UNIVERSITY OF CALIFORNIA, SAN DIEGO

Microphone Array Processing for Speech: Dual Channel Localization, Robust Beamforming, and ICA Analysis

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Signal and Image Processing)

by

Wenyi Zhang

Committee in charge:

Professor Bhaskar D. Rao, Chair
Professor Sanjoy Dasgupta
Professor William Hodgkiss
Professor Kenneth Kreutz-Delgado
Professor Mohan M. Trivedi

2010

The dissertation of Wenyi Zhang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

Chair

University of California, San Diego

2010

DEDICATION

To my family

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

It is a pleasure to take this opportunity to acknowledge many people who helped me during my graduate work to make this thesis possible. First and foremost, I would like to thank my advisor Prof. Bhaskar Rao for his guidance, inspiration, dedication, and his academic and financial support. I benefited a lot from his exceptional insights and rich knowledge in signal processing. His valuable advice contributes to a great extent in this work. I am also thankful for his continuous mental support during my years at UCSD. It is so fortunate to have such a tremendous mentor who really cares for students. I am also grateful for my doctoral committee, and wish to thank Prof. Sanjoy Dasgupta, Prof. William Hodgkiss, Prof. Kenneth Kreutz-Delgado, and Prof. Mohan M. Trivedi for their valuable inputs and suggestions.

Special thanks go to my labmates YuZhe Jin, Alireza Masnadi-Shirazi and Shankar Shivappa for the valuable technical discussions and collaborations on part of this work. I would also like to thank the other members of the DSP lab: Albenzio Cirillo, Zhongren Cao, Ethan Duni, Rajesh Hegde, Yogananda Isukapalli, Aditya K. Jagannatham, Chandra R. Murthy, Tianmin Ren, June C. Roh, David Wipf, Liwen Yu, Chengjin Zhang, Zhilin Zhang, Jun Zheng, and Hairuo Zhuang. It is my great pleasure to work with them and talk with them. I am indebted to many of my colleagues and friends in UCSD, Dashan Gao, Jiahua He, Xianfei He, Min Li, Dan Liu, Zhou Lu, Yanhua Mao, Yushi Shen, Haichang Sui, An Sun, Junwen Wu, Zheng Wu, Shan Yan, Chenmo Yang, Nang Zang, Linyun Zhang, Xin Zhang, Qinhua Zhao, and Ping Zhou, whose fun filled my life.

I would like to express a deep sense of gratitude to my parents, Shangliang Zhang and Yulan Chen for their great support. I owe many thanks to my wife, Jing Sun, for her love, patience, inspiration and companion. Also thank for my little son David Zhang, born during the busiest time of my life, who pushed me to work even harder and brought me so much happiness.

Chapter 2, in full, is a reprint of the material as it appears in Wenyi Zhang, and Bhaskar D. Rao, "A Two Microphone Based Approach for Source Localization of Multiple Speech Sources", submitted to *IEEE Transactions on Audio, Speech and Language Processing* and Wenyi Zhang, and Bhaskar D. Rao, "Two Microphone Based

Direction of Arrival Estimation for Multiple Speech Sources using Spectral Properties of Speech", *Acoustics, Speech, and Signal Processing, 2009. ICASSP 09. IEEE International Conference on*, pp. 2193–2196, Apr. 2009. Chapter 3, in part, is a reprint of the material which has appeared in Wenyi Zhang, and Bhaskar D. Rao, "Robust Broadband Beamformer With Diagonally Loaded Constraint Matrix and Its Application to Speech Recognition", *Acoustics, Speech, and Signal Processing, 2006. ICASSP 06. IEEE International Conference on*, pp. 785-788, May 2006. Chapter 4, in part, is a reprint of the material which has been published in Wenyi Zhang, and Bhaskar D. Rao, "Robust Adaptive Beamformer with Feasibility Constraint on the Steering Vector", *European Signal Processing Conference, 2006, EUSIPCO*, Florence, Italy, 2006. Chapter 5, in part, has been submitted for publication of the material as it may appear in Wenyi Zhang, Alireza Masnadi-Shirazi, and Bhaskar D. Rao, "Insights into the frequency domain ICA/IVA approach", submitted to *IEEE Transactions on Audio, Speech and Language Processing, special issue on reverberant speech: methodologies and applications*. Chapter 7, in full, is a reprint of the material which has appeared in Wenyi Zhang, and Bhaskar D. Rao, "Combining Independent Component Analysis with Geometric Information and its Application to Speech Processing", *Acoustics, Speech, and Signal Processing, 2009. ICASSP 09. IEEE International Conference on*, pp. 3065-3068, Apr. 2009. The dissertation author was the primary investigator and author of these material.

VITA

2000	Bachelor of Science, in Electrical Engineering, Shanghai Jiao-Tong University, China
2001-2002	Intern at Microsoft, China
2002-2003	Intern at Motorola, China Research Center
2003	Master of Science, in Electrical Engineering, Shanghai JiaoTong University, China
2004	Intern at AT&T Labs, Shannon Lab, USA
2003-2009	Research Assistant, Department of Electrical and Computer Engineering, University of California, San Diego, USA
2008	Teaching Assistant, Department of Electrical and Computer Engineering, University of California, San Diego, USA
2009	Software Engineer, Bloomberg L.P., USA
2010	Doctor of Philosophy, in Electrical Engineering, University of California, San Diego, USA

PUBLICATIONS

Wenyi Zhang, Alireza Masnadi-Shirazi, and Bhaskar D. Rao, "Insights into the frequency domain ICA/IVA approach", submitted to *IEEE Transactions on Audio, Speech and Language Processing, special issue on reverberant speech: methodologies and applications*.

Alireza Masnadi-Shirazi, Wenyi Zhang, and Bhaskar D. Rao, "Glimpsing IVA: A Framework for Overcomplete/Complete/Undercomplete Convolutive Source Separation", submitted to *IEEE Transactions on Audio, Speech and Language Processing, special issue on reverberant speech: methodologies and applications*.

Wenyi Zhang, and Bhaskar D. Rao, "A Two Microphone Based Approach for Source Localization of Multiple Speech Sources", *IEEE Transactions on Audio, Speech and Language Processing*. Accepted for publication.

Alireza Masnadi-Shirazi, Wenyi Zhang, and Bhaskar D. Rao, "Glimpsing Independent Vector Analysis: Separating More Sources Than Sensors Using Active and Inactive States", *Acoustics, Speech, and Signal Processing, 2010. ICASSP 10. IEEE International Conference on*.

Wenyi Zhang, and Bhaskar D. Rao, "Two Microphone Based Direction of Arrival Estimation for Multiple Speech Sources using Spectral Properties of Speech", *Acoustics, Speech, and Signal Processing, 2009. ICASSP 09. IEEE International Conference on*, pp. 2193–2196, Apr. 2009.

Wenyi Zhang, and Bhaskar D. Rao, "Combining Independent Component Analysis with Geometric Information and its Application to Speech Processing", *Acoustics, Speech, and Signal Processing, 2009. ICASSP 09. IEEE International Conference on*, pp. 3065-3068, Apr. 2009.

Wenyi Zhang, and Bhaskar D. Rao, "Robust Adaptive Beamformer with Feasibility Constraint on the Steering Vector", *European Signal Processing Conference, 2006, EUSIPCO*, Florence, Italy, 2006.

Wenyi Zhang, and Bhaskar D. Rao, "Robust Broadband Beamformer With Diagonally Loaded Constraint Matrix and Its Application to Speech Recognition", *Acoustics, Speech, and Signal Processing, 2006. ICASSP 06. IEEE International Conference on*, pp. 785-788, May 2006.

ABSTRACT OF THE DISSERTATION

Microphone Array Processing for Speech: Dual Channel Localization, Robust Beamforming, and ICA Analysis

by

Wenyi Zhang

Doctor of Philosophy in Electrical Engineering (Signal and Image Processing)

University of California, San Diego, 2010

Professor Bhaskar D. Rao, Chair

Compared with single channel speech processing, multi-microphone based speech processing methods are capable of high interference suppression in noisy environments because of their spatial filtering capability. This dissertation develops novel microphone array speech processing methods in a variety of configurations and also analyzes and provides insights into existing popular techniques. First we develop a two microphone based source localization technique for multiple speech sources utilizing speech specific properties and the generalized mixture decomposition clustering algorithm. Voiced speech is sparse in the frequency domain and can be represented by sinusoidal tracks via sinusoidal modeling which provides high local SNR. By utilizing the inter-channel phase differences (IPD) between the dual channels on the sinusoidal tracks, the source localization of the mixed multiple speech sources is turned into a clustering problem on the IPD vs. frequency plot. The generalized mixture decomposition algorithm (GMDA) is used to cluster the groups of points corresponding to multiple sources and thus estimate the DOA of the sources.

Our next work considers data dependent adaptive beamformers, which are

known to have high resolution and interference rejection capability when the array steering vector is accurately known. However, these methods degrade severely if steering vector error exists and so robust variants are needy to remedy this sensitivity. We compare and analyze recent developments in adaptive beamforming. We then develop a robust broadband adaptive beamforming algorithm which combined the robustness of the delay-and-sum beamforming in the look direction with the high interference rejection capability of adaptive beamforming algorithm. Based on J. Li and P. Stoica's work on robust Capon beamforming, we develop variants of the constrained robust Capon beamformer that attempt to limit the search in the underlying optimization problem to a feasible set of steering vectors thereby achieving improved performance.

Another class of promising multi-channel signal separation algorithms that complement beamforming methods are blind source separation methods. We analyze and provide insight into one such class of blind source separation methods,independent component analysis (ICA) methods. For separating convolutively mixed source signals, the frequency domain ICA approach is often used because it simplifies the time domain convolutive mixing problem into the instantaneous mixing problem in each frequency bin. We examine and provide insights into the frequency domain ICA methods for source separation in reverberant environments. Concentrating on the bin-wise ICA methods, a significant contribution of this work is to show that signals modeled using Gaussian scale mixtures (GSM) density can be separated using ICA even though they might be dependent on each other as long as the the frame dynamics of the source signals are different almost surely. We also analyze the stability conditions of the complex maximum likelihood ICA/IVA.

Lastly, in an attempt to make the best of ICA and beamforming methods, we propose two approaches for combining geometric information with ICA algorithms to solve the permutation problem in a scenario where approximate information about the direction of the desired source is known.

1 Introduction

1.1 Challenges

Array processing has been developed for many years and has gained important applications in various fields such as radar/sonar systems, wireless communication systems, biomedical apparatus, seismic engineering and microphone array speech processing systems. In recent years, microphone array speech processing has attracted a lot of attention not only because of the recent development of robust adaptive beamforming algorithms and blind source separation algorithms based on independent component analysis (ICA), but also because of the advances in computer engineering and semiconductor industry which provide sufficient computational power and make real time implementation of array processing algorithms possible.

Microphone array processing for speech may be used for different purposes such as source localization, tracking, building situation awareness, speech enhancement and noise suppression, boosting automatic speech recognition system, and source separation. Because of the spatial distribution of multiple microphone sensors, the signals emitting from a sound source may arrive at different sensors with different time delays and magnitudes (or different channels if reverberation exist). These differences are the real factors which make all the array processing methods work.

Despite the many years of development of antenna array processing techniques in radar/sonar applications, microphone array speech processing brings up new challenges because of its specific characteristics. First of all, speech is a broadband signal covering several octaves, while many antenna processing algorithms have been focusing on narrowband signals. Even though a broadband signal can be decomposed into a bank of narrowband signals and be processed therein by narrowband algorithms, there

are many other difficulties such as decomposing and synthesizing without distortion, possible permutation and scaling ambiguities between the different narrowband signals, etc. All these problems should be considered carefully and handled properly before we apply narrowband array processing algorithms to process a broadband signal.

Secondly, speech is not a white signal. It exhibits special spectral characteristics. It is highly dynamic with a rapid varying spectrum. It is usually considered as non-stationary (In chapter 5, we take a different point of view and model it as a dynamic random process which is stationary at the frame and ensemble level and exhibits non-stationarity characteristics at the realization level). What's more, in speech enhancement or speech separation applications, the interference signals (or called noise signals) to be suppressed may also be speech signals which exhibit the same dynamic characteristics as the desired signal. Speech's non-white and non-stationary properties make many array processing algorithms which take Gaussian white signals inapplicable. Hence special algorithms are desired. Although speech proposes new challenges because of the special characteristics, it may also provide specific opportunities which are not available in general signal processing. For example, speech signal's sparsity in the time-frequency domain and the harmonic structure in the spectrum have been used in single channel or dual channel speech separation. Specifically, in chapter 2, we design a two microphone based source localization technique for multiple speech sources utilizing speech specific properties such as the sparsity in the time-frequency domain and sinusoidal modeling.

Another challenge in microphone array speech processing is the high level of room reverberations often encountered in many applications. Many microphone arrays are developed to work in a room environment where there is no reverberation. The reverberation time of a room is usually long such that the room impulse response filter may have a length which is more than one thousand of samples. In such a case, the array processing filters may be even longer. This is different from the antenna array processing where a few taps are enough to model the transfer function of the channel. Furthermore, the room impulse response are usually modeled as a dynamic channel where besides the fixed strong reverberation paths, the rest of the channel is modeled as a random process. All these factors already put forward big challenges, not to mention the room impulse response is often sensitive to environment changes such as subject moving or

facing in different directions. Due tothese factors, learning the transfer function of the channel through a training process usually is not possible for array speech processing. Therefore, the transfer functions of the channels can only be learned through a blind process, which in general is a very hard problem.

Furthermore, microphone arrays for speech processing are limited in deployment freedom. In real applications, microphone arrays are often supposed to be inobtrusive and can only be deployed in specific locations, such as on walls or roofs. When the array is far away from the sound sources, some voices will have low signal to noise ratio (SNR), partly because of the huge dynamic range in human voice volume. Also, the size of a microphone array is usually constrained by the environment and by the budget. Boosting the speech enhancement performance using a limited number of sensors is another challenge.

Based on whether geometric information is utilized explicitly, the microphone array processing techniques can be classified into two general classes. The first class demands prior knowledge about the array geometry, and uses the geometric information directly in the processing algorithms. This class includes localization algorithms, source tracking, and beamforming algorithms. The second class is the blind source separation which includes blind equalization and independent component analysis (ICA). No geometric information is needed in these algorithms.

Motivated by the challenges in microphone array processing for speech, in this dissertation, we focus our efforts in improving the robustness of the beamforming algorithms, analyzing the blind source separation techniques, making the best of both worlds and exploiting speech specific properties. In the next section, we provide some background in these areas and summarize our contributions.

1.2 Geometric Array Processing

1.2.1 Source Localization

Source localization is an important component of a multichannel speech signal processing system which in addition to localization may include other functions such as tracking, speech enhancement and noise suppression. The estimated source direction or

location can then be used by the beamforming algorithms to enhance the desired signal and suppress the interference signals. Source localization using microphone array can also be combined with other systems such as a video camera system to provide more complete information about an environment.

Depending on how localization is achieved, it may be alternatively referred to as time delay estimation (TDE), time difference of arrival (TDOA) estimation and direction of arrival (DOA) estimation, in various fields. For wideband signals, source localization is generally based on inter-channel time difference (ITD) or inter-channel intensity difference (IID). For far field source scenario, which assumes the sources are far away from sensors such that each source's contribution at all sensors have the same intensity, only ITD is used in localization. A number of source localization algorithms have been developed, e.g. MUSIC, ESPRIT, spatial power spectrum based approach, maximum likelihood method, correlation based approach, and adaptive multichannel time delay estimation method based on blind equalization [1, Ch.9] [2]. These algorithms usually assume the number of sensors is greater than the number of sources.

However, in some applications, the number of microphone sensors may be limited. For example, on a hand held device or a laptop computer, we may only be able to install 2 to 4 microphones. In chapter 2 of this dissertation, DOA estimation based on only two microphones is considered, and methods in this context can be broadly categorized into two classes: time domain approaches and frequency domain approaches. The time domain algorithms include time domain cross correlation method, average-magnitude-difference function method [3], LMS-type adaptive TDE algorithm [4], and adaptive eigenvalue decomposition algorithm associated with blind channel identification [5, 6]. The frequency domain algorithms include linear regression method [7, 8], blind channel identification based method [9] and the well known generalized cross-correlation (GCC) family of methods, which includes many variations, the smoothed coherence transform, the phase transform, the maximum likelihood approach, and so forth [10, 11, 12, 13, 14, 15, 16, 17]. However, most of the these algorithms are based on single source signal model and can not locate multiple sources. Surprisingly, the human auditory system as a binaural system has shown its high efficiency in locating multiple speech sources simultaneously [18]. Research on human auditory system shows

speech's specific properties such as the sparsity in the time-frequency domain and the harmonic structure in the voiced speech have been utilized extensively for this purpose. This provides us with new opportunities in developing localization algorithms based on two microphones. Our contribution is summarized in section 1.4 and fully described in chapter 2.

1.2.2 Robust Beamforming

In an array processing system for speech, after the sources are localized, usually a speech enhancement stage is employed to extract the desired speech signal and suppress any interference signal or noise. Compared with single channel speech processing, multi-microphone speech processing has high interference suppression in noisy environment because of its spatial filtering capability. Array processing algorithms usually try to form a spatial response pattern which has a peak in the direction of the desired signal and nulls in the directions of the interference signals. The spatial patterns are referred to as beams and the systems as beamformers or beamforming algorithms.

Beamformers can be classified into data dependent (adaptive) beamformers or data independent (fixed) beamformers. The standard data independent beamformers such as the delay-and-sum (DS) beamformer is robust to the signal of interest (SOI) steering vector errors, which may be due to look direction error, array sensor position error, and small mismatches in the sensor responses. Yet those beamformers suffer from low resolution and high sidelobes, inducing bad interference rejection capability, especially when the number of sensors is limited. In contrast, data dependent adaptive beamformers can learn the environment and places nulls in the interference direction, thus having high resolution and interference rejection capability when the array steering vector is accurately known. However, the performance of the traditional adaptive beamformer can degrade severely in practice when there exist SOI steering vector errors. In such cases, the SOI might be mistaken as an interference signal and be suppressed. Fig.1.1 demonstrates a look direction error in a uniform linear array system. In this example, the look direction is the broadside direction, i.e. 0°. However, the true desired signal incidence angle is 5°, which means there's a 5° look direction error. In this case, the traditional adaptive beamformer will take the desired signal as an interference signal

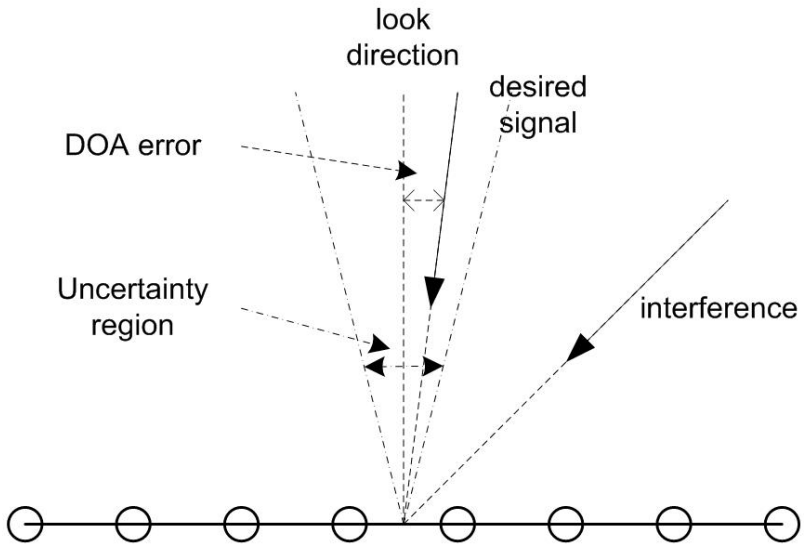


Figure 1.1: A demonstration of the look direction error problem

and suppress it. This SOI steering vector error problem should be addressed before the adaptive beamforming algorithms can be applied to process real world speech data.

Broadband beamforming

Based on the signals processed, beamformers can be categorized into broadband beamformers or narrowband beamformers. Since speech is a broadband signal, the broadband beamformer is a natural choice.

Frost Beamformer: Optimum Solution and LMS Algorithm

Among the broadband adaptive beamformer, the Frost beamformer is one of the most extensively studied [19]. The Frost beamformer has a multichannel tapped-delay-line structure (Fig. 1.2), and a set of linear constraints are introduced to ensure a desired frequency response in the look direction. Thereby it is also called linear constrained minimum variance (LCMV) beamformer. Suppose a Frost beamformer use a linear array with K sensors and J taps. The first step is to pre-steer the array in the look

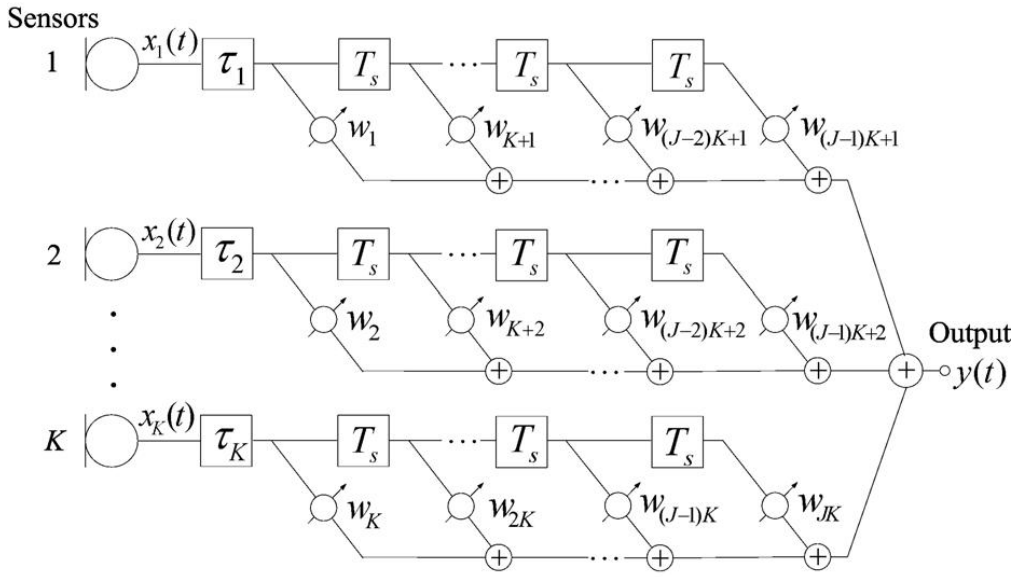


Figure 1.2: A broadband beamformer with K sensors and J taps.

direction. The Frost beamforming problem is then mathematically formulated as,

$$\min_{\mathbf{w}} \mathbf{w}^T R_{XX} \mathbf{w}, \quad \text{subject to} \quad C^T \mathbf{w} = F \tag{1.1}$$

where $\mathbf{w}$ is the KJ dimension concatenated weight vector for all the taps. R_{XX} is the KJ dimension covariance matrix. $F = [f_1, f_2, \cdots, f_J]^T$ is the vector of weights of the look-direction-equivalent tapped delay line, a simple choice for F will be $F = [1, 0, \cdots, 0]^T$. C is the constraint matrix with KJ rows and J columns, in which the j_{th} column picks up the j_{th} column of the array elements in the array matrix. The optimum solution to the the Frost beamforming problem (Eq. (1.1)) can be obtained using the Lagrange multiplier method and corresponding LMS algorithm is developed (section 3.2).

The performance of the Frost beamformer (optimum solution) can degrade severely in practice when the steering vector errors exist. However, experiments also illustrate the Frost LMS algorithm is robust to the look direction error when the adaptation does not go through too many iterations (i.e. the speech source length is not too long). Motivated by these observations, we analyze the effect that the signal incidence angle has on the convergence rate of the Frost LMS beamforming weight vector. Our analysis confirms that the Frost LMS algorithm is robust to the look direction error when the adaptation does not go through too many iterations. Our contributions in this context

are summarized in section 1.4 and fully described in chapter 3.

Narrowband beamforming

Because of the high sampling rate in the time domain, broadband beamforming algorithms usually are computationally intensive. Besides the broadband processing, speech can also be decomposed into narrowband signals by a bank of filters. In each narrowband, the narrowband beamformer is applied to enhance the desired signal. Then the enhanced narrowband signal can be synthesized into the time domain signal. Among the adaptive narrowband beamformers, the minimum variance distortionless response (MVDR) beamformer is probably the most popular one.

Standard MVDR Beamformer (MVDR)

The MVDR beamforming is also called Capon beamforming [20]. The problem is formulated as minimizing the output energy of the beamformer while maintaining a constant response in the look direction, i.e.

$$\min_{\mathbf{w}} \mathbf{w}^H R \mathbf{w}, \quad \text{s.t.} \quad \mathbf{w}^H \mathbf{a} = 1. \tag{1.2}$$

where R is the signal correlation matrix. $\mathbf{a}$ is the SOI steering vector. $\mathbf{w}$ is the beamformer weight vector. The solution to this optimization problem is given by

$$\mathbf{w} = \frac{R^{-1}\mathbf{a}}{\mathbf{a}^H R^{-1} \mathbf{a}}. \tag{1.3}$$

The performance of standard MVDR beamformer can degrade severely in practice when there exist SOI steering vector errors. Many robust beamforming algorithms have been proposed to address this problem [21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39]. Derivative constraint in the look direction is proposed in [22, 23]. Er and Cantoni proposed a robust beamforming algorithm which restricts the error between the desired and actual beam pattern of the array over a small spatial region around the array's look direction, allowing for uncertainty in the look direction [40]. Linearly constrained minimum variance (LCMV) beamforming is proposed in [41, 42]. Bell proposed a Bayesian approach to robust adaptive beamforming in [24]. Norm constrained and white noise gain constrained adaptive beamformer are

studied in [21, 43, 44] and widely used thereafter. These beamformers use diagonal loading to improve robustness, however, the selection of the diagonal loading parameter is subjective and problematic.

Recently some new robust adaptive beamformers with theoretical background have been proposed. Robust adaptive beamforming using worst-case performance optimization is proposed in [25, 45, 46]. The problem is formulated as minimizing a quadratic function subject to infinitely many quadratic constraints, but is reduced to a second-order cone programming problem which can be solved by interior point methods. Li and Stoica proposed the robust Capon beamformer (RCB) [47, 26, 48] where a spherical uncertainty set constraint is enforced on the array steering vector. They also developed a doubly constrained robust Capon beamformer (DCRCB) [26] based on RCB, wherein a norm constraint on the beamformer steering vector is added. A comparison of these two beamformers is given in [49] and a geometrical explanation is provided.

RCB and DCRCB solve the SOI cancellation problem by searching for the MVDR beamformer with maximum energy output in a spherical uncertainty set. If the true SOI steering vector exists in the spherical uncertainty set, RCB and DCRCB is supposed to find that steering vector as the minimizer of a quadratic optimization problem. However, the minimizer of the quadratic optimization problem may not be a feasible steering vector. A feasible steering vector generally satisfies some geometrical constraints and has a physical meaning. Not all the vectors in the spherical uncertainty set have such physical meaning and are set to be valid steering vectors. Our contributions in this context are summarized in section 1.4 and fully described in chapter 4.

1.3 Blind Source Separation

Besides robustness to steering vector errors, another significant problem with adaptive beamforming algorithms is the signal cancelation problem in strong reverberant environments. In a reverberant environment, the signal from a reverberation path is a delayed version of the signal from the direct path. These signals are coherent. If an adaptive beamformer is employed, it will try to cancel the desired signal from the output

using the correlation between its reverberant components. This is the well-known signal cancelation problem in adaptive beamforming [50]. There have been efforts to solve the signal cancelation problem [51, 52, 53], but none of them is successful enough in a strong reverberant environment.

1.3.1 Blind Equalization

Even though multi-channel blind source separation (BSS) does not utilize geometric information explicitly, they are more successful in a reverberant environment compared to the beamforming algorithms. One famous class of BSS is the blind channel identification and blind equalization algorithms [54, 55, 56]. But many algorithms in this class are designed for Gaussian, white or constant modulus signals, and they are only suitable for short channel filter length. As we mentioned before, speech is non-white, non-Gaussian and non-stationary. What's more, the reverberation channel filter length is long which may contain thousands of samples. Therefore, those popular blind equalization algorithms are usually not appropriate to be applied on reverberated speech.

1.3.2 Independent Component Analysis

Another class of promising multi-channel BSS algorithms is independent component analysis (ICA). ICA only utilizes the statistical independence between several sources to separate them from their mixtures [57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83]. Assuming an N by N linear instantaneous mixing system (N sources, N sensors),

$$\mathbf{x} = A\mathbf{s} \tag{1.4}$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix} \quad A = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{bmatrix} \quad \mathbf{s} = \begin{bmatrix} s_1 \\ \vdots \\ s_N \end{bmatrix} \tag{1.5}$$

$\mathbf{x}$ is the observation vector. A is an unknown mixing matrix assumed to have full column rank. $\mathbf{s}$ is a vector of latent variables which are assume to be non-Gaussian (or at most one is Gaussian) and mutually independent. The de-mixing system can be written as,

$$\mathbf{y} = W\mathbf{x} \tag{1.6}$$

where $\mathbf{y} = [y_1, \cdots, y_N]^T$ is the outputs and W is a N by N de-mixing matrix. Assuming source signals $\mathbf{s}$ are independent signals, ICA attempts to solve the BSS problem by determining a de-mixing matrix W such that the outputs $\mathbf{y}$ approximate source signals $\mathbf{s}$ under permutation and scaling ambiguity [81]. Most ICA algorithms try to find a de-mixing matrix W so that the components in $\mathbf{y}$ are statistically as independent as possible.

The ICA signal model has inherent ambiguities in the permutation and scaling of the source signals. Suppose P is a permutation matrix with scalings on non-zero elements, then we get the following equation,

$$\mathbf{x} = A\mathbf{s} = AP^{-1}P\mathbf{s} = A'\mathbf{s}' \tag{1.7}$$

where $A' = AP^{-1}$, $\mathbf{s}' = P\mathbf{s}$. It is easy to see that after the permutation and scaling, the components in $\mathbf{s}$ are still independent. This intrinsic ambiguity problem can not be resolved by ICA itself because ICA only utilize independence.

1.3.3 Frequency domain ICA approach

For separating convolutive mixed source signals, the frequency domain ICA approach is often used because it simplifies the time domain convolutive mixing problem into the instantaneous mixing ICA problem in each frequency bin. The frequency domain ICA approach is elaborated in Table 1.1. It has been widely used and has become a conventional blind source separation (BSS) approach for de-mixing convolutive mixtures. However, permutation problem across different frequency bins is an important problem that needs to be solved before the frequency domain ICA approach can be used to separate mixed sources in real applications [84, 72, 85, 86, 87, 88, 89, 90, 87, 91, 92, 93, 94, 95]. Independent vector analysis (IVA) was proposed to circumvent the permutation problem [96, 97, 98, 99, 100, 101, 102, 103]. Instead of solving the mixing problem in each frequency bin independently, IVA considers the frequency domain source signal linked together as a vector source and attempts to resolve the mixing problem in the frequency bins in an integrated manner.

For convolutive mixed blind source separation, Pham et al. argued that in the frequency domain the source signal at each fixed frame and frequency is a Gaussian signal, therefore ICA is not applicable [104]. Parra, Pham, and Choi et al. all proposed

Table 1.1: The conventional frequency domain ICA approach

(1) Apply short time Fourier transform (STFT) on each time domain mixture based on consecutive frames. Thus the convolutive mixing problem in the time domain is transformed into an instantaneous mixing problem in each frequency bin.
(2) On each frequency bin, apply batch ICA algorithm (instantaneous mixing ICA) to separate the source signals.
(3) Use appropriate permutation and scaling correction method to solve the permutation and scaling ambiguity across all frequency bins, so that the spectrum of the source signals can be recovered.
(4) Apply inverse discrete Fourier transform (DFT) on the recovered source spectrum and use overlap-and-save (or overlap-and-add) method to obtain the time domain signal.

to utilize the non-stationarity of source signals for instantaneous mixing blind source separation [105, 67, 106, 107, 66, 108, 109, 110]. In the frequency domain approach for de-mixing convolutive mixtures, if Parra, Pham, and Choi et al.'s methods are applied in each frequency bin to separate the mixed source signals therein, the source signals in each frequency bin are implicitly assumed to be non-stationary. By simultaneously diagonalizing several covariance matrices (second order statistics) at different times in each frequency bin, the mixed source signals in each frequency bin are separated. However, the permutation problem stills need to be solved in order to obtain each source's spectrum for reconstructing the time domain source signal. One shortcoming of this approach is the need for a large amount of data so that the covariance matrices in each frequency bin can be estimated robustly (remember several covariance matrices have to be estimated in each frequency bin). For instance, for convolutive mixed blind speech signal separation more than 20 to 30 seconds of data is necessary to obtain satisfactory results [105].

However, the conventional frequency domain ICA approach (Table 1.1) is widely used and many experiments on convolutive mixing blind source separation demonstrate good separation performance (assuming the permutation problem is correctly solved) no matter what kind of instantaneous mixing ICA algorithms is employed

in separating the mixed source signals in each frequency bin. For example, maximizing neg-entropy (FastICA) algorithm, maximum likelihood algorithm, and Infomax algorithm [58, 60, 61, 63, 62] have been tested in the frequency domain ICA approach and all of them illustrate good separation results. An advantage of these ICA algorithms is that their performance is not too sensitive to the amount of data. For example, in convolutive mixing speech separation, a few seconds of data is enough to provide satisfactory separation result. Inspired by their broad success, we examine and provide insight into frequency domain ICA methods for source separation in reverberant environments. Our contributions in this context are summarized in section 1.4 and fully described in chapter 5.

1.3.4 Make the Best of Both Worlds

ICA assumes no knowledge about the mixing process except the independence between sources. However, sometimes extra information is available and can be utilized to aid the ICA process. For example, in microphone array speech processing, speech's temporal structure or geometric information with the array may be employed to solve the permutation problem. Many recent works have been developed to combine ICA with geometric information to solve the permutation problem [85, 86, 87, 88, 89, 91, 95]. In [85, 86], beam pattern of the ICA processor is utilized to figure out the directions of the sources to solve the permutation problem. These methods become too complicated and are not robust when the number of sources exceed 2. Parra and Alvino proposed the geometrically constrained (or initialized) ICA algorithm [87], but accurate source number is required and correct permutation is not guaranteed. Knaak and Araki proposed an ICA algorithm with a hard linear geometric constraint [88]. However, accurate source number is mandatory for the algorithm to perform properly.

In microphone array speech processing, the geometry of the array and rough information about the direction of the desired signal may be known a priori, for instance, the direction of the desired signal may be assumed to be the broadside direction for a linear array, or be acquired by some direction of arrival (DOA) estimation algorithms [1]. In an attempt to make the best of ICA and beamforming methods, we propose two approaches for combining DOA information of the desired source with ICA algorithm to

solve the permutation problem. Our contributions in this context are summarized in section 1.4 and fully described in chapter 7.

1.4 Contributions of the Thesis

This thesis contributes to many aspects of microphone array based speech processing; robust beamforming algorithms, insights into blind source separation (BSS) methods, algorithms that make use of speech specific properties and make best use of the features of beamforming and BSS methods. The contributions are summarized below.

1. First we consider DOA estimation for multiple speech signals based on only two microphones. Most of the popular dual channel DOA estimation algorithms are based on single source signal model and can not locate multiple sources. Surprisingly, the human auditory system as a binaural system has shown its high efficiency in locating multiple speech sources simultaneously. Research on human auditory system shows speech's specific properties such as the sparsity in the time-frequency domain and the harmonic structure in the voiced speech have been utilized extensively for this purpose. This provides us with new opportunities in developing localization algorithms based on two microphones. Utilizing speech's specific properties such as the sparsity in the time-frequency domain and the sinusoidal modeling, we propose a two microphone based DOA estimation technique for multiple speech sources using the generalized mixture decomposition clustering algorithm. Voiced speech is sparse in the frequency domain and can be represented by sinusoidal tracks via sinusoidal modeling which provides high local SNR. By utilizing the inter-channel phase differences (IPD) between the dual channels on the sinusoidal tracks, the source localization of the mixed multiple speech sources is turned into a clustering problem on the IPD vs. frequency plot. The generalized mixture decomposition algorithm (GMDA) is used to cluster the groups of points corresponding to multiple sources and thus estimate the DOA of the sources. Experiments illustrate the proposed GMDA algorithm with the Laplacian noise model can estimate the number of sources accurately and exhibits

smaller DOA estimation error than the baseline histogram based DOA estimation algorithm in various scenarios including reverberant and additive white noise environments. Experiments suggest that appropriate power thresholding can be a simple and good approximation to the sinusoidal modeling, for the purpose of selecting time-frequency points with high local SNR, with slight loss in performance. The details are presented in Chapter 2.

2. Despite the high resolution and interference rejection capability, the conventional data dependent adaptive beamformers are very sensitive to the steering vector errors. The performance of the Frost beamformer (optimum solution) can degrade severely in practice when the steering vector errors exist. However, experiments also illustrate the Frost LMS algorithm is robust to the look direction error when the adaptation does not go through too many iterations (i.e. the speech source length is not too long). Motivated by these observations, we analyze the effect that the signal incidence angle has on the convergence rate of the Frost LMS beamforming weight vector. Our analysis confirms that the Frost LMS algorithm is robust to the look direction error when the adaptation does not go through too many iterations.

   We develop a robust broadband adaptive beamforming algorithm which combines the robustness of the DS beamforming in the look direction with the high interference rejection capability of the conventional adaptive beamforming algorithm. A quadratic constraint is employed to deal with the uncertainty in the look direction. In order to address the ill-conditioning associated with the constraint matrix, a diagonal loading (DL) is added to the constraint matrix thereby ensuring a robust solution to the quadratic constraint beamforming problem. The advantage of adding DL to constraint matrix is that the constraint matrix is only determined by the geometry of the array thereby allowing the DL level to be chosen offline. This is superior to adding DL to the signal covariance matrix where the DL level has to be chosen online. It is shown that the diagonal loading is equivalent to an additional norm constraint without introducing it explicitly. We also develop an iterative algorithm (and corresponding adaptive algorithm) to solve for the robust beamformer coefficients. The developed algorithm is applied to the problem of

beamforming using microphone arrays for speech recognition and shown to be superior to existing algorithms. The details are presented in Chapter 3.

3. Motivated by recent work in robust MVDR beamforming (RCB&DCRCB) [26], we develop variants of the constrained robust MVDR beamformer that attempts to limit the search in the underlying optimization problem to a feasible set of steering vectors, thereby achieving improved performance. The robustness against steering vector error is provided through a spherical uncertainty set constraint, while a set of magnitude constraints are enforced on each element of the steering vector to constrain the search in the space of feasible steering vectors in a better fashion. By appropriately changing the variables, the optimization problem is modified such that the need for the magnitude constraints are avoided.

   We also develop adaptive algorithms for the RCB and the time delay based robust MVDR beamformer. The adaptive algorithms have two updating steps. The first step updates the steering vector estimation or the time delay estimation; the second step updates the beamformer's weight vector given an estimated steering vector. The developed algorithms are tested in the context of speech enhancement using a microphone array. The details are presented in Chapter 4.

4. Inspired by the broad success of ICA methods, we examine and provide insight into frequency domain ICA methods for source separation in reverberant environments. For the modeling of the source signals, we develop the concept of a dynamic random process to model the source signals. It formalizes the concept of signals that are stationary in a frame but exhibit dynamics at the frame level. Frame dynamics is an important characteristics of these signals and prove important to the success of the ICA methods. With suitable assumptions, the dynamic random process is stationary in the ensemble sense while a given realization may in an engineering sense exhibit 'non-stationarity'. We show for dynamic random processes, the unconditional distribution of the source signal in each frequency bin is a Gaussian scale mixture (GSM). The non-Gaussianity, which is critical to ICA [57], of the source signal in each frequency bin is shown to be a direct consequence of the frame dynamics. Furthermore, the independence between the

unconditional distributions of the source signals in each frequency bin is related to the independence of the frame dynamics of the mixed time domain source signals. The GSM mathematical modeling is extended to the vector random processes formed by stacking the different frequency components of a source. This provides insights into the mathematical models suitable for the frequency domain independent vector analysis (IVA) type approaches. A special case of the distribution turns out to be the 'spherical distribution' employed in IVA source modeling providing support to their use in source separation. Concentrating on the bin-wise ICA methods, a significant contribution of this dissertation is to show that signals modeled using GSM density can be separated using ICA even though they might be dependent on each other as long as the the frame dynamics of the source signals are different almost surely. In particular, we show that Kurtosis and negentropy type measures can be used to separate variance correlated GSM signals. The details are presented in Chapter 5.

5. In an attempt to make the best of ICA and beamforming methods, we propose two approaches for combining DOA information of the desired source with ICA algorithm to solve the permutation problem. The first approach is a new blind extraction algorithm with a soft quadratic geometric constraint. The quadratic constraint restricts the weighted square error between the desired and actual response of the processor over a small spatial uncertainty region chosen to deal with look direction uncertainty. Thereby the desired source is guaranteed to be conveyed to the output with little distortion and the negentropy maximization criterion is used to ensure that the other sources get suppressed at the output. The second approach employs a quadratic geometric test as a post-processing step to pickup the desired source after ICA processing. In every frequency bin, the ICA algorithm separates instantaneously mixed source signals, then the quadratic geometric test will pick up the desired source. An advantage of the proposed two approaches is that they do not require accurate knowledge of the number of sources in the mixtures to recover the desired source, in contrast, other geometric ICA approaches usually fail if the number of sources is not known accurately. The details are presented in Chapter 7.

2 A Two Microphone Based Approach for Source Localization of Multiple Speech Sources

In this chapter we propose a two microphone based source localization technique for multiple speech sources utilizing speech specific properties and the generalized mixture decomposition clustering algorithm. Voiced speech is sparse in the frequency domain and can be represented by sinusoidal tracks via sinusoidal modeling and provide high local SNR. Furthermore, the sinusoidal tracks of different speech signals are usually disjoint in the time-frequency domain. When multiple speech signals are mixed in the two microphone system, the inter-channel phase differences (IPD) between the dual channels on the sinusoidal tracks will be dominated by the spatial information of the most powerful source at that specific time-frequency point because of the spectrum sparsity and masking effects. Thereby, the source localization problem is turned into a clustering problem on the IPD vs. frequency plot. The generalized mixture decomposition algorithm (GMDA) is used to cluster the groups of points corresponding to multiple sources and thus estimate the DOA of the sources. Two variants, one based on Gaussian modeling of the noise (GMDA_Gauss) and another based on a Laplacian noise model (GMDA_Laplace), are developed. Experiments illustrate the proposed GMDA_Laplace algorithm to be very effective. It estimates the number of sources accurately and exhibits smaller DOA estimation error than the baseline histogram based DOA estimation algorithm in various scenarios including reverberant and additive white noise environments. Experiments suggest that appropriate power thresholding can be a simple and good approximation to the sinusoidal modeling, for the purpose of selecting time-frequency

points with high local SNR, with slight loss in performance.

2.1 Introduction to Source Localization

Source localization is an important component of a multichannel signal processing system which in addition to localization may include other functions such as tracking, signal separation, enhancement and noise suppression. Depending on how localization is achieved, it may be alternatively referred to as time delay estimation (TDE), time difference of arrival (TDOA) estimation and direction of arrival (DOA) estimation, in various fields. For wideband signals, source localization is generally based on inter-channel time difference (ITD) or inter-channel intensity difference (IID). For far field source scenario, which assumes the sources are far away from sensors such that each source's contribution at all sensors have the same intensity, only ITD is used in localization. A number of source localization algorithms have been developed, e.g. MUSIC, ESPRIT, spatial power spectrum based approach, maximum likelihood method, and adaptive multichannel time delay estimation method based on blind equalization [1, Ch.9] [2]. These algorithms usually assume the number of sensors is greater than the number of sources.

In this work, DOA estimation based on only two microphones is considered, and methods in this context can be broadly categorized into two classes: time domain approaches and frequency domain approaches. The time domain algorithms include time domain cross correlation method, average-magnitude-difference function method [3], LMS-type adaptive TDE algorithm [4], and adaptive eigenvalue decomposition algorithm associated with blind channel identification [5, 6]. The frequency domain algorithms include linear regression method [7, 8], blind channel identification based method [9] and the well known generalized cross-correlation (GCC) family of methods, which includes many variations, the smoothed coherence transform, the phase transform, the maximum likelihood approach, and so forth [10, 11, 12, 13, 14, 15, 16, 17]. However, most of the these algorithms are based on single source signal model and can not locate multiple sources. The human auditory systems is an efficient binaural system which can locate multiple speech sources simultaneously [18]. It has been shown

that speech specific attributes, i.e. sparsity in time-frequency domain, can be utilized to locate multiple speech sources using dual channels [111, 112, 113, 114, 115, 116, 117, 118, 119, 120]. For instance, in [113] the presence of gaps in the spectrum of each source at different times and frequencies is exploited and an image processing method is employed to detect vertical segments in the frequency vs. path difference plot to locate two sound sources. In [114, 115], harmonic sound stream segregation using localization is considered and a rough localization method based on IPD, IID and harmonic streams is proposed. The sparse speech assumption is explicitly used in [116] for localization and a histogram based method is proposed to locate multiple sources through the frequency vs. DOA plot. The algorithms in [117, 118] can be considered as variations to [116] wherein the inter-microphone distance is restricted to be small so as to avoid spatial aliasing. In [117], a histogram generated using both ITD and IID is utilized. In [118], the algorithm is extended to multiple microphones to estimate 'the 3-dimensional absolute DOA'. In [120], a statistical noise model on ITD was introduced and a posterior distribution on discretized ITD was estimated to localize the sources.

In this work, we follow this tradition and propose a two microphone based DOA estimation technique for multiple speech sources utilizing the speech's sparsity attribute and the generalized mixture decomposition clustering algorithm. Voiced speech is sparse in the frequency (spectrum) domain and can be represented by sinusoidal tracks via sinusoidal modeling [121]. Furthermore, the sinusoidal tracks of different speech signals are usually disjoint in the time-frequency domain. An advantage of utilizing the sinusoidal tracks is that they represent regions where speech energy is concentrated leading to reliable, i.e. high signal to noise ratio (SNR), data points for further analysis. This also suggests the consideration of power thresholding as a simple alternative to sinusoidal tracks for identifying locally high SNR points. When multiple speech signals are mixed in the two microphone system, the inter-channel phase differences (IPD) between the dual channels on the sinusoidal tracks will be dominated by the spatial information of the most powerful source at that specific time-frequency point because of the spectrum sparsity and masking effects (Sec. 2.2). The error between the IPD between the dual channel signals and the IPD between a source signal's contributions at the dual channels is modeled as a random variable and a statistical signal model for the IPD error based

on speech's sparsity attribute and masking effects is proposed. The source localization problem is shown to be a clustering problem on the IPD vs. frequency plot. Generalized mixture decomposition algorithm (GMDA), one based on Gaussian modeling of the noise (GMDA_Gauss) and another based on a Laplacian noise model (GMDA_Laplace), is used to cluster the groups of points corresponding to multiple sources. The direction of each source is derived from the parameters of each cluster. Generalized hard clustering algorithm (GHCA) is also developed to avoid the need for a probabilistic model in GMDA. Depending on the inter-microphone spacing, spatial aliasing effect is taken into consideration by proper phase unwrapping. A minimum description length (MDL) algorithm is used to determine the number of sources. Experimental results show the scheme to be very effective.

