UNIVERSITY OF CALIFORNIA, SAN DIEGO

Blind Separation and Tracking of Sources with Spatial, Temporal and Spectral Dynamics

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in
Electrical Engineering (Signal and Image Processing)

by

Alireza Masnadi-Shirazi

Committee in charge:

Professor Bhaskar D. Rao, Chair
Professor William Hodgkiss
Professor Kenneth Kreutz-Delgado
Professor Lawrence K. Saul
Professor Nuno Vasconcelos

2012

The dissertation of Alireza Masnadi-Shirazi is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2012

DEDICATION

In the Name of Allāh, the Most Gracious, the Most Merciful

I dedicate this thesis to my mother, Mehri Daneshvar, and my father, Dr. Mohammad Ali Masnadi-Shirazi, as I owe them everything after God.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

It has been said that if you want to thank God you should thank his subjects. And so, I would like to acknowledge many people for helping me during my education and doctoral work. Their help and support was instrumental in the completion of this work.

First of all, I would like to express my deepest gratitude to my supervisor, Professor Bhaskar D. Rao. Apart for providing his exceptional scientific guidance and insight, which is without a doubt, reflected in this thesis, I should also thank him for giving me a chance to explore new ideas and new areas while being patient towards my trial-and-errors and also demonstrating to me what it means to be a true scientific professional. I am also grateful for having an exceptional doctoral committee, and wish to thank Professor Kenneth Kreutz-Delgado, Professor Nuno Vasconcelos, Professor William Hodgkiss and Professor Lawrence Saul for their valuable input, accessibility and informative courses.

I would like to thank my colleagues at the DSP lab, Dr. Wenyi Zhang and Oleg Tanchuk for their collaboration and friendship. Also I would like to thank the other DSP lab members, Dr. Yuzhe Jin, Dr. Seong-Ho (Paul) Hur, Dr. Shankar Shivappa, Dr. Yogananda Isukapalli, Dr. Matthew Pugh, Zhilin Zhang, Eddy (Hwan Joon) Kwon, Yichao Huang and Anh Nguyen for their support and assistance throughout the years.

I would like to thank my friends who have helped me in so many ways both in the good times and the bad: Dr. Khosrow Behbehani and his family, Dr. Mahmoud Tarokh and his family, Dr. Amir Rabiee, Dr. Koohyar Minoo and his family, Dr. Mohammad H. Taghavi and his family, Ali Afsahi and his family, Kamaal Martin and his family, Dr. Ebraheem Fontain and his family, Idris Bekkali, Ali Khaki, Sarmad Ashour, the Mir-Jamali brothers, Hosein Zarei and Ali Abu-Talib and his family.

I would like to thank Mr. Javaheri and Mr. Yazdani who taught me the basics of electrical engineering in both hardware and software when I was in high

school. It was with their guidance and help that I was able to finish my first engineering project for the Kharazmi high school level international science fair which was a sound direction detection system using a circular array of microphones and very simple heuristic algorithms. Interestingly, after all these years, my research interests have not taken a major change as that problem and the problem in this thesis are somewhat related.

I would like to thank my extended family, grandparents, my many aunts, uncles and cousins for being my support and foundation throughout the years and during my most difficult times. I reserve a special thanks for my brother, Dr. Hamed Masnadi-Shiraz, for being my brother-in-arms and my sister, Maryam Masnadi-Shirazi, for her endless kindness and grace. Last but not least, I would like to thank my mother, Mehri Daneshvar, and my father, Dr. Mohammad Ali Masnadi-Shirazi for simply everything. I will not even make an attempt at listing the many things you have done for me, I know I can never repay you. I just hope that you are pleased with your son so far and may God reward you.

The text of Chapter III, in full, is based on the material as it appears in: Alireza Masnadi-Shirazi, Wenyi Zhang and Bhaskar Rao, "Glimpsing IVA: A Framework for Overcomplete/Complete/Undercomplete Convolutive Source Separation," IEEE Transactions on Audio, Speech and Language Processing, vol. 18, no. 7, Sept. 2010. The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter IV, in full, is based on the material as it appears in:Alireza Masnadi-Shirazi and Bhaskar Rao, "Separation and Tracking of Multiple Speakers in a Reverberant Environment using a Multiple Model Particle Filtering Glimpsing method" , in proc. IEEE ICASSP 2011. The dissertation author was a primary researcher and an author of the cited material.

The text of Chapter V, in full, is based on the material as it appears in: Alireza Masnadi-Shirazi and Bhaskar Rao, "An ICA-SCT-PHD Filter Approach for Tracking and Separation of Unknown Time-Varying Number of Sources," to

appear in IEEE Transactions on Audio, Speech and Language Processing The dissertation author was a primary researcher and an author of the cited material.

VITA

2001-2005	Bachelor of Science, *Summa Cum Laude* Electrical Engineering, University of Texas at Arlington
2005–2009	Master of Science Electrical Engineering (Signal and Image Processing), University of California at San Diego
2008–2012	Research Assistant Digital Signal Processing Laboratory Department of Electrical and Computer Engineering University of California at San Diego
2009–2012	Doctor of Philosophy Electrical Engineering (Signal and Image Processing), University of California at San Diego

PUBLICATIONS

Alireza Masnadi-Shirazi and Bhaskar Rao, "An ICA-SCT-PHD Filter Approach for Tracking and Separation of Unknown Time-Varying Number of Sources," to appear in IEEE Transactions on Audio, Speech and Language Processing

Alireza Masnadi-Shirazi, Wenyi Zhang and Bhaskar Rao, "Glimpsing IVA: A Framework for Overcomplete/Complete/Undercomplete Convolutive Source Separation," IEEE Transactions on Audio, Speech and Language Processing, vol. 18, no. 7, Sept. 2010

Wenyi Zhang, Alireza Masnadi-Shirazi and Bhaskar Rao, "Insights into the frequency domain ICA/IVA approach," submitted to IEEE Transactions on Signal Processing

Alireza Masnadi-Shirazi and Bhaskar Rao, "An ICA-Based RFS Approach for DOA Tracking of Unknown Time-Varying Number of Sources," in proc. EUSIPCO 2012

Wenyi Zhang, Alireza Masnadi-Shirazi and Bhaskar Rao, "Insights into the Frequency domain ICA approach," in proc. IEEE Asilomar 2011

Alireza Masnadi-Shirazi and Bhaskar Rao, "Separation and Tracking of Multiple Speakers in a Reverberant Environment using a Multiple Model Particle Filtering Glimpsing method" , in proc. IEEE ICASSP 2011

Alireza Masnadi-Shirazi, Wenyi Zhang and Bhaskar Rao, "Glimpsing Independent Vector Analysis: Separating more Sources than Sensors using Active and Inactive States," in proc. IEEE ICASSP 2010

Alireza Masnadi-Shirazi and Bhaskar Rao, "Independent Vector Analysis Incorporating Active and Inactive States," in proc. IEEE ICASSP 2009

ABSTRACT OF THE DISSERTATION

Blind Separation and Tracking of Sources with Spatial, Temporal and Spectral Dynamics

by

Alireza Masnadi-Shirazi
Doctor of Philosophy in Electrical Engineering
(Signal and Image Processing)
University of California, San Diego, 2012
Professor Bhaskar D. Rao, Chair

The problem of separating mixed signals using multiple sensors, commonly known as blind source separation (BSS), has received much attention in recent years. For many real world sources such as acoustical signals, the signals undergo a convoluted mixing due to reverberation caused by the environment. In this thesis we intend to develop algorithms that are able to separate and track convolutedly mixed acoustical sources when dealing with the following adverse scenarios: 1) the number of sources exceeds the number of sensors (overcomplete case), 2) the number of sources is known but their temporal profile is unknown as each source can experience silence periods intermittently, 3) the number of sources is unknown and time-varying as new sources can appear and existing sources can vanish, 4) the sources are moving in space. Overall, these scenarios reflect the spatial and temporal dynamics that acoustical sources can potentially undertake, complicating the BSS problem. In addition, acoustical sources like speech can exhibit spectral dynamics, where the short time Fourier transform (STFT) of the sources experience a certain sparse pattern due to the pitch frequency and formants of speech phonemes that can differ from source to source and from time interval to time interval. In this thesis we will show that spectral dynamics, unlike the other forms

of dynamics, does not complicate the BSS problem and in fact by exploiting it one can simplify the BSS problem when dealing with the adverse aforementioned scenarios. The contributions of this thesis are three algorithms where each algorithm compared to the previous one deals with a more intricate combination of aforementioned scenarios. The first is a batch algorithm that deals with scenarios 1 and 2 by incorporating a glimpsing strategy which "listens in" the silence gaps to compensate for the global degeneracy (of having more sources than sensors) by making use of segments where it is locally less degenerate. The second is an online algorithm that deals with scenarios 1, 2 and 4 by using a glimpsing multiple model particle filter (MMPF) to switch between the different combinations of silence gaps. The third one is a quasi-online algorithm that deals with scenarios 1, 3 and 4 which contain the most uncertainties when compared to the other combinations. In order to deal with this challenging problem, we synergistically combine two key ideas, one in the front end and the other at the back end. In the front end we employ independent component analysis (ICA) to demix the mixtures and the state coherence transform (SCT) to represent the signals in a direction of arrival (DOA) detection framework. By exploiting the spectral sparsity of the sources, ICA/SCT is even effective when the number of simultaneous sources is greater than the number of sensors therefore allowing for minimal number of sensors to be used. At the back end, the probability hypothesis density (PHD) filter is incorporated in order to track the multiple DOAs and determine the number of sources. The PHD filter is based on random finite sets (RFS) where the multi-target states and the number of targets are integrated to form a set-valued variable with uncertainty in the number of sources. A Gaussian mixture implementation of the PHD filter (GM-PHD) is utilized that solves the data association problem intrinsically, hence providing distinct DOA tracks. The distinct tracks also make the separation task possible by going back and rearranging the outputs of the ICA stage.

Chapter I

Introduction

I.A Spatial, Temporal and Spectral Dynamics in Blind Source Separation and Tracking

The problem of separating mixed signals using multiple sensors, commonly known as blind source separation (BSS), has received much attention in recent years. The earliest and most basic form of BSS problems started with a model of linear and instantaneous mixing of the sources. Independent component analysis (ICA) became a popular and promising method to deal with this issue [34]. ICA separates the mixed signals by assuming the sources are statistically independent and the sources are non-Gaussian distributed. For many real world sources such as acoustical signals, the signals undergo a convoluted mixing due to reverberation introduced by the environment. By transforming the mixture to the frequency domain by applying the short-time Fourier transform (STFT), convolution in the time domain translates to linear mixing in the frequency domain. Subsequently, ICA can be performed on every single frequency bin resulting in an estimated mixing matrix for each bin. Since ICA is indeterminate of source permutation, further post processing methods are necessary to correct for possible permutations of the separated sources in each frequency bin [78, 75]. In summary, the process of performing ICA in the frequency bins is commonly known as frequency domain ICA (FD-ICA).

In the context of FD-ICA, in order to simplify the problem in practice, the following assumptions are usually made (in addition to the theoretical assumptions of ICA that sources should be statistically independent and non-Gaussian distributed):

- The number of sources is less than or equal to the number of sensors.
- The sources are either always active or their temporal activity profile is known
- Sources are spatially static.

In this thesis we intend to progressively relax the aforementioned assump-

tions in order to develop algorithms that are able to separate convolutedly mixed acoustical sources when dealing with the following adverse scenarios:

1. The number of sources exceeds the number of sensors (overcomplete case).
2. The number of sources is known but their temporal profile is unknown as each source can experience silence periods intermittently.
3. The number of sources is unknown and time-varying as new sources can appear (birth) and existing sources can vanish (death).
4. The sources are moving in space.

Overall, the above scenarios reflect the spatial and temporal dynamics that acoustical sources can potentially undertake, complicating the BSS problem. In addition, acoustical sources like speech can exhibit spectral dynamics, where the STFT of the sources experience a certain sparse pattern due to the pitch frequency and formants of speech phonemes that can differ from source to source and from time interval to time interval. In this thesis we will show that spectral dynamics, unlike the other forms of dynamics, does not complicate the BSS problem and in fact by exploiting it one can simplify the BSS problem when dealing with the adverse aforementioned scenarios.

Later on we will show that as the development for the separation task is carried out for when the sources exhibit spatial dynamics, new doors are opened allowing for the localization/tracking task of the sources to be accomplished as well.

I.B Contributions of the Thesis

The contributions of this thesis has three main components summarized in three major sections where each section compared to the previous one deals with a harder combination of aforementioned scenarios. The first section deals with scenarios 1 and 2, the second one deals with scenarios 1, 2 and 4, and the third

one deals with scenarios 1, 3 and 4 which contain the most uncertainties when compared to the other combinations.

I.B.1 Glimpsing Independent Vector Analysis

Independent vector analysis (IVA) is a frequency based method for convolutive blind source separation that normally requires no bin-wise permutation correction post-processing [36, 32]. It extends the ICA concept by treating the data in the frequency bins as one multivariate vector and utilizing the inner dependencies between the frequency bins, therefore, significantly reducing the occurrence of bin-wise permutations. IVA models each individual source as a dependent multivariate symmetric super-Gaussian distribution while still maintaining the fundamental assumption of BSS that each source is independent from the other.

We first investigate the role that the temporal dynamics of the signals play in frequency domain BSS and show that in order for the sources to be separable, they must have a dynamic temporal structure. Fortunately, most signals of interest in BSS like speech, music and EEG/MEG follow such structure. We then clarify how such dynamic structure results in a Gaussian scale mixture (GSM) (super-Gaussian shaped) distribution in the frequency domain, therefore, justifying the selection of such distributions that are used in IVA and other ICA-based frequency domain approaches [36, 32, 78].

We take our investigation of the dynamic temporal structure a step further enabling us to build a general IVA-based framework that can facilitate overcomplete convolutive BSS as an extension to the more trouble-free undercomplete/complete BSS. One common type of temporal dynamics, especially present in speech, is that the signals can have intermittent silence periods, hence varying the set of active sources with time. This feature can be used to improve separation in well-determined undercomplete ($L > M$)/complete ($L = M$) cases, and to deal with the ill-determined overcomplete ($L < M$) case. As the set of active sources for each time period decreases, the degree of overcompleteness ($M - L$) decreases

locally. Hence, by exploiting silence gaps, one is actually compensating for the global degeneracy by making use of segments where it is locally less degenerate. We construct a unifying IVA-based framework that can deal with the challenging overcomplete case as well as the straight-forward complete/undercomplete case for convolutive mixing BSS. Various psycho-acoustic studies have confirmed that human listeners use similar strategies of exploiting silence gaps by "glimpsing" or listening in the gaps to identify target speech in adverse conditions of multiple competing speakers [45, §4.3.2], [21, 22]. Consequently, we name this algorithm "glimpsing independent vector analysis (G-IVA)".

I.B.2 Online Separation and Tracking of Sources with Silence Periods

When the sources are allowed to move/maneuver is space, batch algorithms like regular ICA/IVA are no longer effective. Instead one has to resort to an online or quasi-online method to perform the separation and tracking tasks. In the online method, the mixing matrices are updated with every new STFT frame. In a quasi-online method a sequence of STFT frames are accumulated to form a block and the mixing matrices are updated for each block. By assuming that the overall number of sources is known but their activity profile is unknown as they can undergo silence periods intermittently (similar to the glimpsing model discussed earlier), we intend to both fully separate and track the multiple speakers using an online algorithm. Our assertion is that for moving/maneuvering sources, if one is able to track the mixing matrices accurately in the frequency domain to ensure full separation, the accurate localization of the sources based on directions of arrivals (DOA) can be a straightforward consequence as the DOA information is embedded in the mixing matrices. We note that utilizing a glimpsing strategy is essential in an online algorithm because if a source becomes silent but assumed active by the model, the update to the column of the mixing matrix corresponding to that source can diverge or fluctuate unstably [35]. We track the mixing matrices at each bin in the frequency domain by employing a multiple model particle filter (MMPF)

method that is able to switch between the different combinations of silence gaps present in the sources. The proposed algorithm can also maintain track in the more challenging situation where the sources are silent but move around, under the condition that the silence period does not exceed a certain length. We denote these periods as silence blind zones (SBZ). This method can potentially work for the overcomplete case as it uses the same glimpsing strategy discussed earlier.

I.B.3 Quasi-online Tracking and Separation of Unknown Time-Varying Number of Sources

In this section we are particularly interested in estimating the bearing information of multiple sources or their direction of arrival (DOA) by means of the time difference of arrival (TDOA). TDOA estimation is the first stage for many speaker localization algorithms involving one or more microphone pairs. In the case of a single speaker, TDOA can be reliably estimated using the generalized cross-correlation phase transform (GCC-PHAT) using one microphone pair [38, 65]. GCC-PHAT is a scanning method that computes the correlation of the microphone pair inputs for a range of TDOAs with an arbitrary resolution, resulting in peaks where the correlation is high. In case of multiple speakers, GCC-PHAT does not always provide reliable TDOA for all the sources since one of the sources can dominate over the others [11]. This means that as the concurrent sources increase in number, multiple TDOA estimation using GCC-PHAT becomes less reliable. Also, multipath propagation due to reverberation can cause additional peaks in the GCC-PHAT that correspond to multi-path propagations. This results in the situation where for example in the case of two sources, the first and second peak do not always correspond to the first and second source and sometimes the third or subsequent peaks need to be considered [46].

Multiple TDOA estimation using FD-ICA was first proposed in [79]. In [79], similar to the previous section, multiple TDOAs are calculated directly from the columns of the estimated mixing matrix. However, this method works well

only if the possible source permutations in the frequency bins have been corrected and there are no frequency bins affected by spatial aliasing (hence a minimal microphone spacing). Recently an extension to [79] has been proposed under the name of state coherence transform (SCT) that does not require permutation correction and is insensitive to spatial aliasing [57, 59]. Similar to GCC-PHAT, SCT is a scanning method. However, instead of finding the correlation between the two microphone input signals for TDOA points in the scan, it forms a pseudo-likelihood between a propagation model for the different TDOA scan points and the TDOA observations pertaining to the columns of the mixing matrices, resulting in peaks where the scan points in the model and observations best match. One attractive feature of SCT is that by exploiting the frequency sparsity of the sources, it is effective even when the number of simultaneous sources is larger than the number of sensors. Also, since SCT uses ICA outputs which attempt to separate the sources, it is more suitable for TDOA estimation for multiple sources compared to GCC-PHAT [59].

Assuming that the number of sources is known and fixed in time, some methods exist that track the location information for each source by incorporating a separate tracker for each source [18]. However, in many real world problems, not only do the states of the sources change with time, the number of concurrent sources is unknown and varies with time as new speakers can appear and existing speakers can disappear or undergo long silence periods. Moreover, the measurements can receive a set of spurious peaks (clutter) due to the multi-path propagation caused by reverberation and spatial aliasing, resulting in false alarms. In addition, not all of the sources are detected giving rise to missed detections as well. Therefore, the passive scanning methods discussed earlier result in an assortment of indistinguishable observations where only a subset of them are generated by the sources. Recently, methods based on random finite sets (RFS) have presented promising and mathematically elegant solutions to the problem of multi-target tracking (MTT) for time-varying number of targets [48, 47]. Using RFSs,

the collection of indistinguishable observations in the presence of clutter is treated as a set-valued observation while the multi-target states and the number of targets are integrated to form a set-valued state. The goal becomes to estimate the target states and the target number while rejecting clutter and accounting for missed detections. The RFS formulation allows the problem to be posed in an optimal multi-target Bayesian filtering framework, and is an extension of the well known single target Bayes filter. However, the optimal RFS Bayes filter is computationally intractable as it becomes a combinatorial problem on the number of targets involving high dimensional integrals. The probability hypothesis density (PHD) filter is a suboptimal approximation to the RFS Bayes filter which propagates the first moment of multi-target posterior density rather than the full posterior density [48]. This said, the PHD filter still involves multiple integrals with no closed form solution in general. Also, the PHD filter in itself, does not solve the data association problem indicating which estimate belongs to which target. The Gaussian mixture implementation of the PHD filter (GM-PHD) alleviates these two difficulties: It provides a closed form solution of the PHD filter when the target states and observations follow a linear/Gaussian dynamic model (which is a reasonable model for the problem of interest in this paper) [85]. It also solves the data association problem intrinsically and provides track labels which are imperative to the separation task of interest [67].

The problem of extracting location information of unknown time-varying number of speakers using RFSs and PHD filtering has been proposed before. These methods, however, use GCC-PHAT in the front-end to obtain the measurements and bear the inherent limitations of GCC-PHAT for multiple sources including being inherently incapable of source separation [46, 9]. For the same problem, a method exists that uses ICA/SCT in the front-end and uses a naive thresholding approach to estimate the number of targets [44, 43]. This method, however, is sensitive to the selected thresholds and relies solely on the thresholds to reject clutter. As we discussed in the previous section, we proposed an ICA-based ap-

proach to separate and track multiple sources for when the sources can experience short silence periods [51]. This method, while being able to separate the sources, only estimates the activity patterns and cannot handle new sources being born or completely dying out. In this section we propose a quasi-online algorithm that uses the GM-PHD to filter the measurements obtained from short time blocks using ICA/SCT. By doing so we are able to track the DOA of multiple time-varying number of sources and from the track labels we are able to go back to the ICA outputs and perform the separation task by associating each separated time-frequency block with its estimated corresponding track. The separation scheme exploits the frequency sparsity (i.e. spectral dynamics) of the sources and enables the separation of more concurrent sources than sensors. In contrast to the previous section, this section does not model the short silence periods explicitly but rather focuses on the birth and deaths of the sources. However, since this section incorporates a quasi-online method that considers missed detections, short pauses in the sources are implicitly taken care of. If the pause/silence period exceeds a certain length in time the source will be assigned a new track once it becomes active again.

Overall, this section demonstrates how a mixture/superposition model in the framework of BSS can be easily represented as a standard detection model in the framework of multi-target tracking, assuming that the sources have frequency sparsity. Such an idea of transforming a mixture/superposition model to a detection model, was first presented in [10], where the sources were assumed to be narrowband audio tones and the STFT representation was enough to execute such transformation. As it turns out, the approach in [10] is a special case of the proposed method for when the sources have a super-sparse representation to a degree where they will be non-overlapping and occupy a single frequency bin, making the ICA separation scheme unnecessary. The proposed method offers a solution for executing the transformation from the mixture model to the detection model for broadband signals that have some sort of frequency sparsity, such as speech and communication signals.

I.C Organization of the thesis

The rest of the thesis is organized as follows. In Chapter II, we present some background on basic Maximum likelihood ICA as well as some background on sequential Bayesian state estimation and filtering. In Chapter III we develop a BSS batch algorithm that deals with the ill-determined overcomplete case of having more sensors than sources by glimpsing or "'listening in" the silence gaps in both time and frequency. In Chapter IV we make the problem more challenging by allowing the sources to move/maneuver is space while they can still experience silence periods and use the same combinatorial glimpsing strategy in both time and frequency. In Chapter V we take the previous challenge one step further by allowing the number of sources to be unknown and time-varying where new sources can appear and existing sources can disappear. Finally, in Chapter VI conclusions of the thesis are presented.

Chapter II

Background

II.A Independent Component Analysis

ICA is a method for finding underlying factors or components from multivariate statistical data and is used frequently in the BSS problem often known as the "cocktail party" problem. Suppose there are n speakers in a cocktail party with n microphones located at different places in the room. Denote each speaker with $s_i, i = 1, ..., n$ and each microphone input by $x_i, i = 1, ..., n$. If we assume there is a linear instantaneous mixing of the sources (assuming no reverberation and delays) caused by the distance from each speaker to each microphone, we have:

$$\begin{aligned} x_1(t) &= a_{11}s_1(t) + a_{12}s_2(t) + ... + a_{1n}s_n(t) && \text{(II.1)} \\ x_2(t) &= a_{21}s_1(t) + a_{22}s_2(t) + ... + a_{2n}s_n(t) && \text{(II.2)} \\ &\vdots \\ x_n(t) &= a_{n1}s_1(t) + a_{n2}s_2(T) + ... + a_{nn}s_n(t) && \text{(II.3)} \end{aligned}$$

the above equations can be written in matrix form as follows:

$$X = AS \tag{II.4}$$

where,

$$X = [x_1(t)x_2(t)...x_n(t)]^T \tag{II.5}$$

$$S = [s_1(t)s_2(t)...s_n(t)]^T \tag{II.6}$$

$$A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ \vdots & \vdots & ... & \vdots \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix}. \tag{II.7}$$

We do not observe the sources S and neither do we observe the mixing matrix A. Our goal is to estimate the sources S when just observing X. Because of the lack of information to solve this inverse problem we need to make further assumptions. ICA , in its basic form, only requires two assumptions. The first assumption is that

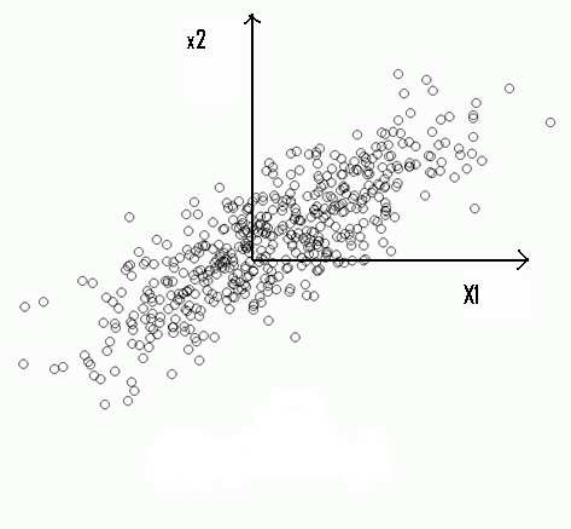

Figure II.1 Joint distribution of the mixture of two Gaussian distributed signals

our sources are statistically independent of each other. This makes sense because the human brain separates signals based on this assumption as well. Before we go to the second assumption which requires some more reasoning we lay the framework for the pre-processing step. This will also serve as a tool to justify the second assumption.

Independence is a much stronger case than uncorrelatedness. This means that two random variables can be uncorrelated but still dependent. In other words, this means that we can make the data uncorrelated without disturbing any underlying factor that controls their independence. As a preprocessing step we choose to center the data (subtract its mean from the data) and then whiten it (make the data uncorrelated or spherical). We keep in mind that these preprocessing steps do not disturb our assumption of having the source signals independent of each other.

The second assumption is that the source densities should be non-Gaussian distributed (more precisely, not more than one of the source densities can be Gaussian). To understand why this should be the case we set the following example. Assume we have two Gaussian distributed signals that are mixed together by a

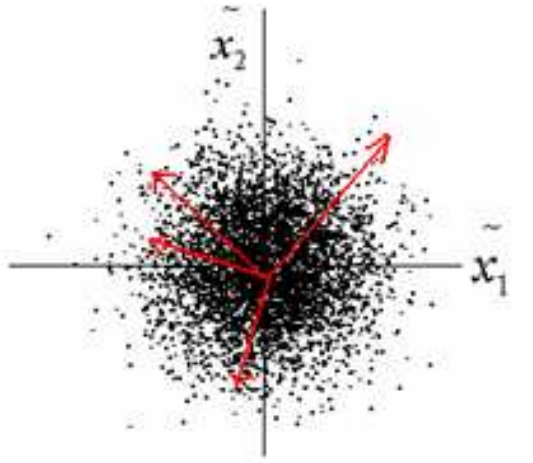

Figure II.2 Joint distribution of the mixture of two Gaussian distributed signals after it has been whitened

2×2 matrix A. Figure II.1 shows the joint distribution after they have been linearly mixed. If we whiten this data as a pre-processing step we will make the data uncorrelated (spherical). Figure II.2 shows the uncorrelated data. The goal of ICA is to iteratively converge towards the directions that yield maximum independence. Because in this case the whitened data is symmetrical from all directions, therefore there are countless directions (some are shown by the arrows in Figure II.2) that yield maximum independence. This is the reason why not more than one of the sources can be Gaussian distributed for ICA to separate the sources. In other words, non-Gaussianity of the sources ensures a non-spherical structure that stores information about the vector direction of each column of the mixing matrix.

To demonstrate how ICA can be performed on sources that are distributed other than Gaussian we refer to Figure II.3 which shows the steps towards finding the independent sources. On the left is the figure of 2 uniformly distributed mixed signals. The middle figure shows the pre-processing step of whitening the data. After pre-processing has been done all we need to do is to find an orthogonal matrix that when multiplied by the whitened data, yields maximum independence. The columns of this orthogonal matrix is shown by the arrows in the middle figure.

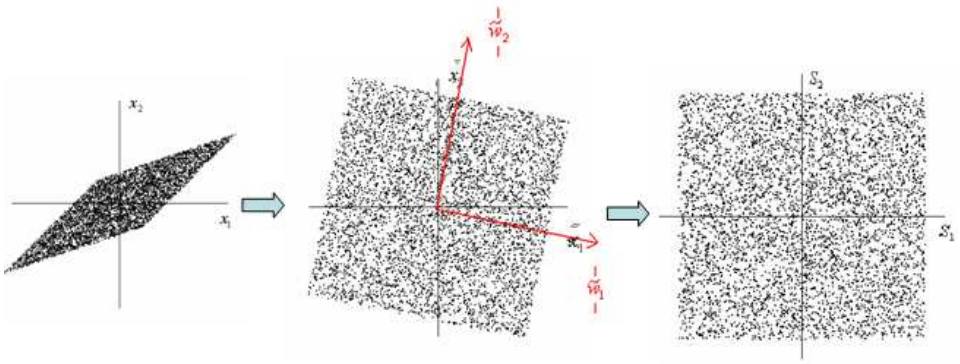

Figure II.3 Left: Joint distribution of the mixture of two uniform distributed signals. Middle: Whitened. Right: Recovered source signals after the whitened data is rotated by an orthogonal matrix that yields maximum independence

The figure on the right shows the independent source distribution as a result of being multiplied by the inverse of the orthogonal matrix, i.e. an estimate of $\hat{A}^{-1}$, where $\hat{A}$ is the mixing matrix in the "whitened space".

