in a reverberant environment. For the separation of the observed mixture signal, we explain the idea of Bayesian nonparametric model so as to cope with the source number uncertainty.

Consider a multichannel convolutive mixture of sound sources in the time domain:

$$x_m^{\text{time}}(\tau) = \sum_{n=1}^{N} b_{mn}^{\text{time}}(\tau) * s_n^{\text{time}}(\tau),$$

where τ is the time index, x_m^{time} is the signal observed by the mth channel ($m = 1 \dots M$) in the time domain, s_n^{time} is the nth source signal ($n = 1 \dots N$), and $b_{mn}^{\text{time}}(\tau)$ denotes the impulse response from the location of source n to microphone m. Operator $*$ denotes convolution. We assume the impulse responses have finite lengths.

The observed audio signal is converted into the TF domain through short-time Fourier transformation (STFT). Let $\mathbf{x}_{tf}$ be the multichannel observation in the TF domain in time frame t and frequency bin f. The audio signal mixture for N sources is observed with M microphones; that is, $\mathbf{x}_{tf} \in \mathbb{C}^M$ is an M-dimensional complex-valued vector. We assume that the N sources are mixed through a convolutive process in the TF domain:

$$\mathbf{x}_{tf} = \mathbf{B}_f^0 \mathbf{s}_{tf} + \sum_{j=1}^{J-1} \mathbf{B}_f^j \mathbf{s}_{t-j,f}, \tag{3.1}$$

where $\mathbf{s}_{tf} \in \mathbb{C}^N$ is the source signal, with each element corresponding to a distinct sound source. The instantaneous mixing coefficients are denoted by $\mathbf{B}_f^0 \in \mathbb{C}^{M \times N}$, where the element in the mth row and nth column, $b_{f,mn}^0$, denotes the propagation coefficient from the nth source to the mth microphone. Similarly, the elements in $\mathbf{B}_f^j \in \mathbb{C}^{M \times N}$ denote the propagation coefficients from the previous source signals with a time lag of j. We assume that the propagation coefficients depend on the direction of arrival of each source. Therefore, each column in $\mathbf{B}_f^j$ ($0 \le j < J$) is determined by the location of each source. Here, we manipulate each frequency bin independently for simplicity. This means that we ignore the spectral leakage between adjacent frequency bins caused by STFT. A discussion of spectral leakage is presented in (Nakatani et al. 2008a).

The task of sound source separation is to estimate source signal vector $\mathbf{s}_{tf}$ on the basis of observations $\mathbf{x}_{tf}$ so as to extract the sound source signals. In our implementation, we extract the image of the kth source $\hat{\mathbf{x}}_{tf}^k = \mathbf{b}_{fk}^0 s_{tf}^k$, where $\mathbf{b}_{fk}^0$ denotes the kth column of instantaneous mixing matrix $\mathbf{B}_f^0$ and s_{tf}^k denotes the kth element of $\mathbf{s}_{tf}$. The difficulty is that we have to retrieve source signals $\mathbf{s}_{tf}$ using only the observed mixture $\mathbf{x}_{tf}$ without knowing propagation $\mathbf{B}_f^j$ and source number N.

3.2.2 Separation without dereverberation

We first formulate the separation and localization problem without considering the reverberation. In this chapter, Eq. (3.1) is approximated as

$$\mathbf{x}_{tf} \approx \mathbf{B}^0_f \mathbf{s}_{tf}. \tag{3.2}$$

Since the size of $\mathbf{B}^0_f$ is $M \times N$, if the number of sources matches the number of microphones ($N = M$), a linear transform can be used to recover the sources from the observed mixture such that $\mathbf{s}_{tf} = \mathbf{W}_f \mathbf{x}_{tf}$ by, for example, using ICA (Hyvärinen et al. 2001, Common and Jutten 2010, Sawada et al. 2007) and its variants (Lee et al. 2007, Ono 2011). If the mixing process is overdetermined, that is, the number of sources is less than the number of microphones ($N < M$), dimensionality reduction based on principal component analysis is often used as a preprocessing to reduce the dimensionality to N (Winter et al. 2006). In this case, the linear transform for the separation is still tractable.

We use a TF masking approach (Yilmaz and Rickard 2004, Mandel et al. 2010, Sawada et al. 2011) to deal with situations in which the source number information is unavailable with a consistent approach. In such situations, we need to consider both the overdetermined and underdetermined conditions, where the underdetermined condition means that $N > M$. The TF masking approach provides a consistent way of dealing with both conditions. This approach is based on two assumptions: the powers of the sound sources are sparsely distributed in the TF domain and the powers of different sources rarely overlap. In other words, at most one sound source is assumed to be dominant for each time t and frequency f (TF point). This enables us to further approximate the observation:

$$\mathbf{x}_{tf} \approx \mathbf{b}^0_{f k_{tf}} s^{k_{tf}}_{tf}, \tag{3.3}$$

where k_{tf} represents the dominant source at time t and frequency f, and $\mathbf{b}^0_{fk}$ is an M-dimensional vector corresponding to the propagation coefficients of the kth source ($1 \leq k \leq N$), which is the kth column vector of $\mathbf{B}^0_f$ in Eq. (3.2), and $s^{k_{tf}}_{tf}$ is the kth source signal of the dominant source. Note that coefficients $\mathbf{b}^0_{fk}$ depends on the direction from which the kth source arrives. The source sparsity assumption implies that $s^{k'}_{tf} = 0$ if $k_{tf} = k$ and $k' \neq k$. Thus, we simply write s_{tf} instead of $s^{k_{tf}}_{tf}$ in the context of source sparsity.

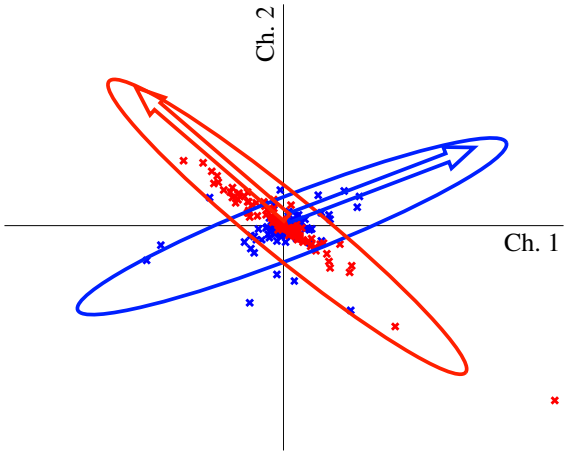


Figure 3.2: Plot of complex-valued multichannel signals for two sources in a frequency bin of 1000 (Hz). X-marks correspond to TF points $\mathbf{x}_{tf}$ in frequency bin. Each axis corresponds to a distinct channel of microphone array: horizontal axis is real part of channel 1 and vertical axis is imaginary part of channel 2. Colors represent distinct sound sources. Arrows along TF points of each source are parallel with propagation coefficients of that source.

TF mask generation is formulated as a clustering problem in the following manner. As illustrated in Figure 3.2, the source signals coming from different directions span distinct subspaces in the TF-domain observation space of each frequency bin. The TF masks for extracting constituent sound sources are estimated through the clustering of each TF point depending on the subspace along $\mathbf{b}^0_{fk}$ for each frequency bin f. We use the covariance model (Duong et al. 2010) defined as follows to achieve this clustering.

$$\mathbf{x}_{tf} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_{tf}|\mathbf{0}, \lambda_{tf}\mathbf{H}_{fk_{tf}}), \tag{3.4}$$

where $\mathcal{N}_{\mathbb{C}}(\mathbf{x}|\boldsymbol{\mu}, \Lambda)$ represents the complex normal distribution of $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and precision matrix (inverse of covariance matrix) Λ. This choice of notation is for precision matrices instead of covariance matrices because we use conjugate prior distributions for precision matrices. In Eq. (3.4), the precision term is factorized into two terms: scale factor λ_{tf} and matrix $\mathbf{H}_{fk_{tf}}$. These two terms are defined as $\lambda^{-1}_{tf} = \mathbb{E}[s_{tf}]$ and $\mathbf{H}^{-1}_{fk} = \mathbb{E}[\mathbf{b}^0_{fk}\mathbf{b}^{0\,H}_{fk}]$, respectively, where $\mathbb{E}[\cdot]$ means the expectation and $\cdot^H$ denotes Hermitian transposition. This factorization is derived by taking the outer product of both sides of Eq. (3.3). The clustering of the observations at each TF point $\mathbf{x}_{tf}$ is achieved by estimating the class index k_{tf} for each TF point on the basis of likelihood defined by Eq. (3.4). The sound source localization (direction of arrival

estimation) can be simultaneously carried out by inspecting matrix $\mathbf{H}_{fk}$ associated with propagation coefficient $\mathbf{b}^0_{fk}$ because the propagation coefficients that form the subspace are related to the direction from which the source is arriving.

3.2.3 Source number uncertainty and Bayesian nonparametrics

The source number uncertainty is closely related to the selection of model complexity. For example, source separation methods using ICA or IVA often reduce the dimensionality of the multichannel observation from the microphone number to the source number by using principal component analysis when the number of sources is available (Winter et al. 2006). PCA is employed in order to reduce the number of latent parameters in the separation matrices as a preprocessing of ICA (Vaseghi and Jetelová 2006, Kovacevic and McIntosh 2007). Similarly, TF masking-based separation methods often use the same number of TF masks as that of sources so that the model complexity should fit the source separation problem (Sawada et al. 2011). In case of source number uncertainty, where an appropriate model complexity is unknown, a simple solution is to use a sufficiently flexible model (Araki et al. 2009, Taghia et al. 2012). For example, if we can assume the source number is at most four, four TF masks are sufficient. This approach is problematic in two ways: first, a model with a finite number of TF masks fails in the separation when the source number exceeds the number of TF masks. Second, using a too flexible model may affect the performance because redundantly flexible models are apt to overfit the data.

Nonparametric Bayesian models are helpful in such situations since we can bypass a careful selection of the mask number K by assuming an infinite number of TF masks in the model. Furthermore, the prior distribution for the TF masks penalizes unnecessary emergence of TF masks. This property helps the inference to avoid an overfitting that may affect the separation and localization performance. Some Bayesian nonparametric models have been related to microphone array processing techniques. Infinite independent component analysis (Knowles and Ghahramani 2007) is a nonparametric counterpart of ICA. Because this model allows only for real-valued variables, the separation is limited to the time domain, which is vulnerable to reverberation. While Nagira et al. extend the model into the time-frequency domain (Nagira et al. 2013, 2012), they cope with the permutation resolution separately after the separation.

This naïve extension into the time-frequency domain is problematic because the inference results in each frequency bin may converge to a different number of sources.

The contribution of this chapter is twofold. (1) We present a nonparametric Bayesian model that unifies sound source localization, separation, and permutation resolution using a hierarchical Dirichlet process (HDP) (Teh et al. 2006). This hierarchical model is advantageous in that the number of sources is globally handled instead of locally for each frequency bin. (2) We derive a collapsed Gibbs sampler (CGS) that promotes the shrinkage of the classes for more accurate sound source estimation. This collapsed inference accelerates the inference by marginalizing out the multichannel precision matrices. While Kameoka et al. develop a similar framework based on Bayesian nonparametrics without an HDP (Kameoka et al. 2012b), the use of this hierarchical structure in our model is convenient to encourage the temporal synchronization of source dominance over frequency bins so as to generate the time-frequency masks. This mechanism gains a robustness against the reverberation because reverberation is apt to obscure the temporal synchronization in the time-frequency domain.

3.3 HDP-Based Sound Source Localization and Separation

As mentioned, the problem of sound source separation and localization is tackled as a clustering problem. The observed multichannel mixture signal is converted into the TF domain by using STFT. The separation is the clustering of multichannel vectors at each TF point while the localization is the matching of each cluster with steering vectors. A separation with permutation resolution has been developed based on latent Dirichlet allocation (LDA) (Blei et al. 2003) in which the time frames are regarded as documents and the TF points are treated as words (Otsuka et al. 2012). In this model, a few sound sources (corresponding to topics in the context of LDA) are preferred in each time frame to help the permutation resolution by synchronizing the appearance of the same source across frequency bins. Because LDA is limited to a finite number of sources, a direct application of LDA fails to take into account the uncertainty about the number of sources. To solve this, we introduce an unbounded model in terms of the number of sources by using HDP (Teh et al. 2006).

Our model is designed to achieve a balance between the capability to deal with an unbounded number of sound sources and tractable inference of the latent parameters. In order to

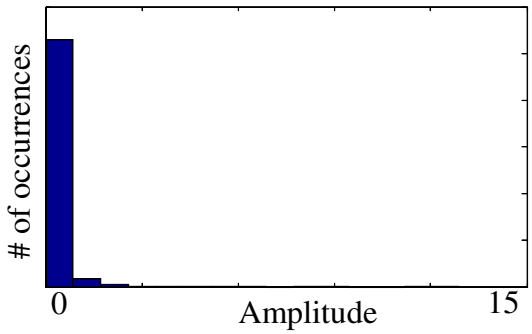


Figure 3.3: Histogram of time-frequency amplitudes $|\mathbf{x}_{tf}|$. The power of each time-freuqency point is close to zero in most cases. This implies the power sparsity of sound sources in the time-frequency domain. That is, at most one source is assumed dominant at each time-frequency point.

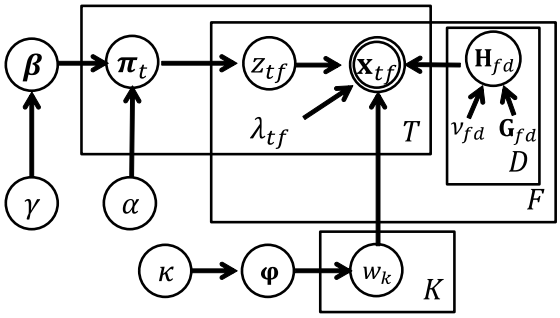


Figure 3.4: Graphical model depicting the generative process. Observed variables are double-circled. Latent random variables are denoted with a single circle. Fixed values are denoted by plain symbols. Variables inside a box with an upper-case symbol are independent and identically distributed with respect to the corresponding lower-case index.

satisfy these properties, we employ a likelihood distribution suitable to model multichannel observation of directional sound sources as well as conjugate prior distributions. The conjugate priors are helpful to develop an efficient inference procedure in that the parameter estimation is accelerated and stabilized with analytic derivation of the posterior distribution and marginalization of some of the latent parameters.

The notations used in this section are listed in Table 3.1. Figure 3.4 shows the graphical representation of our model. The double-circled $\mathbf{x}_{tf}$ is the observation, the circled symbols are latent probability variables, and the plain symbols are fixed values. Section 3.3.1 explains how the multichannel input signal is observed and associated with steering vectors. Section 3.3.2 describes the inference by using CGS. Section 3.3.3 shows how the sound sources are retrieved or localized and how the number of sources is estimated using the sampled latent variables. Finally, Section 3.3.4 shows the initialization procedures. A set of variables is denoted with a tilde without subscripts, e.g., $\tilde{\mathbf{x}} = \{\mathbf{x}_{tf} | 1 \leq t \leq T, 1 \leq f \leq F\}$. As revealed in the subsequent sections, the inference of $\tilde{\mathbf{z}}$ corresponds to the estimation of TF masks for separation, and the inference of $\tilde{\mathbf{w}}$ corresponds to the localization.

Table 3.1: Notations.

Symbol	Meaning
t	Time frame index from 1 to T
f	Frequency bin from 1 to F
k	Class index from 1 to K
d	Direction index from 1 to D
M	Number of microphones
N	Number of sound sources
$\mathbf{x}_{tf}$	Observed M-dimensional complex column vector
z_{tf}	Class indicator at time frame t and frequency bin f
$\boldsymbol{\pi}_t$	Class proportion at time frame t
w_k	Direction indicator for class k
$\boldsymbol{\varphi}$	Direction proportion for all classes
λ_{tf}	Inverse scale parameter for $\mathbf{x}_{tf}$
$\mathbf{H}_{fd}$	Inverse covariance of direction d at frequency bin f
n_{tk}	Number of time-frequency points assigned to class k at time frame t
n_{fk}, n_{fd}	Number of time-frequency points at frequency bin f of class k or direction d, respectively
c_d	Number classes assigned to direction d

3.3.1 Multichannel observation and generative model

This section explains the generative process described in Figure 3.4. As outlined in Section 3.2.2, we use a covariance model (Duong et al. 2010) for the likelihood function of the multichannel observation in the TF domain: each sample follows a complex normal distribution with zero mean and time-varying covariance.

The likelihood function is a complex normal distribution:

$$\mathbf{x}_{tf}|z_{tf}, \tilde{\mathbf{w}}, \lambda_{tf}, \tilde{\mathbf{H}} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_{tf}|\mathbf{0}, \lambda_{tf}\mathbf{H}_{fw_{z_{tf}}}), \tag{3.5}$$

where z_{tf} and w_k indicate the class of $\mathbf{x}_{tf}$ and the direction of class k, respectively. Note that the precision matrix $\mathbf{H}_{fw_{z_{tf}}}$ is indexed with frequency bin f and the direction of arrival of the sound source. Here, $w_{z_{tf}}$ denotes the direction in which $\mathbf{x}_{tf}$ is located. The probability density function (pdf) of a complex normal distribution $\mathcal{N}_{\mathbb{C}}(\mathbf{x}|\boldsymbol{\mu}, \Lambda)$ is defined as $\frac{|\Lambda|}{\pi^M}\exp\left(-\mathbf{x}^H\Lambda\mathbf{x}\right)$ (van den Bos 1995) with mean $\boldsymbol{\mu}$ and precision Λ. $|\Lambda|$ is the determinant of matrix Λ.

The direction matrix $\mathbf{H}_{fd}$ follows the conjugate prior, i.e., complex Wishart distribution (Conradsen et al. 2003).

$$\mathbf{H}_{fd} \sim \mathcal{W}_{\mathbb{C}}(\mathbf{H}_{fd}|\nu_{fd}, \mathbf{G}_{fd}), \tag{3.6}$$

where the pdf of complex Wishart distribution $\mathcal{W}_{\mathbb{C}}(\mathbf{H}|\nu,\mathbf{G})$ is $\frac{|\mathbf{H}|^{\nu-M}\exp\{-\mathrm{tr}(\mathbf{H}G^{-1})\}}{|\mathbf{G}|^{\nu}\pi^{M(M-1)/2}\prod_{m=0}^{M-1}\Gamma(\nu-m)}$; $\mathrm{tr}(\mathbf{A})$ is the trace of $\mathbf{A}$ and $\Gamma(x)$ is the gamma function. The hyperparameters of the complex Wishart distribution are set as $\nu_{fd}=M$ and $\mathbf{G}_{fd}=(\mathbf{q}_{fd}\mathbf{q}_{fd}^{H}+\epsilon\mathbf{I}_M)^{-1}$. $\mathbf{G}_{fd}$ is generated from the given steering vectors $\mathbf{q}_{fd}$, where $\mathbf{q}_{fd}$ is normalized s.t. $\mathbf{q}_{fd}^{H}\mathbf{q}_{fd}=1$ and ϵ is set to 0.01 to enable inverse operation.

An HDP (Teh et al. 2006) is used as the generative process of z_{tf}, which is an infinite extension of an LDA. We introduce this hierarchical generative process to resolve the permutation ambiguity (Otsuka et al. 2012). First, global class proportion $\boldsymbol{\beta}$ is generated, where the dimensionality of $\boldsymbol{\beta}$ is infinitely large. Each element represent the average weights of infinitely-many classes throughout the spectrogram. Then, the time-wise class proportion $\boldsymbol{\pi}_t$ is sampled in accordance with $\boldsymbol{\beta}$. Again, $\boldsymbol{\pi}_t$ is an infinite-dimensional vector where the elements represent the weights of infinite classes at the specific time frame t. Finally, each z_{tf} is sampled in accordance with the time-wise class proportion $\boldsymbol{\pi}_t$. As Figure 3.5 shows the dominance of each source is synchronized across frequency bins. Therefore, we achieve the permutation resolution by introducing $\boldsymbol{\pi}_t$. The stick-breaking construction for an HDP (Teh et al. 2006) is given by:

$$\boldsymbol{\beta}|\gamma\sim\mathrm{GEM}(\gamma),\qquad \boldsymbol{\pi}_t|\alpha,\boldsymbol{\beta}\sim\mathrm{DP}(\alpha,\boldsymbol{\beta}),\qquad z_{tf}|\boldsymbol{\pi}_t\sim\boldsymbol{\pi}_t, \tag{3.7}$$

where $\mathrm{GEM}(\gamma)$ is the Griffiths-Engen-McCloskey distribution with concentration γ; $\mathrm{DP}(\alpha,\boldsymbol{\beta})$ denotes the Dirichlet process with concentration α and base measure $\boldsymbol{\beta}$. Here, the expectation of $\boldsymbol{\pi}_t$ satisfies $\mathbb{E}[\boldsymbol{\pi}_t]=\boldsymbol{\beta}$. We place gamma distribution priors for concentrations $\gamma\sim\mathcal{G}(\gamma|a_\gamma,b_\gamma)$ and $\alpha\sim\mathcal{G}(\alpha|a_\alpha,b_\alpha)$. The hyperparameters are set as $a_\gamma=0.05, b_\gamma=5, a_\alpha=0.01$, and $b_\alpha=1$.

Direction indicator w_k contributes to the sound source localization as well as to the permutation resolution because classes from the same direction are associated with each other across all frequency bins. This variable is drawn from proportion $\boldsymbol{\varphi}$ generated from a flat Dirichlet distribution.

$$\boldsymbol{\varphi}|\kappa\sim\mathcal{D}\left(\boldsymbol{\varphi}\middle|\frac{\kappa}{D}\mathbf{1}_D\right),\qquad w_k|\boldsymbol{\varphi}\sim\boldsymbol{\varphi}, \tag{3.8}$$

where $\mathbf{1}_D$ is a D-dimensional vector in which all elements are 1 and $\mathcal{D}(\cdot|\boldsymbol{\alpha})$ denotes the Dirichlet distribution with parameter $\boldsymbol{\alpha}$. Our model is a finite mixture with regard to direction due

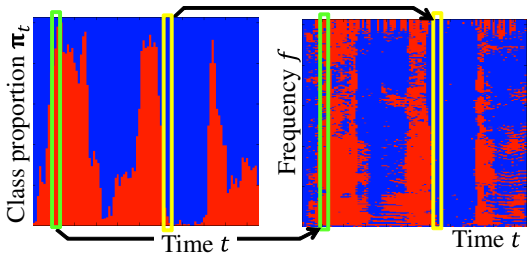


Figure 3.5: Left: class proportion π_t for each time frame. Right: TF mask of two sources denoted by z_{tf}. Each TF point is assigned to a class in accordance with the class proportion of the time frame.

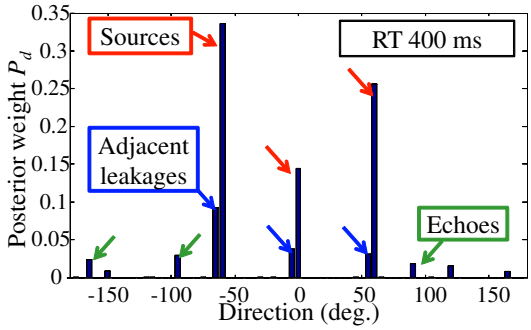


Figure 3.6: Posterior weights for three-source mixture with 400-ms reverberation time. Three salient peaks are found at $-60°$, $0°$, and $60°$ indicated by red arrows. Some echo component (e.g., reflection on the wall) is observed as TF masks with small weights.

to the limitation of the spatial resolution of microphone arrays. We also place a gamma prior over κ as $\mathcal{G}(\kappa|a_\kappa, b_\kappa)$, where $a_\kappa = 1$ and $b_\kappa = 1$.

3.3.2 Inference by collapsed Gibbs sampler

For sound source separation and localization, the inference of $\tilde{\mathbf{z}}$ and $\tilde{\mathbf{w}}$ is important. These variables are inferred by using a CGS with π, φ, and $\tilde{\mathbf{H}}$ marginalized out. The joint distribution of $\tilde{\mathbf{x}}, \tilde{\mathbf{z}}$, and $\tilde{\mathbf{w}}$ is

$$
\begin{aligned}
&p(\tilde{\mathbf{x}}, \tilde{\mathbf{z}}, \tilde{\mathbf{w}} | \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}, \tilde{\mathbf{G}}, \alpha, \beta, \gamma, \kappa) \\
&= \iiint p(\tilde{\mathbf{x}}, \tilde{\mathbf{H}} | \tilde{\mathbf{z}}, \tilde{\mathbf{w}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}, \tilde{\mathbf{G}}) p(\tilde{\mathbf{z}}, \tilde{\pi} | \alpha, \beta, \gamma) p(\tilde{\mathbf{w}}, \varphi | \kappa) d\tilde{\mathbf{H}} d\pi d\varphi \\
&= \prod_{tf} \left(\frac{\lambda_{tf}}{\pi} \right)^M \prod_{fd} \frac{\prod_{m=0}^{M-1} \Gamma(\hat{\nu}_{fd} - m) |\hat{\mathbf{G}}_{fd}|^{\hat{\nu}_{fd}}}{\prod_{m=0}^{M-1} \Gamma(\nu_{fd} - m) |\mathbf{G}_{fd}|^{\nu_{fd}}} \\
&\quad \times \prod_t \left\{ \frac{\Gamma(\alpha)}{\Gamma(\alpha + n_{t\cdot})} \prod_k \frac{\Gamma(\alpha\beta_k + n_{tk})}{\Gamma(\alpha\beta_k)} \right\} \frac{\Gamma(\kappa)}{\Gamma(\kappa + c_\cdot)} \prod_d \frac{\Gamma(\frac{\kappa}{D} + c_d)}{\Gamma(\frac{\kappa}{D})},
\end{aligned}
\tag{3.9}
$$

where a dot in the subscripts denote summation over the index, e.g., $n_{t\cdot} = \sum_k n_{tk}$ and $c_\cdot = \sum_d c_d$. Note that, as explained in (Teh et al. 2006), a finite k can be handled during the inference so that the product over k is valid. The dimensionality of β dynamically changes over sampling

iterations in accordance with the number of classes actually drawn. The posterior parameters of the complex Wishart distribution, $\hat{\nu}_{fd}$ and $\hat{\mathbf{G}}_{fd}$, are updated using sufficient statistics:

$$\begin{aligned}\hat{\nu}_{fd} &= \nu_{fd} + \sum_{t:w_{z_{tf}}=d} 1 = \nu_{fd} + n_{fd}, \\ \hat{\mathbf{G}}_{fd}^{-1} &= \mathbf{G}_{fd}^{-1} + \sum_{t:w_{z_{tf}}=d} \lambda_{tf}\mathbf{x}_{tf}\mathbf{x}_{tf}^{H},\end{aligned} \tag{3.10}$$

where $\sum_{t:w_{z_{tf}}=d} \cdot$ means a summation over the samples assigned to direction d in frequency bin f.

From Eq. (3.9), z_{tf} and w_k are stochastically updated:

$$\begin{aligned}p(z_{tf}=k|\tilde{\mathbf{x}},\vartheta\backslash z_{tf}) \propto (\alpha\beta_k + n_{tk}^{\backslash tf})\frac{\Gamma(\hat{\nu}_{fw_k}^{\backslash tf}+1)}{\Gamma(\hat{\nu}_{fw_k}^{\backslash tf}-M+1)} \\ \times \frac{\left|\mathrm{inv}(\hat{\mathbf{G}}_{fw_k}^{\backslash tf})\right|^{\hat{\nu}_{fw_k}^{\backslash tf}}}{\left|\mathrm{inv}(\hat{\mathbf{G}}_{fw_k}^{\backslash tf})+\lambda_{tf}\mathbf{x}_{tf}\mathbf{x}_{tf}^{H}\right|^{\hat{\nu}_{fw_k}^{\backslash tf}+1}},\end{aligned} \tag{3.11}$$

$$\begin{aligned}p(w_k=d|\tilde{\mathbf{x}},\vartheta\backslash w_k) \propto \left(\frac{\kappa}{D}+c_d^{\backslash k}\right) \\ \times\prod_f\left\{\prod_{m=0}^{M-1}\frac{\Gamma(\hat{\nu}_{fd}^{\backslash k}+n_{fk}-m)}{\Gamma(\hat{\nu}_{fd}^{\backslash k}-m)}\frac{\left|\mathrm{inv}(\hat{\mathbf{G}}_{fd}^{\backslash k})\right|^{\hat{\nu}_{fd}^{\backslash k}}}{\left|\mathrm{inv}(\hat{\mathbf{G}}_{fd}^{\backslash k})+\sum_{t:z_{tf}=k}\lambda_{tf}\mathbf{x}_{tf}\mathbf{x}_{tf}^{H}\right|^{\hat{\nu}_{fd}^{\backslash k}+n_{fk}}}\right\},\end{aligned} \tag{3.12}$$

where $\vartheta\backslash z$ denotes all latent variables except z, superscripts $\backslash tf$ and $\backslash k$ mean the statistics without the sample at t and f or samples of class k, respectively, and $\mathrm{inv}(\mathbf{G})$ is the inverse matrix of $\mathbf{G}$.

Let K be the number of sampled classes. To allow for the probability of z_{tf} taking an unassigned class $K+1$ in Eq. (3.11), $\boldsymbol{\beta}$ has $K+1$ elements, as explained in (Teh et al. 2006). To calculate the probability of $z_{tf}=K+1$, $w_{K+1}=d$ is temporarily drawn with probability $\frac{\kappa/D+c_d}{\kappa+c_\cdot}$. If z_{tf} is chosen to be $K+1$, K is updated as $K\leftarrow K+1$, and the dimensionality of $\boldsymbol{\beta}$ increases by one with $\beta_K \leftarrow b\beta_K$ and $\beta_{K+1}\leftarrow(1-b)\beta_K$, where b is drawn from a beta distribution: $b\sim\mathcal{B}(1,\gamma)$.

The updates of the other parameters, α, β, γ, and κ, follow a procedure described in (Teh et al. 2006, Escobar and West 1995). These parameters are updated using auxiliary variables.

3.3.3 Localization, separation, and source number estimation

The collapsed Gibbs sampler described in Eq. (3.11, 3.12) produces the samples of latent variables indexed by i: $\{\tilde{\mathbf{z}}^{(i)}, \tilde{\mathbf{w}}^{(i)}\}_{i=1}^{I}$. Sound sources are retrieved by applying a TF mask corresponding to a certain direction. The multichannel spectrogram of a sound source in direction d, denoted by $\hat{\mathbf{x}}_{tf}^{d}$, is retrieved using

$$\hat{\mathbf{x}}_{tf}^{d} = \frac{1}{I} \sum_{i=1}^{I} \delta\left(w_{z_{tf}^{(i)}}^{(i)}, d\right) \mathbf{x}_{tf}, \tag{3.13}$$

where $\delta(m,n)$ is the Kronecker delta, i.e., $\delta(m,n) = 1$ if $m = n$, and 0 otherwise. The factor $\frac{1}{I}\sum_{i=1}^{I} \delta(w_{z_{tf}^{(i)}}^{(i)}, d)$ is the estimated TF mask for direction d at time t and frequency f. We can distinguish in which direction sound sources are located by defining the posterior weight for each direction as

$$P_d = \frac{1}{I} \sum_{i=1}^{I} \sum_{tf} \delta\left(w_{z_{tf}^{(i)}}^{(i)}, d\right). \tag{3.14}$$

If we want N sources from the mixture, we choose N directions in descending order of P_d. The sound sources are thereby localized and separated.

The number of sound sources is estimated using the posterior weights defined in Eq. (3.14). Figure 3.6 shows the posterior weights of a three-source mixture with a reverberation time of 400 (ms). We can see three salient peaks (indicated by red arrows) with smaller peaks in the adjacent directions (blue arrows). Reverberation causes additional peaks corresponding to echoes (green arrows). The number of sources is estimated using a three-step process. (1) Ignore the weights adjacent to larger peaks: $P_d \leftarrow 0$, if $P_d < P_{d+1}$ or $P_d < P_{d-1}$. (2) Sort the weights in descending order: $P'_1 > P'_2 > \ldots > P'_D$. (3) Find the number $\hat{N}$ where the weight drops most sharply: $\hat{N} = \text{argmax}_N P'_N / P'_{N+1}$ while $P'_{N+1} > 0$. If P'_N / P'_{N+1} monotonically increases until $P'_{N+1} = 0$, $\hat{N} = N$.