The rest of this chapter is organized as follows. In section 2.2, we describe the two microphone based source localization technique for multiple speech sources based on IPD and sparsity of speech in the time-frequency domain. In section 2.3, a generalized mixture decomposition algorithm (GMDA) is developed for clustering and estimation of the directions of multiple sources. Generalized hard clustering algorithm (GHCA) is also developed to avoid the use of probabilistic models in GMDA. In section 2.4, experiments on localization of multiple speech sources are presented and they provide support to the method developed. The following notation is adopted throughout this chapter: $(.)^T$ for transpose, and $|.|$ for modulus.

2.2 Two Microphone Based Source Localization for Multiple Speech Sources

We first describe a two microphone based DOA estimation method for a single source using inter-channel phase difference. The methodology discussed will then be combined with speech sparsity characteristics to estimate DOA for multiple sources.

2.2.1 Inter-channel Phase Difference Based DOA Estimation

Assuming a far field source scenario, a simple DOA estimation algorithm using two microphones can be developed based on inter-channel phase difference (IPD) or inter-channel time difference (ITD). Denoting the desired source signal by $s[k]$, the two microphone signals $x_1[k]$ and $x_2[k]$ can be expressed as,

$$x_1[k] = s[k] + n_1[k] \tag{2.1a}$$

$$x_2[k] = s[k - \tau] + n_2[k] \tag{2.1b}$$

where $s[k - \tau]$ represents a delayed version of $s[k]$, τ is the time delay of the desired source. $n_1[k]$ and $n_2[k]$ represent ambient noise and more generally also include interference signals at the two channels.

The short time discrete Fourier transform (DFT) of $x_1[k]$ and $x_2[k]$ is denoted by $X_1(\omega)$ and $X_2(\omega)$ respectively. Generally, they are also a function of the frame number. To simplify notations, we ignore this dependence. $X_1(\omega)$ and $X_2(\omega)$ are given by (This is a reasonable approximation if the delay is sufficiently smaller than the frame length),

$$X_1(\omega) = S(\omega) + N_1(\omega) \tag{2.2a}$$

$$X_2(\omega) = S(\omega)e^{-j\omega\tau} + N_2(\omega) \tag{2.2b}$$

where ω represents angular frequency, $S(\omega)$ is the DFT of $s[n]$, $N_1(\omega)$ and $N_2(\omega)$ represent DFT of the noise component $n_1[k]$ and $n_2[k]$ respectively.

In the absence of noise, the inter-channel phase difference can be used for estimating the DOA. The presence of noise introduces errors in the estimates of the phase and hence the phase difference. The IPD $\psi_X(\omega)$ between the two channel signals $X_1(\omega)$ and $X_2(\omega)$ is computed as

$$\psi_X(\omega) = \angle X_1(\omega) - \angle X_2(\omega) \tag{2.3}$$

and is constrained to be in the range $[-\pi, \pi]$ after the $mod(2\pi)$ operation. $\angle X_1(\omega)$ and $\angle X_2(\omega)$ is the phase of $X_1(\omega)$ and $X_2(\omega)$ respectively. If noise components $n_1[k]$ and $n_2[k]$ are much weaker than the source signal $s[k]$, in other words, if $N_1(\omega)$ and $N_2(\omega)$ have much smaller magnitudes than the magnitude of $S(\omega)$, then

$$\psi_X(\omega) = \omega\tau + 2\pi n + v(\omega) \tag{2.4}$$

where τ is the time delay of the desired source. n is an integer number and $2\pi n$ represents possible phase unwrapping. $v(\omega)$ denotes the IPD error, which is dependent on the signal to noise ratio (SNR). If $N_1(\omega)$ and $N_2(\omega)$ are modeled as complex Gaussian random variables, then the IPD error $v(\omega)$ is also a random variable whose probability density function (PDF) is denoted by $p(v(\omega))$. Though it is difficult to find a closed form expression, it is not difficult to deduce that the PDF $p(v(\omega))$ is symmetric and is concentrated around 0. A similar problem is considered in noncoherent detection in communication [122, Ch.4], where a Gaussian distribution is proposed to approximate the PDF $p(v(\omega))$ at high SNR. In this work we also approximate the PDF $p(v(\omega))$ by a Gaussian distribution. In addition, a Laplacian distribution is also considered because of its robustness to outliers, a situation likely to occur in the multi-source scenario. Alternately, the IPD $\psi_X(\omega)$ can be approximated by a Gaussian or Laplacian random variable with mean $\omega\tau + 2\pi n$. The DOA of the desired source can be derived from τ using the following equation,

$$\tau = d \sin\theta / c \tag{2.5}$$

where d represents inter-microphone distance, c the sound speed, and θ the DOA of the desired source (assume broadside direction of the 2 microphone array as 0°).

Using the IPDs at different frequencies, $\omega_l, l = 1, .., L$, and different frames, linear regression [7, 8] can be employed to estimate the slope of curve IPD $\psi_X(\omega)$ with respect to ω in the IPD vs. frequency plot, thereby estimating τ and hence the DOA of the source. To enhance robustness, it is useful to account for the SNR at each frequency and frame and it is preferable to use only those data points with high SNR. The inclusion of only reliable points in the estimation will be an important component of the method proposed in the paper.

2.2.2 Speech's Sparsity Attribute and Inter-channel Phase Difference

The DOA estimation method based on IPD, as discussed in Sec. 2.2.1, performs well when there is only one source and the SNR is high. It is unable to locate multiple sources particularly when there are multiple white sources spectrally flat with similar

power level. Expanding the signal model shown in Eq. (2.2) to multiple sources, it can be readily seen that if there are two white sources with similar power, the IPD $\psi_X(\omega)$ in Eq. (2.3) does not have any simple relationship to the DOA of either source.

In this work, we develop a method that exploits source specific knowledge and attempts to retain the simplicity of the IPD technique. The speech signal is not a white signal. It has a special property, namely sparsity in the time-frequency domain, which enables the possibility of estimating the DOA of multiple sources using dual channels [113, 114, 115, 116]. To elaborate, there are two kinds of sparsity that are often present.

1. Sparsity in time domain. Natural speech generally has many short pauses and silent segments, which may occupy more than half of the total observation interval [123].

2. Sparsity in frequency domain. Speech is not a white signal and exhibits strong short time correlation, especially for voiced speech. In the frequency domain, the signal power is not equally distributed across the whole frequency range even though it is a wideband signal. For voiced speech, the signal power is concentrated on a set of equally distributed discrete frequency points, i.e. harmonics of the pitch frequency [121, 124, 125].

Collectively the above two attributes make the speech sound sparse in the time-frequency domain. When a recording contains multiple speakers, there are segments of time when only one speaker is active and other speakers are inactive (in the state of short pauses or silent breaks). Even though there are segments of time when more than one speaker is active, the signal power of different speakers in the frequency domain may occupy different sets of discrete frequencies.

At a specific time-frequency point, there is a high likelihood that at most one source is dominating (in power) and the contributions from the other sources is negligible. As a consequence, the IPD $\psi_X(\omega)$ (Eq. (2.3)) will be dominated by the IPD $\psi_S(\omega)$ of the dominating source, where $\psi_S(\omega)$ is defined as the IPD between the source's contribution at the two channels, $\psi_S(\omega) = \angle S(\omega) - \angle\{S(\omega)e^{-j\omega\tau}\} = \omega\tau$. So the IPD $\psi_X(\omega)$ contains DOA information of the dominating source at that time-frequency point and

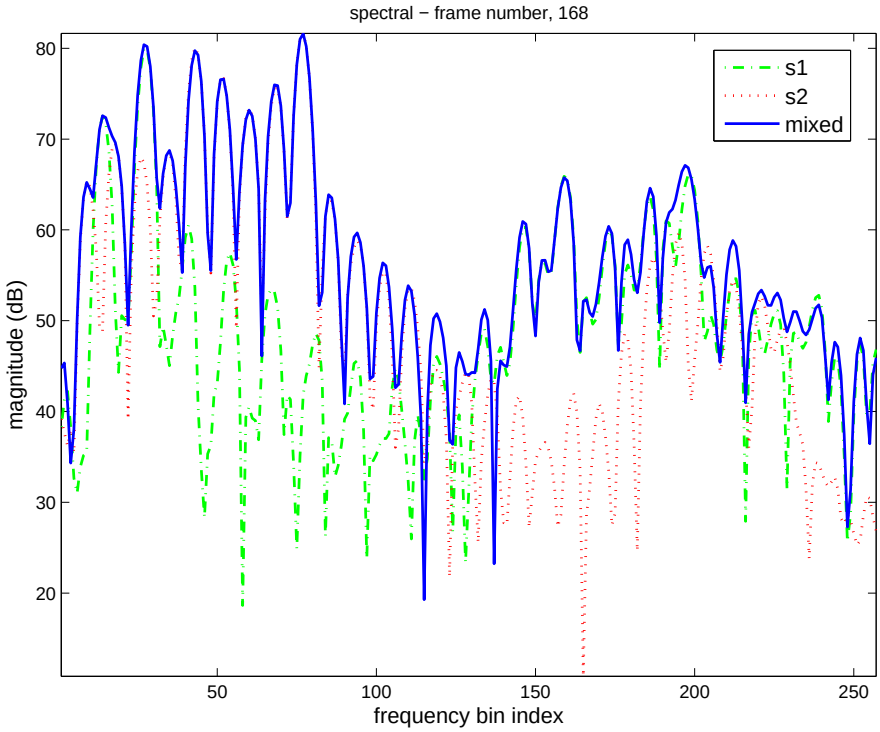


Figure 2.1: Magnitude spectrum of one frame. Dashed, dotted, and real line curve denotes the spectrum for source s1, source s2, and the mixed signal respectively.

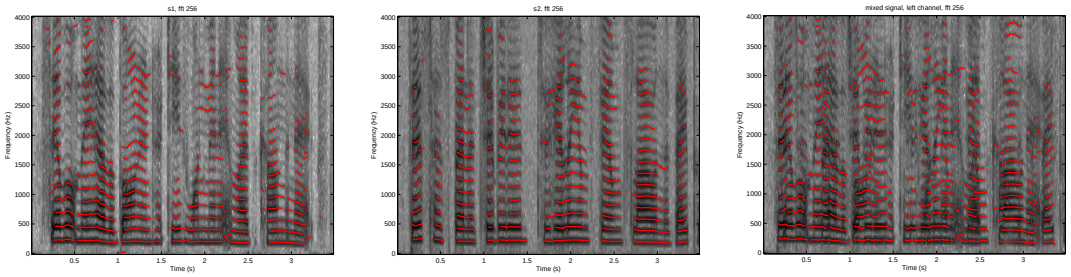
can be used for the DOA estimation. This is denoted as masking effect in this paper (the most powerful source will mask the less powerful interference signals and ambient noise). A similar masking effect defined on the power spectrum is widely used in single microphone speech sources separation [124, 125]. We hasten to add that this masking effect is not to be confused with the masking effects commonly discussed in the context of psychology of hearing where it denotes a powerful signal at one frequency will mask weaker signals at adjacent frequencies [126, Ch.3]. Fig. 2.1 illustrates the magnitude spectrum of one frame for speech sources s_1, s_2 and the mixed signal (the signal received at one microphone, $x_i, i \in \{1, 2\}$). The masking effect is evident. When one source has much bigger power than the other at some frequencies, the spectrum of the mixed signal $X(\omega)$ is approximated by the spectrum of the source with significant power. Only when the two sources has similar power at some frequencies, a rare event, does the mixed signal $X(\omega)$ deviates from either one of the sources.

When masking comes into effect at a specific time-frequency point, the noise

components $N_1(\omega)$ and $N_2(\omega)$ (including the less powerful interference signals and ambient noise) being masked by the dominating source $S(\omega)$ can be modeled as independent and identically distributed complex Gaussian random variables [127, Ch.4] leading to the single source model of the previous section. Consequently, the IPD $\psi_X(\omega)$ at this time-frequency point can be approximated by a Gaussian or Laplacian random variable following the discussion in Sec. 2.2.1.

Speech's sparsity attribute and the masking effect combined leads to the idea of DOA estimation using only data points with high local SNR in the time-frequency domain, e.g. DOA estimation based on sinusoidal modeling [121]. In [121], it is observed that speech usually has power focused on a set of discrete frequencies and can be modeled by a set of sinusoidal tracks. Sinusoidal tracks are defined to be continuous local peaks in the time-frequency domain which satisfy a set of constraints such as power level, continuity in consecutive time frames, starting and ending constraints, etc. By the sparsity and masking properties of the speech signal, the sinusoidal tracks extracted from one channel of a mixed signal can be approximated as a disjoint union of the sinusoidal tracks from each of the different source signals, i.e. a track of the mixed signal can be associated with one of the source signals. This association is not known and will be dealt with in the next section. When two or more sources have similar power level at a time-frequency area, the interaction between source signals will cause the mixed signal $X(\omega)$ to fluctuate frequently resulting in no sinusoidal tracks in the corresponding time-frequency area (the frequency locations of the local spectrum peaks change a lot from frame to frame in this time-frequency area, hence the continuity condition for sinusoidal tracks will not be satisfied). Consequently, the sinusoidal tracks extracted from the mixed signal will be close to the sinusoidal tracks from one source in one time-frequency area, and be close to the sinusoidal tracks from another source in another time-frequency area. The points on the sinusoidal tracks will implicitly have high SNR. Fig. 2.2 shows the spectrum and extracted sinusoidal tracks for one example mixed signal.

We propose to use the IPD $\psi_X(\omega)$ between the two channel signals $X_1(\omega)$ and $X_2(\omega)$ on points of the sinusoidal tracks for the two microphone multiple speech sources DOA estimation. As previously discussed, the IPD $\psi_X(\omega)$ on points of the sinusoidal



(a) source s1 (b) source s2 (c) mixed signal

Figure 2.2: Spectrum and sinusoidal tracks of source s1, s2, and one channel of mixed signal. The red lines overlapping on the spectrum are the sinusoidal tracks.

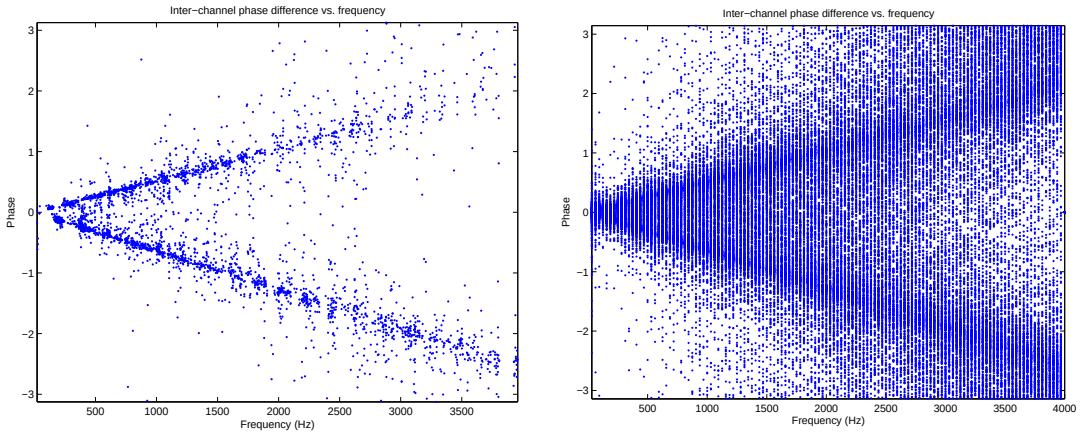
tracks is dominated by the IPD $\psi_S(\omega)$ of one source among multiple sources (it can be any one of the multiple sources depending on which source contributes to the sinusoidal track). As an example, the IPD $\psi_X(\omega)$ on points of the sinusoidal tracks is plotted in a IPD vs. frequency plot (Fig. 2.3 (a)). For comparison, the IPD $\psi_X(\omega)$ for all spectrum points is also plotted (Fig. 2.3 (b)). The sources are two speech sources with sampling rate $8kHz$. The DOA of the two sources are 60° and -45° respectively. The inter-microphone spacing is $4cm$. There is no spatial aliasing [117] in this example (methods for compensating for spatial aliasing are discussed in Sec. 2.3.3). From Fig. 2.3 (a), it is clear there are two clusters of points which can be fitted by two lines. The two clusters represent the two sources and the DOA of the two sources can be derived from the slopes of the two lines. When the IPD $\psi_X(\omega)$ for all spectrum points are plotted on a IPD vs. frequency plot (Fig. 2.3 (b)), the cluster information is obscured and overwhelmed by noise although some cluster information can still be deduced from the plot.

In summary, the steps of the proposed dual channel multiple speech sources DOA estimation method are enumerated in the following.

1. Calculate the time-frequency spectrum $X_1(\omega)$ and $X_2(\omega)$ of the two microphone signals using short time discrete Fourier transform.

2. Extract sinusoidal tracks from the two microphone channels [128], and the final sinusoidal track set employed in the next step is the joint set comprised of the tracks extracted from the two microphone channels.

3. Calculate the IPD $\psi_X(\omega)$ between the two channel signals $X_1(\omega)$ and $X_2(\omega)$ on



(a) IPD $\psi_X(\omega)$ on points of the sinusoidal tracks (b) IPD $\psi_X(\omega)$ on all spectrum points

Figure 2.3: Inter-channel phase difference vs. frequency. There are 2 speech sources, whose DOAs are 60° and −45° respectively. The inter-microphone spacing is 4cm.

points of the sinusoidal tracks.

4. Cluster the points on the IPD vs. frequency plot, employ line fitting techniques to fit set of lines, and derive the DOAs of the sources from the slopes of the lines.

The last step, clustering and line fitting, is discussed in the next section.

In retrospect, though sinusoidal modeling is a natural fit to speech signals, it may be beneficial to view the approach as selecting time-frequency points with high local SNR. This allows one to consider a more general class of methods for identifying locally high SNR points. An alternative to sinusoidal modeling, and potentially simpler, is to use a simple power thresholding to select the time-frequency points in the above proposed algorithm. For example, the time-frequency points can be sorted in power and the top X% selected, where X maybe scenario specific. By the source signal's sparsity attribute and the masking effect, one is expected to acquire time-frequency points with high local SNR even using simple power thresholding. The performance of the DOA estimation algorithm under different levels of simple power thresholding is compared in the experiments section.

2.3 Clustering and Line Fitting

2.3.1 Generalized Mixture Decomposition Algorithm

As previously discussed, the points on the IPD vs. frequency plot is naturally distributed and grouped into several lines based on the DOA information of the sources. Once the grouping is done, a line can be fitted to estimate the DOA. We now describe a procedure that does the clustering and line fitting jointly. For this purpose, the IPD error $v(\omega)$ in Eq. (2.4) (on the IPD vs. frequency plot, it is the distance from a data point to the center of its underlying cluster, i.e. a line) is modeled as a Gaussian or Laplacian random variable. A mixture model is then employed to fit the data and the generalized mixture decomposition algorithm (GMDA) [129, Ch.14] is used to cluster the data.

Assume there are m clusters, $C_j, j = 1, .., m$, i.e. m speech sources, and m is assumed to be known. Assume there are N data points $\mathbf{y}_i$ (i is the data sample index, $i = 1, .., N$). Each data point $\mathbf{y}_i$ is a 2-dimension vector which denotes a point on the IPD vs. frequency plot, $\mathbf{y}_i = [\omega_i, \psi_{X,i}(\omega_i)]^T$. A mixture model can be used to fit the data points. Each component of the mixture is a line, and the distance from each data point to its underlying line is modeled as a random variable with the PDF $p(\mathbf{y}_i|C_j; \boldsymbol{\theta}_j)$, where $\boldsymbol{\theta}_j$ is the parameter vector characterizing the line corresponding to the j^{th} cluster. $\boldsymbol{\theta} = [\boldsymbol{\theta}_1^T, .., \boldsymbol{\theta}_m^T]^T$ encompasses the parameter vectors for all the clusters. Suppose $\mathbf{P} = [P_1, .., P_m]^T$, with P_j being the a-priori probability for the j^{th} cluster.

The parameters of the mixture model are to be learned from the data points. By using the maximum likelihood approach and employing the Expectation-Maximization (EM) algorithm, the conditional expectation of complete data log-likelihood given the observed data under the previous parameter value is,

$$Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}(t)) = \sum_{i=1}^{N} \sum_{j=1}^{m} P(C_j|\mathbf{y}_i; \boldsymbol{\Theta}(t)) \ln \left(p(\mathbf{y}_i|C_j; \boldsymbol{\theta}) P_j\right) \tag{2.6}$$

where $\boldsymbol{\Theta} = [\boldsymbol{\theta}^T, \boldsymbol{P}^T]^T$ is the parameters for the whole mixture model. $P(C_j|\mathbf{y}_i; \boldsymbol{\Theta}(t))$ is the posterior probability for class C_j given data point $\mathbf{y}_i$ and the previous parameter value $\boldsymbol{\Theta}(t)$. t is the iteration number for the learning process. The complete EM algorithm can be written as,

Generalized Mixture Decomposition Algorithm (GMDA)

- $t = 0$. Choose initial estimates for the model parameters, $\boldsymbol{\theta} = \boldsymbol{\theta}(0)$, $\mathbf{P} = \mathbf{P}(0)$.
- Repeat until convergence
  - Compute

$$P(C_j|\mathbf{y}_i; \boldsymbol{\Theta}(t)) = \frac{p(\mathbf{y}_i|C_j; \boldsymbol{\theta}_j(t))P_j(t)}{\sum_{k=1}^{m} p(\mathbf{y}_i|C_k; \boldsymbol{\theta}_k(t))P_k(t)}, \quad i = 1, .., N, j = 1, .., m. \tag{2.7}$$

  - Set $\boldsymbol{\theta}_j(t+1)$ equal to the solution of the equation

$$\sum_{i=1}^{N} P(C_j|\mathbf{y}_i; \boldsymbol{\Theta}(t)) \frac{\partial}{\partial \boldsymbol{\theta}_j} \ln\left(p(\mathbf{y}_i|C_j; \boldsymbol{\theta})P_j\right) = 0 \tag{2.8}$$

    with respect to $\boldsymbol{\theta}_j$, for $j = 1, .., m$.
  - Set

$$P_j(t+1) = \frac{1}{N} \sum_{i=1}^{N} P(C_j|\mathbf{y}_i; \boldsymbol{\Theta}(t)), \quad j = 1, .., m. \tag{2.9}$$

  - $t = t + 1$.
- Convergence Criteria: $\|\boldsymbol{\Theta}(t+1) - \boldsymbol{\Theta}(t)\| < \varepsilon$, where ε is a small threshold.

The algorithm described above is very general. To get more specific update rules for the GMDA, an explicit form for the PDF $p(\mathbf{y}_i|C_j; \boldsymbol{\theta}_j)$ is necessary. As previously discussed in Sec. 2.2, an appropriate form is either a Gaussian distribution or Laplacian distribution. We now specialize the GMDA algorithm for these two forms.

PDF $p(\mathbf{y}_i|C_j; \boldsymbol{\theta}_j)$ chosen to be Gaussian distribution

$$p(\mathbf{y}_i|C_j; \boldsymbol{\theta}_j) = \frac{1}{\sqrt{2\pi}\sigma_j} exp\{-\frac{(\psi_{X,i}(\omega_i) - \alpha_j \cdot \omega_i)^2}{2\sigma_j^2}\} \tag{2.10}$$

recall $\mathbf{y}_i = [\omega_i, \psi_{X,i}(\omega_i)]^T$. The parameter of the cluster is $\boldsymbol{\theta}_j = [\alpha_j, \sigma_j^2]^T$. α_j denotes the slope of the line. σ_j^2 is the variance of the model.

By equation (2.8), we can obtain,

$$\alpha_j(t+1) = \frac{\sum_{i=1}^{N} P(C_j|\mathbf{y}_i; \boldsymbol{\Theta}(t)) \cdot \omega_i \cdot \psi_{X,i}(\omega_i)}{\sum_{i=1}^{N} P(C_j|\mathbf{y}_i; \boldsymbol{\Theta}(t)) \cdot \omega_i^2} \tag{2.11}$$

$$\sigma_j^2(t+1) = \frac{\sum_{i=1}^{N} P(C_j|\mathbf{y}_i;\mathbf{\Theta}(t))(\psi_{X,i}(\omega_i) - \alpha_j(t+1)\cdot\omega_i)^2}{\sum_{i=1}^{N} P(C_j|\mathbf{y}_i;\mathbf{\Theta}(t))} \tag{2.12}$$

If the variance of all clusters are assumed to be the same, i.e. $\sigma_j^2 \equiv \sigma^2$, then

$$\sigma^2(t+1) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{m} P(C_j|\mathbf{y}_i;\mathbf{\Theta}(t))(\psi_{X,i}(\omega_i) - \alpha_j(t+1)\cdot\omega_i)^2}{N} \tag{2.13}$$

PDF $p(\mathbf{y}_i|C_j;\boldsymbol{\theta}_j)$ chosen to be a Laplacian distribution

$$p(\mathbf{y}_i|C_j;\boldsymbol{\theta}_j) = \frac{1}{2b_j} exp\{-\frac{|\psi_{X,i}(\omega_i) - \alpha_j\cdot\omega_i|}{b_j}\} \tag{2.14}$$

The parameter of the cluster $\boldsymbol{\theta}_j = [\alpha_j, b_j]^T$, where α_j denotes the slope of the line. b_j is a parameter related to the variance of the model. By equation (2.8), we obtain the updating equation for α_j,

$$\sum_{i=1}^{N} P(C_j|\mathbf{y}_i;\mathbf{\Theta}(t))\cdot\frac{1}{b_j}\cdot sgn(\psi_{X,i}(\omega_i) - \alpha_j(t+1)\cdot\omega_i)\cdot\omega_i = 0 \tag{2.15}$$

This equation can be solved by Newton type numerical optimization algorithm. The updating equation for b_j is,

$$b_j(t+1) = \frac{\sum_{i=1}^{N} P(C_j|\mathbf{y}_i;\mathbf{\Theta}(t))|\psi_{X,i}(\omega_i) - \alpha_j(t+1)\cdot\omega_i|}{\sum_{i=1}^{N} P(C_j|\mathbf{y}_i;\mathbf{\Theta}(t))} \tag{2.16}$$

If the variance of all clusters are assumed to be the same, i.e. $b_j \equiv b$, then

$$b(t+1) = \frac{\sum_{i=1}^{N}\sum_{j=1}^{m} P(C_j|\mathbf{y}_i;\mathbf{\Theta}(t))|\psi_{X,i}(\omega_i) - \alpha_j(t+1)\cdot\omega_i|}{N} \tag{2.17}$$

2.3.2 Generalized Hard Clustering Algorithm

One of the difficulties associated with the previously discussed GMDA is the involvement of the PDF's, for which a suitable model has to be assumed. To avoid such problems, a generalized hard clustering algorithm (GHCA) is developed. The detailed steps of the GHCA is,

Generalized Hard Clustering Algorithm (GHCA)

- $t = 0$. Choose initial estimates for the clusters' parameters, $\boldsymbol{\theta} = \boldsymbol{\theta}(0)$.

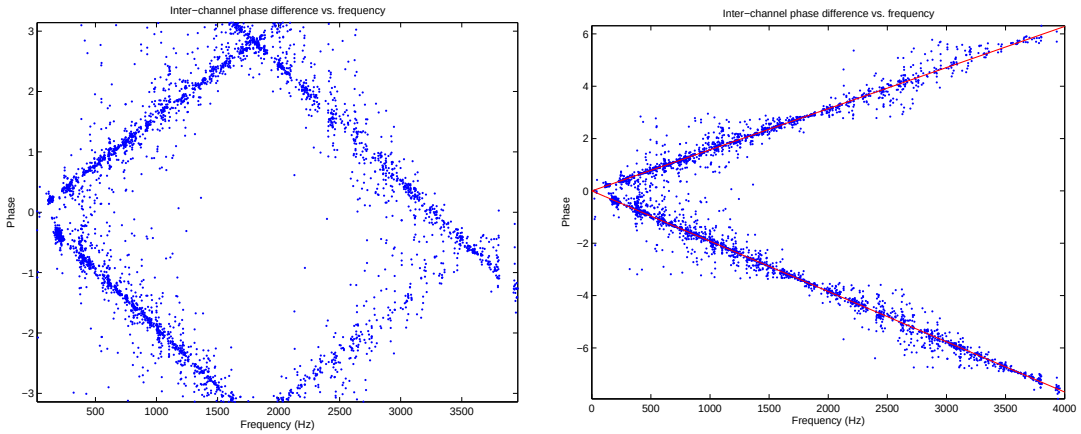
- Repeat until convergence
  - Decide which cluster each data point belongs to.
  - Update each cluster's parameter $\boldsymbol{\theta}_j$ from the data points belonging to that cluster.
  - $t = t + 1$.
- Convergence criteria: $\|\boldsymbol{\theta}(t+1) - \boldsymbol{\theta}(t)\| < \varepsilon$, where ε is a small threshold.

In our line fitting problem, each cluster's parameter $\boldsymbol{\theta}_j$ is only a scalar α_j, which denotes the slope of the line. For deciding which cluster each data point belongs to, the least "point to line distance" is used. The distances of a data point $\mathbf{y}_i$ to each of the lines $d_{ij}(t) = |\psi_{X,i}(\omega_i) - \alpha_j(t) \cdot \omega_i|$ is computed and the data point is determined to belong to the cluster with minimum distance $\arg\min_j d_{ij}(t)$. After each data point is assigned to a cluster, linear regression is used to update the slope α_j for the j^{th} cluster using all data points belonging to the j^{th} cluster. The iterative process continues until convergence.

In reality, GHCA is closely related to GMDA. In Sec. 2.3.1, when the PDF $p(\mathbf{y}_i|C_j;\boldsymbol{\theta}_j)$ is modeled as a Gaussian distribution, if a data point is close to a cluster k, then $P(C_j|\mathbf{y}_i;\boldsymbol{\Theta}(t))$ will be close to 1 for $j = k$ and be close to 0 for all other $j, j \neq k$. This is effectively equivalent to assigning the data point to a single cluster in the hard clustering approach.

2.3.3 Clustering and Line Fitting under Spatial aliasing Scenario

In the previous discussion of clustering and line fitting algorithms (GMDA and GHCA in Sec. 2.3.1 and Sec. 2.3.2), it is assumed the inter-microphone spacing is small such that there is no spatial aliasing [117]. One example IPD vs. frequency plot under such scenario is shown in Fig. 2.3 (a). With the increasing of the inter-microphone distance spatial aliasing may exist. Fig. 2.4 (a) shows the IPD vs. frequency plot for the same scenario as in Fig. 2.3 (a) except that the inter-microphone spacing is increased to $12cm$. Recall the IPD on the original IPD vs. frequency plot is always confined to be in the range of $[-\pi, \pi]$. Two sources can still be observed in Fig. 2.4 (a), however, the two lines corresponding to the two sources are broken into parallel segments because



(a) Original

(b) After adjusting phase

Figure 2.4: Inter-channel phase difference vs. frequency. There are 2 speech sources, whose DOAs are $60°$ and $-45°$ respectively. The inter-microphone spacing is $12cm$ (spatial aliasing scenrio).

of phase wrapping effect. The two lines are easier to be observed if we do phase unwrapping and move the broken line segments parallelly and properly. This is shown in Fig. 2.4 (b). Note that the phase is no longer confined to $[-\pi, \pi]$.

Through analysis of the Cramer-Rao Lower Bounds of the DOA estimation [1], better accuracy in the DOA estimation is expected if high frequency band is used. To utilize the high frequency band, the previously discussed GMDA and GHCA are now modified to handle spatial aliasing properly. Recall the IPD $\psi_X(\omega) = \angle X_1(\omega) - \angle X_2(\omega)$ is confined to be in the range of $[-\pi, \pi]$ after the $mod(2\pi)$ operation (in Sec. 2.2.1). Define a new IPD $\psi'_X(\omega) = \psi_X(\omega) + 2\pi n$, where n is an integer and $2\pi n$ represents possible phase unwrapping. After appropriate phase unwrapping, the new data point $\mathbf{y}'_i = [\omega_i, \psi'_{X,i}(\omega_i)]^T$ will lie around its true underlying line. Fig. 2.4 (b) shows all the new data points $\mathbf{y}'_i$ after proper phase unwrapping for each IPD $\psi_{X,i}(\omega_i)$. For GMDA, we propose to choose the phase unwrapping which yields the highest probability for the observed data point. In Sec. 2.3.1, the PDF of observing data point $\mathbf{y}_i$ given the j^{th} cluster is defined to be Gaussian (Eq. (2.10)). To look for a proper phase unwrapping factor $2\pi n$ for the IPD $\psi_{X,i}(\omega_i)$, this PDF is revised to be,

$$p'(\mathbf{y}_i|C_j;\boldsymbol{\theta}_j) = \max_n \frac{1}{\sqrt{2\pi}\sigma_j} exp\{-\frac{(\psi_{X,i}(\omega_i) + 2\pi n_i - \alpha_j \cdot \omega_i)^2}{2\sigma_j^2}\} \tag{2.18}$$

Denote

$$n_i^j = \arg\max_{n_i} \frac{1}{\sqrt{2\pi}\sigma_j} exp\{-\frac{(\psi_{X,i}(\omega_i) + 2\pi n_i - \alpha_j \cdot \omega_i)^2}{2\sigma_j^2}\}, \quad j = 1, .., m \tag{2.19}$$

Let

$$J = \arg\max_j p'(\mathbf{y}_i|C_j; \boldsymbol{\theta}_j) \tag{2.20}$$

then $\mathbf{y}_i'$ is chosen as,

$$\mathbf{y}_i' = [\omega_i, \psi_{X,i}(\omega_i) + 2\pi n_i^J]^T \tag{2.21}$$

For GHCA, the phase compensation can be done similarly. The smallest "point to line distance" is utilized to select appropriate phase compensation.

2.3.4 Model Selection: Determining the Number of Sources

In the previous discussion of GMDA, the number of sources and hence the number of clusters is assumed to be known. However, the number of clusters is not known in reality and has to be estimated from the data. This is the model selection step. Since GMDA falls into the maximum likelihood framework, minimum description length (MDL) method can be used to estimate the model order [130]. MDL is defined to be,

$$MDL(m) = -\sum_{i=1}^{N} \ln p(\mathbf{y}_i|\hat{\boldsymbol{\Theta}}(m)) + \frac{1}{2}k_p(m) \ln N \tag{2.22}$$

where m is the model order, $\hat{\boldsymbol{\Theta}}(m)$ is the maximum likelihood estimation of $\boldsymbol{\Theta}(m)$, which denotes all the parameters of the model. $k_p(m)$ is the number of freely adjusted parameters in $\boldsymbol{\Theta}(m)$. N is the number of data points. $-\sum_{i=1}^{N} \ln p(\mathbf{y}_i|\hat{\boldsymbol{\Theta}}(m))$ denotes the negative log likelihood of the model and $\frac{1}{2}k_p(m) \ln N$ denotes a penalty term. The model order is chosen to be the value yielding the smallest $MDL(m)$.

$$\hat{m} = \arg\min_m MDL(m) \tag{2.23}$$

Experiments indicate the MDL method may over-estimate the number of clusters. This could be attributed to several factors: first, the Gaussian or Laplacian distribution model used in GMDA (Sec. 2.3.1, Sec. 2.3.1) is an approximation to the true IPD error statistical model; second, the independent and identically distributed assumption

(used in deriving MDL rule) may not be satisfied exactly; third, Fig. 2.3 (a) shows that the IPD error may be frequency dependent, however, frequency independent Gaussian or Laplacian distribution model is used for tractability reasons in this article. To rectify the over-estimation problem, the penalty item $\frac{1}{2}k_p(m)\ln N$ in the MDL function $MDL(m)$ is modified to $\beta k_p(m)\ln N$, where β is a scalar parameter bigger than $\frac{1}{2}$, and whose value is determined by training.

For GHCA approach, likelihood function is not defined, but a model selection function similar to $MDL(m)$ can still be defined based on the summation of all "points to line distance" and a properly designed penalty term.

2.4 Experiments

2.4.1 Implementation Issues

Initialization of Parameters

Initialization is important to GMDA and GHCA for fast and proper convergence. For the GMDA approach, the variance parameter $\sigma_j^2, j = 1, .., m$ does not appear to be sensitive to the initialization value. Only the initialization of the slopes of the lines $\alpha_j, j = 1, .., m$, which is associated with the DOA, is considered here. There are several kinds of initialization methods for α_j.

- Initialization based on ITD histogram. Recall the inter-channel time difference (ITD) is $\tau = \psi_X(\omega)/\omega$, which can be derived from IPD. Therefore, by switching variables, a ITD histogram can be derived from the IPD vs. frequency plot. Fig. 2.5 shows the ITD histogram derived from Fig. 2.3 (a). 20 histogram bins are used in Fig. 2.5. Initialization values for the slope of the line α_j are chosen to be the local peaks in the ITD histogram.

- Random initialization. Choose α_j randomly from its valid range.

- Equally distributed initialization. The DOAs of the m sources are initialized to be uniformly distributed in space ($[-90°, 90°]$).

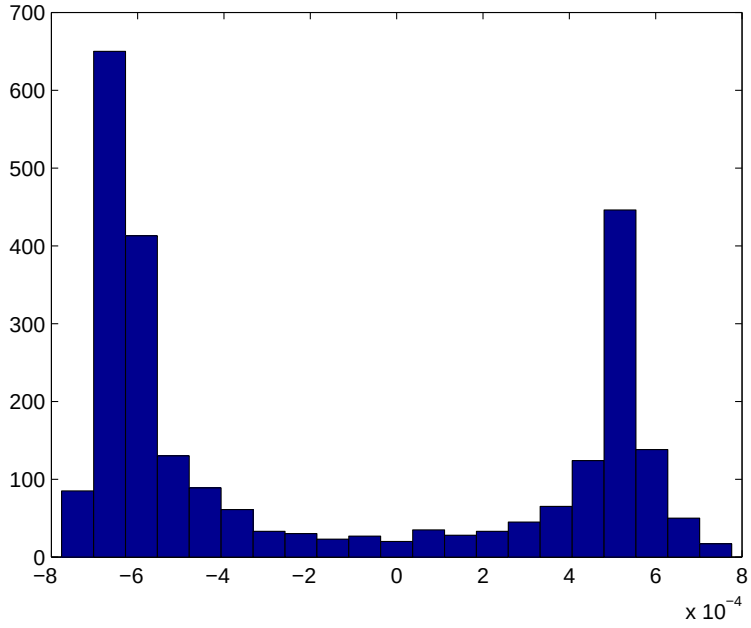


Figure 2.5: Inter-channel Time difference (ITD) histogram

In most instances, for example, when the DOAs of the sources are well separated, the ITD histogram method can estimate the rough directions of the sources, the initialization based on ITD histogram leads to proper convergence for GMDA and GHCA. However, in some instances, e.g. when the DOAs of the sources are too close to each other such that the ITD histogram method fails in estimating the rough directions of the sources, the other two initialization methods are better. In our algorithm, all the three initialization methods are utilized and the result with the method yielding the maximum likelihood score is retained.

Two step clustering for spatial aliasing scenario

Sec. 2.3.3 discussed clustering and line fitting under the spatial aliasing scenario. It is easy to see from Fig. 2.4 (a), even under spatial aliasing scenario, for data points with frequency ω below a threshold ω^d there is no phase unwrapping needed for the IPD $\psi_X(\omega)$. Therefore, phase compensation as described in Sec. 2.3.3 is not necessary for the data points with frequency ω below the threshold ω^d. Consequently, a two step clustering algorithm is developed to learn the clusters from the data. The first step clustering

only utilizes data points with frequency $\omega < \omega^d$ (without considering spatial aliasing) to learn model parameters. Then the second step clustering employs the approach in Sec. 2.3.3 using all the data points, with the parameters learnt in the first step as initialization.

2.4.2 Experimental Setup

Several experiments are conducted to study the performance of the proposed algorithms in various scenarios. All the experiments use dual microphone channels. The source signals are speech sounds from the TIMIT database [131], with $8kHz$ sampling rate. The duration of each speech source is 3 seconds. The short time DFT is applied to the microphone signals to obtain the signal spectra. A window size of 25 ms is used with overlapping frames where the frames are shifted in time by 10 ms resulting in an overlap of 15 ms. The DFT length used is 256 points. The root mean squared (RMS) error in the DOA estimates are used to measure the performance of the various algorithms.

The following acronyms are used for the various algorithms.

- GMDA_Gauss: GMDA with Gaussian model (Sec. 2.3.1).
- GMDA_Laplace: GMDA with Laplacian model (Sec. 2.3.1).
- GHCA: GHCA (Sec. 2.3.2).
- ITDh: Inter-channel time difference (ITD) histogram based algorithm with the stencil filter being used to handle spatial aliasing [116].
- ITDh_NA: ITD histogram based algorithm, only data points without spatial aliasing (phase wrapping) are used to create the histogram.

One shortcoming of the ITDh method is that the stencil filters may create false peaks on the ITD histogram depending on the source signals' directions. The ITDh_NA (_NA stands for non-aliasing) algorithm is developed utilizing the fact that there is no phase wrapping in IPD for data points with frequency ω below some threshold ω^d. Only data points with frequency ω below threshold ω^d are used to create the ITD histogram. The ITDh and ITDh_NA algorithms smooth the ITD histogram with low pass filtering, then

find the local peaks (surpassing a threshold) on the smoothed ITD histogram to estimate the DOAs of the sources. 180 histogram bins are used to construct the histogram in our experiments.

Two experiment setups are frequently used in this section.

- 2-source experiment: The 2-source experiment evaluates the performance of various algorithms using two speech sources. The inter-microphone spacing is 12*cm*. The first source is fixed at the angular direction of 0° (the broadside direction), while the other source's angular direction varies from 5° to 90° (the lateral direction). The signal to white noise ratio (SNR) is 30 dB. The experiment is repeated 100 times with speech sources randomly selected from the TIMIT database.

- 3-source experiment: The 3-source experiment is almost the same as the 2-source experiment except that three speech sources are used. The first source is fixed at the direction of 0°, while the other two sources are positioned symmetrically on both sides of the first source and the DOA varies from 5° to 90°.

2.4.3 Experiment 1: A Study of Power Thresholding Methods

In section 2.2.2, we proposed using sinusoidal modeling to select data points with high local SNR in the time-frequency domain. From a high local SNR perspective, sinusoidal modeling can be viewed as a power thresholding method to select the appropriate time-frequency points. This suggests experimenting and comparing with other simpler power thresholding methods to select the time-frequency points for DOA processing. For example, the time-frequency points can be sorted by power and the top X% is selected, where X is a user selected parameter. The performance of the DOA estimation algorithms with different levels of simple power thresholding is studied in this experiment.

Different power thresholding methods can be utilized such as using a single power threshold across the whole spectrum, or using a different threshold for each frequency bin. Experiments show that the simple power thresholding using a single power threshold across the whole spectrum is consistent, requires less tuning and hence less subjective than using a different threshold for each frequency bin. Only results of the

experiments that select the top X% in power across the whole time-frequency domain, for various choices for X, are shown in the evaluations.

Fig. 2.6, Fig. 2.7, Fig. 2.8 demonstrates the performance of the ITDh_NA, ITDh, and GMDA_Laplace algorithm respectively as a function of the angular separation between the sources with different levels of simple power thresholding in an anechoic environment. In each figure the first row shows the results under the 2-source experiment, while the second row shows the results under the 3-source experiment. The suffix '_PT_XX' in the legend denotes the top 'XX' percent of the spectrum points in power are used. The suffix '_PT_100' in the legend denotes the top 100 percent of the spectrum points, i.e. all spectrum points, are used. The legend 'ITDh_NA', 'ITDh', 'GMDA_Laplace' without any further suffix denotes the sinusoidal modeling is used. In each figure, plot (a) and (d) shows the root mean squared (RMS) DOA estimation error as a function of the source angular separation, plot (b) and (e) (resp. (c) and (f)) demonstrate the percentage of trials in which the number of sources is overestimated (resp. underestimated) vs. source angular separation.

When the number of sources is estimated incorrectly, the DOA estimation error is calculated based on the best match between the estimated and true DOAs of the sources for each true DOA. For example, given 2 sources at locations $r_1 = 30^\circ$ and $r_2 = -20^\circ$ and 3 estimated DOAs at $e_1 = 32^\circ$, $e_2 = 25^\circ$, and $e_3 = -20^\circ$. The DOA estimation RMSE is obtained based only on $dist(r_1, e_1)$ and $dist(r_2, e_3)$, ignoring e_2 altogether. This metric is selected based on the fact that the underestimation of the number of sources is usually considered as more detrimental than the overestimation in many real applications. Thereby, we try to control the error in underestimating the number of sources while keeping the overestimation low at the same time in all the experiments. The best-match DOA RMSE metric combined with the overestimation and underestimation metrics can demonstrate different aspects of a DOA estimation algorithm. One thing to note about the best-match DOA RMSE metric is that it will favor the overestimation of the number of sources since any poor matches will be ignored if a good match exists.

For the ITD histogram based algorithms (ITDh and ITDh_NA), the parameters of the smoothing low pass filter control the estimation of the number of sources. The

parameters of the low pass filter have to be balanced between false alarm (spurious peaks in the histogram) and the DOA estimation resolution. The narrower the bandwidth of the low pass filter, the smoother the histogram curve. This results in less spurious peaks in the histogram and more accurate estimation of the number of sources. However, if the bandwidth of the low pass filter is narrow and the separation between two sources is small, the two peaks in the histogram representing the two sources may merge into one peak, or the two peaks may exhibit bias thus increasing the DOA estimation error and decreasing the resolution. This explains the relatively large percentage of error in underestimating the number of sources and the relatively large DOA estimation error for the ITDh_NA and ITDh algorithm when the source separation is 5° in the experiments.

In Fig. 2.6 and Fig. 2.7, the ITDh_NA and ITDh algorithm tend to overestimate the number of sources. The ITDh_NA algorithm demonstrates similar performance when the simple power thresholding method selects the top 10% to 40% time-frequency points in power. It also shows similar performance when all the spectrum points are used, i.e. no thresholding. This can be understood by noting that ITDh_NA uses only the time-frequency points in the low frequency band ($\omega < \omega^d$), which typically have significant power among all the spectrum points as the power of speech signals mainly resides in the low frequency band. The ITDh_NA and ITDh algorithm show less error in underestimating the number of sources when the sinusoidal modeling is used compared to the simple power thresholding, hence the smaller DOA estimation error when the sinusoidal modeling is used. Fig. 2.7 (b), (e) shows for the ITDh algorithm the error in overestimating of the number of sources is correlated with the directions of the source signals. This is to be expected since the stencil filters may create false peaks on the ITD histogram depending on the directions of the source signals.