II.A.1 Maximum Likelihood approach to ICA

The first step is to derive the likelihood. We assume that we have whitened our data as a pre-processing step. According to $X = AS$ and assuming the independency of the sources(i.e. $p(S) = p(s_1)p(s_2)...p(s_2)$) the density of $p(X)$ is

$$p(X) = |detB|p(S) = |detB| \prod_{i=1}^{n} p(s_i) \tag{II.8}$$

where $B = A^{-1}$, and $p(s_i)$ denotes the densities of the i^{th} independent component. This can be expressed as a function of $B = [b_1, ..., b_n]^T$ and X, giving

$$p(X) = |detB| \prod_{i=1}^{n} p(b_i^T X) \tag{II.9}$$

Assuming that we have T i.i.d. observations of X in time, denoted by $\{X(t)\}_{t=1}^{T}$, then the likelihood can be obtained as the product of the density at the T points in

time. The likelihood is denoted by L and considered as a function of the parameter B:

$$L(B) = \prod_{t=1}^{T} \prod_{i=1}^{n} p(b_i^T X(t)) |detB| \tag{II.10}$$

the Log-Likelihood becomes:

$$\log L(B) = \sum_{t=1}^{T} \sum_{i=1}^{n} \log p(b_i^T X(t)) + Tlog|detB| \tag{II.11}$$

$$\log L(B) = \sum_{t=1}^{T} \sum_{i=1}^{n} \log p(e_i^T BX(t)) + Tlog|detB| \tag{II.12}$$

where e_i is a vector of all zeros except for the i^th position where it is one.

Our goal is find the B that maximizes the likelihood. To do so we move in direction of the gradient of the log-likelihood with respect to B.

$$\frac{\partial logL(B)}{\partial B} = \sum_{t=1}^{T} \sum_{i=1}^{n} g_i(e_i^T BX(t)) e_i X(t)^T + TB^{-T} \tag{II.13}$$

where $g_i = (\log p(s_i))' = \frac{p(si)'}{p(s_i)}$ thus we have:

$$\frac{\partial logL(B)}{\partial B} = \sum_{t=1}^{T} g(BX(t)) X(t)^T + TB^{-T} \tag{II.14}$$

Thus our adaptive algorithm that goes in the direction of the gradient accent looks like:

$$B_{k+1} = B_k + \Delta B \tag{II.15}$$

where ΔB is the gradient in Eq. II.14.

However there is some ambiguity here. Because we do not know the sources yet, we do not have the source densities to use in the gradient equation above. The most basic ICA algorithms pick one standard density function for the case when our source densities are super-Gaussian and pick another density function when our sources are sub-Gaussian. Other adaptive methods exist that learn parameters controlling the shape of the source densities, but will not be discussed here. An

example of a super-Gaussian density is a Laplacian where it has heavier tails and a peakier center compared to the Gaussian. An example of a sub-Gaussian density is a uniform density. Also, One can use the following functions for the super/sub-Gaussian cases

$$\log p_i^+(s) = \alpha_1 - 2\log\cosh(s) \tag{II.16}$$

$$\log p_i^-(s) = \alpha_2 - [s^2/2 - \log\cosh(s)] \tag{II.17}$$

where the $+$ simulates a super-Gaussian density and the $-$ simulates a sub-Gaussian density. α_1 and α_2 are constants to make the densities identifiable. This results in the following:

$$g^+(y) = -2\tanh(y) \tag{II.18}$$

$$g^-(y) = \tanh(y) - y. \tag{II.19}$$

Once we have found the converged B, we can find the estimate of the source signals by computing $\hat{S} = BX$. For a more complete understanding of ICA, we refer the reader to the book in [34].

II.B Sequential Bayesian Filtering

In sequential Bayesian filtering the goal is to iteratively and optimally solve the inverse problem of estimating the state vector x_k from measurements z_k, where k is the discrete time index. The target state evolves according to the following discrete-time stochastic model

$$x_k = a_{k-1}(x_{k-1}) + v_{k-1} \tag{II.20}$$

where a_{k-1} is a known, possibly nonlinear function of the state x_{k-1} and v_{k-1} is the process noise sequence. On the other hand, the measurements are related to the target state via another discrete-time stochastic model known as the measurement equation:

$$z_k = b_k(x_k) + w_k \tag{II.21}$$

where b_k is a known, possibly nonlinear function and w_k the measurement noise sequence. The noise sequences v_{k-1} and w_k are assumed to be white and independent of each other. The initial target state is assumed to have a known pdf $p(x_0)$.

From a Bayesian perspective, the problem is to recursively update the state x_k based on some quantification on degree of belief using the measurements up to time k, i.e. $Z_k = \{z_i\}_{i=1}^k$. Thus, what Bayesian filtering seeks to do is to construct the posterior pdf $p(x_k|Z_k)$. This is usually done by breaking down the recursive pattern of going from time $k-1$ to k in two steps of prediction and update .

Suppose that the density $p(x_{k-1}|Z_{k-1})$ at time $k-1$ is available. The prediction step is to find the prediction posterior via:

$$p(x_k|Z_{k-1}) = \int p(x_k|x_{k-1})p(x_{k-1}|Z_{k-1})dx_{k-1}. \tag{II.22}$$

At time step k when measurement z_k becomes available the updated posterior density can be found via the optimal Bayes rule:

$$p(x_k|Z_k) = p(x_k|z_k, Z_{k-1}) \tag{II.23}$$

$$= \frac{p(z_k|x_k, Z_{k-1})p(x_k|Z_{k-1})}{p(z_k|Z_{k-1})} \tag{II.24}$$

$$= \frac{p(z_k|x_k)p(x_k|Z_{k-1})}{p(z_k|Z_{k-1})} \tag{II.25}$$

where the normalizing constant is

$$p(z_k|Z_{k-1}) = \int p(z_k|x_k)p(x_k|Z_{k-1})dx_k. \tag{II.26}$$

Once the posterior density is found, one can compute the optimal state by finding the mean of the posterior, leading to the minimum mean square error (MMSE) estimate

$$\hat{x_k}^{MMSE} = E[x_k|Z_k], \tag{II.27}$$

or by finding the maximum of the posterior pdf which leads to the maximum a-posteriori (MAP) estimate

$$\hat{x_k}^{MAP} = \underset{x_k}{\text{argmax}}\ p(x_k|Z_k). \tag{II.28}$$

Due to the presence of integrals in Eqs. II.22 and II.26, a closed form solution cannot be analytically determined, in general. However, for the case where the functions a_{k-1} and b_k are linear and the noises v_{k-1} and w_k are Gaussian, a closed form solution can be achieved analytically which coincidently turns out to be the optimal Kalman filter prediction and update equations. For all the other cases a sub-optimal solution needs to be undertaken in order to approximate the integrals. One common method is particle filtering which approximates the integrals using Monte Carlo simulations. For a better understanding on particle filters we refer the reader to the book in [74]. In this thesis we intend to both separate and track multiple speakers in a reverberant environment when dealing with adverse scenarios. We use the main idea of ICA discussed earlier for the separation task and the main idea of sequential Bayesian filtering for the tracking task.

Chapter III

Glimpsing Independent Vector Analysis: A General Framework for Overcomplete, Complete and Undercomplete Source Separation

III.A Introduction

Independent vector analysis (IVA) is a frequency based method for convolutive blind source separation that normally requires no bin-wise permutation correction post-processing[36, 32]. It extends the ICA concept by treating the data in the frequency bins as one multivariate vector and utilizing the inner dependencies between the frequency bins, therefore, significantly reducing the occurrence of bin-wise permutations. IVA models each individual source as a dependent multivariate symmetric super-Gaussian distribution while still maintaining the fundamental assumption of BSS that each source is independent from the other. Other frequency domain methods exist for convolutive BSS that are not ICA-based and perform separation and permutation correction by exploiting properties of source nonstationarity [1] [68, 69, 72, 56, 20, 63].

In this chapter we investigate the role that the dynamics of the signals play in frequency domain BSS and show that in order for the sources to be separable, they must have a dynamic temporal structure. Fortunately, most signals of interest in BSS like speech, music and EEG/MEG follow such structure. We then clarify how such dynamic structure results in a Gaussian scale mixture (GSM) (super-Gaussian shaped) distribution in the frequency domain, therefore, justifying the selection of such distributions that are used in IVA and other ICA-based frequency domain approaches [36, 32, 78]. Lee et al. proposed using a Gaussian mixture model (GMM) for the source distributions by extending independent factor analysis (IFA) to the multivariate case of IVA [40]. IFA is an instantaneous mixture BSS method in the presence of noise which uses a GMM with unknown parameters for the source priors, hence enabling the modeling of a wide range of super-Gaussian, sub-Gaussian and multi-modal distributions [8]. By extending IFA to the multivariate frequency domain case for convoluted mixtures, the same

[1] The notion of "nonstationarity" used in these articles are loose termed and do not follow the definition of a nonstationarity in random processes. Strictly speaking, what makes these algorithms work is not the nonstationarity of the signals, but rather the property that each realization of the source signals has a time varying envelope [71]. In this chapter, we use the same property but we will choose not to use the term "nonstationarity" in order to avoid confusion.

wide range of freedom in the modeling of the sources is allowed. However such general models are unnecessary when knowledge about the general shape of the source distributions can be achieved a priori as a consequence of their dynamics, and could lead to overlearning due to the high number of parameters of the GMM to be estimated. In this chapter, as we intend to model the noise as well, we approximate the GSM super-Gaussian source distributions using a fixed GMM with zero means as they are adequate and tractable.

For standard ICA-based methods, when the number of sources M becomes greater than the number of sensors L ($M > L$), i.e. the matrix is overcomplete, the process of estimating the mixing matrix and the sources are not that straightforward. Various methods in the past with different underlying assumptions have been proposed to deal with overcompleteness (degeneracy) in ICA linear instantaneous mixing. Lee et al. used a maximum likelihood approximation framework for learning the overcomplete mixing matrix and a maximum a posteriori (MAP) estimator with Laplacian source priors, which can be viewed as a ℓ_1 norm minimization problem, to reconstruct the sources[41]. Bofill and Zibulevsky proposed transforming the observations to the frequency domain to increase sparsity, finding the mixing matrix using a geometric method and recovering the sources using the ℓ_1 norm mimization[14]. The ℓ_1 minimization scheme does not guarantee sparse solutions when the sources are not disjoint or nearly disjoint, regardless of whether they are Laplacian distributed or not [14, 80]. In other words when the sources overlap, the reconstruction could yield leakage from other sources during periods when it is actually silent. Other methods incorporate geometric/probabilistic clustering approaches to find the mixing matrix while relying heavily on sparsity to recover the sources, such that it is assumed that at every instant mostly one source is active [55, 23, 61, 62, 25, 1]. Vielva et al. proposed a MAP estimator that seeks the best combination of the columns of the mixing matrix, assuming the mixing matrix is known or estimated beforehand [84]. All such methods, however, do not take into consideration the temporal dynamic structure of the signals for mixing

matrix estimation and, especially, source reconstruction.

Methods for overcomplete BSS have also been proposed for convolutive mixing. Some methods in auditory scene analysis[15] use binary masking/clustering in the time-frequency spectrogram to isolate the sources, assuming that every time-frequency point belongs to one source [89, 7]. Methods that combine ICA(in each frequency bin) with binary masking have also been proposed [77, 6, 70]. Other methods work by performing instantaneous overcomplete BSS on each frequency bin separately, reconstruct the sources in each frequency bin by either using an ℓ_1 minimization approach or only allowing one source component be active at a time, and correct for permutations afterwards [87, 64, 76].

In this chapter we take our investigation of the dynamic temporal structure a step further enabling us to build a general IVA-based framework that can facilitate overcomplete convolutive BSS as an extension to the more trouble-free undercomplete/complete BSS. One common type of temporal dynamics, especially present in speech, is that the signals can have intermittent silence periods, hence varying the set of active sources with time. This feature can be used to improve separation in well-determined undercomplete ($L > M$)/complete ($L = M$) cases, and to deal with the ill-determined overcomplete ($L < M$) case. As the set of active sources for each time period decreases, the degree of overcompleteness ($M - L$) decreases locally. Hence, by exploiting silence gaps, one is actually compensating for the global degeneracy by making use of segments where it is locally less degenerate. An ICA-based approach to model active and inactive intervals for instantaneous linear mixing BSS has been proposed by Hirayama et al. [31]. This method models the sources as a two-mixture of Gaussians with zero means and unknown variances similar to that of IFA, and incorporates a Markov model on a hidden variable that controls state of activity or inactivity for each source. A complicated and inefficient three layered hidden variable (one for the Markov state of activity and two as in normal IFA) estimation algorithm based on variational Bayes is implemented. Extending this to IVA for convoluted mixtures proves to be

even more complicated. In our previous work we proposed a simple and efficient algorithm to model the states of activity and inactivity in the presence of noise for the well determined complete/undercomplete cases of convoluted mixing using a simple mixture model [50]. Unlike the method in [31] where the on/off states were embedded in the sources themselves, they were modeled more naturally as controllers turning on and off the columns of the mixing matrices. In this chapter we build upon our previous work to construct a unifying IVA-based framework that can deal with the challenging overcomplete case as well as the straight-forward complete/undercomplete case for convolutive mixing BSS. The proposed algorithm has the following characteristics: 1) utilizing inner-frequency dependencies to reduce the occurrence of the well-known permutation problem, 2)incorporating active/inactive feature of the dynamic temporal structure of the sources so that the learning is performed on a local level, 3) incorporating a Markovian support on top of the active/inactive dynamics to be used for the ill-determined overcomplete case to allow better separability when the sources overlap, 4) having the capability of separating the sources when the number of sources is possibly unknown, 5) applying an optimal and efficient minimum mean square error (MMSE) estimator for source reconstruction using the outputs from the estimated mixing matrices and state probabilities, 6) including white Gaussian noise in the model framework. Various psycho-acoustic studies have confirmed that human listeners use similar strategies of exploiting silence gaps by "glimpsing" or listening in the gaps to identify target speech in adverse conditions of multiple competing speakers [45, §4.3.2], [21, 22]. Consequently, we name our algorithm "glimpsing independent vector analysis (G-IVA)".

This chapter is organized as follows: Section III.B explains the generative convolutive model and derives the source distributions in the frequency domain as a consequence of the dynamic modulations of the signal in the time domain. Then, estimation procedures for complete/undercomplete and overcomplete convolutive BSS problems are presented and the source reconstruction method is given.

Section III.C gives some pre-processing and post-processing techniques for faster convergence and further improvement. In Section V.E, some results are evaluated. The main focus of the results is on the overcomplete case, since it is more challenging. Finally, in Section III.E, our conclusions are stated and the main contributions of this chapter are summarized.

III.B Convolutive Mixing Model

Assuming L sensors and M sources, with no restriction on the relationship between L and M, the convolutedly mixed observation at the l^{th} sensor is

$$y_l(t) = \sum_{j=1}^{M} \sum_{r=0}^{R-1} h_{lj}(r) s_j(t-r) + w_l(t) \tag{III.1}$$

where $s_j(t)$ is the j^{th} source in the time domain, $h_{lj}(t)$ is the impulse response of duration R linking the j^{th} source to the l^{th} sensor and $w_l(t)$ is zero mean Gaussian white noise. The signals are transformed to the frequency domain using the short time Fourier transform (STFT). The STFT takes the discrete Fourier transform (DFT) of blocks (frames) of the signal using a sliding window, hence creating a time-frequency representation of the signal, commonly known as the spectrogram. We must note that the window length of the STFT should be sufficiently large, ensuring that the conversion from convolution in the time domain, be approximated fairly by multiplication in the frequency domain. Using STFT, the l^{th} sensor observation at time block n and frequency bin k becomes

$$Y_l^{(k)}(n) = \sum_{j=1}^{M} H_{lj}^{(k)} S_j^{(k)}(n) + W_l^{(k)}(n) \tag{III.2}$$

where $S_j^{(k)}(n)$ is the frequency domain representation of the j^{th} source at bin k and frame n, $W_l^{(k)}(n)$ is the frequency domain noise at bin k and frame n added to the l^{th} sensor and having variance σ_{w_l}. We can arrange Eq. V.2 for all frequency bins $k = 1, ..., d$ in matrix form as

$$Y^{(1:d)}(n) = H^{(1:d)} S^{(1:d)}(n) + W^{(1:d)}(n) \tag{III.3}$$

where $Y^{(1:d)} = \left[Y_1^{(1)}...Y_L^{(1)}|...|Y_1^{(d)}...Y_L^{(d)}\right]^T$, $H^{(1:d)} = \begin{pmatrix} H^{(1)} & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & H^{(d)} \end{pmatrix}_{(Ld)\times(Md)}$,
$S^{(1:d)} = \left[S_1^{(1)}...S_M^{(1)}|...|S_1^{(d)}...S_M^{(d)}\right]^T$ and $W^{(1:d)} = \left[W_1^{(1)}...W_L^{(1)}|...|W_1^{(d)}...W_L^{(d)}\right]^T$[2].
$H^{(k)}$ is the $L \times M$ mixing matrix for the k^{th} frequency bin with its entries being $H_{lj}^{(k)}$ from Eq. V.2. Since the noise is assumed white, the covariance of the noise can be written as $\Sigma_W = \begin{pmatrix} \sigma_W & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & \sigma_W \end{pmatrix}_{(Ld)\times(Ld)}$ where $\sigma_W = diag\left(\sigma_{w_1}, ..., \sigma_{w_L}\right)$.

III.B.1 Source Distributions

Let s_j be the j^{th} source in the time domain. By taking the short time Fourier transform (STFT) of source s_j at time block n, the vector of frequency coefficients is

$$S_j^{(1:d)}(n) = \sum_{t=0}^{Q-1} s_j(t + nJ)e^{-i\frac{2\pi(1:d)}{d}t} \tag{III.4}$$

where $S_j^{(1:d)}(n) = \left[S_j^{(1)}(n), ..., S_j^{(d)}(n)\right]^T$, $e^{-i\frac{2\pi(1:d)}{d}t} = \left[e^{-i\frac{2\pi.1}{d}t}, ..., e^{-i\frac{2\pi.d}{d}t}\right]^T$, Q is the STFT sliding window length, d is the DFT length ($d \geq Q$) and J is the sliding window shift size ($J < Q$). We assume that the time domain signal s_j at block n is a realization of a zero mean stationary time series with a power spectrum vector defined as

$$f_{s_j s_j}^{(1:d)}(n) = \sum_u c_n(u)e^{-i\frac{2\pi(1:d)}{d}u} \tag{III.5}$$

where $f_{s_j s_j}^{(1:d)}(n) = [f_{s_j s_j}^{(1)}(n), ..., f_{s_j s_j}^{(d)}(n)]^T$ with $f_{s_j s_j}^{(k)} \in \mathcal{R}^{\geq 0}$, and c_n is an absolutely summable autocorrelation function of the signal for block n defined as

$$c_n(u) = E[s_j(t + nJ)s_j(t + nJ + u)]. \tag{III.6}$$

[2]Throughout this chapter A^T, A^* and A^H denote the transpose, complex conjugate and conjugate transpose of matrix/vector A, respectively.

The spectrum is indexed by the frame index to capture the dynamic nature of the source signal. i.e. the statistics can vary from frame to frame. Using the central limit theorem and noting that the DFT bins are uncorrelated from each other, the frequency domain vector $S_j^{(1:d)}(n)$ of block n, conditioned on the power spectrum for that block is asymptotically distributed as a complex zero mean multivariate Gaussian with diagonal covariance as follows[16, theorem 4.4.1]

$$\begin{aligned} & P\left(S_j^{(1:d)}(n)|f_{s_j s_j}^{(1:d)}(n)\right) \\ & = \mathcal{N}\left(S_j^{(1:d)}(n); 0, diag\left(Qf_{s_j s_j}^{(1)}(n), ..., Qf_{s_j s_j}^{(d)}(n)\right)\right) \end{aligned} \tag{III.7}$$

Similar to hidden Markov models (HMMs) commonly used in speech, to model the frame dynamics we associate the power spectrum at block n with a hidden variable/vector for that block denoted as $\underline{\xi}_n$. Eq. III.7 can be rewritten as

$$\begin{aligned} & P\left(S_j^{(1:d)}(n)|\underline{\xi}_n\right) \\ & = \mathcal{N}\left(S_j^{(1:d)}(n); 0, diag\left(\sigma^{(1)}(\underline{\xi}_n), ..., \sigma^{(d)}(\underline{\xi}_n)\right)\right) \end{aligned} \tag{III.8}$$

From Eq. III.8, the unconditional probability density function (PDF) of the Fourier coefficients vector of the sources for all blocks can be written as

$$\begin{aligned} P\left(S_j^{(1:d)}\right) &= \int_{\underline{\xi}} P\left(S_j^{(1:d)}|\underline{\xi}\right) P\left(\underline{\xi}\right) d\underline{\xi} \\ &= \int_{\underline{\xi}} \mathcal{N}\left(S_j^{(1:d)}; 0, diag\left(\sigma^{(1)}(\underline{\xi}), ..., \sigma^{(d)}(\underline{\xi})\right)\right) P\left(\underline{\xi}\right) d\underline{\xi}. \end{aligned} \tag{III.9}$$

If the source signal has a dynamic power spectrum, modeled by the hidden variable $\underline{\xi}$, Eq. III.9 can be viewed as a mixture of infinite Gaussians with zero means and varying diagonal covariances. This is the well known GSM model [3]. Depending on the distribution of the scaling variable $\underline{\xi}$, $P\left(S_j^{(1:d)}\right)$ (Eq. III.9) may or may not have a closed-form expression. If it is assumed that the diagonal elements of the covariance matrix all have the same values , $\sigma^{(1)}(\underline{\xi}) = ... = \sigma^{(d)}(\underline{\xi}) = \underline{\xi}$ (i.e the signal being a white stationary time series for each block), and for instance, $\underline{\xi}$ follows an inverse Gamma distribution, then $P\left(S_j^{(1:d)}\right)$ is the multivariate spherical Student t-distribution [13, §2.3.7]. A similar spherical GSM model was stated

in the original IVA papers without much discussion on why the distributions in the frequency domain followed such form [36, 26, 32]. In [36, 26, 37], a Gamma prior was employed and the resulting PDF (multivariate K distribution) was approximated in the heavy tails region to be the multivariate spherical Laplacian distribution. Palmer et al. derived the GSM format for IVA independently in [66]. The relationship between non-Gaussianity and the dynamic temporal structure of the sources were also discussed in [69, 71, 87]. For a more rigorous analytic investigation of frequency domain ICA/IVA methods we direct the reader to our companion article in [90].

If the time domain source signal s_j has no temporal dynamics, then its power spectrum is constant over time for all frames. This means that the overall distribution of the variable controlling the power spectrum $P\left(\underline{\xi}\right)$ is a Dirac delta function, $P\left(\underline{\xi}\right) = \delta\left(\underline{\xi} - \alpha\right)$. Consequently, the overall distribution of the source $P\left(S_j^{(1:d)}\right)$ will be Gaussian distributed,

$$P\left(S_j^{(1:d)}\right) = \mathcal{N}\left(S_j^{(1:d)}; 0, diag\left(\sigma^{(1)}(\alpha), ..., \sigma^{(d)}(\alpha)\right)\right). \tag{III.10}$$

Since Gaussian source signals cannot be separated by independence analysis, the above discussion concludes that conventional frequency domain ICA or IVA approaches cannot separate mixed sources without time varying amplitudes.

In this chapter we approximate the GSM in Eq. III.9 with a finite number of Gaussians to form a GMM as follows

$$P\left(S_j^{(1:d)}\right) = \sum_{c_j=1}^{C} \alpha_{j_{c_j}} \mathcal{N}\left(S_j^{(1:d)}; 0, diag\left(\sigma_{j_{c_j}}^{(1)}, ..., \sigma_{j_{c_j}}^{(d)}\right)\right) \tag{III.11}$$

where the variances $\sigma_{j_{c_j}}^{(k)}$ and the mixture coefficients $\alpha_{j_{c_j}}$ are learned and fixed beforehand to approximate a multivariate GSM model (if the model is directly learned from, say, speech signals, we avoid including prolonged silence periods in the data because silence information will be learned separately in the next part of this chapter). For our experiments, the spherical form of Eq. III.11 where $\sigma_{j_{c_j}}^{(1)} = ... = \sigma_{j_{c_j}}^{(d)} = \sigma_{j_{c_j}}$ has been found to be sufficient. This simplifies the density

function to

$$P\left(S_j^{(1:d)}\right) = \sum_{c_j=1}^{C} \alpha_{j_{c_j}} \mathcal{N}\left(S_j^{(1:d)}; 0, \sigma_{j_{c_j}}.I_d\right) \tag{III.12}$$

where I_d is the $d \times d$ identity matrix. This is mainly because whitening is performed on each frequency bin separately as a preprocessing step which makes the sources have roughly unit variance for each frequency bin (see section III.C.1). Nonetheless, for the sake of generality, through the rest of this chapter we express the GMM as in Eq. III.11.

The joint density of the M independent sources is the product of the marginal densities. Hence, we have

$$\begin{aligned} P\left(S^{(1:d)}\right) &= \prod_{j=1}^{M} \sum_{c_j=1}^{C} \alpha_{j_{c_j}} \mathcal{N}\left(S_j^{(1:d)}; 0, diag(\sigma_{j_{c_j}}^{(1)}, ..., \sigma_{j_{c_j}}^{(d)})\right) \\ &= \sum_{q=1}^{C^M} w_q \mathcal{N}\left(S^{(1:d)}; 0, V_q\right) \end{aligned} \tag{III.13}$$

where $\sum_{q=1}^{C^M} = \sum_{c_1=1}^{C} ... \sum_{c_M=1}^{C}$, $w_q = \prod_{j=1}^{M} \alpha_{j_{c_j}}$ and

$$V_q = \begin{pmatrix} v_q^{(1)} & ... & 0 \\ \vdots & \ddots & \vdots \\ 0 & ... & v_q^{(d)} \end{pmatrix}_{(Md)\times(Md)} \qquad \text{with } v_q^{(k)} = diag\left(\sigma_{1_{c_1}}^{(k)}, ..., \sigma_{M_{c_M}}^{(k)}\right).$$

III.B.2 Active and Inactive States

We assume that each source signal will have silence periods and to take advantage of this knowledge we associate two states with each source. At any frame each source can take on two states, either active or inactive. For M sources there will be a total of 2^M states. As a convention throughout this chapter we will arbitrarily encode the states by a number between 1 and $I = 2^M$ with a circle around it. These states are the same for all frequency bins and indicate which column vector(s) of the mixing matrix is(are) present or absent.

Let the source indices form a set $\Omega = \{1, ..., M\}$, then any subset of Ω could correspond to a set of active source indices. For state $\textcircled{i}$, we denote the

subset of active indices in ascending order by $\Omega_i = \{\Omega_i(1), ..., \Omega_i(M_i)\} \subseteq \Omega$, where $M_i \leq M$ is the cardinality of Ω_i (i.e. the number of active sources at a frame). As an example if $M = 2$, their will be a total of four states corresponding to the first source being active, the second source being active, both being active or none being active. From Eq. V.3 and using the source distribution of Eq. III.11, by effectively selecting the columns of the mixing matrices that correspond to each state, it can be easily shown that the observation density function for state $\textcircled{i}$, regardless of being overcomplete/complete/undercomplete is

$$P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right) = \sum_{q_{\textcircled{i}}} w_{q_{\textcircled{i}}} \mathcal{N}\left(Y^{(1:d)}(n); 0, A^{(1:d)}_{q_{\textcircled{i}}}\right) \tag{III.14}$$

where $A^{(1:d)}_{q_{\textcircled{i}}} = \Sigma_W + H^{(1:d)}_{\textcircled{i}} V_{q_{\textcircled{i}}} H^{(1:d)^H}_{\textcircled{i}}$, $\sum_{q_{\textcircled{i}}} = \sum_{c_{\Omega_i(1)}=1}^{C} \cdots \sum_{c_{\Omega_i(M_i)}=1}^{C}$, $w_{q_{\textcircled{i}}} = \prod_{j=1}^{M_i} \alpha_{\Omega_i(j)c_{\Omega_i(j)}}$, $H^{(1:d)}_{\textcircled{i}} = \begin{pmatrix} H^{(1)}_{\textcircled{i}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & H^{(d)}_{\textcircled{i}} \end{pmatrix}_{(Ld)\times(M_i d)}$ with $H^{(k)}_{\textcircled{i}} = \left[h^{(k)}_{\Omega_i(1)} \dots h^{(k)}_{\Omega_i(M_i)}\right]$ being an $L \times M_i$ subset of the full matrix containing only the $\Omega_i(1)^{th}$ to $\Omega_i(M_i)^{th}$ columns, and $V_{q_{\textcircled{i}}} = \begin{pmatrix} v^{(1)}_{q_{\textcircled{i}}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & v^{(d)}_{q_{\textcircled{i}}} \end{pmatrix}_{(M_i d)\times(M_i d)}$ with $v^{(k)}_{q_{\textcircled{i}}} = diag\left(\sigma^{(k)}_{\Omega_i(1)c_{\Omega_i(1)}}, ..., \sigma^{(k)}_{\Omega_i(M_i)c_{\Omega_i(M_i)}}\right)$. When all the sources are active, the observation density in (IV.3) uses the full mixing matrix and when none of the sources are active, the observation density reduces to white Gaussian noise.

III.B.3 Complete/Undercomplete case

Log-likelihood

When there are equal or more sensors L than sources M ($L \geq M$), each observation point in the sensor space generated from a specific state of activity/inactivity is assumed to be independent from the next state in time, establishing a mixture model for the states (i.e. zero-order Markov model, see Section

III.B.4 for further discussion). By introducing an indicator function, $x_i(n)$, defined to be equal to unity when at time n it obeys state $\textcircled{i}$ and zero otherwise, the joint log-likelihood of the sensor observations and hidden variables (indicator variables) of N data points, $\left(X^N, Y^N\right) = \left(\left\{x(1), Y^{(1:d)}(1)\right\}, ..., \left\{x(N), Y^{(1:d)}(N)\right\}\right)$ can be written as

$$\log P\left(X^N, Y^N|\theta\right) = \sum_{n=1}^{N}\sum_{i=1}^{I} x_i(n) \log P_{\textcircled{i}}\left(Y^{(1:d)}(n)|\theta\right) + x_i(n) \log \pi_{\textcircled{i}}(\theta) \tag{III.15}$$

where θ is the collection of all the unknown parameters, consisting of the mixing matrices, the mixing coefficients of the states ($\pi_{\textcircled{i}}$, $i = 1, ...I$) and the noise covariance matrix. Notice that the number of parameters in this model have not changed compared to the previous section. However, the mixing matrices have been broken down into partitions where each will be learned in a more controlled and specialized manner.