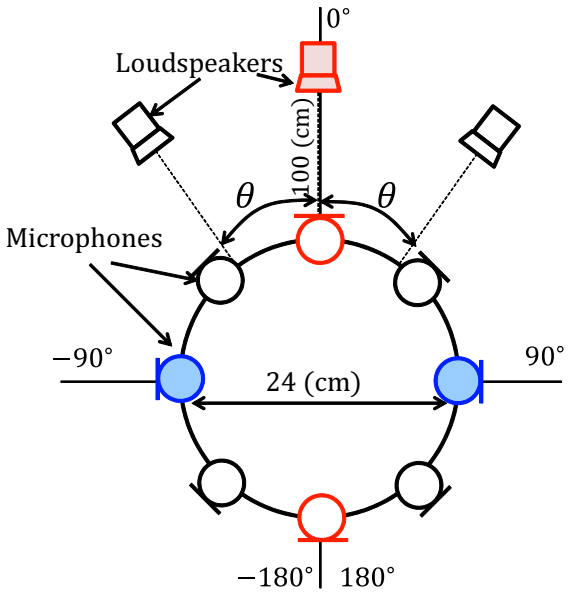


Figure 3.7: Microphone array configuration and positions of sound sources. The number of microphones M is 2, 4, or 8 whereas the number of sources N is set as 2 or 3. When $M = 2$, blue microphones are used. When $M = 4$, blue and red microphones are used. All microphones are used when $M = 8$. When $N = 2$, the center source illustrated in red is omitted.

3.3.4 Initialization of the inference

The inference is initialized in a similar way as previously reported (Otsuka et al. 2012). The inference starts with a certain number of classes K. First, w_k is initialized with a uniform distribution whose support has no overlap with the other classes. Then, each z_{tf} is drawn using the sampled w_k and the hyperparameter of Wishart distribution, $\mathbf{G}_{fd}$, generated from the steering vectors:

$$\begin{aligned} p(w_k = d) &= \mathcal{U}\left(\left\{d \,\middle|\, \frac{k-1}{K}D \leq d < \frac{k}{K}D\right\}\right), \\ p(z_{tf} = k) &\propto \exp\left(-\mathbf{x}_{tf}^H \mathbf{G}_{f w_k} \mathbf{x}_{tf}\right), \end{aligned} \tag{3.15}$$

where $\mathcal{U}(A)$ is a pdf of uniform distribution on set A.

3.4 Experimental Results

We evaluate the sound source separation, localization, and source number estimation performances of our HDP-CGS method using simulated and recorded mixtures. We compare the

source separation performance with those of state-of-the-art sound source separation methods: LDA-VB (Otsuka et al. 2012) and IVA (Ono 2011) for $M \geq N$ and TF masking with permutation resolution (TF-perm) (Sawada et al. 2011) for $M < N$. The localization and source number estimation performance are compared between HDP-CGS and LDA-VB.

3.4.1 Experimental setup

Figure 3.7 illustrates the experimental setup. We used two, four, or eight microphones ($M = 2, 4, 8$) to observe two or three sound source mixtures ($N = 2, 3$) with the interval $\theta = 30$, 60, and 90°. The microphones depicted in shaded blue were used when $M = 2$, those depicted in blue and red were used when $M = 4$, and all microphones in Figure 3.7 were used when $M = 8$. The center speaker (in red) was omitted when $N = 2$. The steering vectors, $\tilde{\mathbf{q}}$, of the microphone array were measured in an anechoic room such that $D = 72$ with 5° resolution. When $M = 2$, we used the steering vectors ranging from $-90°$ to 90° so as to avoid the front-back ambiguity. The steering vectors were generated from a Fourier transform of the first 1024 points of the anechoic impulse responses.

The experiments used both simulated and recorded mixtures in three rooms with reverberation times (RT) of 150, 400, and 600 (ms). The simulated mixtures were generated by convoluting the impulse responses measured in each room. The impulse responses are explained in the appendix. For each condition, 20 mixtures were tested using JNAS phonetically-balanced Japanese utterances. The average length of these mixtures is around 5 (sec). The audio signals were sampled at 16,000 (Hz), and STFT was carried out with a 1024 (pt) hanning window and a 256 (pt) shift size.

We use the signal-to-distortion ratio (SDR) as the metric for separation quality (Vincent et al. 2006). Since this ratio is calculated from the N original signals and the identical number of separated signals, we extracted N sound sources regardless of the source number estimation results. We compare five methods in this experiments: HDP-CGS, LDA-VB, IVA, ICA with permutation resolution (Sawada et al. 2007) (ICA-perm), and TF masking with permutation resolution (TF-perm). HDP-CGS and LDA-VB separate N sources in descending order of the posterior weight, as explained in Section 3.3.3, whereas IVA and ICA-perm take N sources in descending order of the power of the separated audio signals, and TF-perm carries out TF

mask clustering assuming N sources. Note that TF-perm uses the fact of N sources for the inference while the inferences of HDP-CGS and LDA-VB are independent of N. The number of classes K used by LDA-VB was 12, and HDP-CGS was initialized with $K = 12$.

In Section 3.4.4, the source number estimation results are compared between HDP-CGS, LDA-VB, and source separation and source counting method developed by Araki et al. (Araki et al. 2009). Since this method is developed for stereo observation ($M = 2$), we refer to this method as Stereo hereafter. The idea of the source counting of Stereo is similar to our method in that Stereo generates TF masks for each source and then estimates the source number by counting the TF masks the weight of which is above a certain threshold. The TF masks are estimated through the EM algorithm where the observation is based on the phase difference of two microphone, that is, the phase of non-diagonal elements of $\mathbf{x}_{tf}\mathbf{x}_{tf}^H$. In contrast, our method uses both the phase and level difference by considering $\lambda_{tf}\mathbf{x}_{tf}\mathbf{x}_{tf}^H$ in Eq. (3.10), and extends the model to any number of microphones.

The inference (parameter estimation) procedures are configured as follows. The collapsed Gibbs sampler for HDP-CGS was iterated 50 times with the first 20 cycles discarded as a burn-in period. The other methods are iterated until the evaluation function converges. LDA-VB typically converged in about 15 iterations. The iteration of IVA was 50 cycles. ICA-perm carried out 50 iterations for the separation for each frequency bin and 30 iterations for the permutation resolution. TF-perm required 50 iterations for the separation and 30 iterations for the permutation resolution, respectively. Computational complexity of each method is compared in Table 3.2. The number of iterations is the necessary cycles for the convergence. Here, one iteration involves the whole spectrogram; for example, TF masking-based methods updates the weight of TF masks at all TF points in one iteration whereas linear separation methods updates the separation matrices of all frequency bins in each iteration. The class number K for HDP-CGS is the number of instantiated classes during the inference. HDP-CGS requires iterative M^3 operations due to the calculation of matrix determinants in Eqs. (3.11) and (3.12). In practice, we can accelerate the computation by skipping the evaluation of the probability for almost empty classes and directions.

Table 3.2: Computational complexity of each method.

Method	Complexity per iteration	# of iterations until convergence
HDP-CGS	$O(TFKM^3 + FKDM^3)$	50
LDA-VB	$O(TFKDM^2 + FDM^3)$	15
IVA	$O(TFM^3)$	50
ICA-perm.	$O(TFM^2)$	50
TF-perm	$O(TFNM + TFN^2)$	50

3.4.2 Separation results

Figures 3.8–3.11 show the separation results for simulated and recorded mixtures with two or three sources. The bars are grouped by the microphone number M for each method. The SDR scores are averaged over the source interval θ because the interval θ makes little difference in the separation quality. The color represents each method according to the legend in the rightmost figures. In general, a longer reverberation time degrades the SDR of all methods. A comparison of Figures 3.8 with 3.9, and Figures 3.10 with 3.11 shows that a larger number of sources in the observed mixture degrades the separation quality of the respective sources.

Our method is superior to or competitive with the other methods when $M = 4$ and 8. In particular, HDP-CGS tends to produce better SDR than LDA-VB. This is as expected because LDA-VB has more than N masks with non-negligible weights due to local optima, which results in the limited SDR scores. In contrast, the performance of our method is limited especially when $M = 2$. This is explained as follows. Even though the microphone number is small $M = 2$, the proposed approach separates the sources considering a variety of possible numbers of sources with the limited dimensionality of the observation. This source number uncertainty limits the performance of HDP-CGS. On the other hand, linear models including ICA and IVA can assume that the possible source number is two when $M = 2$. The determined problem of two-source and two-microphone is also suitable for the linear models in terms of the model complexity. Thus, the $M = 2$ setup is advantageous for linear models. Similarly, TF-perm uses the same number of TF masks as that of the sources. This assumption improves the separation quality of TF-perm method. Another remaining issue is the reverberation. We can note that the long reverberation (600 ms) in the recorded mixtures deteriorates the separation quality of any methods. To cope with these situations, an explicit model for the reverberation is left for

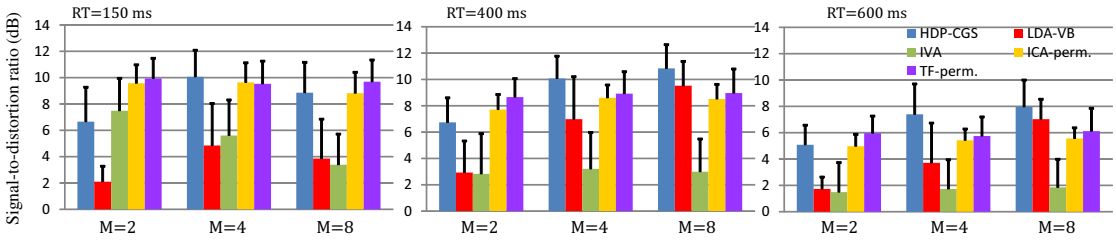


Figure 3.8: Separation results for simulated mixtures with two sources. Larger value means better separation. Bars are the means, and the segments are the standard deviations. Color represents each method. Left: RT = 150 (ms); middle: RT = 400 (ms); right: RT = 600 (ms).

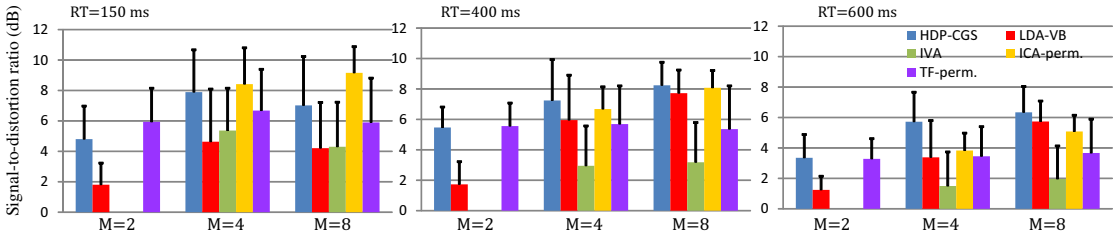


Figure 3.9: Separation results for simulated mixtures with three sources.

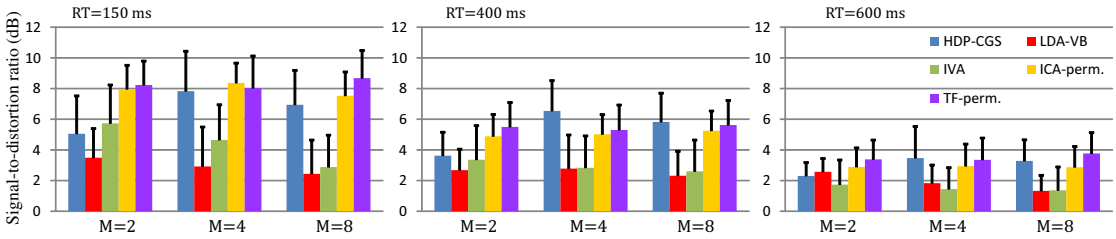


Figure 3.10: Separation results for recorded mixtures with two sources.

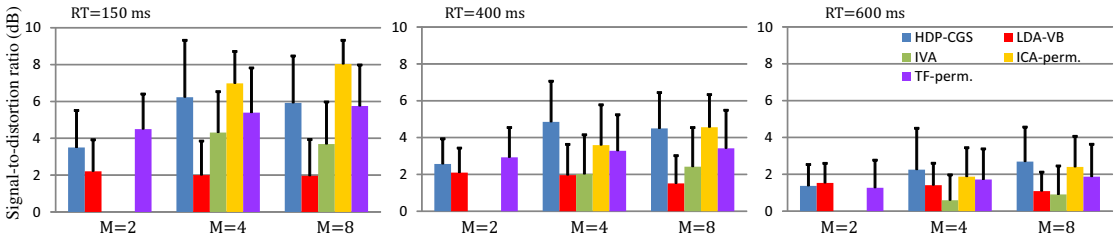


Figure 3.11: Separation results for recorded mixtures with three sources.

future work.

The performance with the recorded mixtures in Figures 3.10 and 3.11 is worse than that

with the simulated mixtures in Figures 3.8 and 3.9. This is because the recorded audio contains more reverberation in the lower frequency region than simulated mixtures. As shown in the appendix, the energy of the reverberation in the impulse responses used to generate the simulated mixtures is attenuated in the lower frequency range. In contrast, the recorded mixtures preserve the lower frequency reverberation of the environments. The intensity of the reverberation in the low frequency region severely affects the separation and localization performance because the subspace structure shown in Figure 3.2 is originally vague and is further disturbed by the reverberation. Furthermore, the SDR score is likely to be influenced from the disturbance of the separation quality in the lower frequency region because speech signals concentrate their power on the lower part in the frequency domain.

3.4.3 Localization results

Figures 3.12–3.15 show the localization results of HDP-CGS and LDA-VB in terms of the absolute errors of the localization results. Similarly to the separation results, the larger number of microphones improves the localization performance while the reverberation tends to affect the localization due to the reflection of the sounds. The errors in LDA-VB is more prominent than those in HDP-CGS because the posterior probability of $w_k = d$ can fall into a local optimum with the variational Bayesian inference of LDA-VB.

For some applications, the localization resolution specified by the steering vectors (5° in our experiment) may be insufficient. We can apply the following post-processing to the separated sound image $\hat{\mathbf{x}}_{tf}^{d}$ to enhance the localization resolution. Let $\mathbf{R}_f^d := \sum_t \hat{\mathbf{x}}_{tf}^{d}$ be the autocorrelation of the sound image and $\hat{\mathbf{q}}_{fd}$ be the eigenvector associated with the largest eigenvalue of $\mathbf{R}_f^d$. The vector $\hat{\mathbf{q}}_{fd}$ is a clue to investigate the direction of the sound source since this vector is parallel to one of the subspaces illustrated in Figure 3.2. The direction that matches $\hat{\mathbf{q}}_{fd}$ is investigated by interpolating the given steering vectors of adjacent directions, $\mathbf{q}_{fd}$ and $\mathbf{q}_{fd\pm1}$ (Matsumoto et al. 2003, Nakamura et al. 2013).

3.4.4 Source number estimation results

Figures 3.16–3.19 show the source number estimation results with HDP-CGS, LDA-VB, and Stereo. Each figure show the histogram of source number estimates for each microphone

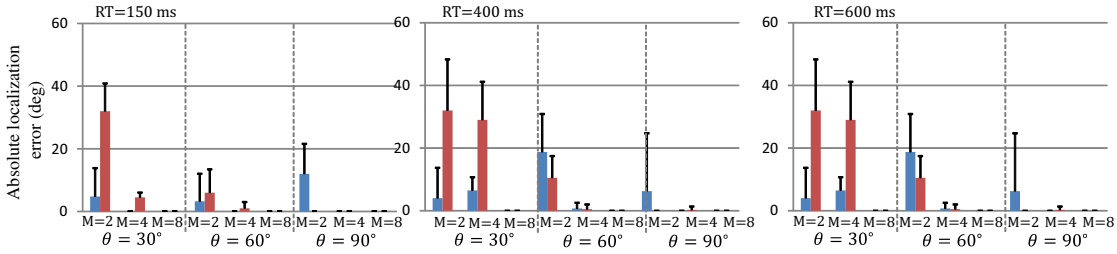


Figure 3.12: Localization results for simulated mixtures with two sources in terms of absolute errors. Smaller value means better localization. Bars are the means, and the segments are the standard deviations. Color represents each method. Left: RT = 150 (ms); middle: RT = 400 (ms); right: RT = 600 (ms).

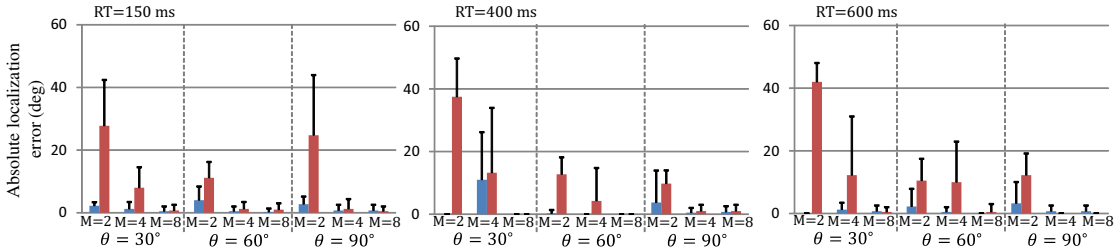


Figure 3.13: Localization results for simulated mixtures with three sources.

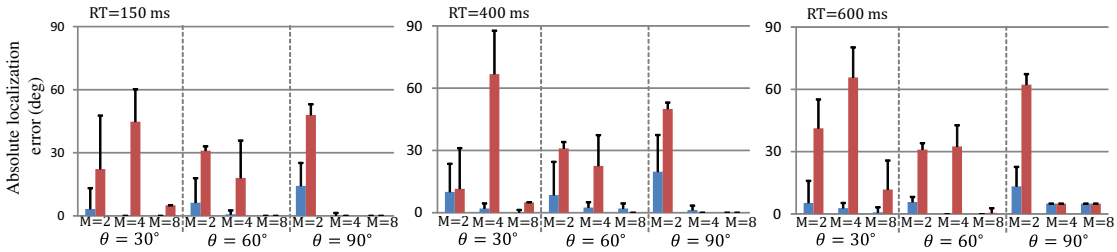


Figure 3.14: Localization results for recorded mixtures with two sources.

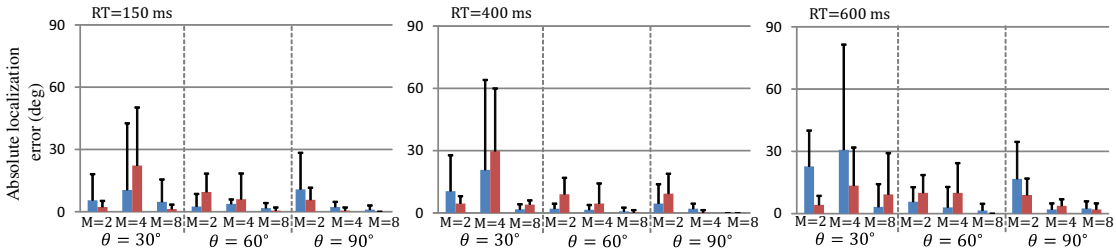


Figure 3.15: Localization results for recorded mixtures with three sources.

number and reverberation. Note that the results of Stereo is only presented for $M = 2$ case. The results are merged in terms of θ for this evaluation because this parameter made little

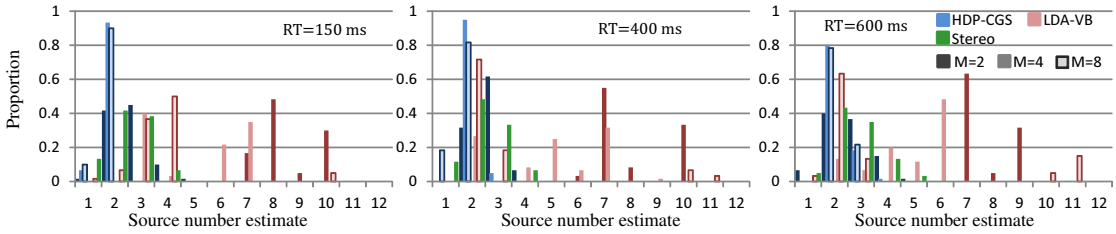


Figure 3.16: Source number estimation results for simulated mixtures with two sources ($N = 2$). Each bar represents the proportion of source number estimates. Color represents each method, and shade represents the number of microphones. Stereo method is only for $M = 2$. Left: RT = 150 (ms); middle: RT = 400 (ms); right: RT = 600 (ms).

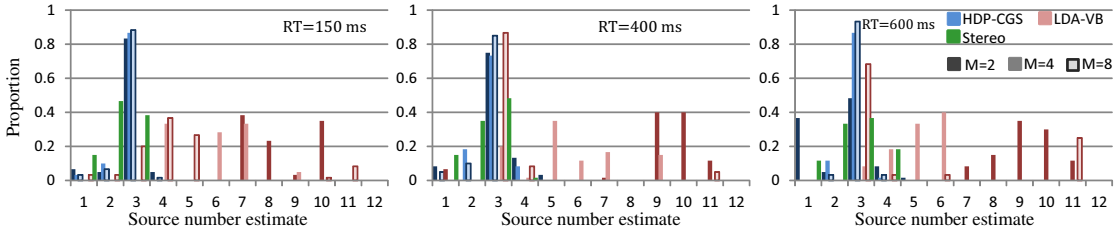


Figure 3.17: Source number estimation results for simulated mixtures with three sources ($N = 3$).

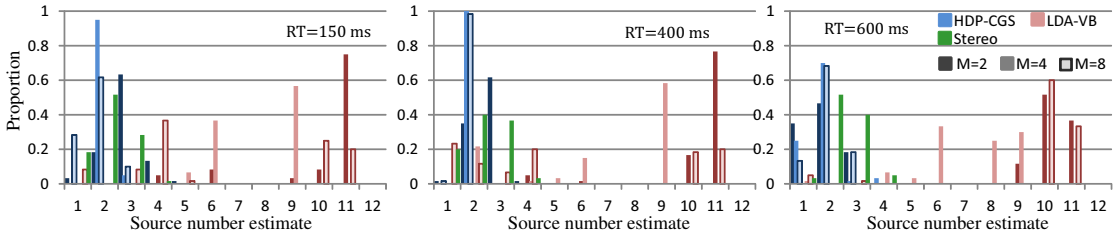


Figure 3.18: Source number estimation results for recorded mixtures with two sources ($N = 2$).

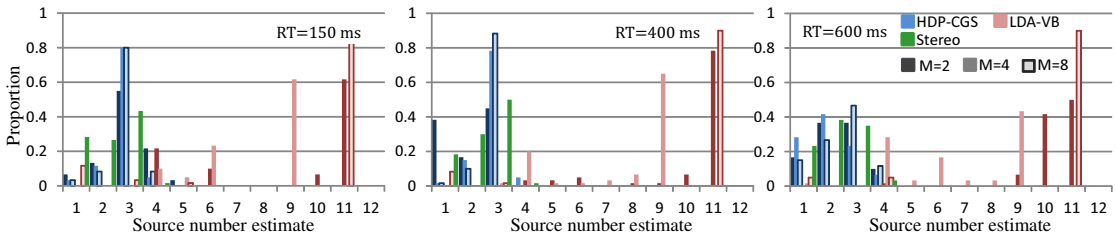


Figure 3.19: Source number estimation results for recorded mixtures with three sources ($N = 3$).

difference to the source number estimation performance. An ideal result of the estimation is that the bar is concentrated at the ground truth source number N.

A comparison of HDP-CGS and LDA-VB reveals that HDP-CGS clearly outperforms LDA-VB because the blue bars are mostly located at the true source number where as the red bars overestimate the number of sources. These results demonstrate that the CGS works well for source number estimation because it avoids local optima of the latent space, unlike variational Bayes inference. The results of Stereo tend to have a larger variance than HDP-CGS with the $M = 2$ case. This is considered because the observation model of Stereo uses only the phase difference between the two microphones and thus the TF mask generation is sometimes unstable. This makes it difficult to set a static threshold for the source counting in general setups. Three points in particular are observed. (1) VB tends to estimate more sources than CGS apparently because local optima obtained by VB have many tails in the posterior weights, which prevents correct source number estimation. (2) A larger number of microphones contributes to a better estimation with HDP-CGS. This means the number of microphones affects source number estimation as well as source separation quality.

HDP-CGS sometimes underestimates the source number when M is small and reverberation time is large. This is because the reverberation component is led to merge with most-weighted TF mask due to HDP prior that encourages a sparsity of activated masks. Thus, the ratio between the largest weight P'_1 and the second largest weight P'_2 is maximized, where the notation P' comes from Section 3.3.3. On the other hand, Stereo can estimate a larger source number as long as the threshold for the TF mask weight is accurately configured. For the improvement of this underestimation of HDP-CGS, more sophisticated source number estimation mechanism may be necessary.

3.4.5 Discussion and future work

The experiments revealed that our method can outperform state-of-the-art methods in terms of separation quality. In addition, our method is capable of robust source number estimation from a multichannel mixture even in a reverberant environment thanks to the CGS.

They also show that reverberation particularly affects the separation quality. We may incorporate a reverberation reduction technique such as (Yoshioka et al. 2011) for further improved

performance. Our method assumes non-moving sound sources. The use of a hidden Markov model would be a natural way to cope with moving sound sources (MacKay 1997) as it would make the direction indicator w_k a time-series sequence. For source number estimation, a model selection approach, such as (Fujimaki and Morinaga 2012), may be useful.

We used the measured impulse responses from the directions we consider as prior information about the microphone array we use. Reducing the necessary prior information about the microphone array can also be enumerated as the future directions. For example, the impulse responses can be simulated from the position of microphones or obtained through more casual and automatic calibration.

3.5 Summary

Our sound source localization and separation method using a microphone array achieves the decomposition function that is essential to CASA systems in a unified manner based on hierarchical Dirichlet process. Source separation experiments using simulated and recorded mixtures under various conditions demonstrated that our method outperforms state-of-the-art methods without a priori source number knowledge.

Chapter 4

Bayesian Nonparametric Multichannel Sound Source Separation, Localization, and Dereverberation

This chapter presents a unified method for sound source separation, localization and dereverberation to overcome the performance degradation due to reverberation in observed mixture signals. Unlike the method presented in Chapter 3 ignoring the reverberation in the observation process, the method developed in this chapter uses autoregressive (AR) processes to model the reverberation for achieving the dereverberation function.

4.1 Introduction

In actual acoustic environments, the microphone array processing for sound source separation and localization often faces the uncertainty about the number of sources as well as reverberation, as illustrated in Figure 4.1. Regarding the source number issue, state-of-the-art source separation methods using independent component analysis (ICA) (Hyvärinen et al. 2001) or independent vector analysis (IVA) (Lee et al. 2007) are limited to cases where the sources do not outnumber the microphones. When this assumption is violated, i.e., on the underdetermined condition, we need to use non-linear models such as time-frequency (TF) masking (Yilmaz and Rickard 2004, Mandel et al. 2010, Araki et al. 2010, Sawada et al. 2011). In addition, the separation and localization performance are degraded when the observed signals are highly reverberant. A common approach to mitigating this problem is to incorporate

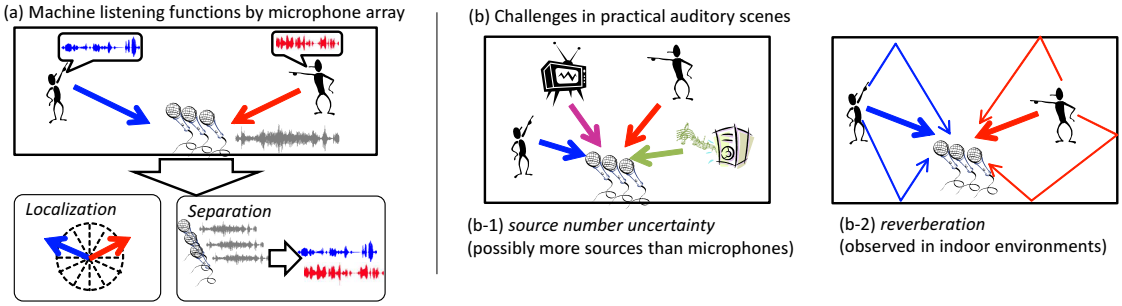


Figure 4.1: (a) Illustration of sound source separation and localization using a microphone array. (b) Acoustical challenges that limit performance and applicability of microphone array processing. Since conventional microphone array algorithms require several assumptions about the environment, we need a priori knowledge about the environment or the ability to identify the environment so as to choose appropriate algorithms, which is generally a difficult task. Our method is designed to handle two situations. (b-1) Source number uncertainty with two elements. First, the number of sources is a critical parameter for most source separation models. Second, the observation can be underdetermined, i.e., the sources outnumber the microphones. This situation is problematic for multichannel dereverberation, as explained in Section 4.2. (b-2) Reverberation leads to separation and localization degradation.

dereverberation processing (Huang et al. 2005, Buchner and Kellermann 2010, Yoshioka et al. 2011, Togami et al. 2013). However, existing methods have limited applicability because they are based on the assumption that the number of sources is known. Furthermore, these methods require that the observation process be overdetermined; that is, the number of sources must be less than the number of microphones. This requirement is possibly unsatisfied in practice. Since a priori knowledge about the environment including the source number is often difficult to obtain, an all-round method applicable to a wide range of environments is desirable.

This chapter presents a Bayesian method for analyzing multi-source reverberant acoustic scenes with an unknown number of sound sources by jointly performing sound source localization, separation, and dereverberation. This method is based on a previously proposed Bayesian model for localization and separation (Otsuka et al. 2012, 2013) that is capable of dealing with any number of sources under possibly underdetermined conditions. The model has been extended to incorporate the dereverberation process in a unified manner. Unlike existing joint source separation and dereverberation methods (Huang et al. 2005, Buchner and Kellermann 2010, Yoshioka et al. 2011, Togami et al. 2013), our model can handle situations in which the source number is unavailable and in which the conditions are underdetermined.

Localization and separation are formulated as two types of clustering problems, and the reverberation signals are modeled as AR processes. We tackle the model selection problem in both separation and dereverberation caused by the source number uncertainty. Bayesian nonparametrics plays a key role in our method. Thanks to the infinitely extensible flexibility of the Bayesian nonparametrics-based model, our model bypasses the model the selection problem and is able to handle any number of sound sources in a consistent manner.

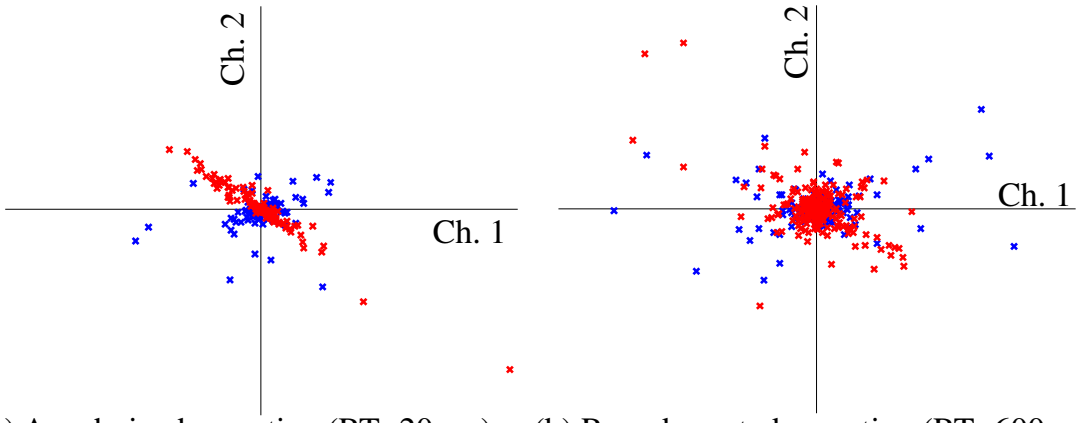
The rest of this chapter is organized as follows. Section 4.2 introduces our Bayesian nonparametric framework for model selection regarding the separation and dereverberation problem for any number of sources even under underdetermined conditions. Following the discussion in Section 4.2, Section 4.3 presents our separation and dereverberation model, which is applicable to an arbitrary number of sources without the source number being specified in advance. Section 4.4 presents the results of our evaluation of separation and dereverberation performance in comparison with a state-of-the-art separation and dereverberation method. Section 4.5 summarizes the key points of this chapter.

4.2 Bayesian Nonparametric Multichannel Dereverberation

This section incorporates Bayesian nonparametrics with multichannel dereverberation technique using AR processes. We explain why conventional multichannel dereverberation methods are limited to overdetermined conditions. Then, an infinite mixture of AR processes based on Bayesian nonparametrics is introduced to overcome this limitation.