Fig. 2.8 illustrates that the GMDA_Laplace algorithm only has a few errors in overestimating and underestimating the number of sources (except when all the spectrum points are used). When all the spectrum points are used, the GMDA_Laplace algorithm may not converge properly and therefore exhibits poor performance. The GMDA_Laplace algorithm using the sinusoidal modeling shows better performance than using the simple power thresholding considering the balance between the overestimation and underestimation of the number of sources.

Table 2.1: Average DOA RMSE (degrees), comparison of power thresholding methods

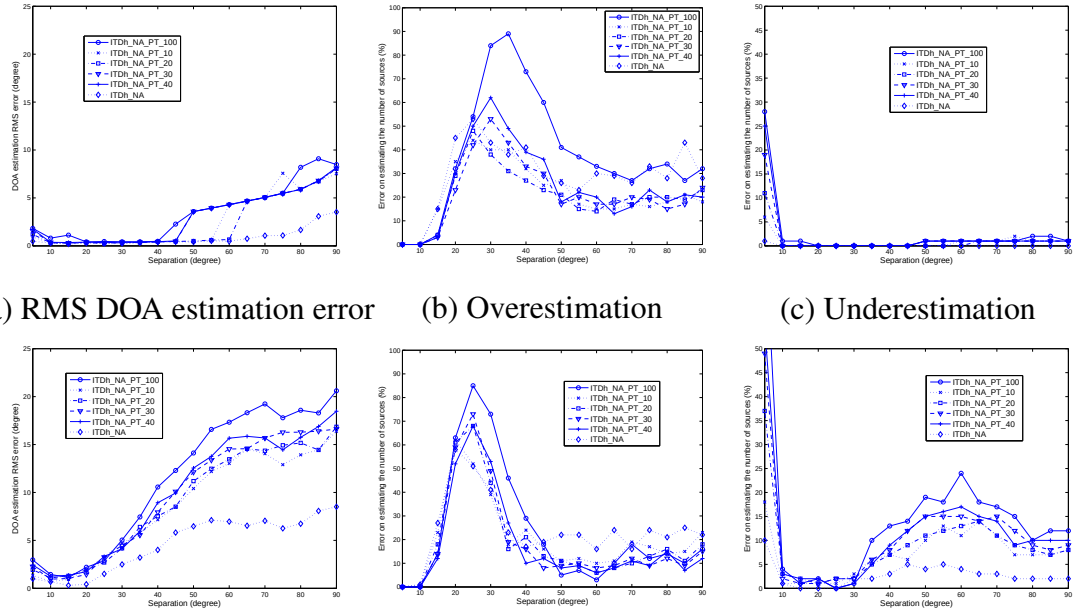
	PT_100	PT_10	PT_20	PT_30	PT_40	Sinusoidal
ITDh_NA	7.44	5.62	5.68	6.25	6.41	2.74
ITDh	2.94	3.95	2.68	2.20	2.08	1.50
GMDA_Laplace	25.04	1.20	1.76	1.88	2.04	0.61

Table 2.1 summarizes the average DOA RMSE performance across all angles of the three different algorithms used in Fig. 2.6, Fig. 2.7, and Fig. 2.8. In summary, we can conclude the simple power thresholding is a good approximation to the sinusoidal modeling in our proposed DOA estimation system, while the sinusoidal modeling achieves better performance at the expense of more complicated implementation. The GMDA_Laplace algorithm shows the best performance among the ITDh_NA, ITDh, and GMDA_Laplace algorithm, i.e. it shows the smallest DOA estimation error and the smallest error in overestimating and underestimating the number of sources among the three.

2.4.4 Experiment 2: Anechoic Environment

Fig. 2.9 illustrates the performance of various algorithms as a function of the angular separation between the sources in an anechoic environment. The first row shows the results in the 2-source experiment, while the second row shows the results in the 3-source experiment. The suffix '_PT_100' in the legend denotes all the spectrum points are used. The legends without the suffix '_PT_100' denotes the sinusoidal modeling is used. Plot (a) and (d) shows the DOA estimation RMS error vs. source angular separation, plot (b) and (e) (resp. (c) and (f)) demonstrate the percentage of trials the number of sources is overestimated (resp. underestimated) vs. source angular separation. Table 2.2 summarizes the average DOA RMSE performance across all angles of the algorithms used in Fig. 2.9.

When the source is close to the lateral direction (90°), the DOA estimation error for all the algorithms increases. This is to be expected because a fixed difference in ITD represents a larger difference in angle when it is close to the broadside direction (0°) than when it is close to the lateral direction (90°). Among the various algo-

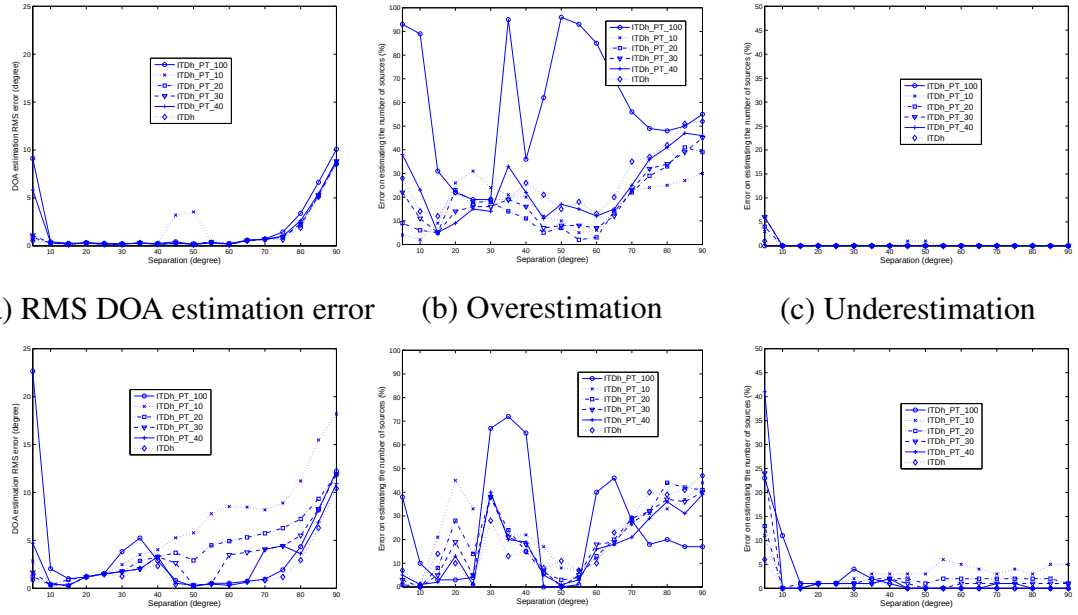


(a) RMS DOA estimation error (b) Overestimation (c) Underestimation

(d) RMS DOA estimation error (e) Overestimation (f) Underestimation

Figure 2.6: Performance of the ITDh_NA algorithms as a function of angular separation of the sources with different levels of power thresholding. The first row shows the results in the 2-source experiment, while the second row shows the results in the 3-source experiment.

rithms, GMDA_Laplace has the smallest DOA estimation error and also estimates the number of sources accurately most of the time. There are still a few occasions when the GMDA_Laplace algorithm may overestimate the number of sources. This happens if one of the two sources is weak (in power), which results in increased outliers in clustering and an extra model order increases the likelihood considerably. A further post-processing step for removing outliers or an improvement on the sinusoidal tracks extraction program [128] may alleviate this problem, but it will complicate the system and we do not pursue it further in this chapter. The GMDA_Gauss algorithm has a high probability of overestimating the number of sources. As is known, the Gaussian distribution as a noise model is sensitive to outliers in model learning [8][132, Ch.2], while the Laplacian distribution has a heavier tail and is more robust to outliers. The GHCA algorithm estimates the number of sources as accurately as the GMDA_Gauss algorithm does, while its DOA estimation error is a little higher.



(a) RMS DOA estimation error (b) Overestimation (c) Underestimation

(d) RMS DOA estimation error (e) Overestimation (f) Underestimation

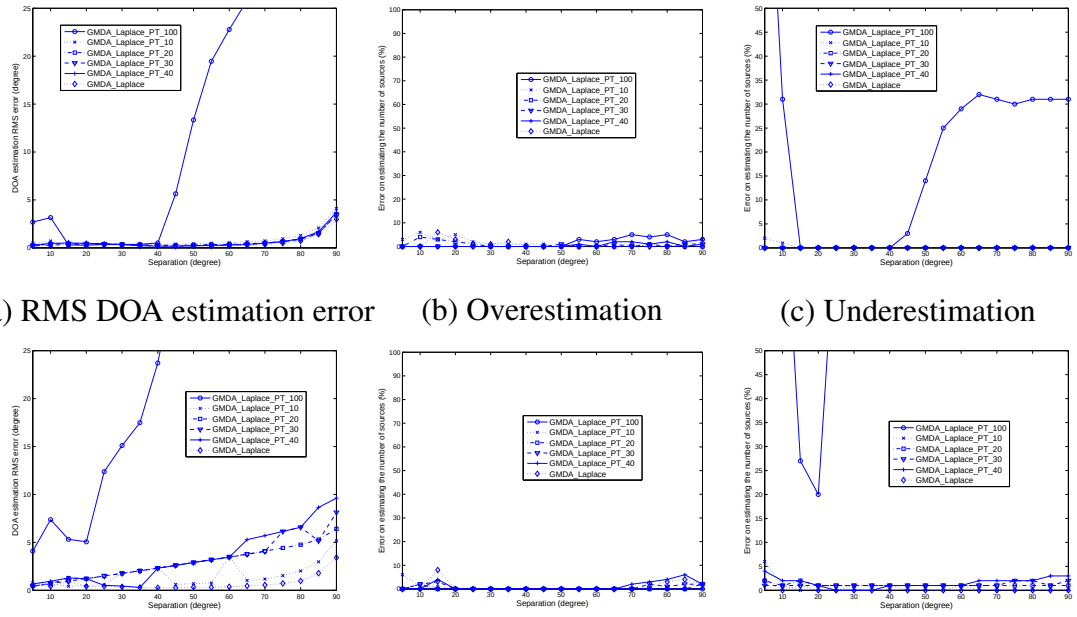
Figure 2.7: Performance of the ITDh algorithms as a function of the angular separation of the sources with different levels of power thresholding. The experiment setup is the same as used in creating Fig. 2.6.

Table 2.2: Average DOA RMSE (degrees), anechoic environment

GMDA_Gauss	GMDA_Laplace	GHCA	ITDh	ITDh_PT_100
1.08	0.61	0.97	1.50	2.94

2.4.5 Experiment 3: Echoic Environment

In this experiment, we study the robustness of the proposed algorithm to room modeling assumption where instead of a simple direct path, the environment is reverberant. Since the proposed algorithm is based on inter-channel time difference, it falls into the same category as GCC and the analysis on GCC under reverberant environment also applies to the proposed algorithm [8, 133, 14, 15]. Therefore, the reverberation paths of the sources can be considered as extra noise components and if the signal to reverberation ratio (SRR) is high, the proposed algorithm can still work to localize the sources. This experiment used the room impulse response generated by the image method [134]. The simulated room dimension is $[8, 5, 3.5]m$, and the reflection coeffi-



(a) RMS DOA estimation error (b) Overestimation (c) Underestimation

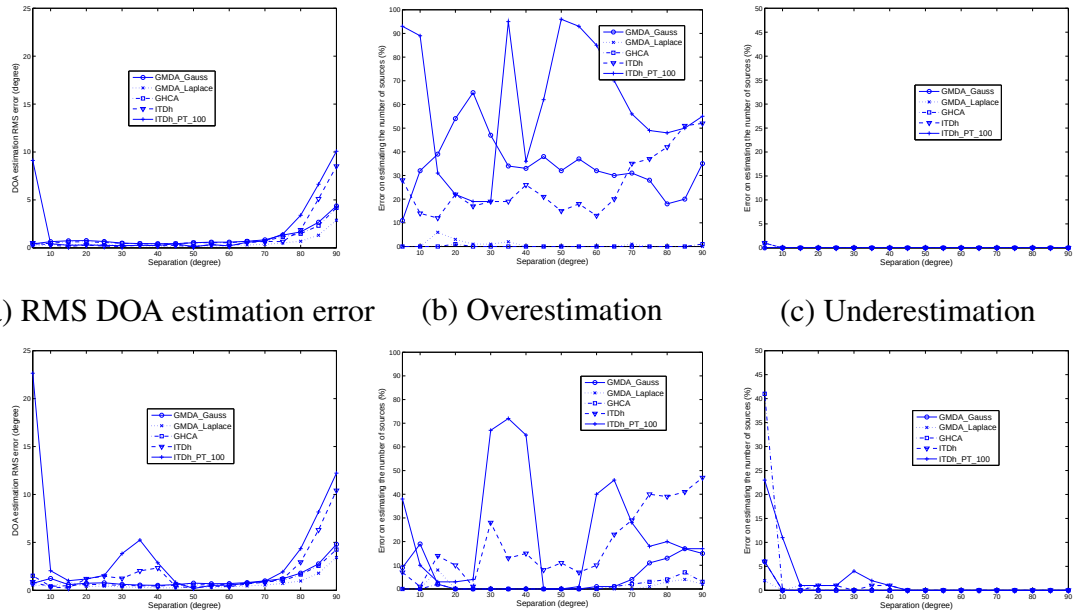
(d) RMS DOA estimation error (e) Overestimation (f) Underestimation

Figure 2.8: Performance of the GMDA_Laplace algorithm as a function of the angular separation of the sources with different levels of power thresholding. The experiment setup is the same as used in creating Fig. 2.6.

cient is -0.4 which results in a reverberation time of approximately $220ms$.

Fig. 2.10 shows one example of the data resulting from the GMDA_Laplace algorithm under reverberation. There are 2 speech sources with DOA of $60°$ and $-45°$ respectively. The inter-microphone spacing is $12cm$. The estimated DOA of the sources are $(57.43°, -42.65°)$. By comparing Fig. 2.4 and Fig. 2.10, it is evident the reverberation causes the data points to be more dispersed on the IPD vs. frequency plot, which makes the DOA estimation task more challenging.

Fig. 2.11 illustrates the performance of the various algorithms as a function of angular separation between the sources in an echoic environment. The experiment setup is the same as used in creating Fig. 2.9. By comparing Fig. 2.9 and Fig. 2.11, it is apparent the performance of various algorithms deteriorate under reverberation. The GMDA_Laplace algorithm still has the best performance among compared algorithms without significantly overestimating or underestimating the number of sources. The DOA estimation error for the GMDA_Laplace algorithm is sometimes worse than the GMDA_Gauss and GHCA algorithm when the separation angle is large. This can be ex-



(a) RMS DOA estimation error (b) Overestimation (c) Underestimation

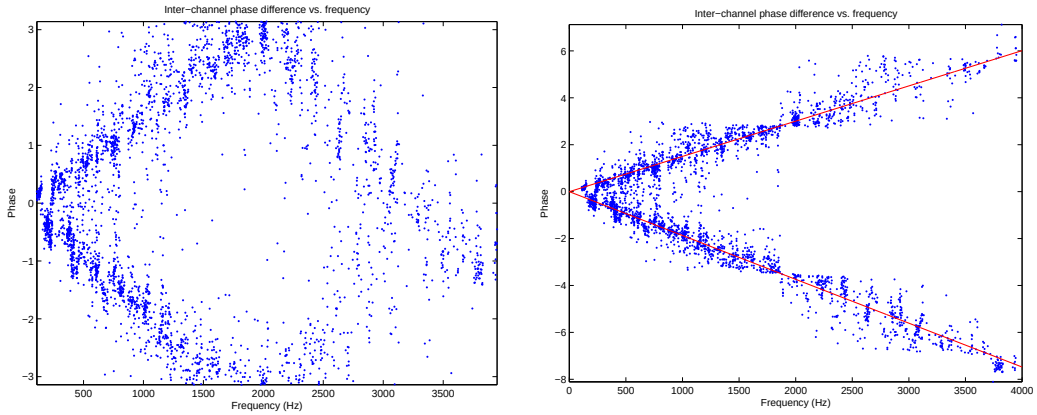
(d) RMS DOA estimation error (e) Overestimation (f) Underestimation

Figure 2.9: Performance of various algorithms as a function of the angular separation of the sources in an anechoic environment. The first row shows the results under the 2-source experiment setup, while the second row shows the results under the 3-source experiment setup.

plained by noting that the GMDA_Gauss and GHCA algorithm overestimate the number of sources and in our analysis the DOA estimation error is calculated based on the best match between the estimated and true DOAs.

2.4.6 Experiment 4: Additive White Noise

In this experiment, the robustness of the proposed algorithms to ambient white noise is studied. High local signal to white noise ratio (SNR) on points of the sinusoidal tracks (or on the points selected by the simple power thresholding) is expected even though the global SNR across the whole spectrum might be low. Therefore, good DOA estimation would be expected even when the white noise level is high. This experiment uses 2 speech sources, and the DOA of the two sources are $30°$ and $-45°$ respectively. The inter-microphone spacing is $12cm$. The white Gaussian noises at the two microphones is independent. Experiments demonstrate that when the SNR is higher than 10



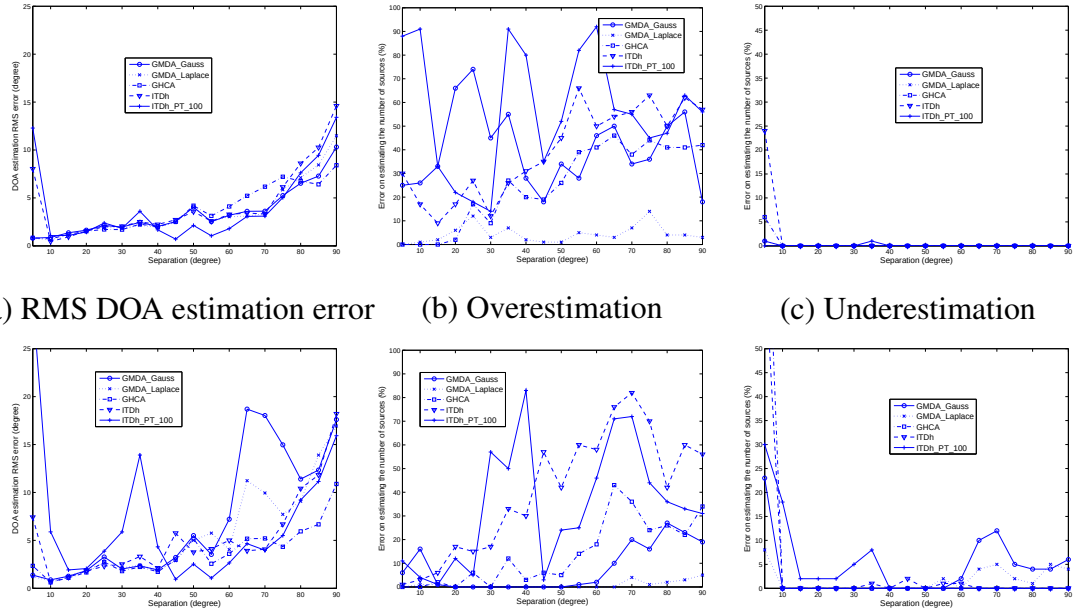
(a) IPD vs. frequency

(b) GMDA_Laplace for line fitting

Figure 2.10: Inter-channel phase difference vs. frequency. Clustering using GMDA_Laplace algorithm. It is in a reverberant environment with 220ms reverberation time. There are 2 speech sources, whose DOAs are 60° and −45° respectively. The inter-microphone spacing is 12cm.

dB, the DOA estimation of the proposed algorithms using sinusoidal modeling is quite accurate with the average DOA estimation error lower than 0.5°. However, when the SNR is lower than 5 dB, the DOA estimation error is large. Closer inspection of the data reveals that the problem is a result of poor sinusoidal track extraction. When the SNR is low, the sinusoidal track extraction program used [128] may fail to pick up the true sinusoidal tracks from the spectrum. By adjusting the parameters of the sinusoidal tracking program, true sinusoidal tracks can be extracted from the spectrum resulting in small average DOA estimation error. The performance of various algorithms under sinusoidal modeling is illustrated in the first row of Fig. 2.12 after adjusting the parameters of the sinusoidal tracking program.

The results show that the ITDh algorithm almost always overestimates the number of sources when the SNR is low. The GMDA_Laplace algorithm has the smallest DOA estimation error and estimates the number of sources accurately when the SNR is higher than 5 dB. There is almost no underestimation of the number of sources except when the SNR is at 0 dB, when the GMDA_Gauss (resp. GMDA_Laplace) algorithm underestimates the number of sources in 13 (resp. 7) percent of the trials. A closer inspection of the data shows that sometimes one source is weak compared to the other source resulting in a disparity in SNR. This results in the weaker source being masked and not



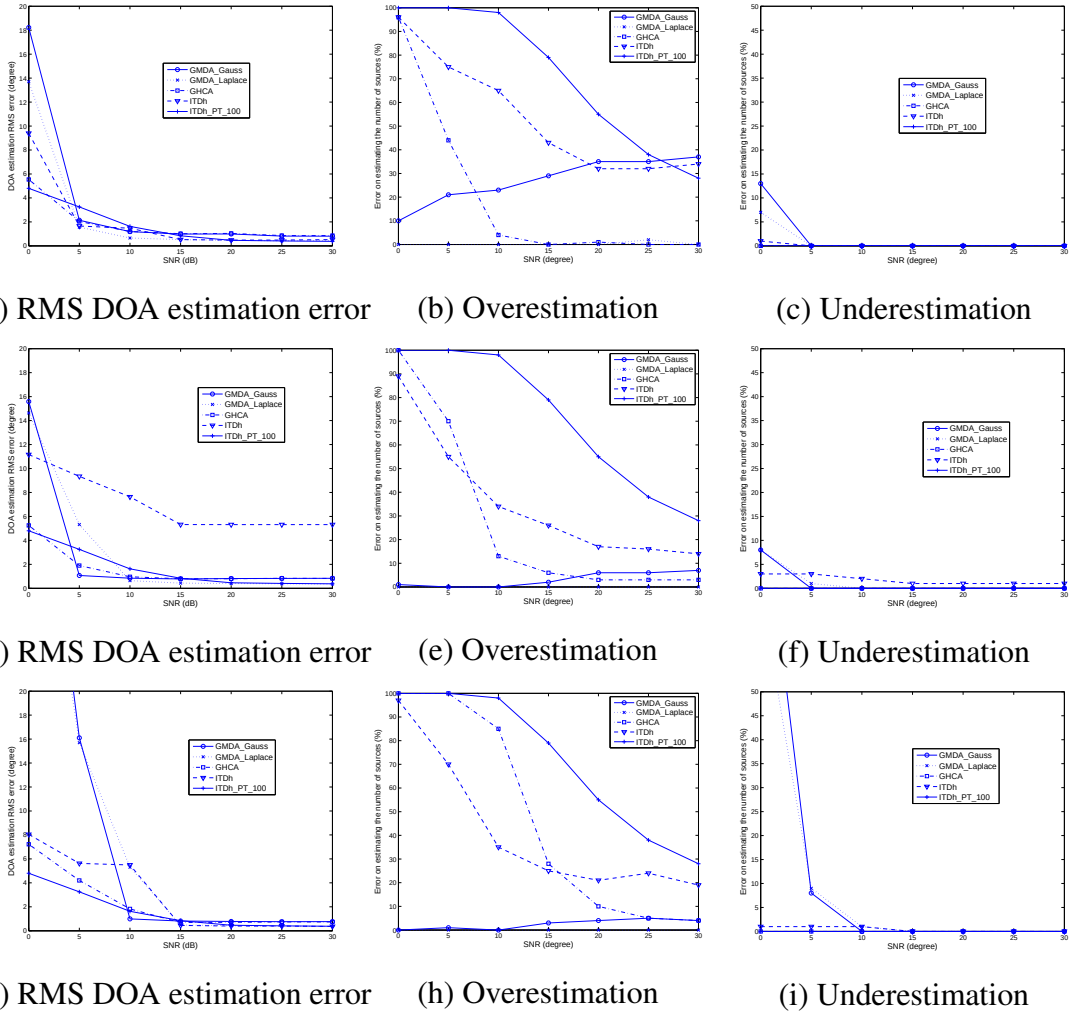
(a) RMS DOA estimation error (b) Overestimation (c) Underestimation

(d) RMS DOA estimation error (e) Overestimation (f) Underestimation

Figure 2.11: Performance of various algorithms as a function of the angular separation of the sources in an echoic environment. The first row shows the results under the 2-source experiment setup, while the second row shows the results under the 3-source experiment setup.

enough sinusoidal tracks can be extracted leading to the underestimation of the number of sources. The relatively large RMS DOA estimation error for the GMDA_Laplace algorithm at 0 dB is contributed by the 7 percent of trials when underestimation of the number of sources occured. Increasing the length of the speech sources alleviates the underestimation problem since there is a higher chance more sinusoidal tracks are extracted thereby increasing the number of data points on the IPD vs. frequency plot. The GHCA algorithm estimates the DOAs accurately and consistently across all SNR level at the expense of overestimation of the number of sources at low SNR. Not surprisingly, the proposed DOA estimation algorithms depend on the successful extraction of sinusoidal tracks from the spectrum. Better sinusoidal track extraction algorithm will enhance the performance of the proposed DOA estimation algorithms.

The second (resp. third) row of Fig. 2.12 shows the results using simple power thresholding which selects the top 10% (resp. 20%) time-frequency points in power. The performance of the GMDA_Laplace algorithm using the simple power thresholding se-



(a) RMS DOA estimation error (b) Overestimation (c) Underestimation

(d) RMS DOA estimation error (e) Overestimation (f) Underestimation

(g) RMS DOA estimation error (h) Overestimation (i) Underestimation

Figure 2.12: Performance of various algorithms versus SNR. There are 2 speech sources, whose DOAs are 30° and −45° respectively. The inter-microphone spacing is 12*cm*.

lecting top 10% points in power is similar to the performance when using the sinusoidal modeling. When the power thresholding is lowered to select the top 20% time-frequency points in power, the GMDA_Laplace algorithm underestimates the number of sources in many trials when the SNR is at 0 dB. When the simple power thresholding selects the top 20% time-frequency points in power, a lot of noisy points are included degrading the GMDA_Laplace algorithm. The experiments suggest that the GMDA_Laplace algorithm with sinusoidal modeling or with simple power thresholding selecting the top 10% time-frequency points in power is the best choice for DOA estimation in low SNR environments.

2.5 Conclusion

This chapter proposes a two microphone based source localization technique for multiple speech sources utilizing speech specific properties and clustering and line fitting techniques. Speech's sparsity in time-frequency domain, the masking effects and sinusoidal modeling are combined together and utilized for dual channel DOA estimation. Two key ideas form the basis of the method developed: one is selecting time-frequency points with high local SNR using sinusoidal modeling and the other is using GMDA clustering algorithm for DOA estimation. Experiments show that power thresholding might be a simple and good approximation to the sinusoidal modeling in selecting points with high local SNR. The sinusoidal modeling however exhibits better performance at the expense of a more complicated implementation. The proposed GMDA_Laplace algorithm performs better than the baseline ITD histogram based DOA estimation algorithm. It estimates the number of sources accurately in different scenarios and shows smaller DOA estimation error than the ITDh algorithm which tends to overestimate the number of sources. Compared with GMDA_Laplace, the GMDA_Gauss algorithm exhibits worse performance and confirms the robustness of the Laplacian distribution to outliers. Experiments confirm that the proposed DOA estimation system is robust to the additive white noise because of the high local SNR on points of the sinusoidal tracks (or on the points selected by the simple power thresholding) even though the global SNR across the whole spectrum is low. The proposed DOA estimation system depends on the successful extraction of the sinusoidal tracks. A robust sinusoidal tracks extraction algorithm would result in the proposed DOA estimation system also being robust. Finally, in the appendix of this chapter, we propose a fast adaptive MVDR beamformer for dual channel speech signal separation based on speech's sparsity attribute and non-stationarity. Experiment shows the method to be very promising.

Acknowledgement

The text of this chapter, in full, is a reprint of the material as it appears in Wenyi Zhang, and Bhaskar D. Rao, "A Two Microphone Based Approach for Source Localization of Multiple Speech Sources", submitted to *IEEE Transactions on Audio, Speech*

and Language Processing and Wenyi Zhang, and Bhaskar D. Rao, "Two Microphone Based Direction of Arrival Estimation for Multiple Speech Sources using Spectral Properties of Speech", *Acoustics, Speech, and Signal Processing, 2009. ICASSP 09. IEEE International Conference on*, pp. 2193–2196, Apr. 2009.

2.6 Appendix for Chapter 2

2.6.1 Fast Adaptive MVDR Beamformer for Source Separation

In a complete dual channel speech processing system, after estimating the DOAs of the speech sources, one may want to separate these speech sources, i.e. enhance the desired source and suppress the interfering signals. In this section, a fast adaptive MVDR (minimum variance distortionless response) beamformer [20, 21, 1] for dual channel speech signal separation based on speech's sparsity attribute and non-stationarity is described. A narrowband beamformer is employed for each frequency bin.

For dual channel microphone array, the weight vector $\mathbf{w}$ of the standard MVDR beamformer (section 4.2.1) has one free dimension and can fully suppress one interfering signal [1]. However, if there are more than one interfering signals, the standard batch algorithm based MVDR beamformer can not fully cancel all interfering signals. Therefore, it may not be an efficient interference suppression algorithm for the dual channel multiple speech sources scenario.

However, as emphasized in Sec. 2.2, speech is a sparse signal in time-frequency domain. At a specific time-frequency point, there is a high likelihood that at most one interference signal is significant (in power) and the contributions from other interfering signals is negligible. Therefore, to enhance the desired speech signal, it is not necessary to put nulls in all the interfering signals' directions. Only the most powerful interference at that time-frequency point is to be suppressed. Since speech is a fast changing quasi-stationary process (each phoneme usually last a few frames), the sparsity property affected by quasi-stationarity of speech implies that the beamformer must be adapted rapidly. In standard batch algorithm based MVDR beamformer, the signal correlation matrix R is calculated based on a long segment of data which can be considered as inde-

pendent of the time index k (frame number). Instead, we propose a fast adaptive MVDR beamformer wherein the correlation matrix is dependent on the time index k (frame number). The beamformer's weight vector $\mathbf{w}_k$ at time index k is

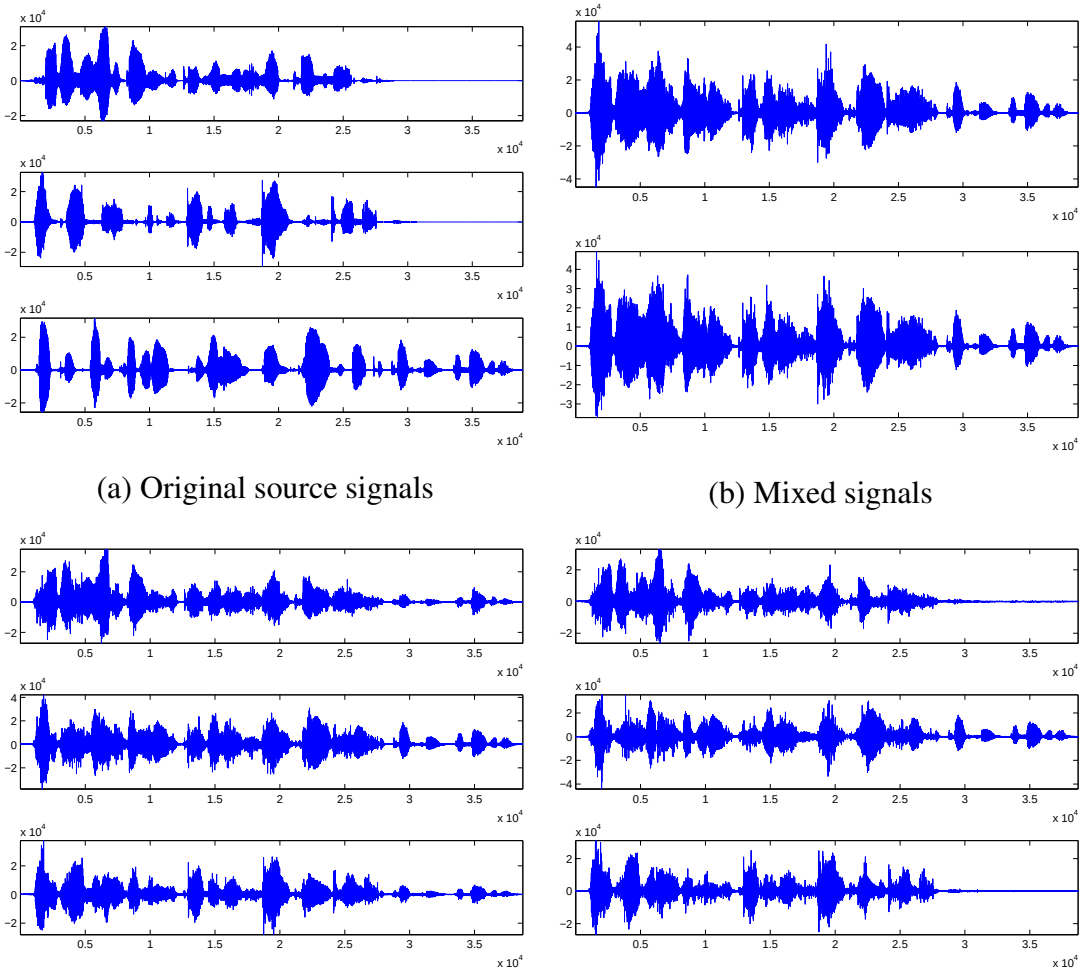
$$\mathbf{w}_k = \frac{R_k^{-1}\mathbf{a}}{\mathbf{a}^H R_k^{-1}\mathbf{a}}. \tag{2.24}$$

$\mathbf{a}$ is the steering vector of the desired source. To adapt to the fast changing statistics of speech signals, the correlation matrix R_k only depends on a few data vector $\mathbf{z}_j$ around current time index k. For example,

$$R_k = \sum_{j=k-L}^{j=k+L} \mathbf{z}_j \mathbf{z}_j^H \tag{2.25}$$

$\mathbf{z}_j = [X_1(\omega, j), X_2(\omega, j)]^T$ denotes the 2 sensor array's data vector at frequency ω and frame number j. Note frequency ω is ignored in above formulas to simplify notation. L is a small integer to control the time window size for calculating R_k. R_k can also be calculated in other ways as long as it represents a short time statistics around current time k. Since the array only has two sensors, the limited data problem and look direction error problem do not pose serious problems for the MVDR beamformer [21, 1].

The following experiment demonstrates the performance of the fast adaptive MVDR algorithm based on speech's sparsity and non-stationarity. The DOA of the three sources are $60°$, $0°$ and $-45°$ respectively. The inter-microphone spacing is $12cm$. Fig. 2.13 (a)(b)(c)(d) shows the original sources, mixed signals, batch algorithm based MVDR beamformer outputs, and fast adaptive MVDR beamformer outputs respectively. The cepstral distance (smaller is better) [135] between original sources and batch algorithm based MVDR beamformer outputs is 0.69, while the cepstral distance between original sources and fast adaptive MVDR beamformer outputs is 0.56. The fast adaptive MVDR beamformer shows better performance than the standard batch algorithm based MVDR beamformer. Informal listening also indicates the output to be perceptually better.



(a) Original source signals

(b) Mixed signals

(c) MVDR BF (batch Alg) outputs

(d) Fast adaptive MVDR BF outputs

Figure 2.13: Performance of the fast adaptive MVDR beamforming algorithm

3 Robust Adaptive Broadband Beamformer

Compared with single channel speech processing, multi-microphone speech processing has high interference suppression in noisy environment because of the spatial filtering capability. The standard data independent beamformers such as delay-and-sum (DS) beamformer is robust to the signal of interest (SOI) steering vector errors, which may be due to look direction error, array sensor position error, and small mismatches in the sensor responses. Yet those beamformers suffer from low resolution and high sidelobes, inducing bad interference rejection capability. In contrast, data dependent adaptive beamformers have high resolution and interference rejection capability when the array steering vector is accurately known. However, the performance of the traditional adaptive beamformer can degrade severely in practice when there exist SOI steering vector errors. In such cases, the SOI might be mistaken as an interference signal and be suppressed. Fig.3.1 demonstrates a look direction error in a linear array system. In this example, the look direction is the broadside direction, i.e. 0°. However, the true desired signal incidence angle is 5°, which means there's a 5° look direction error. In this case, the traditional adaptive beamformer will take the desired signal as an interference signal and suppress it. This SOI steering vector error problem should be addressed before the adaptive beamforming algorithms can be applied to process real world speech data.

In this chapter, we focus on developing broadband beamforming algorithms which are robust to the steering vector errors. Later on in chapter 4, we will concentrate on developing robust narrowband beamforming algorithms. In this work, we analyze the convergence of the Frost iterative and LMS algorithm for signals of different incidence angles. Our analysis exhibits that the Frost LMS algorithm is robust to the look direc-

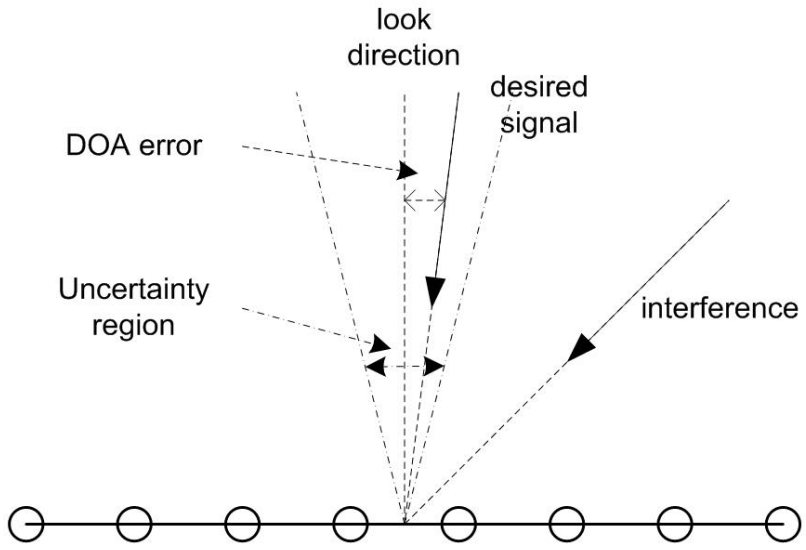


Figure 3.1: A demonstration of the look direction error problem

tion error when the adaptation does not go through too many iterations (i.e. the speech source length is not too long). We develop a robust broadband adaptive beamforming algorithm which combines the robustness of the DS beamforming in the look direction with the high interference rejection capability of the conventional adaptive beamforming algorithm. We test the proposed algorithm on the real world recorded Multi-channel Overlapping Numbers Corpus (MONC) [136] and show that it has better performance compared to other existing algorithms.

3.1 Introduction to Robust Broadband Beamforming

In this work, we consider the development of robust wideband adaptive beamforming algorithms for speech enhancement and noise suppression. An adaptive beamformer is able to adjust its beam pattern based on the input statistics to place deep nulls in the direction of interferences. Among the broadband adaptive beamformer, the Frost beamformer is one of the most extensively studied [19]. The Frost beamformer has a multichannel tapped-delay-line structure (Fig.3.2) with a set of linear

constraints that are introduced to ensure a desired frequency response in the look direction. However, the performance of the Frost beamformer can degrade severely in practice when steering vector errors exist, which may be due to look direction error, array sensor position error, and small mismatches in the sensor response. In such cases, the desired signal might be mistaken as an interference signal and be suppressed [21, 137]. Several robust beamforming algorithms have been proposed to address this problem [137, 138, 23, 139, 140, 141, 142, 143, 144, 37]. In [139], the steering vector errors are modelled by "time-delay errors" and compensated for by self-adjusted interpolation filtering. In [140], a method is proposed to optimize the worst-case performance. The problem is formulated as minimizing a quadratic function subject to infinitely many quadratic constraints. It is reduced to a second-order cone programming problem which can be solved by interior point methods. In [23], derivative constraint are used to provide look direction robustness.

Er and Cantoni proposed a robust broadband beamforming algorithm which restricts the error between the desired and actual beam pattern of the array over the frequency band of interest and over a small spatial region around the array's look direction, allowing for uncertainty in the look direction [137, 138]. The constraint thus obtained can be a quadratic constraint or reduced to a set of linear constraints. Zheng proposed another robust broadband adaptive beamforming algorithm recently [141]. Her paper assumed a near field signal model and took sufficient sampling points in the look direction. For each sampling point, a constraint is constructed. The compact form of all the constraints is $A^T\mathbf{w} = \mathbf{g}$, where every column of matrix A represent a constraint. Then SVD is used to get a low rank matrix A_L approximating A. And the constraint is approximated by $A_L^T\mathbf{w} = \mathbf{g}$. In fact, Zheng's algorithm is close to Er's, as summation can be viewed as an integration if the number of sampling points is large enough. Then Zheng's constraint is the same as Er's constraint. However, Er's linear constraint algorithm is an approximation because the quadratic constraint is not strictly equal to the set of linear constraints. An additional norm constraint is imposed to overcome this limitation and this complicates the optimization problem.

In this work, we analyze the convergence of the Frost iterative and LMS algorithm for signals of different incidence angles. Our analysis exhibits that the Frost LMS

algorithm is robust to the look direction error when the adaptation does not go through too many iterations (i.e. the speech source length is not too long). We also develop a quadratic constraint based robust broadband beamforming algorithm to deal with the uncertainty in the look direction. In order to address the ill-conditioning associated with the constraint matrix, a diagonal loading (DL) is added to the constraint matrix thereby ensuring a robust solution to the quadratic constraint beamforming problem. The advantage of adding DL to constraint matrix is that the constraint matrix is only determined by the geometry of the array thereby allowing the DL level to be chosen offline. This is superior to adding DL to the signal covariance matrix where the DL level has to be chosen online. It is shown that the diagonal loading is equivalent to an additional norm constraint without introducing it explicitly. We also develop an iterative algorithm (and corresponding adaptive algorithm) to solve for the robust beamformer coefficients. The developed algorithm is applied to the problem of beamforming using microphone arrays for speech recognition and shown to be superior to existing algorithms.

This chapter is organized as follows. In section 3.2 we discuss the classical Frost beamforming problem, including the optimum, the iterative and the LMS algorithm. We also analyze the convergence of the Frost iterative and LMS algorithm for signals of different incidence angles. Section 3.3 exhibits the development of the quadratic constraint robust broadband beamforming problem and discusses the mathematical solutions to the problem. Section 3.4 proposes the diagonally loaded quadratic constraint robust broadband beamforming algorithm. Section 3.5 illustrates the experiment results.

3.2 Frost Beamformer

3.2.1 Frost Beamformer: Optimum Solution

Among the broadband adaptive beamformer, the Frost beamformer is one of the most extensively studied [19]. The Frost beamformer has a multichannel tapped-delay-line structure (Fig. 3.2), and a set of linear constraints are introduced to ensure a desired frequency response in the look direction. Thereby it is also called linear constrained minimum variance (LCMV) beamformer.

Suppose a Frost beamformer use a linear array with K sensors and J taps. The

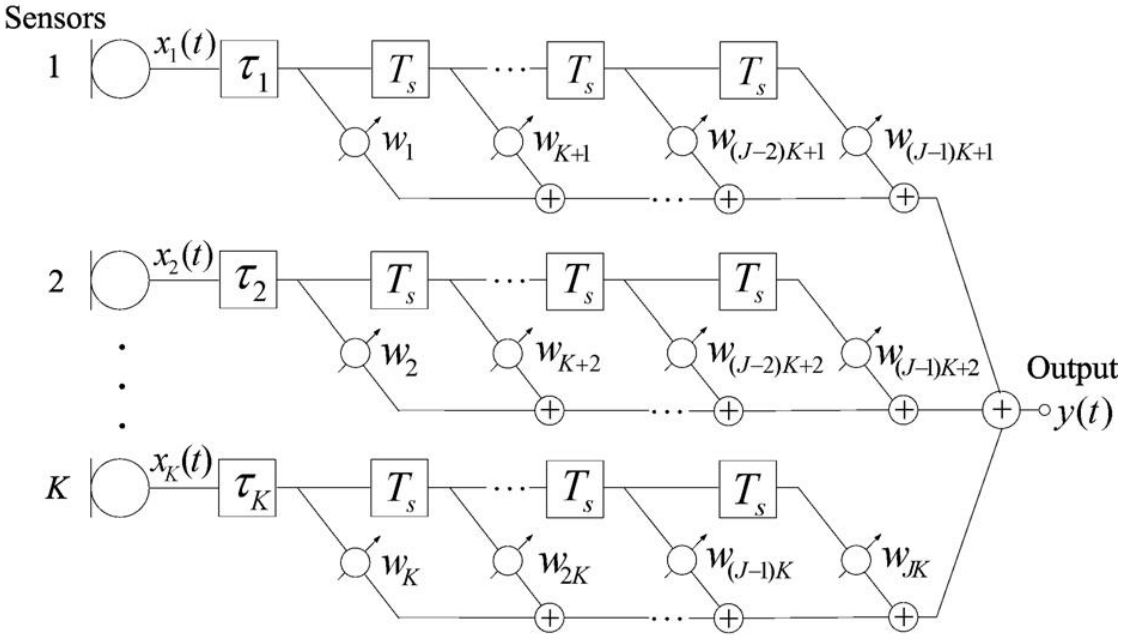


Figure 3.2: A broadband beamformer with K sensors and J taps.

first step is to pre-steer the array in the look direction. The Frost beamforming problem is then mathematically formulated as,

$$\min_{\mathbf{w}} \mathbf{w}^T R_{XX} \mathbf{w}, \quad \text{subject to} \quad C^T \mathbf{w} = F \tag{3.1}$$

where $\mathbf{w}$ is the KJ dimension concatenated weight vector for all the taps. R_{XX} is the KJ dimension covariance matrix. $F = [f_1, f_2, \cdots, f_J]^T$ is the vector of weights of the look-direction-equivalent tapped delay line, a simple choice for F will be $F = [1, 0, \cdots, 0]^T$. C is the constraint matrix with KJ rows and J columns, in which the j_{th} column picks up the j_{th} column of the array elements in the array matrix.