EM Parameter Estimation

The Expectation Maximization (EM) algorithm guarantees to hill-climb the likelihood of observations by taking the expectation of (III.15) with respect to the hidden variables conditioned on the observations and the last update of parameters from the maximization step, indicated as $Q(\theta, \hat{\theta})$ [24]. After some manipulation the E-step becomes

$$\hat{x}_i(n) = \frac{P_{\textcircled{i}}\left(Y^{(1:d)}(n)|\hat{\theta}\right)\pi_{\textcircled{i}}(\hat{\theta})}{\sum_{j=1}^{I} P_{\textcircled{j}}\left(Y^{(1:d)}(n)|\hat{\theta}\right)\pi_{\textcircled{j}}(\hat{\theta})} \tag{III.16}$$

The M-step includes updating the mixture coefficients as

$$\pi^{+}_{\textcircled{i}}(\theta) = \frac{\sum_{n=1}^{N}\hat{x}_i(n)}{N} \tag{III.17}$$

and taking a couple of steps in the gradient direction of the mixing matrices and the noise covariance

$$\nabla_{H^{(k)}} Q(\theta, \hat{\theta}) = \sum_{n=1}^{N} \sum_{i=1}^{I} \hat{x}_i(n) \frac{\left(\frac{\partial}{\partial H^{(k)}_{(i)}} P_{(i)} \left(Y^{(1:d)}(n) \right) \right)^*}{P_{(i)} \left(Y^{(1:d)}(n) \right)} \tag{III.18}$$

$$\nabla_{\sigma_W} Q(\theta, \hat{\theta}) = \sum_{n=1}^{N} \sum_{i=1}^{I} \hat{x}_i(n) \frac{\left(\frac{\partial}{\partial \sigma_W} P_{(i)} \left(Y^{(1:d)}(n) \right) \right)^*}{P_{(i)} \left(Y^{(1:d)}(n) \right)}. \tag{III.19}$$

The derivation of the numerators on the RHS of Eqs. III.18 and III.19 are shown in Appendix III.G.1.

III.B.4 Overcomplete case

Hidden Markov Model

In the overcomplete case $M > L$, since the distribution of the data in the sensor space is lower in dimension than the source space, data points belonging to different states of activity can be overlapping. To illustrate such overlapping, Figure III.1 gives an example of the empirical distribution in the sensor space for an overcomplete representation of 3 sources using 2 sensors such that each point is color-coded to represent the ground truth state of activity. In order to compensate for this overlapping, a first-order Markovian state structure is incorporated using HMMs, enabling us to make use of the temporal dependencies and estimate the states more accurately compared to the mixture model employed for the complete/undercomplete case provided earlier. In order to assure smooth transitions between the states, a non-ergodic HMM is used which assumes that at each new time instant, at most one source can appear or disappear. The HMM transition diagram is depicted in Figure III.2 for the example of $M = 3$. It is clear that for complete/undercomplete case discussed earlier, a similar first-order Markovian structure can be used instead of the zero-order mixture model. However, our experiments show that for this case the Markovian property does not give us an extra

advantage and the simpler mixture model is sufficient to find the correct state estimates. This is naturally due to the fact that as the problem is upgraded to a complete/undercomplete setting, the extra dimension(s) that is(are) added to the sensor space would reduce the overlapping of the states. On the other hand, for the overcomplete case the zero-order mixture model can also be utilized, however, due to the mixture model's discriminative way of state estimation (classification), the overlap between the states is not taken into consideration resulting in a poor state estimation.

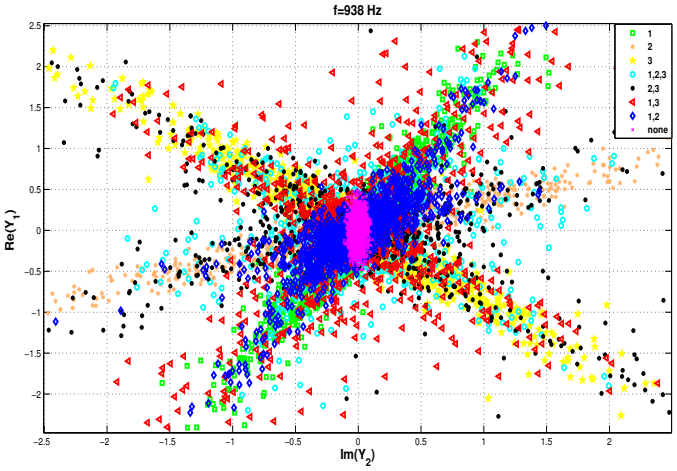

Figure III.1 Data in the sensor space of $f = 938$ Hz(after whitening) for an overcomplete representation of 3 sources using 2 sensors with ground truth states of activity.

HMM EM parameter estimation

Again, EM algorithm is used to learn the HMM initial probabilities π_i, the HMM transition probabilities $a_{ij} = P\left(x(n) = i|x(n-1) = j\right)$, the mixing matrices and noise covariance [73]. The E-step consists of finding the probability $\gamma_n(i) = P\left(x(n) = i|Y^{(1:d)}(1), ..., Y^{(1:d)}(N)\right)$ from the forward/backward probabilities

$$\alpha_n(i) = P\left(Y^{(1:d)}(1), ..., Y^{(1:d)}(n), x(n) = i\right) \tag{III.20}$$

and

$$\beta_n(i) = P\left(Y^{(1:d)}(n+1), ..., Y^{(1:d)}(N) | x(n) = i\right), \tag{III.21}$$

using the relation

$$\gamma_n(i) = \alpha_n(i)\beta_n(i) / \sum_{j=1}^{I} \alpha_n(j)\beta_n(j) \tag{III.22}$$

and the forward/backward recursions of

$$\begin{aligned} \alpha_n(i) =& P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right) \sum_{j=1}^{I} a_{ij}\alpha_{n-1}(j) \\ \beta_n(i) =& \sum_{j=1}^{I} P_{\textcircled{j}}\left(Y^{(1:d)}(n+1)\right) a_{ji}\beta_{n+1}(j). \end{aligned} \tag{III.23}$$

with initial values

$$\begin{aligned} \alpha_1(i) =& \pi_i P_{\textcircled{i}}\left(Y^{(1:d)}(1)\right) \\ \beta_N(i) =& 1 \ , \quad i = 1, ..., I. \end{aligned} \tag{III.24}$$

The M-Step consists of updating the initial and transition probabilities as

$$\hat{\pi}_i^+ = \alpha_1(i)\beta_1(i) / \sum_{j=1}^{I} \alpha_1(j)\beta_1(j) \tag{III.25}$$

$$\hat{a}_{ij}^+ = \frac{\sum_{n=2}^{N} a_{ij}\alpha_{n-1}(j)\beta_n(i) P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)}{\sum_{n=2}^{N} \alpha_{n-1}(j)\beta_{n-1}(j)} \tag{III.26}$$

and taking a couple of steps along the gradient of the auxiliary Q function with respect to the mixing matrices and the noise covariance

$$\nabla_{H^{(k)}} Q(\theta, \hat{\theta}) = \sum_{n=1}^{N}\sum_{i=1}^{I} \gamma_n(i) \frac{\left(\frac{\partial}{\partial H^{(k)}_{\textcircled{i}}} P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)\right)^*}{P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)} \tag{III.27}$$

$$\nabla_{\sigma_W} Q(\theta, \hat{\theta}) = \sum_{n=1}^{N}\sum_{i=1}^{I} \gamma_n(i) \frac{\left(\frac{\partial}{\partial \sigma_W} P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)\right)^*}{P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)} \tag{III.28}$$

The entries in the numerators of (III.27) and (III.28) are found the same way as for the well-determined case (see Appendix III.G.1).

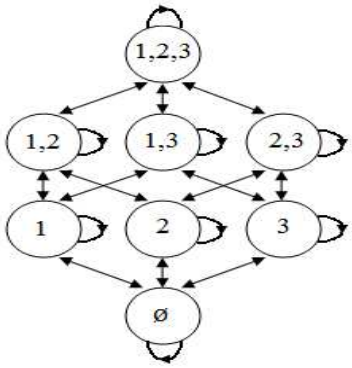

Figure III.2 State transition diagram for $M = 3$ assuming that at most one source can appear or disappear at a time

III.B.5 Source Reconstruction

Once the parameters have been estimated (denoted as $\hat{H}^{(1:d)}$ and $\hat{\Sigma}_W$), we reconstruct the signals using the MMSE estimator through Bayesian inference.

$$\begin{aligned}\hat{S}^{(1:d)}(n) =& E\left[S^{(1:d)}(n)|Y^{(1:d)}(n)\right] \\ =& \sum_{i=1}^{I} \hat{z}_i^{++}(n) E_{\textcircled{i}}\left[S^{(1:d)}(n)|Y^{(1:d)}(n)\right]\end{aligned} \tag{III.29}$$

where $\hat{z}_i^{++}(n)$ is the soft indicator function obtained from the last iteration (converged) of the E-step described as [3]

$$\hat{z}_i^{++}(n) = \begin{cases} \hat{x}_i^{++}(n) & undercomplete/complete \quad L \geq M \\ \gamma_n^{++}(i) & overcomplete \quad L < M \end{cases} \tag{III.30}$$

and

$$E_{\textcircled{i}}\left[S_{\Psi}^{(1:d)}(n)|Y^{(1:d)}(n)\right] = \begin{cases} 0 & \Psi = \Omega - \Omega_i \\ \sum_{q_{\textcircled{i}}} \lambda_{q_{\textcircled{i}}}(n) \Lambda_{q_{\textcircled{i}}}^{(1:d)} \hat{H}_{\textcircled{i}}^{(1:d)^H} \hat{\Sigma}_W^{-1} Y^{(1:d)}(n) & \Psi = \Omega_i \end{cases} \tag{III.31}$$

[3] the superscript ++ denotes that it comes from the last iteration of the E-step

where $\Lambda_{q_{\textcircled{i}}}^{(1:d)} = \left(\hat{H}_{\textcircled{i}}^{(1:d)^H} \hat{\Sigma}_W^{-1} \hat{H}_{\textcircled{i}}^{(1:d)} + {V_q}_{\textcircled{i}}^{-1} \right)^{-1}$ and

$$\lambda_{q_{\textcircled{i}}}(n) = \frac{{w_q}_{\textcircled{i}} \mathcal{N}\left(Y^{(1:d)}(n); 0, \hat{A}_{q_{\textcircled{i}}}^{(1:d)} \right)}{\sum_{q'_{\textcircled{i}}} {w_{q'}}_{\textcircled{i}} \mathcal{N}\left(Y^{(1:d)}(n); 0, \hat{A}_{q'_{\textcircled{i}}}^{(1:d)} \right)}.$$

III.C Pre/Post-Processing

III.C.1 Pre-Processing

Whitening

Prior to learning the mixing matrices, whitening is done on each frequency separately, making it easier for the algorithm to converge to a solution. Because the whitening matrix for each frequency bin is different, the noise covariances are scaled differently from one frequency bin to another. Assuming that the $L \times L$ whitening matrix for bin k is $D^{(k)}$, the noise covariance for bin k after whitening becomes $D^{(k)} \sigma_W {D^{(k)}}^H$. Therefore, some minor modifications need to be made to the gradients in the M-Step to ensure that the noise covariance is scaled properly. The GMM parameters used to model the sources were learned by fitting a spherical multivariate GMM (Eq. III.12) with 3 mixture components, to a 20-min-long continuous speech with no prolonged silence periods and normalized to unit variance speech for each frequency bin. The speech is normalized to unit variance for each frequency bin separately because whitening is preformed on the sensor data for each frequency bin separately as well. Doing so, also, makes the distribution closer to the spherical representation in Eq. III.12.

Initialization of the Mixing Matrices using a Sparser Model

For the overcomplete case where the estimation problem becomes a harder task and more sensitive to initial values of the mixing matrices, simpler and sparser intermediate models can be used to create good initial values to be used in the

proposed EM algorithm that uses the full model (shown in Figure III.2 for $M = 3$). For example, one can start with the sparsest model which assumes that at each time at most one source can be active, and after some iterations, slowly advance to intermediate sparse models that allow more simultaneously active sources. The state dynamic diagram for such a progressive model is illustrated in Figure III.3 for $M = 3$. Running the algorithm using such sparse models as a pre-processing step would attempt to find the star-like legs associated with the columns of the mixing matrices (as seen in Figure III.1) without caring about their overlap when two or more sources are active. This estimate of the mixing matrix is a good initialization for learning the mixing matrix and state probabilities using the full dynamic model which eventually leads to better estimates and faster convergence. This initialization technique is somewhat similar to the bottom-up hierarchical clustering method used in [87] to estimate the mixing matrices, but just like our proposed algorithm, it is done on the vector of frequency bins to significantly reduce permutations in the columns of mixing matrices from one bin to another. In our experiments we use such a technique to initialize the mixing matrices for difficult cases for example when we have 2 sensors and 4 sources (Experiment C in Section III.D.3).

III.C.2 Post-Processing

Adjusting Scales and the Inverse Fourier Transform

One indeterminacy in BSS is that the sources can be multiplied by an arbitrary scalar without violating the underlying assumption. As a consequence, the scaling problem needs to be solved in the frequency bins either by adjusting the source variances or by scaling the estimated mixing matrices. Since the sources are dynamic with varying variances, it would be simpler to scale the estimated mixing matrices using the well-known minimal distortion principle [54] in each frequency bin prior to source reconstruction. After the sources have been reconstructed using the MMSE estimator described in Section III.B.5, the inverse Fourier transform

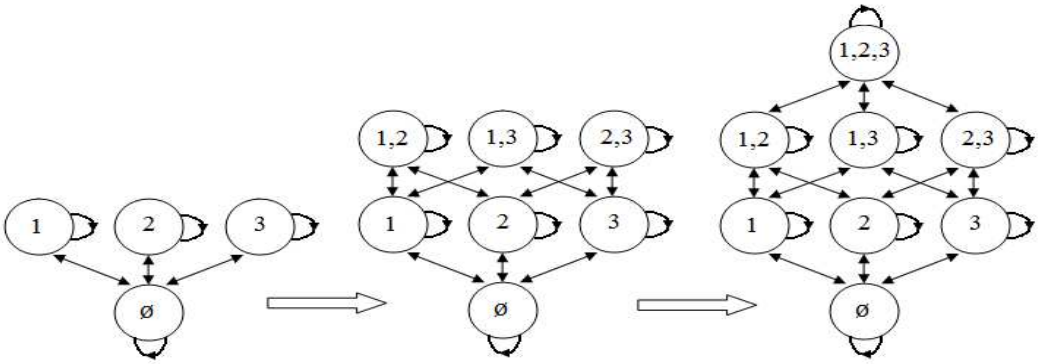

Figure III.3 Bottom-up progressive model for $M = 3$. It starts with the sparsest representation assuming that at each time at most one source can be active (left), then advances to an intermediate case where at most two can be active simultaneously (middle). Finally, the full model is used allowing up to three simultaneously active sources (right). The mixing matrix estimates of each step is used as the initial values for the next step.

using the overlap add method is used to reconstruct the time domain signals.

Glimpsing across Frequency bins

So far our proposed algorithm was based on "glimpsing in time" or taking advantage of the different combination of silence gaps on the local temporal level where the problem could be less degenerate. This means that our estimated states of activity are the same for all frequency bins. However, in reality, when a dynamic signal like speech is active in a time frame, it is not necessarily active across all frequency bins of the same time frame in the spectrogram. Obviously, when the signal is inactive in a time frame, it is also inactive across all frequency bins in that time frame. This means that sparsity in time ("glimpsing in time") comes before sparsity in frequency domain ("glimpsing in frequency"). Therefore if one wants to exploit sparsity in the frequency bins, a rerun of the algorithm can be done for each frequency bin separately as a post-processing step using the estimated parameters from our proposed algorithm as initial values. One can also restrict

Algorithm 1 Glimpsing IVA + Glimpsing in Frequency Post-Processing

Glimpsing IVA: Perform G-IVA described in Section III.B.3/III.B.4 to obtain $\hat{H}^{(k)},\ k = 1, ..., d$ and $\hat{z}_i^{++}(n),\ i = 1, .., I,\ n = 1, .., N$

Glimpsing in Frequency:

for $k = \{1, \ldots, d\}$ **do**

Perform glimpsing algorithm described in Section III.B.3/III.B.4 for each frequency dimension separately using $\hat{H}^{(k)}$ as initial conditions and obtain updates $\hat{H}_{post}^{(k)}$ and $\hat{z}_i^{(k)^{++}}(n),\ i = 1, .., I,\ n = 1, .., N$

if $\hat{z}_i^{(k)^{++}}(n) \leq \hat{z}_i^{++}(n)$ **then**

$\hat{z}_i^{(k)^{++}}(n) \leftarrow \hat{z}_i^{(k)^{++}}(n)$

else

$\hat{z}_i^{(k)^{++}}(n) \leftarrow \hat{z}_i^{++}(n)$

end if

Reconstruct the sources in each bin

end for

Permutation Correction: Use the method in [75] with an option to choose $v_j^{(k)}(n)$ =prob. of source activity in bin k obtained from $\hat{z}_i^{(k)^{++}}(n)$

the corresponding bin-wise-state probabilities at a time-frequency block $\hat{z}_i^{(k)}(n)$ to be less than or equal to the converged state probabilities $\hat{z}_i^{++}(n)$ obtained from the main approach $\left(\hat{z}_i^{(k)}(n) \leq \hat{z}_i^{++}(n)\right)$. This ensures that the states are not declared active for a time-frequency block when it is declared inactive at that time frame. Our experiments show that even though such post-processing is done on each frequency bin separately, little permutation of the sources for different frequency bins takes place which is due to using estimated matrices from the proposed IVA method as initial conditions for the bin-wise rerun of the algorithm. To correct for the permutation that might exist, we use the recent and effective method in [75]. The pseudo-code in Algorithm 1 displays the steps taken for the "glimpsing in frequency" post-processing step. This post-processing method also has a denoising effect which suppresses the noise present in the areas of the spectrogram of the sources where no time-frequency activity is present.

III.D Experimental results

In this section we perform some experiments using real and simulated data. Simulated data was created using the image method in [2] which simulates the impulse response between a source and a sensor for a rectangular room environment. We evaluate the performance for both well-determined complete/undercomplete and ill-determined overcomplete cases. However, since the overcomplete case is more difficult and less straightforward, we will focus most of our experiments on the overcomplete case. For the complete case ($M = L$), the proposed glimpsing IVA algorithm (denoted as G-IVA) is compared to the well-known IVA algorithm [36]. For the overcomplete case, the proposed algorithm is compared to the time-frequency masking algorithm of Sawada et al. [76]. This algorithm uses the clustering along oriented lines method in [61] in each frequency bin which permits only one frequency be active at each time, and then uses the method in [75], which is a simpler and improved version of the method in [78], to effectively correct for permutations of sources in different frequency bins. The performances were evaluated using the signal to disturbance ratio (SDR) described as

$$SDR_{out} = 10\log\left(\frac{\sum_{n,k}\left|\sum_{i}\left(G_n^{(k)}\right)_{ii}S_i^{(k)}(n)\right|^2}{\sum_{n,k}\left|\sum_{i\neq j}\left(G_n^{(k)}\right)_{ij}S_j^{(k)}(n)+\sum_{i',j'}\left(\hat{R}^{(k)}(n)\right)_{i'j'}W_{j'}^{(k)}(n)\right|^2}\right) \tag{III.32}$$

where $G_n^{(k)} = \hat{R}^{(k)}(n)H^{(k)}$ and $\hat{R}^{(k)}(n)$ is the time-varying $M \times L$ reconstruction matrix obtained from the MMSE estimator for bin k and block n described in Section III.B.5. SDR_{out} is the total signal power of direct channels versus the signal power stemming from cross interference and noise combined, therefore giving a reasonable performance measurement for noisy situations. In addition to evaluation using SDR, for the overcomplete case, we also compare the machine intelligibility of the separated sources using a continuous speech recognizer.

We assumed a room size of 8x5x3.5 m with the microphones and the sources having the same height of 1.5 m. Experiments were carried out using different sources with different angles with respect to the microphones. Figure III.4 illustrates the simulated room setting along with the microphones used for each experiment. For all the experiments, we assumed a reverberation time of 200 ms. Each experiment was repeated for four different noise levels measured by the input signal to noise ratio (SNR_{in}) defined as

$$SNR_{in} = 10\log\left(\frac{\sum_{n,k}\left|\sum_{ij} H_{ij}^{(k)} S_j^{(k)}(n)\right|^2}{\sum_{n,k}\left|\sum_j W_j^{(k)}(n)\right|^2}\right). \tag{III.33}$$

To evaluate the performance improvement, a measurement for the input SDR of the convolutive mixture is needed. Since the contribution of each source in the mixture comes from each column of the mixing matrices (rather than the diagonal elements as seen in the output SDR of Eq. III.32), the input SDR needs to be calculated for each source separately based on the columns of the mixing matrices. Therefore, we define the average input SDR as follows

$$SDR_{in} = \frac{1}{M}\sum_{i=1}^{M} 10\log\left(\frac{\sum_{n,k}||h_i^{(k)} S_i^{(k)}(n)||^2}{\sum_{n,k}||\sum_{j\neq i} h_j^{(k)} S_j^{(k)}(n) + W^{(k)}(n)||^2}\right) \tag{III.34}$$

where $||.||$ indicates the vector 2-norm and $h_j^{(k)}$ is the j^{th} column of matrix $H^{(k)}$. A 512-point DFT with a STFT window length of 512 with 75% overlap is used at a sampling rate of 8kHz. The stopping rule for the algorithms was when the log-likelihood of the ratio between the increase in the log-likelihood over the previous value of the log-likelihood did not increase by 10^{-4}. Real data was gathered in a lab/conference room, where loudspeakers were placed on a table in front of the pair of microphones about 1 m away and each playing a female speech signal.

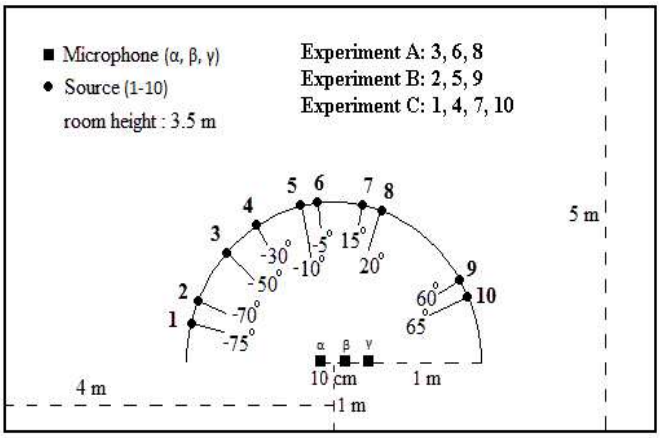

Figure III.4 Simulated room setup. The heights of the microphones and sources are 1.5 m. Three different experiments (A,B,C) using different combinations of sources 1-10 were carried out for the overcomplete case using two microphones(α and β). For the complete case two experiments (A,B) were carried out using three microphones (α, β and γ). Each experiment was repeated for four different noise levels. Experiments A and B have all female speech sources while experiment C has one male and three female sources. Reverberation time for all experiments was 200 ms.

III.D.1 Complete case

Experiments A (sources 3,6,8) and B (sources 2,5,9) in Figure III.4 were performed using the three microphones (α, β, γ). Both Experiments A and B have female voices for all the sources. The evaluation for Experiments A and B are shown in Figure III.5, where the performance of proposed G-IVA for the complete case $M = L = 3$, denoted as G-IVA 3x3, along with the performance after post-processing using glimpsing across frequency bins described in Section III.C.2 , is compared to the performance of the regular IVA method. The SDR_{in} is also given to illustrate the SDR improvement. These panels show that the performance of the proposed algorithm is higher than that of standard IVA, even at the highest SNR_{in}. The advantage of the proposed algorithm is most likely

due to two factors. One is that it exploits the silent regions in the sources to learn the mixing matrices in a more specialized fashion, therefore, resulting in a higher SDR_{out} for even high SNR_{in}. The other is that the proposed algorithm models noise and learns its level, whereas IVA does not. That is why IVA degrades more rapidly for low SNR_{in} compared to G-IVA. Figure III.5 also demonstrates that glimpsing in frequency post-processing boosts the performance of G-IVA. This is mainly due to the de-noising effect that glimpsing in frequency has and listening to the separation results before and after the post-processing verifies this de-noising effect. The advantage of G-IVA over regular IVA comes with a computational cost. The G-IVA 3x3 algorithm was coded in C and run on an Intel 2.5 GHz processor with 4GB RAM with an average computation time of around 4.6 minutes (around 1.2 sec per iteration for 230 iterations). The IVA algorithm was coded in Matlab (in an efficient matrix form structure to reduce computation time) with an average computation time of around 1 min (around 0.24 sec per iteration for 250 iterations).

III.D.2 Unknown Number of Sources using a Complete Setting

In BSS approaches for real world problems, it is usually the case that the total number of sources are unknown. One common approach that is used to deal with such an issue is to assume a large enough number of sources, hoping that the assumed number of sources would be larger than the actual number of sources. Because G-IVA seeks the active and inactive periods of the sources, we expect that the redundant sources be estimated as completely inactive for all times. To explore this situation we set up an example where we assume $M = L = 3$, however, with the actual number of sources being equal to 2. The sources are located in positions 3 and 6 in Figure III.4 using all three microphones with an SNR_{in}=16(dB). The separated sources are shown in Figure III.6 using G-IVA and regular IVA. G-IVA is able to successfully zero out the third redundant source while IVA still outputs some residue from the noise.

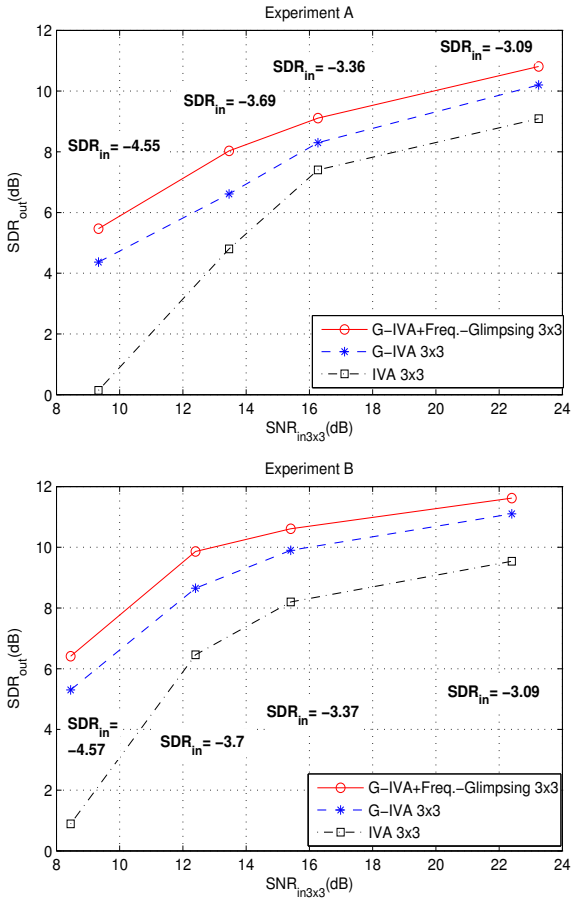

Figure III.5 Performance evaluation for the complete case. Top: Experiment A. Bottom: Experiment B

III.D.3 Overcomplete case

For the ill-determined overcomplete case, Experiments A and B carried out earlier for the complete case are repeated now using only two microphones (α and β). A more difficult setup of four sources in Experiment C (sources 1, 4, 7, 10) using only two microphones (α and β) is also carried out. Overcomplete G-IVA is employed as well as the glimpsing in frequency as a post-processing step. Their performances are then compared to the time-frequency masking method of Sawada et al.. The overcomplete G-IVA algorithm was coded in C and for Experiment A took an average computation convergence time of around 5 minutes (around

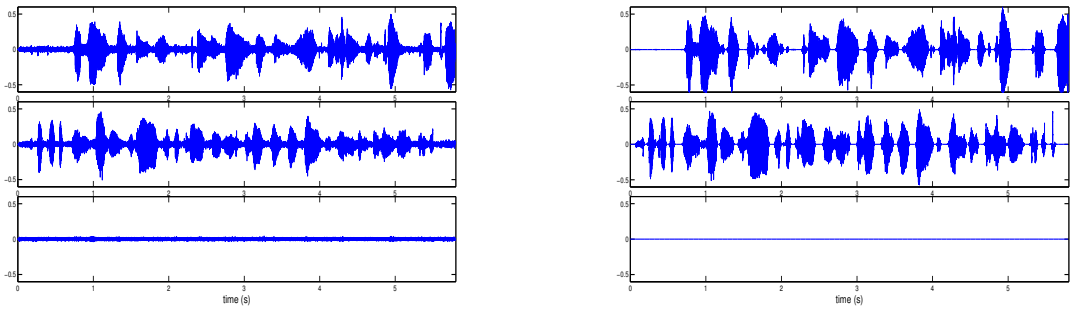

Figure III.6 Case of unknown number of sources. It was assumed that $M = 3$ where the number of sources was actually equal to 2. Left: separated signals using IVA. Right: separated signals using G-IVA.

1.2 sec per iteration for 250 iterations). Sawada et al.'s time-frequency masking method which was implemented efficiently in Matlab took around 45 seconds in total (around 0.35 sec per iteration for all frequency bins combined for an average of 100 iterations per bin and about 10 sec for permutation correction) to converge. As an upper performance measure, cases where extra microphone(s) is(are) added to turn the problem into a complete problem is considered and separated using the complete mode of G-IVA . All these performances are illustrated in Figure III.7 for comparison. These plots show that G-IVA in general performs better than the time-frequency masking method of Sawada et al.. It can also be seen from Figure III.7 that the glimpsing across frequency post-processing increases the SDR of overcomplete G-IVA. However, when listening to the reconstructed sources after this post processing, some synthetic artifacts commonly known as musical noise is introduced due to its greedy de-noisng effect across frequencies (the same was true for the experiments of the complete case in Section III.D.1). Because Sawada et al.'s time-frequency masking method, is a bin-wise separation method similar to glimpsing across frequencies, it too possesses this musical noise after separation. Nevertheless, for even high SNR_{in} this musical noise seems to be still present, especially when using Sawada et al.'s time-frequency masking method for separation.

This is due to it being greedier than the glimpsing across frequencies method as it allows each time-frequency block to be active for only one source.