The goal of dereverberation and separation is explained as follows using Eq. (3.1). Given the reverberant observation of mixture signal $\mathbf{x}_{tf}$ in Eq. (3.1), we would like to estimate the source signals $\mathbf{s}_{tf}$, or equivalently the source image signals extracted in the previous chapter. For the dereverberation, we need to estimate and remove the reverberation component from observed mixture $\mathbf{x}_{tf}$, where the second term in Eq. (3.1) is the reverberation component. For the separation, we have to extract each source in the first term in Eq. (3.1).

This section is organized as follows. First, we explain how multiple sources are observed with a microphone array. Section 4.2.1 explains how reverberation affects the separation performance. Section 4.2.2 reviews the mechanism of existing reverberation modeling based on an AR process and shows that the reverberation is accurately modeled only when the mix-



(a) Anechoic observation (RT 20 ms) (b) Reverberant observation (RT 600 ms)

Figure 4.2: Comparison of TF point subspaces between anechoic and reverberant environments. Axes are the same as those in Figure 3.2. (a) Scatter plot of TF points for reverberation time (RT) of 20 (ms). Subspaces corresponding to signals from different directions are sharp and distinct. (b) Scatter plot of TF points for RT of 600 (ms). Reverberation obscures subspaces, complicating the clustering of TF points for TF mask estimation.

ture is overdetermined (Yoshioka et al. 2011, Gorokhov and Loubaton 1997). Section 4.2.3 presents our novel idea for modeling the reverberation for an arbitrary number of sources, which is to model the reverberation signal as an infinite mixture of AR processes based on Bayesian nonparametrics. Section 4.2.4 compares our dereverberation method with those of several existing methods.

4.2.1 Influence of reverberation on separation methods

In Chapter 3, Bayesian nonparametric sound source separation and localization method is developed to cope with the source number uncertainty issue. While effective in less reverberant environments, the separation performance is significantly degraded when there is severe reverberation. If the observed mixture signal contains reverberation, the subspace structures tend to be indistinguishable from each other, as depicted in Figure 4.2. Thus, separation performance is degraded because the clustering of TF points becomes difficult.

A Bayesian nonparametric model using the likelihood in Eq. (3.4) copes with reverberation in two ways. First, the effect of reverberation is absorbed to some extent by the full-rank covariance matrix (Duong et al. 2010) in Eq. (3.4) because a full-rank covariance matrix can

fit the vague subspace, to some extent. Second, some reflected sound can be separated as a distinct source by using extra masks. In particular, early reflections can be detected with such extra masks, where early reflections are a few strong reflections that are observed sporadically immediately after direct sound arrival. If such a salient reflection is observed from a direction that is different from all the source directions reverberation can partly be removed. Nevertheless, reverberation still affects separation performance, as shown in Section 3.4, which motivates us to explicitly incorporate dereverberation into microphone array processing.

4.2.2 Modeling reverberation component with autoregressive process

Now we return to the original observation model presented in Eq. (3.1) and explain how we model the reverberation component, the second term on the right-hand side of Eq. (3.1), by using an AR process of the multichannel observations. We estimate the reverberant component needed to achieve dereverberation, that is, to remove the reverberant component from the observed signal, thereby uncovering the latent subspace structure formed by the propagation coefficients of direct sounds $\mathbf{B}_f^0$ for accurate separation and localization. The derivation of the model of the reverberant component clarifies why the overdetermined condition has been assumed in previous work to achieve dereverberation (Huang et al. 2005, Buchner and Kellermann 2010, Yoshioka et al. 2011). We extend the AR processes so that the model can fit the reverberant components even in underdetermined situations by taking account of the sparsity of sound sources, as we did for separation in the previous section.

We model the reverberant component as an AR process of order R:

$$\begin{aligned}\sum_{j=1}^{J-1} \mathbf{B}_f^j \mathbf{s}_{t-j,f} &= \sum_{r=1}^{R} \mathbf{A}_{fr}^H \mathbf{x}_{t-r,f} \\ &:= \mathbf{A}_f^H \overline{\mathbf{x}}_{tf},\end{aligned} \tag{4.1}$$

where $\mathbf{A}_{fr} \in \mathbb{C}^{M\times M}$ is the AR coefficients of the rth lag, and $\mathbf{x}_{t-r,f}$ is the observed mixture at time $t-r$. We put a comma in the subscript when some operation is involved to clarify the notation, e.g., between $t-r$ and f (as in $\mathbf{x}_{t-r,f}$), which represent the time frame and frequency bin, respectively, and omit the comma if the correspondence of the notation is clear. In the

second line of Eq. (4.1), we used stacked notation, defined as

$$\mathbf{A}_f := [A_{f1}^H, \ldots, A_{fR}^H]^H \in \mathbb{C}^{MR \times M},$$
$$\overline{\mathbf{x}}_{tf} := [\mathbf{x}_{t-1,f}^H, \ldots, \mathbf{x}_{t-R,f}^H]^H \in \mathbb{C}^{MR}.$$

While effective for overdetermined situations, this model does not work effectively when there are more sources than microphones. To clarify this, we next show that the overdetermined condition is a prerequisite for the above model to accurately represent the reverberation component.

The relationship between propagation coefficients $\mathbf{B}_f^j$ and AR coefficients $\mathbf{A}_f$ is clarified by stacking Eq. (3.1) to remove $\overline{\mathbf{x}}_{tf}$ from Eq. (4.1). By stacking Eq. (3.1), we get

$$\begin{bmatrix} \mathbf{x}_{t-1,f} \\ \mathbf{x}_{t-2,f} \\ \vdots \\ \mathbf{x}_{t-R,f} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_f^0 & \cdots & \mathbf{B}_f^{J-1} & & \mathbf{0} \\ & \mathbf{B}_f^0 & \cdots & \mathbf{B}_f^{J-1} & \\ & \ddots & & \ddots & \\ \mathbf{0} & & \mathbf{B}_f^0 & \cdots & \mathbf{B}_f^{J-1} \end{bmatrix} \begin{bmatrix} \mathbf{s}_{t-1,f} \\ \mathbf{s}_{t-2,f} \\ \vdots \\ \mathbf{s}_{t-R-L+1,f} \end{bmatrix}$$
$$:= \overline{\mathbf{B}}_f \bar{\mathbf{s}}_{t-1,f}. \tag{4.2}$$

Let $\overline{\mathbf{B}}_f \in \mathbb{C}^{MR \times N(R+J-1)}$ be a large matrix consisting of propagation coefficients and $\bar{\mathbf{s}}_{t-1,f} \in \mathbb{C}^{N(R+J-1)}$ be the sequence of source signals on the right side of Eq. (4.2). Using this notation, we can derive the following equation by substituting Eq. (4.2) into the right side of Eq. (4.1):

$$\left[\mathbf{B}_f^1, \ldots, \mathbf{B}_f^{J-1}, \mathbf{0}\right] \bar{\mathbf{s}}_{t-1,f} = \mathbf{A}_f^H \overline{\mathbf{B}}_f \bar{\mathbf{s}}_{t-1,f}. \tag{4.3}$$

We can interpret Eq. (4.3) as follows: if we can set the AR coefficients $\mathbf{A}_f$ such that Eq. (4.3) is satisfied for any source signal $\forall \mathbf{s}_{tf}$, we can fit the reverberant component with the autoregression of the sequence of observations. The equality is fulfilled if the AR coefficients satisfy

$$\mathbf{A}_f^H = \overline{\mathbf{B}}_f^{\dagger} \left[\mathbf{B}_f^1, \ldots, \mathbf{B}_f^{J-1}, \mathbf{0}\right],$$

where $\mathbf{C}^{\dagger}$ is the left inverse of matrix $\mathbf{C}$ such that $\mathbf{C}^{\dagger}\mathbf{C} = \mathbf{I}$. The existence of the left inverse of $\overline{\mathbf{B}}_f$ requires two conditions:

- The number of rows is larger than the number of columns.
- The column vectors of $\overline{\mathbf{B}}_f$ are linearly independent.

These conditions are paraphrased as follows. The first condition is rewritten as $MR > N(R + J - 1)$. This is satisfied when $M > N$ and R is sufficiently large compared with L. For the second condition, the linear independence of the columns in each propagation matrix $\mathbf{B}_f^j$ is the sufficient condition. This is in practice satisfied when the sound sources are located in distinct directions. In summary, the AR process of the observed mixture can fit the reverberant components under three conditions:

1. The mixture is overdetermined ($M > N$).
2. The order of AR process R is sufficiently large compared with the amount of reverberation J.
3. The sources are located in distinct directions.

These conditions have been confirmed elsewhere, e.g., (Yoshioka et al. 2011, Gorokhov and Loubaton 1997). We derived these conditions using the notations above to clarify our idea for extending the AR-based reverberation modeling for any number of sources in the following section.

Note that while the overdetermined mixture process ($M > N$) is required to guarantee the equality in Eq. (4.1), this AR model can, in practice, approximate the reverberant component, as shown by the experimental results in Section 4.4. We can interpret this in two ways. First, we can reduce the difference between the both sides of Eq. (4.3) empirically by using the provided source signals $\mathbf{s}_{tf}, (t = 1, \ldots, T)$ and setting the AR coefficients $\mathbf{A}_f$ such that a certain criterion (for example, the least squared error) is minimized. Second, the estimation of AR coefficients $\mathbf{A}_f$ is carried out by reducing the correlation between previous observations $\overline{\mathbf{x}}_{tf}$ and present observation $\mathbf{x}_{tf}$ or the present constituent source signals $\mathbf{s}_{tf}$ (Takeda et al. 2012). When the sources outnumber the microphones, although complete decorrelation is impossible, existing methods can partly remove the temporal correlation that results from the reverberation.

4.2.3 Infinite mixture of AR processes for dereverberation of arbitrary number of sources

We extend the AR process so that the reverberation component of any number of sources can be constructed from the previous observations of the mixture signal. Our idea is to use the sparsity of source power at each TF point and to incorporate Bayesian nonparametrics again to enhance the flexibility of the model to form the reverberant component.

Since the left inverse of $\overline{\mathbf{B}}_f$ is a key requirement for the AR model to fit the reverberant component and since the rows should outnumber the columns in matrix $\overline{\mathbf{B}}_f$, that is, $MR > N(R+J-1)$, we need some modification to cope with any number of sources, even when $M \leq N$. In a similar way that Eq. (3.2) is approximated by Eq. (3.3), we can reduce the number of columns in $\overline{\mathbf{B}}_f$ by approximating the right side of Eq. (4.2) as

$$\begin{aligned}\overline{\mathbf{B}}_f \overline{\mathbf{s}}_{t-1,f} &\approx \begin{bmatrix} \mathbf{b}^0_{fk_{t-1,f}} & \cdots & \mathbf{b}^{J-1}_{fk_{t-J+1,f}} & & \mathbf{0} \\ & \mathbf{b}^0_{fk_{t-2,f}} & \cdots & \mathbf{b}^{J-1}_{fk_{k-J}} & \\ & \ddots & & \ddots & \\ \mathbf{0} & & \mathbf{b}^0_{fk_{t-R+1,f}} & \cdots & \mathbf{b}^{J-1}_{fk_{t-R-J+1,f}} \end{bmatrix} \begin{bmatrix} s_{t-1,f} \\ s_{t-1,f} \\ \vdots \\ s_{t-R-J+1,f} \end{bmatrix} \\ &:= \overline{\mathbf{B}}'_{t-1,f} \overline{\mathbf{s}}'_{t-1,f}, \end{aligned} \tag{4.4}$$

where $\overline{\mathbf{B}}'_{t-1,f} \in \mathbb{C}^{MR \times (R+J-1)}$ and $\overline{\mathbf{s}}'_{t-1,f} \in \mathbb{C}^{R+J-1}$. This approximation results from the power sparsity of source signals in the TF domain. Thanks to this approximation, the left inverse of $\overline{\mathbf{B}}'_{t-1,f}$ is tractable, so the AR model can fit the reverberant component even when $M \leq N$. However, this approximation undermines the stationarity of the propagation coefficients; $\overline{\mathbf{B}}'_{t-1,f}$ varies depending on time index t. This means that stationary AR coefficients $\mathbf{A}_f$ can no longer form the reverberant component for all time frames. Strictly speaking, we have to switch the AR coefficients on the basis of the history of the source indices $k_{t-R-J+1,f}, \ldots, k_{t-1,f}$, or, more specifically, the history of directions of the dominant sources at each time. However, exact switching on the basis of the history of directions is prohibitive because the possible number of combinations of the previous directions grows exponentially large. This combinatorial explosion would affect the estimation of the AR coefficients because only a small portion of the observed sequence is available for the parameter estimation corresponding to each direction history.

We therefore develop an infinite mixture of AR processes to achieve a balance between the

capability of the model and the feasibility of parameter estimation (Fox et al. 2011). The infinite mixture model is constructed using the Dirichlet process (DP) (Ferguson 1973), which is a widely used Bayesian nonparametrics framework, and a tractable parameter inference procedure is derived using the Chinese restaurant process (Aldous 1985, Neal 2000). We model the AR coefficients as $\mathbf{A}_{fl_{tf}}$, where l_{tf} is an index of the switching, which varies over time. Index l_{tf} is unbounded, so the model can handle an arbitrary switching of the AR processes to fit the reverberant component for $M < N$. At the same time, the prior probability of l_{tf} penalizes the emergence of unnecessary switching, which prevents the overfitting in the parameter estimation. The aim with this model is to stabilize the estimation of AR coefficients while preserving the flexibility of the model rather than to exactly trace the history of dominant directions and to generate the corresponding AR coefficients. Note that this model encompasses the conventional single AR model; when $l_{tf} = 1$ for $\forall t$, the model suffices to remove the reverberation from an overdetermined mixture.

While $\overline{\mathbf{x}}_{tf}$ has been used to denote the previous observations (from $\mathbf{x}_{t-1,f}$ to $\mathbf{x}_{t-R,f}$), in practice, some samples close to time t are omitted from $\overline{\mathbf{x}}_{tf}$ so as to form the reverberation component at time t:

$$\overline{\mathbf{x}}_{tf} = [\mathbf{x}_{t-\delta-1,f}^H, \ldots, \mathbf{x}_{t-\delta-R,f}^H]. \tag{4.5}$$

This is because the samples in adjacent time frames are correlated due to overlapped windows in STFT. If $\delta = 0$, there is a risk of undesirable removal of the direct component in $\mathbf{x}_{tf}$ because of the correlation between $\mathbf{x}_{tf}$ and $\mathbf{x}_{t-1,f}$. Nevertheless, the derivation of an infinite mixture of AR processes above holds even when $D > 0$. Thus, $\overline{\mathbf{x}}_{tf}$ is defined as Eq. (4.5), hereafter.

4.2.4 Comparison to existing dereverberation methods

The discussion of AR-based dereverberation in Section 4.2.3 leads to a mixture of AR processes. The notation $\mathbf{A}_{fl_{tf}}$ can be viewed as a time-varying AR coefficient since the index to specify an AR process l_{tf} depends on time t. Time-varying AR coefficients are equivalent to time-varying propagation of source signals, as represented in Eq. (4.4). While some nonlinear dereverberation methods can be used for the setup of the time-varying propagation (Lebart et al. 2001, Habets et al. 2008), they do not work well in combination with multichannel

Table 4.1: Notations.

Symbol	Meaning
t	Time frame index from 1 to T
f	Frequency bin from 1 to F
k	Class index (positive integer)
K	Instantiated number of classes during inference
d	Direction index from 1 to D
M	Number of microphones
N	Number of sound sources
l	AR coefficient index (positive integer)
$\mathbf{x}_{tf}$	Observed M-dimensional complex column vector
z_{tf}	Class indicator for time t and frequency bin f
$\boldsymbol{\pi}_t$	Class proportion for time frame t
w_k	Direction indicator for class k
$\boldsymbol{\varphi}$	Direction proportion for all classes
λ_{tf}	Inverse scale parameter for $\mathbf{x}_{tf}$
$\mathbf{H}_{fd}$	Precision matrix corresponding to subspace for direction d and frequency bin f
$\bar{\mathbf{x}}_{tf}$	Stacked notation of R previous observations for AR process at time t and frequency f
v_{tf}	AR class indicator for time t and frequency bin f
$\mathbf{A}_{fl}$	lth AR coefficient for frequency f
n_{tk}	Number of TF points assigned to class k at time frame t
n_{fk}, n_{fd}	Number of TF points for frequency bin f of class k or direction d, respectively
c_d	Number of classes assigned to direction d
r_{fl}	Number of TF points assigned to AR class l for frequency bin f

source separation because the nonlinear processing destroys the subspace structures shown in Figure 3.2.

Togami et al. develop a dereverberation method robust against time-varying propagation caused by the speaker head movements (Togami et al. 2013, 2012). Their approach is to treat the AR coefficients as probability variables that absorb fluctuations in the propagation coefficients. This is similar to the approach in our model: our Bayesian model also treats the AR coefficients as probability variables, as described in the following section. The difference is that their method uses a single AR process for each frequency bin and evaluates the dispersion of the reverberation components whereas our method uses multiple AR processes for each frequency bin. In fact, our model subsumes their method in that each AR coefficient is a probability variable and that setting $l_{tf} = 1$ for $\forall t$ reduces it to a single-AR model.

4.3 Unified Model for Dereverberation and Separation of Arbitrary Number of Sources

This section presents our unified model for a dereverberation and separation method capable of handling an unbounded number of sources and underdetermined conditions. Our method uses a TF masking-based separation and localization framework (Otsuka et al. 2013) and the reverberation model incorporating an infinite mixture of AR processes described in Section 4.2.3.

The problem to be solved with our algorithm is specified as follows.

Input: Multichannel observation of sound source mixture denoted by $\mathbf{x}_{tf}$ ($t = 1, \ldots, T, f = 1, \ldots, F$).

Output: Separated source signals arriving from distinct directions denoted by $\mathbf{x}_{tf}^{d}$, where d is direction index.

Assumptions:

(1) Number of sources in observation is unspecified. (2) Steering vector $\mathbf{q}_{fd}$ of microphone array is provided for each frequency f and direction d. (3) Sources stay at the same position.

A steering vector is used for two reasons. First, the steering vector is necessary for sound source localization because the subspace depicted in Figure 3.2 is associated with a certain direction via this steering vector. Second, broadband sound source separation can be attained by implicitly resolving the permutation alignment (Sawada et al. 2004), as previously reported (Otsuka et al. 2012, 2013). We obtain the steering vectors for discrete directions, e.g., 5° resolution on the azimuth plane. Thus, the sound sources are localized with this resolution, and the direction is specified by an integer index $d = 1, \ldots, D$. Note that the steering vectors are determined independently of the environment but only dependent on the microphone array. The steering vectors for our microphone array are generated using impulse responses measured in an anechoic chamber, as noted in Section 4.4.1. We show that these steering vectors are useful for various reverberant setups in Section 4.4. Therefore, the instantaneous propagation coefficients, the column vectors of $\mathbf{B}_{f}^{0}$, are not necessarily equal to one of the steering vectors $\mathbf{q}_{fd}$ partly because the propagation coefficients are affected by the reverberation and partly because the sound may arrive from a direction between the grids on the azimuth

plane. Thus, our problem is a blind source separation problem in the sense that the constituent source signals are extracted using the observed mixture without access to the propagation coefficients and source signals while the steering vectors are provided as knowledge in the form of a microphone array.

The rest of this section has two main parts. Section 4.3.1 presents the generative model, which incorporates the infinite TF masks and infinite AR processes explained in Sections 3.2.2 and 4.2.3, respectively. Section 4.3.2 explains the inference based on the Markov chain Monte Carlo (MCMC) method for the parameter estimation used to obtain the constituent signals. The notations used in this section are summarized in Table 4.1.

4.3.1 Generative process

The likelihood of a multichannel reverberant signal mixture is modeled as a multivariate complex normal distribution:

$$\mathbf{x}_{tf}|\overline{\mathbf{x}}_{tf}, z_{tf}, \tilde{\mathbf{w}}, v_{tf}, \tilde{\mathbf{A}}, \lambda_{tf}, \tilde{\mathbf{H}} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{x}_{tf}|\mathbf{A}_{fv_{tf}}^{H}\overline{\mathbf{x}}_{tf}, \lambda_{tf}\mathbf{H}_{fw_{z_{tf}}}), \tag{4.6}$$

where the mean and precision correspond to the reverberant component and direct sound at time t and frequency f, respectively. A bold symbol with a tilde means the set of all variables, e.g., $\tilde{\mathbf{x}} = \{\mathbf{x}_{tf}|\forall t, f\}$. If $z_{tf} = k$ and $w_k = d$, the kth source signal coming from direction d is dominant at time t and frequency f. Matrix $\mathbf{H}_{fd}$ determines the subspace corresponding to direction d in the frequency bin. Similarly, if $v_{tf} = l$, the lth AR coefficient $\mathbf{A}_{fl}$ is used to form the reverberant component at time t and frequency f.

The sound source separation and localization correspond to the estimation of the latent variables $\tilde{\mathbf{z}}$ for the TF masks and $\tilde{\mathbf{w}}$ for the direction of each mask. The dereverberation is carried out by subtracting the reverberant component $\mathbf{A}_{fv_{tf}}^{H}\overline{\mathbf{x}}_{tf}$ from the observation $\mathbf{x}_{tf}$.

The choice of prior distributions of latent variables related to the separation and localization basically follows those used previously (Otsuka et al. 2013). These priors are selected so that a tractable inference procedure can be derived. The prior of z_{tf} follows a hierarchical Dirichlet process (HDP), allowing for an unbounded number of classes.

$$\boldsymbol{\beta}|\gamma \sim \mathrm{GEM}(\gamma), \qquad \boldsymbol{\pi}_t|\alpha, \boldsymbol{\beta} \sim \mathrm{DP}(\alpha, \boldsymbol{\beta}), \qquad z_{tf}|\boldsymbol{\pi}_t \sim \boldsymbol{\pi}_t, \tag{4.7}$$

where GEM(γ) is the Griffiths-Engen-McCloskey distribution with concentration γ; DP($\alpha, \boldsymbol{\beta}$) denotes the DP with concentration α and base measure $\boldsymbol{\beta}$. We use gamma distribution priors for concentrations $\gamma \sim \mathcal{G}(\gamma|a_\gamma, b_\gamma)$ and $\alpha \sim \mathcal{G}(\alpha|a_\alpha, b_\alpha)$. The hyperparameters are set as $a_\gamma = 0.05, b_\gamma = 5, a_\alpha = 0.01$, and $b_\alpha = 1$. This hierarchical generative process is aimed at implicitly resolving the permutation ambiguity by synchronizing the dominance of a certain source throughout the frequency bin for each time frame (Otsuka et al. 2012). First, infinite-dimensional global class proportion $\boldsymbol{\beta}$ is generated. Each element represents the average weight of infinitely many classes throughout the spectrogram. Then, the time-wise class proportion $\boldsymbol{\pi}_t$ is sampled in accordance with $\boldsymbol{\beta}$. Again, $\boldsymbol{\pi}_t$ is an infinite-dimensional vector in which the elements represent the weights of infinite classes for specific time frame t. Finally, each z_{tf} is sampled in accordance with the time-wise class proportion $\boldsymbol{\pi}_t$.

Direction indicator w_k follows a categorical distribution, and its conjugate prior is a Dirichlet distribution:

$$\boldsymbol{\varphi}|\kappa \sim \mathcal{D}\left(\boldsymbol{\varphi}\middle|\frac{\kappa}{D}\mathbf{1}_D\right), \qquad w_k|\boldsymbol{\varphi} \sim \boldsymbol{\varphi}, \tag{4.8}$$

where $\mathbf{1}_D$ is a D-dimensional vector in which all elements are 1, and $\mathcal{D}(\cdot|\boldsymbol{\alpha})$ denotes a Dirichlet distribution with parameter $\boldsymbol{\alpha}$.

The priors for the parameters in the precision part of the likelihood in Eq. (4.6) are defined as follows so that the inference procedure is made tractable. The scale parameter λ_{tf} follows a gamma distribution, and the matrices depending on direction $\mathbf{H}_{fd}$ follow complex Wishart distributions:

$$\lambda_{tf}|a_{tf}, b_{tf} \sim \mathcal{G}(\lambda_{tf}|a_{tf}, b_{tf}), \tag{4.9}$$

$$\mathbf{H}_{fd}|\nu_{fd}, \mathbf{G}_{fd} \sim \mathcal{W}_{\mathbb{C}}(\mathbf{H}_{fd}|\nu_{fd}, \mathbf{G}_{fd}), \tag{4.10}$$

where $\mathcal{G}(a, b)$ denotes a gamma distribution with shape a and scale b, and $\mathcal{W}_{\mathbb{C}}(\nu, \mathbf{G}_{fd})$ means a complex Wishart distribution (Conradsen et al. 2003) with degree of freedom ν and scale matrix $\mathbf{G}_{fd}$. The hyperparameters are set as $a_{tf} = 1$, $b_{tf} = \mathbf{x}_{tf}^H \mathbf{x}_{tf}$, $\nu_{fd} = M$, and $\mathbf{G}_{fd}^{-1} = \mathbf{q}_{fd}\mathbf{q}_{fd}^H + \epsilon \mathbf{I}_M$, where $\mathbf{I}_M$ is an $M \times M$ identity matrix and $\mathbf{q}_{fd}$ is the anechoic steering vector corresponding to frequency f and direction d. The steering vector is normalized such that $\mathbf{q}_{fd}^H \mathbf{q}_{fd} = 1$. To make $\mathbf{G}_{fd}$ a positive definite matrix, $\epsilon \mathbf{I}_M$ is added, where $\epsilon = 0.01$ in our implementation.

The priors for the indicator of AR coefficients v_{tf} and the AR coefficients themselves are defined for each frequency bin f as

$$\boldsymbol{\varrho}_f|\zeta_f \sim \mathrm{GEM}(\zeta_f), \qquad v_{tf}|\boldsymbol{\varrho}_f \sim \boldsymbol{\varrho}_f, \tag{4.11}$$

$$\mathbf{A}_{fl}|\mathbf{M}_{fl}, \mathbf{K}_{fl}, \mathbf{L}_{fl} \sim \mathcal{MN}_{\mathbb{C}}(\mathbf{A}_{fl}|\mathbf{M}_{fl}, \mathbf{K}_{fl}, \mathbf{L}_{fl}), \tag{4.12}$$

where $\boldsymbol{\varrho}_f$ is the infinite-dimensional weight vector and v_{tf} is generated in proportion to the vector for each time. The prior for AR coefficients $\mathbf{A}_{fl} \in \mathbb{C}^{MR\times M}$ is the complex matrix normal distribution $\mathcal{MN}_{\mathbb{C}}(\mathbf{A}|\mathbf{M}, \mathbf{K}, \mathbf{L})$, where $\mathbf{M} \in \mathbb{C}^{MR\times M}$ is the mean matrix, $\mathbf{K} \in \mathbb{C}^{MR\times MR}$ is the row-wise precision matrix, and $\mathbf{L} \in \mathbb{C}^{M\times M}$ is the column-wise precision matrix. These hyperparameters are set as follows. A DP mixture of AR coefficients is independently constructed for each frequency bin because the complexity of the AR coefficient mixture may differ for each frequency bin in accordance with the acoustic propagation at the corresponding frequency. Therefore, the concentration parameter is also independently generated as $\zeta_f \sim \mathcal{G}(\zeta_f|a_{\zeta_f}, b_{\zeta_f})$, where $a_{\zeta_f} = 1$ and $b_{\zeta_f} = 1$. To assume an absence of reverberation a priori, the prior of $\mathbf{A}_{fl}$ is set as a zero mean variable, that is, $\mathbf{M}_{fl} = \mathbf{0}$. We also assume that the AR coefficients are uncorrelated in the prior. Therefore, $\mathbf{K}_{fl} = \mathbf{I}_{MR}$ and $\mathbf{L}_{fl} = \mathbf{I}_M$.

The matrix normal distribution is equivalently converted into a normal distribution using a vectorization operator and Kronecker product:

$$\mathcal{MN}_{\mathbb{C}}(\mathbf{A}|\mathbf{M}, \mathbf{K}, \mathbf{L}) \equiv \mathcal{N}_{\mathbb{C}}(\mathrm{vec}(\mathbf{A})|\mathrm{vec}(\mathbf{M}), \mathbf{L}\otimes\mathbf{K}), \tag{4.13}$$

where $\otimes$ denotes the Kronecker product operator. This property is used in the inference.

4.3.2 Inference using Markov chain Monte Carlo method

The inference of the posterior distribution of the latent variables ($\tilde{\mathbf{z}}, \tilde{\mathbf{w}}, \tilde{\mathbf{v}}, \tilde{\mathbf{A}}$, and so on) given observation $\tilde{\mathbf{x}}$ is carried out by sampling the values through a Markov chain Monte Carlo (MCMC) method. A partially collapsed Gibbs sampler (PCGS) (Park and van Dyk 2009) is used to accelerate and stabilize the mixing of the Markov chains. To be more precise, we found that collapsing matrix $\mathbf{H}_{fd}$ when possible makes inference more efficient.

The overall inference procedures using PCGS are summarized in Algorithm 1. Gibbs sampling (including PCGS) draws samples from a joint distribution of targeted variables by

Algorithm 1 PCGS-based inference

Initialize variables, as in Section 4.3.3
loop
  for all f, t in random order, update z_{tf} by Eq. (4.14)
  for $k \leftarrow 1, \ldots, K$ in random order, Update w_k by Eq. (4.17)
  for all f, d, update $\mathbf{H}_{fd}$ by Eq. (4.18)
  for all f, t, update v_{tf} by Eq. (4.19)
  for all f, l, update $\mathbf{A}_{fl}$ by Eq. (4.20)
  for all f, t, update λ_{tf} by Eq. (4.21)
  Update other hyperparameters $\alpha, \beta, \gamma, \kappa, \zeta_f$ (for $\forall f$).
end loop

stochastically updating some of the variables using the conditional probability distribution of the variables to be updated conditioned on the rest of variables. In the following, we present the conditional distributions used to update each variable and then explain the initialization of the latent variables.

The assignment of each TF point to a certain class z_{tf} is updated as follows. When the number of current sampled classes is K, the probabilities of $k = 1, \ldots, K+1$ are evaluated for use in updating z_{tf} by Eq. (4.14). Two variables, $\tilde{\boldsymbol{\pi}}$ and $\tilde{\mathbf{H}}$, are marginalized out. The marginalization over the time-wise class proportions $\tilde{\boldsymbol{\pi}}$ is described in (Teh et al. 2006). The marginalization over matrices $\tilde{\mathbf{H}}$ is tractable because the Wishart distribution is the conjugate prior to the normal distribution likelihood.

$$
\begin{aligned}
&p(z_{tf} = k | \tilde{\mathbf{x}}, \boldsymbol{\Theta} \backslash \{z_{tf}, \tilde{\boldsymbol{\pi}}, \tilde{\mathbf{H}}\}) \\
&\propto \int p(z_{tf} = k | \boldsymbol{\pi}_t, \tilde{\mathbf{z}} \backslash z_{tf}) p(\boldsymbol{\pi}_t | \alpha, \beta) d\boldsymbol{\pi}_t \int p(\mathbf{x}_{tf} | \tilde{\mathbf{x}} \backslash \mathbf{x}_{tf}, z_{tf} = k, \boldsymbol{\Theta} \backslash z_{tf}) p(\tilde{\mathbf{H}}) d\tilde{\mathbf{H}} \\
&= (\alpha \beta_k + n_{fk}) \frac{\Gamma\left(\hat{v}_{fw_k}^{\backslash tf} + 1\right)}{\Gamma\left(\hat{v}_{fw_k}^{\backslash tf} - M + 1\right)} \frac{\left|\hat{\mathbf{G}}_{fw_k}^{\backslash tf\ -1}\right|^{\hat{v}_{fw_k}^{\backslash tf}}}{\left|\hat{\mathbf{G}}_{fw_k}^{\backslash tf\ -1} + \lambda_{tf} \mathbf{y}_{tf} \mathbf{y}_{tf}^H\right|^{\hat{v}_{fw_k}^{\backslash tf} + 1}},
\end{aligned}
\tag{4.14}
$$

where $\Gamma(\cdot)$ denotes the gamma function, $|\mathbf{G}|$ is the determinant of matrix $\mathbf{G}$, and $\boldsymbol{\Theta}$ is all the latent variables; $\boldsymbol{\Theta} \backslash \{z_{tf}, \tilde{\mathbf{H}}\}$ means all latent variables except z_{tf} and $\mathbf{H}_{fd}$ (for $\forall f, d$). The calculation of the conditional probability above involves posterior parameters $\hat{v}_{fd}$ and $\hat{\mathbf{G}}_{fd}$ and

dereverberated signal $\mathbf{y}_{tf}$ using the current hypothesis of AR coefficients $\mathbf{A}_{fl}$.

$$\mathbf{y}_{tf} = \mathbf{x}_{tf} - \mathbf{A}_{fv_{tf}}^{H} \bar{\mathbf{x}}_{tf}, \tag{4.15}$$

$$\begin{aligned} \hat{v}_{fd} &= v_{fd} + n_{fd} = v_{fd} + \sum_{t:w_{z_{tf}}=d} 1, \\ \hat{\mathbf{G}}_{fd}^{-1} &= \mathbf{G}_{fd}^{-1} + \sum_{t:w_{z_{tf}}=d} \lambda_{tf} \mathbf{y}_{tf} \mathbf{y}_{tf}^{H}, \end{aligned} \tag{4.16}$$

where n_{fd} is the number of TF points assigned to direction d for frequency bin f. The other counts, n_{fk} and n_{tk}, are defined in a similar way. A posterior parameter with $\cdot^{\backslash tf}$ means that the calculation of the parameter excludes the sample at time t and frequency f.