Using the Lagrange multiplier method, the solution to the Frost beamforming problem (Eq. (3.1)) can be obtained,

$$\mathbf{w}_{opt} = R_{XX}^{-1} C [C^T R_{XX}^{-1} C]^{-1} F \tag{3.2}$$

3.2.2 Frost Beamformer: Iterative Solution

In this subsection, the iterative solution to the Frost beamforming problem is derived [19]. Using constrained gradient-descent optimization at each iteration, the weight

vector is updated as follows,

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \mu\nabla_{\mathbf{w}}H[\mathbf{w}(k)] \tag{3.3}$$

where $H(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T R_{XX}\mathbf{w} + \lambda^T(C^T\mathbf{w} - F)$ is the constrained cost function. Thus we can get,

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \mu[I - C(C^TC)^{-1}C^T]R_{XX}\mathbf{w}(k) + C(C^TC)^{-1}[F - C^T\mathbf{w}(k)] \tag{3.4}$$

The last term is not assumed to be zero. It is kept to prevent arithmetic inaccuracy from accumulation and growth. Define $G = C(C^TC)^{-1}F$ and $P = I - C(C^TC)^{-1}C^T$, the iterative equation can be written as,

$$\mathbf{w}(k+1) = P[\mathbf{w}(k) - \mu R_{XX}\mathbf{w}(k)] + G \tag{3.5}$$

3.2.3 Frost Beamformer: LMS Algorithm

Equation (3.5) defines an iterative constrained gradient descent algorithm. The corresponding LMS algorithm is derived as

$$\mathbf{w}(0) = G \tag{3.6}$$

$$\mathbf{w}(k+1) = P[\mathbf{w}(k) - \mu y(k)X(k)] + G \tag{3.7}$$

3.2.4 Convergence Analysis of the Frost Iterative and LMS Algorithm for Signals of Different Incidence Angles

In this subsection, we analyze the convergence speed of the iterative Frost algorithm for signals of different incidence angles. Define the difference process $\mathbf{v}(k) = \mathbf{w}(k) - \mathbf{w}_{opt}$, where $\mathbf{w}_{opt}$ is defined in equation (3.2). Subtracting $\mathbf{w}_{opt}$ from both sides of equation (3.5), and using the knowledge that $G = (I - P)\mathbf{w}_{opt}$ and $PR_{XX}\mathbf{w}_{opt} = 0$, the equation for the difference process is obtained,

$$\mathbf{v}(k+1) = P\mathbf{v}(k) - \mu PR_{XX}\mathbf{v}(k) \tag{3.8}$$

Since $P^2 = P$,

$$P\mathbf{v}(k+1) = P^2\mathbf{v}(k) - \mu P^2R_{XX}\mathbf{v}(k) = \mathbf{v}(k+1) \tag{3.9}$$

The following can be derived,

$$\mathbf{v}(k+1) = P\mathbf{v}(k) - \mu PR_{XX}\mathbf{v}(k) \tag{3.10a}$$

$$= \mathbf{v}(k) - \mu PR_{XX}P\mathbf{v}(k) \tag{3.10b}$$

$$= [I - \mu PR_{XX}P]\mathbf{v}(k) \tag{3.10c}$$

$$= [I - \mu PR_{XX}P]^{k+1}\mathbf{v}(0) \tag{3.10d}$$

It is obvious that the convergence of the iterative Frost algorithm is controlled by the matrix $PR_{XX}P$. It is easy to prove that $PR_{XX}P$ has J zero eigenvalues, whose corresponding eigenvectors are the column vectors of the constraint matrix C. It can also be shown that $PR_{XX}P$ has $KJ - J$ nonzero eigenvalues σ_i whose values are bounded between the smallest and largest eigenvalues (λ_{min} and λ_{max}) of R_{XX} (see Appendix 3.6.1),

$$\lambda_{min} \leq \sigma_{min} \leq \sigma_{max} \leq \lambda_{max} \tag{3.11}$$

In regard to the robust beamforming problem, we analyze the effect that the signal incidence angle has on the convergence rate of the beamforming weight vector. Suppose only the desired signal is present, and the pre-steering is perfect, the signal vector can be written as,

$$X = C \begin{bmatrix} x[n] \\ \vdots \\ x[n-J+1] \end{bmatrix} \tag{3.12}$$

The correlation matrix R_s for the signal can be written as,

$$R_s = E\{XX^T\} \tag{3.13a}$$

$$= C \begin{bmatrix} r(0) & r(1) & \cdots & r(J-1) \\ r(1) & r(0) & \cdots & r(J-2) \\ \vdots & \vdots & & \vdots \\ r(J-1) & r(J-2) & \cdots & r(0) \end{bmatrix} C^T \tag{3.13b}$$

Since P is a projection which is orthogonal to C, PR_sP is a zero matrix, and $I - \mu PR_sP$ is an identity matrix, which means the desired signal is untouched and passes through the beamforming system no matter how many iterations are taken for the weight vector $\mathbf{w}(k)$.

Table 3.1: Maximum eigenvalue of the projected correlation matrix $PR_{XX}P$ versus the signal incidence angle θ

θ	0°	2°	3°	10°	30°	45°
max eigenvalue	0	$2.4 * 10^5$	$5.4 * 10^5$	$4.9 * 10^6$	$1.7 * 10^7$	$2.5 * 10^7$

If we suppose the pre-steering is not perfect, i.e. there is a look direction error, the correlation matrix R_{XX} will be close to the correlation matrix R_s shown in equation (3.13). This means that the eigenvector of the correlation matrix R_{XX} will be close to the column vectors of C. Since P is orthogonal to C, the eigenvalues of $PR_{XX}P$ will be close to zero. Accordingly, the eigenvalues of $I - \mu PR_{XX}P$ will be close to 1, which means the iteration process described in equation (3.10) will converge very slowly.

Now consider an interference signal coming from a direction far away from the look direction. For such cases, generally, the eigenvectors of the interference signal correlation matrix R_I will be far away from the column vectors of C. So the eigenvalues of the matrix PR_IP are relatively large and the eigenvalues of $I-\mu PR_{XX}P$ will be close to zero. Thus the components corresponding to these eigenvalues in the difference process $\mathbf{v}(k)$ will converge fast. We note that in all the above discussion the parameter μ is assumed to be appropriately chosen so that convergence is assured.

The above discussion is confirmed in the following example. Assume a 8 elements linear array with 20 taps, the look direction is the broadside, i.e. 0°. Only one signal is present at the array. Table 3.1 lists the maximum eigenvalue of the matrix $PR_{XX}P$ with regard to the different signal incidence angle θ.

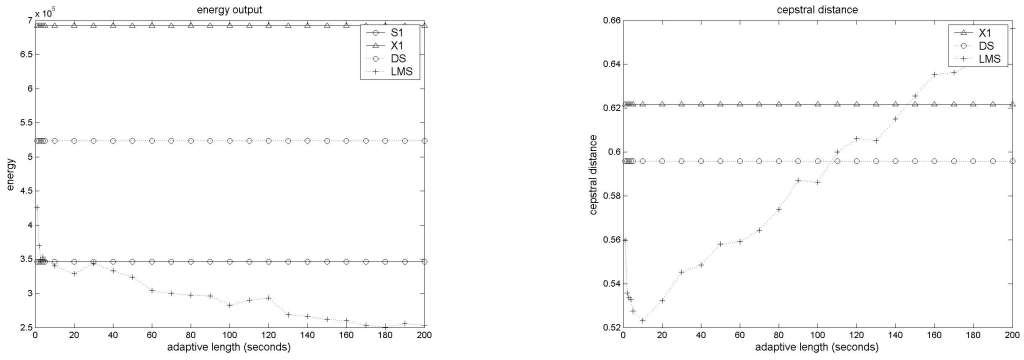
The different convergence rates of the $\mathbf{v}(k)$ process with respect to the signal's incidence angle are important in the analysis of the robustness of iterative Frost algorithm. Fast convergence of the weight vector of the beamformer corresponding to the interference signal which is far away from the look direction means the weight vector is adapted to suppress the interference after a moderate number of iterations. Yet the slow convergence corresponding to the desired signal with a small look direction error assures that the desired signal is not suppressed with a moderate number of adaptation iterations. However, the desired signal will be suppressed if a lot more iterations is done.

The analysis of Frost LMS algorithm's convergence property with regard to the signal incidence angle is similar to that of the Frost iterative algorithm. It can be shown

that these two algorithms will have similar property. This explains why the Frost LMS algorithm is robust to look direction error when the sentence is not too long. When the adaptation of the weight vector of the Frost LMS beamformer goes through a moderate number of adaptation iterations (i.e. the speech source length is not too long), the suppression of the desired signal (which may have a small look direction error) is not significant, while the suppression of the interference signals (which are supposed to be far away from the look direction) is critical. Therefore, the Frost LMS algorithm exhibits good suppression capability for the interference signal while still being robust to small look direction error in the desired signal. However, when the adaptation of the weight vector goes through a large number of adaptation iterations, the desired signal will get suppressed significantly if there is a look direction error.

The above analysis is verified by the following simulation for speech enhancement. We simulate an 8 element linear array with $4cm$ equal spacing between adjacent elements. The sampling rate of the speech sources is $8K$ Hz. We consider the simple case where all the source signals and the linear array exist in the same plane. A Monte Carlo simulation of 20 independent runs was performed for different adaptation lengths. The data for testing performance is 100 seconds long. The result shown in the following is the mean value of those 20 independent Monte Carlo runs. (*To be more specific: for every fixed adaptation length, the same experiment is repeated* 20 *times based on random data selection. That is, for every fixed adaptation length, we pick* 20 *segments of data of that fixed length from the database, run a beamforming algorithm on those segments, get* 20 *weight vectors, then test those* 20 *weight vectors on the test data, which is* 100 *seconds. We got* 20 *cepstral distances for each fixed adaptation length. The final results shown on the performance plots are the mean value of those* 20 *passes.*)

The performances of the Frost LMS beamformer with 20 taps for different adaptation lengths are illustrated in Fig.3.3 and Fig.3.4. Fig.3.3 shows the scenario when there is only one interference signal. The signal of interest (SOI, i.e. the desired signal) direction is 5°, while the interference signal comes from 45°. The look direction is the broadside direction (0°), which means a 5° look direction error. On the sub-figure (a) where the energy output is illustrated, the line with legend $S1$ ($X1$, DS, and LMS respectively) denotes the energy of the desired source signal (the first channel of the

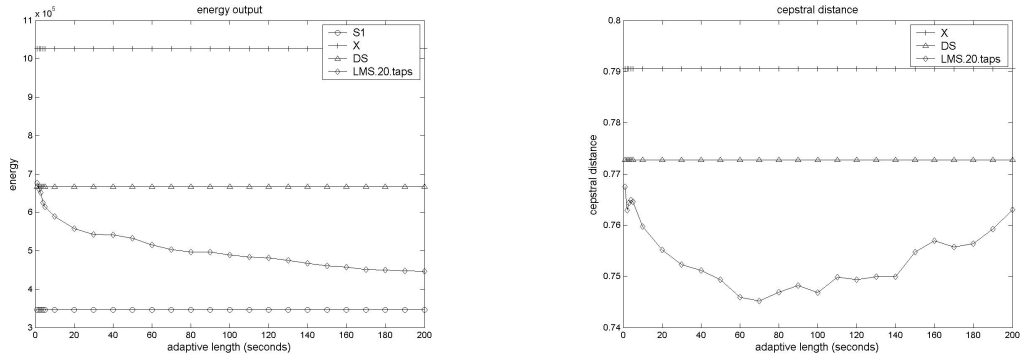


(a) Energy output (b) Cepstral distance

Figure 3.3: The performance of the Frost LMS beamformer vs. adaptation length, one interference signal

mixed signal, the delay-and-sum beamformer output, and the Frost LMS beamformer output respectively). Sub-figure (b) demonstrates the cepstral distance between the SOI and the first channel of the mixed signal (the delay-and-sum beamformer output, and the Frost LMS beamformer output respectively). From Fig.3.3 we can see that the best performance of Frost LMS beamformer is reached when the adaptation length is $10s$. In this case, the output energy matches that of the SOI and the cepstral distance reaches the minimum point. After $10s$, with the adaptation of the weight vector, the performance degrades to a greater extent. The reason for such degradation is that the desired signal begins to get suppressed with the on going adaptation (because of the existence of the look direction error). Fig.3.3 shows the scenario when there exist two interference signals. The observation is similar to that in the case of one interference signal.

The convergence analysis of Frost LMS algorithm can also be shown to be correct in the speech recognition experiment with the real recording MONC database [136]. When the Frost LMS beamforming algorithm is applied to every sentence, which is 1 – 5 seconds long, the speech recognition rate is 76.9. If we concatenate all our test sentences, which is about 1 hour of data, and apply the Frost LMS on that long data file, the speech recognition rate is much lower 57.6. It is apparent that the long-time Frost LMS is not so good as its short-time counterpart. Listening directly to the beamforming outputs also verifies the desired signal was suppressed more in the long data file experiment.



(a) Energy output

(b) Cepstral distance

Figure 3.4: The performance of the Frost LMS beamformer vs. adaptation length, two interference signals

3.3 Quadratic Constrained Robust Beamformer

The idea of Frost beamformer is to minimize output energy of the beamformer subject to a look direction constraint which restricts response of the beamformer to signal coming from the look direction to be always unity. The underlying assumption of Frost beamformer is the desired signal comes exactly from the look direction. Yet in the real world, generally there's a small discrepancy between desired signal's true incidence direction and its assumed direction.

We found in our experiments that a small look direction error can decrease the performance of Frost beamformer greatly. The reason is that Frost beamformer will take the desired signal as an interference signal and try to suppress it when there's a small look direction error.

A quadratic constraint broadband beamformer which is robust to DOA error is proposed by Er [137]. Consider a pre-steered Frost beamformer with K sensors and J taps (Fig.3.2). The weighted square error between desired and actual beam pattern of Frost beamformer over the interested frequency range and a small spatial region, chosen to deal with look direction uncertainty (Fig.3.1 illustrates an example of an uncertainty

region on a linear array), can be written as

$$e^2 = \int_{\bar{\theta}-\Delta\theta}^{\bar{\theta}+\Delta\theta} \int_{\omega} f(\theta) \left| \mathbf{w}^H V(\theta,\omega) - \mathbf{w}_d^H V(\theta,\omega) \right|^2 d\omega d\theta \tag{3.14a}$$

$$= (\mathbf{w} - \mathbf{w}_d)^H \int_{\bar{\theta}-\Delta\theta}^{\bar{\theta}+\Delta\theta} \int_{\omega} f(\theta) V(\theta,\omega) V(\theta,\omega)^H d\omega d\theta (\mathbf{w} - \mathbf{w}_d) \tag{3.14b}$$

$$= (\mathbf{w} - \mathbf{w}_d)^H \Phi (\mathbf{w} - \mathbf{w}_d), \tag{3.14c}$$

$$\text{with} \quad \Phi = \int_{\bar{\theta}-\Delta\theta}^{\bar{\theta}+\Delta\theta} \int_{\omega} f(\theta) V(\theta,\omega) V(\theta,\omega)^H d\omega d\theta. \tag{3.15}$$

$\bar{\theta}$ is the assumed look direction, $\Delta\theta$ is a measure of uncertainty in the assumed look direction, $f(\theta)$ is a spatial weighting function, ω is the frequency variable and $V(\theta,\omega)$ is the array steering vector. Φ is the positive semi-definite constraint matrix which can be calculated by either mathematical integration or by numerical techniques. $\mathbf{w}$ is the pursued beamformer's weight vector, and $\mathbf{w}_d$ is the desired beamformer's weight vector. Generally, Delay and Sum beamformer is used as the desired beamformer because of its robustness in the look direction.

The robust beamforming problem can be mathematically formulated as the following optimization problem

$$\min_{\mathbf{w}} \mathbf{w}^H R_{XX} \mathbf{w}, \quad \text{subject to} \quad (\mathbf{w} - \mathbf{w}_d)^H \Phi (\mathbf{w} - \mathbf{w}_d) \leq \varepsilon \tag{3.16}$$

where R_{XX} is the correlation matrix of the concatenated data vector X as in [19], and ε is a parameter chosen to control the tightness of the quadratic constraint. The weight vector $\mathbf{w}$ is a real vector, and Φ is a positive definite complex matrix. The constraint function $(\mathbf{w} - \mathbf{w}_d)^H \Phi (\mathbf{w} - \mathbf{w}_d)$ will always be a real number. Let $\Phi = \Phi_r + j\Phi_i$, then the constraint function $(\mathbf{w}-\mathbf{w}_d)^H \Phi (\mathbf{w}-\mathbf{w}_d) = (\mathbf{w}-\mathbf{w}_d)^T \Phi_r (\mathbf{w}-\mathbf{w}_d)$. Hence we always replace Φ with its real part Φ_r and assume no difference between Φ and Φ_r. Let $\mathbf{w}_e = \mathbf{w} - \mathbf{w}_d$. Problem (3.16) can be written as

$$\min_{\mathbf{w}_e} (\mathbf{w}_d + \mathbf{w}_e)^T R_{XX} (\mathbf{w}_d + \mathbf{w}_e), \quad \text{subject to} \quad \mathbf{w}_e^T \Phi \mathbf{w}_e \leq \varepsilon \tag{3.17}$$

Two methods will be developed to solve this quadratic constraint beamforming problem.

3.3.1 Lagrangian Multiplier Method

The Lagrange function is defined as,

$$L(\lambda) = (\mathbf{w}_d + \mathbf{w}_e)^T R_{XX}(\mathbf{w}_d + \mathbf{w}_e) + \lambda(\mathbf{w}_e^T \Phi \mathbf{w}_e - \varepsilon) \tag{3.18}$$

Take partial derivative to $\mathbf{w}_e$, the optimal solution is derived,

$$\mathbf{w}_e = -(R_{XX} + \lambda\Phi)^{-1} R_{XX}\mathbf{w}_d \tag{3.19}$$

The parameter λ can be obtained from the following equation,

$$h(\lambda) = \mathbf{w}_e^T \Phi \mathbf{w}_e = \mathbf{w}_d R_{XX}(R_{XX} + \lambda\Phi)^{-1}\Phi(R_{XX} + \lambda\Phi)^{-1} R_{XX}\mathbf{w}_d = \varepsilon \tag{3.20}$$

Using eigenvalue decomposition, suppose,

$$V\Gamma V^T = R_{XX}^{-1/2}\Phi R_{XX}^{-1/2} \tag{3.21}$$

Equation (3.20) can be written,

$$h(\lambda) = \mathbf{z}^T(I + \lambda\Gamma)^{-1}\Gamma(I + \lambda\Gamma)^{-1}\mathbf{z} = \varepsilon \tag{3.22}$$

and

$$h(\lambda) = \sum_{i=1}^{KJ} \frac{|z_i|^2 \gamma_i}{(1 + \lambda\gamma_i)^2} = \varepsilon \tag{3.23}$$

where γ_i is the diagonal elements of matrix Γ, and

$$\mathbf{z} = V^T R_{XX}^{1/2}\mathbf{w}_d \tag{3.24}$$

Note that $h(\lambda)$ is a monotonically decreasing function of λ, and $h(\lambda \to 0) > \varepsilon$, $h(\lambda \to \infty) < \varepsilon$. Hence there is a unique solution to equation (3.23), which can be solved efficiently via Newton's method.

3.3.2 Iterative Algorithm

We develop an iterative algorithm using the approach discussed in [145]. In [145], an iterative solution (also adaptive algorithm) was developed to solve the following constrained optimization problem:

$$\min_{\mathbf{w}\in S}\left\{\frac{1}{2}\mathbf{w}^T R\mathbf{w} - \mathbf{w}^T b\right\}, \quad S = \{\mathbf{w} : \|\mathbf{w}\| \le \varepsilon\} \tag{3.25}$$

The iterative algorithm is given by

$$\mathbf{w}_{k+1} = P[(\mathbf{I} - \mu R)\mathbf{w}_k + \mu b], \tag{3.26}$$

where

$$P(\mathbf{w}) = \begin{cases} \mathbf{w}, & \text{if } \|\mathbf{w}\| \leq \varepsilon \\ \mathbf{w}\frac{\sqrt{\varepsilon}}{\|\mathbf{w}\|}, & \|\mathbf{w}\| > \varepsilon. \end{cases} \tag{3.27}$$

We now manipulate the robust beamforming problem (3.17) into a form compatible with (3.25), enabling development of the iterative algorithm. Using eigenvalue decomposition, the matrix Φ in equation (3.17) can be written as,

$$\Phi = U\Lambda U^T = (U\Lambda^{1/2})(U\Lambda^{1/2})^T \tag{3.28}$$

Let $\tilde{\mathbf{w}}_e = (U\Lambda^{1/2})^T\mathbf{w}_e$. After some manipulation, equation (3.17) can be written as

$$\min_{\tilde{\mathbf{w}}_e}\{\frac{1}{2}\tilde{\mathbf{w}}_e^T\tilde{R}\tilde{\mathbf{w}}_e - \tilde{\mathbf{w}}_e^T\tilde{b}\}, \quad \text{subject to} \quad \tilde{\mathbf{w}}_e^T\tilde{\mathbf{w}}_e \leq \varepsilon \tag{3.29}$$

where $\tilde{R} = \Lambda^{-1/2}U^T R_{XX} U\Lambda^{-1/2}$, and $\tilde{b} = -\Lambda^{-1/2}U^T R_{XX}\mathbf{w}_d$. Now the iterative algorithm (3.26) can be applied to solve this problem [145]. The corresponding adaptive algorithm is obtained by substituting the instantaneous estimates $\hat{R}_k$ and $\hat{b}_k$ into (3.26), where $\hat{R}_k = \Lambda^{-1/2}U^T X_k X_k^T U\Lambda^{-1/2}$ and $\hat{b} = -\Lambda^{-1/2}U^T X_k X_k^T \mathbf{w}_d$, respectively.

3.4 Diagonally Loaded Robust Beamformer

Since the dimension of matrix Φ is large and Φ is obtained by an integration of matrices based on steering vector over a small region, it is generally a low rank matrix and is close to singular. This leads to numerical problems and hence unreliable weights when solving (3.17). Specifically, for the iterative algorithm, the eigenvalue decomposition in equation (3.28) won't be stable because Φ is ill-conditioned. One option is to approximate matrix Φ by a low rank matrix, i.e.

$$\Phi = \begin{bmatrix} U_1; U_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & \mathbf{0} \\ \mathbf{0} & \Lambda_2 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} \tag{3.30a}$$

$$\approx U_1\Lambda_1 U_1^T = (U_1\Lambda_1^{1/2})(U_1\Lambda_1^{1/2})^T \tag{3.30b}$$

where Λ_2 is small compared with Λ_1. Define $\tilde{\mathbf{w}}_1 = \Lambda_1^{1/2} U_1^T \mathbf{w}_e$, $\tilde{\mathbf{w}}_2 = U_2^T \mathbf{w}_e$, $\tilde{\mathbf{w}} = [\tilde{\mathbf{w}}_1^T, \tilde{\mathbf{w}}_2^T]^T$. Then

$$\mathbf{w}_e = U_1 \Lambda_1^{-1/2} \tilde{\mathbf{w}}_1 + U_2 \tilde{\mathbf{w}}_2 \tag{3.31}$$

The original problem (equation (3.17)) can be transformed into the following format,

$$\min_{\tilde{\mathbf{w}}} \{\frac{1}{2} \tilde{\mathbf{w}}^T \tilde{R} \tilde{\mathbf{w}} - \tilde{\mathbf{w}}^T \tilde{b}\}, \quad \text{subject to} \quad \tilde{\mathbf{w}}_1^T \tilde{\mathbf{w}}_1 \leq \varepsilon \tag{3.32}$$

To make the constraints $\mathbf{w}_e^T \Phi \mathbf{w}_e \leq \varepsilon$ and $\tilde{\mathbf{w}}_1^T \tilde{\mathbf{w}}_1 \leq \varepsilon$ be equal, the contribution of $\tilde{\mathbf{w}}_2$ is ignored and $U_2 \tilde{\mathbf{w}}_2$ is set to $\vec{0}$, thus $\mathbf{w}_e^T \Phi \mathbf{w}_e = \tilde{\mathbf{w}}_1^T \tilde{\mathbf{w}}_1$, and $\mathbf{w}_e = U_1 \Lambda_1^{-1/2} \tilde{\mathbf{w}}_1$. Substitute this back into equation (3.17), the obtained problem is only related to $\tilde{\mathbf{w}}_1$. Now Krieger's iterative method can be applied to this problem. Because too much freedom related to $\tilde{\mathbf{w}}_2$ is sacrificed, this method may suffer from insufficient freedom when many interferences are to be suppressed. This is confirmed by our observations in the experiments.

In [137], Er proposed a method wherein the original quadratic constraint is replaced by a set of linear constraints. Assume matrix Φ is a low rank matrix which can be represented by its eigen-decomposition $\Phi = U_1 \Lambda_1 U_1^T$. The original quadratic constraint in equation (3.17) can be approximated by a set of linear constraints $U_1^T \mathbf{w}_e = \mathbf{0}$. The linear constraint algorithm is an approximation because the quadratic constraint $\mathbf{w}_e^T \Phi \mathbf{w}_e \leq \varepsilon$ is not equivalent to the linear constraints $U_1^T \mathbf{w}_e = \mathbf{0}$. Often matrix Φ is only approximated by $U_1 \Lambda_1 U_1^T$ but is not exactly equal to it. In general, $\Phi = U_1 \Lambda_1 U_1^T + U_2 \Lambda_2 U_2^T$. Although the norm of Λ_2 is small, if the norm of $\mathbf{w}$ is big enough, $\mathbf{w}_e^T U_2 \Lambda_2 U_2^T \mathbf{w}_e$ may exceed ε. In such cases, the original quadratic constraint is not satisfied.

In [137], an extra norm constraint on $\mathbf{w}$ was proposed to ensure that the quadratic constraint and the set of linear constraints are equivalent. However, it complicates the optimization problem and was not pursued in detail. The other disadvantage of the algorithm is that to ensure better approximation to matrix Φ, the dimension of Λ_1 has to be increased, which increases the number of linear constraints. Consequently, the degrees of freedom of the weight vector are reduced compromising the ability to suppress interferences.

In view of the low rank property of the matrix Φ, a method which can robustly solve equation (3.17) is now proposed. The idea is to add a diagonal loading to Φ to

restore it to a full rank matrix, i.e. construct a new matrix Φ' such that $\Phi' = \Phi + \lambda I$. The diagonal loading method here is close to the diagonal loading method in [21] except that the diagonal loading is added to the constraint matrix Φ instead of the signal covariance matrix R_{XX}. Substituting Φ' for Φ in (3.17), the problem can be rewritten as

$$\min_{\mathbf{w}_e}(\mathbf{w}_d + \mathbf{w}_e)^T R_{XX}(\mathbf{w}_d + \mathbf{w}_e), \quad \text{s.t.} \quad \mathbf{w}_e^T \Phi' \mathbf{w}_e \leq \varepsilon \tag{3.33}$$

where $\Phi' = \Phi + \lambda I$. The above optimization problem can be solved robustly by the Lagrangian method in section 3.3.1 or by the iterative algorithm in section 3.3.2. Since Φ is totally determined by the array geometry, it can be calculated beforehand as well as the DL level λ. It can be chosen optimally offline with respect to Φ. In contrast, in diagonal loading of the signal covariance matrix R_{XX}, the DL level has to be adjusted online with respect to different values of R_{XX}. Furthermore, the new well-conditioned constraint matrix Φ' ensures a robust eigenvalue decomposition such that the iterative and corresponding adaptive algorithm are stable, while for the method of adding DL to signal covariance matrix no such adaptive algorithm is available.

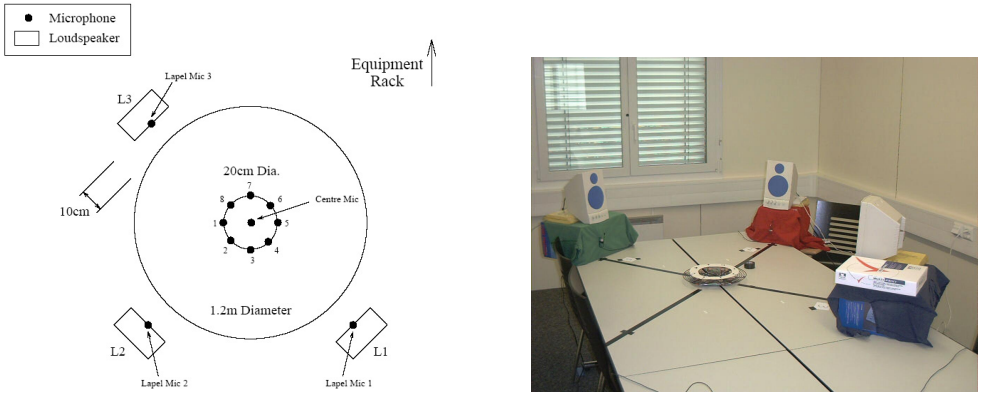
To get more insight into problem (3.33), we expand the constraint

$$\mathbf{w}_e^T \Phi' \mathbf{w}_e = \mathbf{w}_e^T \Phi \mathbf{w}_e + \lambda \mathbf{w}_e^T \mathbf{w}_e. \tag{3.34}$$

Define $\mathbf{w}_e^T \Phi \mathbf{w}_e = \varepsilon_1$ and $\lambda \mathbf{w}_e^T \mathbf{w}_e = \varepsilon_2$, to ensure the quadratic constraint is satisfied, it must be true $\varepsilon_1 + \varepsilon_2 \leq \varepsilon$. In other words, from the constraint $\mathbf{w}_e^T \Phi' \mathbf{w}_e \leq \varepsilon$, two constraints $\mathbf{w}_e^T \Phi \mathbf{w}_e \leq \varepsilon$ and $\lambda \mathbf{w}_e^T \mathbf{w}_e \leq \varepsilon$ can be met. Consequently, if constraint $\mathbf{w}_e^T \Phi' \mathbf{w}_e \leq \varepsilon$ is satisfied, not only is the constraint in our original optimization problem (3.17) satisfied, but also a new norm constraint on the weight vector $\mathbf{w}_e$ is introduced. The norm constraint will introduce further robustness as shown in [26]. And λ can be considered as a parameter which leverages constraint $\mathbf{w}_e^T \Phi \mathbf{w}_e = \varepsilon_1$ and the norm constraint. In conclusion, the diagonal loading to the matrix Φ facilitates robustness in the solution of the original optimization problem (3.17).

3.5 Experiments

A speech recognition experiment is conducted on the Multichannel Overlapping Numbers Corpus (MONC) to test the performances of various beamforming algo-



(a) Loudspeaker and microphone placement

(b) Experimental setup

Figure 3.5: MONC database experiment scenario

rithms. MONC is a multichannel speech database recorded in a moderately reverberant 8.2m*3.6m*2.4m rectangular room [136]. The recording scenario was designed to simulate three speakers seated around a 1.2m diameter circular meeting room table. The loudspeakers (L1, L2, L3) were placed at 90° spacings at an elevation of 35cm. An eight-element, 20cm diameter, circular microphone array was placed in the centre of the table. An additional microphone was placed at the centre of the array. The loudspeaker and microphone placement is illustrated in Fig.3.5(a). A photograph of the experimental setup is included in Fig.3.5(b). The sampling rate for recording is 8kHz.

In applying the beamforming algorithms to the data from the MONC database, the desired source is speaker L1, which is assumed to be a point source coming from the location: angle 180°, radius 70cm, height 35cm. We define the origin of the coordinates to be the center of the circular microphone array, and define angle 0° to be the direction of the 8_{th} microphone. The angle increases counter clockwise, which means the 1_{st} microphone is in the direction of angle 45°. Only microphones 1 to 8 are used in the beamforming.

Speech recognition results using different beamforming algorithms are shown in Fig.3.6. Five beamforming algorithms and one single microphone based approach are compared. CentreMic means using single centre microphone's signal for speech recognition, without beamforming. DS, Frost, robFrost_orig, robFrost_Er and robFrost_DL

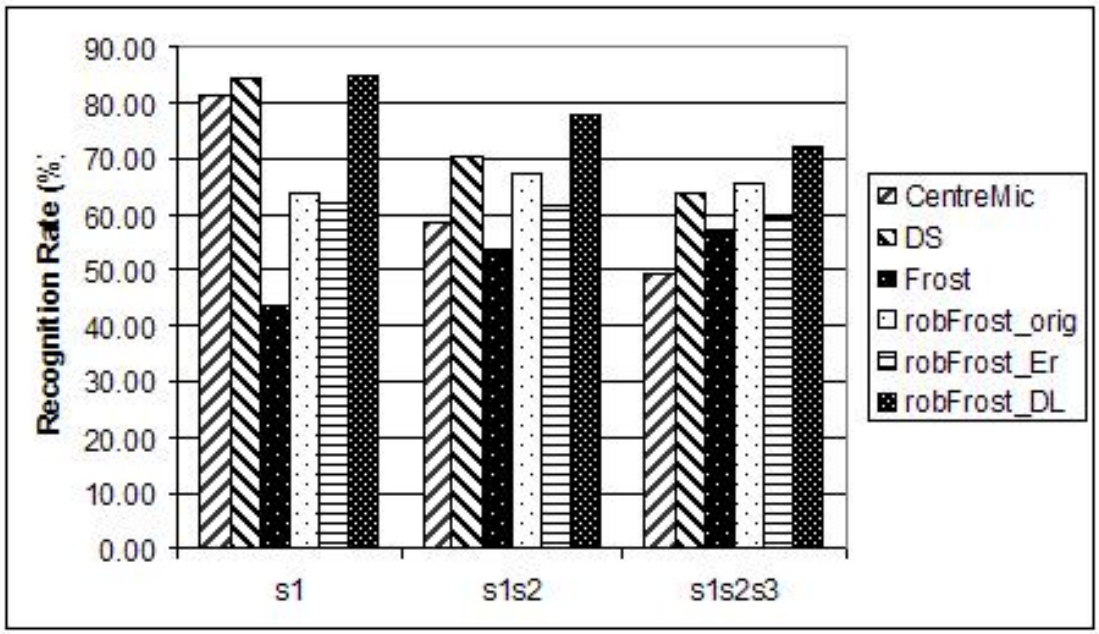


Figure 3.6: Speech recognition rate

represent conventional Delay-and-Sum beamforming, optimal Frost beamforming [19], robust Frost beamforming with quadratic constraint (using original constraint matrix, solved by Lagrangian method (3.19)), Er's robust Frost beamforming with a set of linear constraints [137] and robust Frost beamforming with quadratic constraint (using diagonally loaded constraint matrix) respectively. For the robFrost_DL algorithm, both the Lagrangian multiplier method and the iterative method yield identical experimental results, hence only the results obtained by the Lagrangian method are shown in Fig.3.6. 20 taps is used for the broadband beamforming algorithms and the tap delay between each two taps of the Frost structure is the sampling interval, i.e. $0.125ms$. The constraint matrix of the robust Frost beamforming is calculated through integration over the whole frequency band 0–4000Hz and the uncertainty region: angle $180 \pm 3°$, height $35 \pm 5cm$, and radius $70 \pm 3cm$. s1, s1s2 and s1s2s3 represent three different scenarios separately: only speaker s1 is speaking, both s1&s2 are speaking, and all three speakers are speaking. For the proposed algorithm, the DL level λ is chosen to be 10^{-5} times largest eigenvalue of Φ. From the experiments results, it is evident that the proposed beamforming algorithm is significant better than the other algrithms.

We also plot the beam pattern of different beamformers using one multichannel sample sentence from the MONC database. Only speakers L1 and L2 are present in the sample sentence. Fig.3.7 shows the beam pattern of various beamformers over angle θ and frequency. It is apparent that the robust Frost beamforming with diagonal loaded constraint matrix has the best combination of robustness in the look direction (180°) and good suppression in the interference direction (90°).

Acknowledgement

The text of this chapter, in part, is a reprint of the material which has appeared in Wenyi Zhang, and Bhaskar D. Rao, "Robust Broadband Beamformer With Diagonally Loaded Constraint Matrix and Its Application to Speech Recognition", *Acoustics, Speech, and Signal Processing, 2006. ICASSP 06. IEEE International Conference on*, pp. 785-788, May 2006.

3.6 Appendix for Chapter 3

3.6.1 Eigenvalue Relationship

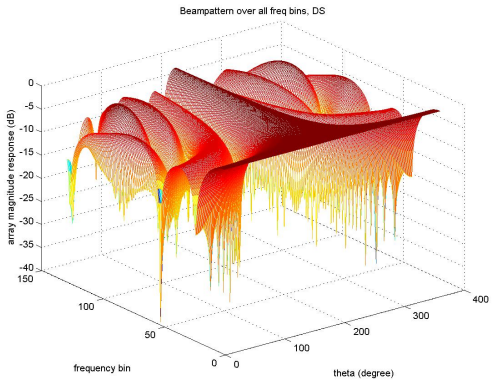
Suppose that the matrix P's eigenvalue decomposition is $U_P\Lambda_P U_P^T$, and R_{XX}'s eigenvalue decomposition is $U_R\Lambda_R U_R^T$. Then we get,

$$PR_{XX}P = (U_P\Lambda_P U_P^T)(U_R\Lambda_R U_R^T)(U_P\Lambda_P U_P^T) \tag{3.35}$$

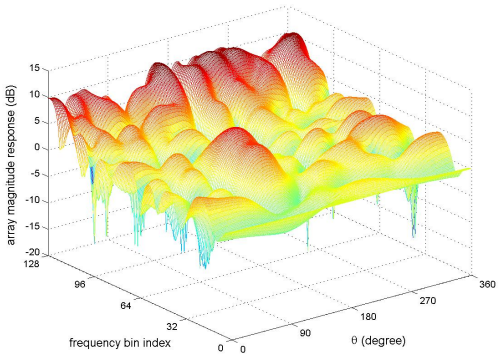
Since U_P and U_R are unitary, $U_P^T U_R$ is unitary. Let $M = U_P^T U_R \Lambda_R U_R^T U_P$. M will have the same eigenvalues as Λ_R.

Since $C(C^TC)^{-1}C^T$ has K eigenvalues equal to 1 (the corresponding eigenvectors are the column vectors of the constraint matrix C) and $KJ - J$ eigenvalues equal to 0, Λ_P is a diagonal matrix with $KJ-J$ ones and J zeros on the diagonal. Therefore $PR_{XX}P$ has J zero eigenvalues whose corresponding eigenvectors are the column vectors of the constraint matrix C. From $\Lambda_P M\Lambda_P$, we just pick the first principle sub-matrix of matrix M. By Cauchy's interlacing theorem, $\Lambda_P M\Lambda_P$ has $KJ-J$ nonzero eigenvalues σ_i whose values are bounded between the smallest and largest eigenvalues (λ_{min} and λ_{max}) of R_{XX},

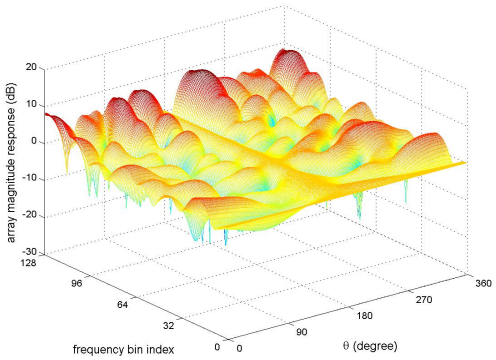
$$\lambda_{min} \le \sigma_{min} \le \sigma_{max} \le \lambda_{max} \tag{3.36}$$



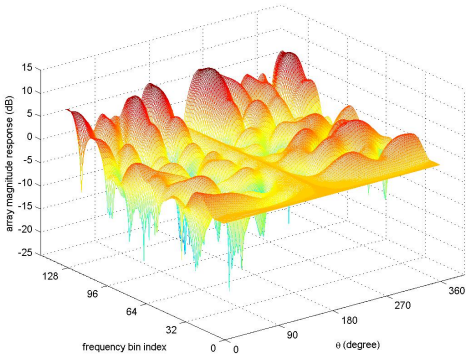
(a) DS beamformer



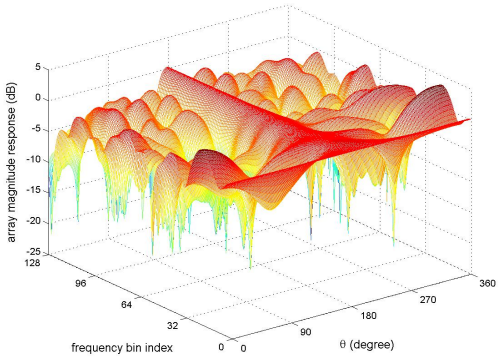
(a) Frost beamformer



(b) robust Frost beamformer (Eq. (3.19))



(c) Er's robust Frost beamformer [137]



(d) robust Frost BF with DL constraint matrix

Figure 3.7: Beam pattern of various beamformers over angle θ and frequency bins. The radius is fixed at 70cm and height fixed at 35cm. The look direction is 180° and the interference direction is 90°

4 Robust Adaptive Narrowband Beamforming Algorithms

Data dependent adaptive beamformer has high resolution and interference rejection capability when the array steering vector is accurately known. However, it is known to degrade severely if steering vector error exists. Motivated by recent work in robust MVDR beamforming (RCB&DCRCB) [26], we develop variants of the constrained robust MVDR beamformer that attempts to limit the search in the underlying optimization problem to a feasible set of steering vectors, thereby achieving improved performance. The robustness against steering vector error is provided through a spherical uncertainty set constraint, while a set of magnitude constraints are enforced on each element of the steering vector to constrain the search in the space of feasible steering vectors in a better fashion. By appropriately changing the variables, the optimization problem is modified such that the need for the magnitude constraints are avoided. We call this newly developed approach as time delay based robust MVDR beamforming.

We also develop adaptive algorithms for the RCB and the time delay based robust MVDR beamformer. The adaptive algorithms have two updating steps. The first step updates the steering vector estimation or the time delay estimation; the second step updates the beamformer's weight vector given an estimated steering vector. The developed algorithms are tested in the context of speech enhancement using a microphone array.

4.1 Introduction to Robust Narrowband Beamforming

Beamforming is widely used in array processing to enhance the desired signal while suppressing interference signals and noise. The standard data independent beamformers such as the delay-and-sum (DS) beamformer is robust to the signal of interest (SOI) steering vector errors, which may be due to look direction error, array sensor position error and small mismatches in the sensor responses. However, those beamformers suffer from low resolution and high sidelobes, inducing degraded interference rejection capability. In contrast, data dependent adaptive beamformers such as the standard MPDR beamformer, has high resolution and interference rejection capability when the array steering vector is accurately known [20]. However, the performance of traditional adaptive beamformer can degrade severely in practice when there exist SOI steering vector errors. In such cases, the SOI might be mistaken as an interference signal and be suppressed.

Many robust beamforming algorithms have been proposed to address this problem [21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39]. Derivative constraint in the look direction is proposed in [22, 23]. Er and Cantoni proposed a robust beamforming algorithm which restricts the error between the desired and actual beam pattern of the array over a small spatial region around the array's look direction, allowing for uncertainty in the look direction [40]. Linearly constrained minimum variance (LCMV) beamforming is proposed in [41, 42]. Bell proposed a Bayesian approach to robust adaptive beamforming in [24]. Norm constrained and white noise gain constrained adaptive beamformer are studied in [21, 43, 44] and widely used thereafter. These beamformers use diagonal loading to improve robustness, however, how the diagonal loading should be selected is not well justified.

Recently some new robust adaptive beamformers with theoretical background have been proposed. Robust adaptive beamforming using worst-case performance optimization is proposed in [25, 45, 46]. The problem is formulated as minimizing a quadratic function subject to infinitely many quadratic constraints, but is reduced to a second-order cone programming problem which can be solved by interior point methods. Li and Stoica proposed the robust Capon beamformer (RCB) [47, 26, 48] where a spherical uncertainty set constraint is enforced on the array steering vector. They

also developed a doubly constrained robust Capon beamformer (DCRCB) [26] based on RCB, wherein a norm constraint on the beamformer steering vector is added. A comparison of these two beamformers is given in [49] and a geometrical explanation is provided.

RCB and DCRCB solve the SOI cancellation problem by searching for the MVDR beamformer with maximum energy output in a spherical uncertainty set. If the true SOI steering vector exists in the spherical uncertainty set, RCB and DCRCB is supposed to find that steering vector as the minimizer of a quadratic optimization problem. However, the minimizer of the quadratic optimization problem may not be a feasible steering vector. A feasible steering vector generally satisfies some geometrical constraints and has a physical meaning. Not all the vectors in the spherical uncertainty set have such physical meaning and are set to be valid steering vectors.

Motivated by the idea of searching for a valid steering vector, in this work, we develop variants of robust MVDR beamformer that attempt to limit the search in the underlying optimization problem to a feasible set of steering vectors thereby achieving improved performance. The robustness to the steering vector error is provided through a spherical uncertainty set constraint, while a set of magnitude constraints is enforced on each element of the steering vector to constrain the search to the space of feasible steering vectors in a better fashion. By appropriately changing the variables, the optimization problem is modified such that the need for the magnitude constraints are avoided.

We also develop adaptive algorithms for the RCB and the time delay based robust MVDR beamformer. The adaptive algorithms have two updating steps. The first step updates the steering vector estimation or the time delay estimation; the second step updates the beamformer's weight vector given an estimated steering vector. The developed algorithms are tested in the context of speech enhancement using a microphone array.

This chapter is organized as follows. Section 4.2 contains background material, it has a quick review of the RCB and DCRCB algorithms and bring forward the problem with RCB and DCRCB. Our proposed time delay based robust MVDR beamformer is developed in section 4.3. Section 4.4 develops the adaptive algorithms for the RCB and the time delay based robust MVDR beamformer. Simulations which compares the per-

formance of various beamformers are given in section 4.5. Finally, section 4.6 presents our conclusions.

The following notations are adopted throughout this chapter: $(.)^T$ denotes transpose, $(.)^*$ denotes conjugate, $(.)^H$ for conjugate transpose, $(.)^{-1}$ denotes inverse matrix, $(.)^{-T}$ denotes inverse and transpose, $\|.\|$ denotes L_2 norm, $|.|$ denotes modulus, $\nabla_{\mathbf{w}}(.)$ denotes gradient to variable $\mathbf{w}$. I is the identity matrix, the size depends on the context. Lower case alphabet denotes scalar, e.g. a; lower case boldface alphabet denotes vector, e.g. $\mathbf{a}$; and upper case alphabet denotes matrix, e.g. R.

4.2 Background

4.2.1 Standard MVDR Beamformer (MVDR)

The MVDR beamforming is also called Capon beamforming [20]. The problem is formulated as minimizing the output energy of the beamformer while maintaining a constant response in the look direction, i.e.

$$\min_{\mathbf{w}} \mathbf{w}^H R \mathbf{w}, \quad \text{s.t.} \quad \mathbf{w}^H \mathbf{a} = 1. \tag{4.1}$$

where R is the signal correlation matrix. $\mathbf{a}$ is the SOI steering vector. $\mathbf{w}$ is the beamformer weight vector. The solution to this optimization problem is given by

$$\mathbf{w} = \frac{R^{-1}\mathbf{a}}{\mathbf{a}^H R^{-1} \mathbf{a}}. \tag{4.2}$$

4.2.2 Robust Capon Beamformer (RCB)

The Robust Capon Beamformer (RCB) is proposed in [47]. Suppose $\mathbf{a}_0$ is the true SOI steering vector and $\overline{\mathbf{a}}$ is the assumed look direction steering vector. $\mathbf{a}_0$ is assumed to be in the vicinity of $\overline{\mathbf{a}}$. This can be expressed mathematically by the following inequality

$$\|\mathbf{a}_0 - \overline{\mathbf{a}}\|^2 \leq \varepsilon, \tag{4.3}$$

where ε is a bound controlling the uncertainty in the assumed look direction.

The Capon beamforming problem can be reformulated as

$$\max_{\sigma^2} \sigma^2, \quad \text{s.t.} \quad R - \sigma^2 \mathbf{a}\mathbf{a}^H \geq 0. \tag{4.4}$$

where R is the signal correlation matrix. σ^2 is the signal power to be estimated.