In order to investigate the effect of musical noise on machine intelligibility, we pass the separated signals using G-IVA (without any post-processing) and Sawada et al.'s time-frequency masking algorithms through an automatic speech recognizer. For our experiments we choose a very high SNR_{in} of around 30(dB) to minimize the effect of the input white noise on the reconstructed sources. In order to simplify training, we perform the recognition on a limited vocabulary set of digits 0-9. Cambridge university hidden Markov toolkit (HTK) is used to train the recognizer. A batch of 36 male English speakers each uttering the digits 0-9 is utilized for the training. A batch of some other 20 male speakers make up the sources to be separated and tested. Each source comprises of two speakers uttering a total of 20 digits in a random order with random silence between each utterances. The average length of the sources in all the experiments was about 11 sec with a sampling rate of 8 kHz. Figure III.8 illustrates the digit error percentage for 10 experiments where each experiment corresponds to a configuration of different source angles. Each experiment is repeated twice with different speakers and the error rate shown in each bar is the average value of the two. Sources were mixed in the simulated room in Figure III.4 with a reverberation time of 200 ms, microphone spacing of 10cm and distance of sources to microphone of 1.5 m. Figure III.8 demonstrates that G-IVA has less recognition errors in all the experiments compared to Sawada et al.'s time-frequency masking. Since, the speech recognizer was trained on clean data, a higher recognition of the former algorithm indicates it has less interference and/or artifacts such as musical noise compared to the latter algorithm.

Figure III.9 shows the true and recovered sources along with the estimated probability of each source being active for the overcomplete case of Experiment A and $SNR_{in} = 11.3(dB)$. The probability of source m being active for frame n can be found by adding the estimated states $\gamma_n^{++}(i)$ that correspond to inclusion

of matrix column m. Also, the estimated state probabilities $\gamma_n^{++}(i)$ as well as the local SDR_{out} for each frame are shown next to the true sources in Figure III.10. Figure III.11 and Figure III.12 illustrate the same information for the harder case of Experiment C with $SNR_{in} = 21.7$. These figures show that the proposed algorithm is able to reconstruct the sources successfully and effectively detect the silence gaps by incorporating the best model based on the different combinations of silence gaps. Finally, we recorded real data in an ordinary lab/conference room setting. The sources consisted of three loudspeakers positioned on a table about 1m away from the two microphones. The sources were also recorded separately by one of the microphones when played one at a time, and synchronized with the original recording. This was done in order to create a perceptual comparison measure. The separation results yielding good perceptual separation are presented in Figure III.13. Furthermore, the estimated state probabilities $\gamma_n^{++}(i)$ are shown next to the sources in Figure III.14. These audio files along with more information are available at our website [4].

III.E Summary and Discussion

We have proposed a novel approach that can solve for the intricate overcomplete convolutive BSS as an extension to the more straight-forward complete/undercomplete case, using a unifying framework that incorporates the temporal structure of silent gaps present in many dynamic signals, especially speech. Our proposed method extends the main concept behind IVA which exploits the inner-frequency dependencies of each source while maintaining the same underlying assumption of independence from one source to another, therefore significantly reducing the occurrence of wrong permutations. By mimicking the separation strategy of the human hearing system, this algorithm is able to exploit the local decrease of degeneracy during the different combinations of silent gaps of the sources allowing it to cover all possible states from when all sources are active to when only

[4] http://dsp.ucsd.edu/∼ali/glimpsing/

one is active at each instant, therefore doing its best to compensate for the apparent global degeneracy. The algorithm works naturally by learning the columns of the mixing matrices in a specialized fashion based on the probability of being in each state and reconstructs the sources using an efficient and optimal (in the mean square sense) MMSE estimator incorporating the converged state estimates. The algorithm was able to outperform IVA in the classical complete/undercomplete cases of convolutive BSS (albeit with longer computation times), especially in environments with high noise levels (due to it having the extra feature of modeling additive noise). Furthermore, for the more challenging overcomplete case, improved separation results were achieved compared to a robust sparsity-based time-frequency masking method, using both SDR and machine intelligibility of a speech recognizer as the performance measurements. The hard on-off switching of the source activities is a good benefit for automatic speech recognition systems since it avoids wrong insertions due to residual interfering noise. On the other hand, if the BSS system is intended for human listeners the on-off switching effect could make the speech sound choppy and perceptually undesirable, hence solutions to this issue is worth being investigated.

A drawback of the proposed algorithm is that the number of states, and along with it the computational cost, will grow exponentially as the number of sources increases. This intractability for large number of sources, of course, is not unique to G-IVA and is shared by other state-based models. For large number of sources, in general, approximations can be made to make it computationally tractable. One way is to reduce the maximum number of active sources (in each frequency bin separately or for all bins) based on the sparsity present in the signals. For example if there are $M = 10$ sources but the activity patterns present in the sources are sparse enough that, roughly speaking, at most 2 sources are active simultaneously, then by limiting the model to allow maximum of 2 sources active at each time, the number of permissible states reduces dramatically and the problem becomes somewhat tractable. Similar constraints on sparsity is used in the domain

of dictionary learning for sparse signal recovery [17]. In the general case, in order to ease the complex patterns possible in the state transitions for the overcomplete case, we have utilized a non-ergodic trellis that allows for at most one source to appear or disappear at each transition. This, to some limited degree, reduces the complexity as well. Other approximations using variational methods (as in [8]) might also be useful.

Another issue that deserves a discussion is the problem of unknown number of sources. The algorithm showed it has the ability of effectively zeroing out redundant sources in case the true number of sources is unknown. In order to do this an overkill strategy of using a large enough number of sources with equal number of sensors ($L = M$ complete setting) is assumed. The experiment that was carried out assumed an overkill of three sources and three sensors while the number of true sources was two. Similar experiments were also carried out for a more challenging problem of unknown sources using an overcomplete setting. For example, the true number of sources is two while we assume three sources but recorded using only two sensors. This problem becomes extremely hard and sensitive to initial values, and even intelligent initializations using the bottom-up progressive model that we proposed, does not often result in zeroing out the redundant source. This indicates that some refinements are needed to make the approach more robust for such situations.

III.F Acknowledgments

The text of Chapter III, in full, is based on the material as it appears in: Alireza Masnadi-Shirazi, Wenyi Zhang and Bhaskar Rao, "Glimpsing IVA: A Framework for Overcomplete/Complete/Undercomplete Convolutive Source Separation," IEEE Transactions on Audio, Speech and Language Processing, vol. 18, no. 7, Sept. 2010. The dissertation author was a primary researcher and an author of the cited material.

III.G Appendix

III.G.1 Derivation of the gradients

$$\frac{\partial P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)}{\partial H^{(k)}_{\textcircled{i}}} = \sum_{q_{\textcircled{i}}} w_{q_{\textcircled{i}}} G\left(Y^{(1:d)}(n), 0, A^{(1:d)}_{q_{\textcircled{i}}}\right) \times \rightarrow$$

$$\left[\frac{-0.5}{|A^{(k)}_{q_{\textcircled{i}}}|}\frac{\partial}{\partial H^{(k)}_{\textcircled{i}}}|A^{(k)}_{q_{\textcircled{i}}}| - \frac{0.5\partial}{\partial H^{(k)}_{\textcircled{i}}}\left(Y^{(k)^H}(n)A^{(k)^{-1}}_{q_{\textcircled{i}}}Y^{(k)}(n)\right)\right] \tag{III.35}$$

with $A^{(k)}_{q_{\textcircled{i}}} = \sigma_W + H^{(k)}_{\textcircled{i}} v^{(k)}_{q_{\textcircled{i}}} H^{(k)^H}_{\textcircled{i}}$. The entries of Eq. III.35 can be found by

$$\frac{-\partial}{\partial H^{(k)}_{\textcircled{i}_{ij}}}\left(Y^{(k)^H}(n)A^{(k)^{-1}}_{q_{\textcircled{i}}}Y^{(k)}(n)\right) =$$

$$\sum_{l,k}\left[\left(A^{(k)^{-T}}_{q_{\textcircled{i}}}Y^{(k)^*}(n)Y^{(k)^T}(n)A^{(k)^{-T}}_{q_{\textcircled{i}}}\right)_{lk}\frac{\partial}{\partial H^{(k)}_{\textcircled{i}_{ij}}}\left(A^{(k)}_{q_{\textcircled{i}}}\right)_{lk}\right] \tag{III.36}$$

and

$$\frac{\partial}{\partial H^{(k)}_{\textcircled{i}_{ij}}}|A^{(k)}_{q_{\textcircled{i}}}| = \sum_{l,k}\left[\left(|A^{(k)}_{q_{\textcircled{i}}}|A^{(k)^{-T}}_{q_{\textcircled{i}}}\right)_{lk}\frac{\partial}{\partial H^{(k)}_{\textcircled{i}_{ij}}}\left(A^{(k)}_{q_{\textcircled{i}}}\right)_{lk}\right] \tag{III.37}$$

where

$$\frac{\partial}{\partial vec(H^{(k)}_{\textcircled{i}})} vec\left(A^{(k)}_{q_{\textcircled{i}}}\right) = \left(H^{(k)^*}_{\textcircled{i}} v_{q_{\textcircled{i}}}\right) \otimes I_M \tag{III.38}$$

where $\otimes$, *vec* and $|.|$ stand for Kronecker product, column-wise vectorization and absolute value of the determinant, respectively. Similarly, the gradient with respect to the noise covariance is

$$\frac{\partial P_{\textcircled{i}}\left(Y^{(1:d)}(n)\right)}{\partial \sigma_W} = \sum_{q_{\textcircled{i}}} w_{q_{\textcircled{i}}} G\left(Y^{(1:d)}(n), 0, A^{(1:d)}_{q_{\textcircled{i}}}\right) \times \rightarrow$$

$$\left[\frac{-0.5}{|A^{(k)}_{q_{\textcircled{i}}}|}\frac{\partial}{\partial \sigma_W}|A^{(k)}_{q_{\textcircled{i}}}| - \frac{0.5\partial}{\partial \sigma_W}\left(Y^{(k)^H}(n)A^{(k)^{-1}}_{q_{\textcircled{i}}}Y^{(k)}(n)\right)\right] \tag{III.39}$$

$$\frac{\partial}{\partial \sigma_{W_{ij}}} \left({Y^{(k)}}^H(n) {A_q^{(k)}}^{-1} Y^{(k)}(n) \right) =$$
$$\sum_{l,k} \left[- \left({A_q^{(k)}}^{-T} {Y^{(k)}}^*(n) {Y^{(k)}}^T(n) {A_q^{(k)}}^{-T} \right)_{lk} \frac{\partial}{\partial \sigma_{W_{ij}}} \left(A_q^{(k)} \right)_{lk} \right] \tag{III.40}$$

and

$$\frac{\partial}{\partial \sigma_{W_{ij}}} \left| A^{(k)} \right| = \sum_{l,k} \left[\left(|A_q^{(k)}| {A_q^{(k)}}^{-T} \right)_{lk} \frac{\partial}{\partial \sigma_{W_{ij}}} \left(A_q^{(k)} \right)_{lk} \right] \tag{III.41}$$

and

$$\frac{\partial}{\partial \sigma_{W_{ij}}} \left(A_q^{(k)} \right)_{lk} = \begin{cases} 0 & i \neq j \\ 1 & i = j = l = k \\ 0 & i = j, l \neq k \end{cases} \tag{III.42}$$

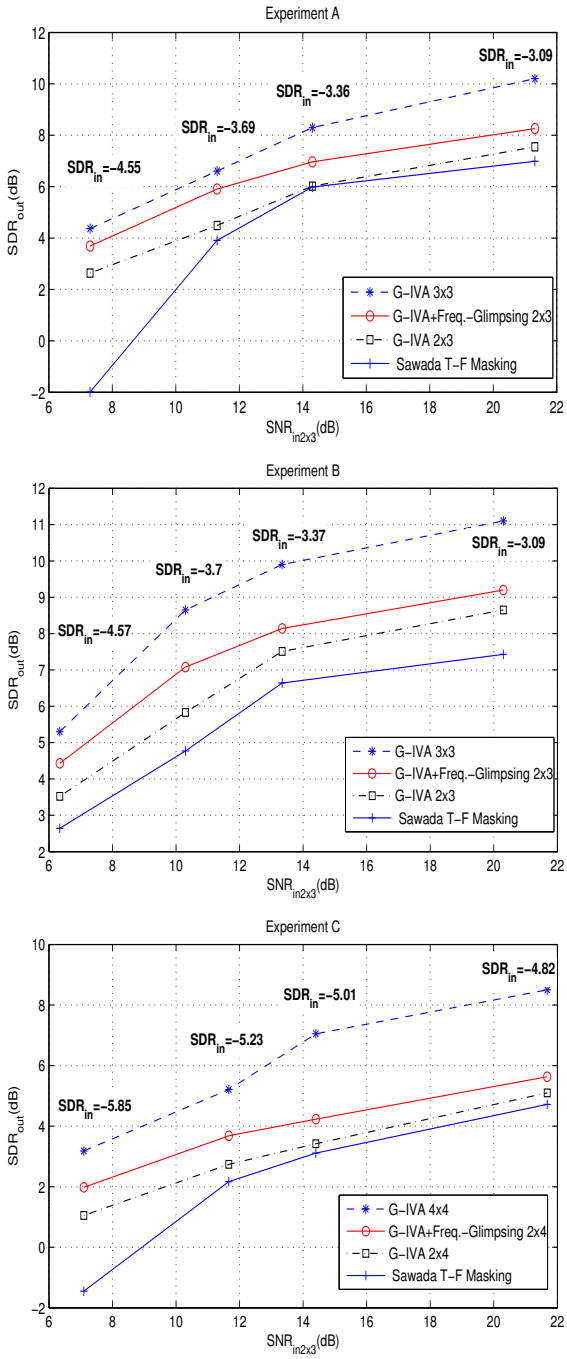

Figure III.7 Performance evaluation for the overcomplete case. Top: Experiment A. Middle: Experiment B. Bottom: Experiment C

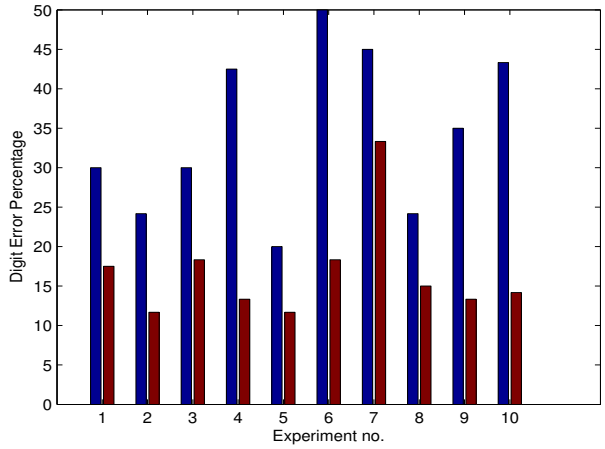

Figure III.8 Digit error percentage of separated sources in an overcomplete setting of three speakers and two microphones using a continuous speech recognizer. The left bar is the error rate after separating using Sawada's time-frequency masking algorithm and the right bar is the error rate after separating using G-IVA algorithm. Each source comprises of two speakers uttering a total of 20 digits in a random order with random silence between each utterances. The average length of the sources in all the experiments is about 11 sec with a sampling rate of 8 kHz. Each experiment is repeated twice with different speakers and the error rate shown in each bar is the average value of the two. The Sources were mixed in the simulated room in Figure III.4 with a reverberation time of 200 ms, microphone spacing of 10 cm and distance of sources to microphone of 1.5 m. The error percentage of the original sources before mixing was around %1. Each experiment refers to a different configuration of the sources with respect to the vertical centerline between the microphones. 1: $[-50^o\ 5^o\ 20^o]$; 2: $[-55^o\ -5^o\ 45^o]$; 3: $[-60^o\ 0^o\ 25^o]$; 4: $[-45^o\ -20^o\ 5^o]$; 5: $[-10^o\ 10^o\ 30^o]$; 6: $[-50^o\ -20^o\ 0^o]$; 7: $[-10^o\ 5^o\ 20^o]$; 8: $[-45^o\ 2^o\ 45^o]$; 9: $[-60^o\ 5^o\ 40^o]$; 10: $[-50^o\ -25^o\ 40^o]$.

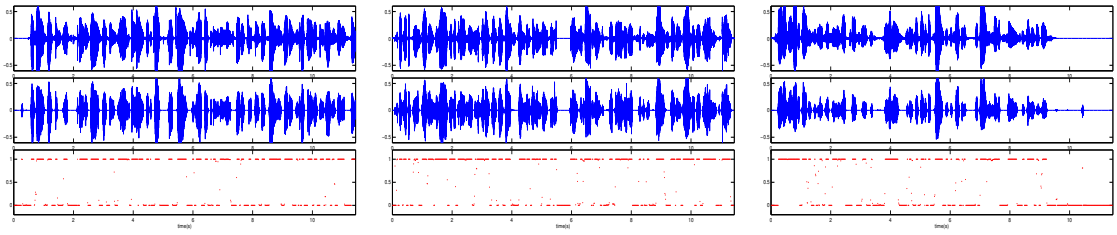

Figure III.9 Experimental results of a simulated room mixing of 3 sources using 2 microphones (Experiment A, $SNR_{in} = 11.3dB$). Top: true sources. Middle: separated sources. Bottom: estimated probability of source activity

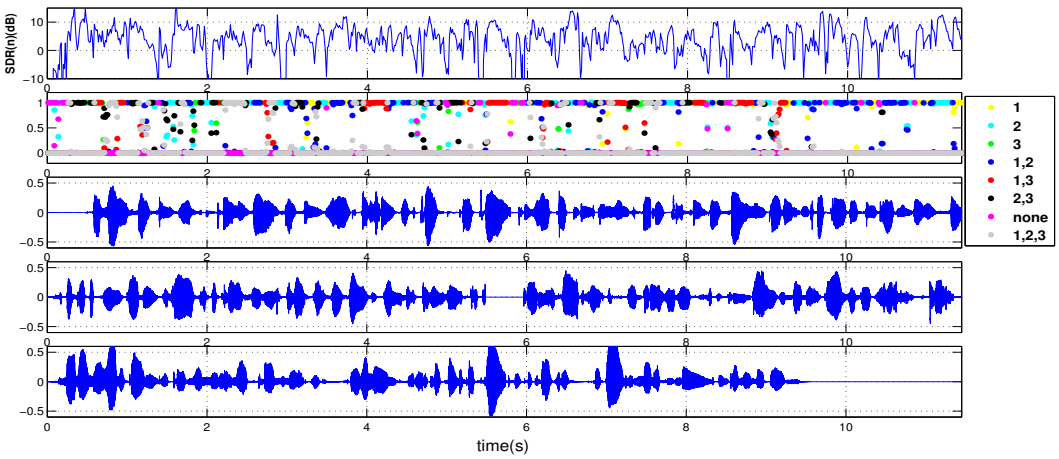

Figure III.10 Experimental results of a simulated room mixing of 3 sources using 2 microphones (Experiment A, $SNR_{in} = 11.3dB$). First row: local block-wise SDR_{out}. Second row: estimated state probabilities. Third to fifth row: true sources

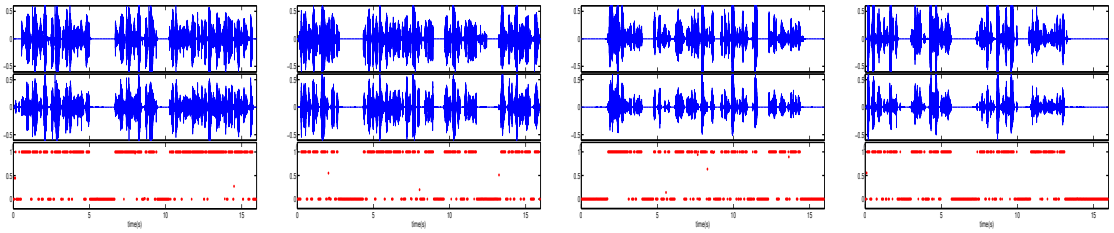

Figure III.11 Experimental results of a simulated room mixing of 4 sources using 2 microphones (Experiment C, $SNR_{in} = 21.7dB$). Top: true sources. Middle: separated sources. Bottom: estimated probabilities of source activity

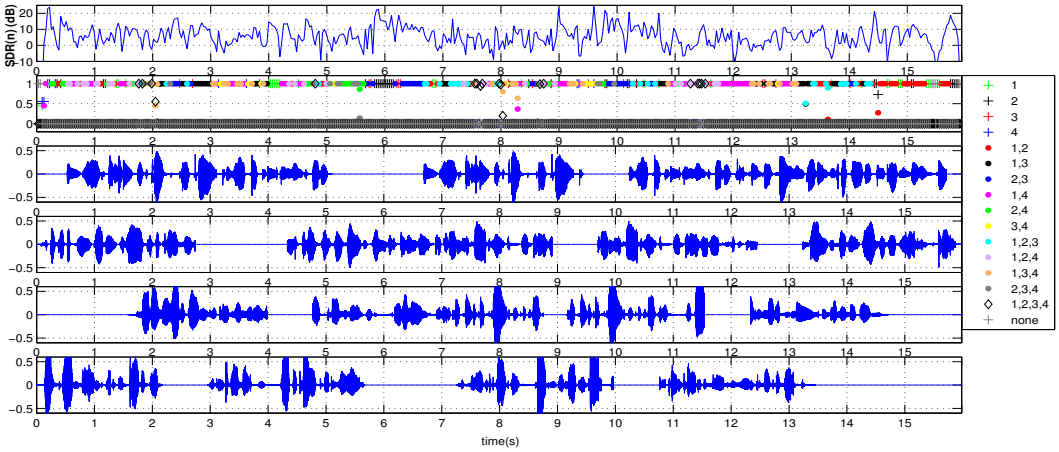

Figure III.12 Experimental results of a simulated room mixing of 4 sources using 2 microphones (Experiment C, $SNR_{in} = 21.7dB$). First row: local block-wise SDR_{out}. Second row: estimated state probabilities. Third to sixth row: true sources

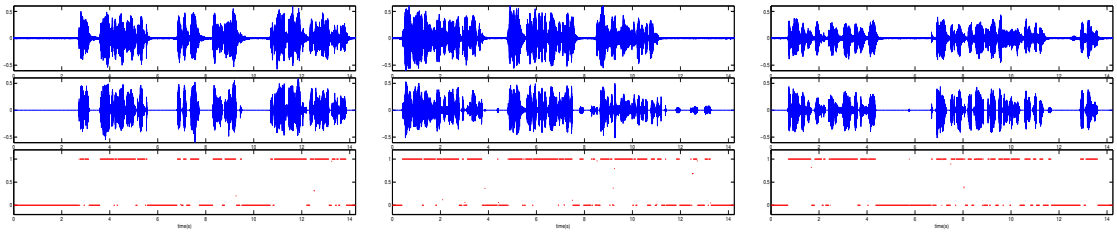

Figure III.13 Experimental results of real recording of 3 sources in a lab room using 2 microphones. Top: true sources (recorded separately). Middle: separated sources. Bottom: probability of source activity

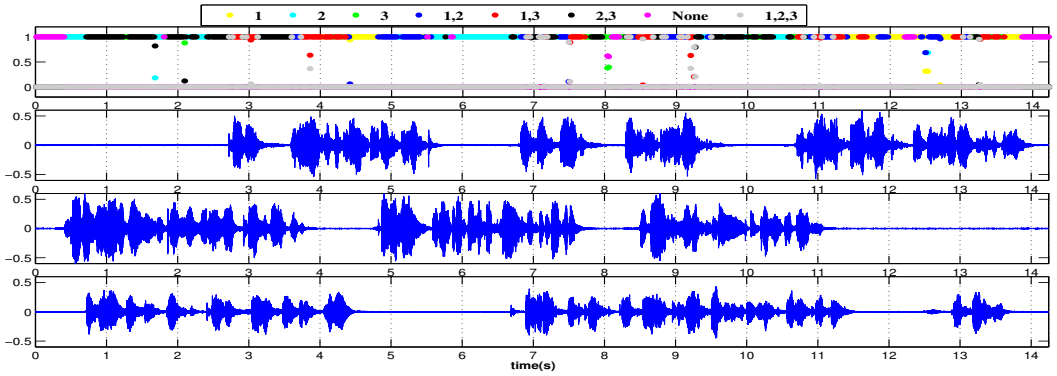

Figure III.14 Estimated state probabilities (top) with the true sources for a real room recording of 3 sources using 2 microphones

Chapter IV

Online Separation and Tracking of Known Number of Sources with Silence Periods

IV.A Introduction

Passive localization and tracking of multiple acoustical sources is of great interest in the field of microphone arrays. Antonacci et al. first proposed a method that uses a quasi-online time-domain-based convolutive ICA method that attempts to (partially) separate the speakers and from that estimate the time difference of arrival (TDOA) for each speaker using the extrema of the demixing filters. The same is repeated for different microphone arrays in different positions in the room resulting in multiple TDOAs for each source. The location of the sources is then estimated by triangulation of the TDOA hyperbola locus points. However, because the sources could be permuted for each array and because they might have not been fully separated, TDOAs from the arrays can't be matched to a specific source, resulting in false localizations. Therefore, more than one extra array (or auxiliary arrays using microphones within different array sets) is needed in order to reject such false positions. This localization method is accompanied by a particle filter on the source dynamics in order to track the sources [5, 81].

In this chapter we intend to both fully separate and track the multiple speakers using an online algorithm. Sawada et. al has shown that for static sources, estimating correct mixing matrices in the frequency domain using ICA can lead to accurate estimations of the directions of arrival (DOA) [79]. The assertion of this chapter is that for maneuvering sources, if one is able to track the mixing matrices accurately in the frequency domain to ensure full separation instead of partial separation, the accurate localization of the sources based on DOAs can be a straightforward consequence. This is in contrast to the aforementioned methods based on TDOA where the focus was on localization using already existing online BSS algorithms that might be able to only achieve partial separation. Moreover, similar to the previous chapter, we exploit a common form of temporal dynamics, especially present in speech, wherein the signals have silence periods intermittently, hence varying the set of active sources with time. By doing so we

enable the algorithm to "glimpse" or listen in the gaps [52]. Utilizing such glimpsing strategy is essential in an online algorithm because if a source becomes silent but assumed active by the model, the update to the column of the mixing matrix corresponding to that source can diverge or fluctuate unstably [35]. In this chapter we track the mixing matrices at each bin in the frequency domain by employing a multiple model particle filter (MMPF) method that is able to switch between the different combinations of silence gaps present in the sources. The proposed algorithm can also maintain track in the more challenging situation where the sources are silent but move around. We denote these periods as silence blind zones (SBZ). Moreover, in the localization phase using triangulation, because the sources can be fully separated, a correlation based method can be used to match the separated sources of each microphone array avoiding the problem of false positions due to permutations. Therefore, the need for and an auxiliary or more than one extra microphone array is eliminated.

IV.B Generative Model

We assume there are L microphones in the array and M sources. After taking the short time Fourier transform (STFT) (with d frequency bins) of the convolutedly mixed (due to reverberance) signals corrupted with white Gaussian noise, the observations would end up having a linear mixture in each frequency bin $k = 1, .., d$ described as

$$Y^{(k)}(t) = H^{(k)}(t)S^{(k)}(t) + W^{(k)}(t) \tag{IV.1}$$

where $H^{(k)}(t)$ is the $L \times M$ time varying mixing matrix for the k^{th} frequency bin. Since the noise is white it will have the same energy in all frequency bins. Hence the covariance of the noise can be written as $\sigma_W = diag(\sigma_{w_1}, ..., \sigma_{w_L})$. Throughout the rest of this chapter all operations are carried out in each individual frequency bin, therefore the superscript (k) is omitted for brevity. Each source is modeled as a multivariate GMM with C mixtures. The joint density of the sources is the product of the marginal densities, based on independency. Hence, we have

$$P_S(S) = \prod_{j=1}^{M} \sum_{c_j=1}^{C} \alpha_{j_{c_j}} G\left(S_j, 0, \sigma_{j_{c_j}}\right)$$
$$= \sum_{q=1}^{C^M} w_q G\left(S, 0, v_q\right) \tag{IV.2}$$

where $\sum_{q=1}^{C^M} = \sum_{c_1=1}^{C} ... \sum_{c_M=1}^{C}$, $w_q = \prod_{j=1}^{M} \alpha_{j_{c_j}}$ and $v_q = diag\left(\sigma_{1_{c_1}}, ..., \sigma_{M_{c_M}}\right)$. The parameters $\alpha_{j_{c_j}}$ and $\sigma_{j_{c_j}}$ for $j = 1, ..., M$ are fixed beforehand corresponding to a Gaussian mixture model (GMM) with zero means and varying variances (Gaussian scaled mixtures), hence having the shape of a symmetric multivariate super-Gaussian density [52]. Since each source can take on two states, either active or inactive, for M sources there will be a total of 2^M states. As a convention throughout this chapter we will encode the states by a number between 1 and $I = 2^M$ with a circle around it. Because we consider each bin separately, these states can vary across the frequency bins and indicate which column vector(s) of the mixing matrix is(are) present or absent for each bin.

Let the source indices form a set $\Omega = \{1, ..., M\}$, then any subset of Ω could correspond to a set of active source indices. For state $x(t) = \textcircled{i}$, we denote the subset of active indices in ascending order by $\Omega_i = \{\Omega_i(1), ..., \Omega_i(M_i)\} \subseteq \Omega$, where $M_i \leq M$ is the cardinality of Ω_i. It can be easily shown that the observation density function for state $\textcircled{i}$ is

$$P_{\textcircled{i}}(Y(t)|H(t)) = \sum_{q_{\textcircled{i}}} w_{q_{\textcircled{i}}} G\left(Y(t), 0, A_{q_{\textcircled{i}}}(t)\right) \tag{IV.3}$$

where $A_{q_{\textcircled{i}}}(t) = \sigma_W + H_{\textcircled{i}}(t) v_{q_{\textcircled{i}}} H_{\textcircled{i}}(t)^H$, $\sum_{q_{\textcircled{i}}} = \sum_{c_{\Omega_i(1)}=1}^{C} ... \sum_{c_{\Omega_i(M_i)}=1}^{C}$, $w_{q_{\textcircled{i}}} = \prod_{j=1}^{M_i} \alpha_{\Omega_i(j)_{c_{\Omega_i(j)}}}$, $H_{\textcircled{i}} = \left[h_{\Omega_i(1)} ... h_{\Omega_i(M_i)}\right]$being a subset of the full matrix containing only the $\Omega_i(1)^{th}$ to $\Omega_i(M_i)^{th}$ columns and $v_{q_{\textcircled{i}}} = diag\left(\sigma_{\Omega_i(1)_{c_{\Omega_i(1)}}}, ..., \sigma_{\Omega_i(M_i)_{c_{\Omega_i(M_i)}}}\right)$. When all the sources are active, the observation density in (IV.3) uses the full mixing matrix and when none of the sources are active, the observation density reduces to white Gaussian noise [52].