The time-wise class proportion parameter $\boldsymbol{\pi}_t$ is collapsed out.This marginalization results in the first term $(\alpha\beta_k + n_{fk})$ in Eq. (4.14) (Teh et al. 2006). Similar marginalization is applied to the subsequent conditional probabilities of the discrete latent variables. The purpose of this manipulation is to make the inference tractable and to accelerate the mixing.

To allow for the probability of z_{tf} taking an unassigned class $K+1$ in Eq. (4.14), $\boldsymbol{\beta}$ is given $K+1$ elements, as explained in (Teh et al. 2006, Otsuka et al. 2013). To calculate the probability of $z_{tf} = K+1$, $w_{K+1} = d$ is temporarily drawn with probability $\frac{\kappa/D + c_d}{\kappa + c_\cdot}$. If z_{tf} is set to $K+1$, K is updated as $K \leftarrow K+1$, and the dimensionality of $\boldsymbol{\beta}$ increases by one with $\beta_K \leftarrow b\beta_K$ and $\beta_{K+1} \leftarrow (1-b)\beta_K$, where b is drawn from a beta distribution: $b \sim \mathcal{B}(1, \gamma)$.

The conditional probability of $w_k = d$ is derived as follows.

$$\begin{aligned} &p(w_k = d | \tilde{\mathbf{x}}, \boldsymbol{\Theta} \backslash \{w_k, \boldsymbol{\varphi}, \tilde{\mathbf{H}}\}) \\ &\propto \int p(w_k = d | \boldsymbol{\varphi}, \tilde{\mathbf{w}} \backslash w_k) p(\boldsymbol{\varphi} | \kappa) d\boldsymbol{\varphi} \int p(\tilde{\mathbf{x}}_k | \tilde{\mathbf{x}}_{\backslash k}, w_k = d, \boldsymbol{\Theta} \backslash w_k) p(\tilde{\mathbf{H}}) d\tilde{\mathbf{H}} \\ &= \left(\frac{\kappa}{D} + c_d^{\backslash k} \right) \prod_f \left\{ \prod_{m=0}^{M-1} \frac{\Gamma\left(\hat{v}_{fd}^{\backslash k} + n_{fk} - m\right)}{\Gamma\left(\hat{v}_{fd}^{\backslash k} - m\right)} \frac{\left| \hat{\mathbf{G}}_{fd}^{\backslash k\ -1} \right|^{\hat{v}_{fd}^{\backslash k}}}{\left| \hat{\mathbf{G}}_{fd}^{\backslash k\ -1} + \mathbf{R}_{fk}^{yy} \right|^{\hat{v}_{fd}^{\backslash k} + n_{fk}}} \right\}, \end{aligned} \tag{4.17}$$

where $\cdot^{\backslash k}$ means that the posterior parameter is calculated while excluding the samples assigned to source class k. For example, $n_{fd}^{\backslash k} = \sum_{t:w_{z_{tf}}=d \text{ and } z_{tf} \neq k} 1$. In the second line of Eq. (4.17), the sets of observed TF points assigned to class k and not assigned to class k are denoted by $\tilde{\mathbf{x}}_k$ and $\tilde{\mathbf{x}}_{\backslash k}$, respectively. Matrix $\mathbf{R}_{fk}^{yy}$ represents the correlation of dereverberated signal $\mathbf{y}_{tf}$

assigned to class k for frequency bin f. Therefore,

$$\mathbf{R}_{fk}^{yy} = \sum_{t:z_{tf}=k} \lambda_{tf} \mathbf{y}_{tf} \mathbf{y}_{tf}^{H}.$$

In Eq. (4.17), $\boldsymbol{\varphi}$ is collapsed out using the conjugacy of the Dirichlet distribution to the discrete probability variable w_k, which results in the term $(\frac{\kappa}{D} + c_d^{\backslash k})$, where c_d is the number of classes assigned to direction d. $c_d^{\backslash k}$ is the number of classes for direction d without class k.

Matrix $\mathbf{H}_{fd}$, which represents the subspace corresponding to direction d, is updated using the distribution

$$p(\mathbf{H}_{fd}|\tilde{\mathbf{x}}, \boldsymbol{\Theta} \backslash \mathbf{H}_{fd}) = \mathcal{W}_{\mathbb{C}}(\mathbf{H}_{fd}|\hat{v}_{fd}, \hat{\mathbf{G}}_{fd}), \tag{4.18}$$

where $\mathcal{W}_{\mathbb{C}}(\cdot)$ means the probability density function of the complex Wishart distribution. The posterior parameters $\hat{v}_{fd}$ and $\hat{\mathbf{G}}_{fd}$ are given by Eq. (4.16).

Similarly to the update of z_{tf}, the class of the AR coefficient at each TF point v_{tf} is updated to $1, \ldots, L_f + 1$, where L_f is the number of instantiated AR coefficients. The following conditional probabilities are calculated for $l = 1, \ldots, L_f + 1$, and v_{tf} is chosen in proportion to the probability.

$$\begin{aligned} &p(v_{tf} = l|\tilde{\mathbf{x}}, \boldsymbol{\Theta} \backslash v_{tf}) \\ &\quad \propto (\zeta_f + r_{fl}^{\backslash tf}) \exp\left\{-(\mathbf{x}_{tf} - \mathbf{A}_{fl}^{H} \overline{\mathbf{x}}_{tf})^{H} \mathbf{H}_{fw_{z_{tf}}} (\mathbf{x}_{tf} - \mathbf{A}_{fl}^{H} \overline{\mathbf{x}}_{tf})\right\}, \end{aligned} \tag{4.19}$$

where r_{fl} is the number of TF points at which the reverberant component is formed by AR coefficient $\mathbf{A}_{fl}$ for frequency bin f; $r_{fl} = \sum_{t:v_{tf}=l} 1$. The first term is derived from the Chinese restaurant process (Aldous 1985), where $\boldsymbol{\varrho}$ is marginalized out, and the second term results from the likelihood given by Eq. (4.6). To allow for unassigned class index v_{f,L_f+1}, $\mathbf{A}_{f,L_f+1}$ is drawn from the prior of the AR coefficients given by Eq. (4.12).

The conditional probability distribution of the AR coefficients is derived using the vectorized notations in Eq. (4.13). A similar derivation was presented in (Yoshioka et al. 2011).

$$p(\mathrm{vec}(\mathbf{A}_{fl})|\tilde{\mathbf{x}}, \boldsymbol{\Theta} \backslash \mathbf{A}_{fl}) = \mathcal{N}_{\mathbb{C}}(\mathrm{vec}(\mathbf{A}_{fl})|\hat{\mathbf{m}}_{fl}, \hat{\Sigma}_{fl}), \tag{4.20}$$

where the posterior parameters $\hat{\mathbf{m}}_{fl}$ and $\hat{\Sigma}_{fl}$ are given as

$$\hat{\Sigma}_{fl}^{-1} = \sum_{t:v_{tf}=l} \mathbf{H}^{*}_{fw_{z_{tf}}} \otimes \lambda_{tf} \bar{\mathbf{x}}_{tf} \bar{\mathbf{x}}_{tf}^{H},$$

$$\hat{\mathbf{m}}_{fl} = \hat{\Sigma}_{fl}^{-1} \mathrm{vec}(\mathbf{Q}_{fl}), \qquad \mathbf{Q}_{fl} = \sum_{t:v_{tf}=l} \lambda_{tf} \bar{\mathbf{x}}_{tf} \mathbf{x}_{tf}^{H} \mathbf{H}_{fw_{z_{tf}}},$$

where $\mathbf{H}^{*}$ denotes the complex conjugate of matrix $\mathbf{H}$.

The probability of λ_{tf} is a gamma distribution:

$$p(\lambda_{tf}|\tilde{\mathbf{x}}, \boldsymbol{\Theta} \backslash \lambda_{tf}) = \mathcal{G}(\lambda_{tf}|\hat{a}_{tf}, \hat{b}_{tf}), \tag{4.21}$$

$$\hat{a}_{tf} = a_{tf} + M, \qquad \hat{b}_{tf} = b_{tf} + \mathbf{y}_{tf}^{H} \mathbf{H}_{fw_{z_{tf}}} \mathbf{y}_{tf},$$

where $\mathcal{G}(\cdot)$ is the probability distribution of the gamma distribution.

The hyperparameters related to the discrete latent variables are updated so that the clustering algorithm can reflect the variability of the observed data. The parameters related to HDP, α, β, and γ, are updated as described in (Teh et al. 2006). The update of the other parameters such as κ and ζ_f (for $\forall f$) follows the steps explained in (Escobar and West 1995). These parameters are updated using auxiliary variables.

We start the MCMC inference by preparing a single AR coefficient for each frequency bin. That is, we initially try to find the reverberant component with a single AR coefficient. If the single-AR model is insufficient to analyze the observation, e.g., due to an underdetermined mixture, additional AR coefficients are instantiated to fit the observation.

4.3.3 Initialization of inference

The initialization procedure comprises five steps:

1. Use one AR class for each frequency bin: $v_{tf} = 1, \forall t, f$.
2. Initialize AR coefficients $\mathbf{A}_{f1}$ using Eq. (4.22).
3. Calculate dereverberated signal $\mathbf{y}_{tf}$ using Eq. (4.15) and initial AR coefficients.
4. Initialize z_{tf} and w_k using dereverberated signal $\mathbf{y}_{tf}$ and Eq. (4.23).
5. Initialize $\mathbf{H}_{fd}$ and λ_{tf} using their posterior distributions Eq. (4.18) and Eq. (4.21).

The initial AR coefficients are calculated using

$$\mathbf{A}_{f1} = \left(\mathbf{K}_{f1} + \sum_t \lambda_{tf} \overline{\mathbf{x}}_{tf} \overline{\mathbf{x}}_{tf}^H \right)^{-1} \left(\sum_t \lambda_{tf} \overline{\mathbf{x}}_{tf} \mathbf{x}_{tf} \right). \tag{4.22}$$

This value is equivalent to the maximum a posteriori (MAP) and minimum mean square erro (MMSE) estimate of the subsequent generative process. Since this dereverberation is an approximation with low computational cost, we use it for the initialization.

$$\begin{aligned} \mathbf{x}_{tf} | \mathbf{A}_{f1}, \overline{\mathbf{x}}_{tf} &\sim \mathcal{G}_{\mathbb{C}}(\mathbf{x}_{tf} | \mathbf{A}_{f1}^H, \lambda_{tf}^{-1} \mathbf{I}_M), \\ \mathbf{A}_{f1} | \mathbf{K}_{f1} &\sim \mathcal{MN}_{\mathbb{C}}(\mathbf{A}_{f1} | \mathbf{0}, \mathbf{K}_{f1}, \mathbf{I}_M), \end{aligned}$$

where $\mathbf{K}_{f1} = \mathbf{I}_{MR}$ is the same parameter as in Eq. (4.12). Here, we use $\lambda_{tf} = \frac{1}{\mathbf{x}_{tf}^H \mathbf{x}_{tf}}$.

The initialization of z_{tf} and w_k is carried out as follows. The inference starts with a certain number of classes K. In our implementation, the initial numberis set to 12. First, w_k is initialized with a uniform distribution for which the support does not the other classes. Then, each z_{tf} is drawn using the sampled w_k and the hyperparameter of the Wishart distribution, $\mathbf{G}_{fd}$, generated from the steering vectors:

$$\begin{aligned} p(w_k = d) &= \mathcal{U}\left(\left\{ d \middle| \frac{k-1}{K} D \leq d < \frac{k}{K} D \right\} \right), \\ p(z_{tf} = k) &\propto \exp\left(-\mathbf{y}_{tf}^H \mathbf{G}_{f w_k} \mathbf{y}_{tf} \right), \end{aligned} \tag{4.23}$$

where $\mathcal{U}(A)$ is a probability density function of uniform distribution on set A.

4.3.4 Extraction of source signals

The PCGS described in the previous section is iterated to draw I samples of the latent variables indexed by i: $\{\tilde{\mathbf{z}}^{(i)}, \tilde{\mathbf{w}}^{(i)}, \tilde{\mathbf{v}}^{(i)}, \tilde{\mathbf{A}}^{(i)}\}_{i=1}^I$. Since we have no knowledge as to which directions the source signals come from or the number of sources, we first calculate the weight of the TF mask for each direction to determine which directions have dominant sound sources. The weight for each direction is given by

$$P_d = \frac{1}{I} \sum_{i=1}^{I} \sum_{tf} \delta\left(w_{z_{tf}^{(i)}}^{(i)}, d \right).$$

Source number estimation using this weight was previously presented in Chapter 3.

The source signal from direction d in the observed mixture is extracted in descending order of weights P_d:

$$\hat{\mathbf{x}}_{tf}^{d} = \frac{1}{I}\sum_{i}\delta(w^{(i)}_{z^{(i)}_{tf}}, d)\left(\mathbf{x}_{tf} - \mathbf{A}^{(i)\,H}_{f v^{(i)}_{tf}}\overline{\mathbf{x}}_{tf}\right), \tag{4.24}$$

where $\delta(m,n)$ is the Kronecker delta; i.e., $\delta(m,n) = 1$ if $m = n$, and 0 otherwise. The delta factor corresponds to the TF masking for the separation, and the subtraction using the AR coefficients corresponds to the dereverberation process.

4.4 Experimental Results

The performance of source separation and dereverberation with our method is experimentally evaluated. The experiments consist of two parts. The first experiment considers various numbers of microphones and sources. The second experiment investigates the performance sensitivity to the choice of AR process order R.

We compare the following four methods in our experiment:

DPAR: Our proposed method with a DP mixture of AR processes for dereverberation.

Single: A simplified version of DPAR–only a single AR process was used to model the reverberation. That is, the AR class indicator was fixed; i.e., $v_{tf} = 1$ (for $\forall t, f$) and the update of v_{tf} in Eq. (4.19) was skipped during the inference. AR coefficient $\mathbf{A}_{f1}$ was updated in accordance with Eq. (4.20) for the dereverberation. Inference was jointly carried out with TF masking estimation and AR coefficient estimation.

Cascade: The observed mixture was first dereverberated using the AR coefficients in Eq. (4.22). Then, HDP-based TF masking method (Otsuka et al. 2013) was used to separate the sound sources.

Linear: An existing method combining dereverberation using a single AR process and separation using ICA (Yoshioka et al. 2011). Due to the limitation of ICA, this method is applicable only when $M \geq N$[1], and assumes the source number, N, to be available. This method dereverberates the observed signal using M channels, while N channels are selected to estimate the separation matrices, as explained in (Yoshioka et al. 2011).

[1]While an "exact" dereverberation requires $M > N$, we can dereverberate the input signal in practice even when $M = N$, as mentioned in the last part of Section 4.2.2

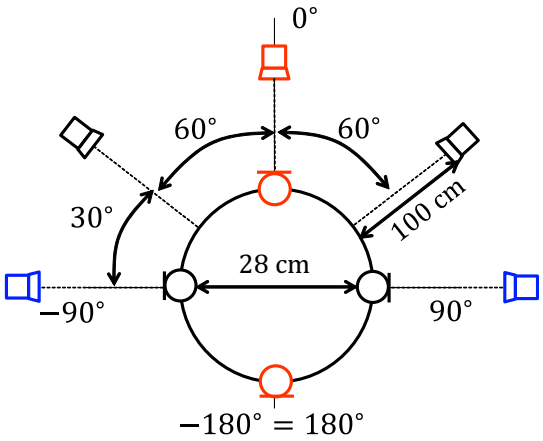


Figure 4.3: Configuration of microphone array and sound sources. Microphones were positioned at 90° intervals. When $M = 2$, microphones at ±90° (depicted in black) were used whereas all microphones were used when $M = 4$. All sound sources were 100 (cm) from microphone array. Center source in red was at 0°. Black sources were at ±60°, and blue sources were at ±90°. When $N = 2$, black sources were presented, when $N = 3$, black and red sources were presented, and when $N = 5$, all sources were presented.

Through the comparison of DPAR and Single, we investigate the effectiveness of an infinite mixture of AR processes. We evaluate the effectiveness of joint optimization of AR coefficients and mixing coefficients by comparing the results of Cascade and those of Single or DPAR. We also show the performance of Linear as it represents the state-of-the-art.

While Single uses a conventional dereverberation model based on a single AR process, this method is novel in that it combines AR process-based dereverberation and TF masking-based separation. Unlike existing joint models that incorporate a linear separation method (Buchner and Kellermann 2010, Yoshioka et al. 2011, Takeda et al. 2012) requiring $N \leq M$, Single manages underdetermined mixtures. Single is similar to the joint separation and dereverberation method proposed by Togami et al. (Togami et al. 2013) in two ways. First, their method uses a single AR process using Gaussian AR coefficients. Second, they model the direct component as a time-varying covariance matrix (Duong et al. 2010) as in Eq. (4.6). However, in order to perform the source separation, they apply a Wiener filter to the observed mixtures instead of using TF masks, where the Wiener filter estimated based on the source covariance matrices that correspond to the subspace of each source. The Wiener filter-based separation requires the source number information and thus limits the applicability of the method.

4.4.1 Experimental setups

Figure 4.3 illustrates the configuration of the microphone array and sound sources. We used two or four microphones ($M = 2, 4$) and two, three, or five sources ($N = 2, 3, 5$), including an underdetermined condition: $(M, N) = (2, 3), (2, 5), (4, 5)$. The steering vectors, $\tilde{\mathbf{q}}$, were calculated from impulse responses measured in an anechoic room such that $D = 72$ with 5° resolution. The steering vectors were generated from a Fourier transform of the first 512 points of the anechoic impulse responses. To avoid front-back ambiguity, the steering vectors from $-90°$ to $90°$ were used when $M = 2$.

The mixture signals were synthesized by convolving impulse responses measured in indoor environments with recorded speech utterances excerpted from JNAS corpus containing phonetically balanced Japanese utterances. We used three environments with different reverberation times (RTs): 150, 400, and 600 (ms). Reference signals for separation and dereverberation were generated by convoluting impulse responses truncated at the first 512 points to eliminate the reverberation component from the reference. For each condition, 20 mixtures were tested. The average length of these mixtures was around 5 (sec). The sampling rate was 16,000 (Hz), with a 512 (pt) hanning window and a 256 (pt) shift for STFT. To avoid the suppression of direct components by the AR-based dereverberation, parameter δ in Eq. (4.5) was set to 1.

We used two metrics to evaluate the separation and dereverberation performance. The signal-to-distortion ratio (SDR) (Vincent et al. 2006) was used to measure both separation and dereverberation quality. The direct-to-reverberation ratio (DRR) was used to measure the dereverberation performance. Two types of reference signals, the direct component and reverberant component of each source signal, were used to calculate to these ratios. The SDR and DRR (in dB) for the nth signal are given in the time domain with time index τ:

$$\mathrm{SDR}_n = 10\log_{10}\frac{\sum_\tau s_n^{\mathrm{dir}}(\tau)^2}{\sum_\tau(\hat{x}_n(\tau) - s_n^{\mathrm{dir}}(\tau))^2},$$
$$\mathrm{DRR}_n = 10\log_{10}\frac{\sum_\tau s_n^{\mathrm{dir}}(\tau)^2}{\sum_\tau \hat{r}_n(\tau)^2},$$

where $\hat{x}_n$, s_n^{dir}, and $\hat{r}_n$ are the separated and dereverberated output corresponding to the nth source, the reference direct component for source n, and the reverberant components of the

sources in the mixture remaining in output $\hat{x}_n$, respectively. The N separated results $\{\hat{x}_n\}_{n=1}^N$ and the direct components for the N sources $\{s_n\}_{n=1}^N$ were matched so that the interference of the other sources was minimized (Yoshioka et al. 2011, Vincent et al. 2006). When the output signal had multiple channels, we calculated these ratios using the first channel. For the evaluation of Linear, a multichannel sound image of each sound source was extracted, and the first channel was evaluated using the criteria above. The separation and dereverberation qualities are considered to be better when SDR and DRR are larger. Reverberant component $\hat{r}_n$ in output $\hat{x}_n$ was obtained using the sum of the synthesized reverberant components:

$$
\begin{aligned}
s^{\mathrm{rev}}(\tau) &= \sum_{n'=1}^{N} s_{n'}^{\mathrm{rev}}(\tau), \\
\hat{r}_n(\tau) &= G s^{\mathrm{rev}}(\tau - \Delta) \text{ such that } \min_{G,\Delta} \sum_{\tau} (\hat{x}_n - G s^{\mathrm{rev}}(\tau - \Delta))^2,
\end{aligned}
$$

where s_n^{rev} is the synthesized reverberant component for source n. We extracted the reverberant components for all sources in each separated signal $\hat{x}_n$ to evaluate the DRR because the separated signal for source n may contain the reverberant component for other sources $n' \neq n$. The reverberant components were extracted up to a linear amplification of gain G and a time shift Δ of 16 (ms).

The direct component s_n^{dir} and reverberant component s_n^{rev} were synthesized as follows. Let $s_n(\tau)$ be the nth source signal in the time domain and $b(\tau)$ be the impulse response used to generate the simulated signal. The impulse response was further decomposed into two parts: $b(\tau) = b^{\mathrm{dir}}(\tau) + b^{\mathrm{rev}}(\tau)$, where $b^{\mathrm{dir}}(\tau)$ is the truncated impulse response corresponding to the direct component. After measuring $b(\tau)$ in each environment, we obtained $b^{\mathrm{dir}}(\tau)$ by truncating $b(\tau)$ at 32 (ms), which corresponds to the length of the window for STFT. The reverberant impulse response $b^{\mathrm{rev}}(\tau)$ was derived by padding zeros to the first 32 (ms) of $b(\tau)$. Then, s_n^{dir} and s_n^{rev} were generated using

$$
\begin{aligned}
s_n^{\mathrm{dir}}(\tau) &= b^{\mathrm{dir}}(\tau) * s_n(\tau), \\
s_n^{\mathrm{rev}}(\tau) &= b^{\mathrm{rev}}(\tau) * s_n(\tau).
\end{aligned}
$$

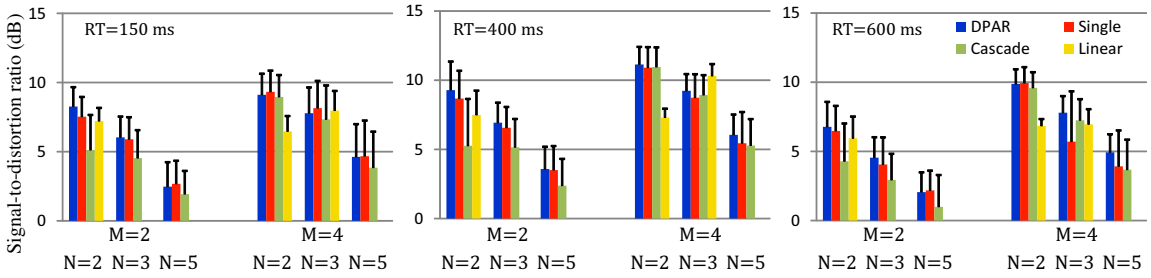


Figure 4.4: SDR of each method for various numbers of microphones ($M = 2, 4$) and sources ($N = 2, 3, 5$). Bars represent means, and segments are standard deviations.

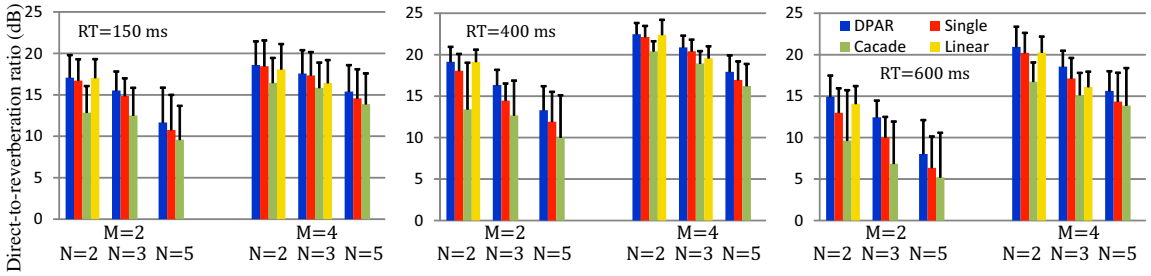


Figure 4.5: DRR of each method for various numbers of microphones ($M = 2, 4$) and sources ($N = 2, 3, 5$). Bars represent means, and segments are standard deviations.

4.4.2 Result 1: separation and dereverberation performance

We first present the separation and dereverberation performances when AR order R was 10. The SDRs and DRRs for the three environments are shown in Figures 4.4 and 4.5. Note that the ones of Linear method, shown with yellow bars, are omitted for $M < N$ because Linear is not applicable to underdetermined setups.

The TF masking-based methods (DPAR, Single, and Cascade) produces larger ratios when the number of microphones is larger and the number of sources is smaller. Unlike the other three methods, Linear has a higher SDR with $M = 4$ and $N = 3$ than with $M = 4$ and $N = 2$. This is because the linear separation process changed the amplitude of the direct components of the constituent sources, while the evaluation criterion is affected by the amplification effect[2]. Comparison of DPAR and Linear shows that our method performs better than a state-of-the-art method, except when $M = 4$, $N = 3$, and RT $= 400$ (ms). Note that the TF masking-

[2] The interference of the direct component of the other sources was smaller when $N = 2$ than when $N = 3$.

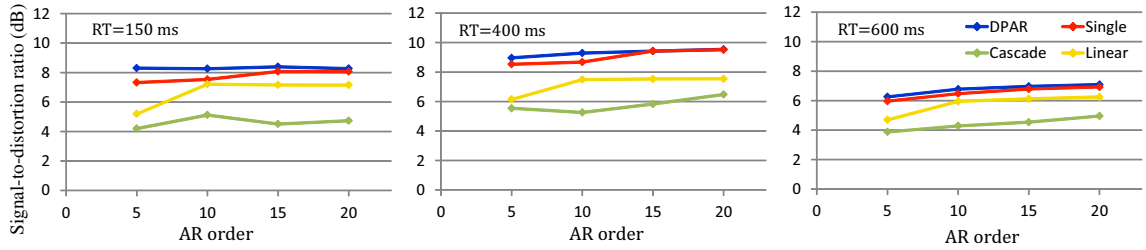


Figure 4.6: SDR of each method for various AR orders ($R = 5, 10, 15, 20$) for $M = 2$ and $N = 2$.

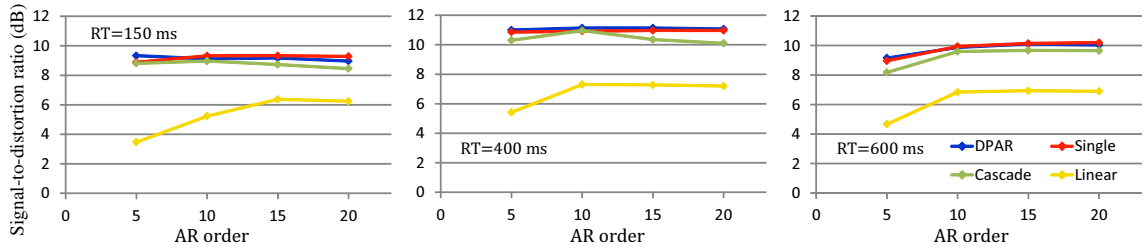


Figure 4.7: SDR of each method for various AR orders for $M = 4$ and $N = 2$.

based methods dispense with the source number information while Linear makes use of the information during the separation process.

The joint optimization approaches (DPAR and Single) have higher DRRs than Cascade. Similar results are obtained for SDRs, although the differences in the scores are less. This indicates that the joint optimization improves the estimation accuracy of AR coefficients for dereverberation, which results in improved separation performance.

The DRRs for DPAR is slightly but consistently higher than those for Single because DPAR has a more flexible AR model of reverberation. The flexibility of DPAR contributes to higher SDRs especially when $M = 2$, where the AR model has a limited fitting capability due to a small size of AR coefficient matrices ($\mathbf{A}_{fl} \in \mathbb{C}^{MR \times M}$), and $N = 5$, where the mixture process is underdetermined.

4.4.3 Result 2: performance for various AR orders

The SDRs for the three environments for $R = 5, 10, 15$, and 20 and $N = 2$ are shown in Figures 4.6 and 4.7 for $M = 2$ and 4, respectively. The DRRs for $M = 2$ and 4 are shown in Figures 4.8 and 4.9, respectively.

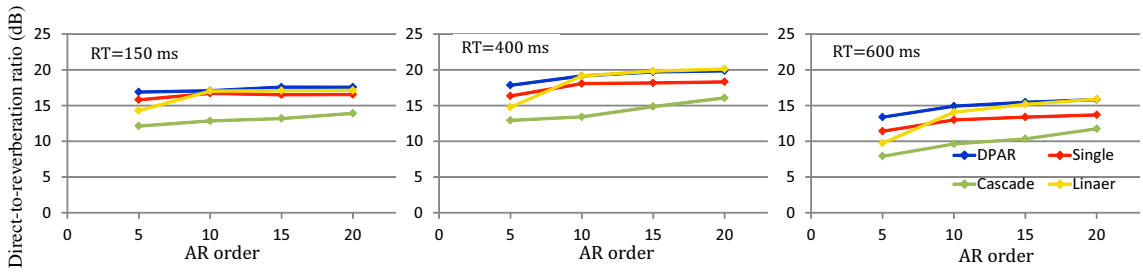


Figure 4.8: DRR of each method for various AR orders for $M = 2$ and $N = 2$.

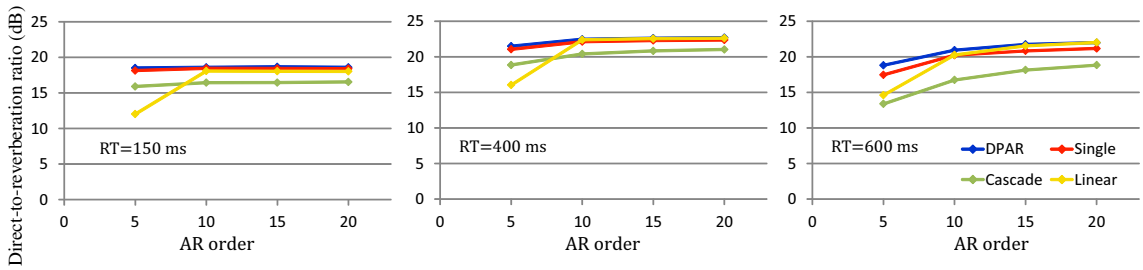


Figure 4.9: DRR of each method for various AR orders for $M = 4$ and $N = 2$.

Overall, Linear is more sensitive to the AR order value than the other methods. In particular, while the DRRs of Linear method with $R = 5$ is low, the score is substantially improved when R is increased to 10. The Bayesian approaches is less sensitive to the AR order value in the SDRs and DRRs. This is an advantage of the Bayesian approaches: the inference results are insensitive to the model complexity with regard to the AR order value.

In Figures 4.8 and 4.9, we can see that the impact of the AR order on the DRR varies with the reverberation time: when the amount of reverberation is the smallest, RT = 150 (ms), the DRRs almost saturate when $R = 5$ while the DRRs continue to increase with the AR order when RT = 600 (ms). A similar tendency is evident for the SDR curves (Figures 4.6 and 4.7). This indicates that the AR order should be chosen carefully so as to maximize the separation and dereverberation performances.

The DRRs for DPAR are larger than those for Single due to the flexible AR processes, especially when the size of AR coefficients is limited ($M = 2$). We can also observe a slight difference in the SDRs. While Cascade yields degraded DRRs compared with DPAR and Single, the SDRs for Cascade are as large as those of DPAR and Single when $M = 4$. This is probably because the degradation in dereverberation performance can be compensated for by

Table 4.2: Computational complexity of each method. Time complexity is per iteration. Total computation time is product of time complexity and No. of iterations. K and L denote instantiated number of TF masks and average number of instantiated AR classes over frequency bins, respectively.