Use the new formulation, one can write the RCB problem as

$$\max_{\sigma^2, \mathbf{a}} \sigma^2, \quad \text{s.t.} \quad R - \sigma^2 \mathbf{a}\mathbf{a}^H \geq 0 \quad \text{and} \quad \|\mathbf{a} - \bar{\mathbf{a}}\|^2 \leq \varepsilon. \tag{4.5}$$

Using the fact that, for any fixed $\mathbf{a}$, the solution to (4.4) with regard to σ^2 is obtained by

$$\hat{\sigma}^2 = 1/(\mathbf{a}^H R^{-1} \mathbf{a}) \tag{4.6}$$

the optimization problem (4.5) can be written as

$$\min_{\mathbf{a}} \mathbf{a}^H R^{-1} \mathbf{a}, \quad \text{s.t.} \quad \|\mathbf{a} - \bar{\mathbf{a}}\|^2 \leq \varepsilon. \tag{4.7}$$

The solution can be found using Lagrange multiplier method as

$$\hat{\mathbf{a}}_0 = \bar{\mathbf{a}} - U(I + \lambda\Gamma)^{-1} U^H \bar{\mathbf{a}} \tag{4.8}$$

where $R = U\Gamma U^H$ is the eigenvalue decomposition of R, and λ is the Lagrange multiplier. Once the SOI steering vector is estimated, the signal power can be estimated as in (4.6) and the beamformer weight vector is easily obtained as in MVDR beamforming (4.2).

One difficulty with this approach is that it tends to overestimate the signal power σ^2, because both the SOI power and the SOI steering vector are taken as unknowns in problem (4.5). Thus, $(\sigma^2, \mathbf{a})$ and $(\sigma^2/\alpha, \alpha^{1/2}\mathbf{a}), \forall \alpha > 0$ will give the same item $\sigma^2 \mathbf{a}\mathbf{a}^H$. Suppose $(\sigma_0^2, \mathbf{a}_0)$ is the true solution to be found, the formulation of (4.5) will prefer the pair $(\sigma_0^2/\alpha, \alpha^{1/2}\mathbf{a}_0)$ if only $\alpha < 1$ and $\alpha^{1/2}\mathbf{a}_0$ is still in the uncertainty set. By the deduction above, we can be certain that the solution to (4.5) will make the inequality constraint in (4.5) active, i.e. $\|\hat{\mathbf{a}}_0 - \bar{\mathbf{a}}\|^2 = \varepsilon$. This problem is solved in [47] by a normalization step such that $\|\hat{\mathbf{a}}_0\|^2 = N$, where N is the number of sensor elements.

4.2.3 Doubly Constrained Robust Capon Beamformer (DCRCB)

To avoid the signal power overestimation problem discussed above in section 4.2.2, the doubly constrained robust Capon beamformer (DCRCB) is proposed [26].

The problem is formulated in a similar way as in (4.7) except that an extra norm constraint on the steering vector $\mathbf{a}$ is added.

The problem is formulated as

$$\min_{\mathbf{a}} \mathbf{a}^H R^{-1} \mathbf{a}, \quad \text{s.t.} \quad \|\mathbf{a} - \bar{\mathbf{a}}\|^2 \leq \varepsilon \quad \text{and} \quad \|\mathbf{a}\|^2 = N \tag{4.9}$$

The solution can be found using the Lagrange multiplier method

$$\hat{\mathbf{a}} = (N - \frac{\varepsilon}{2}) \frac{U(I + \lambda\Gamma)^{-1} U^H \bar{\mathbf{a}}}{\bar{\mathbf{a}}^H U(I + \lambda\Gamma)^{-1} U^H \bar{\mathbf{a}}} \tag{4.10}$$

where $R = U\Gamma U^H$ is the eigenvalue decomposition of R, and λ is the Lagrange multiplier. Then the beamformer weight vector is easily obtained as in MVDR beamforming (4.2).

In both RCB and DCRCB, the bound ε is chosen such that all possible SOI steering vectors $\mathbf{a}_0$ are included in the uncertainty set described by (4.3). Also, it is made clear in [26] that the choice of ε should be made as small as possible since when ε is too large, the ability to suppress interferences degrades.

4.2.4 Problems with RCB and DCRCB

RCB and DCRCB solve the SOI cancellation problem by searching for the MVDR beamformer with maximum energy output in a spherical uncertainty set. If the true SOI steering vector exists in the spherical uncertainty set, RCB and DCRCB are supposed to find that steering vector as the minimizer of the quadratic optimization problem (4.7) or (4.9). However, the RCB (section 4.2.2) and DCRCB (section 4.2.3) beamforming algorithms may fail because the optimum solution $\hat{\mathbf{a}}$ to the optimization problem (4.7) or (4.9) may not be a valid steering vector. A valid steering vector is usually structured and is not any element in the constrained set (4.3). It usually satisfies some constraints and has a physical meaning. Not all the vectors in the spherical uncertainty set (4.3) have such physical meaning and are valid steering vectors.

The following simple simulations illustrate the problems with RCB and DCRCB. We assume a uniform linear array with 8 sensors and half-wavelength sensor spacing, and there is spatially white Gaussian noise which keep the SNR at 30dB.

The first example illustrates how RCB and DCRCB will be affected by the uncertainty bound ε. Suppose there are two uncorrelated source signals: one SOI comes from 0.5°; one interference signal comes from 45°, both signals have power level 1. The assumed look direction is 0°. Fig.4.1 and Fig.4.2 demonstrates the uncertainty bound ε and the power estimate as a function of the uncertainty angle range θ. A larger uncertainty angle range corresponds to a larger uncertainty in the look direction. With the increase of the uncertainty angle range θ, the uncertainty bound ε increases, and the power estimate of RCB and DCRCB increases as well. Since the SOI has a power level of 1, the overestimation of power is evident for RCB and DCRCB, and the overestimation increases with the increases of uncertainty bound. It is demonstrated that RCB has better power estimation than DCRCB under large uncertainty bound, this is consistent with the analysis and simulations shown in [26]. A further inspection of the estimated steering vector by (4.8) and (4.10) shows that they are far from a valid steering vector when the power overestimation is severe. The magnitude of each element of the estimated steering vector is far from 1, while for a valid steering vector, the magnitude of each element should always be 1. We will revisit this point in the following section.

The second example will illustrate that the power overestimation is not only dependent on the uncertainty bound but also affected by the interference signals. The experiment scenario is the same as above except that we add a second interference signal which is injected into the array from $-60°$. The second interference signal also has a power level 1. Fig.4.3 shows the power estimate as a function of the uncertainty angle range θ. Compared to Fig.4.2, the power overestimation for RCB and DCRCB is more severe in this case.

The power overestimation problem is not significant when the uncertainty bound ε is small. However, the problem becomes more serious with the increase of the uncertainty bound. In a real application, there are many cases that have a large uncertainty bound. For example, when using a microphone array for speech processing, given an assumed look direction, it is reasonable to assume a look direction error of 5° to 10°.

There are two ways to improve the performance of RCB and DCRCB under such an environment. The first approach is to use a set of RCBs (or DCRCBs) with small uncertainty bound ε to cover a large uncertainty range. Suppose the true SOI steering

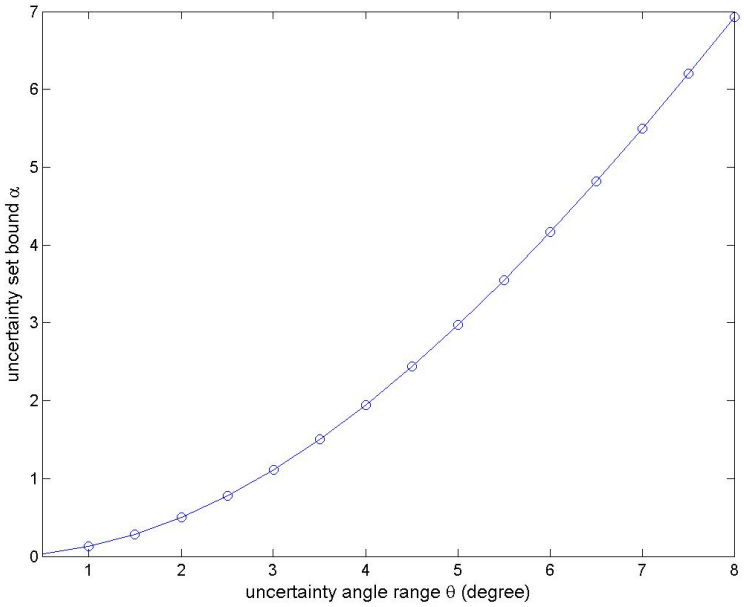


Figure 4.1: Uncertainty bound ε versus uncertainty angle range.

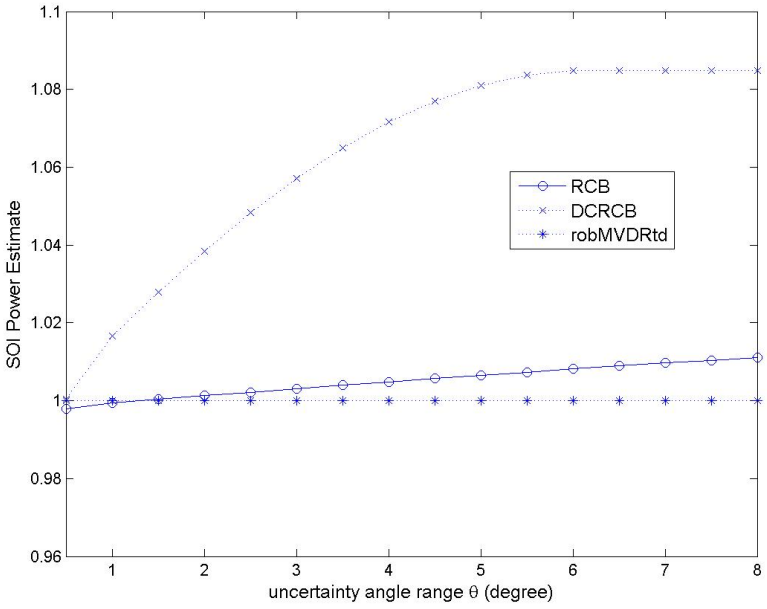


Figure 4.2: Power estimate versus uncertainty angle range, one interference signal.

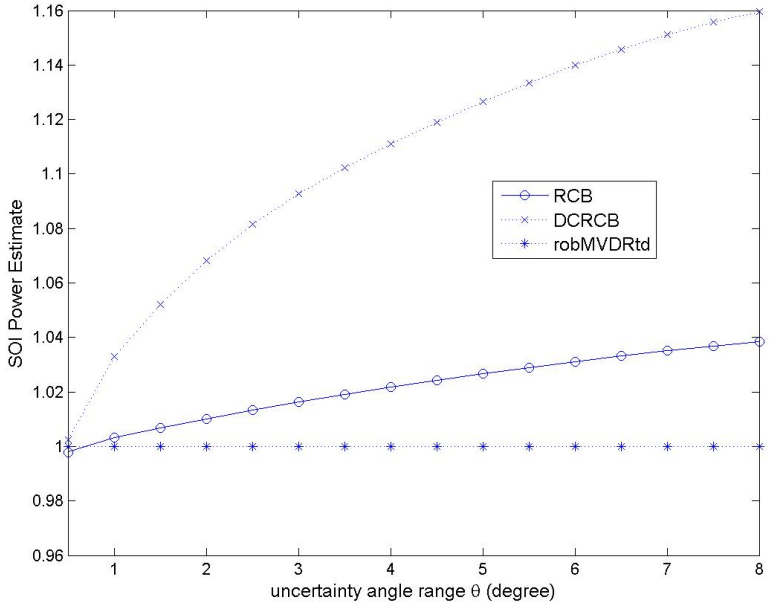


Figure 4.3: Power estimate versus uncertainty angle range, two interference signals.

vector falls into the uncertainty set of one of those RCBs (or DCRCBs). Such specific RCB (or DCRCB) will have a good estimation of the SOI's power and a significant power output. Instead, other RCBs (or DCRCBs) which do not include the true SOI steering vector in their uncertainty sets will have none or small power output. Thereby we can choose the maximum power estimate from that set of RCBs (or DCRCBs) as the final power estimate for the SOI. Since each of those RCBs (or DCRCBs) has a small uncertainty bound ε, the power estimate will be more accurate than that of a single RCB (or DCRCB) with a big uncertainty range. One shortcoming of this approach is that to cover a large uncertainty range in a real application, we may need a large number of RCBs (or DCRCBs). What's more, the power overestimation problem will always exist and we can never claim that we estimate the true steering vector.

The second approach to improve the performance of RCB and DCRCB under a large uncertainty bound is to add extra constraints so that the solution is a valid steering vector. By constraining the estimated steering vector to be a valid steering vector, the robust MVDR beamformer is assured to attain the true SOI steering vector and obtain an accurate power estimation. We will propose some methods using such constraints in

the following section.

4.3 Magnitude Constrained Robust MVDR Beamformer

A valid steering vector is usually structured and is not any element in the constrained set (4.3). We develop variants of the constrained robust adaptive beamformer that attempt to limit the search in the underlying optimization problem to a feasible set of steering vectors thereby achieving improved performance. For an array with identical omnidirectional sensors, a valid steering vector $\mathbf{a}$ can be expressed as $\mathbf{a} = [e^{-j\omega\tau_1}, e^{-j\omega\tau_2}, ..., e^{-j\omega\tau_N}]^T$ for the far field sources. We observe that each element of the the steering vector $\mathbf{a}$ has magnitude 1. Thereby we propose to enforce a set of magnitude constraints on each element of the steering vector $\mathbf{a}$ based on RCB (4.7) thereby making the search space smaller and more feasible. The new optimization problem can be formulated as

$$\min_{\mathbf{a}} \mathbf{a}^H R^{-1} \mathbf{a}, \quad \text{s.t.} \quad \|\mathbf{a} - \overline{\mathbf{a}}\|^2 \leq \varepsilon \quad \text{and} \quad |a_k| = 1, k = 1..N \tag{4.11}$$

where a_k is the k_{th} element of the steering vector $\mathbf{a}$, i.e. $\mathbf{a} = [a_1, a_2, ..., a_N]^T$. Unfortunately, a closed form solution to this optimization problem is not available and an optimization routine has to be utilized.

4.3.1 Time Delay Based Robust MVDR Beamformer (robMVDRtd)

By manipulating the variables, we can create a similar robust beamforming problem to problem (4.11). In particular, we use the form of the steering vector $\mathbf{a}_i$ for a specific frequency ω_i as

$$\mathbf{a}_i = [e^{-j\omega_i\tau_1}, e^{-j\omega_i\tau_2}, ..., e^{-j\omega_i\tau_N}]^T \tag{4.12}$$

As $|e^{j\omega_i\tau_k}| \equiv 1$, optimizing over the time delay variables τ_i ensures the magnitude constraint in (4.11) is automatically satisfied and thus can be omitted. The new robust beamforming problem is formulated as

$$\min_{\tau} \mathbf{a}_i^H R_i^{-1} \mathbf{a}_i, \quad \text{s.t.} \quad |\tau_k - \overline{\tau}_k| \leq \delta_k, \quad k = 1..N \tag{4.13}$$

where R_i is the signal correlation matrix for frequency ω_i. $\tau = [\tau_1, \tau_2, ..., \tau_N]^T$, and $\overline{\tau} = [\overline{\tau}_1, \overline{\tau}_2, ..., \overline{\tau}_N]^T$ is the assumed look direction time delay vector. $\delta_k, k = 1..N$ is a set of bounds controlling the uncertainty in the look direction. The new problem (4.13) also can only be solved by using an appropriate optimization routine.

We use a subspace trust region method which is based on interior-reflective Newton algorithm to find the solution to problem (4.13). We need the gradient and Hessian of the objective function $h_i(\tau) = \mathbf{a}_i^H R_i^{-1} \mathbf{a}_i$, where $\mathbf{a}_i$ is specified by (4.12). It is straightforward to obtain gradient as

$$\nabla_\tau h_i = A R_i^{-T} \mathbf{a}_i^* + A^* R_i^{-1} \mathbf{a}_i \tag{4.14a}$$

$$= real(A^* R_i^{-1} \mathbf{a}_i) \tag{4.14b}$$

where

$$A = (-j\omega_i) \begin{bmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_N \end{bmatrix} \tag{4.15}$$

where a_k is the k_{th} element of the steering vector $\mathbf{a}_i$, i.e. $\mathbf{a}_i = [a_1, a_2, ..., a_N]^T$.

Also, the Hessian is obtained as

$$\nabla_\tau^2 h_i = real((-j\omega_i) diag(\mathbf{a}_i^H R_i^{-1}) A + A R_i^{-T} A^*) \tag{4.16}$$

A simple simulation shows that this time delay based robust MVDR beamformer can estimate the SOI steering vector and power accurately. The results are illustrated in Fig.4.2 and Fig.4.3. This beamformer is denoted as **robMVDRtd** in the plots. All the experiment scenarios are aforementioned.

In the context of broadband signals, for each frequency component of the signal one has to solve a problem like (4.13) based on the specific frequency ω_i. However, the objective minimizer $\hat{\tau}$ is the true time delay from the SOI to each microphone element, which doesn't depend on the frequency ω_i. In other words, we want to find a common minimizer to all the optimization problems (4.13) on various frequencies. This is not automatic and has to be enforced. It can be achieved by combining the series of beamforming problems on individual frequency bins into a single problem to provide

robustness. The broadband beamforming problem can be formulated as

$$\min_{\tau} \sum_{i} \mathbf{a}_i^H R_i^{-1} \mathbf{a}_i, \quad \text{s.t.} \quad \|\tau - \overline{\tau}\|^2 \leq \delta \tag{4.17}$$

4.3.2 Angle Based Robust MVDR Beamformer (robMVDRangle)

The RCB (section 4.2.2), DCRCB (section 4.2.3) and robMVDRtd (section 4.3.1) algorithms assume uncertainty in the steering vector, which takes both the SOI look direction error and the array sensor's position error into consideration. The problem can be simplified when only SOI look direction error exists. For instance, in the case of 2-dimensional space the sources' incidence directions can be represented by only one parameter θ. Hence, we can use $\mathbf{v}(\theta)$ to substitute for the steering vector $\mathbf{a}$ in (4.13). The new robust beamforming problem can be written as

$$\min_{\theta} \mathbf{v}(\theta)^H R^{-1} \mathbf{v}(\theta), \quad \text{s.t.} \quad |\theta - \overline{\theta}| \leq \varepsilon \tag{4.18}$$

where $\mathbf{v}(\theta) = [e^{-j\omega\tau_1}, e^{-j\omega\tau_2}, ..., e^{-j\omega\tau_N}]^T$, and $\tau_i, i = 1, .., N$ is functions of θ based on the geometry of the array. $\overline{\theta}$ is the assumed look direction. ε is a bound controlling the uncertainty in the assumed look direction. The problem (4.18) can be solved by one dimensional numerical optimization algorithm such as the golden section search method.

4.4 Adaptive Robust Beamforming

The beamforming algorithms we have discussed in the above sections are off-line algorithms. In other words, they are batch algorithms. This means that to run the algorithms, we need a batch of data to estimate the correlation matrix R and solve for the weight vector $\mathbf{w}$, then the weight vector $\mathbf{w}$ will be used to filter the batch of data. In this section, we are going to develop the following online algorithms: adaptive robust Capon beamformer and adaptive time delay based robust MVDR beamformer. The weight vector $\mathbf{w}$ is updated upon every new data sample, which is more like LMS-based adaptive filtering. That's why we call these algorithms adaptive algorithms.

4.4.1 Adaptive Robust Capon Beamformer

We now develop an adaptive algorithm for the robust Capon beamformer. This adaptive algorithm can be separated into two adaptive steps, the first step is for updating the steering vector estimation $\hat{\mathbf{a}}$, the second step is for updating the beamformer's weight vector $\mathbf{w}$.

For simple notation, let Q denote R^{-1}, then problem (4.7) can be written as,

$$\min_{\mathbf{a}} \mathbf{a}^H Q\mathbf{a}, \quad \text{s.t.} \quad \|\mathbf{a} - \overline{\mathbf{a}}\|^2 \leq \varepsilon. \tag{4.19}$$

Suppose $\mathbf{a} = \mathbf{a}_r + j\mathbf{a}_i$ and $\overline{\mathbf{a}} = \overline{\mathbf{a}}_r + j\overline{\mathbf{a}}_i$, where $\mathbf{a}_r$, $\overline{\mathbf{a}}_r$ and $\mathbf{a}_i$, $\overline{\mathbf{a}}_i$ are the real and image parts of the steering vector $\mathbf{a}$ and $\overline{\mathbf{a}}$ respectively. We can write $Q = Q_r + jQ_i$ as well, where Q_r and Q_i are the real and image parts respectively. We construct new concatenated real vectors and matrices.

$$\breve{\mathbf{a}} = [\mathbf{a}_r^T \quad \mathbf{a}_i^T]^T \tag{4.20}$$

$$\breve{Q} = \begin{bmatrix} Q_r & -Q_i \\ Q_i & Q_r \end{bmatrix} \tag{4.21}$$

Since Q is positive definite,

$$\mathbf{a}^H Q\mathbf{a} = real(\mathbf{a}^H Q\mathbf{a}) \tag{4.22a}$$

$$= \breve{\mathbf{a}}^T \breve{Q}\breve{\mathbf{a}} \tag{4.22b}$$

$$\|\mathbf{a} - \overline{\mathbf{a}}\|^2 = \|\breve{\mathbf{a}} - \breve{\overline{\mathbf{a}}}\|^2 \tag{4.23}$$

Hereby the problem (4.19) can be written as

$$\min_{\breve{\mathbf{a}}} \breve{\mathbf{a}}^T \breve{Q}\breve{\mathbf{a}}, \quad \text{s.t.} \quad \|\breve{\mathbf{a}} - \breve{\overline{\mathbf{a}}}\|^2 \leq \varepsilon. \tag{4.24}$$

This is a real optimization problem equivalent to the original complex optimization problem. Let $\breve{\mathbf{a}}_e = \breve{\mathbf{a}} - \breve{\overline{\mathbf{a}}}$, problem (4.24) can be written as

$$\min_{\breve{\mathbf{a}}_e} (\breve{\mathbf{a}}_e + \breve{\overline{\mathbf{a}}})^T \breve{Q}(\breve{\mathbf{a}}_e + \breve{\overline{\mathbf{a}}}), \quad \text{s.t.} \quad \|\breve{\mathbf{a}}_e\|^2 \leq \varepsilon. \tag{4.25}$$

Since Q is a Hermitian matrix, which implies $\breve{Q} = \breve{Q}^T$, thus problem (4.25) is equivalent to

$$\min_{\breve{\mathbf{a}}_e} \breve{\mathbf{a}}_e^T \breve{Q}\breve{\mathbf{a}}_e + 2\breve{\mathbf{a}}_e^T \breve{Q}\breve{\overline{\mathbf{a}}}, \quad \text{s.t.} \quad \|\breve{\mathbf{a}}_e\|^2 \leq \varepsilon. \tag{4.26}$$

This is a quadratic optimization problem with inequality norm constraint. In [145], an iterative solution (also adaptive algorithm) was developed to solve an optimization problem in exactly the same format. The iterative solution can be written as follows.

$$\breve{\mathbf{a}}_{e,k+1} = P[(\mathbf{I} - \mu\breve{Q})\breve{\mathbf{a}}_{e,k} - \mu\breve{Q}\breve{\overline{\mathbf{a}}}] \tag{4.27a}$$

$$P(\breve{\mathbf{a}}_e) = \begin{cases} \breve{\mathbf{a}}_e, & \|\breve{\mathbf{a}}_e\| \leq \sqrt{\varepsilon} \\ \breve{\mathbf{a}}_e \frac{\sqrt{\varepsilon}}{\|\breve{\mathbf{a}}_e\|}, & \|\breve{\mathbf{a}}_e\| > \sqrt{\varepsilon}. \end{cases} \tag{4.27b}$$

where μ is a parameter controlling the convergence speed. To assure convergence, we must choose $0 < \mu < 2/(\lambda_1 + \lambda_K)$, where λ_1 and λ_K is the maximum and the minimum eigenvalues of matrix $\breve{Q}$ respectively. The corresponding adaptive algorithm is obtained by substituting the instantaneous estimates of $\breve{Q}$ into the algorithm (4.27).

After some tedious mathematical manipulations, we can render the iterative algorithm (4.27) as the following iterative algorithm which operate on the original complex vector variable.

$$\mathbf{a}_{e,k+1} = P[(\mathbf{I} - \mu Q)\mathbf{a}_{e,k} - \mu Q\overline{\mathbf{a}}] \tag{4.28a}$$

$$P(\mathbf{a}_e) = \begin{cases} \mathbf{a}_e, & \|\mathbf{a}_e\| \leq \sqrt{\varepsilon} \\ \mathbf{a}_e \frac{\sqrt{\varepsilon}}{\|\mathbf{a}_e\|}, & \|\mathbf{a}_e\| > \sqrt{\varepsilon}. \end{cases} \tag{4.28b}$$

Although equation (4.27) and equation (4.28) show the same format, they are actually different. Equation (4.27) operates on concatenated real vector variable while equation (4.28) operates on complex vector variable.

The only little revision needed for applying the corresponding adaptive algorithm is to provide the instantaneous estimation of Q, i.e. R^{-1}. To lower the computation complexity, the inverse matrix lemma is used for updating R_k^{-1} when every new data sample is available, here k is the time stamp. For estimation of the correlation matrix R_k at time k, we have,

$$R_k = \alpha R_{k-1} + (1 - \alpha)\mathbf{x}_k\mathbf{x}_k^H \tag{4.29}$$

where α is the forgetting factor, $\mathbf{x}_k$ is the new data sample at time k. Using inverse matrix lemma, we can get

$$R_k^{-1} = \frac{1}{\alpha}R_{k-1}^{-1} - \frac{1-\alpha}{\alpha^2}\mathbf{u}(I + \frac{1-\alpha}{\alpha}\mathbf{x}_k^H\mathbf{u})^{-1}\mathbf{u}^H \tag{4.30}$$

where $\mathbf{u} = R_{k-1}^{-1}\mathbf{x}_k$.

Thereby we get the adaptive algorithm for updating the steering vector estimation $\hat{\mathbf{a}}$ when every new data sample is available. We now develop the adaptive step for updating beamformer's weight vector $\mathbf{w}$ at every time stamp. Since we are assuming the steering vector $\mathbf{a}$ is known in this step, we are actually developing an adaptive algorithm for the standard MVDR beamforming problem (4.1). Similar to Frost's process for developing the Frost LMS beamformer [19], we do the following deduction.

Using complex gradient-descent optimization, at each time stamp k the weight vector is updated as follows

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \mu\nabla_{\mathbf{w}}H[\mathbf{w}_k] \tag{4.31}$$

where $H(\mathbf{w}) = \mathbf{w}^H R\mathbf{w} + Re\{\lambda^H(\mathbf{w}^H a - 1)\}$ is the constrained cost function. Using complex gradient, similar to [19], we can get

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \mu[I - \mathbf{a}(\mathbf{a}^H\mathbf{a})^{-1}\mathbf{a}^H]R\mathbf{w}_k + \mathbf{a}(\mathbf{a}^H\mathbf{a})^{-1}[1 - \mathbf{a}^H\mathbf{w}_k] \tag{4.32}$$

The last term is not assumed to be zero. It is kept to prevent arithmetic inaccuracy from accumulation and growth.

Define $G = \mathbf{a}(\mathbf{a}^H\mathbf{a})^{-1}$ and $P = I - \mathbf{a}(\mathbf{a}^H\mathbf{a})^{-1}\mathbf{a}^H$, the iterative equation can be written as

$$\mathbf{w}_{k+1} = P[\mathbf{w}_k - \mu R\mathbf{w}_k] + G \tag{4.33}$$

By substituting the instantaneous estimation $R_k = \mathbf{x}_k\mathbf{x}_k^H$ into equation (4.33), we get the adaptive algorithm for updating $\mathbf{w}$.

4.4.2 Adaptive Time Delay Based Robust MVDR Beamformer

Similar to what we have done for the RCB (section 4.4.1), we develop a two-step adaptive algorithm for the time delay based robust MVDR beamformer. The first step is for updating the time delay estimation $\hat{\tau}$, while the second step is for updating the beamformer's weight vector $\mathbf{w}$.

We use complex gradient descent algorithm for adapting the time delay estimation $\hat{\tau}$. The gradient for the narrowband objective function $h_i(\tau) = \mathbf{a}_i^H R_i^{-1}\mathbf{a}_i$ has

been derived in equation (4.14). Its extension for broadband signal's objective function $f(\tau) = \sum_i \mathbf{a}_i^H R_i^{-1} \mathbf{a}_i$ in problem (4.17) is quite straight forward,

$$\nabla_\tau f = \sum_i \nabla_\tau h_i \tag{4.34a}$$

$$= \sum_i real(A^* R_i^{-1} \mathbf{a}_i) \tag{4.34b}$$

After we update the estimation of time delay $\hat{\tau}$ at every time stamp, the steering vector $\mathbf{a}_i$ can be easily acquired by equation (4.12). The adaptive step for updating beamformer's weight vector $\mathbf{w}$ at every time stamp can be derived in the same way as in adaptive RCB beamformer (section 4.4.1).

4.5 Simulation

4.5.1 Beamforming Algorithms Notation

We use the following notation for each beamforming algorithm.

- OMVDR: the optimal MVDR beamformer which assumes we know the true SOI steering vector
- MVDR: standard MVDR beamformer based on the assumed steering vector
- DS: conventional delay and sum beamformer
- RCB: robust Capon beamformer (section 4.2.2)
- DCRCB: doubly constrained robust Capon beamformer (section 4.2.3)
- robMVDRtd: time delay based robust MVDR beamformer (section 4.3.1, equation (4.13))
- robMVDRtdmultif: time delay based multi-frequency robust MVDR beamformer (section 4.3.1, equation (4.17))
- robMVDRangle: angle based robust MVDR beamformer (section 4.3.2)
- adaptrobMVDR: adaptive robust Capon beamformer (section 4.4.1)

- adaptrobMVDRtd: adaptive time delay based robust MVDR beamformer (section 4.4.2)

4.5.2 Simulation Scenario

In this section, we provide numerical examples on speech enhancement using a microphone array to compare the performances of various beamformers. We simulate an 8 element uniform linear array with 4cm equal spacing between adjacent elements. The sources, both the SOI and interference signals, are plane waves which exist in the same plane as the linear array. In the experiments, we always assume the look direction is the broadside of the array, i.e. 0°. The uncertainty bound used in RCB, DCRCB (ε in equation (4.7)) and robMVDRtd beamformer (δ in equation (4.13)) are calculated based on the uncertainty range -8° to 8°.

In the simulation, every source signal is a speech wave signal. The sampling rate is $8k$Hz. Since we are discussing narrowband beamforming algorithms while speech is a wideband signal, we use short time Fourier transform (STFT) to map the time domain speech signal into frequency domain based on consecutive frames. Then the narrowband beamforming algorithms are applied on every frequency bin. The frame length is $0.25s$ (200 samples), with a step length of $0.125s$ (100 samples). A 256 points FFT is used on each frame. The performance of various beamformers is measured by the cepstral distance between the recovered signal's spectrum and the original SOI's spectrum. The cepstral distance is used because it is a perceptual metric commonly used in speech processing to measure distortion. Note that we ignore the experiments on angle based robust MVDR beamformer (robMVDRangle) since its performance has been illustrated and explained in [30].

4.5.3 Simulation with Batch Algorithms

Example 1

Fig.4.4 demonstrates the beamformers' performance versus SNR (signal to white noise ratio). Only one SOI and one interference signal exist in this experiment. The interference signal and SOI have similar level of energy. The interference signal comes

from direction 45°. The assumed look direction is broadside, i.e. 0°, while the true SOI direction is 5°, which means a 5° look direction error. There's no sensor position error in this experiment.

The OMVDR beamformer gives the optimal performance and bounds the performance that can be attained by these class of adaptive beamformers. Our simulation results clearly demonstrate that the proposed robMVDRtdmultif beamformer consistently performs well and approximates the OMVDR beamformer. It outperforms the conventional fixed DS beamformer and other adaptive beamformers such as MVDR, RCB and DCRCB. The robMVDRtd beamformer has better performance than RCB and DCRCB if the SNR is high; and has a little inferior performance compared to RCB if the SNR is low.

The RCB and DCRCB demonstrate similar performance, and RCB is always a little better among the two. RCB and DCRCB can suppress most of the interference energy and perform better than DS beamformer. By a detailed inspection of the beamformers' performance in every frequency bin, we find that the estimated steering vector by RCB and DCRCB is close to the true SOI steering vector when it is in the low frequency range. When it is in the high frequency range, the steering vector estimated by RCB and DCRCB is far away from the true SOI steering vector. Actually, the magnitude of each element of the estimated steering vector is far from 1. Of course, the estimated steering vector is not a valid steering vector anymore. This explains why RCB and DCRCB has inferior performance in high frequency range.

The performance of various beamformers can be certified by examining the beampatterns in Fig.4.8 and Fig.4.9. Fig.4.8 and Fig.4.9 illustrates the magnitude beam pattern of various beamformers on one sample data when there is only one interference signal. This sample data is selected from the data set used to generate Fig.4.4. The SNR is 37dB. The MVDR beamformer forms two deep nulls, one in the interference direction, the other in the SOI direction, which implies SOI cancellation. The RCB and DCRCB method can steer a null in the interference direction (45°) at low frequency range, while at middle to high frequency range, their beam patterns are similar to that of DS beamformer. This can be explained by the choice of uncertainty bound ε. The bound ε is chosen such that all possible SOI steering vectors $\mathbf{a}_0$ are included in the prescribed

uncertainty set. This usually brings on a big value of ε at high frequency, which results in many infeasible steering vectors being included in the uncertainty set (4.3). Thereby the minimizer to the optimization problem (4.7) and (4.9) is no longer a valid steering vector in the high frequency range. A close inspection of the estimated steering vector by RCB and DCRCB at high frequency range confirms the above reasoning. The element magnitudes of those steering vectors have been found to be far away from 1. The robMVDRtd beamformer can put a deep null in the interference direction consistently over all frequencies. It is evident that the beam pattern of the robMVDRtdmultif method is close to that of the OMVDR beamformer.

Example 2

Fig.4.5 shows the beamformers' performance versus SNR when two interference signals exist. The second interference signal injects from direction $-60°$. All the other settings are the same as those of the aforementioned experiment.

The robMVDRtd beamformer demonstrates much better performance than RCB and DCRCB under all different SNR level. But it is worse than the robMVDRtdmultif beamformer. This is reasonable since robMVDRtd is not utilizing the information that speech is broadband signal and processes each frequency bin independently. In those frequency bins where the speech signal does not have high energy level, corresponding to the vallies on speech spectrum, the local SNR is lower than the global averaged SNR, thereby the robMVDRtd will have worse performance on those frequency bins. On the contrary, the robMVDRtdmultif beamformer consider the broadband speech signal as one integrated part and just solve one time delay τ across all the frequency bins, since the time delay τ should be the same across all the frequencies for a broadband signal. This is apparently more noise robust. What's more, we can estimate and track the local SNR in every frequency bin and only use those frequency bins with high local SNR to update the time delay τ estimation in equation (4.17). This is especially useful if we have a broadband signal and part of the spectrum is corrupted by a band limited noise. We can use those frequency bins not being corrupted to estimate a time delay τ for applying beamforming in all the frequencies.

The RCB and DCRCB demonstrate similar performance, and RCB is always

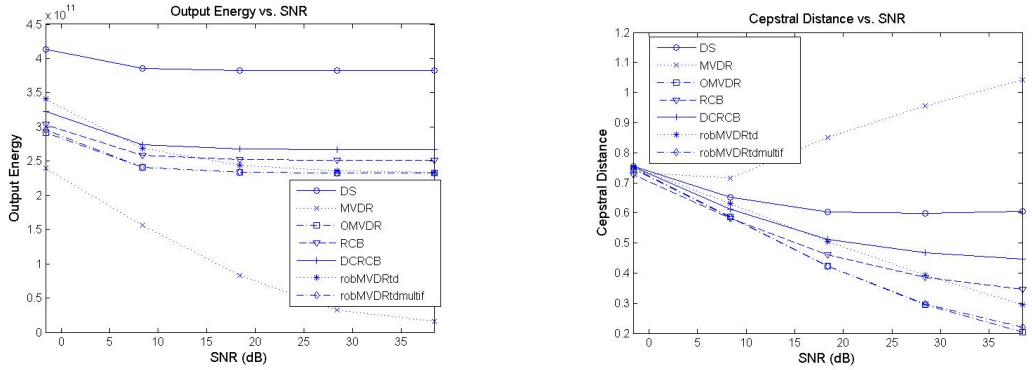
a little better among the two. The performance of RCB and DCRCB is much worse than that of robMVDRtd and robMVDRtdmultif beamformer and is close to that of DS beamformer. Detailed inspection of the beamformers' performance in every frequency bin shows that the steering vector estimated by RCB and DCRCB is far away from the true SOI steering vector. The magnitude of each element of the estimated steering vector is far from 1. Comparing Fig.4.5 with Fig.4.4, we see that RCB and DCRCB have much worse performance when there are two interference signals compared to the single interference scenario. When there are two interferences, the correlation matrix R has more signal components, thereby RCB and DCRCB have more freedom to minimize the criterion function $\mathbf{a}^H R^{-1}\mathbf{a}$ (equation (4.7)), this makes the estimated steering vector to move even far away from a valid steering vector compared to the single interference scenario.

The performance of various beamformers can be certified by examining the beampatterns in Fig.4.10. Fig.4.10 illustrates the magnitude beam pattern of various beamformers on one sample data when there exist two interference signals. This sample data is selected from the data set used to generate Fig.4.5. RCB and DCRCB shows a beam pattern similar to DS beamformer. The robMVDRtd beamformer can put a deep null in the interference directions consistently over all frequencies. However, on some frequency bins, the robMVDRtd beamformer partially suppresses the SOI. The robMVDRtdmultif beamformer shows a beam pattern which is very close to that of the OMVDR beamformre.

Example 3

Fig.4.6 demonstrates the beamformers' performance versus SNR when there exist not only the look direction error, but also sensor position error. The displacement error for each sensor is generated by an uniformly distributed random variable whose maximum value is 3mm. Only one interference from 45° exists in this experiment. Fig.4.7 demonstrates the beamformers' performance versus SNR when there exist both look direction error and sensor position error. Two interference speech from 45° and −60° exist in this experiment.

By this example we show that the RCB, DCRCB, robMVDRtd and robMVDRt-



(a) Output Energy

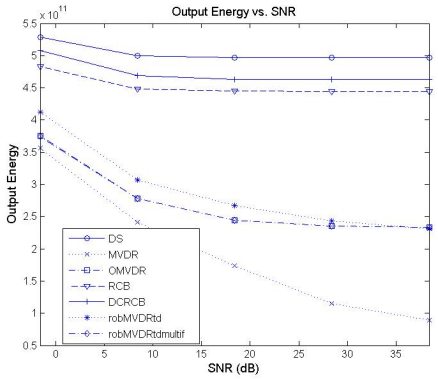
(b) Cepstral distance

Figure 4.4: The performance of beamformers vs. SNR, only look direction error exists, one interference.

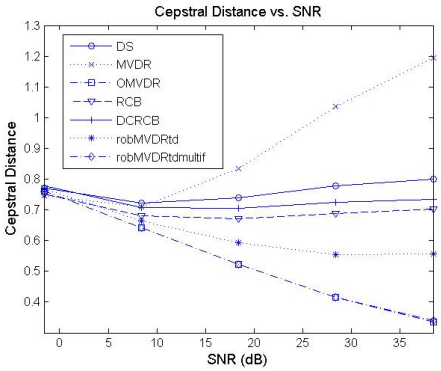
dmultif beamformers are not sensitive to the sensor position error. The results shown in Fig.4.6 and Fig.4.7 are very close to the results in Fig.4.4 and Fig.4.5.

4.5.4 Simulation with Adaptive Algorithms

We use computer simulations to demonstrate the performance of the adaptive robust Capon beamformer (adaptrobMVDR) and the adaptive time delay based robust MVDR beamforming (adaptrobMVDRtd). We assume two sources: the SOI direction is 5°, while the interference signal comes from 45°. The look direction is the broadside direction (0°), which means a 5° look direction error. A Monte Carlo simulation of 20 independent runs was performed for different adaptation lengths. The data for testing performance is 100 seconds long. The result shown in the following is the mean value of those 20 independent Monte Carlo runs. (*To be more specific: for every fixed adaptation length, the same experiment is repeated* 20 *times based on random data selection. That is, for every fixed adaptation length, we pick* 20 *segments of data of that fixed length from the database, run a beamforming algorithm on those segments, get* 20 *weight vectors, then test those* 20 *weight vectors on the test data, which is* 100 *seconds. We got* 20 *cepstral distances for each fixed adaptation length. The final results shown on the performance plots are the mean value of those* 20 *passes.*) The step size for adaptation, i.e. the frame rate for adaptation is $2ms$ (16 samples under $8K$Hz sampling rate).

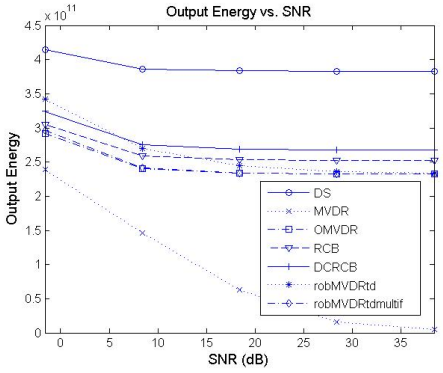


(a) Output Energy

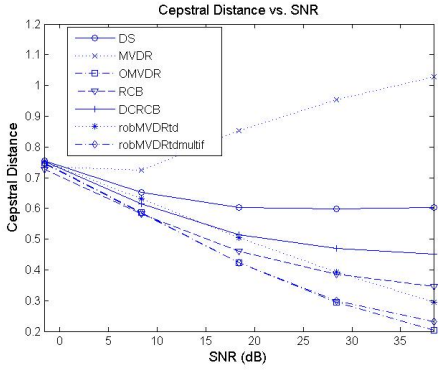


(b) Cepstral distance

Figure 4.5: The performance of beamformers vs. SNR, only look direction error exists, two interferences.

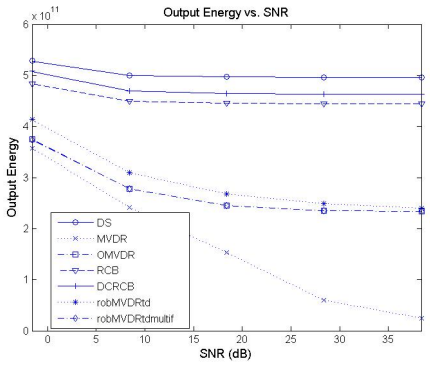


(a) Output Energy

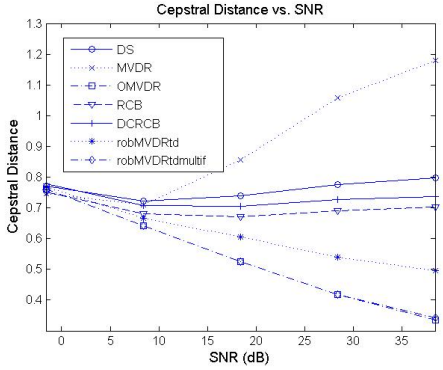


(b) Cepstral distance

Figure 4.6: The performance of beamformers vs. SNR, both look direction error and sensor position error exist, one interference.



(a) Output Energy



(b) Cepstral distance

Figure 4.7: The performance of beamformers vs. SNR, both look direction error and sensor position error exist, two interferences.

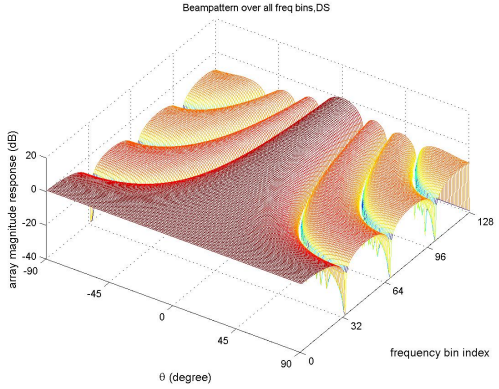


Figure 4.8: Beam pattern of the delay-and-sum beamformer over angle θ and frequency bins. The look direction is 0°.

Result with adaptive RCB beamformer

The performance of the adaptive RCB beamformer versus different adaptation lengths is illustrated in Fig.4.11. S1 is the desired signal. On the sub-figure (a) where the energy output is illustrated, the line with legend $S1$ (DS, and $adaptiverobMVDR$ respectively) denotes the energy of the desired source signal (the delay-and-sum beamformer output, and the adaptive RCB beamformer output respectively). Sub-figure (b) demonstrates the cepstral distance between the desired signal and the delay-and-sum beamformer output (the adaptive RCB beamformer output respectively). Fig.4.11 illustrates the performance of the adaptive RCB beamformer improves as the adaptation goes on.

Result with adaptive time delay based robust MVDR beamformer

The performance of the adaptive time delay based robust MVDR beamformer versus different adaptation lengths is illustrated in Fig.4.12. Compared with Fig.4.11, the adaptive time delay based robust MVDR beamformer demonstrates a better performance compared to the adaptive RCB beamformer for the same adaptation length.

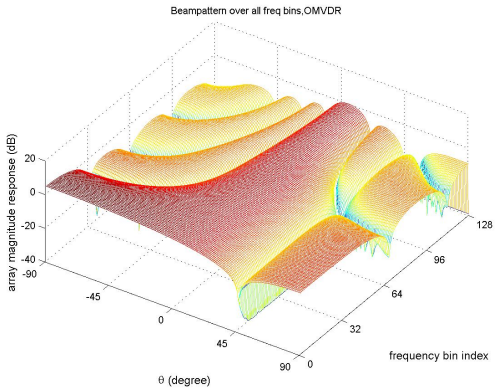
4.6 Conclusions

We analyzed the RCB and DCRCB beamforming algorithms and demonstrated that RCB and DCRCB may overestimate the SOI power since the estimated steering vector may not be a valid steering vector. The problem becomes more serious when the uncertainty bound is large. We proposed two solutions to this problem. One approach is to use a set of RCBs (or DCRCBs) with small uncertainty bounds to cover a large uncertainty range. The other approach is to enforce some extra constraints on the optimization problem so that the solution is a valid steering vector. We developed variants of the constrained robust MVDR beamformer that attempt to limit the search in the underlying optimization problem to a feasible set of steering vectors thereby achieving improved performance. The robustness against steering vector error was provided through a spherical uncertainty set constraint, while a set of magnitude constraints was enforced on each element of the steering vector to better constrain

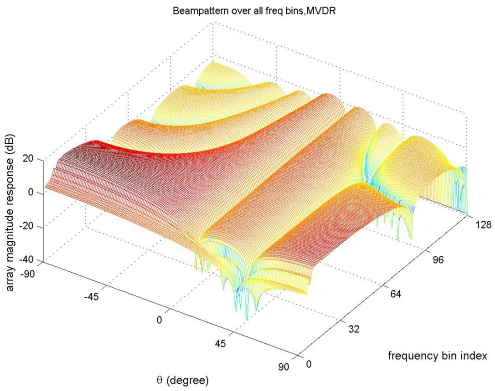
the search in the space of feasible steering vectors. By appropriately changing the variables, the optimization problem was modified such that the need for the magnitude constraints are avoided. Our numerical simulations demonstrated that our time delay based robust MVDR beamformer performs better than RCB and DCRCB under large uncertainty bounds. We also developed adaptive algorithms for the RCB and the time delay based robust MVDR beamformer. The adaptive algorithms have two updating steps. The first step updates the steering vector estimation or the time delay estimation; the second step updates the beamformer's weight vector given an estimated steering vector. The adaptive time delay based robust MVDR beamformer performs better than the adaptive RCB beamformer.