We represent the evolution of the columns of the mixing matrices with indices $m = 1, ..., M$ as a random walk model of

$$h_m(t) = h_m(t-1) + \nu_m(t-1) \tag{IV.4}$$

where $\nu_m(t)$ is a white Gaussian random variable with a diagonal covariance. We assume that the transition from one state to another follows a Markovian property with transition matrix $\Pi = [\pi_{ij}]$ where $\pi_{ij} = Pr\left[x(t) = \textcircled{j} | x(t-1) = \textcircled{i}\right]$. In the next section we describe a MMPF algorithm capable of tracking the mixing matrices and the sources' activity pattern.

IV.C Multiple Model Particle Filtering

Particle filtering is an online Bayesian state estimation technique widely used for nonlinear/nonGaussian state estimation. From Eqs. IV.1 and IV.3, it is clearly evident that the relationship between the observations and the states is nonlinear and nonGaussian. Particle filtering in junction with ICA for time-variant mixing has been proposed before for linear instantaneous mixing while assuming the sources were active at all times [53, 28]. In this section we describe a frequency domain MMPF for convolutive mixing capable of switching between states corresponding to different source activity patterns. Assuming N particles are used, the main steps of the MMPF is summarized as follows [74]:

1. Initialize the state particles $\{h_m^n(0), m = 1, ...M\}_{n=1}^N$ and $\{x^n(0)\}_{n=1}^N$ based on a intial prior and using uniform weights $\{w_m^n(0) = 1/N, m = 1, ..., M\}_{n=1}^N$ and $\{r^n(0) = 1/N\}_{n=1}^N$.

2. Classify the particles to sets corresponding to different activity states, denoting $n_i = \{n | x^n(t) = \textcircled{i}\}$ for $i = 1, ..., I$. Next predict the new set of particles by drawing a new set of samples at time t according to state transitions described by

$$\begin{cases} h_m^{n_i}(t) = h_m^{n_i}(t-1) + \nu_m^{n_i}(t-1) & \text{state } \textcircled{i} \text{ contains column } m \\ h_m^{n_i}(t) = h_m^{n_i}(t-1) & \text{state } \textcircled{i} \text{ excludes column } m \end{cases} \tag{IV.5}$$

for $m = 1, ..., M$, $i = 1, ..., I$ and $n = 1, ...N$. This model assumes that columns of

the mixing matrices vary only when the corresponding sources are active. The reason for this is to avoid the particles from drifting when no information is available and the sources are silent. However, by keeping a memory of the silence patterns of the sources based on previous frames, the covariance of the cloud of particles can be increased virtually. This way the clouds of particles during the SBZs would be large enough to find the track once the sources become active again. After that if the sources remain active enough the covariances can be decreased back to normal (similar to recovering track in blind Doppler zones in radar signal processing [74]). Also, the activity pattern is predicted according to the rule that if $x^n(t-1) = \textcircled{j}$ then $x^n(t) = \textcircled{i}$ with probability π_{ji}.

3. In this step the weights for the culumns of the mixing matricies are updated as

$$\begin{cases} w_m^{n_i}(t) = w_m^{n_i}(t-1) P_{\textcircled{i}}(Y(t)|H^{n_i}(t)) & \textcircled{i} \text{ contains column } m \\ w_m^{n_i}(t) = w_m^{n_i}(t-1) & \textcircled{i} \text{ excludes column } m \end{cases} \tag{IV.6}$$

for $m = 1, ..., M$, $i = 1, ...I$ and $n = 1, ...N$. The activity weights are updated as

$$r^{n_i}(t) = r^{n_i}(t-1) P_{\textcircled{i}}(Y(t)|H^{n_i}(t)) \tag{IV.7}$$

for $i = 1, ...I$ and $n = 1, ...N$.

4. Normalize the activity weights so their sum is unit value

$$r^n(t) \leftarrow \frac{r^n(t)}{\sum_n r^n(t)} \tag{IV.8}$$

and from that obtain the probability of each activity state

$$p(x(t) = \textcircled{i}|Y(1, ..., t)) = \sum_{n_i} r^{n_i}(t) \tag{IV.9}$$

The column weights is then normalized as

$$w_m^{n_i}(t) \leftarrow w_m^{n_i}(t) p(x(t) = \textcircled{i}|Y(1, ..., t)) / \sum_{n_i} w_m^{n_i}(t) \tag{IV.10}$$

5. If the particles become degenerate resample them and reassign the weights to uniform.

6. Estimate the matrix columns using

$$\hat{h}_m(t) = \sum_n w_m^n(t) h_m^n(t) \tag{IV.11}$$

and from that the sources can be reconstruced using a minimum mean square error (MMSE) estimator [52].

7. Permutation in the frequency bins is corrected using a correlation method on the activity patterns by keeping a memory of the past estimates of the sources in each frequency bin [75]. After that the sources are converted to the time domain.

IV.D Localization and tracking

Once an estimate of the time varying mixing matrices is found, the DOAs can be found using the method in [79]. If the same procedure is repeated in parallel for another microphone array placed at a different position in the room the sources can be located and tracked using triangulation (similar to a multiple bearings-only framework with static sensors in radar signal processing [74]). For simplicity we assume that the secondary DOA estimates are synchronous to the primary estimates and that zero delay transmission delay exists between the two. Because the DOA estimates can be jittering especially when the sources are silent for some time and suddenly become active, we propose to smooth the localization process of triangulation by incorporating kinematic dynamics for the motion of the sources. Therefore, another tracking stage is added where the DOA estimates are treated as measurements and the positions and velocities of the sources are treated as states. Because the relationship between the DOAs and the position

of the sources is nonlinear [74], a method based on particle filtering is proposed again. Moreover, in order to model maneuvering sources a MMPF (similar to the tracking of the mixing matrices in the previous section) incorporating constant velocity and constant acceleration models in the x and y directions is employed. We note that because the algorithm is able to to first fully separate the sources and then localize them, the DOA estimates from the two arrays can be matched to a specific source by evaluating the correlation of the separated sources activity patterns [75], avoiding the problem of false positions due to possible permutations.

IV.E Computer Simulations

The proposed algorithm was put to test in a simulated room settings using the image method. We assumed a room size of 8x5x3.5m with a reverberation time of 200ms. We picked the simple case of $M = 2$ speakers and $L = 2$ microphones for each array. The two arrays were placed facing each other. The sampling frequency for speech sources was 8kHz and the spacing between the microphones in both arrays was 4cm in order to avoid spatial aliasing. The two sources moved in the same direction (one chasing the other) in a maneuvering radial pattern with an angular speed of around 6.4 deg/sec with respect to the primary microphone array. The total duration of the sources was 12.5 seconds with them being active only for an average of about 5.5 seconds. The data was corrupted with white Gaussian noise, with the noise level resulting in an input signal to noise ratio (SNR_{in}) of 14(dB). Signal to disturbance ratio (SDR) is used as the performance measure for the separation phase. SDR is the total signal power of direct channels versus the signal power stemming from cross interference and noise combined. Number of particles used was $N = 1000$. Position root mean square error (RMSE) is used as a performance measure for the tracking phase.

In order to evaluate the results, the proposed method was compared to an online independent vector analysis (IVA) method with a normalized natural

gradient nonholonomic constraint (NNGNC) [35]. Both algorithms were initialized the same way by performing batch IVA [52] on the first 2 seconds of data. The SDR using the proposed method measured to be 11.4 (dB) while the SDR of the online IVA algorithm came out to be 6.8 (dB). Figure IV.1 shows the true positions of the sources along with the estimated positions using the proposed method. Figure IV.2 compares the average position RMSE for the sources using the proposed algorithm and the online IVA algorithm. Because of the high jitter in the DOAs when using the online IVA method, tracking with a motion model failed to work. Therefore the positions found based on simple intersections of DOAs, without applying any motion model on the sources, was used to compute the RMSE of the online IVA algorithm. The audio files along with the video of the tracking phase are available at our website[1].

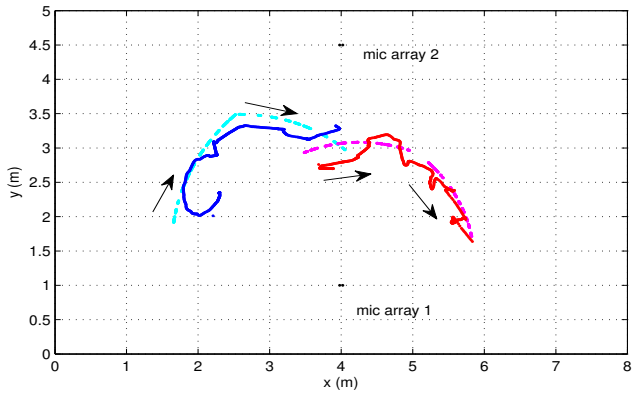

Figure IV.1 True (cyan and magenta) trajectories and estimated (blue and red) trajectories using the proposed method.

IV.F Summary and Discussion

We have proposed a novel frequency domain particle filtering method capable of tracking the mixing matrices of maneuvering sources in a reverberant environment. A glimpsing approach is also incorporated to switch between differ-

[1] `http://dsp.ucsd.edu/~ali/tracking/`

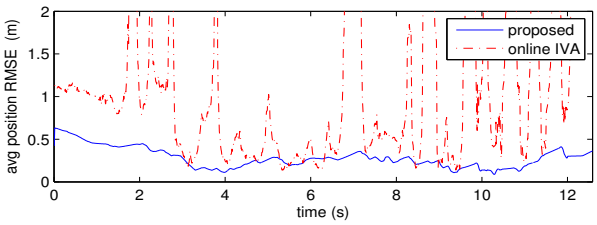

Figure IV.2 Average position RMSE of trajectories.

ent combinations of tracks when the source(s) become inactive, therefore avoiding losing track during such periods. The algorithm is also capable of recovering tracks during silence blind zones (SBZ) where the sources are moving while silent. Once the mixing matrices are correctly estimated, obtaining the directions of arrival (DOA) becomes a straightforward post-processing step. Using a secondary array positioned elsewhere in the room, the DOAs are matched up and triangulated by incorporating a multiple motion model on the source trajectories. Improved separation and tracking results were achieved in the simulations.

IV.G Acknowledgments

The text of Chapter III, in full, is based on the material as it appears in: Alireza Masnadi-Shirazi and Bhaskar Rao, "Separation and Tracking of Multiple Speakers in a Reverberant Environment using a Multiple Model Particle Filtering Glimpsing method" , in proc. IEEE ICASSP 2011. The dissertation author was a primary researcher and an author of the cited material.

Chapter V

Quasi-online Tracking and Separation of Unknown Time-Varying Number of Sources

V.A Introduction

As discussed in the previous chapter, passive localization and tracking of multiple acoustical sources is of great interest in the field of microphone arrays which is driven by applications such as automatic camera steering for teleconferencing and surveillance. Speaker localization is also very useful in aiding systems achieving the task of separating concurrent speakers or a desired speaker from background interference with applications such as high-quality hearing aids, speech enhancement and noise reduction for smart phones. By localization, one can refer to finding the bearings of the speakers or their Cartesian coordinate. In this chapter we are particularly interested in estimating the bearing information of multiple sources or their direction of arrival (DOA) by means of the time difference of arrival (TDOA).

TDOA estimation is the first stage for many speaker localization algorithms involving one or more microphone pairs. In the case of a single speaker, TDOA can be reliably estimated using the generalized cross-correlation phase transform (GCC-PHAT) using one microphone pair [38, 65]. GCC-PHAT is a scanning method that computes the correlation of the microphone pair inputs for a range of TDOAs with an arbitrary resolution, resulting in peaks where the correlation is high. In case of multiple speakers, GCC-PHAT does not always provide reliable TDOA for all the sources since one of the sources can dominate over the others [11]. This means that as the concurrent sources increase in number, multiple TDOA estimation using GCC-PHAT becomes less reliable. Also, multipath propagation due to reverberation can cause additional peaks in the GCC-PHAT that correspond to multi-path propagations. This results in the situation where for example in the case of two sources, the first and second peak do not always correspond to the first and second source and sometimes the third or subsequent peaks need to be considered [46]. Extensions of the GCC-PHAT for multiple sources have been proposed [19, 83, 39, 88]. However, they require microphone pair redundancy and

high sampling rates to increase the reliability of the TDOA estimates.

Multiple TDOA estimation using frequency domain independent component analysis (ICA) was first proposed in [79]. In the context of blind source separation (BSS), ICA is a well known tool for the separation of linear and instantaneous mixed signals picked up by multiple sensors [34]. ICA estimates a de-mixing matrix for the separation task and does so by assuming the sources are statistically independent and non-Gaussian distributed. For many real world problems, the signals undergo a convoluted mixing due to reverberation. By transforming the mixture to the frequency domain by applying the short-time Fourier transform (STFT), convolution in the time domain translates to linear mixing in the frequency domain. Subsequently, ICA can be performed on every single frequency bin. Since ICA is indeterminate of source permutation, further post processing methods are necessary to correct for possible permutations of the separated sources in each frequency bin [78, 75]. In [79], multiple TDOAs are calculated directly from the columns of the estimated mixing matrix. However, this method works well only if the possible source permutations in the frequency bins have been corrected and there are no frequency bins affected by spatial aliasing (hence a minimal microphone spacing). Recently an extension to [79] has been proposed under the name of state coherence transform (SCT) that does not require permutation correction and is insensitive to spatial aliasing [57, 59]. Similar to GCC-PHAT, SCT is a scanning method. However, instead of finding the correlation between the two microphone input signals for TDOA points in the scan, it forms a pseudo-likelihood between a propagation model for the different TDOA scan points and the TDOA observations pertaining to the columns of the mixing matrices, resulting in peaks where the scan points in the model and observations best match. One attractive feature of SCT is that by exploiting the frequency sparsity of the sources, it is effective even when the number of simultaneous sources is larger than the number of sensors. Also, since SCT uses ICA outputs which attempt to separate the sources, it is more suitable for TDOA estimation for multiple sources

compared to GCC-PHAT [59].

Assuming that the number of sources is known and fixed in time, some methods exist that track the location information for each source by incorporating a separate tracker for each source [18]. However, in many real world problems, not only do the states of the sources change with time, the number of concurrent sources is unknown and varies with time as new speakers can appear and existing speakers can disappear or undergo long silence periods. Moreover, the measurements can receive a set of spurious peaks (clutter) due to the multi-path propagation caused by reverberation and spatial aliasing, resulting in false alarms. In addition, not all of the sources are detected giving rise to missed detections as well. Therefore, the passive scanning methods discussed earlier result in an assortment of indistinguishable observations where only a subset of them are generated by the sources. Recently, methods based on random finite sets (RFS) have presented promising and mathematically elegant solutions to the problem of multi-target tracking (MTT) for time-varying number of targets [48, 47]. Using RFSs, the collection of indistinguishable observations in the presence of clutter is treated as a set-valued observation while the multi-target states and the number of targets are integrated to form a set-valued state. The goal becomes to estimate the target states and the target number while rejecting clutter and accounting for missed detections. The RFS formulation allows the problem to be posed in an optimal multi-target Bayesian filtering framework, and is an extension of the well known single target Bayes filter. However, the optimal RFS Bayes filter is computationally intractable as it becomes a combinatorial problem on the number of targets involving high dimensional integrals. The probability hypothesis density (PHD) filter is a suboptimal approximation to the RFS Bayes filter which propagates the first moment of multi-target posterior density rather than the full posterior density [48]. This said, the PHD filter still involves multiple integrals with no closed form solution in general. Also, the PHD filter in itself, does not solve the data association problem indicating which estimate belongs to which target. The Gaussian

mixture implementation of the PHD filter (GM-PHD) alleviates these two difficulties: It provides a closed form solution of the PHD filter when the target states and observations follow a linear and Gaussian dynamic model (which is a reasonable model for the problem of interest in this chapter) [85]. It also solves the data association problem intrinsically and provides track labels which are imperative to the separation task of interest [67].

The problem of extracting location information of unknown time-varying number of speakers using RFSs and PHD filtering has been proposed before. These methods, however, use GCC-PHAT in the front-end to obtain the measurements and bear the inherent limitations of GCC-PHAT for multiple sources including being inherently incapable of source separation [46, 9]. For the same problem, a method exists that uses ICA/SCT in the front-end and uses a naive thresholding approach to estimate the number of targets [44, 43]. This method, however, is sensitive to the selected thresholds and relies solely on the thresholds to reject clutter. Another class of methods uses a steered beamformer for acquiring the measurements and then applies a variable-dimension particle filter or track-before-detect filtering scheme for the tracking and source activity detection [30, 29]. These methods cannot perform the separation task inherently and don't quite estimate the number of targets but estimate the activity pattern of a limited few number of sources. In the previous chapter, we have proposed an ICA-based approach to separate and track multiple sources for when the sources can experience short silence periods [51]. This method, while being able to separate the sources, only estimates the activity patterns and cannot handle new sources being born or completely dying out. In this chapter we propose the use of the GM-PHD to filter the measurements obtained from short time blocks using ICA/SCT. By doing so we are able to track the DOA of multiple time-varying number of sources and from the track labels we are able to go back to the ICA outputs and perform the separation task by associating each separated time-frequency block with its estimated corresponding track. The separation scheme exploits the frequency sparsity of the

sources and enables the separation of more concurrent sources than sensors. Computer simulations on the DOA tracking using the proposed method is compared with the first two aforementioned existing approaches and the results are favorable and promising.

Overall, this chapter demonstrates how a mixture/superposition model in the framework of BSS can be easily represented as a standard detection model in the framework of multi-target tracking, assuming that the sources have frequency sparsity. Such an idea of transforming a mixture/superposition model to a detection model, was first presented in [10], where the sources were assumed to be narrowband audio tones and the STFT representation was enough to execute such transformation. As it turns out, the approach in [10] is a special case of the proposed method for when the sources have a super-sparse representation to a degree where they will be non-overlapping and occupy a single frequency bin, making the ICA separation scheme unnecessary. The proposed method offers a solution for executing the transformation from the mixture model to the detection model for broadband signals that have some sort of frequency sparsity, such as speech and communication signals. Recently, in the context of multi-target tracking, other methods have been proposed that deal with the mixture/superposition model directly and perform a moment-based RFS filtering [49, 82, 12, 4]. These methods, however, are either computationally intractable or do not enjoy the relative simplicity of the PHD filter.

This chapter is organized as follows: Section V.B explains the front end of the system consisting of ICA in junction with SCT where the mixture representation of the sources are transformed to a DOA multi-target detection representation, regardless of permutation, spatial aliasing and the number of sources being more than the sensors. Section V.C explains the back-end of the system and gives a background theory on multi-target filtering using a RFS framework along with implementations and appropriate extensions of the PHD filter. We present all the formulations of this section in a summarized way while maintaining a consistent

context. In Section V.D, we present how the front-end and back-end are synergistically combined to perform both the tracking and separation tasks. In Section V.E, some experimental results are evaluated. Finally, in Section V.F, our conclusions are stated and the main contributions of this chapter are summarized.

V.B Frequency Domain BSS and SCT

Assuming L sensors and M sources, the convolutedly mixed observation at the l^{th} sensor at time u is

$$y_l(u) = \sum_{j=1}^{M} \sum_{r=0}^{R-1} h_{lj}(r) s_j(u-r) \tag{V.1}$$

where $s_j(u)$ is the j^{th} source in the time domain, h_{lj} is the finite impulse response (FIR) approximation of duration R linking the j^{th} source to the l^{th} sensor. The signals are transformed to the frequency domain using the short time Fourier transform (STFT). The STFT takes the discrete Fourier transform (DFT) of frames of the signal using a sliding window, hence creating a time-frequency representation of the signal, commonly known as the spectrogram. We must note that the window length of the STFT should be sufficiently large, ensuring that the conversion from convolution in the time domain, is approximated fairly by multiplication in the frequency domain. Using STFT, the l^{th} sensor observation at time frame n and frequency bin $k = 1, ..., K$ becomes

$$Y_l(n,k) = \sum_{j=1}^{M} H_{lj}(k) S_j(n,k) \tag{V.2}$$

where $S_j(n,k)$ is the frequency domain representation of the j^{th} source at bin k and frame n. Omitting n for simplicity, we can arrange Eq. V.2 for frequency bin k in matrix form as

$$Y(k) = H(k)S(k) \tag{V.3}$$

where $Y(k) = [Y_1(k)...Y_L(k)]^T$, $S(k) = [S_1(k)...S_M(k)]^T$ and $H(k)$ is the $L \times M$ mixing matrix corresponding to the k^{th} frequency bin.

For the case of $L = M$, any complex-valued ICA algorithm [34] can be applied to each frequency bin to estimate the inverse of the mixing matrix $H(k)$. Denoting the estimate of the separated sources at the k^{th} bin as $\hat{S}(k)$, from ICA we get

$$\hat{S}(k) = \hat{W}(k)Y(k) \tag{V.4}$$

where $\hat{W}(k)$ denotes the estimate of the demixing matrix up to scaling and permutation ambiguities:

$$\hat{W}(k) = \Lambda(k)\Pi(k)\hat{H}^{-1}(k) \tag{V.5}$$

where $\Lambda(k)$ is a diagonal scaling matrix, $\Pi(k)$ is a permutation matrix and $\hat{H}(k)$ is the estimate of the true mixing matrix $H(k)$.

Without loss of generality, for simplicity, we consider a configuration of two sources and two sensors. In an ideal anechoic setting the true mixing matrix can be modeled as

$$H(k) = \begin{pmatrix} |h_{11}(k)|e^{-j2\pi f_k T_{11}} & |h_{12}(k)|e^{-j2\pi f_k T_{12}} \\ |h_{21}(k)|e^{-j2\pi f_k T_{21}} & |h_{22}(k)|e^{-j2\pi f_k T_{22}} \end{pmatrix} \tag{V.6}$$

where T_{qp} is the propagation time from the p^{th} source to the q^{th} microphone and f_k is the frequency in Hz for the k^{th} frequency bin. By neglecting the permutation problem for now but taking into account the scaling ambiguity, the estimate of the inverse of the demixing matrix becomes

$$\begin{aligned} \hat{W}^{-1}(k) &= \begin{pmatrix} |\hat{h}_{11}(k)|e^{-j2\pi f_k \hat{T}_{11}} & |\hat{h}_{12}(k)|e^{-j2\pi f_k \hat{T}_{12}} \\ |\hat{h}_{21}(k)|e^{-j2\pi f_k \hat{T}_{21}} & |\hat{h}_{22}(k)|e^{-j2\pi f_k \hat{T}_{22}} \end{pmatrix} \begin{pmatrix} \frac{1}{\eta_1(k)} & 0 \\ 0 & \frac{1}{\eta_2(k)} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\eta_1(k)}|\hat{h}_{11}(k)|e^{-j2\pi f_k \hat{T}_{11}} & \frac{1}{\eta_2(k)}|\hat{h}_{12}(k)|e^{-j2\pi f_k \hat{T}_{12}} \\ \frac{1}{\eta_1(k)}|\hat{h}_{21}(k)|e^{-j2\pi f_k \hat{T}_{21}} & \frac{1}{\eta_2(k)}|\hat{h}_{22}(k)|e^{-j2\pi f_k \hat{T}_{22}} \end{pmatrix} \end{aligned} \tag{V.7}$$

where $\eta_i(k)$ represents the diagonal entries of the arbitrary scaling matrix $\Lambda(k)$ in (V.5). Neglecting reverberation, the TDOA information emerges when taking the ratios of the entries of each column in (V.7)

$$r_1(k) = \frac{|\hat{h}_{11}(k)|}{|\hat{h}_{21}(k)|}e^{-j2\pi f_k \hat{\Delta} t_1}, \quad r_2(k) = \frac{|\hat{h}_{12}(k)|}{|\hat{h}_{22}(k)|}e^{-j2\pi f_k \hat{\Delta} t_2} \tag{V.8}$$

where $\hat{\Delta}t_i$ are the TDOAs of the sources with respect to the microphone pair. As it can be seen from (V.8), such ratios are invariant to the scaling ambiguities of the estimation process. Since the TDOA information resides only in the phase of the ratios in (V.8) and is invariant to scaling and magnitude, the ratios can be simplified as

$$\bar{r}_1(k) = \frac{r_1(k)}{|r_1(k)|}, \quad \bar{r}_2(k) = \frac{r_2(k)}{|r_2(k)|} \tag{V.9}$$

If the permutation of the sources can be somehow corrected and if the mixing does not undergo spatial aliasing, the TDOAs of the sources can be estimated directly from phase information of (V.9) by exploiting the linear relationship between the TDOAs and the true frequencies along the different bins [79]. However, solving the permutation problem and dealing with spatial aliasing can prove to be difficult in practice. SCT is a method that can sidestep these issues by forming a pseudo-likelihood between the TDOA observations in (V.8) and a propagation model that can intrinsically account for both permutations and spatial aliasing [57]. The propagation model that results in TDOA of a source with respect to the microphones, denoted as τ, is assumed to be

$$c(k, \tau) = e^{-j2\pi f_k \tau} \tag{V.10}$$

The SCT for the configuration of two sources and two microphones is formulated to be

$$SCT(\tau) = \sum_{k} \sum_{m=1}^{2} \left[1 - g\left(\frac{\|c(k, \tau) - \bar{r}_m(k)\|}{2} \right) \right] \tag{V.11}$$

where the transform is scanned for different values of τ with an arbitrary resolution and $g(.)$ is a function of the Euclidian distance. A good option for $g(.)$ is shown to have a sigmoidal shape such as the following

$$g(\xi) = tanh(\alpha \xi) \tag{V.12}$$

where α is a real positive constant that defines the inter-source resolution of the spatial likelihood, i.e. the capability of the system to spatially discriminate TDOAs

related to different sources and is usually set empirically. The sigmoidal shape gives more emphasis when the observations $\bar{r}_m(k)$ are close to the model $c(k,\tau)$ while ignoring the other values. It can be easily understood from (V.11) that one could expect to see higher mappings of SCT for values of τ which $\bar{r}_m(k)$ and the model $c(k,\tau)$ are closer in some Euclidian form of distance, thus creating peaks for values of τ matching the TDOAs. One important feature of SCT is that it is invariant to source permutations since it jointly utilizes the TDOA information of all the ratios in (V.9) across all frequencies. On the other hand, since the model $c(k,\tau)$ incorporates the 2π phase wrap-arounds (i.e. it is periodic for 2π shifts) caused by spatial aliasing, its sensitivity towards spatial aliasing is greatly reduced. Moreover, the most important feature of SCT that makes it an attractive platform for tracking unknown time-varying number of sources is that it is able to map the TDOA peaks for the underdetermined or overcomplete case which involves having more sources than microphones. This is achieved by partitioning the data (STFT frames) into small blocks and performing ICA/SCT on each data block. For example for the case explored so far of two microphones, by exploiting the frequency sparsity of the sources (which is typical of speech) in each data block, and assuming that at each frequency bin and each data block at most two sources are active, a complete TDOA mapping of all the sources (whose number in total can be greater than two) becomes possible. From the far-field assumption, one can convert TDOA detections into DOA using

$$\theta = cos^{-1}(c\Delta t/\Delta q) \tag{V.13}$$

where c is the speed of sound and Δq is the distance between the microphone pair.

In case the number of microphones is greater than two, the generalized state coherence transform (GSCT), which is a multi-dimensional extension to SCT, is used. In GSCT each dimension of the domain pertains to a τ_p variable, where p is the index of the microphone pair [57]. In this chapter we use two microphones for our experiments, hence the multi-dimensional TDOA mapping using GSCT is not discussed. It is noteworthy to say that even though the SCT propagation

model only considers the direct path in an anechoic setting, nonetheless, it is still shown to be effective for multi-path propagation due to reverberation. The reason for this is that in a reverberant environment the direct path between the source and the microphone is usually dominant over other multi-path propagations. As the amount of reverberation increases the chance of multi-paths creating peaks in the SCT increases as well. Consequently, for dealing with unknown time-varying number of sources, as considered in this chapter, a suitable filtering technique is needed to reject clutter caused by multi-path propagations.

V.C Bayesian Multi-Target Tracking and PHD Filtering

In the previous section we explained how to effectively transform a mixture representation of multiple concurrent sources in the framework of ICA into a detection representation of the source DOAs by extracting significant peaks of the SCT. In the detection framework, hereafter, we will call the SCT peaks that originate from a source as "target", assuming that each source only gives rise to one target. Let's assume that at time t, the sensor makes N_t observations (detections) $z_{t,1}, ..., z_{t,N_t}$ each taking values in the state space $\mathcal{Z}$. These detections are ambiguous in the sense that it is not known whether they have originated from targets or are false detections (clutter). Moreover, due to the imperfections in the sensor resolution it is possible that an arbitrary subset of targets do not get detected (missed detections). Our goal is to process such detections in order to reject clutter, account for missed detections, identify the number of sources and track the target states. Now let's consider the multi-target scenario where at time $t-1$ there exist M_{t-1} targets with states $x_{t-1,1}, ..., x_{t-1,M_{t-1}}$ taking values in the state space $\mathcal{X}$. At the next instance of time, t, some of the targets can die, some new targets can be born and the surviving targets can evolve according to some dynamic model. This results in M_t targets at time t with states $x_{t,1}, ..., x_{t,M_t} \in \mathcal{X}$. Assuming that the respective ordering of the measurements and the state estimates

have no significance, the multi-target states and observations can be represented as finite sets such as

$$X_t = \{x_{t,1}, ..., x_{t,M_t}\} \in \mathcal{F}(\mathcal{X}) \tag{V.14}$$

$$Z_t = \{z_{t,1}, ..., z_{t,N_t}\} \in \mathcal{F}(\mathcal{Z}) \tag{V.15}$$

where $\mathcal{F}(\mathcal{X})$ and $\mathcal{F}(\mathcal{Z})$ are finite subsets of the spaces of $\mathcal{X}$ and $\mathcal{Z}$, respectively. By assuming that the multi-target RFS state $X(t)$ is the union of surviving targets, spontaneous births and spawned targets, and the multi-target detection RFS state $Z(t)$ is the union of target-generated detections and clutter, the goal of Mahler's RFS multi-target filtering [48] is to estimate the number of targets and their states while rejecting clutter and accounting for missed detections. The RFS formulation for multi-target Bayesian filtering is the extension of the well known single-target Bayesian filtering which can be computed sequentially via the prediction and update steps as following

$$f_{t|t-1}(X_t|Z_{1:t-1}) = \int f_{t|t-1}(X_t|X_{t-1}) f_{t-1|t-1}(X_{t-1}|Z_{1:t-1}) \delta X_{t-1} \tag{V.16}$$

$$f_{t|t}(X_t|Z_{1:t}) = \frac{f_{t|t}(Z_t|X_t) f_{t|t-1}(X_t|Z_{1:t-1})}{\int f_{t|t}(Z_t|X'_t) f_{t|t-1}(X'_t|Z_{1:t-1}) \delta X'_t} \tag{V.17}$$

where $f_{t|t-1}(X_t|Z_{1:t-1})$ is the multi-target predictive density, $f_{t|t}(X_t|Z_{1:t})$ is the multi-target posterior density, $Z_{1:t}$ is the concatenation of all previous measurements up to time t and δ is an appropriate reference measure on $\mathcal{F}(\mathcal{X})$ which indicates that the integrals are set-integrals. A set-integral is a non-trivial extension of a regular integral which is defined as a mixture of regular integrals over all different subsets of the multi-target states. This accounts for the uncertainty in the target number which can vary over time as new targets enter and old ones vanish. The exact definitions of set-integrals and set-derivatives is part of Mahler's Finite Set Statistics (FISST) [48] which provides a systematic calculus-based approach to multi-target filtering using RFSs.