Method	Space complexity	Time complexity	# of iterations
DPAR	$O(LFM^2R+M^2FD+TFK)$	$O(TFMRL+FLR^3M^6$ $+TFKM^3+FKDM^3)$	30
Single	$O(FM^2R+M^2FD+TFK)$	$O(FR^3M^6+TFKM^3+FKDM^3)$	30
Cascade derev.	$O(FM^2R)$	$O(TFM^2R^2+FM^3R^3)$	1
Cascade separation	$O(M^2FD+TFK)$	$O(TFKM^3+FKDM^3)$	50
Linear	$O(FMNR+FN^2+TFN)$	$O(TFMNR+FM^3N^3R^3+TFN^2)$	3

the separation process if the number of microphones is not too small.

4.4.4 Discussion

The experimental results presented above demonstrate that DPAR is effective even when there is uncertainty about the degree of reverberation and the number of sources, although the performance still depends on environmental conditions to some extent. DPAR method outperforms the state-of-the-art Linear method especially when the number of microphones is small ($M = 2$). Comparison between the joint optimization approaches (DPAR and Single) with the Cascade method indicates that the joint optimization methods are better than the Cascade method. This is because the iterative update of the AR coefficients contributes to better dereverberation and thus better separation.

DPAR has better dereverberation performance than Single due to the flexibility of its infinite AR model based on Bayesian nonparametrics. This leads to better separation in terms of SDR when $M = 2$. On the other hand, the SDRs of DPAR and Single are similar when $M = 4$ because the infinite mixture of AR coefficients is more effective when the dimensionality of a single AR coefficient is restricted. As shown below, DPAR requires more computational resources than Single. Therefore, in practice, the use of DPAR is more attractive when the number of microphones is limited, in which case the mixing process is more likely to be underdetermined.

The computational complexity of each method is summarized in Table 4.2. The complex-

ity of the separation (and localization) process, presented as Cascade separation in Table 4.2, is the same as our previous method (Otsuka et al. 2013). The first and second terms of time complexity for DPAR, $O(TFMRL)$ and $O(FLR^3M^6)$ come from the sampling of v_{tf} in Eq. (4.19) and calculation of parameters $\hat{\mathbf{m}}_{fl}$ and $\hat{\Sigma}_{fl}$ in Eq. (4.20), respectively. One of the limitations of the DPAR method is its high computational cost due to many matrix operations. We derived an inference algorithm based on the MCMC method, which is generally computationally expensive. Instead, we may develop a variational inference framework for our generative model presented in Section 4.3.1. Since a typical variational inference procedure is similar to expectation maximization (EM) algorithms, we can expect acceleration of parameter inference due to parallelizing the computations, e.g., parallelization with regard to the frequency bins. We chose MCMC-based inference because MCMC avoids local optima while variational methods susceptible to a local optima issue. Therefore, a MCMC method is a reasonable way to avoid the optimization issue. Separation and dereverberation of moving sound sources are additional limitations of current microphone array processing because most microphone array-based algorithms are based on the assumption of the stationary observation process in Eq. (3.1). Our method may be able to solve this dynamic source problem due to its Bayesian nonparametric formulation. The switching of AR coefficients can be estimated using the idea expressed in Eq. (4.4) for dereverberation. The infinite TF masking separation may be able to handle moving sound sources by generating several TF masks for directions that track the paths of the sources. In the current framework, a moving source is split into multiple TF masks for different directions. Merging the TF masks for the same sound source to retrieve the moving source signal will be a problem. Such a source tracking mechanism is a future direction for addressing the moving sound source problem.

4.5 Summary

We have presented a unified sound source separation and dereverberation model that is applicable to mixture signals containing an unknown number of constituent sources signals, even for underdetermined conditions, and that is robust against reverberation. We modeled the reverberation as switching AR processes based on the sparsity of the source dominance in the TF domain. Our unified model uses Bayesian nonparametrics to achieve both the switching

AR processes for the dereverberation and TF masking-based separation for an arbitrary number of sources. The experimental results reveal the efficacy of our method for underdetermined setups as well as superiority of the dereverberation and separation performances to those of a state-of-the-art method, which is inapplicable to underdetermined conditions. This Bayesian nonparametrics-based model will contribute to machine listening functions as a general framework that works in our daily environments with auditory uncertainties.

Chapter 5

Multichannel Dereverberation using Infinite-Order Autoregressive Model

Multichannel dereverberation methods on the basis of autoregressive (AR) processes need to select an appropriate AR order value in accordance with the amount of reverberation (reverberation time) of the environment. Reverberation time is an auditory uncertainty regarding the reverberation because reverberation time is dependent on various factors: the width and height of the room, materials of the walls and floor, locations of sound sources etc. To robustly achieve the dereverberation in various environments, the selection of AR order value is discussed in this chapter.

5.1 Introduction

Many phenomena in our life are regarded as a time series sequence where the sequence of observations are temporally correlated. The AR process is a widely used model to analyze such sequential data (Box et al. 2008). Applications include an analysis of fMRI measurements (Harrison et al. 2003), enhancement of speech signals (T. Yoshioka and Miyoshi 2009), econometrics (Dickey and Fuller 1981), and modeling of the dance of honey bees (Fox et al. 2011). Recent incorporation of Bayesian nonparametric models with AR processes enables us to further handle complex time series data by enriching their fitting capability (Fox et al. 2011, Lucca et al. 2013).

For reliable comprehension and analysis of the temporal structure, an appropriate choice of the order of AR processes has become a critical issue. A motivative example is the removal

of reverberation from audio signals, also known as dereverberation (T. Yoshioka and Miyoshi 2009, Nakatani et al. 2008b), to improve audio intelligibility. Reverberations are caused by the reflection of sound waves on the walls in indoor environments. AR processes can be used to model reverberation because sound reflections can be modeled as the propagation of past observations. Here, the AR order is dependent on how reverberant the room is. The choice of an appropriate order is therefore critical to achieve a good dereverberation performance. If we have no knowledge about the environment, a sensible way is to estimate the order from observations.

A number of order selection methods have been developed in the literature. These methods can be categorized into two classes: model selection approaches and Bayesian inference of the order. The model selection approaches (Hannan 1980, Ing et al. 2012) use various criteria such as the Akaike information criterion (AIC) (Akaike 1974) or the Bayesian information criterion (BIC) (Schwarz 1978). An order selection using a variational lower bound of the marginal likelihood (Penny and Roberts 2002, Roberts and Penny 2002) may also be categorized here. These selection approaches suffer from two limitations, both of which undermine the flexibility of AR models. First, these order selection approaches require the determination of a particular order from a finite set of candidate orders. Second, these methods are incompatible with other Bayesian models, such as infinite mixture (Lucca et al. 2013), that can be applied to a wider range of sequential data. This is because it is difficult to carry out order selection and mixture inference at once; the model selection methods require that all the samples share the same AR structure to compute the selection criteria whereas Bayesian mixture models assume each sample stochastically belongs to different AR processes.

Alternatively, we can infer the order from the given data in a Bayesian framework by introducing a prior on the order (Godsill 2001, Vermaak et al. 2004). The posterior belief of the orders is inferred for given observations to investigate which order fits the observed data. In contrast to the selection approaches above, this is superior because the flexibility of the model can easily be enhanced by incorporating other Bayesian models. The limitations of the existing approaches (Godsill 2001, Vermaak et al. 2004) are twofold; (1) the possible order is limited to a finite range, which ristricts the modeling capability, and (2) the inference becomes inefficient. Since the analytic derivation of the posterior of the order and AR coefficients is of-

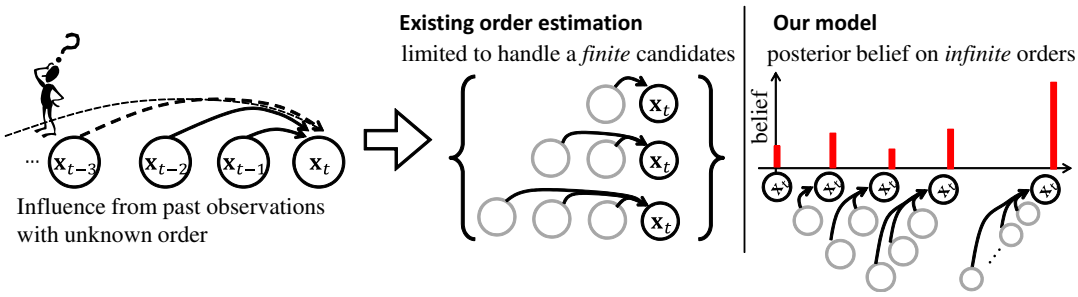


Figure 5.1: Order uncertainty problem and our approach. We are uncertain how influential the past samples are. Existing methods cope with only a finite set of candidate orders. Our model obtains the posterior belief on infinitely many orders to find an appropriate order for the data.

ten intractable, Markov chain Monte Carlo (MCMC) can be used to draw samples of the order and AR coefficients from their posterior distribution. Because the number of AR coefficients varies with different orders, reversible jump (RJ) MCMC (Green 1995) is employed to compensate for the dimensionality difference. This sampler can suffer from slow mixing because a change in the dimensionality of latent parameters is carried out through the Metropolis-Hastings algorithm, where the parameters are stochastically updated or unchanged.

This chapter presents Bayesian AR processes with infinite orders, as illustrated in Fig. 5.1. The order prior follows the infinite Markov model used to construct ∞-gram language models (Mochihashi and Sumita 2008). We also present an efficient MCMC inference that uses slice sampling (Neal 2003) to update the order whereas the AR coefficients are collapsed out to accelerate the mixing, or partially updated using a partitioned matrix notation to avoid RJMCMC. Another advantage of our Bayesian model is the capability of enhancing the modeling flexibility through combinations with other Bayesian models (Fox et al. 2011, Lucca et al. 2013).

5.2 Preliminaries: Bayesian AR process with a fixed order

This section outlines the Bayesian formulation of an AR process with a fixed order. Here, we assume the observation is multivariate in general, which is also known as a vector or multivariate autoregressive process, and introduce conjugate priors that facilitate the computation. Similar models have been presented in the literature (Penny and Roberts 2002, Godsill 2001).

Suppose a time series observation $\{\mathbf{x}_t\}_{t=1}^T$ is generated from the AR process of order r. Then, the likelihood of D-dimensional column vector $\mathbf{x}_t \in \mathbb{R}^D$ conditioned on the past r observation denoted by $\bar{\mathbf{x}}_t^{(r)} = [\mathbf{x}_{t-1}{}^\top, ..., \mathbf{x}_{t-r}{}^\top]^\top \in \mathbb{R}^{Dr}$ is given by

$$\begin{aligned} \mathbf{x}_t | \bar{\mathbf{x}}_t^{(r)}, \mathbf{A}^{(r)}, \mathbf{L} &\sim \mathcal{N}(\mathbf{A}^{(r)\top} \bar{\mathbf{x}}_t^{(r)}, \mathbf{L}), \\ \mathbf{A}^{(r)} | \mathbf{M}^{(r)}, \mathbf{K}^{(r)}, \mathbf{L} &\sim \mathcal{MN}(\mathbf{M}^{(r)}, \mathbf{K}^{(r)}, \mathbf{L}), \quad \mathbf{L} | \nu, \Lambda \sim \mathcal{W}(\nu, \Lambda), \end{aligned} \tag{5.1}$$

where $\mathbf{A}^{(r)} \in \mathbb{R}^{Dr \times D}$ is the AR coefficient and $\mathbf{L}$ is the precision matrix. $\mathbf{C}^\top$ is the tranpose of matrix $\mathbf{C}$. The likelihood is the normal distribution denoted by $\mathcal{N}(\boldsymbol{\mu}, \Lambda)$ with the mean $\boldsymbol{\mu}$ and precision Λ. We can equivalently represent the time series as $\mathbf{x}_t = \mathbf{A}^{(r)\top} \bar{\mathbf{x}}_t + \mathbf{e}_t$, where $\mathbf{e}_t$ is independent and identically distributed (i.i.d.) from $\mathcal{N}(\mathbf{0}, \mathbf{L})$. This component is called the excitation, meaning the energy source of the observation. The prior distribution for $\mathbf{A}^{(r)}$ is the matrix normal distribution $\mathcal{MN}(\mathbf{M}^{(r)}, \mathbf{K}^{(r)}, \mathbf{L})$ with the mean $\mathbf{M}^{(r)} \in \mathbb{R}^{Dr \times D}$, column-wise precision $\mathbf{K}^{(r)} \in \mathbb{R}^{Dr \times Dr}$, and row-wise precision $\mathbf{L} \in \mathbb{R}^{D \times D}$. Finally, the precision matrix $\mathbf{L}$ follows Wishart distribution $\mathcal{W}(\nu, \Lambda)$ with the degree of freedom ν and scale matrix Λ.

The attractive property of these prior distributions is their conjugacy; we can analytically derive the parameters of posterior distributions as well as the marginal likelihood through matrix operations. The posterior distributions of $\mathbf{A}^{(r)}$ and $\mathbf{L}$ given the observation $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^T$ become $\mathcal{MN}(\hat{\mathbf{M}}^{(r)}, \hat{\mathbf{K}}^{(r)}, \mathbf{L})$ and $\mathcal{W}(\hat{\nu}, \hat{\Lambda})$ with the following parameters:

$$\begin{aligned} \hat{\mathbf{K}}^{(r)} &= \mathbf{K}^{(r)} + \mathbf{R}_{\bar{x}\bar{x}}^{(r)}, & \hat{\mathbf{M}}^{(r)} &= \hat{\mathbf{K}}^{(r)-1} \left(\mathbf{K}^{(r)} \mathbf{M}^{(r)} + \mathbf{R}_{\bar{x}x}^{(r)} \right), \\ \hat{\nu} &= \nu + T, & \hat{\Lambda}^{-1} &= \Lambda^{-1} + \mathbf{R}_{xx} + \mathbf{M}^{(r)\top} \mathbf{K}^{(r)} \mathbf{M}^{(r)} - \hat{\mathbf{M}}^{(r)\top} \hat{\mathbf{K}}^{(r)} \hat{\mathbf{M}}^{(r)}, \end{aligned} \tag{5.2}$$

where the matrices $\mathbf{R}_{\{\bar{x}\bar{x}, \bar{x}x, xx\}}$ are the correlation of observations $\mathbf{R}_{\bar{x}\bar{x}}^{(r)} = \sum_t \bar{\mathbf{x}}_t^{(r)} \bar{\mathbf{x}}_t^{(r)\top}$, $\mathbf{R}_{\bar{x}x}^{(r)} = \sum_t \bar{\mathbf{x}}_t^{(r)} \mathbf{x}_t{}^\top$, and $\mathbf{R}_{xx} = \sum_t \mathbf{x}_t \mathbf{x}_t{}^\top$. The marginal likelihood is given as

$$\iint p(\mathbf{X} | \mathbf{A}^{(r)}, \mathbf{L}) p(\mathbf{A}^{(r)} | \mathbf{L}) p(\mathbf{L}) d\mathbf{A}^{(r)} d\mathbf{L} = \left(\frac{1}{\pi} \right)^{\frac{DT}{2}} \prod_{d=0}^{D} \frac{\Gamma(\frac{\nu+T-d}{2})}{\Gamma(\frac{\nu-d}{2})} \left(\frac{|\mathbf{K}^{(r)}|}{|\hat{\mathbf{K}}^{(r)}|} \right)^{\frac{D}{2}} \frac{|\Lambda^{-1}|^{\frac{\nu}{2}}}{|\hat{\Lambda}^{-1}|^{\frac{\nu+T}{2}}}, \tag{5.3}$$

where $\Gamma(\cdot)$ is the gamma function and $|\mathbf{C}|$ is the determinant of matrix $\mathbf{C}$.

To make the prior distributions uninformative, the following choice of the hyperparameters is reasonable. We can configure the hyperparameters of $\mathbf{L}$ as $\nu = D$ and $\Lambda = \frac{\mathbf{I}}{D}$, where $\mathbf{I}$ is the identity matrix. This is because the lower degree of freedom ν increases the uncertainty of $\mathbf{L}$ and the diagonal scale matrix Λ means each dimensionality is uncorrelated a priori. For $\mathbf{M}$,

a zero matrix is a preferable choice to assume no temporal correlation before observing the time series. For $\mathbf{K}$, setting $\mathbf{K} = \mathbf{I}$ is sensible for the same reason; the lagged autocorrelation is ignored a priori.

5.2.1 Model variations

While the basic model presented above provides a straightforward inference procedure thanks to the conjugate priors, the flexibility of the model is somewhat limited. For example, the excitation at each time follows a stationary normal distribution with precision $\mathbf{L}$. This stationarity cannot satisfactorily describe distributions of many classes of natural signals and phenomena. To alleviate this restriction in the excitation term, we explain two variations in an effort to preserve the efficient inference.

First, to account for an excitation distribution with a heavier tail than the normal distribution, we introduce a scale parameter $s_t > 0$ for each time as $\mathcal{N}(\mathbf{A}^{(r)\top}\bar{\mathbf{x}}_t^{(r)}, s_t\mathbf{L})$. When s_t follows a gamma distribution, we can interpret that the excitation follows the Student's t-distribution. With this auxiliary scale variable s_t, the modifications in the posterior estimation is the calculation of autocorrelation matrices: $\mathbf{R}_{\bar{x}\bar{x}}^{(r)} = \sum_t s_t \bar{\mathbf{x}}_t^{(r)} \bar{\mathbf{x}}_t^{(r)\top}$, $\mathbf{R}_{\bar{x}x}^{(r)} = \sum_t s_t \bar{\mathbf{x}}_t^{(r)} \mathbf{x}_t{}^\top$, and $\mathbf{R}_{xx} = \sum_t s_t \mathbf{x}_t \mathbf{x}_t{}^\top$. If we fix this scale with a certain function as $s_t = f(\mathbf{x}_t)$, we can use Eqs. (5.2, 5.3) for the inference with the modified autocorrelation matrices. We use the fixed scales for a speech dereverberation task in Section 5.4.3 because the excitation, the pure speech signal, better fits a non-Gaussian distribution with a heavy tail.

Second, we consider a case in which the precision matrix of excitation is different from the row-wise precision of AR coefficients $\mathbf{A}$, for example, time-varying precision of the excitation $\mathbf{L}_t$. Here, we omit $\cdot^{(r)}$ for simplicity. Note that this is a more general case compared to the case discussed in the previous paragraph, where the modification of the precision is only up to the scalar scaling by s_t. If the observation $\mathbf{x}_t$ and AR coefficients $\mathbf{A}$ do not have a common precision matrix, the posterior computations in Eq. (5.2) no longer hold. We can still analytically derive the posterior of $\mathbf{A}$ using the vectorized notation $\mathrm{vec}(\mathbf{A})$. The model and inference of time-varying precision $\mathbf{L}_t$ is beyond the focus of this chapter, but interested readers can refer to the Wishart processes (Wilson and Ghahramani 2011) for examples.

Finally, we mention the complex-valued observations. Complex values appear in many

signal processing applications, typically with the use of Fourier-transformed representations. We can naturally extend the model into the complex region. In this case, we need to replace the distributions with complex-valued distributions such as those in (van den Bos 1995, Conradsen et al. 2003). We can still adopt the update rules in Eq. (5.2).

5.3 Bayesian Autoregressive Models with Infinite Orders

This section presents our autoregressive processes that bypass the truncation of possible orders to consider. The basic idea is shared with existing approaches (Godsill 2001, Vermaak et al. 2004); we place a prior on the order r and discuss the posterior of r given the observation. The novelty of our model is that the order is unbounded, unlike the exising models that truncate the maximum order at some point. Our model is able to exhibit a powerful flexibility thanks to Bayesian nonparametric modeling over the order of AR processes. For the prior of the order, we employ a similar approach used for ∞-gram language models where the observations are discrete words (Mochihashi and Sumita 2008). Our model can be viewed as an extension into the infinite order model on continuous variables through the linear Gaussian AR processes.

The prior of the order is constructed through the following stick-breaking procedures:

$$p(r|\{\theta_l\}_{l=0}^{\infty}) = \theta_r \prod_{l=0}^{r-1}(1-\theta_l), \qquad \theta_l \sim \mathcal{B}(\alpha_l, \beta_l), \tag{5.4}$$

where $\mathcal{B}(\alpha, \beta)$ denotes the beta distribution. We can marginalize out the beta parameter θ_l in Eq. (5.4) as explained in (Mochihashi and Sumita 2008). Note that this construction sums up to unity $\sum_{r=0}^{\infty} p(r) = 1$ and penalizes a larger order, which acts like a regularization to avoid overfitting. With this prior, we can view our model as a mixture model indexed by r:

$$\mathbf{x}_t | r, \mathbf{A}^{(r)}, \mathbf{L} \sim \mathcal{N}(\mathbf{A}^{(r)\top} \bar{\mathbf{x}}_t^{(r)}, \mathbf{L}), \qquad \mathbf{A}^{(r)} \sim \mathcal{MN}(\mathbf{M}^{(r)}, \mathbf{K}^{(r)}, \mathbf{L}) \ \text{ for } \forall r \geq 0, \tag{5.5}$$

where $\mathbf{L}$ is the same as in Eq. (5.1). The case $r = 0$ means that $\mathbf{x}_t$ is i.i.d. from $\mathcal{N}(\mathbf{0}, \mathbf{L})$.

Figure 5.2 shows some variations. We basically use *shared order model* that shares the same order for one AR proceess, as in Fig. 5.2 (a), though the original model (Mochihashi and Sumita 2008) varies the order for each point as r_t, as in Fig. 5.2 (b). This is because the *point-wise order model* often overfits data generated from a stationary AR process, as we

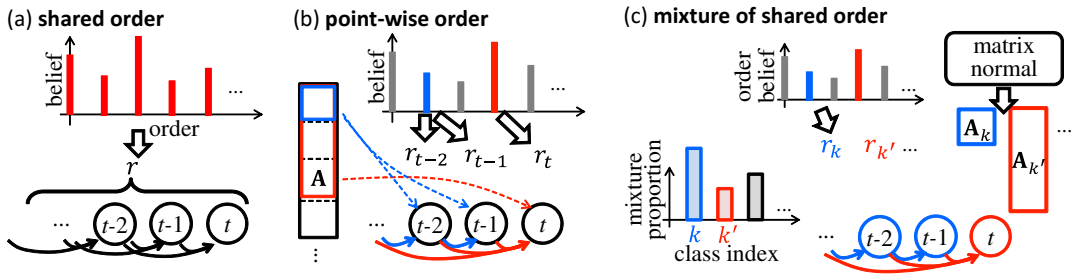


Figure 5.2: Variation of AR models with infinite orders. (a) A single order r drawn from the order belief is shared among the samples. (b) Point-wise order model draws orders for each time r_t whereas the AR coefficients are shared. (c) We can also construct a mixture of shared order AR processes. As a mixture component, the order r_k, AR coefficients $\mathbf{A}_k{}^{(r_k)}$, and excitation precision $\mathbf{L}_k$ are generated.

confirm in our experiments in Section 5.4.1. We can extend the model to a mixture of AR processes for non-linear regressions, as shown in Fig. 5.2 (c). The order may vary as denoted by r_k, where k is the index for each AR process.[1]

5.3.1 Inference

We first present the inference of the shared order model where we can analytically collapse out the AR coefficients $\mathbf{A}^{(r)}$ and excitation precision matrix $\mathbf{L}$. We then discuss the point-wise order case and the sampling of AR coefficients. Though the model considers infinitely large orders, the inference involves only finite-order computations. We use slice sampling (Neal 2003) to numerically generate the samples from the following posterior distribution:

$$p(r|\mathbf{X}) \propto p(r)p_r(\mathbf{X}), \qquad p_r(\mathbf{X}) := \iint p(\mathbf{X}|\mathbf{A}^{(r)},\mathbf{L})p(\mathbf{A}^{(r)},\mathbf{L})d\mathbf{A}^{(r)}d\mathbf{L}. \tag{5.6}$$

In concrete terms, the sampler moves from the current order r to the new order r' in the following steps. (1) Draw a height of slice u from the uniform distribution $\mathcal{U}([0: \ p(r)p_r(\mathbf{X})])$. (2) Make a slice $[0: \ r^+]$, where r^+ is extended from the current position r with a certain interval as $r^+ \leftarrow r + \Delta$. (3) If the marginal likelihood of r^+ comes above u, extend the slice, namely, repeat $r^+ \leftarrow r^+ + \Delta$ until $u > p(r^+)p_{r^+}(\mathbf{X})$. (4) Randomly choose the candidate order as $\tilde{r} \sim \mathcal{U}([0: \ r^+])$. If $p(\tilde{r})p_{\tilde{r}}(\mathbf{X}) > u$, then $r' = \tilde{r}$. If not, shrink the slice such that $r^+ \leftarrow \tilde{r}$ and

[1] Due to the limited space, we leave the detailed presentation of the mixture models as future work.

repeat until r' is determined. In practice, step (3) stops at a certain order because both $p(r)$ and $p_r(\mathbf{X})$ decrease rapidly with a large r.

We use the following Gibbs sampler for the point-wise order model to update r_t of each point $\mathbf{x}_t$:

$$p(r_t|\mathbf{x}_t, \mathbf{X}^{-t}, \mathbf{r}^{-t}) \propto p(r_t|\mathbf{r}^{-t}) \iint p(\mathbf{x}_t|\mathbf{X}^{-t}, \mathbf{A}^{(r_t)}, \mathbf{L}) d\mathbf{A}^{(r_t)} d\mathbf{L}, \tag{5.7}$$

where $\mathbf{X}^{-t}$ and $\mathbf{r}^{-t}$ is a set of variables excluding the sample at time t. We can use the same slice sampler to generate each r_t. The computation of $p(r_t|\mathbf{r}^{-t})$ is presented in (Mochihashi and Sumita 2008). The marginalization above is carried out using a correlation matrix calculated through the instantiated orders $\mathbf{r}^{-t}$. For example, the top-left $Dr_t \times Dr_t$ block of $\mathbf{R}_{\bar{x}\bar{x}} = \sum_{s \neq t} \bar{\mathbf{x}}_s^{(r_s)} \bar{\mathbf{x}}_s^{(r_s)\top}$ is used to marginalize $\mathbf{A}^{(r_t)}$.

5.3.2 Construction of AR coefficients by partitioned matrices

This section provides a way to handle the AR coefficients of the infinite order. The inference presented above enjoys an efficient mixing of orders because $\mathbf{A}^{(r)}$ and $\mathbf{L}$ are collapsed out. In case we have to explicitly sample the AR coefficients, we need an efficient and tractable way to handle a possibly infinite-dimensional variable $\mathbf{A}^{(\infty)}$. For example, we should sample the AR coefficients when the precision matrix of $\mathbf{x}_t$ and row-wise precision of $\mathbf{A}$ are different, as noted in Section 5.2.1. Existing Bayesian models (Godsill 2001, Vermaak et al. 2004) use RJMCMC (Green 1995) to sample AR coefficients of different dimensionalities. They avoids considering the infinite-order case by truncating the model up to a certain maximum order. Furthermore, as in Fig. 5.3 (c), RJMCMC is sometimes inefficient because the samples are not necessarily updated all time due to the acceptance or rejection procedure. To overcome these issues, we introduce the partitioned matrices of AR coefficients depicted in Fig. 5.3.

We assume the hyperparameter $\mathbf{K}^{(\infty)}$ is block diagonal consisting of $D \times D$ matrices $\mathbf{K}_l$ so that AR coefficients are partitioned into $D \times D$ matrices. Then, generating $\mathbf{A}^{(\infty)}$ is equivalent to independnetly generating each partitioned matrix $\mathbf{A}_l \in \mathbb{R}^{D \times D}$, as shown in Fig. 5.3 (a):

$$\begin{aligned} &\mathbf{A}^{(\infty)\top} = [\mathbf{A}_1^\top, \mathbf{A}_2^\top, ..., \mathbf{A}_l^\top, ...]^\top, \quad \mathbf{K}^{(\infty)} = \mathrm{diag}[\mathbf{K}_1, \mathbf{K}_2, ..., \mathbf{K}_l, ...], \\ &\mathbf{A}^{(\infty)} \sim \mathcal{MN}(\mathbf{M}^{(\infty)}, \mathbf{K}^{(\infty)}, \mathbf{L}) \Leftrightarrow \quad \mathbf{A}_l \overset{i.i.d.}{\sim} \mathcal{MN}(\mathbf{M}_l, \mathbf{K}_l, \mathbf{L}) \ (l = 1, 2, \ldots). \end{aligned} \tag{5.8}$$

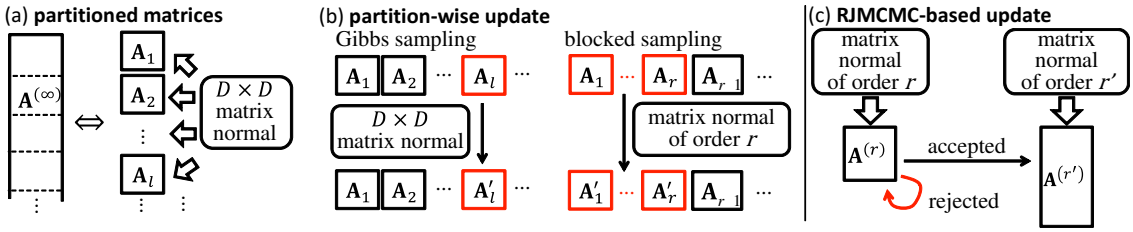


Figure 5.3: (a) Partitioned matrices sharing the same prior. (b) Partition update with a Gibbs sampler. Updating multiple partitions at once is also tractable using the equivalent matrix normal distribution. (c) RJMCMC involves acceptance or rejection of newly proposed sample to move to different dimensionality, for example, order r to r'. The risk of rejection slows the mixing of MCMC.

This construction comes from the property of the multivariate normal distribution; the uncorrelated elements become independent of each other. The assumption that $\mathbf{K}^{(\infty)}$ is block diagonal is not too limiting in practice because we often choose the identity matrix for $\mathbf{K}$ to reflect the uninformative prior knowledge, as discussed in Section 5.2.

We then introduce a sequence of binary variables $\{z_l\}_{l=1}^{\infty}$ as an equivalent representation of the order r. These variables are set as $z_l = 1$ for $l \leq r$ and $z_l = 0$ otherwise. These binary variables are used to mask out unnecessary past samples. Using this notation, we can rewrite the model in Eq. (5.5) into

$$\mathbf{x}_t | \{z_l\}, \{\mathbf{A}_l\}, \mathbf{L} \sim \mathcal{N}\left(\sum_{l=1}^{\infty} z_l \mathbf{A}_l^\top \mathbf{x}_{t-l}, \mathbf{L} \right), \qquad r \sim p(r), \qquad z_l = \mathbf{1}(l \leq r), \tag{5.9}$$

where $p(r)$ comes from Eq. (5.4) and $\mathbf{1}(y) = 1$ if y is true or 0 otherwise. We can consider a partition-wise Gibbs sampler for the inference such as $p(\mathbf{A}_l | \mathbf{X}, \{z_l\}, \{\mathbf{A}\}^{-l})$, which dispenses with the risk of rejection like RJMCMC. See Fig. 5.3 (b) for the graphical explanation. We can also recover the same posterior distribution with the parameters in Eq. (5.4) with a blocked inference such as $p(\mathbf{A}_1, \ldots, \mathbf{A}_r | \mathbf{X}, \{z_l\}, \{\mathbf{A}\}^{-(1:r)}, \mathbf{L}) = \mathcal{MN}(\hat{\mathbf{M}}^{(r)}, \hat{\mathbf{K}}^{(r)}, \mathbf{L})$, if the instantiated order represented by $\{z_l\}$ equals r. $\{\mathbf{A}\}^{-(1:r)}$ is a set of partitioned matrices excluding $\mathbf{A}_1, \ldots, \mathbf{A}_r$. This udpate of mutiple partitions accelerates the MCMC-based inference. The rest of the matrices $\{\mathbf{A}\}^{-1:r}$ follow the prior distribution in Eq. (5.8). We can still use the slice sampler to update r, or equivalently $\{z_l\}$, by using the instantiated matrices $\{\mathbf{A}_l\}$. If the slice requires a larger order than the instantiated AR coefficients, we can draw them from the prior distribution $p(\mathbf{A}_l)$ on demand. The same applies to the vectorized cases in Section 5.2.1 by considering

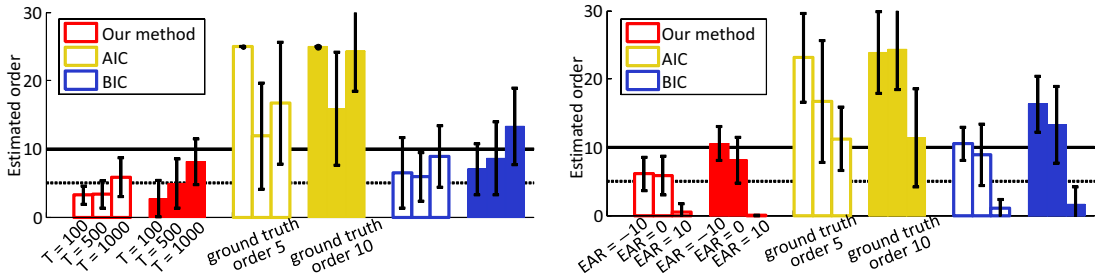


Figure 5.4: Comparison of order selection performance between our method, AIC, and BIC. Left: results for different lengths with EAR = 0 (dB). Right: different EARs with length 1000. White-face bars and color-face bars represent the ground truth order $r = 5, 10$, respectively. Vertical segments at each bar are the standard deviation. Black horizontal lines indicate the ground truth orders.

the arrangement of vectorized elements corresponding to each partition matrix $\mathbf{A}_l$.