Acknowledgement

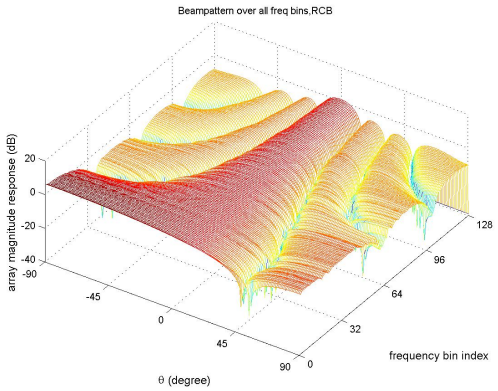
The text of this chapter, in part, is a reprint of the material which has been published in Wenyi Zhang, and Bhaskar D. Rao, "Robust Adaptive Beamformer with Feasibility Constraint on the Steering Vector", *European Signal Processing Conference, 2006, EUSIPCO*, Florence, Italy, 2006.



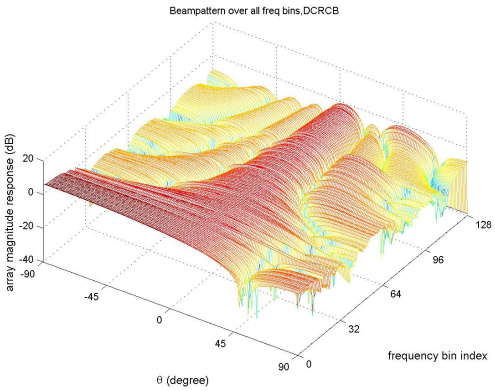
(a) OMVDR beamformer



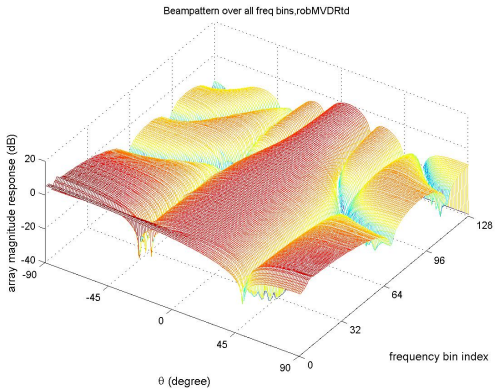
(b) standard MVDR beamformer



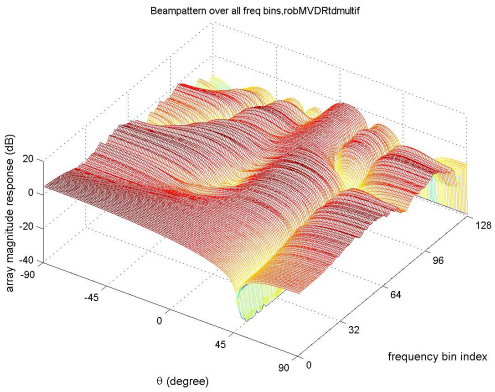
(d) RCB beamformer



(e) DCRCB beamformer

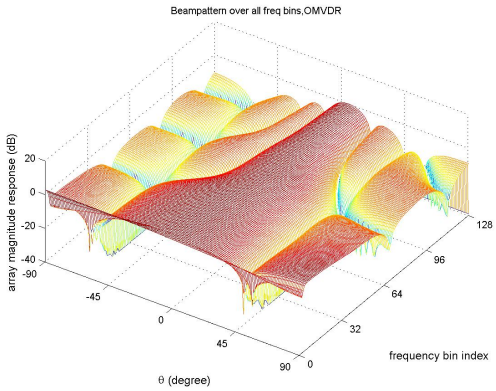


(f) robMVDRtd beamformer

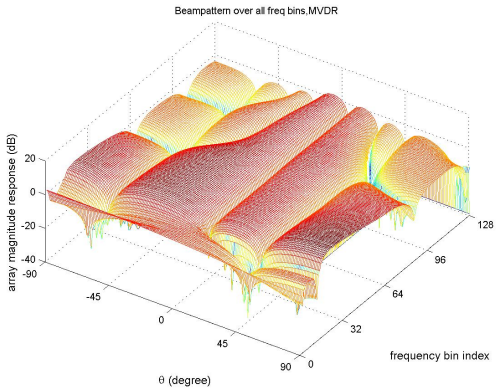


(e) robMVDRtdmultif beamformer

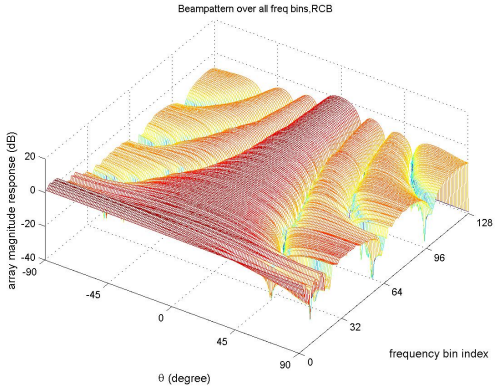
Figure 4.9: Beam pattern of various beamformers over angle θ and frequency bins. The look direction is 0°, the true SOI direction is 5°, and the interference direction is 45°.



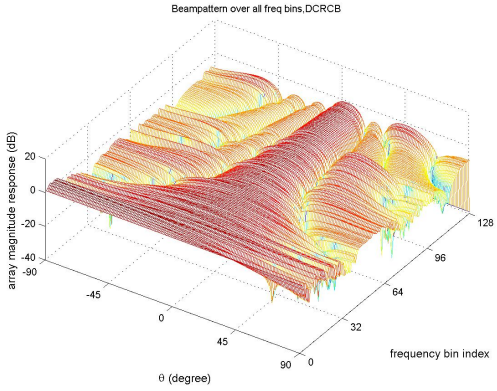
(a) OMVDR beamformer



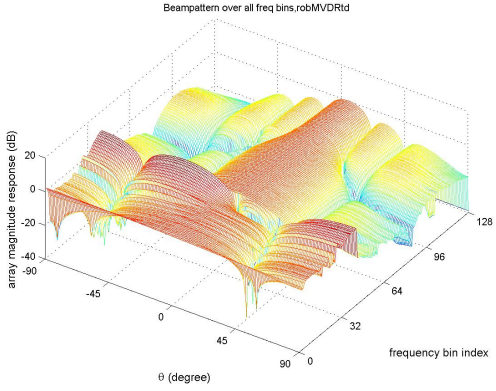
(b) standard MVDR beamformer



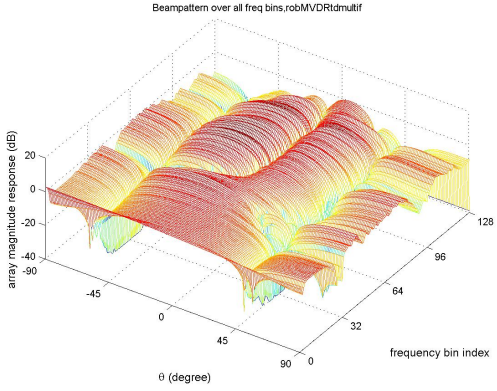
(d) RCB beamformer



(e) DCRCB beamformer

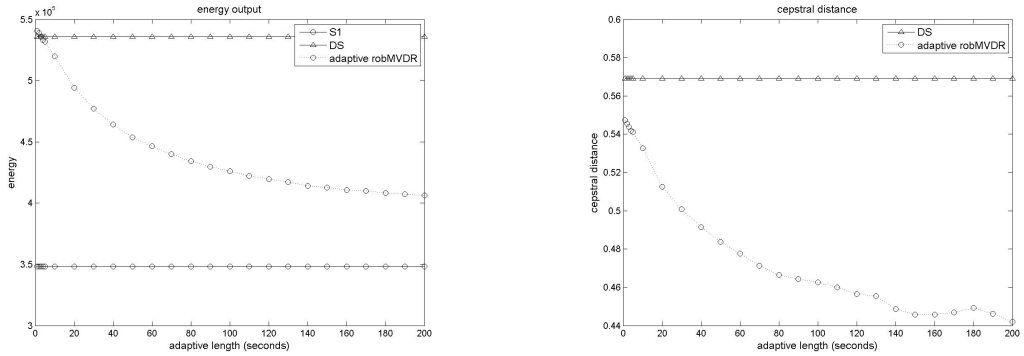


(f) robMVDRtd beamformer



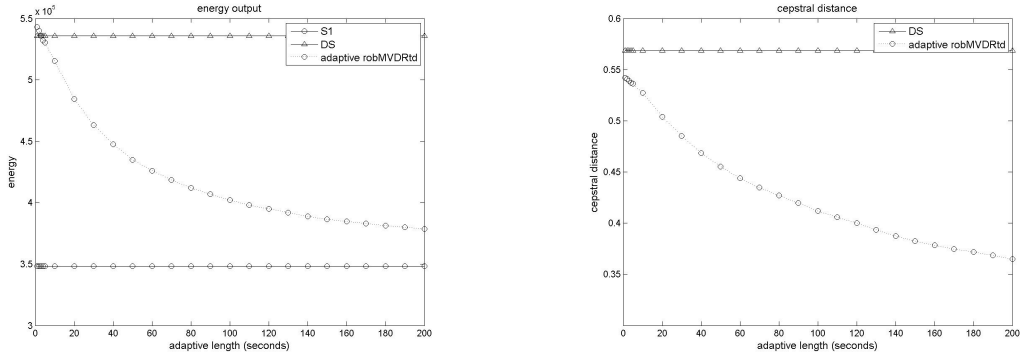
(e) robMVDRtdmultif beamformer

Figure 4.10: Beam pattern of various beamformers over angle θ and frequency bins. The look direction is 0°, the true SOI direction is 5°, one interference comes from 45°, another interference comes from −60°.



(a) Energy output (b) Cepstral distance

Figure 4.11: The performance of the adaptive RCB beamformer vs. adaptation length, 2ms step size



(a) Energy output (b) Cepstral distance

Figure 4.12: The performance of the adaptive robMVDRtd beamformer vs. adaptation length, 2ms step size

5 Insights into the Frequency Domain ICA/IVA Approach

Another class of promising multi-channel signal separation algorithms besides beamforming is independent component analysis (ICA). ICA is a blind source separation (BSS) technique which only utilizes the statistical independence between several sources to separate them out from their mixtures [57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83]. For separating convolutive mixed source signals, the frequency domain ICA approach is often used because it simplifies the time domain convolutive mixing problem into the instantaneous mixing ICA problem in each frequency bin. However, permutation problem across different frequency bins is an important problem that needs to be solved before the frequency domain ICA approach can be used to separate mixed sources in real applications [84, 72, 85, 86, 87, 88, 89, 90, 87, 91, 92, 93, 94, 95]. Independent vector analysis (IVA) was proposed to circumvent the permutation problem [96, 97, 98, 99, 100, 101, 102, 103]. Instead of solving the mixing problem in each frequency bin independently, IVA considers the frequency domain source signal linked together as a vector source and attempts to resolve the mixing problem in the frequency bins in an integrated manner.

In this work, we examine and provide insight into frequency domain ICA methods for source separation in reverberant environments. For the modeling of the source signals, we develop the concept of a dynamic random process to model the source signals. It formalizes the concept of signals that are stationary in a frame but exhibit dynamics at the frame level. Frame dynamics is an important characteristic of these signals and prove important to the success of the ICA methods. With suitable assumptions, the

dynamic random process is stationary in the ensemble sense while a given realization may in an engineering sense exhibit 'non-stationarity'. We show for dynamic random processes, the unconditional distribution of the source signal in each frequency bin is a Gaussian scale mixture (GSM). The non-Gaussianity, which is critical to ICA, of the source signal in each frequency bin is shown to be a direct consequence of the frame dynamics. Furthermore, the independence between the unconditional distributions of the source signals in each frequency bin is related to the independence of the frame dynamics of the mixed time domain source signals. The GSM mathematical modeling is extended to the vector random processes formed by stacking the different frequency components of a source. This provides insights into the mathematical models suitable for the frequency domain independent vector analysis (IVA) type approaches. A special case of the distribution turns out to be the 'spherical distribution' employed in IVA source modeling providing support to their use in source separation. Concentrating on the bin-wise ICA methods, a significant contribution of this work is to show that signals modeled using GSM density can be separated using ICA even though they might be dependent on each other as long as the the frame dynamics of the source signals are different almost surely.

5.1 Introduction to the Frequency Domain ICA Approach

Ever since the frequency domain independent component analysis (ICA) approach was proposed [146, 147, 148, 84, 72, 91, 86, 87, 90, 83, 75, 78], it has been widely used and has become a conventional blind source separation (BSS) approach for de-mixing convolutive mixtures. It simplifies the time domain convolutive mixing problem by turning it into instantaneous mixing problem in each frequency bin. In this work, we will focus on the discussion of the applicability of ICA algorithm in each frequency bin – when can ICA separate the mixed source signals in each frequency bin.

For convolutive mixed blind source separation, Pham et al. argued that in the frequency domain the source signal at each fixed frame and frequency is a Gaussian signal, therefore ICA is not applicable [104]. Parra, Pham, and Choi et al. all proposed

to utilize the non-stationarity of source signals for instantaneous mixing blind source separation [105, 67, 106, 107, 66, 108, 109, 110]. In the frequency domain approach for de-mixing convolutive mixtures, if Parra, Pham, and Choi et al.'s methods are applied in each frequency bin to separate the mixed source signals therein, the source signals in each frequency bin are implicitly assumed to be non-stationary. By simultaneously diagonalizing several covariance matrices (second order statistics) at different times in each frequency bin, the mixed source signals in each frequency bin are separated. However, the permutation problem stills need to be solved in order to obtain each source's spectrum for reconstructing the time domain source signal. One shortcoming of this approach is the need for a large amount of data so that the covariance matrices in each frequency bin can be estimated robustly (remember several covariance matrices have to be estimated in each frequency bin). For instance, for convolutive mixed blind speech signal separation more than 20 to 30 seconds of data is necessary to obtain satisfactory results [105].

However, the conventional frequency domain ICA approach (Table 5.1) is widely used and many experiments on convolutive mixing blind source separation demonstrate good separation performance (assuming the permutation problem is correctly solved) no matter what kind of instantaneous mixing ICA algorithms is employed in separating the mixed source signals in each frequency bin. For example, maximizing neg-entropy (FastICA) algorithm, maximum likelihood algorithm, and Infomax algorithm [58, 60, 61, 63, 62] have been tested in the frequency domain ICA approach and all of them illustrate good separation results. An advantage of these ICA algorithms is that their performance is not too sensitive to the amount of data. For example, in convolutive mixing speech separation, a few seconds of data is enough to provide satisfactory separation result.

In this work, we examine and provide insight into frequency domain ICA methods for source separation in reverberant environments. For the modeling of the source signals, we develop the concept of a dynamic random process to model the source signals. It formalizes the concept of signals that are stationary in a frame but exhibit dynamics at the frame level. Frame dynamics is an important characteristics of these signals and prove important to the success of the ICA methods. With suitable assumptions, the

dynamic random process is stationary in the ensemble sense while a given realization may in an engineering sense exhibit 'non-stationarity'. We show for dynamic random processes, the unconditional distribution of the source signal in each frequency bin is a Gaussian scale mixture (GSM). The non-Gaussianity, which is critical to ICA [57], of the source signal in each frequency bin is shown to be a direct consequence of the frame dynamics. Furthermore, the independence between the unconditional distributions of the source signals in each frequency bin is related to the independence of the frame dynamics of the mixed time domain source signals. The GSM mathematical modeling is extended to the vector random processes formed by stacking the different frequency components of a source. This provides insights into the mathematical models suitable for the frequency domain independent vector analysis (IVA) type approaches. A special case of the distribution turns out to be the 'spherical distribution' employed in IVA source modeling providing support to their use in source separation. Concentrating on the bin-wise ICA methods, a significant contribution of this work is to show that signals modeled using GSM density can be separated using ICA even though they might be dependent on each other as long as the the frame dynamics of the source signals are different almost surely. In particular, we show that Kurtosis and neg-entropy type measures can be used to separate variance correlated GSM signals.

The rest of the chapter is organized as follows. Section 5.2 proposes the signal model and analyzes the statistical distributions of the source signals in the frequency domain for both scalar and vector sources. First, the unconditional distribution of a single source signal in the frequency domain is shown to be a GSM. Second, the unconditional joint distribution of the multiple source signals is analyzed. Section 5.3 proves that the ICA batch algorithms may separate mixed variance correlated GSM signals. Section 5.4 provides the analysis on the applicability of ICA/IVA online algorithms.

5.2 Signal Modeling

We consider a reverberant environment wherein N time domain source signals $s_i(t), i \in \{1, .., N\}$ are measured after they are convolved with the corresponding channel impulse response filters. Mathematically, the measured signal at the N sensors can be

Table 5.1: The conventional frequency domain ICA approach

(1) Apply short time Fourier transform (STFT) on each time domain mixture based on consecutive frames. Thus the convolutive mixing problem in the time domain is transformed into an instantaneous mixing problem in each frequency bin.
(2) On each frequency bin, apply batch ICA algorithm (instantaneous mixing ICA) to separate the source signals.
(3) Use appropriate permutation and scaling correction method to solve the permutation and scaling ambiguity across all frequency bins, so that the spectrum of the source signals can be recovered.
(4) Apply inverse discrete Fourier transform (DFT) on the recovered source spectrum and use overlap-and-save (or overlap-and-add) method to obtain the time domain signal.

represented as

$$\mathbf{x}(t) = \sum_{\tau=0}^{P} A^{td}(\tau)\mathbf{s}(t-\tau) \tag{5.1}$$

where

$$\mathbf{x}(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_N(t) \end{bmatrix} \quad A^{td}(\tau) = \begin{bmatrix} a_{11}(\tau) & \cdots & a_{1N}(\tau) \\ \vdots & \ddots & \vdots \\ a_{N1}(\tau) & \cdots & a_{NN}(\tau) \end{bmatrix} \quad \mathbf{s}(t) = \begin{bmatrix} s_1(t) \\ \vdots \\ s_N(t) \end{bmatrix}$$

$\mathbf{x}(t)$ represents the N mixtures, $A^{td}(\tau), \tau = 1, .., P$ are the matrices of mixing filters with $a_{ij}(\tau)$ representing the impulse response filter from source j to sensor i.

The conventional frequency domain ICA approach segments the sensor measurements (data) into frames and applies the short time Fourier transform (STFT) on the consecutive frames. This transforms the time domain convolutive mixing problem (Eq. 5.1) into an instantaneous mixing problem in each frequency bin. The mixing problem in the k_{th} frequency bin can be written as,

$$\mathbf{X}(k,n) = A(k)\mathbf{S}(k,n) \tag{5.2}$$

where $n \in \{1, .., L\}$ is the frame index, L is the total number of frames. $\mathbf{S}(k,n) = [S_1(k,n), ..., S_N(k,n)]^T$, where $S_i(k,n), k = 1, .., K$ is the Discrete Fourier transform

(DFT) coefficients of the time domain source signal $s_i(t)$ at the n_{th} frame. k is the frequency bin index, K is the DFT length, i.e. the total number of frequency coefficients. $\mathbf{X}(k,n) = [X_1(k,n),...,X_N(k,n)]^T$, where $X_i(k,n), k = 1,..,K$ is the DFT of the time domain mixture signal $x_i(t)$ at the n_{th} frame. $A(k)$, the fourier transform of the matrix channels $A^{td}(\tau)$, is the N by N mixing matrix at the k^{th} frequency bin. As a notational convention we will use i for the source index, k for the frequency bin index, and n for the frame index in this chapter. The steps of the conventional frequency domain ICA approach are summarized in Table 5.1.

We now make some assumptions about the signals leading to the development of signal models that are then employed to evaluate and understand frequency domain methods. An assumption that is commonly made about speech to capture its dynamic nature is that it is stationary at the frame level with the statistics changing from frame to frame. The time varying nature of the signals was also identified to be a key attribute contributing to the success of blind source separation methods [67, 106]. Motivated by the piecewise stationary (quasi-stationary) models used in speech modeling, we first start by making the following assumptions on a single realization of the source signal.

Assumption 1. *(Block-stationary assumption) The time domain samples of each source signal $s_i(t)$ in a frame are a realization of a zero mean wide-sense stationary ergodic random process.*

Assumption 2. *(Dynamic signal assumption) The statistics of each source signal $s_i(t)$ varies from frame to frame.*

The above Assumptions 1 and 2 is equivalent to the quasi-stationary modeling commonly used in speech processing. By Assumption 2, we explicitly emphasize the dynamic nature of the signal model which turns out to play a significant role in the frequency domain ICA approach. Strictly speaking, Assumptions 1 and 2 are not satisfied for most real applications but are a reasonable mathematical model for analysis and for garnering insight.

Lemma 1. *[127, Ch.4] Under Assumption 1 and 2, for sufficiently long frame length, $S_i(k,n)$ can be approximated as a zero mean circular complex Gaussian random variable. Also, $S_i(k,n)$ is independent of $S_i(l,n)$ for $k \neq l$.*

This result is a direct consequence of applying the properties of the Fourier transform of a stationary random process [127, Ch.4]. Using Lemma 1, the distribution of $S_i(k,n)$ can be written as,

$$p_{S_i(k,n)}(z;n) = \mathcal{N}(z;0,\gamma_i(k,n)) \tag{5.3}$$

$\mathcal{N}(\cdot;0,\gamma_i(k,n))$ denotes a circular complex Gaussian distribution PDF (probability density function) with mean 0 and variance $\gamma_i(k,n)$. The variance of $S_i(k,n)$, denoted as $\gamma_i(k,n)$, is dependent on the spectra at the n_{th} frame and represents the dynamic nature of the source signal. Defining $\Gamma_i(n) = [\gamma_i(1,n),\gamma_i(2,n),\cdots,\gamma_i(K,n)]^T$ and $\Sigma_i(n) = diag(\Gamma_i(n))$, the distribution of the spectral vector $\underline{S}_i(n) \triangleq [S_i(1,n),...,S_i(K,n)]^T$ for the i_{th} source is given by

$$p_{\underline{S}_i(n)}(\underline{z};n) = \mathcal{N}(\underline{z};0,\Sigma_i(n)) \tag{5.4}$$

For further analytical tractability, an additional assumption on the frames can be made which is described below.

Assumption 3. *(In frame i.i.d. assumption) The time domain samples of each source signal $s_i(t)$ in a frame are a realization of a zero mean i.i.d. random process.*

This assumption is made sparingly. With this additional assumption, $\Gamma_i(n) = K\xi_i(n)\cdot\underline{1}$ (where K is the DFT length, $\xi_i(n)$ is the variance of the samples of the time domain signal $s_i(t)$ in frame n, and $\underline{1}$ is a vector whose elements are all 1s). Consequently, $\Sigma_i(n) = K\xi_i(n)\cdot I$ (I is the identity matrix) and the distribution in (5.4) is parameterized by a single parameter $\xi_i(n)$.

In order to characterize the ensemble of waveforms, additional assumptions are needed on the temporal dynamics $\Gamma_i(n)$. If we assume that $\Gamma_i(n)$ is a deterministic time series, then by Eq.(5.4) we have a nonstationary time series (see section 5.4 for the analysis of online ICA algorithms where $\Gamma_i(n)$ is modeled as a deterministic time series). Since batch ICA algorithms process a collection of frames one can employ time averaging as is done in the context of cyclostationary processes to derive meaningful statistics. However this is cumbersome and instead we model the spectral evolution as a random process. This is a common approach in the development of speech recognition systems. In speech recognition, a hidden Markov model (HMM) is widely used to capture the

time varying statistics of speech. Motivated by this, we associated each frame with a hidden state/latent variable. Since there is a one to one correspondence between the states and the spectral dynamics $\Gamma_i(n)$, assumptions typically on the states/latent variables result in assumptions on the spectral dynamics. Here, as in HMMs, we associate a Markov chain with the spectral dynamics $\Gamma_i(n)$.

Assumption 4. *(variance assumption) The spectral dynamics $\Gamma_i(n)$ is modeled as an irreducible ergodic [149, Ch.15] Markov chain.*

An irreducible ergodic Markov chain possesses an invariant (stationary) probability distribution [149, Ch.15]. Hence under Assumption 4, the state, i.e. $\Gamma_i(n)$, has a stationary distribution which in turn induces a stationary distribution on the observations, i.e. $\underline{S}_i(n)$.

5.2.1 Source Signal's Distribution in One Frequency Bin

In this subsection, the distribution of the i_{th} source signal in the k_{th} frequency bin, i.e. $S_i(k,n)$, is analyzed. Instead of Assumption 4, one can make a simpler and compatible assumption which is listed next.

Assumption 5. *(variance assumption) The spectral dynamics $\Gamma_i(n)$ is a stationary ergodic vector random process.*

Since we are interested in the distribution of $S_i(k,n)$, we need to only concern ourselves with the temporal dynamics $\gamma_i(k,n)$, i.e. the variance of the power spectrum $S_i(k,n)$ at the n_{th} frame and the k_{th} frequency bin. With the above modeling assumption, $\gamma_i(k,n)$ is a stationary ergodic random process. One can then derive the form for the distribution of $S_i(k,n)$.

Theorem 2. *Under the block-stationary assumption (Assumption 1), the dynamic assumption (Assumption 2), and the variance assumption (Assumption 5), the distribution of the i_{th} source signal in the k_{th} frequency bin, i.e. $S_i(k,n)$, is a Gaussian scale mixture.*

Proof. Assumption 5 results in the following consequences; a) the distribution of $\gamma_i(k,n)$ is independent of the frame index n, i.e. $p_{\gamma_i(k,n)}(r) = p_{\gamma_i(k)}(r)$. b) The distribution of

$S_i(k,n)$ in Eq.(5.3) is now interpreted as a conditional distribution conditioned on the spectral parameter $\gamma_i(k,n)$, i.e.

$$p_{S_i(k,n)|\gamma_i(k,n)}(z|r) = \mathcal{N}(z;0,r) \tag{5.5}$$

As a result, the stationary distribution of the spectral component $S_i(k,n)$ is given by averaging over the spectral parameter

$$p_{S_i(k,n)}(z) = E_{\gamma_i(k,n)}\left\{p_{S_i(k,n)|\gamma_i(k,n)}(z|r)\right\} = \int \mathcal{N}(z;0,r)p_{\gamma_i(k)}(r)dr \tag{5.6}$$

Based on the expression for the distribution, the PDF $p_{S_i(k,n)}(z)$ is frame invariant and so n will be dropped for notational simplicity. $p_{S_i(k)}(z)$ will be used instead of $p_{S_i(k,n)}(z)$ and the distribution can be interpreted as a mixture of Gaussians. This is the well known scale mixture of Gaussians model, or referred to as Gaussian scale mixtures (GSM) [150, 151].

□

A similar problem has been considered in image processing [152, 153]. A closely related signal model is the spherical invariant random process (SIRP) [154, 67] where the conditional distribution can be more general zero mean distribution other than Gaussian distribution. In [67], Parra et al. proved (for real random variables) that the kurtosis of the unconditional distribution is larger than the kurtosis of the conditional distribution unless the PDF $p_{\gamma_i(k)}(r)$ is a Dirac delta function (the unconditional/conditional distribution are named as 'long term'/'instantaneous' distribution in [67]). If Assumption 4 is used instead of Assumption 5, then $p_{\gamma_i(k)}(r)$ is a delta function if and only if the Markov chain $\gamma_i(k,n), n = 1..L$ has a single absorbing state [149, Ch.15]. For the GSM (Eq. (5.6)), using the kurtosis definition for complex variables in [60], it can be shown that for the complex case the kurtosis of the PDF $p_{S_i(k)}(z)$ is larger than 0 (the kurtosis of a circular complex Gaussian random variable is 0) unless the PDF $p_{\gamma_i(r)}(r)$ is a Dirac delta function. Hence the random variable $S_i(k)$ is in general super-Gaussian. Note this conclusion is not dependent on the distribution of the time domain source signal $s_i(t)$ in each frame (Assumption 1). Even if the time domain source signal $s_i(t)$ in each frame follows a Gaussian distribution, the source signal in the frequency domain $S_i(k)$ is still super-Gaussian unless the PDF $p_{\gamma_i(r)}(r)$ is a Dirac delta

function. In summary, the non-Gaussianity of the unconditional distribution $p_{S_i(k)}(z)$ comes from the fact that the random process $\gamma_i(k,n), n = 1..L$ is not constant (Assumption 2).

To clarify the modeling, we note that in the literature the quasi-stationary model is usually used with the point of view that the random process is non-stationary over the long run. But here we see that under Assumption 5, the whole process is still a stationary process at the frame level. To avoid confusion and to distinguish the model from non-stationary signals, we use the terminology 'dynamic random process' for describing the signal model which satisfies Assumption 1, 2 and 5. Even though the terminology 'non-stationarity' has been used in these contexts, in Pham [106] and Parra [67] the authors note that, "'what makes the algorithms work' is – strictly speaking – not the nonstationarity itself but rather the property that each realization of the source signals has a time-varying envelope". We will expand on this and show later that actually the source signals should have different time-varying envelopes for the frequency domain ICA approach to be able to separate convolutive mixtures. We also note the ergodic assumption is significant in the sense that it relates the time average to the ensemble average, thus making the discussion in this work meaningful and applicable to one realization.

We discuss a few special examples here. If the time domain source signal $s_i(t)$ is a stationary random process, its time domain variance $\xi_i(n)$ will be constant over n, then in the k_{th} frequency bin, $\gamma_i(k,n)$ is constant over n. This means (the Markov chain $\gamma_i(k,n), n = 1..L$ has a single absorbing state if Assumption 4 is used) the unconditional distribution $p_{\gamma_i(k)}(r)$ is a Dirac delta function, $p_{\gamma_i(k)}(r) = \delta(r-\alpha)$, α is a constant. Consequently, the unconditional distribution $p_{S_i(k)}(z)$ will be a zero mean circular complex Gaussian distribution, $p_{S_i(k)}(z) = \mathcal{N}(z;0,\alpha)$. Our experiments on stationary time domain source signals confirm that in each frequency bin the kurtosis of $p_{S_i(k)}(z)$ is close to 0 regardless of the statistical distribution of the time domain source signal. Since ICA can not separate mixed Gaussian source signals [57], the above discussion clarifies why the conventional frequency domain ICA approach can not separate mixed stationary time domain source signals. If the PDF $p_{\gamma_i(k)}(r)$ is the inverse Gamma distribution (which is the conjugate prior to the Gaussian distribution), then $p_{S_i(k)}(z)$ takes the form of the Student's t-distribution. For the general case where $p_{\gamma_i(k)}(r)$ is an arbitrary distribution over

Table 5.2: The conventional frequency domain IVA approach

(1) Apply the short time Fourier transform (STFT) on each time domain mixture based on consecutive frames. Thus the convolutive mixing problem in the time domain is transformed into an instantaneous mixing problem in each frequency bin.
(2) Use batch algorithm based IVA to separate mixed spectral vectors in the frequency domain and recover the spectrum of each source signal. Permutation problem is solved intrinsically.
(3) Apply inverse DFT on the recovered source spectrum to obtain the time domain signal.

$[0, +\infty)$, the PDF $p_{S_i(k)}(z)$ is a super-Gaussian distribution, which may be approximated by a Laplacian distribution. This provides supports for the use of the super-Gaussian distribution which is commonly used for frequency domain speech modeling [75].

5.2.2 Frequency Domain Spectral Vector's Distribution

In this section we analyze the unconditional distribution of the spectral vector $\underline{S}_i(n)$. This is desired by the analysis of the frequency domain independent vector analysis (IVA) approach. In the frequency domain ICA approach, correcting permutation (the third step in Table 5.1) is critical to the recovery of each source signal's spectrum and thus the time domain source signal. IVA was proposed to circumvent the permutation problem [96, 97, 98, 99, 100, 101, 102, 103]. Instead of solving the mixing problem in each frequency bin independently, IVA considers the frequency domain source signal as a vector source and attempts to resolve the mixing problem in the frequency bins in an integrated manner. The procedure of the conventional frequency domain IVA approach is listed in Table 5.2.

In Hiroe, Kim and Lee et al.'s paper [96, 97, 98, 99], 'spherical distribution' is employed empirically as the distribution for the K-dimensional vector source $\underline{S}_i(n)$. The PDF has the form $p(\underline{z}) = \beta f(\|\underline{z}\|_2)$, where β is a normalization factor, $f(\cdot)$ is an appropriately chosen function depending on the statistical model being employed, e.g. Laplacian distribution, and $\|\underline{z}\|_2 = (\sum_k |z(k)|^2)^{1/2}$ is the 2-norm of the vector source. In

this subsection we develop the vector GSM model which naturally leads to the 'spherical distribution' as a special case for modeling the frequency domain vector source.

Similar to the analysis in the subsection 5.2.1, under Assumption 1, 2, and 5, the distribution of $\Gamma_i(n)$ is independent of the frame index n, i.e. $p_{\Gamma_i(n)}(\underline{r}) = p_{\Gamma_i}(\underline{r})$. The distribution of $\underline{S}_i(n)$ in Eq.(5.4) is now interpreted as a conditional distribution conditioned on the spectral parameter $\Gamma_i(n)$, i.e.

$$p_{\underline{S}_i(n)|\Gamma_i(n)}(\underline{z}|\underline{r}) = \mathcal{N}(\underline{z}; \underline{0}, diag(\underline{r})) \tag{5.7}$$

The PDF $p_{\underline{S}_i(n)}(\underline{z})$ can be derived,

$$p_{\underline{S}_i(n)}(\underline{z}) = E_{\Gamma_i(n)}\left\{p_{\underline{S}_i(n)|\Gamma_i(n))}(\underline{z}|\underline{r})\right\} \tag{5.8a}$$

$$= \int \mathcal{N}(\underline{z}; \underline{0}, diag(\underline{r}))p_{\Gamma_i}(\underline{r})d\underline{r} \tag{5.8b}$$

With the additional in-frame i.i.d. assumption (Assumption 3), $\Gamma_i(n) = K\xi_i(n)\cdot\underline{1}$. Because of the stationarity, the distribution of $\xi_i(n)$ is independent of the frame index n, i.e. $p_{\xi_i(n)}(q) = p_{\xi_i}(q)$. The conditional distribution in Eq.(5.7) simplifies to

$$p_{\underline{S}_i(n)|\xi_i(n)}(\underline{z}|q) = \mathcal{N}(\underline{z}; \underline{0}, Kq \cdot I) \tag{5.9a}$$

$$= \frac{1}{\sqrt{\pi(Kq)^K}} \exp\left\{-\frac{\|\underline{z}\|_2^2}{Kq}\right\} \tag{5.9b}$$

Therefore, the PDF $p_{\underline{S}_i(n)}(\underline{z})$ can be derived,

$$p_{\underline{S}_i(n)}(\underline{z}) = E_{\xi_i(n)}\left\{p_{\underline{S}_i(n)|\xi_i(n)}(\underline{z}|q)\right\} \tag{5.10a}$$

$$= \int \frac{1}{\sqrt{\pi(Kq)^K}} \exp\left\{-\frac{\|\underline{z}\|_2^2}{Kq}\right\} p_{\xi_i}(q)dq \tag{5.10b}$$

The above distribution is a stationary distribution and is not dependent on the frame index n, and therefore it can be written as $p_{\underline{S}_i}(\underline{z})$ by ignoring the frame index n. It is evident the distribution of $\underline{S}_i$ is a function of $\|\underline{z}\|_2$, $p_{\underline{S}_i}(\underline{z}) = \beta f(\|\underline{z}\|_2)$, i.e. it takes a 'spherical distribution' format. This supports the usage of 'spherical distribution' in the IVA algorithms. Eq. (5.8) shows the PDF $p_{\underline{S}_i}(\underline{z})$ can be interpreted as a mixture of multivariate Gaussians. This is the well known scale mixture of Gaussians model, or referred to as Gaussian scale mixtures (GSM) [150, 151, 155].

In summary, we have the following theorem.

Theorem 3. *Under the block-stationary assumption (Assumption 1), the dynamic assumption (Assumption 2), and the variance assumption (Assumption 5), the unconditional/marginal distribution of the spectral vector $\underline{S}_i(n)$ in the frequency domain is a Gaussian scale mixture. With the additional in-frame i.i.d. assumption (Assumption 3), it is a 'spherical distribution'.*

Depending on the distribution of ξ_i, $p_{\underline{S}_i(n)}(\underline{z})$ (Eq. (5.10)) may or may not have a tractable closed-form expression. For instance, if $p_{\xi_i}(q)$ is inverse Gamma distributed (which is the conjugate prior to the Gaussian distribution), then $p_{\underline{S}_i(n)}(\underline{z})$ is the multivariate Student's t-distribution. Palmer also derived a GSM format in [156] based on the point of view that the time domain source signal is 'piecewise stationary' and "each window (the STFT window where the DFT is taken on) will contain a random size segment of a given stationary state". Note that our dynamic random process model (Assumption 1, 2, and 5) is totally different from the so-called 'piecewise stationary' model used in [156].

We now discuss the special case as we did in section 5.2.1. If the time domain source signal $s_i(t)$ is a stationary process, then its variance $\xi_i(n)$ is constant over time. This means, the unconditional distribution $p_{\xi_i}(q)$ is a Dirac delta function, $p_{\xi_i}(q) = \delta(q - \alpha)$, α is a constant. Consequently, the unconditional/marginal distribution $p_{\underline{S}_i(n)}(\underline{z})$ will be a Gaussian distribution, $p_{\underline{S}_i(n)}(\underline{z}) = \mathcal{N}(\underline{z}; 0, K\alpha \cdot I)$. Since Gaussian source signals cannot be separated by independence analysis, the above discussion provides insight into why the conventional frequency domain IVA approach cannot separate convolutively mixed stationary time domain source signals.

5.2.3 Independence between the Scalar Source Signals in One Frequency Bin

The previous subsections derived the unconditional/marginal distribution of a source signal (scalar or vector) in the frequency domain. In the next we analyze the independence between the multiple scalar source signals in one frequency bin by inspecting their unconditional joint distributions. The following new assumptions are to be used.

Assumption 6. *(joint stationary assumption) The spectral dynamics $\Gamma_i(n)$ and $\Gamma_j(n)$, $i \neq j$ are jointly stationary [157, Ch.10] ergodic random processes.*

This allows for obtaining joint stationary distributions without time averaging. Assumption 6 implies the spectral dynamics in the k_{th} frequency bin, $\gamma_i(k,n)$ and $\gamma_j(k,n)$, $i \neq j$ are jointly stationary random processes. Note Assumption 5 is implied by Assumption 6.

Assumption 7. *(bin-wise conditional independence assumption) Under Assumption 1 and 2, the source signals in each frequency bin, $S_i(k,n)$ and $S_j(k,n)$, are independent conditioned on the variances $\gamma_i(k,n)$ and $\gamma_j(k,n)$ $(i \neq j)$.*

This can also be viewed as assuming that conditioned on the state sequence for both the sources, the time domain samples of the sources for the frame are independently generated.

Theorem 4. *Under the block-stationary assumption (Assumption 1), the dynamic assumption (Assumption 2), the joint stationary assumption (Assumption 6), and the bin-wise conditional independence assumption (Assumption 7), the unconditional distribution of the multiple scalar source signals in the k_{th} frequency bin, $S_i(k,n)$ and $S_j(k,n)$, are independent if and only if the variance series $\gamma_i(k,n)$ for the different source signals are independent random processes.*

Proof. Mathematically, Assumption 7 results in

$$p_{S_i(k,n),S_j(k,n)|\gamma_i(k,n),\gamma_j(k,n)}(z_i,z_j|r_i,r_j) = p_{S_i(k,n)|\gamma_i(k,n)}(z_i|r_i)p_{S_i(k,n)|\gamma_i(k,n)}(z_j|r_j) \tag{5.11}$$

Therefore we can derive the unconditional joint distribution for the multiple source signals in the k_{th} frequency bin,

$$p_{S_i(k,n),S_j(k,n)}(z_i,z_j) = \int p_{S_i(k,n),S_j(k,n),\gamma_i(k,n),\gamma_j(k,n)}(z_i,z_j,r_i,r_j)dr_i dr_j \tag{5.12a}$$

$$= \int p_{S_i(k,n),S_j(k,n)|\gamma_i(k,n),\gamma_j(k,n)}(z_i,z_j|r_i,r_j)p_{\gamma_i(k,n),\gamma_j(k,n)}(r_i,r_j)dr_i dr_j \tag{5.12b}$$

$$= \int p_{S_i(k,n)|\gamma_i(k,n)}(z_i|r_i)p_{S_i(k,n)|\gamma_i(k,n)}(z_j|r_j)p_{\gamma_i(k,n),\gamma_j(k,n)}(r_i,r_j)dr_i dr_j \tag{5.12c}$$

If $\gamma_1(\omega, n)$ and $\gamma_2(\omega, n)$ are independent random processes, i.e.

$$p_{\gamma_i(k,n),\gamma_j(k,n)}(r_i, r_j) = p_{\gamma_i(k,n)}(r_i)p_{\gamma_j(k,n)}(r_j) \tag{5.13}$$

Then Eq. (5.12) simplifies to

$$p_{S_i(k,n),S_j(k,n)}(z_i, z_j) = \int p_{S_i(k,n)|\gamma_i(k,n)}(z_i|r_i)p_{\gamma_i(k,n)}(r_i)dr_i \tag{5.14a}$$

$$\cdot \int p_{S_i(k,n)|\gamma_i(k,n)}(z_j|r_j)p_{\gamma_j(k,n)}(r_j)dr_j \tag{5.14b}$$

$$= p_{S_i(k,n)}(z_i)p_{S_j(k,n)}(z_j) \tag{5.14c}$$

By Assumption 6, $S_i(k, n)$ and $S_j(k, n)$ are stationary random processes, hence the notation can be simplified by dropping the frame index n in Eq. (5.14).

$$p_{S_i(k),S_j(k)}(z_i, z_j) = p_{S_i(k)}(z_i)p_{S_j(k)}(z_j) \tag{5.15}$$

Therefore, $S_i(k)$ and $S_j(k)$ are independent.

If $\gamma_i(k, n)$ and $\gamma_j(k, n)$ are not independent[1],

$$p_{S_i(k),S_j(k)}(z_i, z_j) \neq p_{S_i(k)}(z_i)p_{S_j(k)}(z_j) \tag{5.16}$$

and $S_i(k)$ and $S_j(k)$ are not independent.

In conclusion, the unconditional distribution of the multiple source signals in the k_{th} frequency bin are independent if and only if the variance series $\gamma_i(k, n)$ for the different source signals are independent random processes.

□

Under the block-stationary assumption (Assumption 1), the dynamic assumption (Assumption 2), and with the stronger in-frame i.i.d. assumption (Assumption 3), because $\gamma_i(k, n) = K\xi_i(n), \forall k \in \{1, .., K\}$ ($i \in \{1, .., N\}$), we can use the following simpler assumption to substitute Assumption 6 and 7.

Assumption 8. *(joint stationary assumption) The dynamics $\xi_i(n)$ and $\xi_j(n)$, $i \neq j$ are jointly stationary ergodic random processes.*

[1]One exception is that when both $\gamma_i(k, n)$ and $\gamma_j(k, n)$ are constant over n, $S_i(k)$ and $S_j(k)$ will be independent but both are Gaussians. In this case, ICA can not separate the mixed source signals because of the Gaussianity even though they are independent.

Assumption 9. *(time domain independent sources assumption) Under Assumption 1 and 2, in each frame n, the time domain source signals $s_i(t_1)$ is independent of $s_j(t_2)$ conditioned on the variances $\xi_i(n)$ and $\xi_j(n)$ ($\forall t_1, t_2 \in$ frame n, $i \neq j$).*

Therefore we obtain the following corollary.

Corollary 5. *Under the block-stationary assumption (Assumption 1), the dynamic assumption (Assumption 2), the in-frame i.i.d. assumption (Assumption 3), the joint stationary assumption (Assumption 8), and the time domain independent sources assumption (Assumption 9), the unconditional distribution of the multiple source signals in each frequency bin are independent if and only if the variances of the multiple time domain source signals are independent.*

An experiment verifying this corollary is as follows: generate two dynamic random processes with independent time domain white source signals (with Laplacian distribution in any fixed frame), and with the energy contours of the two time domain white source signals being correlated. The mutual information between the two source signals, i.e. $S_i(k, n)$ and $S_j(k, n)$, in each frequency bin is significant indicating that the source signals in each frequency bin are not independent.

Theorem 4 and Corollary 5 illustrate the significance of the dynamic randomness of the time domain source signals in adopting the conventional frequency domain ICA approach (Table 5.1). In summary, if the time domain source signals are dynamic random processes and independent in each frame, and their variance series change independently, then the unconditional joint distribution of the multiple source signals in each frequency bin will be independent and non-Gaussian. Thereby the conventional frequency domain ICA approach can be adopted to separate the convolutive mixed source signals. However, if the variance series $\gamma_i(\omega, n)$ for the different source signals are not independent, it is not clear if the conventional frequency domain ICA approach can be employed to separate the convolutive mixed source signals. Many complex models have been proposed to recover source signals when their variances are not independent [158, 159, 160, 161]. However, they are more complex than instantaneous linear mixing ICA model and are not consistent with the ordinary ICA model anymore.

In section 5.3.1, we will prove that give certain conditions, the conventional frequency domain ICA approach can be employed to separate convolutive mixed source

signals if only the variance random processes $\gamma_i(k,n), n = 1..L$ for the different source signals are not the same in each frequency bin.

5.2.4 Independence between the Frequency Domain Spectral Vectors

Frequency domain IVA approach is supposed to separate the time domain convolutive mixtures based on the independence between the frequency domain spectral vectors $\underline{S}_i(n)$ and $\underline{S}_j(n)$, $i \neq j$. In this subsection we analyze the independence between the multiple frequency domain spectral vectors by inspecting their unconditional joint distributions. This is very much along the same lines as before except the assumptions are made on the spectral vectors. We need the following assumptions for analysis.

Assumption 10. *(conditional independence assumption) Under Assumption 1 and 2, for any fixed frame n, the source signals in the frequency domain* $\underline{S}_i(n)$ *and* $\underline{S}_j(n)$, $i \neq j$, *are independent conditioned on the variances* $\Gamma_i(n)$ *and* $\Gamma_j(n)$.