Due to the use of combinatorial set-integrals in the optimal Bayesian recursions of (V.16-V.17), they involve multiple high dimensional integrals on the

space $\mathcal{F}(\mathcal{X})$ rendering it computationally intractable. The PHD filter is a suboptimal approximation to the mutli-target Bayesian recursions of (V.16-V.17) which instead of propagating the full posterior density, propagates the FISST-based first moment of multi-target posterior density, known as the posterior intensity [47, 48]. This is analogous to the well known constant-gain Kalman filter in single-target tracking, which also only propagates the first moment (mean) of the target.

Let $D_{t|t-1}(x_t|Z_{1:t-1})$ and $D_{t|t}(x_t|Z_{1:t})$ denote the respective PHD intensities of the multi-target predictive posterior $f_{t|t-1}(X_t|Z_{1:t-1})$ and the multi-target posterior $f_{t|t}(X_t|Z_{1:t})$ of equations (V.16-V.17). It is worthy to note that due to the first order moment mapping of the PHD filter, the finite set-valued random variable state $X_t \in \mathcal{F}(\mathcal{X})$ of the multi-target posterior is represented by an ordinary random variable $x_t \in \mathcal{X}_0$ with dimensions pertaining to the dimensions of a single target, i.e. $D_{t|t}(x_t|Z_{1:t})$ is an intensity function on the single target space $\mathcal{X}_0$. This PHD intensity function is not in the form of a probability density function (pdf) as its integral does not equate to unity. Under certain assumptions and using FISST [48, 47], the PHD intensities can be recursively estimated as follows

$$D_{t|t-1}(x_t|Z_{1:t-1}) = b_t(x_t) + \int F_{t|t-1}(x_t|x_{t-1}) D_{t-1|t-1}(x_{t-1}|Z_{1:t-1}) dx_{t-1} \tag{V.18}$$

$$D_{t|t}(x_t|Z_{1:t}) = [1 - p_D(x_t)]\, D_{t|t-1}(x_t|Z_{1:t-1}) + \sum_{z_t \in Z_t} \frac{\psi_{z_t}(x_t) D_{t|t-1}(x_t|Z_{1:t-1})}{\kappa_t(z_t) + \int \psi_{z_t}(\zeta) D_{t|t-1}(\zeta|Z_{1:t-1}) d\zeta} \tag{V.19}$$

In the prediction equation (V.18)

$$F_{t|t-1}(x_t|x_{t-1}) = p_S(x_{t-1}) f_{t|t}(x_t|x_{t-1}) + \beta_{t|t-1}(x_t|x_{t-1}) \tag{V.20}$$

where $f_{t|t}(x_t|x_{t-1})$ is the single target transition pdf, p_S is the probability of target survival and $\beta_{t|t-1}$ is the intensity of target spawned from targets at time $t-1$. Also in (V.18), b_t is the intensity of spontaneous new births at time t. In the

update equation (V.19),

$$\psi_{z_t}(x_t) = p_D(x_t)g(z_t|x_t) \tag{V.21}$$

where p_D is the probability of detection, $g(z_t|x_t)$ is the single target detection likelihood model (i.e. observation model in the space of $\mathcal{X}_0$) and the intensity of clutter points $\kappa_t(z_t)$ is given as

$$\kappa_t(z_t) = \lambda c_t(z_t) \tag{V.22}$$

where λ is the average number of Poisson-distributed false alarms and $c_t(z)$ is the spatial distribution of clutter. As we mentioned before the PHD intensity function is not a pdf and in fact it turns out that the integral of the PHD intensity gives the expected number of targets as follows [48]

$$\hat{M}_{t|t} = \int D_{t|t}(x_t|Z_{1:t})dx_t \tag{V.23}$$

At the end, the state estimates for each target are extracted by finding the $\hat{M}_{t|t}$ peaks of intensity $D_{t|t}(x_t|Z_{1:t})$. In the case where only a single target is present, the formulations above reduces to the constant-gain Kalman filter.

Even though the PHD filter is much less computationally expensive compared to the multi-target recursions of (V.16-V.17), due to the fact that it operates in the space of a single target $\mathcal{X}_0$, the integrals present in the PHD recursions of (V.18-V.19) result in it not having a closed form solution in general. Therefore, Sequential Monte Carlo (SMC) methods are usually used to approximate the integrals in general [86]. However, for the special case where the target dynamics follow a linear Gaussian Model, a Gaussian mixture (GM) implementation can provide a closed form solution to the PHD filter [85]. The GM-PHD does not suffer from the complexities of sampling and resampling in SMC methods and due to its closed form solution, it is more accurate. In this chapter, since it is reasonable to assume that our measurements and target state dynamics follow a linear and Gaussian model, GM-PHD is used for the multi-source filtering.

V.C.1 GM-PHD Implementation

Assuming that the target dynamics and sensor model follows a linear and Gaussian form, we have

$$f_{t|t-1}(x_t|x_{t-1}) = \mathcal{N}(x_t; A_{t-1}x_{t-1}, Q_{t-1}) \tag{V.24}$$

$$g(z_t|x_t) = \mathcal{N}(z_t; B_t x_t, R_t) \tag{V.25}$$

where $\mathcal{N}(.; a, C)$ denotes a Gaussian pdf with mean a and covariance C, A_{t-1} is the state transition matrix, B_t is the observation matrix, Q_{t-1} is the transition process noise covariance and R_t is the observation noise covariance. The GM-PHD requires that the survival and detection probabilities be state independent, therefore $p_S(x_t) = p_S$ and $p_D(x_t) = p_D$. Another assumption is that the birth and spawn intensities are Gaussian mixtures [85]. For simplicity we neglect the spawning of new targets from previous targets and just rely on spontaneous births to model new targets. Therefore we have

$$b_t(x_t) = \sum_{i=1}^{J_{b,t}} \omega_{b,t}^{(i)} \mathcal{N}\left(x_t; m_{b,t}^{(i)}, P_{b,t}^{(i)}\right) \tag{V.26}$$

where $J_{b,t}$, $\omega_{b,t}^{(i)}$, $m_{b,t}^{(i)}$, $P_{b,t}^{(i)}$, $i = 1, ..., J_{b,t}$, are given model parameters that determine the shape of the birth intensity. Usually one adapts these parameters to model regions in the state space which correspond to detection persistences. Again, we make note that equation (V.26) is not a pdf, in general. That is because there is no restriction on the coefficients $\omega_{b,t}^{(i)}$,$i = 1, ..., J_{b,t}$, adding to unity.

Assuming that the posterior PHD at time $t-1$ is a Gaussian mixture of the form

$$D_{t-1|t-1}(x_{t-1}|Z_{1:t-1}) = \sum_{i=1}^{J_{t-1}} \omega_{t-1}^{(i)} \mathcal{N}\left(x_{t-1}; m_{t-1}^{(i)}, P_{t-1}^{(i)}\right), \tag{V.27}$$

then the predicted intensity at time t is a Gaussian mixture given by

$$D_{t|t-1}(x_t|Z_{1:t-1}) = b_t(x_t) + p_{S,t} \sum_{j=1}^{J_{t-1}} \omega_{t-1}^{(j)} \mathcal{N}\left(x_t; m_{S,t|t-1}^{(j)}, P_{S,t|t-1}^{(j)}\right) \tag{V.28}$$

where

$$m_{S,t|t-1}^{(j)} = A_{t-1} m_{t-1}^{(j)} \tag{V.29}$$

$$P_{S,t|t-1}^{(j)} = Q_{t-1} + A_{t-1} P_{t-1}^{(j)} A_{t-1}^{T} \tag{V.30}$$

As the predicted intensity for time t can be rearranged to have a Gaussian mixture of the form

$$D_{t|t-1}(x_t|Z_{1:t-1}) = \sum_{i=1}^{J_{t|t-1}} \omega_{t|t-1}^{(i)} \mathcal{N}\left(x_t; m_{t|t-1}^{(i)}, P_{t|t-1}^{(i)}\right), \tag{V.31}$$

the posterior intensity at time t also becomes a Gaussian mixture as follows

$$D_{t|t}(x_t|Z_{1:t}) = (1 - p_{D,t}) D_{t|t-1}(x_t|Z_{1:t-1}) + \sum_{z_t \in Z_t} D_{D,t}(x_t; z_t) \tag{V.32}$$

where

$$D_{D,t}(x_t; z_t) = \sum_{j=1}^{J_{t|t-1}} \omega_t^{(j)}(z_t) \mathcal{N}\left(x_t; m_{t|t}^{(j)}(z_t), P_{t|t}^{(j)}\right) \tag{V.33}$$

$$\omega_t^{(j)}(z_t) = \frac{p_{D,t} \omega_{t|t-1}^{(j)} q_t^{(j)}(z_t)}{\kappa_t(z_t) + p_{D,t} \sum_{l=1}^{J_{t|t-1}} \omega_{t|t-1}^{(l)} q_t^{(l)}(z_t)} \tag{V.34}$$

$$q_t^{(j)}(z_t) = \mathcal{N}\left(z_t; B_t m_{t|t-1}^{(j)}, R_t + B_t P_{t|t-1}^{(j)} B_t^T\right) \tag{V.35}$$

$$m_{t|t}^{(j)}(z_t) = m_{t|t-1}^{(j)} + K_t^{(j)}\left(z_t - B_t m_{t|t-1}^{(j)}\right) \tag{V.36}$$

$$P_{t|t}^{(j)} = \left[I - K_t^{(j)} B_t\right] P_{t|t-1}^{(j)} \tag{V.37}$$

$$K_t^{(j)} = P_{t|t-1}^{(j)} B_t^T \left(B_t P_{t|t-1}^{(j)} B_t^T + R_t\right)^{-1} \tag{V.38}$$

Similar to the Gaussian sum filter in single target tracking [74], as time progresses, the number of Gaussian components in GM-PHD increases without bound. To fix this, a simple pruning and merging technique can be used to limit the growth of number of Gaussians [85]. It works by discarding Gaussians whose weight $\omega_t^{(i)}, i = 1, ..., J_t$ falls below some threshold and then normalizes the weights of the surviving Gaussians so that the sum of the weights, which from (V.23) is the expected number of targets, remains the same. Then it uses a Mahalanobis distance measure to merge Gaussians that are close to each other.

Once the expected number of targets $\hat{M}_{t|t}$ is found, estimating the multi-target states at first glance appears to be straightforward since the peaks in the posterior intensity $D_{t|t}(x_t|Z_{1:t})$ correspond to the means of the Gaussians, given that they are well-separated. However, since the height of peaks in the posterior intensity depends on both weight and covariance, selecting the $\hat{M}_{t|t}$ highest peaks may result in state estimates that correspond to Gaussians with weak weights. This is not desirable since the expected number of targets due to these peaks is small even though the magnitudes of the peaks are large. A better alternative is to select the means of the Gaussians with weights greater than some threshold, say 0.5 [85].

V.C.2 Data Association using the GM-PHD

The PHD filter, in itself, does not solve the data association problem, therefore one cannot tell which state estimates belong to which target. However, by associating tags to the mixture components of the GM-PHD filter, a data association scheme can be utilized providing us with distinct tracks on the sources. The tag labeling steps for GM-PHD filter with track management is as follows [67]:

Initialization

At time $t = 0$, J_0 Gaussians are distributed across the state space to form the intensity

$$D_0(x_t) = \sum_{j=1}^{J_0} \omega_0^{(j)} \mathcal{N}\left(x_t; m_0^{(j)}, P_0^{(j)}\right) \tag{V.39}$$

A unique tag is assigned to each Gaussian to form the set

$$\mathcal{T}_0 = \{\Upsilon_0^{(1)}, ..., \Upsilon_0^{(J_0)}\} \tag{V.40}$$

Prediction

After predicting forward the PHD intensity, Gaussians associated with new births receive new tags and Gaussians that are associated with surviving ones

retain previous tags, i.e. the set of tags is as follows

$$\mathcal{T}_{t|t-1} = \mathcal{T}_{t-1|t-1} \cup \{\Upsilon_{b,t}^{(1)}, ..., \Upsilon_{b,t}^{(J_{b,t})}\} \tag{V.41}$$

where $\Upsilon_{b,t}^{(j)}$ is the j^{th} new tag associated with the spontaneous birth intensity in equation (V.26) and $\mathcal{T}_{t-1|t-1}$ contains the tags of targets at time $t-1$ such that the predicted Gaussian with mean $m_{S,t|t-1}^{(j)}$ in equations (V.28-V.29) retains the tag of the Gaussian with mean $m_{t-1|t-1}^{(j)}$.

Update

The predicted intensity is updated according to equation (V.32). Hence, each Gaussian component of the predicted intensity gives rise to $1+|Z_t|$ components in the updated intensity, where $|\mathcal{A}|$ denotes the cardinality of set $\mathcal{A}$. Hence, the tags of the predicted Gaussian components get propagated to the updated Gaussian components, i.e. the Gaussian with mean $m_{t|t}^{(j)}(z_t)$ in equation (V.33) gets the same tag that the Gaussian with mean $m_{t|t-1}^{(j)}$ had in equation (V.31).

Pruning and merging

At this step the tags of the Gaussians that get pruned vanish. For the Gaussian that are merged, the tag of the one with the largest weight is retained.

Multi-target state estimation

At this step the means and tags of the Gaussians with weights higher than the aforementioned threshold (see end of Section V.C.1) are reported as state targets and track labels, respectively. Hence, there is an identifying tag associated with each estimate. If the target is a new born one, it has a new tag and if it is a surviving target it retains its previous track label.

V.C.3 Incorporating Amplitude Information in the PHD Likelihood

In target tracking applications, the detection step consists of extracting local peaks in the observations that are higher than some certain threshold. These detection points either come from targets or from clutter. The standard PHD filter discussed treats all detections equally and relies on the track continuity of the targets to reject clutter. However, in most cases the amplitude of detections generated from targets are higher than clutter and carry reliability information. This information about the amplitudes can be incorporated in the PHD tracking algorithm to further assist the discrimination of targets from clutter [20]. This is done by introducing an augmented measurement vector $\bar{z}_t = [z_t^T a]^T$, where $a \geq 0$ is the detection amplitude. Assuming that the amplitudes are independent of the target states, the respective target and clutter likelihood functions ψ_{z_t} and $\kappa_t(z_t)$ in equation (V.19) are modified to become

$$\bar{\psi}_{\bar{z}_t}(x_t) = g(z_t|x_t)g_a(a|d) \tag{V.42}$$

$$\bar{\kappa}_t(\bar{z}_t) = \lambda c_t(z_t)c_a(a) \tag{V.43}$$

where d is related to the signal-to-noise ratio (SNR), i.e. the ratio between target amplitude and clutter amplitude. SNR is defined in the log scale as

$$\text{SNR(dB)} = 10log_{10}(1 + d) \tag{V.44}$$

and is assumed to be the same for all targets. For the case of $d = 0$ the amplitude of the targets and clutter become the same, hence it is reduced to the standard PHD filter. In equations (V.42-V.43), $g_a(a|d)$ and $c_a(a)$ are the amplitude likelihood densities for targets and clutter, respectively. Assuming that the detection threshold is τ_o in which all peaks above τ_o are reported, the amplitude likelihoods for measurements that exceed τ_o are denoted as $g_a^{\tau_o}(a|d)$ and $c_a^{\tau_o}(a)$. Hence, due to normalization we have

$$g_a(a|d) = g_a^{\tau_o}(a|d)p_D^{\tau_o}(d) \tag{V.45}$$

$$c_a(a) = c_a^{\tau_o}(a)p_{FA}^{\tau_o} \tag{V.46}$$

where

$$p_D^{\tau_o}(d) = \int_{\tau_o}^{\infty} g_a(a|d)da \tag{V.47}$$

$$p_{FA}^{\tau_o} = \int_{\tau_o}^{\infty} c_a(a)da \tag{V.48}$$

are the probability of detection and probability of false alarm, respectively. By incorporating the amplitude likelihoods, the PHD update of (V.19) becomes

$$\begin{aligned} D_{t|t}(x_t|Z_{1:t}) = &\left[1 - p_D^{\tau_o}(d)\right] D_{t|t-1}(x_t|Z_{1:t-1}) + \\ &\sum_{\bar{z}_t \in \bar{Z}_t} \frac{\bar{\psi}_{\bar{z}_t}(x_t) D_{t|t-1}(x_t|Z_{1:t-1})}{\bar{\kappa}_t(\bar{z}_t) + \int \bar{\psi}_{\bar{z}_t}(\zeta) D_{t|t-1}(\zeta|Z_{1:t-1})d\zeta} \end{aligned} \tag{V.49}$$

For the case of known d, it is common to model the amplitude likelihoods with Rayleigh distributions

$$g_a^{\tau_o}(a|d) = \frac{a}{1+d}\exp\left(\frac{\tau_o^2 - a^2}{2(1+d)}\right), \quad p_D^{\tau_o}(d) = \exp\left(\frac{-\tau_o^2}{2(1+d)}\right) \tag{V.50}$$

$$c_a^{\tau_o}(a) = a\exp\left(\frac{\tau_o^2 - a^2}{2}\right), \quad p_{FA}^{\tau_o} = \exp\left(\frac{-\tau_o^2}{2}\right) \tag{V.51}$$

Note that the Rayleigh parameter for the clutter model in (V.51) is assumed to be unity which might not be true in general. However, given that the clutter level is known, the amplitudes of the detections can be scaled so that the parameter for the clutter Rayleigh distribution becomes unity while the parameter for the target Rayleigh distribution conforms with the SNR level corresponding to $(1 + d)$. On the other hand, for the case of unknown d, one can marginalize equation (V.50) over a range of possible values $[d_1\ d_2]$ and find a distribution for g_a that is not conditional on d , hence

$$g_a(a) = \int_{d_1}^{d_2} p(\gamma) g_a(a|\gamma) d\gamma \tag{V.52}$$

$$p_D^{\tau_o} = \int_{d_1}^{d_2} p(\gamma) p_D^{\tau_o}(\gamma) d\gamma \tag{V.53}$$

By picking a suitable prior distribution $p(d)$ and assuming $g_a(a|d)$ is Rayleigh distributed with parameter $(d + 1)$, one can obtain a closed form solution to (V.52). The probability of detection $p_D^{\tau_o}$ in (V.53) can then be found using numerical integration offline since it does not need to be computed for every iteration [20].

V.D System Integration

V.D.1 Tracking Task

In the previous two sections we described the front-end (ICA/SCT) and the back-end (PHD filtering) of our system model, respectively. The front-end uses the output of ICA to perform the SCT mapping where peaks that are above some detection threshold are selected. These peaks are declared as DOA measurements or detections and are fed into the PHD filter. The PHD filter then filters the measurements and estimates the DOA and number of targets using the GM-PHD filter assuming that the state dynamics and sensor model for a single source follow a linear and Gaussian model according to equations (V.24-V.25) so that

$$x_t = A_{t-1}x_{t-1} + \nu_{t-1} \tag{V.54}$$

$$z_t = B_t x_t + \vartheta_t \tag{V.55}$$

where $\nu_t \sim \mathcal{N}(0, Q_t)$, $\vartheta_t \sim \mathcal{N}(0, R_t)$ and x_t is the vector of states at time t for a single target model. The dimensions of x_t depends on the number of microphone pairs since each microphone pair has a separate TDOA. Also information about the velocity information of the DOA (denoted as $\dot{\theta}_t$) can be incorporated in the state to represent a constant velocity model. In the most simple case where only one microphone pair is present ($L = 2$) and the velocity information is not considered, the state reduces to $x_t = \theta_t$ with a dimension of unity and the model parameters in equations (V.54-V.55) reduce to $A_{t-1} = 1$ and $B_t = 1$.

Figure V.1 illustrates the system model incorporating ICA/SCT with PHD filtering. As depicted in Figure V.1, ICA is performed on blocks of data in which each block is a collection of a certain number of STFT frames. Note that the time index of the sensor raw data is u, the frame index after converting to the frequency domain using STFT is n and the block index for a collection of frames is t. Any complex-valued ICA algorithm can be used on the blocks. An important note is that the initialization of the ICA iterations for each block should be done from

scratch and not based on the previous block converged values. This is to encourage diversity in the ICA estimates so that if a source dies out or a new source is born, such dynamics can be picked up by ICA and translated to meaningful location information via SCT. To better distinguish between clutter and targets, the GM-PHD filter incorporates the detection amplitudes as described in section V.C.3. Also, the GM-PHD tracker enables data association and track labels as described in section V.C.2. The track labeling is crucial for the separation task, since the track labels will be used to stitch together the ICA outputs from the blocks enabling a separated source for each track.

V.D.2 Separation Task

The key notion that allows us to perform the separation task even for the overcomplete/underdetermined case is assuming block-frequency sparsity of the sources in which the number of source components at each frequency-block segment does not exceed the number of sensors even though the overall number of sources can exceed the number of sensors. However, the estimated mixing matrices, i.e. the immediate output of the ICA stage, contain no valuable separated information of the mixed sources. This is due to the fact that at that stage no inference on number of sources is achieved and the ordering of the columns of the mixing matrices across the frequency bins and time blocks are indeterminate. Since, SCT is invariant to such mismatch in ordering, it is able to translate the mixture model of ICA into a detection model similar to that commonly used in radar/sonar (hence the use of the term 'sources' in the mixture model and 'targets' in the detection model) and from there the GM-PHD filter determines the expected number of sources and provides distinct tracks on the DOA of the sources. Now one can use this information obtained from the output of the PHD filter and feed it back to the output of the ICA stage to effectively carry out the separation task in the following two steps:

Permutation correction in each block across frequencies

The expected number of sources and the estimates of TDOAs obtained from the PHD filter is used to correct for possible permutations. Let's consider the case where we have L sensors and at time block t the PHD filter has declared $\hat{M}_{t|t} \geq L$ sources to be active with corresponding TDOAs $\hat{\tau}_1, ..., \hat{\tau}_{\hat{M}_{t,t}}$. The mixing matrix in each bin for block t has dimensions $L \times L$. From the sparsity assumption, this means that at block t, each frequency bin has at most L columns of its mixing matrix that are active, however at this stage, the ordering of which L out of $\hat{M}_{t|t}$ sources being active is not known. We introduce $\hat{M}_{t|t} - L$ virtual null columns of zeros to represent inactivity. Therefore, for each frequency bin there are a total of $\frac{\hat{M}_{t|t}!}{(\hat{M}_{t|t}-L)!}$ possible permutations of the mixing matrix columns with augmented null columns. We can now use the PHD filter estimates of the TDOA as reference to align the columns of the null augmented mixing matrix for frequency bins $k = 1, ..., K$ as following (similar to [58])

$$\bar{\Pi}(k) = \underset{\Pi}{\operatorname{argmin}} \sum_{m=1}^{\hat{M}_{t|t}} ||c(k, \hat{\tau}_m) - \bar{r}_{\Pi_m}(k)|| \tag{V.56}$$

where Π is a permutation of the mixing matrix as described in (V.5) and $\bar{r}_{\Pi_m}(k)$ is the normalized ratio of the m^{th} column of the matrix affected by permutation Π as described in (V.9). We make the note that for the null columns, the value of such ratios are not defined and one can replace them with a single constant as it does not effect the minimization of (V.56). Once the ordering of the sources is estimated from (V.56), one can perform the separation in the time block for each frequency bin by rearranging the rows of the demixing matrix (inverse of the aforementioned $L \times L$ mixing matrix) to align with their corresponding L active source components and forcing the remainder of inactive source components to zero. Next step would be to determine whether the separated sources at the current block t are newborn sources or surviving ones, and if surviving, to stitch it to the corresponding segments of the same source from the previous block $t-1$.

Stitching segments across blocks

In the previous step we explained how for each time block, the mixing matrix for each frequency bin can be aligned so that each column is linked to a single DOA obtained from the PHD filter. In this step we explain how the components from one time block are stitched to the components from the previous block. If the DOAs of the sources do not undergo any dynamics, then one can use the DOAs themselves to link the ICA components of one block to the previous block. In the case where the DOAs undergo dynamics in terms of both values and birth/death occurrences, then some kind of data association scheme is required to link the DOAs of surviving sources and initiate a new track for newborn sources. The track labeling algorithm described in section V.C.2 using the GM-PHD implementation effectively accomplishes the task of data association, therefore enabling the stitching of sources from one block to another. We note that in such a separation scheme, any newborn source is declared as new source even though, for example, it might be coming from a previous source that underwent a silence period. The feedback arrows in Figure V.1 illustrate the separation task where the PHD tracks are used to go back and perform the alignment of the mixing matrices across the frequencies and the stitching of the source components across the blocks. At the end, in order to regularize the scaling ambiguity of the ICA outputs we use the well-known minimal distortion principle [54] for each block and frequency bin. Once the stitching and scaling of the ICA outputs are performed, the inverse Fourier transform using the overlap add method is used to reconstruct the time domain signals.

V.E Experimental Results

In this section we present some experiments on simulated data for both tasks of interest: Quasi-online tracking and separation of multiple moving sources with birth/death dynamics. The simulated data was obtained using Lehmann's image

method [42] which simulates the impulse response between a source and a sensor for a rectangular room environment. For each task we use a different experimental set-up since each task has a different level of difficulty with the separation task being more difficult in general than the tracking task. Thus we try to introduce experiments so that it would push the complexity envelope for each task in its own context independently.

V.E.1 Tracking Results

We evaluate the performance for the DOA tracking task and compare the proposed method mainly with the two alternative approaches discussed earlier. One method uses the GCC-PHAT at the front-end to acquire detection measurements and the same PHD filter as the proposed back-end for the filtering. The other method uses the same ICA/SCT of the proposed front-end and a naive thresholding method to post-process the detections in the back-end. For the first experiment, the room dimensions used in the simulation are $8m \times 5m \times 2.5m$ with a reverberation time of $T_{60} = 600ms$. Signals were sampled at $f_s = 16kHz$ and the STFT frequency-frame segments were obtained using a Hanning window of size 2048 samples with 87.5% overlap. The blocks in which the ICA was conducted on had a 50% overlap with each block being 0.64 seconds (40 frames) in length. The experiment lasted for a total duration of 15.04 seconds. Only $L = 2$ microphones were used which were placed $36cm$ apart. The speakers could appear and disappear at any time. There were a total of 7 different speakers with the maximum number of 6 concurrent speakers in this experiment. The speakers all moved along a semi-circular path about $1.5m$ from the microphone pair as depicted in Figure V.2. The ICA algorithm carried out for each block was a standard complex valued maximum likelihood Infomax algorithm [34]. Figure V.3 shows the DOA detections and the true source DOAs along with their estimated tracks using the proposed method: front-end (ICA/SCT) + back-end (GM-PHD with amplitude information). The tag or label for each track is represented using a

unique colored shape. Figure V.4 illustrates the results using the "GCC-PHAT + proposed back-end" approach (similar to [46, 9]) while Figure V.5 and Figure V.6 show the results using two variations of the "proposed front-end + naive thresholding" method ([44, 43]). The first variation uses a higher detection threshold compared to the second variation, hence resulting in fewer clutter but bearing the risk of more missed detections. In contrast the second variation results in more clutter but with less missed detections. In addition to [44, 43], we also disregard non-persistent peaks using some distance measure in order to reject isolated clutter. We note that the "proposed front-end + naive thresholding" method is not a tracking technique (but a peak selection scheme), and thus does not solve the data association problem inherently and requires an additional module to do so. That is why the estimates in Figure V.5 and Figure V.6 are not color-shape coded. In order to highlight the importance of incorporating amplitude information in our method, we also run our proposed method without considering any amplitude information: "proposed front-end + GM-PHD without considering amplitude information" and show the result in Figure V.7.