At first sight, the likelihood in Eq. (5.9) resembles the latent feature models (Griffiths and Ghahramani 2006) in that infinitely many features are either activated or suppressed through the binary variables z_l. The essential difference is that the standard latent feature models assume the exchangeability between each z_l, and our model does not. Rather, our model partitions the binary variables to all-one region $z_1 \ldots z_r = 1 \ldots 1$ and all-zero region $z_{r+1} \ldots = 0 \ldots$ with the stick-breaking construction in Eq. (5.4).

5.4 Experimental Results

This section presents the experimental results using synthetic data and real data. Using synthetic data, we compare the order selection performance of our method with conventional model selection methods based on AIC (Akaike 1974) and BIC (Schwarz 1978). We also investigate the prediction performance of a point-wise model where the order r_t may vary for each time t, as in Eq. (5.7), compared with the shared order model where the order is the same throughout the sequence, as in Eq. (5.6).

As a real dataset, we use audio recordings to perform a multichannel dereverberation task (T. Yoshioka and Miyoshi 2009). In this experiment, we compare the shared order model with the fixed-order Bayesian AR model presented in Section 5.2 and the model selection approaches based on AIC and BIC.

5.4.1 Synthetic data: order estimation

We first present the order selection results comparing our shared order model with AIC and BIC. These methods use the maximum likelihood estimation for the parameters to calculate the selection criterion. AIC and BIC calculate their criteria from order 0 to 30 and choose the best order.

The objective of this experiment is to investigate the influence of the length of observations and the power of the excitation component. We use the excitation-to-autoregression ratio (EAR) to measure the power of excitation. EAR ξ is defined as $\xi = 10\log_{10}\frac{\sum_t \|\mathbf{e}_t\|^2}{\sum_t \|\mathbf{x}_t - \mathbf{e}_t\|^2}$ (dB), where $\mathbf{e}_t$ is the excitation component generated from $\mathcal{N}(\mathbf{0}, \mathbf{L})$. When EAR is large, the data behaves like i.i.d. samples. The order estimation may become more difficult with larger EARs because the AR component is dominated by the excitation, and vice versa. We can modify EARs by adjusting the scale of the AR coefficients. The lengths T were set to 100, 500, and 1000 while EAR was -10, 0, and 10. The order of the synthetic data was 5 or 10 with the dimensionality $D = 2$. For each condition, we randomly generated 10 sequences with random AR coefficients and excitation precision matrix, and fed them into the algorithms to estimate the order. The hyperparameters of our model in Eq. (5.4) were set to $\alpha_l = 1, \beta_l = 1$. Other parameters were set uninformative, as noted in Section 5.2.

Figure 5.4 illustrates the order selection results with various data lengths and EARs. AIC is apt to choose larger orders in most configurations mainly because the increase in the likelihood with a larger order is scarcely penalized by the number of free parameters. In the left figure, both our method and BIC increase the order estimates as the data length becomes large. While BIC exceeds the true order with an increasing amount of data, our method robustly refrains from increasing the order estimates. The right figure shows the results with various EARs. The results confirm that the estimated order decreases with a larger EAR. Furthermore, our method outperforms the others when EAR $= -10$ (dB), where the autoregression is saliently observed.

5.4.2 Synthetic data: point-wise model

In this experiment, we compare the prediction performances of the point-wise model and the shared order model. For each given observation sequence, the first 90% $\mathbf{X}_{\text{train}}$ is used for the

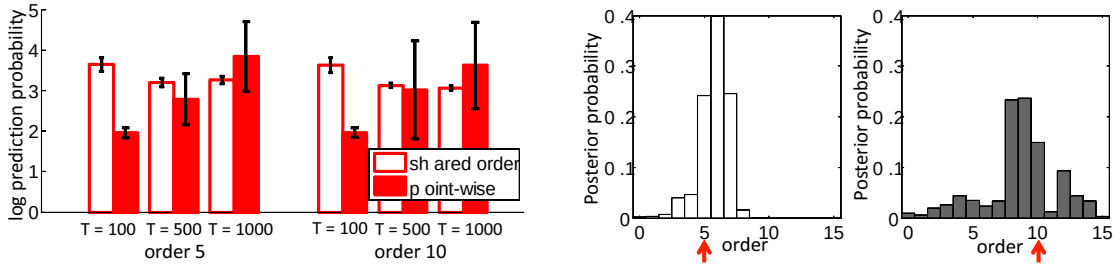


Figure 5.5: Left: Log prediction probabilities of held-out data by shared order model and point-wise order model. Larger value implies better prediction performance. Right: Posterior probability of sampled orders by point-wise model where the ground truth order is indicated by red arrows.

inference and the log prediction probability (LPP) of the last 10% sequence $\mathbf{X}_{\text{test}} = \{\mathbf{x}_t\}_{t=1}^{T'}$ is computed using MCMC samples in a similar manner to the computation of perplexity per word.

$$
\begin{aligned}
LPP &= \frac{1}{T'} \log p(\mathbf{X}_{\text{test}} | \mathbf{X}_{\text{train}}) \\
&\approx \frac{1}{T'N} \sum_{t=1}^{T'} \sum_{n=1}^{N} \log \iint p\left(\mathbf{x}_t | r^n, \mathbf{A}^{(r^n)}, \mathbf{L}\right) p\left(r^n, \mathbf{A}^{(r^n)}, \mathbf{L} | \mathbf{X}_{\text{train}}\right) d\mathbf{A}^{(r^n)} d\mathbf{L},
\end{aligned}
$$

where r^n is the nth MCMC sample of the order of the shared order model. The LPP of the point-wise model is similarly calculated.

The synthesized data was configured as follows: the order was set to 5 and 10 and the length was 100, 500, and 1000. EAR was fixed at 0 (dB), and the dimensionality $D = 2$. For each configuration, we randomly generated 10 sequences and LPP was calculated by both models. The left of Figure 5.5 shows the LPPs of both methods for each condition. When the length of the data is short, the point-wise model has an inferior prediction performance. When $T = 1000$, the mean LPP of the point-wise model is better than that of the shared order model, however, the variance is rather high. Taking into account that the point-wise model requires more computational resources for the inference due to a larger latent space, the use of the point-wise model may be costly. The right of Figure 5.5 shows the posterior probability of the order obtained by the MCMC samples of the point-wise model. The graph exhibits a high variance in the posterior order, which results in high-variance LPPs.

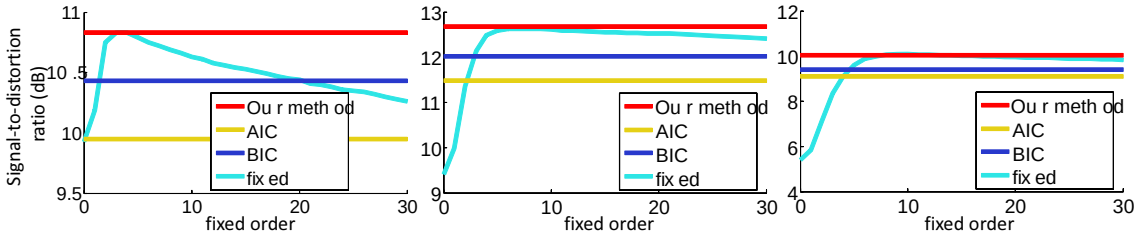


Figure 5.6: Dereverberation performance in terms of signal-to-distortion ratio (dB) in three rooms. Left: RT 150 ms, center: RT 400 ms, right: RT 600 ms. Larger values indicate better performance.

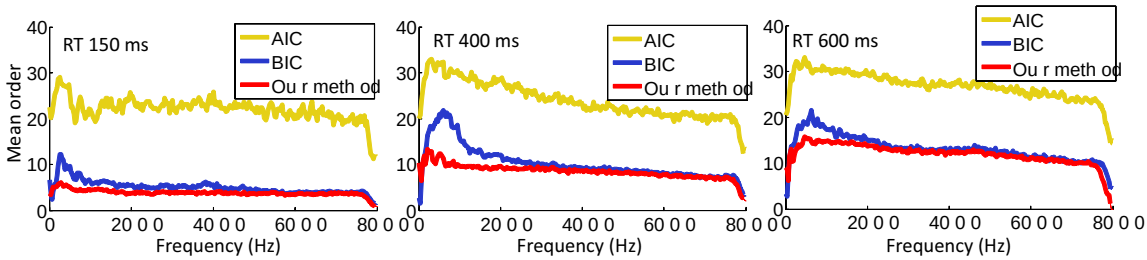


Figure 5.7: Mean values of estimated order at each frequency bin with three methods. Left: RT 150 ms, center: RT 400 ms, right: RT 600 ms.

5.4.3 Multichannel dereverberation

This section presents the results of multichannel dereverberation task, where multichannel means the sound is observed with multiple microphones. The dereverberation problem is formulated using an AR process as follows. First, we make the observed waveform into the time-frequency domain through the short-time Frourier transform. In our setup, the waveform was sampled at 16000 (Hz). We used a 512-point Hanning window with 256-point shift for the processing and we obtained multichannel vectors $\{\mathbf{x}_{tf}\}_{1\leq t\leq T, 1\leq f\leq F}$, where t is the time frame index, f is the frequency bin index. The dimensionality of $\mathbf{x}_{tf}$ is the number of microphones. As a result of Fourier transform, $\mathbf{x}_{tf}$ is a complex number. The reverberation is modeled as follows. For each frequency bin f, a multichannel reverberant signal follows an AR process $\mathbf{x}_{tf} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{A}_f{}^{(r)H}\bar{\mathbf{x}}_{tf}, s_{tf}\mathbf{L}_f)$, where $\mathcal{N}_{\mathbb{C}}(\cdot,\cdot)$ denotes the complex normal distribution and $\mathbf{C}^H$ is a Hermitian transpose. As discussed in Section 5.2.1, we introduce a scale factor $s_{tf} = \|\mathbf{x}_{tf}\|^{-2}$ to adapt to the distribution of speech signals with a heavy tail. The dereverberated signal is restored as $\hat{\mathbf{x}}_{tf} = \mathbf{x}_{tf} - \hat{\mathbf{A}}_f^{(r)H}\bar{\mathbf{x}}_{tf}$, where $\hat{\mathbf{A}}_f^{(r)}$ is the estimate of the AR coefficients. The Bayesian

methods use the MMSE estimate in Eq. (5.2) whereas AIC and BIC-based methods use maximum likelihood estimates. The shared order model averages the dereverberated signals $\hat{\mathbf{x}}_t$ with the sampled orders.

We recorded the impulse responses of three rooms with reverberation times (RTs) of 150, 400, 600 (ms). If RT is large, a larger order is generally required to cope with the long reverberation. The reverberation length are also dependent on the frequency bin, and even the relative position between the sound source and the microphones. Therefore, the order is a critical parameter for the dereverberation procedure. We used four microphones to record the sound located at five positions in each room. For each position, 10 utterances of a distinct speaker's clean speech signal from JNAS phonetically-balanced Japanese utterances were convolved to generate a reverberant speech signal. The speakers consisted of three males and two females, corresponding to each position. The average length was 6 seconds, around $T = 350$ in the time-frequency domain. Signal-to-distortion ratio (SDR) (Vincent et al. 2006) was used to measure the dereverberation quality. This measure is maximized when the algorithm removes the reverberant component while preserving the original speech component.

Figure 5.6 summarizes the SDR scores of our shared order model, the fixed-order Bayesian AR, the AIC, and the BIC. Our method, AIC, and BIC estimated the order for each frequency bin. Meanwhile, the fixed Bayesian model used the same order for all frequency bins. The fixed order ranged from 0 to 30, whereas AIC and BIC chose their order from 0 to 40. Our method outperforms AIC and BIC partly because the Bayesian estimation of the AR coefficients stabilizes the dereverberation. We also find that the score of our method is the maximum point of the fixed AR models while the performance of the fixed-order model is influenced by too small or large order values. This indicates that our method can efficiently find an optimum order depending on different conditions.

Figure 5.7 shows the estimated order at each frequency bin. We can confirm a simliar tendency as in Figure 5.4: the AIC prefers far larger orders. Comparison of our model and BIC again shows that our method prefers smaller orders. We can isolate two factors that contribute to the growth of orders: (1) as the RT increases, the order grows accordingly, and (2) larger orders are selected at lower frequency regions. This is partly because a lower frequency component is less attenuated on the reflection of the walls. While a similar configuration has

been manually developed in related studies, for example (Yoshioka et al. 2011), our method automatically achieves a reasonable order selection.

5.5 Summary

This chapter presented Bayesian autoregressive processes that allows for infinite orders. Our generative model and corresponding inference scheme aim to determine the appropriate order given the observed sequential data. Our model can be noted as an extension of the infinite Markov model of discrete variables (Mochihashi and Sumita 2008) into the continuous domain through linear Gaussian modeling. One of the future work is to verify our model on complex time series beyond the linear Gaussian structure by incorporating nonparametric Bayesian models (Fox et al. 2011, Lucca et al. 2013). Other work includes a model extension into AR variants such as autoregressive-moving-average (ARMA) models and application to time series analyses such as anomaly detection or missing value compensation.

Chapter 6

Sound Source Separation in Dynamic Environments

This chapter addresses the third uncertainty in auditory scenes–dynamic sound sources. In particular, we tackle the sound source separation problem where the relative positions between the microphone array and the sound sources change over time.

6.1 Introduction

One of the most important functions for mobile robots is the capability to perceive and analyze the information in the environment with the sensors equipped on the robots. This function is essential for mobile robots to probe an unknown environment in the both cases when a stand-alone robot is in operation and when a robot is operated by a remote human operator. Simultaneous localization and mapping (SLAM) (Thrun et al. 2004, Se et al. 2005) based on visual information captured by cameras on the robot has recently been developed as a perception capability of mobile robots. In addition to these visual information processing, the auditory processing function may enable the robots to acquire the following functions: (1) the robust perception against occlusions of objects, (2) the detection of changes in objects, and (3) the spoken communication. For example, (1) while a robot may have difficulty in detecting an object or a certain event over a wall using only the visual information, the robot may be capable of listening to them by using its auditory function. (2) A movement of an object or a change in the situation often accompanies some sounds. Figure 6.1 illustrates an example of a glass that falls from a table, which emits a cracking sound. Robots with the

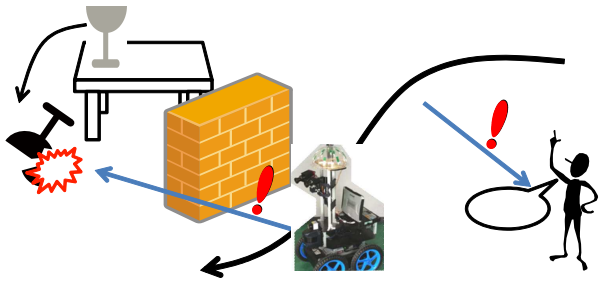


Figure 6.1: Auditory functions by a mobile robot.

auditory function are expected to detect this sort of event. (3) Needless to say, the auditory function can establish a communication channel between robots and humans such that robots can accepts the humans' oral commands.

Since many environments contain multiple sounds, a robot that probes the environment must be capable of extracting the sources from the observed mixture audio signal. A microphone array is often employed to cope with the separation of sound sources (Benesty et al. 2008). An application of microphone array technique to mobile robots involves the following two issues:

1. the suppression of the noise caused by the mobility of the robot itself, and
2. the changes in the relative positions between the microphone array and the sound sources.

We cope with the suppression of the self noise within the framework of an ordinary microphone array processing: the motor noise of our robot is separated as a sound source arriving from a distinct direction. The second issue is a critical challenge that we have to overcome in the application of microphone array-based techniques to mobile robots. In fact, most separation methods using a microphone array assume that the propagation of sounds is time-invariant, whereas the sound propagation is determined by the relative position between the microphones and sound sources.

The basic idea to cope with this changing propagation issue is to segment the observed mixture signal along the time axis such that each source is assumed to keep a stable position, that is, to have a time-invariant sound propagation in each segment. The segmentation of the observed audio signal is automatically carried out by applying the algorithm presented in Chapter 3. Using the property of the formulation using Bayesian nonparametrics that enables

the source separation algorithm to handle any number of sources, this algorithm separates the sound sources in the mixture while splitting one moving source into several segments localized in different directions. After this separation and localization process, several separated segmentations are integrated to retrieve the audio signal of moving sources.

6.2 Problem Statement and Method Overview

This section describes the problem of the sound source separation in a dynamic situation, and outlines our method with comparison to existing methods. Figure 6.2 illustrates the assumption of the separation problem in (a), and the idea of our method using the separation and localization algorithm based on Bayesian nonparametrics.

The following two points are assumed in the problem of moving sound source separation.

1. Each sound source may change its direction from the microphone array within a disjoint range from each other, as shown in Figure 6.2 (a).
2. The direction range of each source is known.

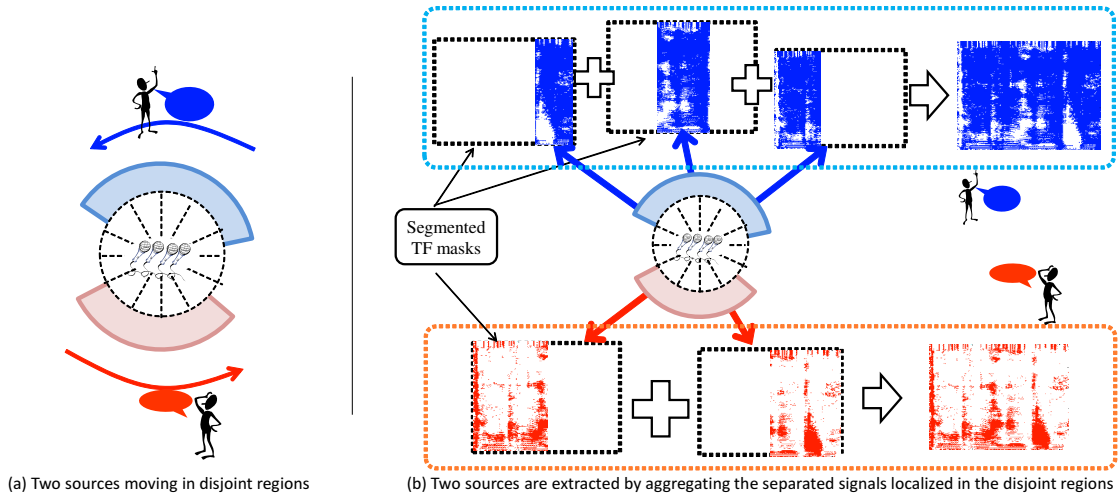
These assumptions are used to facilitate the integration of segmented signals into each moving source signals. As depicted in Figure 6.2 (b), our method estimates some segmented TF masks from the observed mixture of moving sound sources. The segmentation is simultaneously carried out during the separation and localization process such that the sources in each segment is regarded as stable and distinct sources. For example, in the experiment presented in Section 6.3, two sources are located in the left and right sides of the microphone array equipped on a mobile robot.

6.2.1 Algorithm

According to the assumptions, we have a set of directions $\mathcal{D}_n$ for each source n corresponding to the direction range in which source n may move. Note that the sets are disjoint as

$$\mathcal{D}_n \cap \mathcal{D}_{n'} = \phi \qquad (\text{if } n \neq n'), \tag{6.1}$$

where ϕ denotes the empty set.



(a) Two sources moving in disjoint regions

(b) Two sources are extracted by aggregating the separated signals localized in the disjoint regions

Figure 6.2: Assumed situation and our method. (a) We assume that the microphone array observes multiple sound sources, depicted in blue and red, with their range of direction from the microphone array is disjoint from each other. (b) Our Bayesian nonparametrics-based algorithm separates and localizes the moving sources through the TF mask estimation while segmenting the TF-masks along the time axis. This segmentation is automatically achieved such that the constituent sources are regarded as stable sources in each segment. Then, we can reconstruct each moving source by integrating the segmented signals localized in each disjoint range of direction.

Let $\mathbf{x}_{tf}$ be the multichannel observation of the mixture of sound sources and $\hat{\mathbf{x}}^d_{tf}$ be the separation result from direction d, which is obtained by the algorithm presented in Chapter 3. After obtaining the segmented signals $\hat{\mathbf{x}}^d_{tf}$, we can integrate these segments into the audio signal $\hat{\mathbf{x}}^n_{tf}$ corresponding to source n by the following post processing.

$$\hat{\mathbf{x}}^n_{tf} = \sum_{d \in \mathcal{D}_n} \mathbf{x}^d_{tf}. \tag{6.2}$$

6.3 Experimental Results with A Mobile Robot

This section first describes the setup of the recording using a microphone array on a mobile robot, and then presents the separation results in two environments. Figure 6.3 depicts the location of the sound sources in the recording environment with the path of the robot movement, and the microphone array equipped to the robot. As shown in the left part of Figure 6.3, the mobile robot moved along the path between the two sound sources to record the audio

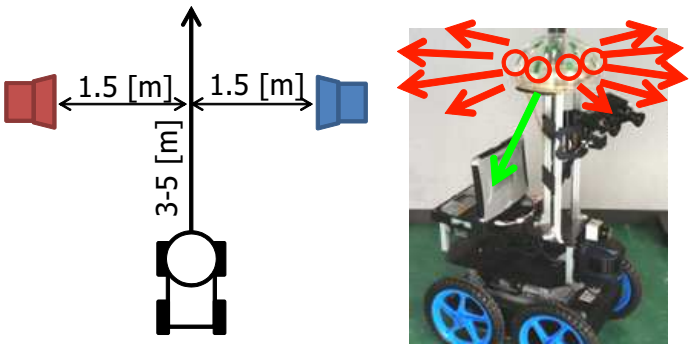


Figure 6.3: Left: positions of sound sources and the path of the mobile robot. Right: Microphone array embedded at the top of the robot. The right arrows represent the directions for azimuth localization with 5° resolution, whereas the green arrow corresponds to the direction of wheels.

mixture. One source was located on the left side, whereas the other source was located on the right side of the robot. This configuration meets the assumption in Section 6.2.

6.3.1 Experimental setup

The experiment was carried out in two environments using an eight-channel microphone array embedded at the top of the robot. One environment is an outdoor situation where the reverberation time is RT_{60} = 150 (ms), and the other is an indoor environment with the reverberation time RT_{60} = 800 (ms). In order to demonstrate the robustness of our source separation algorithm against the type of sound sources, a musical audio signal containing guitar or piano performance was played from the loudspeaker at the right, illustrated in blue in Figure 6.3, and various sounds such as human speech signal and calls of crickets or frogs were played from the loudspeaker at the left side, illustrated in red in Figure 6.3. The length of the recording was 20 (s). In both environments, the robot moved on a flat floor, with a few undulations, so that the robot can move smoothly.

The steering vectors for the separation and localization algorithm were configured as follows. For the azimuth localization, 72 steering vectors were used with 5° resolution (the red arrows in Figure 6.3). In order to separate the noise from the moving wheels, the steering vector toward the mobile unit (the green arrow in Figure 6.3) was also added. In total, D = 73 directions were taken into account for the separation and localization algorithm.

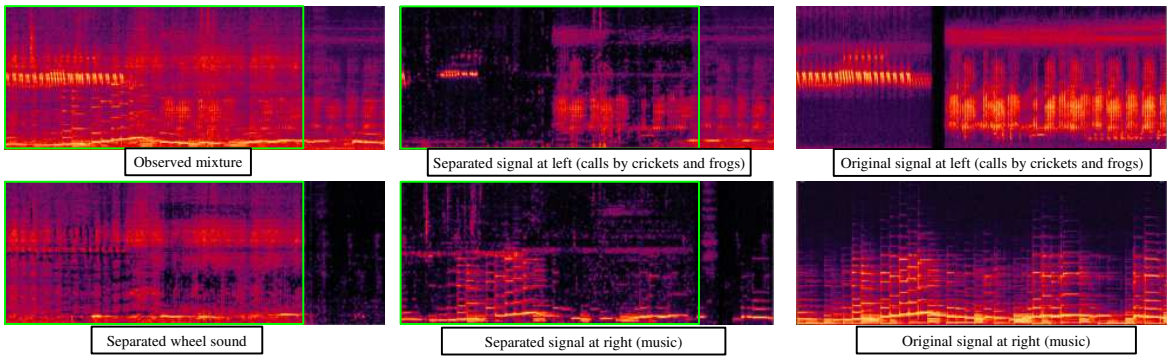


Figure 6.4: Separation result of outdoor recording with the reverberation time RT_{60} = 150 (ms).

6.3.2 Results

Figures 6.4 and 6.5 shows the spectrograms of the observed mixture signal, separated signals, and the original signals in indoor and outdoor environments, respectively. The green frames in the observed mixture and separated signals indicate the period when the robot was moving. The separated signals were obtained by integrating the separated segments localized on the left or right side of the robot. The sound emitted by the moving wheels were removed because the wheel noise was localized in the direction toward the mobile unit.

The separated signal corresponding to the wheel noise in the outdoor environment shown in Figure 6.4 is mute when the robot is halting, whereas the separated wheel sound in the indoor environment shown in Figure 6.5 contains other signals such as human speech even when the robot is not moving. Furthermore, the separated wheel sound contains the lower frequency region of the music audio signal, which should be segregated into the right side signal. Generally, in a reverberant environment, the separation quality, in particular in the lower frequency region, degrades severely when we use a model that assumes only direct sounds without an explicit model of reverberation components.

The sound source separation in the outdoor environment in Figure 6.4 is mostly achieved even during the robot is moving, though the first part of cricket calls on the left side is attenuated. We speculate the reason why the extraction of cricket calls was difficult is that the energy of the cricket calls is concentrated on a small range of frequency band and that the temporal duration of each call is also limited. If the energy of the other sound sources conflict

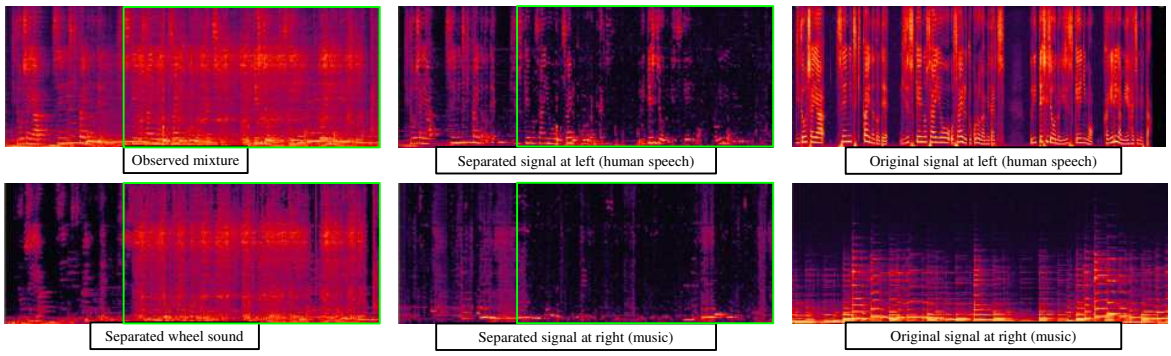


Figure 6.5: Separation result of indoor recording with the reverberation time RT_{60} = 800 (ms).

in this range, the estimation of TF masks to separate the cricket calls may be severely affected. On the other hand, the separation quality is degraded in the indoor environment, as shown in Figure 6.5, especially when the robot is moving. The reason for this deterioration is, on top of the severe reverberation, the power level of the right signal (musical audio signal) was lower than the other sounds.

Through this experiment, we have confirmed the following advantages of our method:

1. The separation of sound sources while the relative positions of the sources are time-variant.
2. The capability to suppress the noise signal caused by the movement of the robot itself.

We have also found some limitations that may degrade the separation quality such as the reverberation, a restricted frequency range of the target audio signal, and the low power level of the target signals.

6.4 Summary

This chapter tackled the source separation problem in a dynamic environment where the relative positions between the microphones and the sound sources may change over time. This problem is motivated by the realization of auditory functions on mobile robots in multi-source environments. Our method was designed based on the assumption that each source has disjoint range of directions. Thus, the Bayesian nonparametrics-based source separation and localization algorithm is helpful because this algorithm simultaneously separates and segments the

observed mixture such that each source can be retrieved by integrating the segmented signals in accordance with the direction range of each sound source. The experimental result showed the possibility of the separation of sound sources where the sound propagation is time-variant as well as the capability to suppress the noise caused by the movement of the robot itself. The result also indicated some limitations of our method such as the degradation of the separation quality due to the reverberation or the low power of the target signals.

We can consider several future directions to cope with the microphone array processing in dynamic environments. The first direction is to handle more general configuration unlike the disjoint range assumption. If the mobile robot moves along a complex path around many sound sources, the localization information would be unreliable information to integrate the separated segments to reconstruct each sound source. In order to achieve this integration process without the localization information, the identification of the sound sources by using their timbre (Sasaki et al. 2009b) should be carried out.

Another future direction is the suppression of the self noise. Our method successfully suppressed the wheel noise because the direction of the noise source was fixed from the microphone array. This may not be the case when complex movements with many motors cause the noise signal such as motions of a humanoid robot. Possible approaches are the noise suppression incorporating additional sensors attached to the noise source (Sawada et al. 2010), and the noise subtraction using noise templates estimated from the motor commands from the motor control module of the robot (Ince et al. 2011).

Chapter 7

Sound Source Localization in Dynamic Environments

In Chapter 6, sound source separation problem in a dynamic environment was tackled under the assumption that sound sources move but stay in disjoint ranges of direction. In this chapter, the problem is reduced to sound source localization where general movements of sound sources or microphones are allowed.

7.1 Introduction

Auditory information holds an important place in the human perception. Needless to say oral speech as a communication channel, humans perceive audio signals emitted by surrounding objects to understand their situation. For example, the sound of footsteps may inform people that somebody is approaching or moving away without any glance. Achieving a computational auditory function will help people, especially hearing-impaired people, to have enhanced auditory awareness by showing trajectories of audio sources or presenting enhanced speech signals (Kubota et al. 2008).

Sound source localization is the most fundamental and important function for distant speech enhancement and simultaneous speech separation using a microphone array (Nakadai et al. 2010), presentation of sound sources to the operator of a tele-presence robot (Mizumoto et al. 2011), and the detection and mapping of sound sources by a mobile robot (Sasaki et al. 2010). Figure 7.1 illustrates a robot standing in an auditory dynamic environment. There may be moving and multiple sound sources surrounding the robot or the microphone array system.

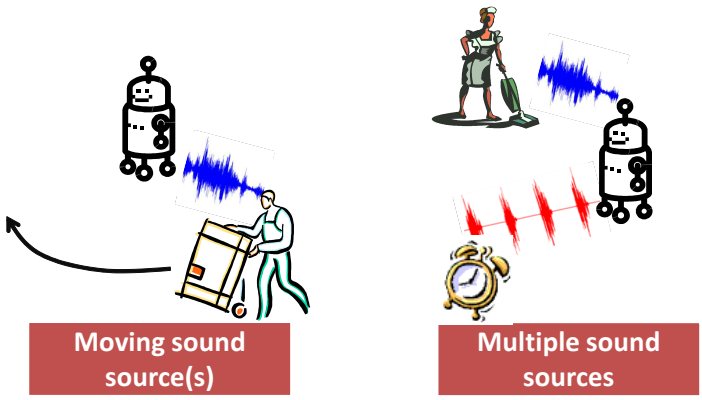


Figure 7.1: Sound source localization in a dynamic environment.

These system should robustly localize and track each sound source without a time-consuming parameter tuning.