Considering two time domain source signals $s_i(t)$ and $s_j(t)$ $(i \neq j)$, which maps to the spectral vectors $\underline{S}_i(n)$, and $\underline{S}_j(n)$, $n = 1..L$ respectively in the frequency domain. Under Assumption 6, $\left\{\underline{S}_i(n), \underline{S}_j(n)\right\}, n = 1..L$ is a joint ergodic stationary process. The PDF of the unconditional joint distribution of this stationary process is denoted as $p_{\underline{S}_i(n),\underline{S}_j(n)}(\underline{z}_i, \underline{z}_j)$. Because of the stationarity, the dependence on the frame index n can be dropped and we replace the notation $p_{\underline{S}_i(n),\underline{S}_j(n)}(\underline{z}_i, \underline{z}_j)$ with $p_{\underline{S}_i,\underline{S}_j}(\underline{z}_i, \underline{z}_j)$, where $\underline{S}_i$ can be written as $\underline{S}_i \triangleq [S_i(1), ..., S_i(K)]^T$.

In Theorem 4, we have seen that under certain assumptions $S_i(k)$ is independent of $S_j(k)$ if and only if the the variance series $\gamma_i(k,n), \gamma_j(k,n), n = 1..L$ are independent random processes. Following the same approach as used in proving Theorem 4, we can prove $\underline{S}_i(n)$ is independent of $\underline{S}_j(n)$ if and only if the the variance series $\Gamma_i(n), \Gamma_j(n)$ are independent random processes. Consequently we have the following theorem.

Theorem 6. *Under the block-stationary assumption (Assumption 1), the dynamic assumption (Assumption 2), the joint stationary assumption (Assumption 6), and the conditional independence assumption (Assumption 10), the unconditional distribution of*

the spectral vectors $\underline{S}_i(n)$ *and* $\underline{S}_j(n)$ *are independent if and only if the spectral dynamics* $\Gamma_i(n), \Gamma_j(n)$ *are independent random processes.*

Similar to section 5.2.3, we have the following corollary.

Corollary 7. *Under the block-stationary assumption (Assumption 1), the dynamic assumption (Assumption 2), the in-frame i.i.d. assumption (Assumption 3), the joint stationary assumption (Assumption 8) and the time domain independent sources assumption (Assumption 9), the unconditional distribution of the spectral vectors in the frequency domain are independent if and only if the variances of the time domain source signals are independent.*

5.3 Separability

5.3.1 ICA Batch Algorithm May Separate Variance Correlated GSM Signals

Theorem 4 shows that under certain assumptions the unconditional distribution of the multiple scalar source signals in the k_{th} frequency bin are independent if and only if the variance series $\gamma_i(k,n), n = 1..L$ for the different source signals are independent random processes. Now that raises an interesting question: Can a ICA batch algorithm separate mixed source signals in one frequency bin even if the unconditional distribution of the multiple source signals are not independent? The answer is affirmative as we show in the next.

In Section 5.2.3, in each frequency bin the signal $S_i(k,n)$ was shown to have density which could be modeled as a Gaussian scale mixture. Motivated by these observations, we propose the following signal model.

Assumption 11. *We assume N source signals* $u_i(n)$*,* $i \in \{1,..,N\}$*, where the distribution of each of the time series is a Gaussian scale mixture (GSM).*

A signal with a GSM density can be modeled as follows:

$$u_i(n) = r_i(n)\tilde{u}_i(n), i \in \{1,..,N\} \tag{5.17}$$

Since the following discussion is not dependent on the time variable n, we drop the time symbol n hereafter for notational simplicity. Then Eq. (5.17) *can be written as,*

$$u_i = r_i \tilde{u}_i, i \in \{1, .., N\} \tag{5.18}$$

where r_i is a positive real random variable controlling the variance modulation, $\tilde{u}_i$ is a zero mean unit variance circular complex Gaussian random variable. $\tilde{u}_i$s are assumed to be independent from each other for $i \in \{1, .., N\}$. r_i is assumed to be independent of $\tilde{u}_j$ for all $i, j \in \{1, .., N\}$. r_is however are not necessarily independent of each other.

Under Assumption 11, u_is are variance dependent GSM random variables. Based on the assumption, it is easy to show u_i is zero mean, i.e. $E\{u_i\} = 0$, and u_i is uncorrelated of u_j for $i \neq j$, i.e. $E\{u_i u_j^*\} = 0$. Also, the real and imaginary parts of u_i are uncorrelated and have equal variances. Since $\tilde{u}_i$ is a zero mean unit variance circular complex Gaussian random variable, and $\tilde{u}_i$ is independent from each other for $i \in \{1, .., N\}$, we have

$$E\{\tilde{u}_i \tilde{u}_j^*\} = \begin{cases} 1, & \text{if} \quad i = j \\ 0, & \text{otherwise} \end{cases} \tag{5.19}$$

and

$$E\{\tilde{u}_i \tilde{u}_j\} = 0, \quad \forall i, j \in \{1, .., N\} \tag{5.20}$$

Assumption 11 proposes a general signal model which may fit different applications. Note the signal model in Assumption 11 can be used to generate both temporally dependent[2], or temporally independent stationary random process. The source signals $S_i(k, n), n = 1, .., L$ in each frequency bin k discussed in Section 5.2.3 can be generated by the signal model proposed in Assumption 11. The variance series $\gamma_i(k, n), n = 1..L$ is related to r_i. The condition that r_i and r_j are coherent[3], i.e. $r_i \equiv r_j$, is equivalent to say $\gamma_i(k, n) = \gamma_j(k, n), n = 1..L$, i.e. the time series $\gamma_i(k, n), n = 1..L$ is the same as $\gamma_j(k, n), n = 1..L$.

Without loss of generality, we can assume u_i has unit variance. This is a commonly used assumption in ICA because of the scaling indeterminacy inherent in ICA.

[2]for a temporally dependent time series $u_i(n), n = 1, .., L$, $u_i(n_1)$ is dependent on $u_i(n_2)$, $n_1 \neq n_2$

[3]In this paper, coherence means $r_i \equiv r_j$ almost surely. That is, r_i equals to r_j with probability 1, or in other words, r_i and r_j are allowed to be different on a set of measure zero.

Since $\tilde{u}_i$ is assumed to have unit variance, scaling of u_i just means scaling of r_i. The non-Gaussianity of u_i is not affected if we scale it. The non-Gaussianity of u_i is solely determined by the distribution of r_i.

It is easy to see the source signals u_is ($i \in \{1, .., N\}$) in Assumption 11 are not independent from each other if the r_is ($i \in \{1, .., N\}$) are not independent. A question we now address is: Can ICA algorithms (like FastICA, maximum likelihood ICA, Infomax, etc.) separate the mixed source signals u_is given N instantaneous linear mixtures of u_is? ICA algorithms are designed for independent source separation and since the u_is are not independent, the answer is not clear. However, after a closer examination and some analysis, we show that the answer to the above question is actually 'Yes'.

In Hyvarinen, Karhunen and Oja's book [81] and Comon's work [57], they discussed in detail the connection between non-Gaussianity and ICA. Without non-Gaussianity, ICA would be impossible. By the central limit theorem, the distribution of a sum of two or more independent non-Gaussian random variables tends to be closer to Gaussian than any single one of the original random variables. Hence maximization of non-Gaussianity is motivated as an ICA approach. Kurtosis is a commonly used measure of non-Gaussianity because of its simplicity. Therefore, maximization of Kurtosis is developed as an ICA approach [81, 58, 60].

The kurtosis of a complex random variables y can be defined as [60],

$$kurt\{y\} = E\{|y|^4\} - 2(E\{|y|^2\})^2 - |E\{y^2\}|^2 \tag{5.21}$$

which is simplified to

$$kurt\{y\} = E\{|y|^4\} - 2 \tag{5.22}$$

if the real and imaginary parts of y are uncorrelated and have zero mean and equal variances, and the unit variance constraint ($E\{|y|^2\} = 1$) is enforced. Note that the kurtosis for zero mean unit variance circular complex Gaussian random variable is 0. The Kurtosis maximization criterion can be expressed in the form of the following optimization problem,

$$\max_{\mathbf{w}} kurt\{y\}, \quad \text{s.t.} \quad E\{|y|^2\} = 1 \tag{5.23}$$

In general, maximization of Kurtosis can not separate mixed dependent source signals from their mixtures. However, we will prove it can separate the mixed source

signals u_is that satisfies Assumption 11 if only r_i is not coherent with r_j, i.e. $r_i \not\equiv r_j$, for any $i, j \in \{1, .., N\}, i \neq j$.

Theorem 8. *Assume N source signals* $\mathbf{u} = [u_1, .., u_N]^T$ *which satisfy Assumption 11. Given N instantaneous linear mixtures* x_i, $i \in \{1, .., N\}$, $\mathbf{x} = [x_1, .., x_N]^T$, $\mathbf{x} = A\mathbf{u}$, *where A is a N×N mixing matrix, the kurtosis maximization criterion (Eq.* (5.23)*) can separate the mixed source signals* u_i*s if only* r_i *is not coherent with* r_j*, i.e.* $r_i \not\equiv r_j$*, for any* $i, j \in \{1, .., N\}, i \neq j$.

Proof. Without loss of generality, we assume u_i has unit variance for all $i \in \{1, .., N\}$ and the data $\mathbf{x}$ has been whitened, i.e. $E\{\mathbf{x}\mathbf{x}^H\}$ equals identity matrix and the mixing matrix A is orthogonal.

Let $y = \mathbf{w}^H\mathbf{x} = \mathbf{w}^H A\mathbf{u} = \mathbf{q}^H\mathbf{u}$, where $\mathbf{w}$ is the weight vector used to estimate a source and $\mathbf{q} = A^H\mathbf{w}$. Since u_is are uncorrelated and have unit variance, the constraint $E\{|y|^2\} = 1$ is equivalent to $\|\mathbf{q}\|^2 = 1$. Then the Kurtosis maximization problem (Eq. (5.23)) can be rewritten as

$$\max_{\mathbf{q}} h = E\{|y|^4\} = E\{|\mathbf{q}^H\mathbf{u}|^4\}, \quad \text{s.t.} \quad \|\mathbf{q}\|^2 = 1 \tag{5.24}$$

We use the method of induction in the following proof. First consider the case where $N = 2$, i.e. only two source signals. We have,

$$h = E\{|y|^4\} = E\{|q_1^* u_1 + q_2^* u_2|^4\} \tag{5.25}$$

After some manipulation and under Assumption 11, denoting $p_i = |q_i|^2$, $i \in \{1, .., N\}$, the Kurtosis maximization problem (Eq. (5.24)) can be written as,

$$\max_{\mathbf{p}} h = 2\mathbf{p}^T Q\mathbf{p} \quad \text{s.t.} \quad \sum_i p_i = 1, \quad 0 \leqslant p_i \leqslant 1, \quad i \in \{1, 2\} \tag{5.26}$$

where

$$\mathbf{p} = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix} \quad Q = \begin{bmatrix} E\{r_1^4\} & E\{r_1^2 r_2^2\} \\ E\{r_1^2 r_2^2\} & E\{r_2^4\} \end{bmatrix} \tag{5.27}$$

It is easy to see Q is a positive definite matrix if the random variables r_1, r_2 are not coherent, i.e. $r_1 \not\equiv r_2$. Therefore, the function h is a convex function. Consequently, the Kurtosis maximization problem (Eq. (5.24)) is expressed as maximizing a convex

function on a simplex set (convex set). The maximum of such a problem is reached on the boundaries of the convex set [162]. Therefore, the solution for the Kurtosis maximization problem (Eq. (5.24)) is $\mathbf{p} = [p_1, p_2]^T = [1, 0]^T$ if $E\{r_1^4\} > E\{r_2^4\}$, and the solution is $\mathbf{p} = [0, 1]^T$ if $E\{r_1^4\} < E\{r_2^4\}$. In the case where $E\{r_1^4\} = E\{r_2^4\}$, the optimization may either converge to $\mathbf{p} = [1, 0]^T$ or $\mathbf{p} = [0, 1]^T$ depending on initialization. After the Kurtosis maximization problem (Eq. (5.23)) yields an estimate of the first source signal u_1 (assuming $E\{r_1^4\} > E\{r_2^4\}$), the other source signal can be obtained by utilizing the fact that all source signals u_is are uncorrelated. Adding the uncorrelatedness constraint $\mathbf{w_i}^H\mathbf{w_j} = 0$ for $i \neq j$, the source signals can be estimated in the deflationary (sequential) or symmetric manner [81, 58, 60].

If the random variables r_1, r_2 are coherent, i.e. $r_1 \equiv r_2$, then the Q matrix is rank deficient and the objective function $h = 2E\{r_1^4\} = 2E\{r_2^4\}$ which is not dependent on $\mathbf{p} = [p_1, p_2]^T$ and the Kurtosis maximization criterion (Eq. (5.24)) can not separate the mixed source signals u_is.

Using the method of induction, suppose the theorem is satisfied for N source signals, i.e. we are able to estimate one source using a linear combining set of weights in the presence of N sources. We now show the same is possible in the case of $N + 1$ source signals. The objective function is,

$$h = E\left\{|y|^4\right\} = E\left\{\left|\sum_{i=1}^{N+1} q_i^* u_i\right|^4\right\} \tag{5.28}$$

After some tedious manipulation and under Assumption 11, denoting $p_i = |q_i|^2$, $i \in \{1, .., N + 1\}$, the Kurtosis maximization problem (Eq. (5.24)) can be written as,

$$\max_{\mathbf{p}} h = 2\mathbf{p}^T Q\mathbf{p} \quad \text{s.t.} \quad \sum_{i=1}^{N+1} p_i = 1, \quad 0 \leqslant p_i \leqslant 1, \quad i \in \{1, .., N + 1\} \tag{5.29}$$

where $\mathbf{p} = [p_1, .., p_{N+1}]^T$ and Q is a $(N + 1) \times (N + 1)$ matrix whose element on the i^{th} row and j^{th} column is $E\left\{r_i^2 r_j^2\right\}$. Since for any non-zero vector $\mathbf{v} = [v_1, .., v_{N+1}]^T$, $\mathbf{v}^T Q\mathbf{v} = E\left\{\left(\sum_{i=1}^{N+1} v_i r_i^2\right)^2\right\} > 0$ if random variable r_i is not coherent with r_j, so Q is a positive definite matrix. Thus the function h is a convex function. Consequently, the Kurtosis maximization problem (Eq. (5.24)) is maximizing a convex function on a simplex set (convex set) [162]. The maximum of such a problem is reached on the

boundaries of the convex set. Without loss of generality, suppose the maximum of h is reached on the boundary of the convex set where $\sum_{i=1}^{N} p_i = 1$ and $p_{N+1} = 0$, the problem reduces to the degenerate Kurtosis maximization problem with a mixture of N source signals. Therefore, the proposed theorem also holds for a mixture with $N + 1$ source signals.

□

Since Kurtosis has the disadvantage of being sensitive to outliers, neg-entropy is proposed as a robust non-Gaussianity measure [81]. Neg-entropy is motivated by a fundamental theorem from information theory: a Guassian random variable has the largest entropy among all random variables with the same variance. Thereby, maximizing neg-entropy will select the most non-Gaussian direction in the mixtures and thus recover one source signal. In practice, neg-entropy is usually approximated by a nonlinear function $E\{G(y)\}$ [81]. For complex signals, it takes the format $E\{G(|y|^2)\}$ [60] (where $G : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}$ is a smooth even function) so that the maximization of the objective function is a well defined problem. For example, if $G(y) = y^2$, then $E\{G(|y|^2)\} = E\{|y|^4\}$ which is the kurtosis measure.

The FastICA algorithm is developed based on neg-entropy maximization [81, 58, 60]. Assume N source signals, $\mathbf{x}$ is N-dimensional mixtures, $\mathbf{w}$ is N-dimensional complex weight vector. Define $J(\mathbf{w}) = E\{G(|\mathbf{w}^H\mathbf{x}|^2)\}$, the neg-entropy criterion results in the following optimization problem,

$$\min_{\mathbf{w}} J(\mathbf{w}), \quad s.t. \quad E\{|\mathbf{w}^H\mathbf{x}|^2\} = 1 \tag{5.30}$$

More than one source signals can be obtained by utilizing the fact that all source signals u_is are uncorrelated. Adding the uncorrelatedness constraint $\mathbf{w_i}^H\mathbf{w_j} = 0$ for $i \neq j$, the source signals can be estimated in the deflationary (sequential) or symmetric manner [81, 58, 60].

The FastICA algorithm is designed for independent signal separation. However, we will show that it can separate the mixed source signals u_is that satisfy Assumption 11. Our approach is similar to the stability analysis in [81, 58, 60].

Theorem 9. *Assume N source signals which satisfy Assumption 11 and have unit variance. Given N prewhitened mixtures* x_i, $i \in \{1, .., N\}$, $\mathbf{x} = A\mathbf{u}$, *where* $\mathbf{x} = [x_1, .., x_N]^T$,

$\mathbf{u} = [u_1, .., u_N]^T$, *and A is a* $N \times N$ *orthonormal mixing matrix, the local maxima (resp. minima) of* $E\{G(|\mathbf{w}^H\mathbf{x}|^2)\}$ *under the constraint* $E\{|\mathbf{w}^H\mathbf{x}|^2\} = 1$ *include those rows of* A^{-1} *such that the corresponding independent components* u_k *satisfy*

$$E\left\{|u_j|^2 g(|u_k|^2) + |u_j|^2|u_k|^2 g'(|u_k|^2) - |u_k|^2 g(|u_k|^2)\right\} < 0 \quad (> 0, resp.), \quad \forall j \neq k \tag{5.31}$$

where $g(\cdot)$ *is the derivative of* $G(\cdot)$ *and* $g'(\cdot)$ *is the derivative of* $g(\cdot)$.

Proof. Define $J(\mathbf{w}) = E\{G(|\mathbf{w}^H\mathbf{x}|^2)\}$ and denote $\mathbf{q} = A^H\mathbf{w}$. Then we have $J(\mathbf{q}) = E\{G(|\mathbf{q}^H\mathbf{u}|^2)\}$. If $\mathbf{w}^H$ coincides with one row of A^{-1}, then $\mathbf{q} = [0, .., 0, v, 0, .., 0]^T$, where v is a complex value with unit modulus ($|v| = 1$), a consequence of the phase ambiguity inherent in complex ICA problems even though the variance of the estimated source is constrained to be one. Then the stability of such $\mathbf{q}$ is analyzed by Taylor expansion.

We use the derivative with respect to the real and imaginary components of $\mathbf{q}$ for the stability analysis. Suppose $q_i = q_{Ri} + Iq_{Ii}$, where q_{Ri} and q_{Ii} denotes the real and imaginary parts of q_i respectively, and I denotes the imaginary symbol. The gradient of $J(\mathbf{q})$ is,

$$\nabla J(\mathbf{q}) = \begin{pmatrix} \frac{\partial}{\partial q_{R1}} \\ \frac{\partial}{\partial q_{I1}} \\ \vdots \\ \frac{\partial}{\partial q_{RN}} \\ \frac{\partial}{\partial q_{RN}} \end{pmatrix} J(\mathbf{q}) = 2 \begin{pmatrix} E\left\{\Re\left\{u_1(\mathbf{q}^H\mathbf{u})^*\right\} g(|\mathbf{q}^H\mathbf{u}|^2)\right\} \\ E\left\{\Im\left\{u_1(\mathbf{q}^H\mathbf{u})^*\right\} g(|\mathbf{q}^H\mathbf{u}|^2)\right\} \\ \vdots \\ E\left\{\Re\left\{u_N(\mathbf{q}^H\mathbf{u})^*\right\} g(|\mathbf{q}^H\mathbf{u}|^2)\right\} \\ E\left\{\Im\left\{u_N(\mathbf{q}^H\mathbf{u})^*\right\} g(|\mathbf{q}^H\mathbf{u}|^2)\right\} \end{pmatrix} \tag{5.32}$$

where $\Re\{\cdot\}$ and $\Im\{\cdot\}$ denotes real and imaginary parts respectively. Define

$$\begin{aligned} J_{Ri} &= E\left\{\Re\left\{u_i(\mathbf{q}^H\mathbf{u})^*\right\} g(|\mathbf{q}^H\mathbf{u}|^2)\right\} \\ J_{Ii} &= E\left\{\Im\left\{u_i(\mathbf{q}^H\mathbf{u})^*\right\} g(|\mathbf{q}^H\mathbf{u}|^2)\right\} \end{aligned} \tag{5.33}$$

The Hessian of $J(\mathbf{q})$ is defined as,

$$\nabla^2 J(\mathbf{q}) = 2 \begin{pmatrix} \frac{\partial J_{R1}}{\partial q_{R1}} & \frac{\partial J_{R1}}{\partial q_{I1}} & \cdots & \frac{\partial J_{R1}}{\partial q_{RN}} & \frac{\partial J_{R1}}{\partial q_{IN}} \\ \frac{\partial J_{I1}}{\partial q_{R1}} & \frac{\partial J_{I1}}{\partial q_{I1}} & \cdots & \frac{\partial J_{I1}}{\partial q_{RN}} & \frac{\partial J_{I1}}{\partial q_{IN}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial J_{RN}}{\partial q_{R1}} & \frac{\partial J_{RN}}{\partial q_{I1}} & \cdots & \frac{\partial J_{RN}}{\partial q_{RN}} & \frac{\partial J_{RN}}{\partial q_{IN}} \\ \frac{\partial J_{IN}}{\partial q_{R1}} & \frac{\partial J_{IN}}{\partial q_{I1}} & \cdots & \frac{\partial J_{IN}}{\partial q_{RN}} & \frac{\partial J_{IN}}{\partial q_{IN}} \end{pmatrix} \tag{5.34}$$

Without loss of generality, we analyze the stability of the point $\mathbf{q} = v\mathbf{e}_1$, where $\mathbf{e}_1 = [1, 0, .., 0]^T$, and $|v| = 1$. $v = v_R + Iv_I$, v_R and v_I are real and imaginary parts of v respectively. Using Assumption 11 and the constraint that u_i has unit variance, after some tedious manipulations, we obtain,

$$\nabla J(\mathbf{q})|_{\mathbf{q}=v\mathbf{e}_1} = 2\begin{pmatrix} v_R E\left\{|u_1|^2 g(|u_1|^2)\right\} \\ v_I E\left\{|u_1|^2 g(|u_1|^2)\right\} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{5.35}$$

$$\nabla^2 J(\mathbf{q})|_{\mathbf{q}=v\mathbf{e}_1} = 2$$

$$\begin{pmatrix} G1 & 2v_R v_I E\left\{|u_1|^4 g'(|u_1|^2)\right\} & 0 & 0 & \cdots & 0 & 0 \\ 2v_R v_I E\left\{|u_1|^4 g'(|u_1|^2)\right\} & G2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \alpha_2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \alpha_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \alpha_N & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 & \alpha_N \end{pmatrix} \tag{5.36}$$

where

$$G1 = E\left\{|u_1|^2 g(|u_1|^2) + 2v_R^2 |u_1|^4 g'(|u_1|^2)\right\}$$

$$G2 = E\left\{|u_1|^2 g(|u_1|^2) + 2v_I^2 |u_1|^4 g'(|u_1|^2)\right\}$$

$$\alpha_j = E\left\{|u_j|^2 g(|u_1|^2) + |u_j|^2 |u_1|^2 g'(|u_1|^2)\right\}$$

Consider a small perturbation $\varepsilon = [\varepsilon_{R1}, \varepsilon_{I1}, .., \varepsilon_{RN}, \varepsilon_{IN}]^T$ at the point $\mathbf{q} = v\mathbf{e}_1$, ε_{Ri} and ε_{Ii} represent the real and imaginary parts of ε_i (which is complex) respectively.

Take the Taylor expansion,

$$
\begin{aligned}
&J(v\mathbf{e}_1 + \varepsilon) \\
&= J(v\mathbf{e}_1) + \varepsilon^T \nabla J(v\mathbf{e}_1) + \frac{1}{2}\varepsilon^T \nabla^2 J(v\mathbf{e}_1)\varepsilon + o(\|\varepsilon\|^2) \\
&= J(v\mathbf{e}_1) + 2(\varepsilon_{R1}v_R + \varepsilon_{I1}v_I)E\left\{|u_1|^2 g(|u_1|^2)\right\} + \varepsilon_{R1}^2 E\left\{|u_1|^2 g(|u_1|^2) + 2v_R^2|u_1|^4 g'(|u_1|^2)\right\} \\
&+ 4v_R v_I \varepsilon_{R1}\varepsilon_{I1} E\left\{|u_1|^4 g'(|u_1|^2)\right\} + \varepsilon_{I1}^2 E\left\{|u_1|^2 g(|u_1|^2) + 2v_I^2|u_1|^4 g'(|u_1|^2)\right\} \\
&+ \sum_{j>1} E\left\{|u_j|^2 g(|u_1|^2) + |u_j|^2|u_1|^2 g'(|u_1|^2)\right\}(\varepsilon_{Rj}^2 + \varepsilon_{Ij}^2) + o(\|\varepsilon\|^2)
\end{aligned}
\tag{5.37}
$$

Because of the constraint that each estimated source has unit variance $E\{|\mathbf{w}^H\mathbf{x}|^2\} = 1$, we can derive $\|\mathbf{q}\|^2 = 1$ (since the source signals u_is are uncorrelated and A is orthogonal). Hence $\|v\mathbf{e}_1 + \varepsilon\|^2 = 1$. Then we can get,

$$
2(\varepsilon_{R1}v_R + \varepsilon_{I1}v_I) = -\sum_{i=1}^{N}(\varepsilon_{Ri}^2 + \varepsilon_{Ii}^2) \tag{5.38}
$$

so $(\varepsilon_{R1}v_R + \varepsilon_{I1}v_I)^2$ is $o(\|\varepsilon\|^2)$. Therefore we can get,

$$
\begin{aligned}
&J(v\mathbf{e}_1 + \varepsilon) \\
&= J(v\mathbf{e}_1) + \sum_{j>1} E\left\{|u_j|^2 g(|u_1|^2) + |u_j|^2|u_1|^2 g'(|u_1|^2) - |u_1|^2 g(|u_1|^2)\right\}(\varepsilon_{Rj}^2 + \varepsilon_{Ij}^2) + o(\|\varepsilon\|^2)
\end{aligned}
\tag{5.39}
$$

If $E\left\{|u_j|^2 g(|u_1|^2) + |u_j|^2|u_1|^2 g'(|u_1|^2) - |u_1|^2 g(|u_1|^2)\right\} < 0 \quad (> 0, resp.)$, $\forall j > 1$, then $\mathbf{q} = v\mathbf{e}_1$ is a maximum (resp. minimum) point.

□

Remark 1: If we further assume that r_i is independent of r_j, $\forall i, j \in \{1, .., N\}$, $i \neq j$, in the signal model proposed in Assumption 11, then u_i is independent of u_j, $\forall i, j \in \{1, .., N\}$, $i \neq j$. Therefore,

$$
\begin{aligned}
&E\left\{|u_j|^2 g(|u_k|^2) + |u_j|^2|u_k|^2 g'(|u_k|^2) - |u_k|^2 g(|u_k|^2)\right\} && \text{(5.40a)} \\
=&E\left\{|u_j|^2\right\} E\left\{g(|u_k|^2)\right\} + E\left\{|u_j|^2\right\} E\left\{|u_k|^2 g'(|u_k|^2)\right\} - E\left\{|u_k|^2 g(|u_k|^2)\right\} && \text{(5.40b)} \\
=&E\left\{g(|u_k|^2) + |u_k|^2 g'(|u_k|^2) - |u_k|^2 g(|u_k|^2)\right\} && \text{(5.40c)}
\end{aligned}
$$

This is exactly the same stability condition derived in [60] when each source is assumed to be independent.

Remark 2: For the special case where $G(y) = \frac{1}{2}y^2$, then $g(y) = y$ and $g'(y) = 1$. $J(y) = E\{G(|y|^2)\}$ is the kurtosis. Suppose $E\{r_1^4\} \geq E\{r_j^4\}$ for $j \neq 1$, then we have (recall $u_i = r_i\tilde{u}_i$, $E\{|u_i|^4\} = 2E\{r_i^4\}$),

$$E\left\{|u_j|^2 g(|u_1|^2) + |u_j|^2|u_1|^2 g'(|u_1|^2) - |u_1|^2 g(|u_1|^2)\right\} = 2(E\{r_j^2 r_1^2\} - E\{r_1^4\}) \tag{5.41a}$$

$$\leq 2\sqrt{E\{r_1^4\}}(E\{r_j^4\} - E\{r_1^4\}) \tag{5.41b}$$

$$\leq 0 \tag{5.41c}$$

The first inequality in the above equation follows from Cauchy-Schwartz inequality. Thereby $\mathbf{q} = v\mathbf{e}_1$ is a stable stationary point of the optimization problem. This conclusion is consistent with the result presented in Theorem 8.

Theorem 8 and 9 illustrate that the non-Gaussianity maximization criterion (maximization of neg-entropy) can separate the mixed source signals u_is modeled in Assumption 11 if only the variance random variables r_is are not coherent, i.e. $r_i \not\equiv r_j$, $\forall i \neq j$. Since maximization of neg-entropy is equivalent to minimization of mutual information under the constraint that the estimates are uncorrelated [81], and because the source signals u_is are uncorrelated, it can be concluded that any ICA algorithm based on mutual information minimization is able to separate the mixed source signals u_is modeled in Assumption 11 if only $r_i \not\equiv r_j$, $\forall i \neq j$.

Summarizing the results from Section 5.2.1, 5.2.3, and 5.3.1, the following theorem can be drawn.

Theorem 10. *Under the block-stationary assumption (Assumption 1), the dynamic assumption (Assumption 2), the joint stationary assumption (Assumption 6), and the binwise conditional independence assumption (Assumption 7), in the conventional frequency domain ICA approach (table 5.1), the source signals in the k_{th} frequency bin can be separated by the ICA batch algorithms if only the variance series $\gamma_i(k,n)$ is different from $\gamma_j(k,n)$ almost surely ($i \neq j$).*

5.3.2 IVA Separability and Discussion

The stability analysis for IVA is much more complicated than that for ICA because of the vector variables involved. We analyze the stability conditions for IVA in

section 6.2. Unfortunately, the mathematical formulas of the stability conditions are so complicated that they do not give us any insight into the problem. Therefore we can not discuss the separability of the frequency domain spectral vectors by IVA algorithms as we do in ICA analysis (section 5.3.1). Instead, we will try to provide some intuitive arguments.

In the frequency domain IVA algorithm, the separation of the mixed source signals in the k_{th} frequency bin, i.e. the updating of the de-mixing weight vectors in the k_{th} frequency bin, is similar to that of the ICA algorithm in the k_{th} frequency bin. The following is the updating rule of the de-mixing weight vectors $\mathbf{W}(k)$ in the k_{th} frequency bin for the frequency domain IVA algorithm [96].

$$\Delta\mathbf{W}(k) = \mu\left[I - \underline{\psi}_k^*(\mathbf{Y}(n))\mathbf{Y}(k,n)^H\right]\mathbf{W}(k) \tag{5.42}$$

where $\mathbf{Y}(t)$ is the vector outputs of the de-mixing system. The updating of all the de-mixing weight vectors in different frequency bins are coupled through $\mathbf{Y}(t)$. However, thinking in a sequential, suboptimal optimization approach, where the de-mixing weight vectors in all other frequency bins have been fixed at the true separating solution and the permutations have been aligned correctly except for the k_{th} frequency bin, considering the updating of the de-mixing weight vectors in the k_{th} frequency bin. The updating is similar to that of ICA algorithm in the k_{th} frequency bin except the objective functions are different, therefore the stability analysis in Theorem 9 should also apply here. By the same argument as in section 5.3.1, the source signals in each frequency bin may be separated if the variance series of different source signals in that frequency bin are different.

Note a vector source signal's components in different frequency bins may be dependent on each other. In fact, this is exactly where the power of IVA comes from and the reason why IVA can solve the permutation problem. IVA use the dependency between a vector source signal's components in different frequency bins to solve the permutation problem. If there is no dependency between the vector source signal's components in different frequency bins, IVA degrades to the frequency domain ICA approach and can not solve the permutation problem [96]. We can prove that the dependency between a vector source signal's components in different frequency bins, e.g. $S_i(k,n)$ and $S_i(m,n)$, $k \neq m$, actually comes from the dependency between the random pro-

cesses $\gamma_i(k,n)$ and $\gamma_i(m,n), n = 1..L$ following the recipe used in proving Theorem 4. In speech processing, it is observed there are strong correlations between the variance series $\gamma_i(k,n)$ for the neighboring frequency bins, and weak correlations for the frequency bins that are far apart. A a consequence, there are strong dependencies for a spectral vector's components in the neighboring frequency bins and almost no dependency for those in the frequency bins that are far apart [101]. In summary, the advantages of IVA arise from the dependency between the variance series $\gamma_i(k,n)$ and $\gamma_i(m,n), n = 1..L$, $k, m \in \{1, .., K\}, k \neq m$.

5.4 ICA or IVA Online (stochastic) Algorithms

In the previous sections, we discussed the applicability of the frequency domain ICA (or IVA) approach. We verified under certain conditions the frequency domain ICA (or IVA) approach can be applied to separate convolutively mixed source signals. Actually, what we really discussed is the applicability of the batch algorithm based ICA (or IVA) (the second step in Table 5.1 or Table 5.2). We have taken the point of view of 'dynamic random process' and made the Assumption 1, 2 and 5. Under those assumptions, the whole process is still a stationary process at the frame level (page 110). The dynamic random process is stationary in the ensemble sense while a given realization may in an engineering sense exhibit 'non-stationarity' (page 104). However, if instead of Assumption 5, we make the following assumption,

Assumption 12. *(deterministic variance assumption) The spectral dynamics* $\Gamma_i(n)$ *is a deterministic time series.*

Then under Assumption 1, 2 and 12, the signal process in the frequency domain $S_i(k,n)$ is actually a non-stationary process. Under such circumstances, batch algorithm based ICA (or IVA) is not appropriate to be applied anymore (since batch algorithm is only reasonable to be applied on stationary signal processing, otherwise you are implicitly inducing a stationary process from the underlying non-stationary process by time averaging). Instead, online algorithm could be used on non-stationary signal processing. Therefore, in this section, we discuss the application of the online ICA (or IVA)

algorithm in the second step of Table 5.1 or Table 5.2 under the deterministic variance assumption (Assumption 12).

5.4.1 On Applicability of ICA Online (Stochastic) Algorithm

Most of the previously mentioned ICA batch algorithms have their corresponding online (stochastic) version [58, 63, 62]. Moreover, specific online ICA algorithms have been developed for separating non-stationary source signals [163, 164, 165, 76, 106, 108].

We prove the following theorem 11 under Assumption 12. Note if the spectral dynamics $\Gamma_i(n)$ is assumed to be a deterministic time series, then the variance series $\xi_i(n), n = 1..L$ is implicitly assumed to be a deterministic time series. In [163], it is proved that for its proposed stochastic ICA algorithm based on second order statistics to be able to recover the source signals, $\gamma_i(k,n)/\gamma_j(k,n)(n = 1..L, i, j \in \{1,..,N\}, i \neq j)$ should not be constant over n. Theorem 11 in the following uses a more general proof than the proof in [163] which is only based on second order statistics.

Theorem 11. *Under the block-stationary assumption (Assumption 1), the dynamic assumption (Assumption 2), the deterministic variance assumption (Assumption 12), and the time domain independent sources assumption (Assumption 9), in the k_{th} frequency bin, if the variances satisfy $\gamma_i(k,n) = \alpha\gamma_j(k,n), n = 1..L$, $(i, j \in \{1,..,N\}, i \neq j)$, where α is a constant, then the online (stochastic) algorithm based ICA is not eligible to recover the source signals $S_i(k,n), S_j(k,n), n = 1..L$.*

Proof. Without loss of generality, we assume only two time domain source signals $s_1(t)$ and $s_2(t)$. Fixing the frame index n, $S_i(k,n), i \in \{1,2\}$ follows Gaussian distribution by Lemma 1 (Eq. (5.3)).

$$p_{S_1(k,n)}(z;n) = \mathcal{N}(z;0,\gamma_1(k,n)) \tag{5.43a}$$

$$p_{S_2(k,n)}(z;n) = \mathcal{N}(z;0,\gamma_2(k,n)) \tag{5.43b}$$

$$\tag{5.43c}$$

Since the time domain source signal $s_1(t)$ is independent of $s_2(t)$ at the frame n, and $S_1(k,n), S_2(k,n)$ are functions of $s_1(t)$ and $s_2(t)$ respectively at frame n, then $S_1(k,n)$ should be independent of $S_2(k,n)$ for any fixed $n \in \{1,..,L\}$.

Supposing the mixing matrix in the k_{th} frequency bin is $A(k)$. The mixing system can be expressed as,

$$\underline{x}(k,n) = A(k)\underline{S}(k,n) \tag{5.44}$$

$$\underline{x}(k,n) = \begin{bmatrix} x_1(k,n) \\ x_2(k,n) \end{bmatrix} \quad \underline{S}(k,n) = \begin{bmatrix} S_1(k,n) \\ S_2(k,n) \end{bmatrix}$$

where $\underline{x}(k,n)$ is the mixed signals.

Fixing $n = n_1$. Supposing matrix $Q(k)$ is the whitening matrix for the system at the time point n_1.

$$\underline{z}(k,n_1) = Q(k)\underline{x}(k,n_1) \tag{5.45}$$

$$\underline{z}(k,n) = \begin{bmatrix} z_1(k,n) \\ z_2(k,n) \end{bmatrix}$$

Then $E\{\underline{z}(k,n_1)\underline{z}^H(k,n_1)\} = I$ (I is the identity matrix) by the definition of whitening matrix. And we have,

$$E\{\underline{z}(k,n_1)\underline{z}^H(k,n_1)\} = Q(k)A(k)E\{\underline{S}(k,n_1)\underline{S}^H(k,n_1)\}A(k)^H Q(k)^H \tag{5.46a}$$

$$= Q(k)A(k)\begin{bmatrix} \gamma_1(k,n_1) & 0 \\ 0 & \gamma_2(k,n_1) \end{bmatrix} A(k)^H Q(k)^H \tag{5.46b}$$

$$= \gamma_1(k,n_1)Q(k)A(k)\begin{bmatrix} \alpha & 0 \\ 0 & 1 \end{bmatrix} A(k)^H Q(k)^H \tag{5.46c}$$

So we obtain,

$$\gamma_1(k,n_1)Q(k)A(k)\begin{bmatrix} \alpha & 0 \\ 0 & 1 \end{bmatrix} A(k)^H Q(k)^H = I \tag{5.47}$$

Because $\underline{z}(k,n)$ are linear transformation of Gaussians, $\underline{z}(k,n)$ will be Gaussians. Since $E\{\underline{z}(k,n_1)\underline{z}^H(k,n_1)\} = I$, $z_1(k,n_1)$ is independent of $z_2(k,n_1)$.

For any fixed time $n = n_2$ ($n_2 \neq n_1$), define

$$\underline{z}(k,n_2) = Q(k)\underline{x}(k,n_2) \tag{5.48}$$

We can derive,

$$E\{\underline{z}(k,n_2)\underline{z}^H(k,n_2)\} = Q(k)A(k)E\{\underline{S}(k,n_2)\underline{S}^H(k,n_2)\}A(k)^HQ(k)^H \tag{5.49a}$$

$$= Q(k)A(k)\begin{bmatrix}\gamma_1(k,n_2) & 0\\ 0 & \gamma_2(k,n_2)\end{bmatrix}A(k)^HQ(k)^H \tag{5.49b}$$

$$= \gamma_1(k,n_2)Q(k)A(k)\begin{bmatrix}\alpha & 0\\ 0 & 1\end{bmatrix}A(k)^HQ(k)^H \tag{5.49c}$$

$$= \frac{\gamma_1(k,n_2)}{\gamma_1(k,n_1)}\cdot I \tag{5.49d}$$

Since $\underline{z}(k,n)$ are Gaussian, $z_1(k,n_2), z_2(k,n_2)$ are independent. This is true for any fixed n_2, so $z_1(k,n), z_2(k,n)$ are independent for any n. It is easy to see that for any orthogonal matrix U, elements of $U\underline{z}(k,n)$ are independent (for fixed n). This means there are multiple random processes which are independent based on different demixing matrices. Therefore ICA can not be employed to recover the original source signals.

□

Under the block-stationary assumption (Assumption 1), the dynamic assumption (Assumption 2), and with the stronger in-frame i.i.d. assumption (Assumption 3), because $\gamma_i(k,n) = K\xi_i(n), \forall k \in \{1,..,K\}$ ($i \in \{1,..,N\}$), we obtain the following corollary.

Corollary 12. *Under the block-stationary assumption (Assumption 1), the dynamic assumption (Assumption 2), the in-frame i.i.d. assumption (Assumption 3), the deterministic variance assumption (Assumption 12), and the time domain independent sources assumption (Assumption 9), if the variances of the time domain source signals satisfy $\xi_i(n) = \alpha\xi_j(n), i \neq j$, then in the frequency domain ICA approach, online algorithm based ICA is not eligible to recover the source signals in each frequency bin.*

5.4.2 On Applicability of IVA Online (Stochastic) Algorithm

In this subsection, we discuss the applicability of the online IVA algorithm in the second step of Table 5.2 under the deterministic variance assumption (Assumption 12). Most of the previously mentioned batch algorithm based IVA algorithms have their corresponding online (stochastic) version [96, 97, 98, 99]. We prove the following theorem under Assumption 12.

Theorem 13. *Under the block-stationary assumption (Assumption 1), the dynamic assumption (Assumption 2), the deterministic variance assumption (Assumption 12), and the time domain independent sources assumption (Assumption 9), in the frequency domain, if the variances satisfy* $\gamma_i(k,n) = \alpha(k)\gamma_j(k,n)$, $n = 1..L$, $i, j \in \{1,..,N\}$, $i \neq j$, *where* $\alpha(k)$ *is a constant, then the online algorithm based IVA is not eligible to recover the source signals* $\underline{S}_i(n), \underline{S}_j(n), n = 1..L$, *where* $\underline{S}_i(n) = [S_i(1,n), ..., S_i(K,n)]^T$ *(L is the total number of frames, K is the DFT length, N is the number of sources).*

Proof. Without loss of generality, we assume only two time domain source signals $s_1(t)$ and $s_2(t)$. In the proof of Theorem 11, we have seen $z_1(k,n)$ is independent of $z_2(k,n)$ under the condition $\gamma_1(k,n) = \alpha(k)\gamma_2(k,n)$ for any fixed n, where $\underline{z}(k,n) = [z_1(k,n), z_2(k,n)]^T$ is the whitened mixture in the k_{th} frequency bin at frame n.

Fixing n, since $S_i(k,n)$ is functions of $s_i(t)$, $i \in \{1,2\}$, and $s_1(t)$ is independent of $s_2(t)$ in frame n, $S_1(k,n)$ is independent of $S_2(m,n)$, $\forall k, m$. Under the block-stationary assumption, $S_i(k,n)$ is independent of $S_i(m,n)$ by Lemma 1, $\forall k \neq m$. Because $z_1(k,n)$ is a linear combination of $S_1(k,n)$ and $S_2(k,n)$ ($i \in \{1,2\}, \forall k$), $z_1(k,n)$ is independent of $z_2(m,n)$, $\forall k \neq m$. Therefore, $\underline{z}_1(n) \triangleq [z_1(1,n), ..., z_1(K,n)]^T$ is independent of $\underline{z}_2(n) \triangleq [z_2(1,n), ..., z_2(K,n)]^T$.

For any orthogonal matrix $U(k)$, we construct,

$$\underline{y}(k,n) = U(k)\underline{z}(k,n) \tag{5.50}$$

$$\underline{y}(k,n) = \begin{bmatrix} y_1(k,n) \\ y_2(k,n) \end{bmatrix}$$

Since $z_i(k,n)$ are Gaussians, it is easy to see $\underline{y}_i(n) \triangleq [y_i(1,n), ..., y_i(K,n)]^T, i \in \{1,2\}$ are Gaussians and $\underline{y}_1(n)$ is independent of $\underline{y}_2(n)$. This means there are multiple vector random processes which are independent based on different demixing matrices. Therefore IVA is not capable to recover the original frequency domain vector source signals. □

Under the block-stationary assumption (Assumption 1), the dynamic assumption (Assumption 2), and with the stronger in-frame i.i.d. assumption (Assumption 3), because $\gamma_i(k,n) = K\xi_i(n), \forall k \in \{1,..,K\}$ ($i \in \{1,..,N\}$), we obtain the following corollary.

Corollary 14. *Under the block-stationary assumption (Assumption 1), the dynamic assumption (Assumption 2), the in-frame i.i.d. assumption (Assumption 3), the deterministic variance assumption (Assumption 12), and the time domain independent sources assumption (Assumption 9), if the variances of the time domain source signals satisfy* $\xi_i(t) = \alpha\xi_j(t)$, $i \neq j$, *then in the frequency domain IVA approach, online algorithm based IVA is not eligible to recover the source signals in the frequency domain.*

5.5 Conclusion and Discussion

In this work, we provide insight into why frequency domain independent component analysis (ICA) approach can separate convolutive mixed source signals. For the modeling of the source signals, we develop the concept of a dynamic random process to model the source signals. It formalizes the concept of signals that are stationary in a frame but exhibit dynamics at the frame level. Frame dynamics is an important characteristic of these signals and prove important to the success of the ICA based separation methods. With suitable assumptions, the dynamic random process is stationary in the ensemble sense while a given realization may in an engineering sense exhibit 'non-stationarity'. This modeling makes rigorous often used terminology such as time varying envelope, quasi-stationary etc. We show for dynamic random processes, the unconditional distribution of the source signal in each frequency bin is a Gaussian scale mixture (GSM). The non-Gaussianity, which is critical to ICA [57], of the source signal in each frequency bin is shown to be a direct consequence of the frame dynamics. Furthermore, the independence between the unconditional distributions of the source signals in each frequency bin is related to the independence of the frame dynamics of the mixed time domain source signals. The GSM mathematical modeling is extended to the vector random processes formed by stacking the different frequency components of a source. This provides insights into the mathematical models suitable for the frequency domain independent vector analysis (IVA) type approaches. A special case of the distribution turns out to be the 'spherical distribution' employed in IVA source modeling providing support to their use in source separation. Concentrating on the bin-wise ICA methods, a significant contribution of the work is to show that signals modeled using

GSM density can be separated using ICA even though they might be dependent on each other as long as the the frame dynamics of the source signals are different.

When we discuss batch algorithm based ICA in each frequency bin, we make the assumption that the source signals are ergodic stationary processes. Assumption 5 or 4 is made for this purpose. However, instead of using Assumption 5, we can also assume the variance series $\gamma_i(k,n)$ is a deterministic time series. Under such an assumption, the source signals in each frequency bin are actually non-stationary random process, then online (stochastic) ICA algorithms can be applied in each frequency bin to separate the source signals. We discuss the applicability of the online ICA (or IVA) algorithm under the deterministic variance assumption (Assumption 12).