Wasserstein miss distance which is an optimal multi-target error metric for time-varying number of targets is used to evaluate the performances of such experiments [33]. Wasserstein miss distance is optimal in the sense that it intrinsically considers the mismatch in target number and state values. Assuming that $X_t = \{x_1, ..., x_n\}$ is the subset of true states at time t and $\hat{X}_t = \{\hat{x}_1, ..., \hat{x}_m\}$ is the estimated subset of states, the Wasserstein miss distance is defined as

$$d(X_t, \hat{X}_t) = \min_{C} \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} C_{i,j} ||x_i - \hat{x}_j||^2} \quad , \tag{V.57}$$

where the minimum is taken over all $n \times m$ transportation matrices $C = \{C_{i,j}\}$. An $n \times m$ matrix C is a transportation matrix if for all $i = 1, ..., n$ and $j = 1, ..., m$

$$C_{i,j} \geq 0, \quad \sum_{i=1}^{n} C_{i,j} = \frac{1}{m}, \quad \sum_{j=1}^{m} C_{i,j} = \frac{1}{n} \ . \tag{V.58}$$

The minimization in (V.57) means that it gives the distance for the best association between true and estimated set of states, and can be done using standard linear programming algorithms. For the aforementioned experiment with results illustrated using the different methods in Figure V.3-Figure V.7, we present the point-wise and mean-valued Wasserstein distance in Figure V.8. In addition to the methods discussed earlier, Figure V.8 also shows the Wasserstein distance using the raw peaks of the proposed front-end as the DOA estimates. For this experiment, Figure V.8 shows that the proposed algorithm outperforms the other methods. In order to get a better quantification of the robustness and versatility of the proposed method compared to the other methods, the mean Wasserstein distance error is computed for other experiments varying the input signal-to-noise ratio ($\text{SNR}_{\text{input}}$[1]), T_{60} and maximum number of concurrent speakers. Two different noise types were considered. One being additive white Gaussian noise (WGN) and the other being Babble noise. Babble noise was simulated using 9 speakers speaking from 9 different locations distributed across the room, at least $3m$ away from microphone pair, as depicted in Figure V.2. The probability of activity of each one of these babble sources was 80%. The $\text{SNR}_{\text{input}}$ was computed as follows

$$\text{SNR}_{\text{input}}(dB) = 10log_{10}\left(\frac{\sum_{i=1}^{2}\sum_{u}|y_{i,target}(u)|^2}{\sum_{i=1}^{2}\sum_{u}|y_{i,noise}(u)|^2}\right) \tag{V.59}$$

where $y_{i,target}(u)$ and $y_{i,noise}(u)$ are the microphone inputs due to the target sources and the noise (additive WGN or babble sources), respectively. Two different reverberation times $T_{60} = 600ms$ and $300ms$ were considered along with two different scenarios with maximum number of concurrent sources being 6 and 4. The trajectories for the maximum 6 concurrent sources being the same as those depicted in Figure V.3 and the trajectories for the maximum 4 concurrent sources being the "blue", "dark green", "green", "magenta" and "cyan" solid lines of Figure V.3. The results of the experiments for all such different variations in input noise values/types, T_{60} and maximum number of concurrent speakers are presented

[1] note that $\text{SNR}_{\text{input}}$ is different than the SNR in (V.44) which defines the detection amplitude of the targets compared to clutter

in Figure V.9 and Figure V.10. These figures show that the proposed method outperforms the other methods for most scenarios. As the problem becomes least challenging, for example when $T_{60} = 300ms$ with maximum 4 concurrent speakers and $\text{SNR}_{\text{input}} \geq 26dB$ (Figure V.10-Right), the method that uses the "proposed front-end + naive thresholding (high)" performs slightly better compared to the proposed method. In such scenarios the amplitudes of the SCT peaks originating from clutter and targets are more discriminative/separable compared to the more challenging scenarios. Thus by choosing the right threshold, one could effectively reject clutter while keeping the peaks belonging to the targets. As a result, simple peak selection methods, like that of naive thresholding (high), can perform pretty well in such cases and sometimes even perform better compared to multi-target tracking methods. The reason the better performance in such scenarios is that multi-target tracking methods can lose some accuracy in the track initiation and termination (can lag behind 1-3 time updates when initiating and terminating a track), while peak selection methods can promptly identify a target's initiation and termination given that the threshold is correct and the amplitudes of targets and clutter are separable.

We note that the parameters d_1 and d_2 in (V.52) which characterize the amplitude information of the targets in the proposed back-end and the thresholding parameters for the naive thresholding methods were fixed and obtained by inspection based on the physical properties such as the room's T_{60}, STFT frame window size and percentage overlap. Also, as explained in Section V.C.3, it is assumed that the clutter level is known and amplitudes of the detections are scaled so that the parameter for the clutter Rayleigh distribution in (V.51) becomes unity. The assumed clutter level effects the sensitivity of the algorithm: the lower it is, the more likely the algorithm is in declaring a detection as a target and vice versa. Moreover, for all experiments the average number of Poisson-distributed false alarm was $\lambda = 10$, the probability of detection was $p_D = 0.5$, the probability of target survival was $p_S = 0.99$, birth rate was $\sum_i \omega_{b,t}^{(i)} = 0.1$, the process and observation

noise covariances Q_t and R_t were both set to $10I$, where I is the identity matrix.

V.E.2 Separation Results

In this section the separation capabilities of the proposed method is investigated. The room dimensions used in all the simulations in this section were $6m \times 4m \times 2.5m$ and the signals were sampled at $f_s = 16kHz$. We experimented with a total of four scenarios and increased the level of difficulty with each scenario. For the first scenario a reverberation time of $T_{60} = 200ms$ was considered. The STFT frequency-frame segments were obtained using a Hanning window of 1024 taps with 75% overlap. The blocks in which the ICA was conducted on had a 50% overlap with each block being 0.64 seconds (40 frames) in length. Again, only $L = 2$ microphones were used which were placed $6cm$ apart. Since separating the sources is our focus here, we use the recursively regularized ICA algorithm [60] for our ICA module as it has been shown to achieve better separation results for short duration mixtures when compared to regular infomax ICA. A total of 3 speakers were involved in this experiment. Speaker 1 and 3 are active for the total experiment which lasted for 15.04 seconds. Speaker 2 enters the conversation at around the 3 second mark and leaves the conversation at around the 11 second mark. All three speakers moved along a semi-circular path $1m$ from the microphone pair, i.e. with radii $r_i = 1m,\ i = 1, ..., 3$. Figure V.11 illustrates the DOA detections and the true source DOAs along with the estimated tracks for all three speakers. The PEASS toolbox which is a perceptual BSS evaluation software, was used for the performance evaluation of the separated sources [27]. The signal to disturbance ratio (SDR), source image to spatial distortion ratio (ISR), signal to interference ratio (SIR) and signal to artifact ratio (SAR) metrics with the decomposition estimated by the PEASS tool was used for the evaluation of our separated sources. At first no assumption on the number of sources was made. As seen in Figure V.11, the proposed method was able to identify the correct number of sources at the appropriate intervals. Next, in order to create a benchmark for comparison of the

separation results of our proposed algorithm we conduct another experiment where we mute speaker 2 for the entire duration and assume the number of sources is known and equal to two, while all the other settings remain the same. Since now the number of sources is assumed to be known, we use a separate gated Kalman filter for each source to track the DOAs (similar to [18]) which is necessary for permutation correction and stitching across blocks in the separation task. Table V.1 shows the separation performances for the controlled benchmark experiment and the main experiment. The performances indicate that even though the main experiment compared to the benchmark one had an uncertainty factor in the number of sources and also had to deal with more sources than sensors for most of the duration of the experiment, the performance only degraded marginally, suggesting robustness and versatility of the algorithm in coping with dynamic scenarios. For our second scenario, the same is repeated for a set-up where the radii of the sources increase with the source number e.g. $r_1 = 1m$, $r_2 = 1.7m$, $r_3 = 2m$ while all the other settings remain the same, with results shown in Table V.2. In this scenario the separation of the more distant sources becomes more challenging as the tracking algorithm assumes all sources have roughly the same detection amplitude compared to clutter. Table V.2 shows the farthest source experiences the largest drop in performance from the benchmark experiment as expected, nevertheless it does not fail in carrying out the separation task.

Next we consider four sources using two microphones $8cm$ apart and the total experiment lasting for 16.32 seconds. We start with a reverberation time of $T_{60} = 200ms$ for this scenario and jump to $T_{60} = 300ms$ for the last scenario. The radii of the sources were $r_i = 1m,\ i = 1, ..., 4$ with the DOA trajectories depicted in Figure V.12. All the other elements in this scenario were the same as the previous scenarios. Similar to the previous cases we start with the benchmark case of known 2 sources with S_1 and S_4 being only active, then progress upwards to a setting with unknown/time-varying number of sources with S_1, S_3 and S_4 active and finally to a case with unknown/time-varying number of sources with all

Table V.1 Separation results for $r_i = 1m,\ i = 1, ..., 3$, using 2 microphones, $T_{60} = 200ms$

2 Sources, Known

	S_1	N/A	S_3	mean
SDR (dB)	5.0	–	5.6	5.3
ISR (dB)	5.9	–	6.2	6.05
SIR (dB)	13.8	–	17.4	15.6
SAR (dB)	17.1	–	18.6	17.85

3 Sources, Unknown/Time-varying

	S_1	S_2	S_3	mean
SDR (dB)	4.6	5.4	4.3	4.77
ISR (dB)	5.8	8.8	5.9	6.83
SIR (dB)	12.2	7.8	10.5	10.17
SAR (dB)	17.1	16.8	17.1	17.0

4 sources $S_i,\ i = 1, ..., 4$ active. Figure V.12 shows the estimated DOA tracks for the latter case using the proposed method. Table V.3 presents the performance evaluation results for this scenario. In the last scenario the same is repeated but for a more challenging scenario with $T_{60} = 300ms$. In order to better cope with the increase in reverberation, a larger STFT window size of 2046 samples with 87.5% overlap was utilized. The results of the last scenario are shown in Table V.4. Table V.3 and Table V.4 demonstrate the flexibility of the proposed algorithm in separating unknown time-varying number of moving sources for overcomplete situations with twice as many concurrent sources as sensors.

V.F Summary and Discussion

In this chapter we present a novel framework to solve the problem of tracking and separation of unknown time-varying number of speakers using minimal number of microphones in a reverberant environment. We proposed the integration of a powerful and versatile ICA-based scanning method for multiple DOA estimation with a well known method in multi-target tracking. Such combination

Table V.2 Separation results for $r_1 = 1m$, $r_2 = 1.7m$, $r_3 = 2m$, using 2 microphones, $T_{60} = 200ms$

2 Sources, Known

	S_1	N/A	S_3	mean
SDR (dB)	5.9	–	3.5	4.7
ISR (dB)	6.8	–	5.3	6.05
SIR (dB)	15.8	–	8.9	12.35
SAR (dB)	17.3	–	14.7	16.0

3 Sources, Unknown/Time-varying

	S_1	S_2	S_3	mean
SDR (dB)	5.5	4.3	2.2	4.0
ISR (dB)	6.7	6.2	4.6	5.83
SIR (dB)	13.2	8.3	5.3	8.93
SAR (dB)	18.3	15.7	15.2	16.4

Table V.3 Separation results for $r_i = 1m,\ i = 1, ..., 4$ using 2 microphones, $T_{60} = 200ms$

2 Sources, Known

	S_1	N/A	N/A	S_4	mean
SDR (dB)	3.6	–	–	3.9	3.75
ISR (dB)	4.8	–	–	4.5	4.65
SIR (dB)	11.7	–	–	13.6	12.65
SAR (dB)	14.6	–	–	16.5	15.55

3 Sources, Unknown/Time-varying

	S_1	N/A	S_3	S_4	mean
SDR (dB)	3.4	–	4.0	3.5	3.63
ISR (dB)	4.7	–	7.4	4.5	5.53
SIR (dB)	11.1	–	6.4	10.7	9.4
SAR (dB)	15.1	–	14.7	15.6	15.13

4 Sources, Unknown/Time-varying

	S_1	S_2	S_3	S_4	mean
SDR (dB)	2.8	4.4	3.3	3.4	3.46
ISR (dB)	4.6	7.3	6.6	4.4	5.73
SIR (dB)	8.4	7.3	5.4	10.4	7.86
SAR (dB)	15.4	15.2	14.4	15.9	15.23

Table V.4 Separation results for $r_i = 1m,\ i = 1, ..., 4$ using 2 microphones, $T_{60} = 300ms$

2 Sources, Known

	S_1	N/A	N/A	S_4	mean
SDR (dB)	2.4	–	–	3.7	3.05
ISR (dB)	3.8	–	–	4.9	4.35
SIR (dB)	9.2	–	–	11.5	10.35
SAR (dB)	10.2	–	–	12.1	11.15

3 Sources, Unknown/Time-varying

	S_1	N/A	S_3	S_4	mean
SDR (dB)	1.6	–	2.5	2.6	2.23
ISR (dB)	3.6	–	5.0	4.2	4.27
SIR (dB)	7.0	–	4.8	7.6	6.47
SAR (dB)	9.8	–	10.9	10.8	10.5

4 Sources, Unknown/Time-varying

	S_1	S_2	S_3	S_4	mean
SDR (dB)	1.0	1.1	1.9	2.3	1.58
ISR (dB)	3.5	3.3	4.2	4.0	3.75
SIR (dB)	4.3	2.0	3.4	7.1	4.2
SAR (dB)	10.1	10.2	11.2	11.7	10.8

showed promising results in both the tracking and separation tasks using only two microphones for relatively high reverberant environments and in challenging dynamic scenarios involving moving sources and spontaneous births/deaths.

This chapter demonstrates how a mixture/superposition model in the framework of BSS can be easily represented as a standard detection model in the framework of multi-target tracking, assuming that the sources have block-frequency sparsities. The solution involves, first, performing frequency-domain ICA on the sensor measurements, then utilizing a permutation-invariant TDOA scanning method such as SCT on the ICA outputs, therefore enabling the mixture model observations to be represented as source location observations in a detection model. The PHD filter, which has proven to be a highly effective method for multi-target tracking when observations are posed in a detection model, is then used for the tracking of the location detections. The post-filtered DOAs are then used to align and stitch the ICA outputs across frequencies and blocks, respectively.

As part of our future work we would like to extend our work performing tracking and separation using multiple sensor pairs hence representing the TDOA measurements in a multidimensional framework using GSCT. One advantage of incorporating multiple dimensional TDOAs is that one can provide more detailed location information and possibly extract the Cartesian location information of the sources. Also the extra dimensions in the measurement model can provide better discrimination in track labeling when sources cross over or get very close. The other advantage would be in the separation task since the extra sensors used can allow for the improvement in the separation performance.

V.G Acknowledgments

The text of Chapter III, in full, is based on the material as it appears in: Alireza Masnadi-Shirazi and Bhaskar Rao, "An ICA-SCT-PHD Filter Approach for Tracking and Separation of Unknown Time-Varying Number of Sources," to

appear in IEEE Transactions on Audio, Speech and Language Processing. The dissertation author was a primary researcher and an author of the cited material.

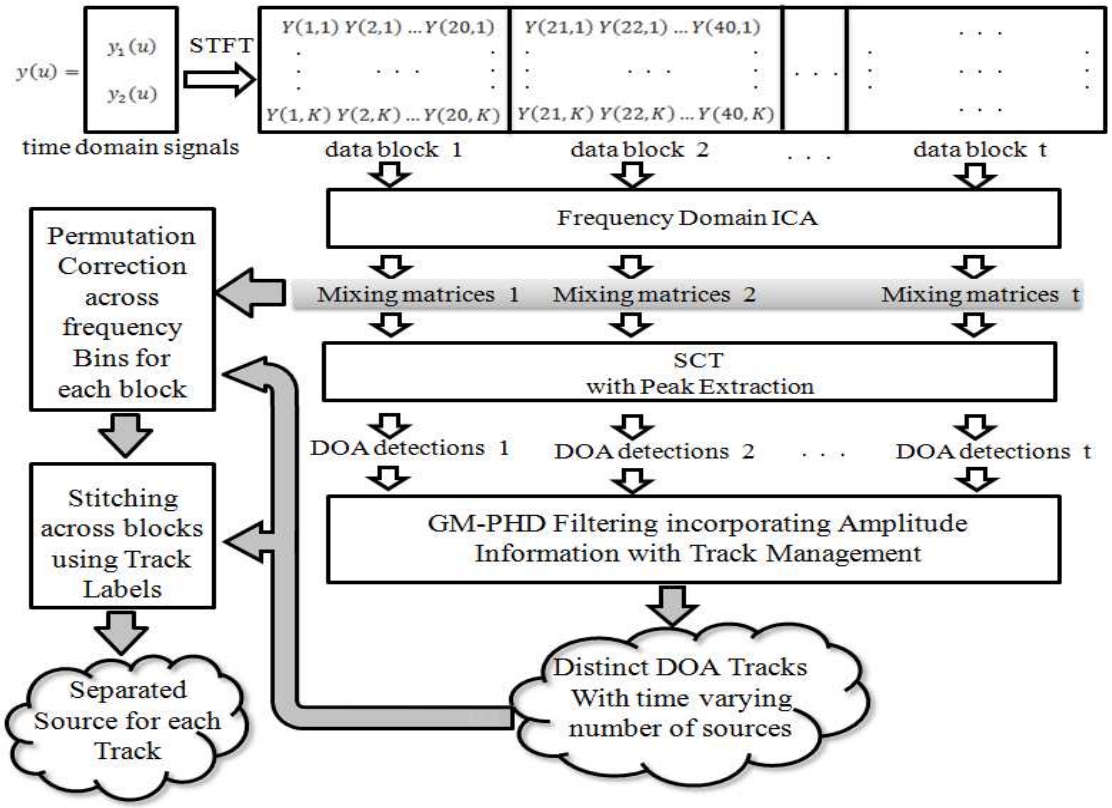

Figure V.1 Block diagram of proposed method: STFT, ICA and SCT segments form the front-end and the PHD filtering segment form the back-end. The feedback from the back-end to the front-end describes the separation task which uses the distinct estimated tracks to perform permutation correction across frequencies for each block and to stitch together the separated components from one block to another.

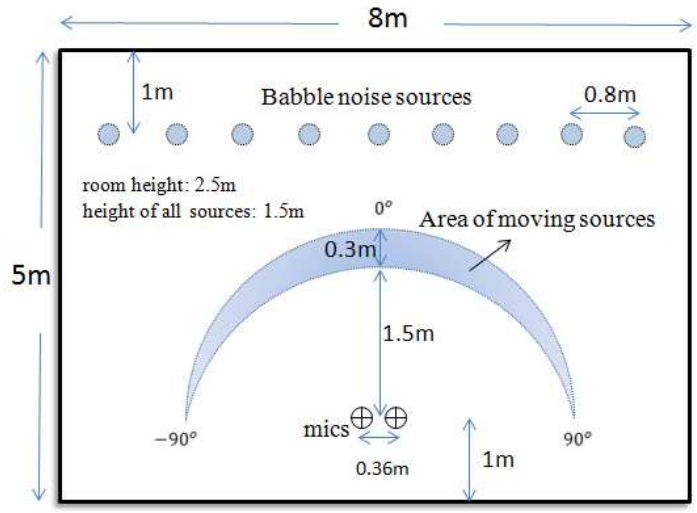

Figure V.2 Room set-up (not drawn to scale). Note that the source trajectories are not shown but rather the area of motion is illustrated. The reason for this is that their activities are time-varying. Refer to Figure V.3 for their true activities and trajectories in terms of DOA.

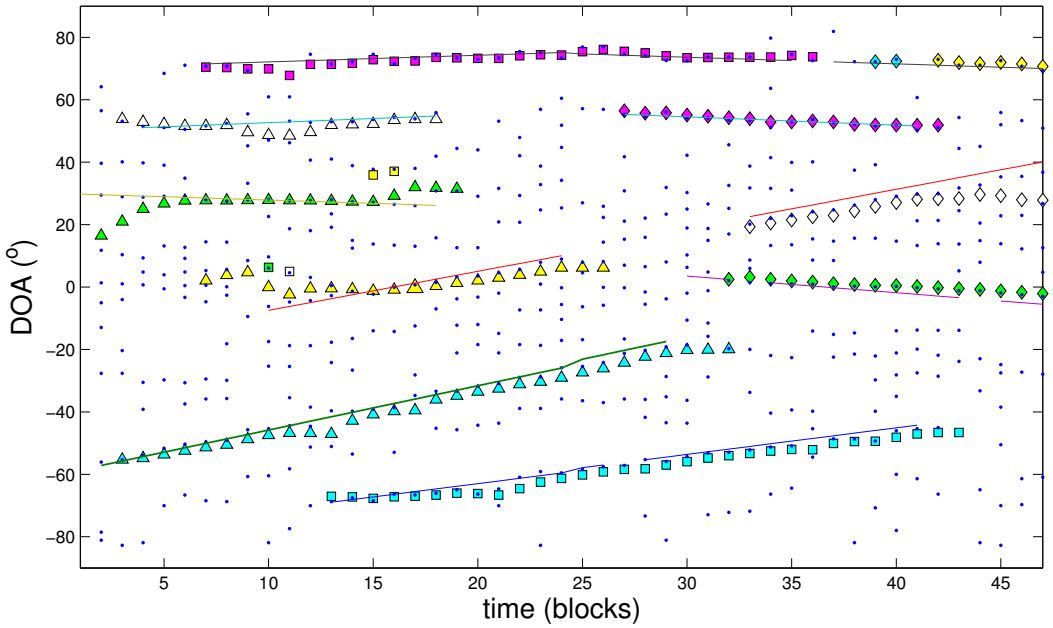

Figure V.3 Proposed method: front-end (ICA/SCT) + back-end (GM-PHD with amplitude information). True DOA (colored lines), SCT peaks (dots) and estimated DOA tracks(colored shapes).

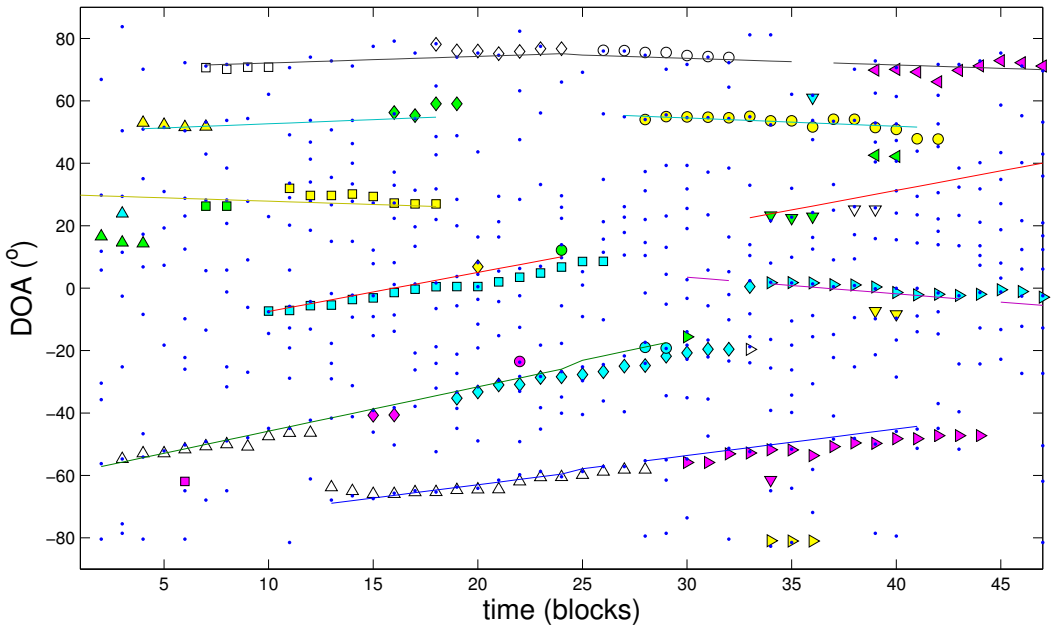

Figure V.4 GCC-PHAT+proposed back-end: True DOA (colored lines), GCC-PHAT peaks (dots) and estimated DOA tracks (colored shapes).

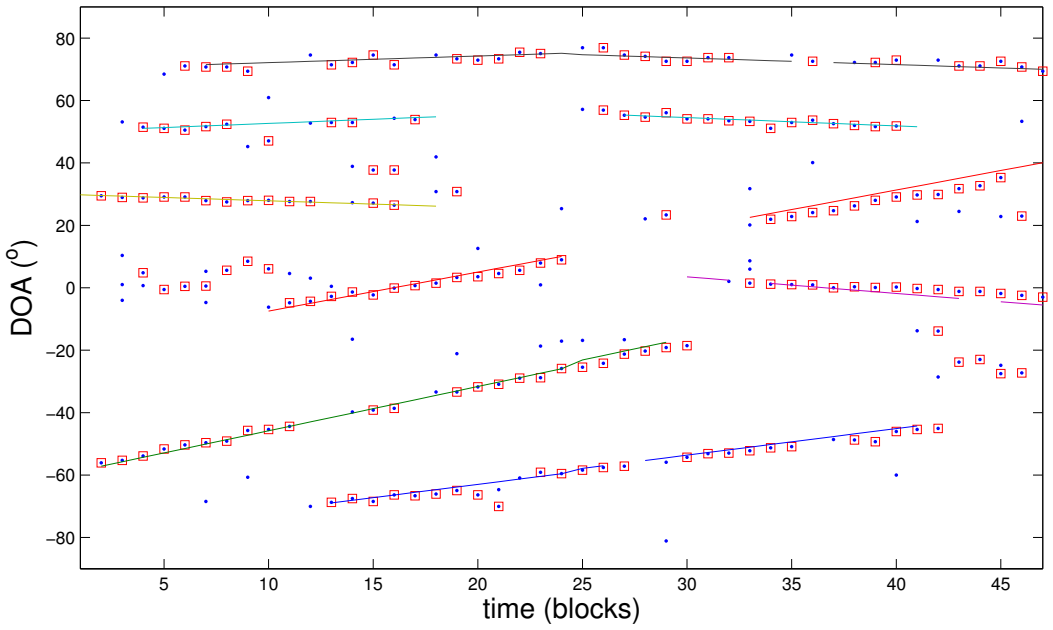

Figure V.5 proposed front-end + naive thresholding (high): True DOA (colored lines), SCT peaks after thresholding (dots) and selected peaks (squares).

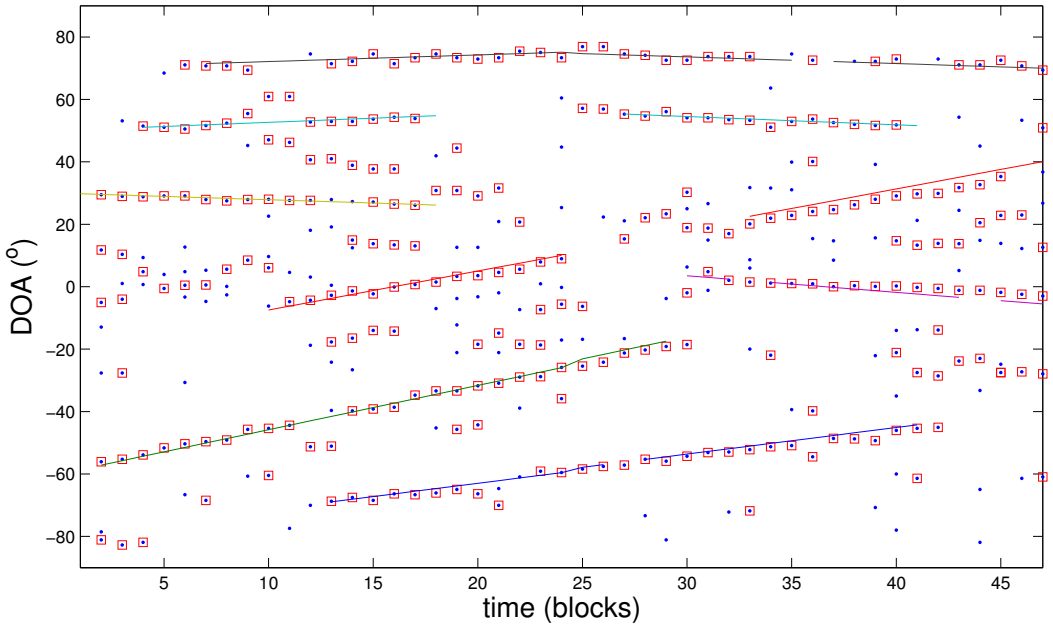

Figure V.6 proposed front-end + naive thresholding (low): True DOA (colored lines), SCT peaks after thresholding (dots) and selected peaks (squares).

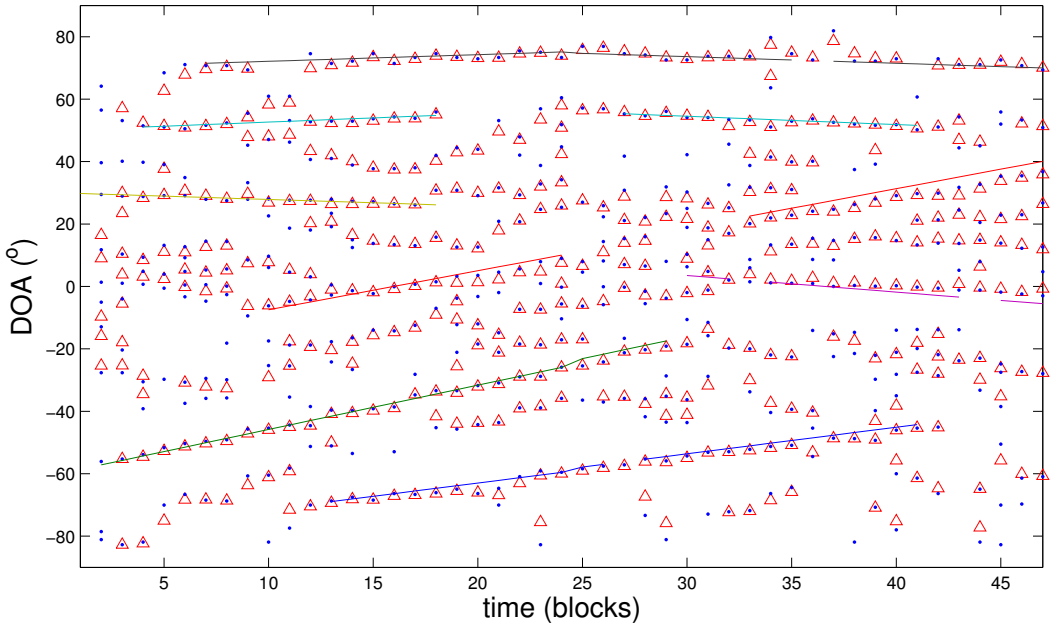

Figure V.7 proposed front-end + GM-PHD filtering without considering amplitude information: True DOA (colored lines), SCT peaks (dots) and DOA estimates (triangles).