For localizing sound sources with a microphone array, two methods have been widely exploited; beamforming (Doclo and Moonen 2001) and multiple signal classification (MUSIC) (Schmidt 1986, Asano et al. 2001, Danès and Bonnal 2010). Between these two methods, MUSIC is said to produce better localization performances because the evaluation function for the direction of arrival detection called MUSIC spectrum has much sharper peaks at the directions of sound sources than the evaluation function of the beamforming method. Furthermore, MUSIC is capable of detecting multiple sound sources on condition that the number of sound sources is less than that of microphones.

In the frame work of MUSIC-based sound source localization, the threshold should be carefully set for the MUSIC spectrum to detect active sound sources. The problem is that this threshold is inevitably dependent on the number of sound sources and the reverberation time of the environment. The estimation of the number of sound sources have so far been tackled with Akaike information criterion (Danès and Bonnal 2010) or a support vector machine (Yamamoto et al. 2006). However, an elaborate setting of the threshold is still necessary for the robust detection and tracking of sound sources. Typically, the threshold should be empirically tuned by looking into the MUSIC spectrum of the recording of the environment in question.

This paper presents a Bayesian extension of MUSIC-based multiple sound source localization and tracking. This method dispenses with most parts of the manual and empirical

parameter tuning that is critical to existing frameworks. Our method consists of two stages: (1) The parameters for the localization and tracking are automatically estimated using a pre-recorded audio signal as the learning data based on the variational Bayesian hidden Markov model (VB-HMM) (Beal 2003). (2) Our method incrementally localizes and tracks multiple sound sources with previously estimated parameters based on a particle filter (Arulampalam et al. 2002).

7.2 MUSIC-Based Sound Source Localization

This section specifies the problem and explains the MUSIC algorithm in general. We assume the sound source localization problem on the azimuth plane because we use a circular microphone array, as illustrated by Figure 7.2. In our configuration, the localization resolution is set 5 (deg). The problem statement is given as follows:

Input: M-channel audio signal, D steering vectors [1] for each direction and frequency bin,
Output: N directions where sound sources exist,
Assumption: the maximum number of sources is less than the number of microphones ($N \leq N_{max} < M$).

Here, we briefly outline the procedures of an ordinary MUSIC algorithm. Detailed explanation is provided in (Schmidt 1986, Danès and Bonnal 2010). The MUSIC algorithm is applied in the time-frequency domain. The short-time Fourier transform is carried out with the sampling rate 16,000 (Hz), the window length 512 (pt), the hop size 160 (pt).

The notation is defined in a different way from the previous chapters. In particular, the time frame index t in the TF domain is replaced with τ because the localization method presented in this chapter is carried out at longer intervals, as we see later on. The time index t denotes unit time index for this localization interval. Let $\mathbf{x}_{\tau f}$ denote the amplitude of input M-channel audio signal at time τ and frequency bin f. For each frequency f and time t at ΔT interval, (1) the correlation matrix $\mathbf{R}_{tf}$ of the input signal is calculated. (2) Then, the eigenvalue decomposition of $\mathbf{R}_{tf}$ is obtained. (3) Finally, the MUSIC spectrum is calculated using the eigenvectors and

[1] $D = 72$ in our implementation since the localization resolution is 5 (deg) in Fig. 7.2

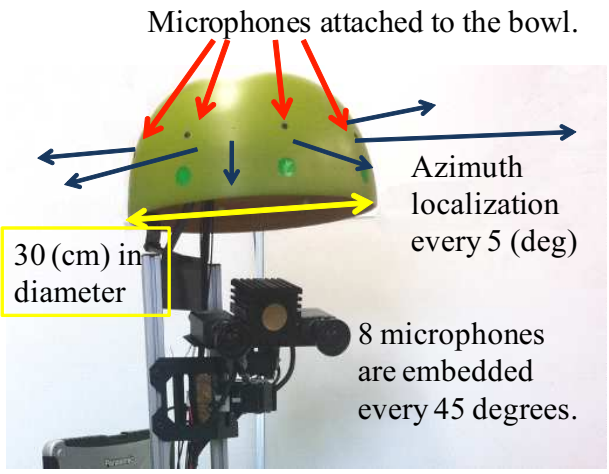


Figure 7.2: Sound source localization on the azimuth plane using an 8-channel MEMS microphone array on a mobile robot called "Kappa". The directions of sound sources are localized as blue arrows show.

the steering vectors.

(1) The correlation matrix is calculated by averaging over observed samples for ΔT (sec) as follows:

$$\mathbf{R}_{tf} = \frac{1}{\Delta T} \sum_{\tau=t-\Delta T}^{t} \mathbf{x}_{\tau f} \mathbf{x}_{\tau f}^{H}, \tag{7.1}$$

where $\cdot^H$ is the conjugate transpose operator. The vector $\mathbf{x}_{\tau f}$ has M elements corresponding to each channel.

(2) The eigenvalue decomposition of $\mathbf{R}_{tf}$ is given by

$$\mathbf{R}_{tf} = \mathbf{E}_{tf} \mathbf{V}_{tf} \mathbf{E}_{tf}, \tag{7.2}$$

where $\mathbf{E}_{tf}$ is the eigenvector matrix and $\mathbf{V}_{tf}$ is the eigenvalue matrix. The column vectors of $\mathbf{E}_{tf}$ are the eigenvectors of $\mathbf{R}_{tf}$, that is, $\mathbf{E}_{tf} = [\mathbf{e}_{tf}^1, \ldots, \mathbf{e}_{tf}^M]$. $\mathbf{V}_{tf}$ is a diagonal matrix with eigenvalues of $\mathbf{R}_{tf}$, that is, $\mathbf{V}_{tf} = \mathrm{diag}(v_{tf}^1, \ldots, v_{tf}^M)$. The eigenvectors are arranged in descending order.

When we observe N sound sources, eigenvalues from v_{tf}^1 to v_{tf}^N have larger values corresponding to the power of each sound source; whereas the rest v_{tf}^{N+1} to v_{tf}^M have smaller values corresponding to the power of microphone measurement noises. The important feature of the

eigenvalues is that the noise eigenvectors $\mathbf{e}_{tf}^{N+1}, \ldots, \mathbf{e}_{tf}^{M}$ are orthogonal to the steering vector vectors that correspond to the directions of sound sources (Schmidt 1986).

(3) The MUSIC spectrum is calculated as:

$$P(t,d,f) = \frac{\left|\mathbf{q}_{fd}^{H}\mathbf{q}_{fd}\right|}{\sum_{m=N_{max}+1}^{M}\left|\mathbf{q}_{fd}^{H}\mathbf{e}_{tf}^{m}\right|}, \tag{7.3}$$

where $\mathbf{q}_{fd}$ is the M-dimensional steering vector for the dth direction and frequency bin f. These steering vectors are measured in advance as a calibration of the microphone array. When we assume the maximum number of sound sources N_{max}, the eigenvectors $\mathbf{e}_{tf}^{N_{max}+1}$ through $\mathbf{e}_{tf}^{M}$ are orthogonal to the steering vectors $\mathbf{a}_d(f)$ with the direction d where sound source exists. Thus, the denominator in Eq. (7.3) becomes close to zero at d; in other words, a salient peak of the MUSIC spectrum $P(t,d,f)$ is observed at d. However, in practice, the peaks in the MUSIC spectrum are smoothed partly because the reverberation in the environment virtually adds sound sources from all directions.

To account for a range of frequency bins, we integrate the MUSIC spectrum for each f as follows:

$$P'(t,d) = \sum_{f=F_{min}}^{F_{max}} \sqrt{v_{tf}^{1}}P(t,d,f), \tag{7.4}$$

where v_{tf}^{1} is the largest eigenvalue at frequency f, which roughly corresponds to the sound pressure level of the observed signal. To target at speech signals, we set the range of frequency bins such that F_{min} and F_{max} correspond to 500 and 2800 (Hz), respectively.

Basically, we can carry out a localization by detecting the direction d with $P'(t,d) > P_{thres}$, where P_{thres} is the threshold to determine whether a sound source is active. Since P_{thres} is dependent on N_{max} and the reverberation, this threshold is set empirically.

7.3 Bayesian Sound Source Localization And Tracking

This section presents our Bayesian extension of MUSIC-based sound source localization and tracking method. Our method consists of two steps: (1) off-line posterior estimation with the variational Bayesian hidden Markov model (VB-HMM), (2) on-line tracking of multiple

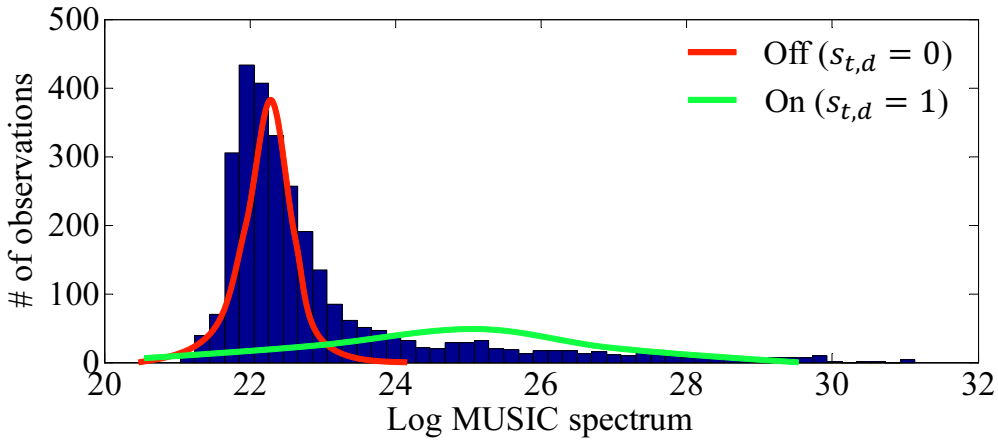


Figure 7.3: Blue: histogram of logarithmic MUSIC spectrum; Red: a Gaussian for non-active direction; Green: a Gaussian for active direction.

sound sources using a particle filter. The state vector of the HMM is a D-dimensional binary vector whose element indicates whether the sound source at direction d is active or not. The counterpart of P_{thres} is automatically obtained through the training of the VB-HMM.

For the observation model, we employ a Gaussian mixture model. We approximate the MUSIC spectrum by a Gaussian distribution partly because the spectrum is a sum over the frequency bins specified in Eq. (7.4) and partly because analytic computation is possible for this distribution. Figure 7.3 shows the histogram of MUSIC spectrum values on the logarithmic scale. As illustrated in Fig. 7.3, a cluster of non-active MUSIC spectrum is found in a lower area; on the other hand, active sources are likely to have higher spectrum values. Through the training of HMM, we obtain the posterior distribution of the parameters for Gaussian distributions in Figure 7.3.

We use a particle filter for efficient incremental localization and tracking. The reasons why we use a particle filter are: (1) The number of active sound sources in a state vector is easily capped at N_{max}. (2) Only local peaks of $P'(t,d)$ can be activated as a sound source using a proposal distribution. Further explanation is given in Section 7.3.2

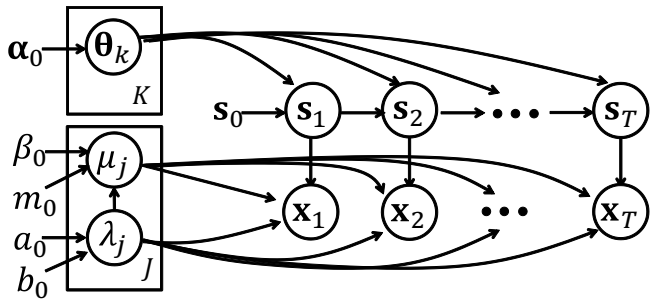


Figure 7.4: Graphical model for VB-HMM

7.3.1 Off-line parameter learning

We use a logarithmic MUSIC spectrum as an observation vector defined as:

$$x_{td} = 10 \log_{10} P'(t, d). \tag{7.5}$$

Let s_{td} be a binary variable. When $s_{td} = 1$, the sound source at direction d and time t is active.

Figure 7.4 shows the graphical model for the VB-HMM. The difference from the ordinary HMM is that the parameters for the state transition $\boldsymbol{\theta}_k$ and the observation $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ are probability variables instead of deterministic values. By taking account of many possibility of the parameters as probability variables, the training and the subsequent tracking produce better results than maximum likelihood-based HMM.

Observation model

The observation model is defined as:

$$p(\mathbf{x}_t | \mathbf{s}_t, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \prod_{d=1}^{D} \prod_{j=0}^{1} \mathcal{N}(x_{td} | \mu_j, \lambda_j)^{\delta_j(s_{td})}, \tag{7.6}$$

where $\delta_j(s_{td}) = 1$ iff $s_{td} = j$, and $\mathcal{N}(\cdot|\mu, \lambda)$ denotes the Gaussian distribution with the mean μ and precision λ. The vector notations such as $\mathbf{x}_t$ or $\mathbf{s}_t$ denote a set of variables for all directions, i.e., $\mathbf{x}_t = [x_{t1}, \ldots, x_{tD}]$. We use the Gaussian-gamma distribution for $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ which is the conjugate prior distribution of the Gaussian distribution:

$$p(\boldsymbol{\mu}, \boldsymbol{\lambda} | \beta_0, m_0, a_0, b_0) = \prod_{j=0}^{1} \mathcal{N}(\mu_j | m_0, \beta_0 \lambda_j) \mathcal{G}(\lambda_j | a_0, b_0), \tag{7.7}$$

where $\mathcal{G}(\cdot|a, b)$ denotes the gamma distribution with the shape a and rate b.

State transition model

To account for moving sound sources, the state transition model can be divided into 4 cases as summarized in Table 7.1. These 4 cases are the combination of previous states, that is, they depend on whether the previous state $s_{t-1,d}$ is active and whether the previous adjacent states are both inactive ($s_{t-1,d-1}s_{t-1,d+1} = 0$) or not. The state transition probability is defined as:

$$p(\mathbf{s}_t|\mathbf{s}_{t-1},\boldsymbol{\theta}) = \prod_{d=1}^{D}\prod_{k=1}^{4}\prod_{j=0}^{1}\left(\theta_k^{s_{td}}(1-\theta_k)^{1-s_{td}}\right)^{f_k(\mathbf{s}_{t-1},d)}, \tag{7.8}$$

where $f_k(\mathbf{s}_{t-1},d)$ is a classifier that returns 1 if when k matches the condition of the previous state values from $s_{t-1,d-1}$ to $s_{t-1,d+1}$ as specified in Table 7.1, and returns 0 otherwise. As the initial state, s_{0d} is set 0 for all d.

$\boldsymbol{\theta} = [\theta_1,\ldots,\theta_4]$ conforms to the beta distribution which is the conjugate prior distribution of Eq. (7.8).

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}_0) = \prod_{k=1}^{4}\mathcal{B}(\theta_k|\alpha_{0,1},\alpha_{0,0}), \tag{7.9}$$

where $\mathcal{B}(\cdot|c,d)$ is the beta distribution with parameters c and d.

Table 7.1: State transition probabilities with adjacent values.

previous state $s_{t-1,d}$	adjacent states $1 - s_{t-1,d-1}s_{t-1,d+1}$	transition probability $p(s_{td} = 1\|s_{t-1,d-1:d+1})$
0 (off)	0	θ_1
0 (off)	1	θ_2
1 (on)	0	θ_3
1 (on)	1	θ_4

Variational inference of posterior distribution

Here, the off-line training of the VB-HMM parameters means the estimation of posterior distribution $p(\mathbf{s}_{1:T},\theta,\mu,\lambda|\mathbf{x}_{1:T})$. We approximate this posterior by a factorized distribution:

$$p(\mathbf{s}_{1:T},\theta,\mu,\lambda|\mathbf{x}_{1:T}) \approx q(\mathbf{s}_{1:T},\theta,\mu,\lambda), \tag{7.10}$$

$$q(\mathbf{s}_{1:T},\theta,\mu,\lambda) = q(\mathbf{s}_{1:T})q(\theta)q(\mu,\lambda), \tag{7.11}$$

where $\cdot_{1:T}$ denotes a set of values with the time from 1 to T. (Beal 2003) explains the general inference algorithm in detail. The update equations are derived as follows. $q(\boldsymbol{\theta})$ is updated to the beta distribution with parameters $\hat{\alpha}_{k,0}$ and $\hat{\alpha}_{k,1}$ for each k, while $q(\mu,\lambda) = \prod_j q(\mu_j,\lambda_j)$ is updated to the Gaussian-gamma distribution with parameters $\hat{\beta}_j, \hat{m}_j, \hat{a}_j, \hat{b}_j$.

$$\hat{\alpha}_{k,j} = \alpha_{0,j} + \sum_{t,d} \langle s_{tdj} f_k(\mathbf{s}_{t-1}, d) \rangle, \tag{7.12}$$

$$\hat{\beta}_j = \beta_0 + n_j, \hat{m}_j = (\beta_0 m_0 + n_j \bar{x}_j)/(\beta_0 + n_j), \tag{7.13}$$

$$\hat{a}_j = a_0 + \frac{n_j}{2}, \hat{b}_j = b_0 + \frac{n_j S_j^2}{2} + \frac{\beta_0 n_j \left(\bar{x}_j - m_0\right)^2}{2(\beta_0 + n_j)}, \tag{7.14}$$

where s_{tdj} is equal to s_{td}, if $j = 1$ and $1 - s_{td}$, if $j = 0$. The quantities in Eqs. (7.13,7.14) are $n_j = \sum_{t,d} \langle s_{tdj} \rangle$, $\bar{x}_j = \frac{\sum_{t,d} \langle s_{tdj} \rangle x_{td}}{n_j}$, and $S_j^2 = \frac{\sum_{t,d} \langle s_{tdj} \rangle (x_{td} - \bar{x}_j)^2}{n_j}$. $\langle \cdot \rangle$ is the expectation over Eq. (7.11). $\langle s_{tdj} \rangle$ and $\langle s_{tdj} f_k(\mathbf{s}_{t-1}, d) \rangle$ are calculated as:

$$\langle s_{tdj} \rangle \propto \alpha(s_{tdj}) \beta(s_{tdj}), \tag{7.15}$$

$$\langle s_{tdj} f_k(\mathbf{s}_{t-1}, d) \rangle \propto \tilde{\alpha}(s_{t-1,d,k}) \tilde{p}(s_{t,d} | \mathbf{s}_{t-1}) \tilde{p}(x_{t,d} | s_{t,d}) \beta(s_{tdj}), \tag{7.16}$$

where $\alpha(s_{tdj})$ and $\beta(s_{tdj})$ are forward backward recursions.

$$\alpha(s_{tdj}) \propto \sum_{k=1}^{4} \tilde{\alpha}(s_{t-1,d,k}) \tilde{p}(s_{td} | \mathbf{s}_{t-1}) \tilde{p}(x_{td} | s_{td}), \tag{7.17}$$

$$\beta(s_{tdj}) = \sum_{j'=0}^{1} \beta(s_{t+1,d,j'}) \tilde{p}(s_{t+1,d,j'} | s_{tdj}) \tilde{p}(x_{td} | s_{td}). \tag{7.18}$$

The smoothed transition probability is

$$\tilde{p}(s_{td} = j | \mathbf{s}_{t,d-1:d+1}) \propto \prod_{k=1}^{4} \exp\left\{\psi(\hat{\alpha}_{k,j}) - \psi(\hat{\alpha}_{k,0} + \hat{\alpha}_{k,1})\right\}^{f_k(\mathbf{s}_{t-1},d)},$$

where $\psi(\cdot)$ denotes the digamma function. The smoothed observation probability is

$$\tilde{p}(x_{td} | s_{td}) \propto \prod_j \exp\left\{(\psi(\hat{a}_j) - \log \hat{b}_j - 1/\hat{\beta}_j)/2 - a_j (x_{t,d} - \hat{m}_j)^2 / 2\hat{b}_j\right\}^{s_{tdj}}$$

Eqs. (7.15, 7.16) are normalized such that the summation over j or j,k becomes 1. $\tilde{\alpha}(s_{t-1,d,k})$ is the probability regarding the condition k. Parameters are iteratively updated by Eqs. (7.12–7.16) until convergence. We start this iteration by setting $\langle s_{tdj} \rangle$ and $\langle s_{tdj} f_k(\mathbf{s}_{t-1}, d) \rangle$ as 1 or 0 with a threshold m_0 on the observation $x_{t,d}$.

7.3.2 Online localization and tracking using particle filter

This section explains an incremental tracking using a particle filter (Arulampalam et al. 2002) with the parameters obtained Eqs. (7.12–7.14). Here, the posterior distribution of the sound source activation vector given a sequence of MUSIC spectrum is approximated by P particles:

$$p(\mathbf{s}_t|\mathbf{x}_{1:t}) \approx \sum_{p=1}^{P} w_p \mathbf{s}_t^p, \tag{7.19}$$

where w_p and $\mathbf{s}_t^p$ are the weight and state vector of particle p, respectively. The number of particle is denoted by P. These w_p and $\mathbf{s}_t^p$ are obtained with two steps.

(1) Draw $\mathbf{s}_t^p$ from the proposal distribution:

$$\mathbf{s}_t^p \sim p'(\mathbf{s}_t|\mathbf{x}_t, m, a, b), \tag{7.20}$$

$$p'(\mathbf{s}_t^p|\mathbf{x}_t, \hat{m}, \hat{a}, \hat{b}) \propto \prod_d \prod_{j=0}^{1} C(x_{td})^{s^p_{td1}} \exp\left(-\Delta_{dj}^2/2\right)^{s^p_{tdj}}, \tag{7.21}$$

where $C(x_{td}) = 1$ if x_{td} is a local peak with respect to the direction d and $C(x_{td}) = 0$, otherwise. We define $C(x_{td})^{s^p_{td1}} = 1$ when $C(x_{td}) = 0$ and $s^p_{td1} = 0$. Therefore, $C(x_{td})^{s^p_{td1}}$ is equivalent to the exclusive or operation between $C(x_{td})$ and s^p_{td1}. The proposal weight is given by the Mahalanobis distance $\Delta_{dj}^2 = (x_{td} - \hat{m}_j)^2 \hat{a}_j / \hat{b}_j$.

(2) Calculate the weight w_p for each particle p as:

$$w_p \propto \frac{\bar{p}(\mathbf{x}_t|\mathbf{s}_t^p)\bar{p}(\mathbf{s}_t^p|\mathbf{s}_{t-1}^p)}{p'(\mathbf{s}_t^p|\mathbf{x}_t, \hat{m}, \hat{a}, \hat{b})}, \tag{7.22}$$

$$\bar{p}(\mathbf{x}_t|\mathbf{s}_t^p) = \prod_d C(x_{td})^{s^p_{td1}} \int p(\mathbf{x}_t|\mathbf{s}_t^p, \boldsymbol{\mu}, \boldsymbol{\lambda}) q(\boldsymbol{\mu}, \boldsymbol{\lambda}) d\boldsymbol{\mu} d\boldsymbol{\lambda}, \tag{7.23}$$

$$\bar{p}(\mathbf{s}_t^p|\mathbf{s}_{t-1}^p) = \int p(\mathbf{s}_t^p|\mathbf{s}_{t-1}^p, \boldsymbol{\theta}) q(\boldsymbol{\theta}). \tag{7.24}$$

The observation and state transition probabilities in Eqs. (7.23,7.24) are given by marginalizing out the parameters with the posterior distributions from those in HMM in Eqs. (7.6,7.8). These are analytically calculated as

$$\bar{p}(\mathbf{x}_t|\mathbf{s}_t^p) = \prod_d C(x_{td})^{s^p_{td1}} St(x_{td}|\hat{m}_j, \frac{\hat{\beta}_j \hat{a}_j}{(1+\hat{\beta}_j)\hat{b}_j}, 2\hat{a}_j)^{s^p_{tdj}}, \tag{7.25}$$

$$\bar{p}(\mathbf{s}_t^p|\mathbf{s}_{t-1}^p) = \prod_d \prod_k \left(\frac{\hat{\alpha}_{k,s_{td}}}{\hat{\alpha}_{k,0} + \hat{\alpha}_{k,1}}\right)^{f_k(\mathbf{s}^p_{t-1}, d)}, \tag{7.26}$$

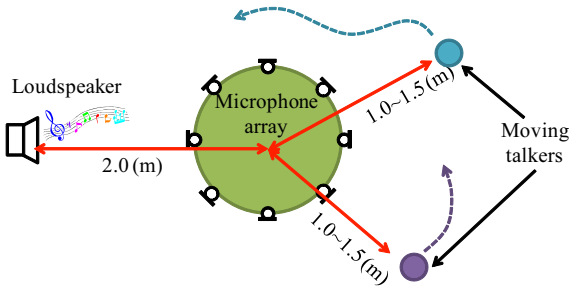


Figure 7.5: Moving talkers and a loudspeaker around the microphone array.

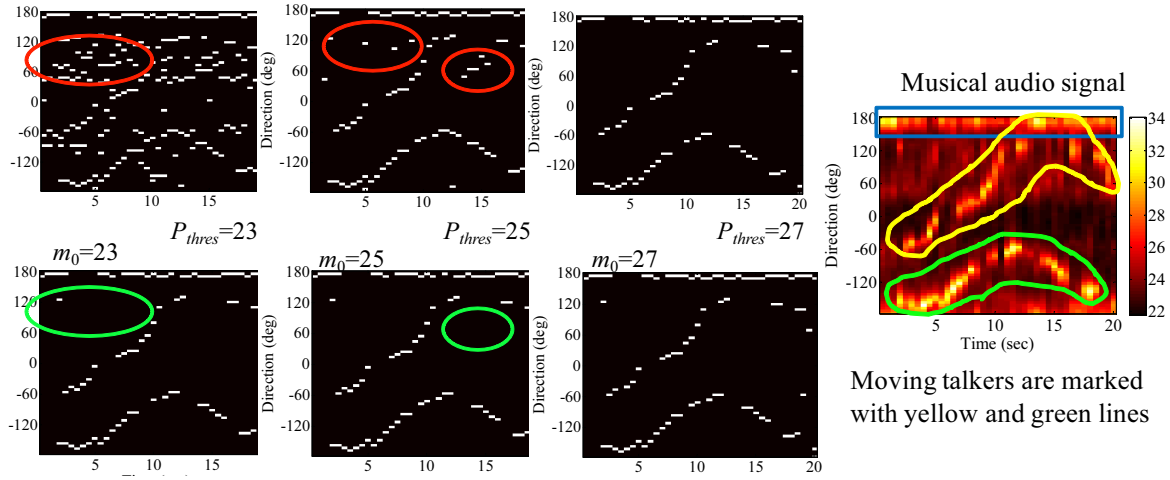


Figure 7.6: Plots of trajectories of sound sources. White plots are active audio sources. Top: static thresholding with P_{thres}. Bottom: Our method with m_0. Right: Observed logarithmic MUSIC spectrum. Musical audio signal is observed at close to 180 (deg).

where $\mathcal{S}t(\cdot|m,\lambda,\nu)$ denotes the Student's t-distribution with the mean m, precision λ, and the degree of freedom ν. To keep the number of active sources under N_{max}, the observation probability is set 0 if that of active sources in $\mathbf{s}_t^p$ exceeds N_{max}.

After calculating the weight of all particles, the weights w_p are normalized s.t. $\sum_{p=1}^{P} w_p = 1$. Then, the posterior is obtained as Eq. (7.19). In our implementation, the particles are re-sampled for each time t in proportion to the weight of each particle.

7.4 Experimental Results

This section presents the experimental results. Our method is compared to the fixed threshold approach. The experimental setup is shown in Figure 7.5. For the off-line learning of VB-HMM, only one talker moves around the microphone array while talking. During the on-line tracking using the particle filter, two talkers move around while talking and a musical audio signal is played from the loudspeaker. Both signals are 20 seconds in length. The parameters are set as follows: $N_{max} = 3$, $\alpha_0 = [1, 1]$, $\beta_0 = 1$, $a_0 = 1$, $b_0 = 500$. The number of particles is set as $P = 500$. The reverberation time of the experiment chamber is $RT_{20} = 840$ (msec).

Figure 7.6 shows the results with threshold $P_{thres} = 23, 25, 27$ and $m_0 = 23, 25, 27$. Particle filtering results show the trajectories where the posterior probability exceeds 0.95. The fixed threshold approach with a low threshold produces enormous false detections, as red circles show in Figure 7.6. On the other hand, our method produces much more stable results as green circles show. We also confirmed that the resulting trajectories are almost the same as long as the threshold of the posterior is in 0.95–1.0. These results confirm that our method automatically converges to good parameters for the sound source localization and tracking. Furthermore, our method stably tracks multiple sound sources even if only one sound source is used in the training phase.

7.5 Summary

This paper presented a Bayesian extension of MUSIC-based sound source localization. Our method consists of (1) an automatic parameter learning in the VB-HMM framework, and (2) an incremental localization and tracking using a particle filter. Future work includes the integration with the robot audition system HARK (Nakadai et al. 2010), and the application to auditory scene analysis with a mobile robot.

Chapter 8

Discussion

This chapter summarizes the observations on the developed multichannel methods so as to clarify the contributions of this work. Then, some open problems are discussed as future directions of the research.

8.1 Observations

In order to realize various auditory processing functions in multisource environments, we have developed multichannel methods that overcome the following auditory uncertainties:

1. uncertainty of the number of sources in the environment,
2. degradation of sound source separation performance by reverberation, and
3. dynamic acoustic environments such as moving sound sources.

Chapter 3 tackled the source number uncertainty in the multichannel sound source separation and localization problem. The method is designed on the basis of the TF masking approach, which is consistently applicable to both overdetermined and underdetermined mixtures. We incorporated Bayesian nonparametrics so that the model is capable of handling an arbitrary number of sound sources. The experimental results confirmed the superiority of our method to existing source separation methods in terms of the separation performance as well as its applicability to various number of sound sources regardless of the number of microphones. On the other hand, the experiments also revealed that the reverberation degrades the separation and localization performance.

Chapter 4 addressed the second reverberation issue. The dereverberation function is also related to the source number uncertainty: existing dereverberation methods assume that the number of sources is known in advance and that the number of microphones is larger than that of sources. A joint model for sound source separation and dereverberation was developed on the basis of Bayesian nonparametrics to cope with the source number uncertainty. We used the infinite mixture of AR processes to model the reverberation of the arbitrary number of sound sources. The experimental results confirmed the efficacy of our method for both underdetermined and overdetermined conditions.

In Chapter 5, a Bayesian nonparametric model to adaptively determine the AR order value for the dereverberation was developed. This model is motivated by the fact that the AR order value should be set depending on the amount of the reverberation in the environment. Our experiments confirmed that the AR order values were automatically adjusted in accordance with the reverberation time of the observation.

In Chapter 6, the Bayesian nonparametric sound source separation and localization method presented in Chapter 3 was applied to the sound source separation in a dynamic environment. In our setup, multiple sound sources are observed by a moving microphone array embedded on a mobile robot. The separation was carried out where the sound sources were assumed to be located in disjoint direction ranges. The results confirm the separation capability in an outdoor environment with low reverberation. On the other hand, the experimental result in an indoor environment showed that the separation quality was degraded compared to the outdoor results. This is considered because the reverberation in the indoor environment affected the separation performance.

Chapter 7 also tackled the issue of dynamic environment by developing a robust sound source localization method. A Bayesian HMM model was used to automatically choose a threshold of the existence of sound sources on MUSIC spectra. In the experiment, we confirmed the efficacy of this Bayesian method in that our method bypasses an elaborate tuning of the threshold of MUSIC spectra in a reverberant environment.

8.2 Contribution

This dissertation presented multichannel methods to realize fundamental auditory processing functions for CASA and machine listening. Bayesian nonparametric modeling played the key role in the development of these methods to overcome the auditory uncertainties. The contributions of this work are twofold: (1) the theoretical aspect and (2) practical aspect. The theoretical contribution is the development of a generic computational model for microphone array processing. The model is designed based on a Bayesian nonparametric perspective. The practical contribution of this work is the wide applicability of the model for an arbitrary number of sound sources or in dynamic environments.

The theoretical side can be regarded as a contribution for the signal processing community in that this work develops a model for microphone array processing in general situations. Our model provides the fundamental three functions implemented by microphone array processing: 1) sound source separation, 2) sound source localization, and 3) dereverberation. These three functions are unified into a single generative process and the corresponding inference procedure. The model is consistently applicable without any modification depending on the number of sound sources in the environment.

As the practical side of the contribution, the generic models with robustness against the auditory uncertainties in real environments are the contribution for CASA or robot audition systems. For example, the audio intelligibility of the operator of telepresence robot (Mizumoto et al. 2011) by presenting the location and separated signals of surrounding sound sources. This technology may also constitute the auditory functions for probing robots (Sasaki et al. 2009a). Our models are robustly usable in uncertain or dynamic situations without a elaborate tuning of the model depending on specific acoustic environments.

8.3 Open Problems

This section remarks several open problems regarding the methods developed in this dissertation as future research directions. These open problems include

1. fast and online computation for real time processing,
2. an automatic acquisition of microphone positions,

3. improvement in intelligibility of separated audio signals,
4. more general movements of sound sources in dynamic environments, and
5. high-level auditory processing after separating multisource observations.

While some applications such as spoken dialogue systems (McTear 2004) in a noisy environment may require a fast computation for achieving quick responses, our method currently has difficulty to output the result in real time. One of the reasons for the considerable computation time is the use of MCMC method, or Gibbs sampling, for the parameter inference because the parallelization of the computation is difficult for Gibbs sampling. One way to accelerate the parameter inference is to use variational inference methods (Attias 2000) since variational methods in practice require less iterations before the convergence of the parameters and are readily parallelized. Another advantage of the variational inference is online inference frameworks (Hoffman et al. 2010b, Wang et al. 2011). These inference schemes may realize a fast and online processing of our multichannel methods.

The only prior knowledge that our method requires is the positions of microphones in the microphone array. While our methods use steering vectors obtained through the calibration of the microphone array in use, this information may be automatically acquired by blind alignment approaches (Raykar et al. 2005, Ono et al. 2009, Miura et al. 2011, Miyabe et al. 2013). These methods may facilitate the calibration process of the microphone array.

The TF masking-based separation often produces musical noise in the resultant audio signals, especially when the observation contains a large number of sound sources or intensive reverberation. This musical noise impairs the intelligibility of the separated signals. The use of musical noise reduction as a post-processing can improve the degradation of the intelligibility (Araki et al. 2005, Esch and Vary 2009). The imperfect separation partly results from the assumption that at most one sound source is dominant at each time-frequency point. While this assumption justifies the TF masking approach to separate the sound sources, this assumption may be violated when a large number of sound sources are simultaneously active. To address this problem, the use of factorial models such as (Hoffman et al. 2010a, Nagira et al. 2013) may be helpful because these models assumes several source signals can coexist in a time-frequency point. The trade-off should be kept in mind between the complexity of the observation process and simplicity of the parameter inference.

The separation of the moving sources in a dynamic environment has been halfway developed. In this dissertation, each moving source are assumed to fall in a disjoint direction range so that the separated segments can easily be merged to reconstruct the sound source. To cope with general movements of sound sources, we need to manage the tracking of sound sources and identification of separated segments. For a robust tracking, a multimodal approach may be helpful (Naqvi et al. 2010, Nickel et al. 2005).

In order to enable computers and robots to manage complex tasks on the basis of auditory information acquired in real environments, high-level auditory processing, such as speech recognition, prosody or emotion recognition, sound source identification and so on, is essential after the separation of multisource observations. These auditory pattern recognition tasks often use Mel-frequency cepstrum coefficients (MFCC) as a feature representation. Though MFCCs are widely used and proved effective for many pattern recognition problems in the literature (Li et al. 2001, Aucouturier et al. 2007), the weakness of MFCC feature is its vulnerability to the distortion and interference of other audio signals; namely, an MFCC-based decoding, such as classification, is severely affected when the source separation result is imperfect. Since a perfect source separation is difficult in the face of various auditory uncertainties, these techniques should be considered to manage the high-level auditory processing and complex tasks. To alleviate this problem, some speech recognition and audio patter recognition techniques models the uncertainty resulting from the speech enhancement in the extracted MFCC vectors (Cooke et al. 2001, Deng et al. 2005, Vincent 2012).

An alternative way for handling the uncertainty in MFCC vectors to achieve a robust auditory patter recognition is to use deep neural networks. Deep neural networks have recently gained much interest of researchers due to their high performance in various patter recognition tasks (Bengio 2009, Hinton et al. 2012, Dahl et al. 2012). Deep neural networks automatically construct feature representations suitable for the given task through the training of the deep architecture. When separated audio signals are provided as a training dataset, these approaches may be able to manage the distortions and remaining interferences in the separated signals for the high-level audio processing.

Chapter 9

Conclusion

This dissertation presented multichannel methods so as to achieve the decomposition function essential to deal with multisource environments. These methods provided all or a part of these three functions: sound source separation, localization and dereverberation. Through this dissertation, three auditory uncertainties were addressed, which are commonly observed in actual environments: the uncertainty in the number of sound sources, reverberation, and dynamic environments.

These auditory uncertainties were overcome by using Bayesian nonprametric models. In particular, the selection of model complexity caused by the source number uncertainty is bypassed by the infinite flexibility of the Bayesian nonparametric model. The contributions of this work have two aspects. For the theoretical side, the three functions of microphone array processing is unified in to a single model on the basis of a Bayesian nonparametric perspective. For the practical side, the developed method is widely usable in the three auditory uncertainties without environment-specific tunings. The Bayesian nonparametrics-based framework constitutes a fundamental building block of CASA systems and robot audition architectures that work in our daily environments.

The decomposition of sound sources with microphone arrays is carried out on the basis of direction of arrivals of sound sources. This is because multichannel methods are formulated to handle the subspace structures in the time-frequency domain, as depicted in Figure 3.2. These subspaces corresponds to distinct directions of respective sound sources. This granularity of decomposition may be insufficient for some applications. For example, a robot wants to listen to a narrative coming from a TV set, though the sound contains a background music as well.

In this case, a straightforward application of multichannel methods is difficult to separate the human voice and music audio signal since the targeted voice signal and music signal arrive from the same direction. A possible approach to mitigate this limitation is to use a hybrid method of multichannel and single channel frameworks (Ozerov and Fevotte 2010, Sawada et al. 2013). For achieving hybrid approaches, a trade-off may occur between the granularity of decomposition function and scalability of the type of sound sources. A key question is what is an elemental representation of sounds–a TF mask on the spectrogram, a harmonic structure with a fundamental frequency, or base segment waveforms such as mother wavelets? How to implement an elemental representation of sounds can arise as a research question.

Another interesting future direction for CASA systems is an incorporation of human intelligence for auditory information processing (Quinn and Bederson 2011, Bryan and Mysore 2013). Two major ways are considered to incorporate human intelligence into computational algorithms. The first way is to build computational models that reflects the knowledge as to how auditory processing is carried out in a human brain (Bregman 1994, Blauert 1997). Understanding how humans analyzes auditory events may contribute to the realization of richer auditory information processing by computers. The second approach is to treat humans as a system component where the way how humans internally perceive audio information is regarded as a black box (Pareek and Ravikumar 2013). A successful example is an interactive sound source separation system (Bryan and Mysore 2013) where the sound separation algorithm is aided by human annotations. Here, the annotations are used as a regularization term in the evaluation function of the separation method. Development of interfaces that encourage humans to precisely respond to the CASA system, and algorithms that incorporate the response of humans in a natural way into a mathematical form is future work for this direction.

Bibliography

D. R. Griffin, *Listening in the Dark*. New York: Yale University Press, 1958.

W. E. Evans, "Echolocation by marine delphinids and one species of fresh-water dolphin," *Journal of the Acoustical Society of America*, vol. 54, no. 1, pp. 191–199, 1973.

W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 34, no. 3, pp. 334–352, 2004.

S. Thrun, "Probabilistic robotics," *Communications of the ACM*, vol. 45, no. 3, pp. 52–57, 2002.

S. Thrun, "Toward robotic cars," *Communications of the ACM*, vol. 53, no. 4, pp. 99–106, 2010.

C. R. Weisbin and G. Rodriguez, "NASA robotics research for planetary surface exploration," *IEEE Robotics & Automation Magazine*, vol. 7, no. 4, pp. 25–34, 2000.

K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

D. F. Rosenthal and H. G. Okuno, *Computational Auditory Scene Analysis*. New Jersey: Lawrence Erlbaum, 1998.

D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. New York: Wiley-IEEE Press, 2006.

E. Vincent, J. Barker, S. Watanabe, J. L. Roux, F. Nesta, and M. Matassoni, Eds., *The 2nd International Workshop on Machine Listening in Multisource Environments*, 2013.

E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *Journal of the acoustical society of America*, vol. 25, no. 8, pp. 975–979, 1953.

A. W. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions," *Acta Acustica united with Acustica*, vol. 86, no. 1, pp. 117–128, 2000.

D. Wood, "The physiological basis of selective attention: Implications of event-related potential studies," in *Event-related brain potentials: Basic issues and applications*, J. W. Rohrbaugh, R. Parasuraman, and J. R. Johnson, Eds. Oxford University Press, 1990, pp. 178–209.

Y. Sasaki, S. Kagami, and H. Mizoguchi, "Online short-term multiple sound source mapping for a mobile robot by robust motion triangulation," *Advanced Robotics*, vol. 23, no. 1–2, pp. 145–164, 2009.

T. Mizumoto, T. Yoshida, K. Nakadai, R. Takeda, T. Otsuka, T. Takahashi, and H. G. Okuno, "Design and implementation of selectable sound separation on a Texai telepresence system using HARK," in *Proc. of IEEE/RAS International Conference on Robotics and Automation*, 2011, pp. 2130–2137.

J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*, ser. Springer Topics in Signal Processing. Springer, 2008.

C. M. Bishop, "Bayesian regression and classification," in *Advances in Learning Theory: Methods, Models and Applications*, 2003, pp. 267–285.

C. Fraley and A. E. Raftery, "Bayesian regularization for normal mixture estimation and model-based clustering," *Journal of Classification*, vol. 24, no. 2, pp. 155–181, 2007.

T. S. Ferguson, "A bayesian analysis of some nonparametric problems," *Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

C. E. Antoniak, "Mixtures of Dirichlet processes with applications to bayesian nonparametric estimation," *Annals of Statistics*, vol. 2, no. 6, pp. 1152–1174, 1974.

S. J. Gershman and D. M. Blei, "A tutorial on Bayesian nonparametric models," *Journal of Mathematical Psychology*, vol. 56, no. 1, pp. 1–12, 2012.

N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. MIT Press, 1949.

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

R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

S. Roweis, "One microphone source separation," *Advances in Neural Information Processing Systems*, vol. 13, pp. 793–799, 2000.

M. E. Davies and C. J. James, "Source separation using single channel ICA," *IEEE Trans. on Speech Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

G.-J. Jang and T.-W. Lee, "A maximum likelihood approach to single-channel source separation," *Journal of Machine Learning Research*, vol. 4, pp. 1365–1392, 2003.

L. Benaroya and F. Bimbot, "Wiener based source separation with HMM/GMM using a single sensor," in *Proc. of International Symposium on Independent Component Analysis and Blind Signal Separation*, 2003, pp. 957–961.

B. A. Pearlmutter and A. M. Zador, "Monaural source separation using spectral cues," in *Proc. of International Conference on Independent Component Analysis*, 2004, pp. 478–485.

P. Smaragdis, M. Shashanka, and B. Raj, "A sparse non-parametric approach for single channel separation of known sounds," in *Neural Information Processing Systems*, 2009.

M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. of Interspeech*, 2006, pp. 957–961.

T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1564–1578, 2007.

M. D. Hoffman, D. M. Blei, and P. R. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. of the International Conference on Machine Learning*, 2010.

H. Kameoka, M. Nakano, K. Ochiai, Y. Imoto, K. Kashino, and S. Sagayama, "Constrained and regularized variants of non-negative matrix factorization incorporating music-specific constraints," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 5365–5368.

Y. Sasaki, S. Thompson, M. Kaneyoshi, and S. Kagami, "Map-generation and identification of multiple sound sources from robot in motion," in *Proc. of International Conference on Intelligent Robots and Systems*, 2010, pp. 437–443.

T.-W. Lee, M. Girolami, and T. J. Sejnowski, "Independent component analysis using an extended infomax algorithm for mixed sub-Gaussian and super-Gaussian sources," *Neural Computation*, vol. 11, pp. 417–441, 1999.

A. Hyvärinen, J. Karhunen, and E. Oja, Eds., *Independent Component Analysis*. Wiley, 2001.

P. Common and C. Jutten, Eds., *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 2010.

H. Sawada, S. Araki, and S. Makino, "MLSP 2007 data analysis competition: Frequency-domain blind source separation for convolutive mixtures of speech/audio signals," in *IEEE International Workshop on Machine Learning for Signal Processing*, 2007, pp. 45–50.

I. Lee, T. Kim, and T.-W. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Processing*, vol. 87, no. 8, pp. 1859–1871, 2007.

N. Ono, "Stable and Fast Update Rules for Independent Vector Analysis Based on Auxiliary Function Technique," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011, pp. 189–192.

D. Knowles and Z. Ghahramani, "Infinite sparse factor analysis and infinite independent components analysis," in *Proc. of International Conference on Independent Component Analysis and Signal Separation*, 2007, pp. 381–388.

K. Nagira, T. Otsuka, and H. G. Okuno, "Nonparametric Bayesian sparse factor analysis for frequency domain blind source separation without permutation ambiguity," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 4, 2013.

O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation maximization source separation and localization," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 2, 2010.

S. Araki, T. Nakatani, and H. Sawada, "Simultaneous clustering of mixing and spectral model parameters for blind sparse source separation," in *Proc. of IEEE International Conference on Acoustics Speech and Signal Processing*, 2010, pp. 5–8.

H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

S. Araki, T. Nakatani, H. Sawada, and S. Makino, "Stereo source separation and source counting with MAP estimation with Dirichlet prior considering spatial aliasing problem," in *Proc. of the 8th International Conference on Independent Component Analysis and Signal Separation*, 2009, pp. 742–750.

B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *Proc. of Latent Variable Analysis and Signal Separation*, 2010.

J. Taghia, N. Mohammadia, and A. Leijon, "A variational Bayes approach to the underdetermined blind source separation with automatic determination of the number of sources," in *Proc. of International Conference on Acoustics Speech and Signal Processing*, 2012, pp. 253–256.

K. Yamamoto, F. Asano, W. F. G. van Rooijen, E. Y. L. Ling, T. Yamada, and N. Kitawaki, "Estimation of the number of sound sources using support vector machines and its application to sound source separation," in *Proc. of International Conference on Acoustics Speech and Signal Processing*, 2003, pp. V–485.

B. Loesch and B. Yang, "Source number estimation and clustering for underdetermined blind source separation," in *International Workshop on Acoustic Echo and Noise Control*, 2008.

K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica United with Acustica*, vol. 87, pp. 359–366, 2001.

E. A. P. Habets, S. Gannot, I. Cohen, and P. C. W. Sommen, "Joint dereverberation and residual echo suppression of speech signals in noisy environments," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 8, pp. 1433–1451, 2008.

H. W. Löllmann and P. Vary, "Low delay noise reduction and dereverberation for hearing aids," *EURASIP Journal Advances in Signall Processing*, vol. 2009, 2009.

L. Wang, K. Odani, and A. Kai, "Dereverberation and denoising based on generalized spectral subtraction by multi-channel LMS algorithm using a small-scale microphone array," *EURASIP Journal Advances in Signall Processing*, vol. 2012, 2012.

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. on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1746–1765, 2010.

M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 36, no. 2, pp. 145–152, 1988.

M. Triki and D. T. M. Slock, "Delay and predict equalization for blind speech dereverberation," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, 2006, pp. V–97–V–100.

M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 430–440, 2007.

T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear predictor," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

K. Kinoshita, M. Delcroix, T. Nakatani, and M. Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 534–545, 2009.

T. Yoshioka and T. Nakatani, "Generalization of multi-channel linear prediction methods for blind MIMO impulse response shortening," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 882–895, 2005.

H. Buchner and W. Kellermann, "TRINICON for dereverberation of speech and audio signals," in *Speech Dereverberation*, ser. Signals and Communication Technology, P. A. Naylor and N. D. Gaubitch, Eds. Springer London, 2010, pp. 311–385.

T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 19, no. 1, pp. 69–84, 2011.

R. Takeda, K. Nakadai, T. Takahashi, T. Ogata, and H. G. Okuno, "Efficient blind dereverberation and echo cancellation based on independent component analysis for actual acoustic signals," *Neural Computation*, vol. 24, no. 1, pp. 234–272, 2012.

M. Togami, Y. Kawaguchi, R. Takeda, Y. obuchi, and N. Nukaga, "Optimized speech dereverberation from probabilistic perspective for time varying acoustic transfer function," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 21, no. 7, pp. 1369–1380, 2013.

A. Gorokhov and P. Loubaton, "Subspace-based techniques for blind separation of convolutive mixtures with temporally correlated sources," *IEEE Trans. on Circuits and Systems I*, vol. 44, no. 9, 1997.

S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *Computer Speech and Language*, vol. 9, no. 2, pp. 153–169, 1995.

M. S. Brandstein, J. E. Adcock, and H. F. Silverman, "A practical time-delay estimator for localizing speech sources with a microphone array," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 5, pp. 249–252, 1997.

J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003, pp. 1228–1233.

M. Murase, S. Yamamoto, J.-M. Valin, K. Nakadai, K. Yamada, K. Komatani, T. Ogata, and H. G. Okuno, "Multiple moving speaker tracking by microphone array on mobile robot," in *Proc. of INTERSPEECH*, 2005, pp. II–1133–II–1136.

W.-K. Ma, B.-N. Vo, S. S. Singh, and A. Baddeley, "Tracking an unknown time-varying number of speakers using TDOA measurements: a random finite set approach," *IEEE Trans. on Signal Processing*, vol. 54, no. 9, pp. 3291–3304, 2006.

A. Koutvas, E. Dermatas, and G. Kokkinakis, "Blind speech separation of moving speakers in real reverberant environments," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, 2000, pp. II–1133–II–1136.

R. E. Prieto and P. Jinachitra, "Blind source separation for time-variant mixing systems using piecewise linear approximations," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, 2005, pp. V–301–V–304.

W. Addison and S. Roberts, “Blind source separation with non-stationary mixing using wavelet,” in *Proc. of International Conference on independent component analysis and Blind Source Separation*, 2006.

S. M. Naqvi, M. Yu, and J. A. Chambers, “A multimodal approach to blind source separation of moving sources,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 895–910, 2010.

M. I. Mandel, D. P. W. Ellis, and T. Jebara, “An EM Algorithm for Localizing Multiple Sound Sources in Reverberant Environments,” *Advances in Neural Information Processing Systems*, vol. 19, pp. 953–960, 2007.

K. Nakadai, T. Takahashi, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, “Design and Implementation of Robot Audition System “HARK”,” *Advanced Robotics*, vol. 24, no. 5–6, pp. 739–761, 2010.

T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, “Bayesian unification of sound source localization and separation with permutation resolution,” in *Proc. of AAAI Conference on Artificial Intelligence*, 2012, pp. 2038–2045.

P. Bofill and M. Zibulevsky, “Underdetermined Blind Source Separation Using Sparse Representations,” *Signal Processing*, vol. 81, no. 11, pp. 2353–2362, 2001.

M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, “A Survey of Convolutive Blind Source Separation Methods,” in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. Springer Press, 2007.

H. Sawada, R. Mukai, S. Araki, and S. Makino, “A robust and precise method for solving the permutation problem of frequency-domain blind source separation,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 12, no. 5, pp. 530–538, 2004.

T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, “Speech dereverberation in short time Fourier transform domain with crossband effect compensation,” in *Proc. of Hands-Free Speech Communication and Microphone Arrays*, 2008, pp. 220–223.

S. Winter, H. Sawada, and S. Makino, "Geometrical interpretation of the PCA subspace approach for overdetermined blind source separation," *EURASIP Journal on Advances in Signal Processing*, vol. 2006, pp. 1–11, 2006, article ID 71632.

N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.

S. Vaseghi and H. Jetelová, "Principal and independent component analysis in image processing," in *Proc. of the 14th International Conference on Mobile Computing and Networking*, 2006, pp. 1–5.

N. Kovacevic and A. R. McIntosh, "Groupwise independent component decomposition of EEG data and partial least square analysis," *Neuroimage*, vol. 35, no. 3, pp. 1103–1112, 2007.

K. Nagira, T. Takahashi, T. Ogata, and H. G. Okuno, "Complex extension of infinite sparse factor analysis for blind speech separation," in *Proc. of International Conference on Latent Variable Analysis and Signal Separation*, 2012, pp. 388–396.

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

H. Kameoka, M. Sato, T. Ono, N. Ono, and S. Sagayama, "Blind separation of infinitely many sparse sources," in *Proc. of International Workshop on Acoustic Signal Enhancement*, 2012, pp. 1–4.

D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

A. van den Bos, "The multivariate complex normal distribution–a generalization," *IEEE Trans. on Information Theory*, vol. 41, no. 2, pp. 537–539, 1995.

K. Conradsen, A. A. Nielsen, J. Schou, and H. Skiriver, "A test statistic in the complex wishart distribution and its application to change detection in polarimetric SAR data," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 41, no. 1, pp. 4–19, 2003.

M. Escobar and M. West, “Bayesian density estimation and inference using mixtures,” *Journal of the American Statistical Association*, vol. 90, pp. 577–588, 1995.

E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

M. Matsumoto, M. Tohyama, and H. Yanagawa, “A method of interpolating binaural impulse responses for moving sound images,” *Acoustical Science and Technology*, vol. 24, no. 5, pp. 284–292, 2003.

K. Nakamura, K. Nakadai, and H. G. Okuno, “A real-time super-resolution robot audition system that improves the robustness of simultaneous speech recognition,” *Ad*, vol. 27, no. 12, pp. 933–945, 2013.

D. MacKay, “Ensemble Learning for Hidden Markov Models,” Department of Physics, Cambridge University, Tech. Rep., 1997.

R. Fujimaki and S. Morinaga, “Factorized asymptotic bayesian inferencce for mixture modeling,” in *Proc. Artificial Intelligence and Statistics*, 2012.

T. Otsuka, K. Ishiguro, H. Sawada, and H. G. Okuno, “Bayesian nonparametrics for microphone array processing,” *IEEE/ACM Trans. on Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 493–504, 2013.

E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, “Bayesian nonparametric inference of switching dynamic linear models,” *IEEE Trans. on Signal Processing*, vol. 59, no. 4, pp. 1569–1585, 2011.

D. Aldous, “Exchageability and related topics,” *École d’Été de Probabilités de Saint-Flour XIIIŮ1983*, pp. 1–198, 1985.

R. M. Neal, “Markov chain sampling methods for Dirichlet process mixture models,” *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 249–265, 2000.

M. Togami, Y. Kawaguchi, R. Takeda, Y. Obuchi, and N. Nukaga, "Multichannel speech dereverberation and separation with optimized combination of linear and nonlinear filtering," in *Proc. of IEEE International Conference on Acoustics Speech and Signal Processing*, 2012, pp. 4057–4060.

T. Park and D. A. van Dyk, "Partially collapsed Gibbs samplers: Illustrations and applications," *Journal of Computational and Graphical Statistics*, vol. 18, pp. 283–305, 2009.

G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 4th ed. Wiley, 2008.

L. Harrison, W. D. Penny, and K. Friston, "Multivariate autoregressive modeling of fMRI time series," *NeuroImage*, vol. 19, pp. 1477–1491, 2003.

T. N. T. Yoshioka and M. Miyoshi, "Integrated speech enhancement method using noise suppression and dereverberation," *IEEE Trans. ASLP*, vol. 17, no. 2, pp. 231–246, 2009.

D. A. Dickey and W. A. Fuller, "Likelihood ratio statistics for autoregressive time series with a unit root," *Econometrica*, vol. 49, no. 4, pp. 1057–1072, 1981.

M. A. D. Lucca, A. Guglielmi, P. Müller, and F. A. Quintana, "A simple class of Bayesian nonparametric autoregression models," *Bayesian Analysis*, vol. 8, no. 1, pp. 63–88, 2013.

T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Blind speech dereverberation with multi-channel linear prediction based on short time Fourier transform representation," in *Proc. ICASSP*, 2008, pp. 85–88.

E. J. Hannan, "The estimation of the order of an ARMA process," *Annals of Statistics*, vol. 8, no. 5, pp. 1071–1081, 1980.

C.-K. Ing, C.-Y. Sin, and S.-H. Yu, "Model selection for integrated autoregressive processes of infinite order," *Journal of Multivariate Analysis*, vol. 106, pp. 57–71, 2012.

H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

W. D. Penny and S. J. Roberts, "Bayesian multivariate autoregressive models with structured priors," *IEE Proceedings Vision, Image and Signal Processing*, vol. 149, no. 1, pp. 33–41, 2002.

S. J. Roberts and W. D. Penny, "Variational Bayes for generalized autoregressive models," *IEEE Trans. Signal Processing*, vol. 50, no. 9, pp. 2245–2257, 2002.

S. J. Godsill, "On the relationship between Markov chain Monte Carlo methods for model uncertainty," *Journal of Computational and Graphical Statistics*, vol. 10, no. 2, pp. 230–248, 2001.

J. Vermaak, C. Andrieu, A. Doucet, and S. J. Godsill, "Reversible jump Markov chain Monte Carlo strategies for Bayesian model selection in autoregressive processes," *Journal of Time Series Analysis*, vol. 25, no. 6, pp. 785–809, 2004.

P. J. Green, "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, pp. 711–732, 1995.

D. Mochihashi and E. Sumita, "The infinite Markov model," in *Proc. NIPS*, 2008, pp. 1017–1024.

R. M. Neal, "Slice Sampling," *Annals of Statistics*, vol. 31, no. 3, pp. 705–767, 2003.

A. G. Wilson and Z. Ghahramani, "Generalised Wishart processes," in *Proc. UAI*, 2011.

T. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Indian buffet process," in *Proc. NIPS*, 2006, pp. 475–482.

S. Thrun, M. Montemerlo, D. Koller, B. Wegbreit, J. Nieto, and E. Nebot, "FastSLAM: An Efficient Solution to the Simultaneous Localization and Mapping Problem with Unknown Data Association," *Journal of Machine Learning Research*, 2004.

S. Se, D. G. Lowe, and J. J. Little, "Vision-Based Global Localization and Mapping for Mobile Robots," *IEEE Trans. on Robotics*, vol. 21, no. 3, pp. 364–375, 2005.

Y. Sasaki, M. Kaneyoshi, S. Kagami, H. Mizoguchi, and T. Enomoto, "Daily sound recognition using pitch-cluster-maps for mobile robot audition," in *Proc. of International Conference on Intelligent Robots and Systems*, 2009, pp. 2724–2729.

H. Sawada, J. Even, H. Saruwatari, K. Shikano, and T. Takatani, "Improvement of Speech Recognition Performance for Spoken-Oriented Robot Dialog System using End-fire Array," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 970–975.

G. Ince, K. Nakamura, F. Asano, H. Nakajima, and K. Nakadai, "Assessment of General Applicability of Ego Noise Estimation," in *Proc. of International Conference on Robotics and Automation*, 2011, pp. 3517–3522.

Y. Kubota, M. Yoshida, K. Komatani, T. Ogata, and H. G. Okuno, "Design and Implementation of 3D Auditory Scene Visualizer towards Auditory Awareness with Face Tracking," in *Proc. of IEEE International Symboposium on Multimedia (ISM-2008)*, 2008, pp. 468–476.

S. Doclo and M. Moonen, *Microphone arrays*. Springer, 2001, ch. GSVD-based optimal filtering for multi-microphone speech enhancement, pp. 111–132.

R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

F. Asano, M. Goto, K. Itou, and H. Asoh, "Real-time Sound Source Localization and Separation System and Its Application to Automatic Speech Recognition," in *Proc. of Eurospeech2001*, 2001, pp. 1013–1016.

P. Danès and J. Bonnal, "Information-theoretic detection of broadband sources in a coherent beamspace MUSIC scheme," in *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2010)*, 2010, pp. 1976–1981.

K. Yamamoto, F. Asano, T. Yamada, and N. Kitawaki, "Detection of overlapping speech in meeting using support vector machines and support vector regression," *IEICE Trans. Fundamentals*, vol. E89-A, no. 8, pp. 2158–2165, 2006.

M. J. Beal, "Variational Algorithms for Approximate Bayesian Inference," Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University Colledge London, 2003.

M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A Tutorial on Particle Filters for Online Nonlinear/Non-Gaussian Bayesian Tracking," *IEEE Transactions on Signal Proc.*, vol. 50, no. 2, pp. 174–189, 2002.

M. McTear, *Spoken Dialogue Technology*. London: Springer Verlag, 2004.

H. Attias, "A variational Bayesian framework for graphical models," in *Advances in Neural Information Processing Systems 12*, 2000, pp. 209–215.

M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent Dirichlet allocation," *Advances in Neural Information Processing Systems*, vol. 23, pp. 856–864, 2010.

C. Wang, J. Paisley, and D. M. Blei, "Online variational inference for the hierarchical Dirichlet process," in *Proc. of International Conference on Artificial Intelligence and Statistics*, 2011.

V. C. Raykar, I. Kozintsev, and R. Lienhart, "Position calibration of microphones and loudspeakers in distributed computing platforms," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 1, pp. 70–83, 2005.

N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2009, pp. 161–164.

H. Miura, T. Yoshida, K. Nakamura, and K. Nakadai, "SLAM-based online calibration of asynchronous microphone array for robot audition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 524–529.

S. Miyabe, N. Ono, and S. Makino, "Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in STFT domain," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, 2013, pp. 674–678.

S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *Proc.*

of International Conference on Acoustics, Speech, and Signal Processing, 2005, pp. III–81–III84.

T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement system," in *Proc. of International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 4409–4412.

K. Nickel, T. Gehrig, R. Stiefelhagen, and J. McDonough, "A joint particle filter for audio-visual speaker tracking," in *Proc. of International Conference on Multimodal Interfaces*, 2005, pp. 61–68.

D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 533–544, 2001.

J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, 2001.

L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 13, no. 3, pp. 412–421, 2005.

E. Vincent, "Advances in audio source separation and multisource audio content retrieval," in *SPIE Defense, Security, and Sensing*, 2012.

G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 18, no. 3, pp. 550–563, 2010.

H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 21, no. 5, pp. 971–982, 2013.

A. J. Quinn and B. B. Bederson, "Human computation: A survey and taxonomy of a growing field," in *Proc. of SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 1403–1412.

N. Bryan and G. Mysore, "An efficient posterior regularized latent variable model for interactive sound source separation," in *Proc. of International Conference on Machine Learning*, 2013, pp. 208–216.

A. S. Bregman, *Auditory Scene analysis: the Perceptual Organization of Sound*. MIT Press, 1994.

J. Blauert, *Spatial Hearing: the Pshychophysics of Human Sound Localization*. MIT Press, 1997.

H. Pareek and P. Ravikumar, "Human boosting," in *Proc. of International Conference on Machine Learning*, 2013, pp. 338–346.

List of Selected Publications

Major Publications

Journal Papers

1) **Takuma Otsuka**, Katsuhiko Ishiguro, Hiroshi Sawada, and Hiroshi G. Okuno: Bayesian Nonparametrics for Microphone Array Processing, IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 22, No. 2, pp. 493-504, 2014. → **Chapter 3**.

2) **Takuma Otsuka**, Katsuhiko Ishiguro, Takuya Yoshioka, Hiroshi Sawada, and Hiroshi G. Okuno: Multichannel Sound Source Dereverberation and Separation for Arbitrary Number of Sources based on Bayesian Nonparametrics, IEEE/ACM Transactions on Audio, Speech and Language Processing, under review. → **Chapter 4**.

Other 2 publications as the first author.

International Conference Papers (Peer-reviewed)

1) **Takuma Otsuka**, Kazuhiro Nakadai, Tetsuya Ogata, and Hiroshi G. Okuno: Bayesian Extension of MUSIC for Sound Source Localization and Tracking, *Proceedings of International Conference on Spoken Language Processing (Interspeech 2011)*, pp. 3109–3112, 2011. → **Chapter 7**.
2) **Takuma Otsuka**, Katsuhiko Ishiguro, Hiroshi Sawada, and Hiroshi G. Okuno: Bayesian Unification of Sound Source Localization and Separation with Permutation Resolution, *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12)*, pp. 2038–2045, 2012. → **Chapter 3**.
3) **Takuma Otsuka**, Katsutoshi Ishiguro, Hiroshi Sawada, and Hiroshi G. Okuno: Unified

Auditory Functions based on Bayesian Topic Model, *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS-2012)*, pp. 2370–2376, 2012. → **Chapter 3**.

Other 6 publications as the first author.

Domestic Conference Papers (w/o review)

1) **Takuma Otsuka**, Katsuhiko Ishiguro, Hiroshi Sawada, and Hiroshi G. Okuno: Bayesian Microphone Processing and Its Application to Mobile Robot Audition (in Japanese), *JSAI Technical Report SIG-Challenge*, B202-9, 2012. → **Chapter 6**.