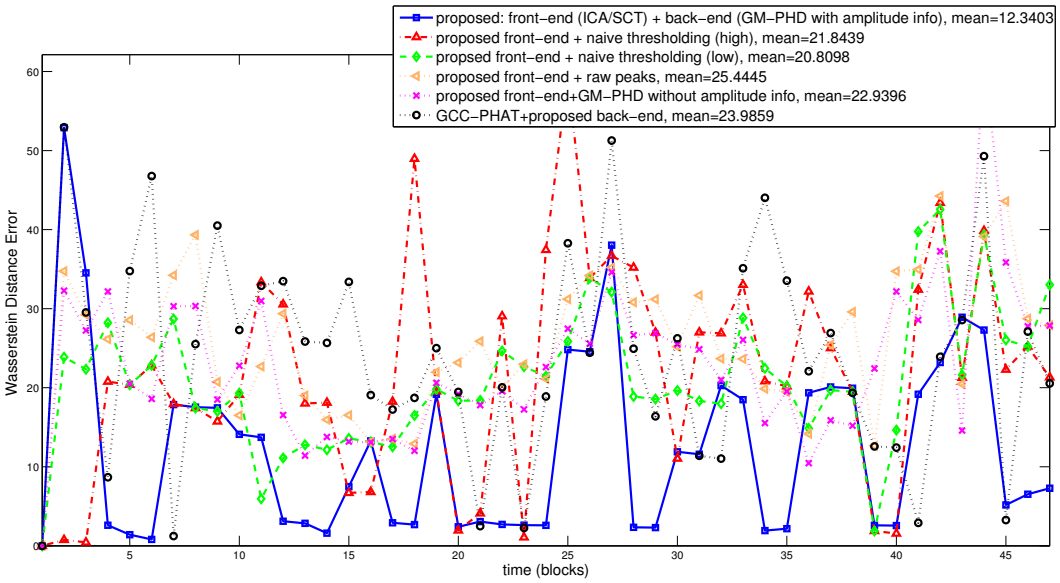

Figure V.8 Wasserstein miss distance error for different methods of Figure V.3-Figure V.7, maximum 6 concurrent sources, $T_{60} = 600ms$.

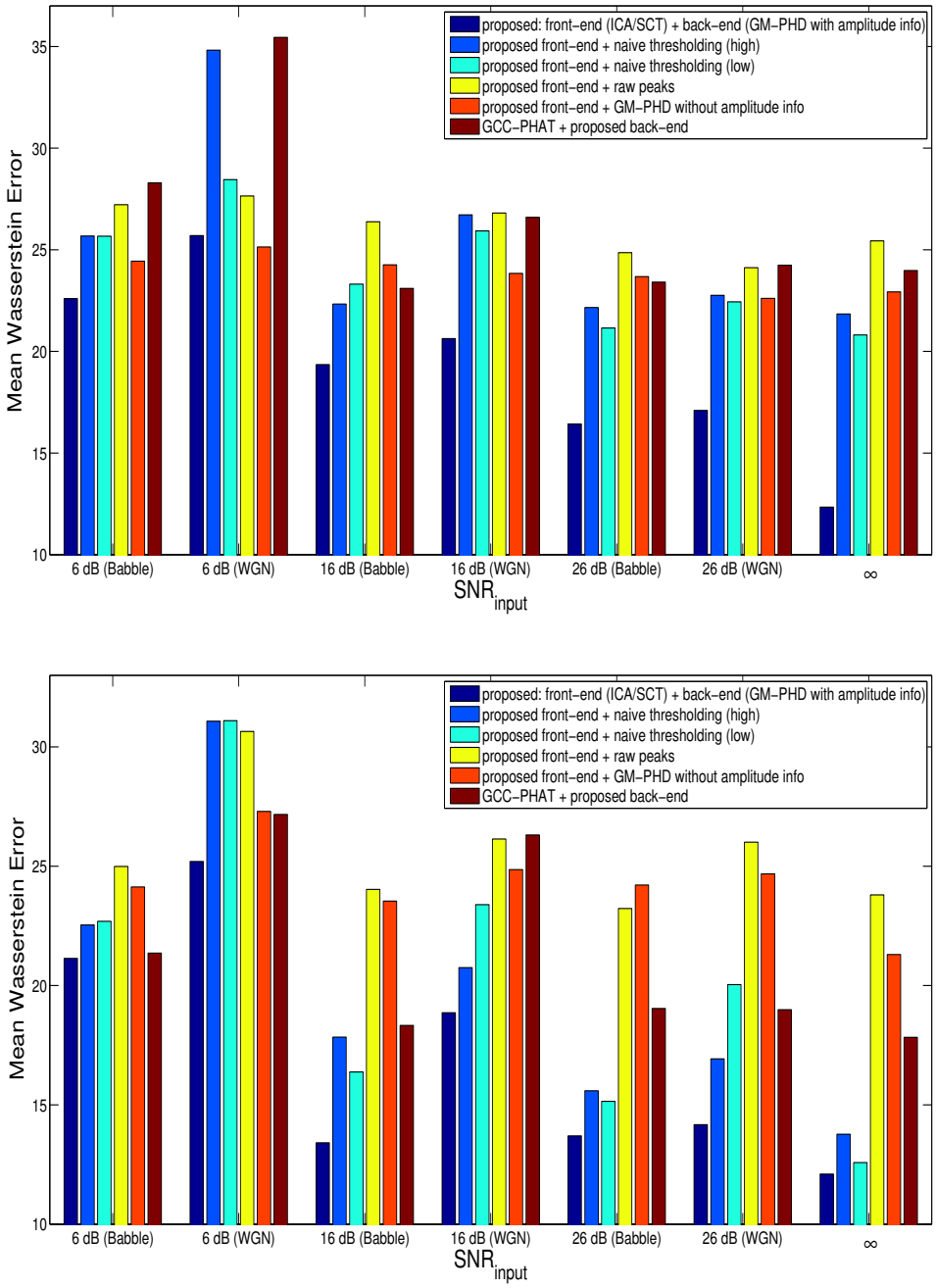

Figure V.9 Performance evaluation for different noise values/types for maximum 6 concurrent sources. Top: $T_{60} = 600ms$, Bottom: $T_{60} = 300ms$.

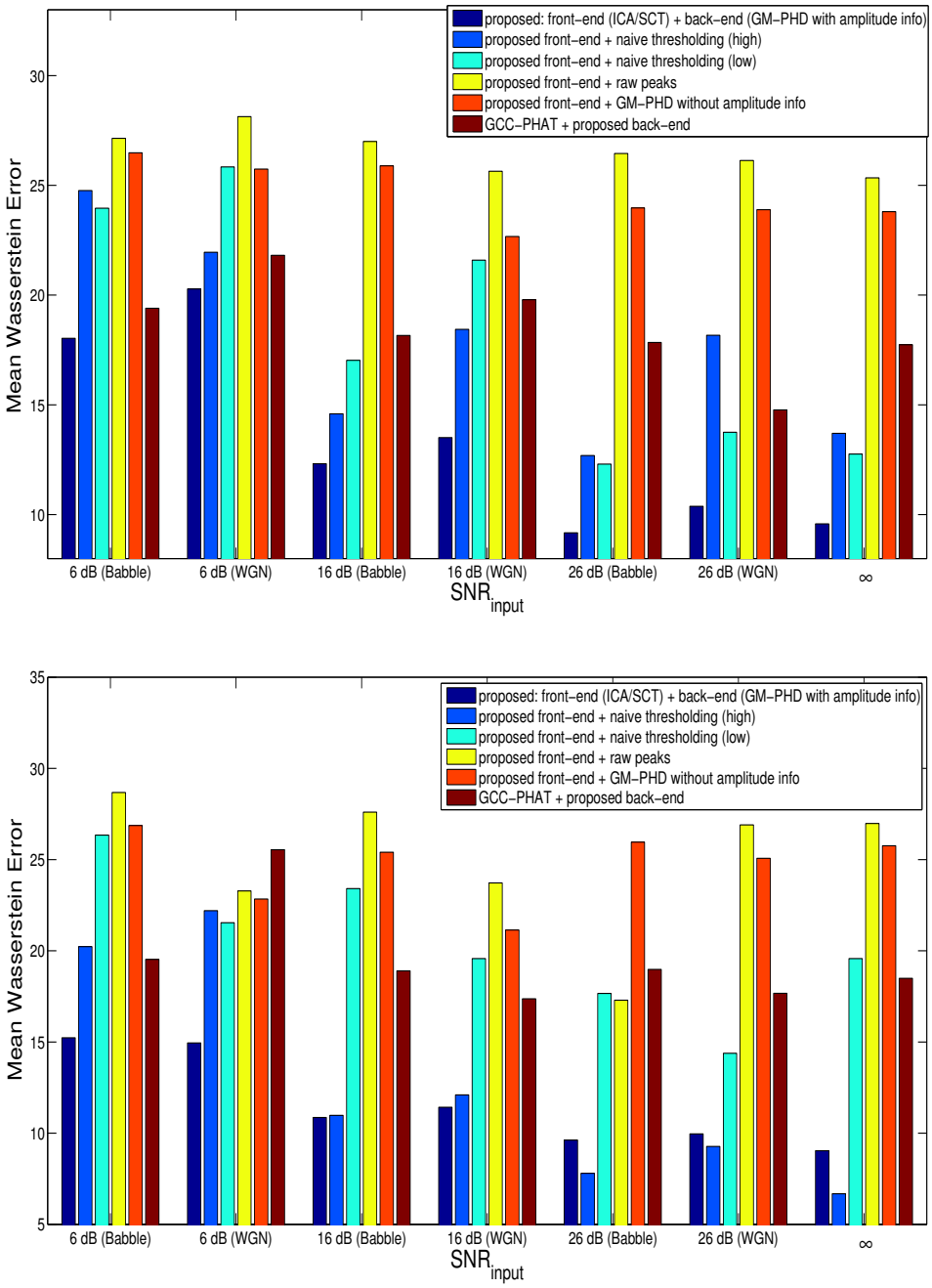

Figure V.10 Performance evaluation for different noise values/types for maximum 4 concurrent sources. Top: $T_{60} = 600ms$, Bottom: $T_{60} = 300ms$.

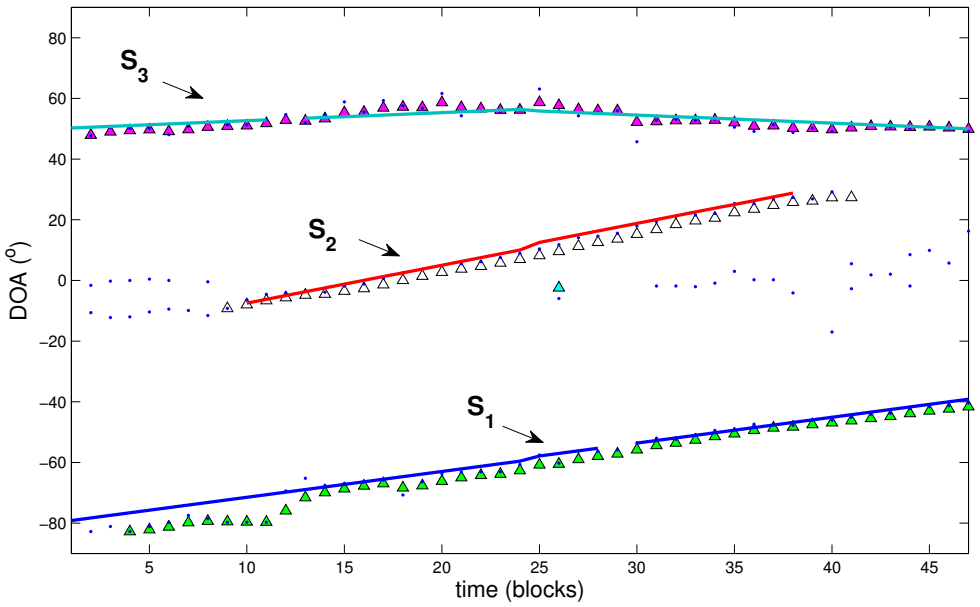

Figure V.11 Separation experiment with 3 unknown time-varying sources and 2 microphones, $T_{60} = 200ms$: True DOA (colored lines), SCT peaks (dots) and estimated tracks (colored shapes).

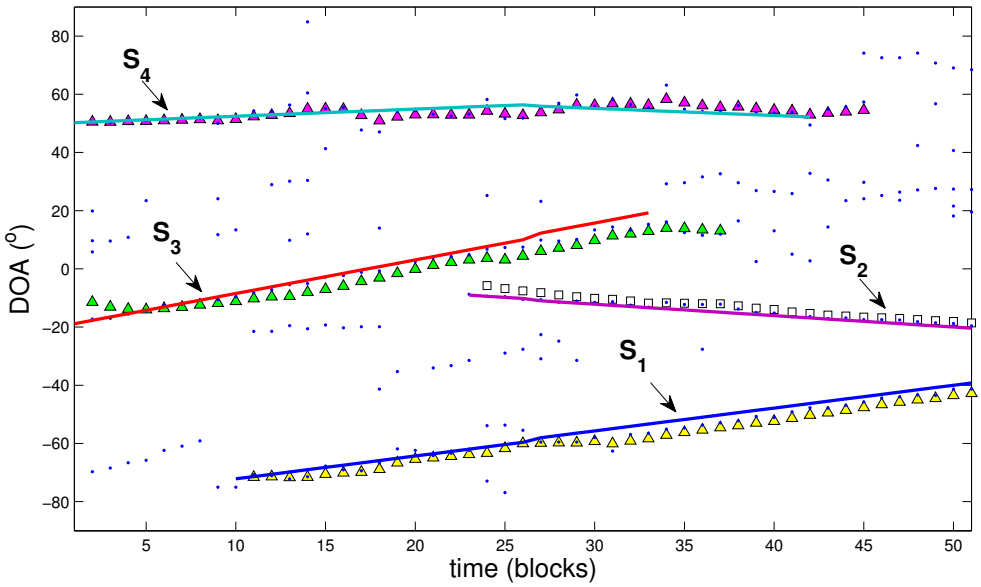

Figure V.12 Separation experiment with 4 unknown time-varying sources and 2 microphones, $T_{60} = 200ms$: True DOA (colored lines), SCT peaks (dots) and estimated tracks (colored shapes).

Chapter VI

Conclusions

In this thesis we have presented three methods based on frequency domain independent component analysis (FD-ICA) that tackle the problem of separation and tracking of multiple speakers in typical real world but adverse scenarios such as having more speakers than microphones (overcomplete case), speakers experiencing temporal dynamics (silence periods in the speech or new speakers entering the room and existing speakers leaving the room) and spatial dynamics (speakers moving/maneuvering). The first method deals with the case where the number of speakers exceeds the number of microphones and speakers can have silence periods intermittently but are spatially static. By mimicking the separation strategy of the human hearing system, it is able to exploit the local decrease of degeneracy during the different combinations of silent gaps of the speakers allowing it to cover all possible states from when all speakers are active to when only one is active at each instant, therefore doing its best to compensate for the apparent global degeneracy. The algorithm works naturally by learning the columns of the mixing matrices in a specialized fashion based on the probability of being in each state. One downside of the first method is that the number of states, and along with it the computational cost, will grow exponentially as the number of speakers increases. This issue is fixed by shifting to a new paradigm in the third method.

The second method, is an online extension of the first algorithm for when the speakers are moving in space. It uses a multiple model particle filter (MMPF) to track the mixing matrices while being able to switch between different combinations of states when the speaker(s) become inactive, therefore avoiding losing track during such periods. The algorithm is also capable of recovering tracks during silence blind zones (SBZ) where the speakers are moving while silent under the condition that the silence gaps are not too long. Once the mixing matrices are correctly estimated, obtaining the directions of arrival (DOA) becomes a straightforward post-processing step. Using a secondary array positioned elsewhere in the room, the DOAs are matched up and triangulated by incorporating a motion model on the speaker trajectories. A drawback of the second method is that it if

the silence period of a speaker is too long it can lose track once the speaker comes back on. Also, it has to know the maximum number of speakers it is tracking otherwise it completely breaks down. In many real world problems the number of speakers at each given time is not known as speakers can either enter the scene or leave. These issues are solved in the third method.

The third method deals with the case where not only the speakers can be moving in space, but also the number of speakers is unknown in an overcomplete setting and can vary with time as new speakers can be born and existing speakers can die out. This is done by introducing a new paradigm where we transform the mixture/superposition model in the framework of ICA to a standard detection model in the framework of multi-target tracking, by exploiting the sparse spectral dynamics of speech. More precisely, the solution involves, first, performing frequency-domain ICA in a quasi-online manner in blocks of data, then utilizing a permutation-invariant TDOA scanning method such as state coherence transform (SCT) on the ICA outputs, therefore enabling the mixture model observations to be represented as speaker location observations in a detection model. The probability hypothesis density (PHD) filter, which is proven to be a highly effective method for multi-target tracking when observations are posed in a detection model, is then used for the tracking of the location detections. The post-filtered DOAs are then used to align and stitch the ICA outputs across frequencies and blocks, respectively, thus enabling the separation task to be carried out. It is worthy to note that the PHD filter, unlike the previous two methods discussed earlier, does not suffer from a combinatorial problem where the computational costs grows exponentially as the number of speakers increases. Also, in contrast to the second method, the third method does not model the silence periods explicitly but rather focuses on the birth and deaths of the speakers. However, since it incorporates a quasi-online method that considers missed detections in the PHD filter phase, short pauses in the speakers are implicitly taken care of and if the pause/silence period become too long, the speaker will be assigned a new track once it becomes

active again.

Bibliography

[1] F. Abrard and Y. Deville, "a timefrequency blind signal separation method applicable to underdetermined mixtures of dependent sources," *Signal Processing*, vol. 85, no. 7, pp. 1389–1403, 2005.

[2] J. Allen and D. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, 1979.

[3] D. Andrews and C. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society*, vol. 36, 1974.

[4] D. Angelosante and M. L. E. Biglieri, "Multiuser detection in a dynamic environment: part II: Joint user identification and parameter estimation," *IEEE Trans. on Information Theory*, vol. 55, no. 5, 2009.

[5] F. Antonacci, D. Riva, D. Saiu, A. Sarti, M. Tagliasacchi, and S. Tubaro, "Tracking multiple acoustic sources using particle filtering," in *Proc. of EUSIPCO*, 2006.

[6] S. Araki, S. Makino, A. Blin, R. Mukai, and H. Sawada, "Underdetermined blind separation for speech in real environments with sparseness and ICA," in *Proc. ICASSP*, 2004, pp. 881–884.

[7] S. Araki, H. Sawada, R. Mukai, and S. Makino, "A novel blind source separation method with observation vector clustering," in *Proc. IWAENC*, 2005, pp. 117–120.

[8] H. Attias, "Independent factor analysis," *Neural Comp.*, vol. 11, 1999.

[9] S. S. B.-N. Vo and W. K. Ma, "Tracking multiple speakers using random sets," in *Proc. of ICASSP*, 2004, pp. 357–360.

[10] B. Balakumar, A. Sinha, T. Kirubarajan, and J. P. Reilly, "PHD filtering for tracking an unknown number of sources using an array of sensors," in *IEEE Workshop on Stat. Sign. Proc.*, 2005, pp. 43–48.

[11] D. Bechler and K. Kroschel, "Considering the second peak in the GCC function for multi-source TDOA estimation with microphone array," in *Proc. International Workshop on Acoustic Echo and Noise Control*, 2003, pp. 315–318.

[12] E. Biglieri and M. Lops, "Multiuser detection in a dynamic environment part I: User identification and data detection," *CoRR*, 2007.

[13] C. Bishop, *Pattern Recognition and Machine Learning*. New York, Springer, 2006.

[14] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations," *Sign. Process.*, vol. 81, no. 11, pp. 2353–2362, 2001.

[15] A. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[16] D. Brillinger, *Time Series: Data Analysis and Theory*. Holt, Rinehart and Winston, Inc., 1975.

[17] A. Bruckstein, D. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, 2009.

[18] A. Brutti and F. Nesta, "Multiple source tracking by sequential posterior kernel density estimation through GSCT," in *Proc. of EUSIPCO*, 2011, pp. 259–263.

[19] A. Brutti, M. Omologo, and P. Svaizer, "Multiple source localization based on acoustic map de-emphasis," *EURASIP Journal on Audio, Speech, and Music Processing*, 2010.

[20] D. Clark, B. Ristic, B.-N. Vo, and B.-T. Vo, "Bayesian multi-object filtering with amplitude feature likelihood for unknown object SNR," *IEEE Trans. on Signal Processing*, vol. 58, no. 1, 2010.

[21] M. Cooke, "Glimpsing speech," *J. phoetics*, vol. 31, pp. 579–584, 2003.

[22] ——, "Making sense of everyday speech: a glimpsing account," in *Speech Separation by Humans and Machines, P. Divenyi (Ed.)*. Kluwer Acedemic Publishers, 2005, pp. 305–314.

[23] M. Davies and N. Mitianoudis, "Simple mixture model for sparse overcomplete ICA," *Proc. IEE Vision, Image, Signal Process.*, vol. 151, no. 1, pp. 35–43, 2004.

[24] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[25] Y. Deville, "Temporal and time-frequency correlation-based blind source separation methods," in *Proc. ICA*, 2003, pp. 1059–1064.

[26] T. Eltoft, T. Kim, and T.-W. Lee, "Multivariate scale mixture of gaussians modeling," in *Proc. ICA*, 2006, pp. 799–806.

[27] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 7, 2011.

[28] R. Everson and S. Roberts, "Particle filters for non-stationary ICA," in *Independent Component Analysis, Principles and Practice*, 2001.

[29] M. Fallon and S. Godsill, "Multi target acoustic source tracking using track before detect," in *Proc. of WASPAA*, 2007, pp. 102–105.

[30] ——, "Multi target acoustic source tracking with an unknown and time varying number of targets," in *Proc. of HSCMA*, 2008, pp. 77–80.

[31] J.-I. Hirayama and S.-I. M. S. Ishii, "Markov and semi-markov switching of source appearances for nonstationary independent component analysis," *IEEE Trans. on Neural Networks*, vol. 18, no. 5, pp. 1326–1342, 2007.

[32] A. Hiroe, "Solution of permutation problem in frequency domain ica, using multivariate probability density functions," in *Proc. ICA*, 2006, pp. 601–608.

[33] J. Hoffman and R. Mahler, "Multi-target miss distance via optimal assignment," *IEEE Trans. Syst. Man Cybermetics*, vol. 3, no. 9, 2004.

[34] A. Hyvarinen, J. Karhunen, and E. Oja, *Indepedent Component Analysis*. New York, Wiley Interscience, 2001.

[35] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Trans. on Circuits and Systems*, vol. 57, no. 7, 2010.

[36] T. Kim, H. Attias, and T.-W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Trans. Speech, Audio and Language Processing*, vol. 15, no. 1, pp. 70–79, 2007.

[37] T. E. T. Kim and T.-W. Lee, "On the multivariate laplace distribution," *IEEE Signal Processing letters*, vol. 13, no. 5, pp. 300–303, 2006.

[38] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, no. 4, 1976.

[39] G. Lathoud and J. Odobez, "Short-term spatio-temporal clustering applied to multiple moving speakers," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 15, no. 5, 2007.

[40] I. Lee, J. Hao, and T.-W. Lee, "Adaptive independent vector analysis for the seperation of convoluted mixtures using EM algorithm," in *IEEE Proc. of ICASSP*, 2008, pp. 803–806.

[41] T. Lee, M. Lewicki, M. Girolami, and T. Sejnowski, "Blind source separation of more ssources than mixtures using overcomplete representations," *IEEE Sign. Process. Letters*, vol. 6, no. 4, 1999.

[42] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image source model," *J. of the Acoustical Soc. of America*, vol. 124, no. 1, 2008.

[43] B. Loesch and B. Yang, "Adaptive segmentation and separation of determined convolutive mixtures under dynamic conditions," in *Proc. of LVA/ICA*, 2010, pp. 41–48.

[44] ——, "Blind source separation based on time-frequency sparseness in the presence of spatial aliasing," in *Proc. of LVA/ICA*, 2010, pp. 1–8.

[45] P. Loizou, *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[46] W.-K. Ma, B.-N. Vo, S. Singh, and A. Baddeley, "Tracking and unknown time-varying number of speakers using TDOA measurements: a random finite set approach," *IEEE Trans. Signal Process.*, vol. 54, no. 9, 2006.

[47] R. Mahler, "Multi-target Bayes filtering via first-order multi-target moments," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 39, no. 4, 2003.

[48] ——, *Statistical multisource multitarget information fusion*. Norwood, MA, Artech House, 2007.

[49] ——, "CPHD filters for superpositional sensors," *O. E. Drummond (ed.), Sign. and Data Proc. of Small Targets 2009, SPIE Proc.*, vol. 7445, 2009.

[50] A. Masnadi-Shirazi and B. Rao, "Independent vector analysis incorporating active and inactive states," in *IEEE Proc. of ICASSP*, 2009, pp. 1837–1840.

[51] ——, "Separation and tracking of multiple speakers in a reverberant environment using a multiple model particle filter glimpsing method," in *Proc. of ICASSP*, 2011, pp. 2516–2519.

[52] A. Masnadi-Shirazi, W. Zhang, and B. Rao, "Glimpsing IVA: a framework for overcomplete/complete/undercomplete convolutive source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18, no. 7, 2010.

[53] M. A. Masnadi-Shirazi, S. Banani, A. Masnadi-Shirazi, and R. Rezaie, "Separation and tracking of maneuvering sources with ica and particle filters using a new switching dynamic model," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 46, no. 3, 2010.

[54] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source seperation," in *Proc. ICA*, 2001, pp. 803–806.

[55] N. Mitianoudis and T. Stathaki, "Batch and online underdetermined source separation using laplacian mixture models," *IEEE Tran. on Speech, Audio and Language Processing*, vol. 15, no. 6, pp. 1818–1832, 2007.

[56] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, pp. 1–24, 2001.

[57] F. Nesta and M. Omologo, "Generalized state coherence transform for multidimensional TDOA estimation of multiple sources," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 1, 2012.

[58] F. Nesta, M. Omologo, and P. Svaizer, "Multiple TDOA estimation by using a state coherence transform for solving the permutation problem in frequency-domain BSS," in *Proc. of MLSP*, 2008.

[59] F. Nesta, P. Svaizer, and M. Omologo, "Robust two-channel TDOA estimation for multiple speaker localization by using recursive ICA and a state coherence transform," in *Proc. of ICASSP*, 2009.

[60] ——, "Convolutive BSS of short mixtures by ICA recursively regularized across frequencies," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, no. 3, 2011.

[61] P. D. O'Grady and B. A. Pearlmutter, "Soft-lost: EM on a mixture of oriented lines," in *Proc. ICA*, 2004, pp. 430–436.

[62] P. O'grady and B. A. Pearlmitter, "Hard-lost: modified K-means for oriented lines," in *Proc. Irish Signals Syst. Conf.*, 2004.

[63] R. Olsson and L. Hansen, "Probabilistic blind deconvolution of non-stationary sources," in *Proc. EUSIPCO*, 2004, pp. 1697–1700.

[64] ——, "Blind separation of more sources than sensors in convolutive mixtures," in *Proc. ICASSP*, 2006, pp. V657–V660.

[65] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum based technique," in *Proc. of ICASSP*, 1994, pp. 273–276.

[66] J. A. Palmer, K. Kreutz-Delgado, and S. Makeig, "Probabilistic formulation of independent vector analysis using complex gaussian scale mixtures," in *Proc. ICA*, 2009, pp. 90–97.

[67] K. Panta, D. Clark, and B.-N. Vo, "Data association and track management for the Gaussian mixture probability hypothesis density filter," *IEEE Trans. on Aerospace and Electronic systems*, vol. 45, no. 3, 2009.

[68] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 3, pp. 320–327, 2000.

[69] ——, "Separation of nonstationary natural signals," in *Independent Components Analysis: Principles and Practice, C. Roberts and R. Everson (Eds.).* Cambridge Univ. Press, 2001, pp. 135–157.

[70] M. Pedersen, D. Wang, J. Larsen, and U. Kjems, "Separating underdetermined convolutive speech mixtures," in *Proc. ICA*, 2006, pp. 674–681.

[71] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of nonstationary sources," *IEEE Trans. Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001.

[72] D.-T. Pham, C. Serviere, and H. Boumaraf, "Blind separation of speech mixtures based on nonstationarity," in *Proc. Seventh International Symposium on Signal Processing and Its Applications*, vol. 2, 2003, pp. 73–76.

[73] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[74] B. Ristik, S. Arulampalam, and N. Gordon, *Beyond the Kalman filter: Particle filters for tracking applications.* Artech House, 2004.

[75] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency domain BSS," in *Proc. ISCAS*, 2007, pp. 3247–3250.

[76] ——, "A two stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2007, pp. 139–142.

[77] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of dominant target sources using ICA and time-frequency masking," *IEEE Trans. on Speech, Audio and Language Processing*, vol. 14, no. 6, pp. 2165–2173, 2006.

[78] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem," *IEEE Trans. on Speech, Audio and Language Processing*, vol. 12, no. 5, pp. 530–538, 2004.

[79] H. Sawada, R. Mukai, and S. Makino, "Direction of arrival estimation for multiple source signals using independent component analysis," in *Proc. of ISSPA*, 2003.

[80] I. Takigawa, M. Kudo, and J. Toyama, "Performance analysis of minimum l_1 norm solutions for underdetermined source separation," *IEEE Trans. Signal Processing*, vol. 52, no. 3, pp. 582–591, 2004.

[81] P. Teng, A. Lambard, and W. Kellermann, "Disambiguation in multidimensional traking of multiple acoustic sources using a gaussian likelihood criterion," in *IEEE Proc. of ICASSP*, 2010, pp. 145–148.

[82] F. Thouin, S. Nannuru, and M. Coates, "Multi-target tracking for measurement models with additive contributions," in *Proc. of Int'l Conf. on Information Fusion*, 2011.

[83] J.-M. Valin, F. Michaud, and J. Rouat, "Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *Robotics and Autonomous Systems Journal (Elsevier)*, vol. 55, no. 3, 2007.

[84] L. Vielva, D. Erdogmus, , and J. C. Principe, "Underdetermined blind source separation using a probabilistic source sparsity model," in *Proc. ICA*, 2001, pp. 675–679.

[85] B.-N. Vo and W. K. Ma, "The Gaussian mixture probability hypothesis density filter," *IEEE Trans. Signal Process.*, vol. 54, no. 11, 2006.

[86] B.-N. Vo, S. Singh, and A. Doucet, "Sequential Monte Carlo methods for multi-target filtering with random finite sets," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 41, no. 4, 2005.

[87] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "Underdetermined blind source separation of convolutive mixtures by hierarchical clustering and L1-norm minimization," in *Blind Speech Separation, S. Makino, T.-W. Lee and H. Sawada (Eds.)*. Springer Netherlands, 2007.

[88] J. A. wnd G. Lathoud and L. McCowan, "Clustering and segmenting speakers and their locations in meetings," in *Proc. ICASSP*, 2004, pp. 605–608.

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

[90] W. Zhang, A. Masnadi-Shirazi, and B. Rao, "Insights into the frequency domain ICA/IVA approach," *submitted to IEEE Trans. on Audio, Speech and Language Processing*.