Our inspiration in this work is drawn from speech processing. Consequently. the terminologies used and the models are borrowed from the speech processing area. However, the modeling idea is general and the results can be applied in many different areas such as EEG or image processing among others.

Acknowledgement

The text of this chapter, in part, has been submitted for publication of the material as it may appear in Wenyi Zhang, Alireza Masnadi-Shirazi, and Bhaskar D. Rao, "Insights into the frequency domain ICA/IVA approach", submitted to *IEEE Transactions on Audio, Speech and Language Processing, special issue on reverberant speech: methodologies and applications*.

6 Stability Analysis of Complex Maximum Likelihood ICA/IVA

In this work, we analyze the stability conditions for the complex maximum likelihood ICA/IVA.

6.1 Stability Analysis of Complex Maximum Likelihood ICA

6.1.1 Complex Maximum Likelihood ICA

For convolutive mixing blind source separation problem, usually the frequency domain ICA approach is employed, which simplifies the time domain convolutive mixing problem by transforming it into instantaneous mixing problem in each frequency bin. In each frequency bin, the instantaneous mixing system is a complex mixing system, both the mixing matrix and the source signals therein are complex signals. In each frequency bin, assuming an N by N linear instantaneous complex mixing system (N sources, N sensors),

$$\mathbf{x} = A\mathbf{s} \tag{6.1}$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_M \end{bmatrix} \quad A = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{bmatrix} \quad \mathbf{s} = \begin{bmatrix} s_1 \\ \vdots \\ s_N \end{bmatrix} \tag{6.2}$$

$\mathbf{x}$ is the observation vector. A is an unknown mixing matrix assumed to have full column rank. $\mathbf{s}$ is a vector of latent variables which are assume to be non-Gaussian (or at most

one is Gaussian) and mutually independent. The de-mixing system can be written as,

$$\mathbf{y} = W\mathbf{x} \tag{6.3}$$

where $\mathbf{y} = [y_1, \cdots, y_N]^T$ is the outputs and W is a N by N de-mixing matrix. Assuming source signals $\mathbf{s}$ are independent signals, ICA attempts to solve the BSS problem by determining a de-mixing matrix W such that the outputs $\mathbf{y}$ approximate source signals $\mathbf{s}$ under permutation and scaling ambiguity [81]. Most ICA algorithms try to find a de-mixing matrix W so that the components in $\mathbf{y}$ are statistically as independent as possible.

The maximum likelihood ICA attempt to recover the source signals by minimizing the negative log-likelihood function [63].

$$\ell = -\log|\det(W)| - \sum_{k=1}^{N} \log p_k(y_k) \tag{6.4}$$

where $p_k(\cdot)$ is the assumed PDF for the k_{th} source signal. The adaptive natural gradient updating rule for the demixing matrix is,

$$\Delta W = \mu \left[I - \underline{\psi}^*(\mathbf{y})\mathbf{y}^H \right] W \tag{6.5}$$

where $\underline{\psi}(\mathbf{y}) = [\psi_1(y_1), ..., \psi_N(y_N)]^T$, $\psi_k(y_k) = -\frac{\partial \log p_k(y_k)}{\partial y_k}$ is the so called activation function [63].

6.1.2 Stability Analysis

Stability analysis of real signal ICA has been considered in [68, 71, 166, 167, 168]. Li analyzed the stability of complex maximum likelihood ICA using Wirtinger calculus [169, 170, 171]. The stationary point of the updating rule (Eq. (6.5)) satisfies,

$$E\left\{ I - \underline{\psi}^*(\mathbf{y})\mathbf{y}^H \right\} = 0 \tag{6.6}$$

It is easy to see the true separating solution W_o with which y_i and y_j are independent is a solution of Eq. (6.6), because when y_i and y_j are independent the off-diagonal term $E\left\{\psi_i^*(y_i)y_j^*\right\} = 0, i \neq j$. The diagonal term is $E\left\{\psi_i^*(y_i)y_i^*\right\} = 1$, which determines the scaling of the recovered signals. To analyze the stability of the true separating solution W_o, we need to determine the second order differential $d^2\ell$. The stationary point is

stable if and only if the expectation of $d^2\ell$ is positive to any small perturbations. Define $dZ \triangleq (dW)W^{-1}$.

$$d\ell = -tr(dZ) - tr(dZ^*) + \underline{\psi}^T(\mathbf{y})dZ\mathbf{y} + \underline{\psi}^H(\mathbf{y})dZ^*\mathbf{y}^* \tag{6.7}$$

$$d^2\ell = 2Re\left\{\mathbf{y}^T dZ^T \Phi(\mathbf{y})dZ\mathbf{y} + \mathbf{y}^T dZ^T \Theta(\mathbf{y})dZ^*\mathbf{y}^* + \underline{\psi}^T(\mathbf{y})dZdZ\mathbf{y}\right\} \tag{6.8}$$

where $\Phi(\mathbf{y})$ is a diagonal matrix with the i_{th} diagonal element $\phi_i(y_i) = -\frac{\partial \log p_i(y_i)}{\partial y_i \partial y_i}$, $\Theta(\mathbf{y})$ is a diagonal matrix with the i_{th} diagonal element $\theta_i(y_i) = -\frac{\partial \log p_i(y_i)}{\partial y_i \partial y_i^*}$. At the stationary point W_o (the true separating solution), we can simplify the expectation of $d^2\ell$ as,

$$E\left\{d^2\ell\right\} = \sum_{j>i} \begin{bmatrix} dz_{ij}^* \\ dz_{ji}^* \\ dz_{ij} \\ dz_{ji} \end{bmatrix}^H \begin{bmatrix} H_1(i,j) & H_2(i,j) \\ H_2^*(i,j) & H_1(i,j) \end{bmatrix} \begin{bmatrix} dz_{ij}^* \\ dz_{ji}^* \\ dz_{ij} \\ dz_{ji} \end{bmatrix} + \sum_i \begin{bmatrix} dz_{ii}^* \\ dz_{ii} \end{bmatrix}^H H_3(i) \begin{bmatrix} dz_{ii}^* \\ dz_{ii} \end{bmatrix} \tag{6.9a}$$

where

$$H_1(i,j) = \begin{bmatrix} \beta_j\delta_i & 0 \\ 0 & \beta_i\delta_j \end{bmatrix} \quad H_2(i,j) = \begin{bmatrix} \alpha_j\gamma_i & 1 \\ 1 & \alpha_i\gamma_j \end{bmatrix} \quad H_3(i) = \begin{bmatrix} v_i & u_i + 1 \\ u_i^* + 1 & v_i \end{bmatrix} \tag{6.10}$$

$$\begin{aligned} \alpha_i = E\left\{y_i^2\right\} \quad \beta_i = E\left\{|y_i|^2\right\} \\ \gamma_i = E\left\{\phi_i(y_i)\right\} \quad \delta_i = E\left\{\theta_i(y_i)\right\} \\ u_i = E\left\{y_i^2\phi_i(y_i)\right\} \quad v_i = E\left\{|y_i|^2\theta_i(y_i)\right\} \end{aligned}$$

To the end, the stability conditions are stated in the following theorem.

Theorem 15. *[169] The true separating solution W_o is a stable stationary point of the updating rule* (6.5) *if and only if* $\begin{bmatrix} H_1(i,j) & H_2(i,j) \\ H_2^*(i,j) & H_1(i,j) \end{bmatrix}$, $i \neq j, i, j \in \{1, ..., N\}$ *and* $H_3(i), i \in \{1, ..., N\}$ *is positive definite.*

However, we will prove in the following that for complex ICA, Theorem 15 is often not complete. As we know, ICA analyzes a blind mixing system through the criterion of independence between source signals. Generally, a complex ICA algorithm

requires each recovered source signal satisfy a variance constraint (such as the variance equals to 1) to avoid scaling ambiguity, but it does not specify the phase of the recovered source signals. That is, supposing each of the recovered complex source signals is multiplied by a unimodular scalar (this means a phase change), they are still independent and this is still a valid solution of the complex ICA algorithm. For example, for the complex maximum likelihood ICA, the assumed PDF $p_k(\cdot)$ for the k_{th} complex source signal y_k may be chosen as a Laplacian distribution $p_k(y_k) = \frac{1}{2b}\exp\left\{-\frac{|y_k|}{b}\right\}$ (b is a scalar controlling the variance of the distribution). It is easy to see this distribution is invariant to any unimodular scalar multiplication on y_k. There are many other formats of PDF $p_k(y_k)$ used in various applications, but usually they are invariant to unimodular scalar multiplication. In summary, we conclude the loss function of the complex maximum likelihood ICA (Eq. (6.4)) is invariant to any unimodular scalar multiplication to the separating matrix W. Thereby, any minimizer $\hat{W}$ to the maximum likelihood ICA loss function (Eq. (6.4)) is not unique, but is associated with a connected family of solutions given by,

$$\left[\hat{W}\right] = \left\{\hat{W}_{\underline{\rho}} | \hat{W}_{\underline{\rho}} = diag(\underline{\rho}) \cdot \hat{W}, \underline{\rho} = [\rho_1, ..., \rho_N]^T, \rho_i \in \mathbb{C}, |\rho_i| = 1\right\} \tag{6.12}$$

where $diag(\underline{\rho})$ denotes a diagonal matrix constructed by using the elements of $\underline{\rho}$ as the diagonal elements of the matrix. A similar non-unique minimizer problem is considered in Constant Modulus based blind adaptive equalization (see [172] and references therein).

Supposing $w_{i\cdot}$ and $dw_{i\cdot}$ denote the i_{th} row of the matrix W and dW respectively, the i_{th} output $y_i = w_{i\cdot}\mathbf{x}$. As we know, if W is a stationary point of the ICA loss function (Eq. (6.4)), $diag(\underline{\rho}) \cdot W$ ($|\rho_i| = 1$) is also a stationary point of the loss function, and $\ell(W) = \ell(diag(\underline{\rho}) \cdot W)$. Consider a small perturbation $dW = diag(\underline{\rho}) \cdot W - W$, $dw_{i\cdot} = \rho_i w_{i\cdot} - w_{i\cdot} = (\rho_i - 1)w_{i\cdot}$, where $|\rho_i - 1| \leqslant \epsilon$, ϵ is a small number to make sure $dw_{i\cdot}$ is a small perturbation. Because $|\rho_i| = 1$ and $|\rho_i - 1| \leqslant \epsilon$, we can have the approximation $\rho_i - 1 \simeq \pi_i J$, where π_i is a small real number $|\pi_i| \leqslant \epsilon$, J is the symbol denoting imaginary number. That is, $\rho_i - 1$ is a pure imaginary number. Because $\ell(W) = \ell(W + dW)$ and W is a stationary point, $dW = diag(\underline{\rho}) \cdot W - W$ represents a direction which keeps $\ell(\cdot)$ constant and the second order variation $d^2\ell = 0$. In summary, when $dw_{i\cdot} = (\rho_i - 1)w_{i\cdot}$, $\rho_i - 1$ denotes a pure imaginary perturbation, then $d^2\ell = 0$.

By definition the perturbation term $dZ \triangleq (dW)W^{-1}$, $dW = (dZ)W$. Supposing all the elements of dZ is 0 except the i_{th} diagonal elements $dz_{ii} = \pi_i J$ (π_i is a small real number $|\pi_i| \leqslant \epsilon$ and $\pi_i J$ denotes a pure imaginary number), then $dw_{i\cdot} = dz_{ii} w_{i\cdot}$, and $d^2\ell = 0$. By Eq. (6.9) (recall all the elements of dZ is 0 except dz_{ii}),

$$E\left\{d^2\ell\right\} = \begin{bmatrix} dz_{ii}^* \\ dz_{ii} \end{bmatrix}^H H_3(i) \begin{bmatrix} dz_{ii}^* \\ dz_{ii} \end{bmatrix} \tag{6.13}$$

Thereby, $H_3(i)$ should have at least one zero eigenvalue. This phenomenon is caused by the non-unique minimizer property of the ICA loss function, i.e. any minimizer $\hat{W}$ to the ICA loss function (Eq. (6.4)) is not unique, but is associated with a connected family of solutions. This phenomenon is not displayed in real signal ICA problem since therein each minimizer is isolated. As a consequence, Theorem 15 should be revised as following,

Theorem 16. *The true separating solution W_o is a stable stationary point of the updating rule* (6.5) *if and only if* $\begin{bmatrix} H_1(i,j) & H_2(i,j) \\ H_2^*(i,j) & H_1(i,j) \end{bmatrix}$, $i \neq j, i, j \in \{1, ..., N\}$ *is positive definite and the non-zero eigenvalue of $H_3(i), i \in \{1, ..., N\}$ is positive.*

6.2 Stability Analysis of Complex Maximum Likelihood IVA

6.2.1 Complex Maximum Likelihood IVA

IVA is proposed to solve the permutation ambiguity problem [96, 97, 98, 99]. IVA does not solve the mixing problem in each frequency bin independently, instead, it considers the frequency domain source signal as a vector source and attempt to solve the mixing problems in all frequency bins as a whole problem. Thereby, each vector source signal's components in all frequency bins are aligned together and the permutation ambiguity across frequency domain is avoided.

Assuming an N by N time domain convolutive mixing system (N sources, N sensors) and the total number of frequency bins is K. The time domain and the corresponding frequency domain mixing system is expressed in Eq. 5.1 and Eq. 5.2 respectively.

The IVA de-mixing system can be written as,

$$\begin{bmatrix} Y_1(1,t) \\ \vdots \\ Y_1(K,t) \\ \vdots \\ Y_N(1,t) \\ \vdots \\ Y_N(K,t) \end{bmatrix} = \begin{bmatrix} w_{11}(1) & \mathbf{0} & \cdots & w_{1N}(1) & \mathbf{0} \\ & \ddots & & \ddots & \\ \mathbf{0} & w_{11}(K) & \mathbf{0} & & w_{1N}(K) \\ & \vdots & \ddots & \vdots & \\ w_{N1}(1) & \mathbf{0} & & w_{NN}(1) & \mathbf{0} \\ & \ddots & \cdots & \ddots & \\ \mathbf{0} & w_{N1}(K) & \mathbf{0} & & w_{NN}(K) \end{bmatrix} \begin{bmatrix} X_1(1,t) \\ \vdots \\ X_1(K,t) \\ \vdots \\ X_N(1,t) \\ \vdots \\ X_N(K,t) \end{bmatrix} \tag{6.14}$$

$$\Leftrightarrow \mathbf{Y}(t) = \mathbf{W}\mathbf{X}(t) \tag{6.15}$$

$$\Leftrightarrow \begin{bmatrix} \mathbf{Y}_1(t) \\ \vdots \\ \mathbf{Y}_N(t) \end{bmatrix} = \begin{bmatrix} \mathbf{W}_{11} & \cdots & \mathbf{W}_{1N} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{N1} & \cdots & \mathbf{W}_{NN} \end{bmatrix} \begin{bmatrix} \mathbf{X}_1(t) \\ \vdots \\ \mathbf{X}_N(t) \end{bmatrix} \tag{6.16}$$

where $\mathbf{W}$ is the whole separating matrix (demixing matrix). K is the total number of frequency bins. $\mathbf{X}_i(t) = [X_i(1,t), ..., X_i(K,t)]^T$ is the mixed signal's spectrum at the i_{th} channel. $\mathbf{Y}_i(t) = [Y_i(1,t), ..., Y_i(K,t)]^T$ is the spectrum of the i_{th} recovered source signal.

The maximum likelihood IVA attempts to recover the vector source signals by minimizing the negative log-likelihood function.

$$\ell = -\log|\det(\mathbf{W})| - \sum_{k=1}^{N} \log p_k(\mathbf{Y}_k(t)) \tag{6.17}$$

where $p_k(\mathbf{Y}_k(t))$ is the assumed PDF for k_{th} vector source signal $\mathbf{Y}_k(t)$. The adaptive natural gradient updating rule for the demixing matrix is,

$$\Delta\mathbf{W} = \mu\left\{I - \frac{\partial}{\partial\mathbf{W}}\left[-\sum_{k=1}^{N}\log p_k(\mathbf{Y}_k(t))\right]\right\}\mathbf{W} \tag{6.18}$$

However, the updating rule for the whole separating matrix $\Delta\mathbf{W}$ is difficult to express in a simple formula. Hence we write the updating rule for the separating matrix in each frequency bin $\Delta\mathbf{W}(\omega)$ instead of $\Delta\mathbf{W}$.

$$\mathbf{W}(\omega) \triangleq \begin{bmatrix} w_{11}(\omega) & \cdots & w_{1N}(\omega) \\ \vdots & \ddots & \vdots \\ w_{N1}(\omega) & \cdots & w_{NN}(\omega) \end{bmatrix} \tag{6.19}$$

The updating rule of $\Delta\mathbf{W}(\omega)$ is,

$$\Delta\mathbf{W}(\omega) = \mu\left[I - \underline{\psi}^*_{\omega}(\mathbf{Y}(t))\mathbf{Y}(\omega,t)^H\right]\mathbf{W}(\omega) \tag{6.20}$$

where $\mathbf{Y}(\omega,t) \triangleq [Y_1(\omega,t),...,Y_N(\omega,t)]^T$, $\underline{\psi}_{\omega}(\mathbf{Y}(t)) = [\psi_{1\omega}(\mathbf{Y}_1(t)),...,\psi_{N\omega}(\mathbf{Y}_N(t))]^T$, $\psi_{k\omega}(\mathbf{Y}_k(t)) = -\frac{\partial}{\partial Y_k(\omega,t)}\log p_k(\mathbf{Y}_k(t))$ is the so called activation function.

6.2.2 Stability Analysis

This section analyzes the stability conditions for the complex maximum likelihood IVA. The stationary point of the updating rule (Eq. (6.20)) satisfies,

$$E\left\{I - \underline{\psi}^*_{\omega}(\mathbf{Y}(t))\mathbf{Y}(\omega,t)^H\right\} = 0 \tag{6.21}$$

It is easy to see the true separating solution $\mathbf{W}_o$ with which $\mathbf{Y}_i(t)$ and $\mathbf{Y}_j(t)$ are independent is a solution of Eq. (6.21), because when $\mathbf{Y}_i(t)$ and $\mathbf{Y}_j(t)$ are independent the off-diagonal term $E\left\{\psi^*_{i\omega}(\mathbf{Y}_i(t))Y_j(\omega,t)^*\right\} = 0, i \neq j$. The diagonal term is $E\left\{\psi^*_{i\omega}(\mathbf{Y}_i(t))Y_i(\omega,t)^*\right\} = 1$, which determines the scaling of the recovered signals. To analyze the stability of the true separating solution $\mathbf{W}_o$, we need to determine the second order differential $d^2\ell$. The stationary point is stable if and only if the expectation of $d^2\ell$ is positive to any small perturbations.

Define $d\mathbf{Z} \triangleq (d\mathbf{W})\mathbf{W}^{-1}$.

$$d\ell = -tr(d\mathbf{Z}) - tr(d\mathbf{Z}^*) + \underline{\psi}^T(\mathbf{Y}(t))d\mathbf{Z}\mathbf{Y}(t) + \underline{\psi}^H(\mathbf{Y}(t))d\mathbf{Z}^*\mathbf{Y}(t)^* \tag{6.22}$$

$$d^2\ell = 2Re\left\{\mathbf{Y}(t)^T d\mathbf{Z}^T\Phi(\mathbf{Y}(t))d\mathbf{Z}\mathbf{Y}(t) + \mathbf{Y}(t)^T d\mathbf{Z}^T\Theta(\mathbf{Y}(t))d\mathbf{Z}^*\mathbf{Y}(t)^* + \underline{\psi}^T(\mathbf{Y}(t))d\mathbf{Z}d\mathbf{Z}\mathbf{Y}(t)\right\} \tag{6.23}$$

where $\underline{\psi}(\mathbf{Y}(t)) = [\psi_{11}(\mathbf{Y}_1(t)),...,\psi_{1K}(\mathbf{Y}_1(t))),...,\psi_{N1}(\mathbf{Y}_N(t)),...,\psi_{NK}(\mathbf{Y}_N(t))]^T$ is a $N \cdot K$ dimension vector, $\psi_{kj}(\mathbf{Y}_k(t)) = -\frac{\partial}{\partial Y_k(j,t)}\log p_k(\mathbf{Y}_k(t))$. At the true separating solution point $\mathbf{W}_o$, $\Phi(\mathbf{Y}(t)) = \frac{\partial}{\partial\mathbf{Y}(t)}\underline{\psi}(\mathbf{Y}(t))$ is a block diagonal matrix with the i_{th} diagonal block denoted by $\Phi_i(\mathbf{Y}(t))$ (the block has K dimensions), $\Theta(\mathbf{Y}(t)) = \frac{\partial}{\partial\mathbf{Y}^*(t)}\underline{\psi}(\mathbf{Y}(t))$ is also a block diagonal matrix with the i_{th} diagonal block denoted by $\Theta_i(\mathbf{Y}(t))$ (the block has K dimensions).

Since $\mathbf{W}$ is a square block matrix with $N{\cdot}N$ blocks, each block is a K dimensional diagonal matrix, by inverse of block matrix lemma, it is easy to see $\mathbf{W}^{-1}$ has the same

structure as $\mathbf{W}$. $d\mathbf{W}$ also has the same structure. Thereby $d\mathbf{Z} = (d\mathbf{W})\mathbf{W}^{-1}$ has the same structure as $\mathbf{W}$, i.e. $d\mathbf{Z}$ is a square block matrix with $N \cdot N$ blocks, each block is a K dimensional diagonal matrix. Let $dz_{ij,k}$ denotes the k_{th} diagonal element of the block of $d\mathbf{Z}$ at the i_{th} row and the j_{th} column. At a stationary point of the updating rule (Eq. (6.20)), $E\left\{\psi^*_{i\omega}(\mathbf{Y}_i(t))Y_i(\omega,t)^*\right\} = 1$. Combining the above mentioned properties, after a lengthy derivation, we can simplify the expectation of $d^2\ell$ at the stationary point $\mathbf{W}_o$ (the true separating solution),

$$E\left\{d^2\ell\right\} = \sum_{j>i} \begin{bmatrix} dz_{ij,1} \\ \vdots \\ dz_{ij,K} \\ dz^*_{ij,1} \\ \vdots \\ dz^*_{ij,K} \\ dz_{ji,1} \\ \vdots \\ dz_{ji,K} \\ dz^*_{ji,1} \\ \vdots \\ dz^*_{ji,K} \end{bmatrix}^H Q_2(i,j) \begin{bmatrix} dz^*_{ij,1} \\ \vdots \\ dz^*_{ij,K} \\ dz_{ij,1} \\ \vdots \\ dz_{ij,K} \\ dz^*_{ji,1} \\ \vdots \\ dz^*_{ji,K} \\ dz_{ji,1} \\ \vdots \\ dz_{ji,K} \end{bmatrix} + \sum_i \begin{bmatrix} dz_{ii,1} \\ \vdots \\ dz_{ii,K} \\ dz^*_{ii,1} \\ \vdots \\ dz^*_{ii,K} \end{bmatrix}^H Q_1(i) \begin{bmatrix} dz^*_{ii,1} \\ \vdots \\ dz^*_{ii,K} \\ dz_{ii,1} \\ \vdots \\ dz_{ii,K} \end{bmatrix} \tag{6.24}$$

$$Q_1(i) = \begin{bmatrix} P_1(i) & P_2(i) \\ P^*_2(i) & P^*_1(i) \end{bmatrix} \tag{6.25}$$

$$P_1(i) = \begin{bmatrix} E\{|Y_i(1,t)|^2\theta_{i,11}\} & \cdots & E\{Y_i(1,t)Y^*_i(K,t)\theta_{i,1K}\} \\ \vdots & \ddots & \vdots \\ E\{Y_i(1,t)Y^*_i(K,t)\theta_{i,1K}\}^* & \cdots & E\{|Y_i(K,t)|^2\theta_{i,KK}\} \end{bmatrix} \tag{6.26}$$

where $\theta_{i,kl}$ is the k_{th} row and the l_{th} column element of $\Theta_i(\mathbf{Y}(t))$.

$$P_2(i) = \begin{bmatrix} E\{Y_i(1,t)^2\phi_{i,11}\} + 1 & \cdots & E\{Y_i(1,t)Y_i(K,t)\phi_{i,1K}\} \\ \vdots & \ddots & \vdots \\ E\{Y_i(1,t)Y_i(K,t)\phi_{i,1K}\} & \cdots & E\{|Y_i(K,t)|^2\phi_{i,KK}\} + 1 \end{bmatrix} \tag{6.27}$$

where $\phi_{i,kl}$ is the k_{th} row and the l_{th} column element of $\Phi_i(\mathbf{Y}(t))$. $P_1(i)$ is Hermitian, $P_2(i)$ is symmetrical, hence $Q_1(i)$ is Hermitian.

$$Q_2(i,j) = \begin{bmatrix} P_3(i,j) & P_4(i,j) & 0 & I_K \\ P_4^*(i,j) & P_3^*(i,j) & I_K & 0 \\ 0 & I_K & P_3(j,i) & P_4(j,i) \\ I_K & 0 & P_4^*(j,i) & P_3^*(j,i) \end{bmatrix} \tag{6.28}$$

where I_K denotes K dimension identity matrix.

$$P_3(i,j) = \begin{bmatrix} E\{|Y_j(1,t)|^2\}E\{\theta_{i,11}\} & \cdots & E\{Y_j(1,t)Y_j^*(K,t)\}E\{\theta_{i,1K}\} \\ \vdots & \ddots & \vdots \\ E\{Y_j(1,t)Y_j^*(K,t)\}^*E\{\theta_{i,1K}\}^* & \cdots & E\{|Y_j(K,t)|^2\}E\{\theta_{i,KK}\} \end{bmatrix} \tag{6.29}$$

$$P_4(i,j) = \begin{bmatrix} E\{Y_j(1,t)^2\}E\{\phi_{i,11}\} & \cdots & E\{Y_j(1,t)Y_j(K,t)\}E\{\phi_{i,1K}\} \\ \vdots & \ddots & \vdots \\ E\{Y_j(1,t)Y_j(K,t)\}E\{\phi_{i,1K}\} & \cdots & E\{|Y_j(K,t)|^2\}E\{\phi_{i,KK}\} \end{bmatrix} \tag{6.30}$$

$P_3(i,j)$ is Hermitian, $P_4(i,j)$ is symmetrical, hence $Q_2(i,j)$ is Hermitian. To the end, the stability conditions are stated in the following theorem.

Theorem 17. *The true separating solution* $\mathbf{W}_o$ *is a stable stationary point of the updating rule* (6.20) *if and only if* $Q_2(i,j), i \neq j, i, j \in \{1,...,N\}$ *and* $Q_1(i), i \in \{1,...,N\}$ *is positive definite.*

If the PDF $p_k(\mathbf{Y}_k(t))$ (in Eq. 6.17) is assumed to be the commonly used spherical distribution [96] (see section 5.2.2 for the insight in choosing spherical distribution), because of the invariance of the spherical distribution to any unimodular scalar multiplication on $Y_i(k,t)$, we can demonstrate that $Q_1(i)$ has K zero eigenvalues by the same argument as used in proving Theorem 16. Specifically, if $p_k(\mathbf{Y}_k(t)) = h\exp\{-K\|\mathbf{Y}_k(t)\|_2\}$, we can induce directly that $Q_1(i)$ has K zero eigenvalues. Therefore, we have the following theorem,

Theorem 18. *The true separating solution* $\mathbf{W}_o$ *is a stable stationary point of the updating rule* (6.20) *if and only if* $Q_2(i,j), i \neq j, i, j \in \{1,...,N\}$ *is positive definite and the* K *non-zero eigenvalues of* $Q_1(i), i \in \{1,...,N\}$ *is positive.*

One weakness of the stability condition described in Theorem 18 is that the mathematical formulas are so complicated that it does not give us any insight into into the problem. We can not use the stability condition to guide the design of the right assumed PDF $p_k(\mathbf{Y}_k(t))$ (in Eq. 6.17). Theorem 18 can only be used to test whether a given PDF $p_k(\mathbf{Y}_k(t))$ is appropriate to be used for separating specific source signals. We test the stability conditions on using $p_k(\mathbf{Y}_k(t)) = h \exp\{-K\|\mathbf{Y}_k(t)\|_2\}$ for speech signal separation, and the experiment confirmed $Q_2(i, j)$ is positive definite and $Q_1(i)$ is semi-positive definite with K zero eigenvalues. Therefore, $p_k(\mathbf{Y}_k(t)) = h \exp\{-K\|\mathbf{Y}_k(t)\|_2\}$ is appropriate to be used in IVA for speech signal separation.

7 Combining Independent Component Analysis with Geometric Information

As a study on combing ICA and beamforming, in this work, we propose two approaches for combining geometric information with ICA algorithm to solve permutation problem under the scenario where a rough information about the direction of the desired source is known. The first approach is a new blind extraction algorithm with a soft quadratic geometric constraint. The desired source is guaranteed to be conveyed to the output with little distortion by the quadratic constraint and the negentropy maximization criterion is used to ensure that the other sources get suppressed at the output. The second approach employs a quadratic geometric test as a post-processing step to pickup the desired source after ICA processing. An advantage of the proposed two approaches is that they do not require accurate knowledge of the number of sources in the mixtures to recover the desired source, in contrast, other geometric ICA approaches usually fail if the number of sources is not known accurately.

7.1 Introduction to the permutation problem

Independent component analysis (ICA) is a statistical method for extracting independent components from a group of mixtures [81]. For convolutive mixtures, frequency domain ICA approach is mostly used since it simplifies the problem into instantaneous mixing problem in every frequency bin and can be solved therein by simple instantaneous mixing ICA algorithms. However the intrinsic scaling and permutation

ambiguities need to be addressed in applying the frequency domain ICA approach.

ICA assumes no knowledge about the mixing process except the independence between sources. However, sometimes extra information is available and can be utilized to aid the ICA process. For example, in microphone array speech processing, speech's temporal structure or geometric information with the array may be employed to solve the permutation problem. Many recent works have been developed to combine ICA with geometric information to solve the permutation problem [85, 86, 87, 88, 89, 91, 95]. In [85, 86], beam pattern of the ICA processor is utilized to figure out the directions of the sources to solve the permutation problem. These methods become too complicated and are not robust when the number of sources exceed 2. Parra and Alvino proposed the geometrically constrained (or initialized) ICA algorithm [87], but accurate source number is required and correct permutation is not guaranteed. Knaak and Araki proposed an ICA algorithm with a hard linear geometric constraint [88]. However, accurate source number is mandatory for the algorithm to perform properly.

In microphone array speech processing, the geometry of the array and rough information about the direction of the desired signal may be known a priori, for instance, the direction of the desired signal may be assumed to be the broadside direction for a linear array, or be acquired by some direction of arrival (DOA) estimation algorithms [1]. In this work, we propose two approaches for combining DOA information of the desired source with ICA algorithm to solve the permutation problem. The first approach is a new blind extraction algorithm with a soft quadratic geometric constraint. Given a rough estimate of the direction of the desired source, the proposed algorithm will extract the desired source from the mixed signals. The quadratic constraint restricts the weighted square error between the desired and actual response of the processor over a small spatial uncertainty region chosen to deal with look direction uncertainty. Thereby the desired source is guaranteed to be conveyed to the output with little distortion and the negentropy maximization criterion is used to ensure that the other sources get suppressed at the output. This method solves the permutation problem in the frequency domain ICA approach since the desired signal is extracted consistently across all the frequency bins. The second approach employs a quadratic geometric test as a post-processing step to pickup the desired source after ICA processing. In every frequency bin, the ICA algo-

rithm separates instantaneously mixed source signals, then the quadratic geometric test will pick up the desired source. An advantage of the proposed two approaches is that they do not require accurate knowledge of the number of sources in the mixtures to recover the desired source, in contrast, other geometric ICA approaches usually fail if the number of sources is not known accurately.

In section 7.2.1 and 7.2.2, we describe the complex FastICA algorithm and the linear constrained ICA algorithm. Section 7.3 discusses the proposed two approaches combining ICA with geometric information. The simulation results and discussion are presented in Section 7.4.

7.2 Background

7.2.1 Complex FastICA

Hyvarinen and Oja proposed a fast fixed point algorithm (FastICA) for solving the real variable instantaneous mixing ICA problem [58]. FastICA maximizes the negentropy of the output $y_i, i = 1..N$, subject to the constraints that all the $y_i, i = 1..N$ are uncorrelated and have unit variances. Bingham and Hyvarinen extended it to complex variables and developed a complex FastICA algorithm [60]. The mathematical problem to be solved is,

$$\min_{\mathbf{w}_j, j=1..N} \sum_{j=1}^{N} J_G(\mathbf{w}_j), \quad s.t. \quad E\{(\mathbf{w}_k^H \mathbf{x})(\mathbf{w}_j^H \mathbf{x})^*\} = \delta_{jk}$$

where $\delta_{jk} = 1$ for $j = k$ and 0 otherwise. $\mathbf{w}_j$ is M-dimensional complex weight vector. $\mathbf{x}$ is the observation vector. The contrast function is defined as,

$$J_G(\mathbf{w}) = E\{G(|\mathbf{w}^H \mathbf{x}|^2)\} \tag{7.1}$$

where $G : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}$ is a smooth even function. For example, $G(z) = \log(a + z)$ is a good choice for speech separation task.

When the observed data $\mathbf{x}$ is preliminary whitened, Bingham derived the fixed-

point algorithm for one unit,

$$\begin{aligned} \mathbf{w}^+ &= E\{\mathbf{x}(\mathbf{w}^H\mathbf{x})^* g(|\mathbf{w}^H\mathbf{x}|^2)\} \\ &\quad - E\{g(|\mathbf{w}^H\mathbf{x}|^2) + |\mathbf{w}^H\mathbf{x}|^2 g'(|\mathbf{w}^H\mathbf{x}|^2)\}\mathbf{w} \\ \mathbf{w} &= \frac{\mathbf{w}^+}{\|\mathbf{w}^+\|} \end{aligned} \tag{7.2}$$

To estimate N independent components, we can either use a deflation scheme based on Gram-Schmidt decorrelation, or estimate all the independent components simultaneously and use a symmetric decorrelation [60].

7.2.2 ICA with a Linear Geometric Constraint

Knaak, Araki and Makino studied the similarities between adaptive beamforming and ICA, and proposed a geometrically constrained ICA algorithm (CICA) [88]. CICA combined a linear look direction constraint with the ICA criterion. It is claimed that the CICA converges to the right solution as long as its look direction is closer to the target signal than to the jammer signal. Knaak developed the look direction constrained ICA algorithm based on Bingham's complex FastICA. The problem to be solved is formulated mathematically as follows.

$$\arg\min_{\mathbf{w}} E\{G(|\mathbf{w}^H\mathbf{x}|^2)\} \equiv \arg\min_{\mathbf{t}} E\{G(|\mathbf{t}^H\mathbf{z}|^2)\} \tag{7.3a}$$

$$s.t. \quad \mathbf{w}^H\mathbf{a} = \mathbf{t}^H V\mathbf{a} = 1. \tag{7.3b}$$

where $\mathbf{w}^H = \mathbf{t}^H V$, $\mathbf{w}$ is the weight vector for original observed data $\mathbf{x}$, and $\mathbf{t}$ is the weight vector for the sphered data $\mathbf{z}$. V is the sphering matrix determined by principle component analysis (PCA). $\mathbf{a}$ is the estimated steering vector for the desired source. The algorithm is initialized with the MVDR beamformer (section 4.2.1) on the sphered data,

$$\mathbf{t}_0 = V\mathbf{a} \tag{7.4}$$

Knaak developed the solution to the above optimization problem based on Bingham's Newton-type iterative algorithm. The Newton solution is projected onto the linear constraint (7.3b) after each iteration.

$$\mathbf{t}_{k+1_{new}} = \frac{\mathbf{t}_{k+1}}{|\mathbf{t}_{k+1}^H V\mathbf{a}|}$$

7.3 Combine ICA with Geometric Information

7.3.1 ICA with Quadratic Geometric Constraint

One shortcoming of the CICA algorithm is that it requires accurate knowledge of the number of the sources in the mixtures. Also the performance of the CICA algorithm depends heavily on the initialization of the weight vector. A careful chosen initialization weight vector ensures the algorithm converge to the desired source. However, such initialization weight vector is available only when the source number is accurately known and the PCA preprocessing is properly done based on that.

Assuming a M sensor, N source system. When the PCA preprocessing is performed based on accurate source number, we effectively have a mixing system with a square mixing matrix, i.e. $M' = N$, where M' is the effective sensor number. As is known, ICA is closely related to beamforming [85, 86, 88]. When all the original sources have similar variances, the MVDR beamformer $\mathbf{t}_0$ (7.4) on the sphered data is close to the optimum ICA solution because of the limited free spatial dimensions. Thereby, the CICA is inclined to converge to the right solution. However, when the PCA preprocessing is performed based on an overestimated source number, we are effectively working on a mixing system with more sensors than sources, i.e. $M' > N$. In this case, there is enough free spatial dimensions in the weight vector such that the MVDR beamformer $\mathbf{t}_0$ (Eq.(7.4)) on the sphered data is far from the right solution (Here we assume the estimated steering vector $\mathbf{a}$ is close to but not exactly the true steering vector for the desired source since DOA error always exists in real applications). Thereby, the CICA algorithm may not converge to the right solution and it may amplify an undesired source and suppress the desired source. Furthermore, the CICA algorithm fails when the PCA preprocessing is performed based on underestimated source number. These drawbacks are confirmed in the experiments. Another possibility is to initialize the CICA algorithm with the Delay-and-Sum beamformer. However, experiments show that this still can not ensure the CICA algorithm to converge to the right solution consistently.

In this section, we propose a new ICA algorithm with a quadratic geometric constraint which combines the geometric information of the array with the ICA criterion. Instead of using a hard linear constraint as in CICA algorithm, here we propose to use a

soft quadratic constraint which can accommodate uncertainty in the look direction information. In [142], a quadratic constraint is used to ensure the robustness of the adaptive beamformer in the look direction. The quadratic constraint restricts the weighted square error between desired and actual beam pattern of the beamformer over a small spatial region chosen to deal with look direction uncertainty. This error item can be written as,

$$\begin{aligned} e^2 &= \int_{\bar{\theta}-\Delta\theta}^{\bar{\theta}+\Delta\theta} f(\theta) \left| \mathbf{w}^H \mathbf{a}(\theta,\omega) - \mathbf{w}_d^H \mathbf{a}(\theta,\omega) \right|^2 d\theta \\ &= (\mathbf{w}-\mathbf{w}_d)^H \int_{\bar{\theta}-\Delta\theta}^{\bar{\theta}+\Delta\theta} f(\theta)\mathbf{a}(\theta,\omega)\mathbf{a}(\theta,\omega)^H d\theta (\mathbf{w}-\mathbf{w}_d) \\ &= (\mathbf{w}-\mathbf{w}_d)^H \Phi (\mathbf{w}-\mathbf{w}_d), \end{aligned}$$

$$\text{with} \quad \Phi = \int_{\bar{\theta}-\Delta\theta}^{\bar{\theta}+\Delta\theta} f(\theta)\mathbf{a}(\theta,\omega)\mathbf{a}(\theta,\omega)^H d\theta.$$

$\bar{\theta}$ is the assumed look direction, $\Delta\theta$ is a measure of uncertainty in the assumed look direction, $f(\theta)$ is a spatial weighting function, ω is a fixed frequency and $\mathbf{a}(\theta,\omega)$ is the array steering vector. Φ is a positive definite constraint matrix which can be calculated by either mathematical integration or by numerical techniques. $\mathbf{w}$ is the beamformer's weight vector of interest. $\mathbf{w}_d^H \mathbf{a}(\theta,\omega)$ is the desired response in the direction θ, and it is expressed as the inner product between a desired beamformer's weight vector $\mathbf{w}_d$ and steering vector $\mathbf{a}(\theta,\omega)$ to simplify computation. Generally, Delay-and-Sum beamformer is used as the desired beamformer because of its robustness in the look direction. Then the quadratic constraint is written as,

$$(\mathbf{w}-\mathbf{w}_d)^H \Phi (\mathbf{w}-\mathbf{w}_d) \leq \varepsilon$$

We propose to combine this quadratic constraint with the negentropy maximization criterion. The new optimization problem is stated as,

$$\min_{\mathbf{w}} E\{G(|\mathbf{w}^H\mathbf{x}|^2)\}, \quad s.t. \quad (\mathbf{w}-\mathbf{w}_d)^H \Phi (\mathbf{w}-\mathbf{w}_d) \leq \varepsilon \tag{7.6}$$

The quadratic constraint will ensure the solution of interest has a flat response close to 1 in the uncertainty region. When the desired source lies in the uncertainty region, it will be conveyed to the output with little distortion. The negentropy maximization criterion then will suppress the undesired sources to ensure distribution of the output be as far

as possible from the Gaussian distribution. In other words, maximizing the negentropy, as an ICA criterion, will converge to recover one of the N sources, while the quadratic geometric constraint will ensure it converge to recover the desired source. The little distorted conveyance of the desired source to the output is an attractive attribute. It avoids the scaling ambiguity intrinsic in ICA. This is most useful for frequency domain ICA approach, enabling proper scaling in every frequency bin and escaping the re-scaling headache in conventional frequency domain ICA approach. Define

$$\tilde{\mathbf{w}} = \Phi^{1/2}(\mathbf{w} - \mathbf{w}_d)$$
$$f(\tilde{\mathbf{w}}) = E\{G(|(\Phi^{-1/2}\tilde{\mathbf{w}} + \mathbf{w}_d)^H \mathbf{x}|^2)\}.$$

Problem (7.6) can be written as,

$$\min_{\tilde{\mathbf{w}}} f(\tilde{\mathbf{w}}), \quad s.t. \quad \|\tilde{\mathbf{w}}\|^2 \leq \varepsilon$$

We use the following iterative conjugate complex gradient and projection method to solve the above optimization problem (see [142] and refs therein), where μ is a step size parameter.

$$\tilde{\mathbf{w}}_k = P[\tilde{\mathbf{w}}_{k-1} - \mu \nabla_{\tilde{\mathbf{w}}} f(\tilde{\mathbf{w}})]$$
$$P(\tilde{\mathbf{w}}) = \begin{cases} \tilde{\mathbf{w}}, & \|\tilde{\mathbf{w}}\| \leq \sqrt{\varepsilon} \\ \tilde{\mathbf{w}} \frac{\sqrt{\varepsilon}}{\|\tilde{\mathbf{w}}\|}, & \|\tilde{\mathbf{w}}\| > \sqrt{\varepsilon}. \end{cases}$$

7.3.2 Use Geometric Test as Post-processing to ICA

The second approach for combining ICA with geometric information is to employ a quadratic geometric test as a post-processing step to pick up the desired source after ICA processing. In every frequency bin, the ICA algorithm separates instantaneously mixed source signals, then the quadratic geometric test will pick up the desired source.

Assuming a M sensor, N source system. Suppose the estimated source number is N' (N' is not necessary to be N but should satisfies $N' \geqslant N$). Suppose the ICA weight vectors are $\mathbf{w}_i, i = 1, .., N'$. We can calculate a set of response scores, $\mathbf{w}_i^H \Phi \mathbf{w}_i, i = 1, .., N'$ (Φ is defined in sec.7.3.1) based on a rough direction information of the desired source.

Because of the similarity between ICA and beamforming, the ICA weight vector which recover the desired source should have a response close to 1 around the desired source's direction, while other ICA weight vectors should have a response close to 0 around that direction. Consequently, one of the response scores should be close to 1 while others should be close to 0. We will pick the ICA weight vector which yields the biggest response score to recover the desired source.

7.4 Simulation

We provide examples on microphone array speech processing to compare the performances of various algorithms. The image method is used to generate artificial room impulse response. Simulated room dimension is $[8, 5, 3.5]m$. We simulate an 8 element uniform linear array with $4cm$ inter-microphone spacing. The sources, both the signal of interest (SOI) and interference signals, are plane waves which exist in the same plane as the linear array. In the experiments, it is always assumed the look direction is the broadside direction of the array, i.e. 0°. Every source signal is a speech wave signal. The frequency domain ICA approach is employed. We use short time Fourier transform (STFT) to map the time domain signals into frequency domain based on consecutive frames. The ICA algorithms are applied on every frequency bin. The frame length is $0.3s$ (240 samples), with a step length of $0.125s$ (100 samples). A 256 points FFT is used on each frame. The performance of various algorithms is measured by the average performance factor across all frequency bins and the cepstral distance between the recovered signal's spectrum and the original desired source's spectrum. The performance factor is defined as, $\sum_i \frac{|p_i|}{\max_j |p_j|} - 1$, where $\mathbf{p} = \mathbf{w}^H A$, A is the mixing matrix on one frequency bin. The performance factor measures the degree the algorithm enhances the desired source and suppresses interference signals on one frequency bin. The cepstral distance is used because it is a perceptual metric commonly used in speech processing to measure distortion. Not only does it account for the interference and noise level, but it also detects spectrum shape distortion. Thereby, both permutation and scaling problem are taken into consideration by a single metric.

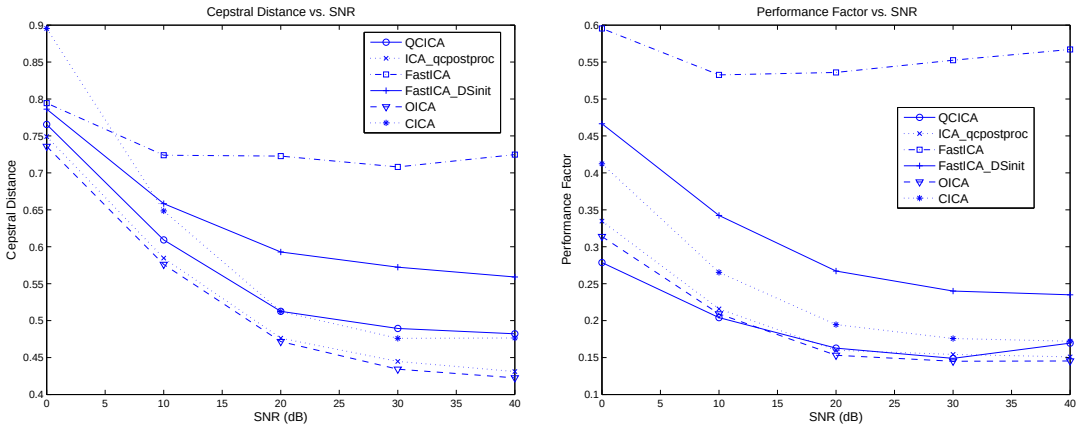
We use the following notation for each algorithm.

- QCICA: ICA with quadratic geometric constraint (sec.7.3.1)
- ICA_qcpostproc: Use geometric test as post-processing (sec.7.3.2)
- FastICA: FastICA with random initialization (sec.7.2.1)
- FastICA_DSinit: FastICA with Delay-and-Sum beamformer as initialization.
- OICA: optimum ICA, the permutation problem is solved manually assuming we know the correct permutation.
- CICA: ICA with linear geometric constraint (sec.7.2.2)

Example 1

There are two sources for this scenario, one is the desired source, the other is the interference signal. The desired source and the interference signal comes from direction 5° and 45° respectively. The assumed look direction is broadside, i.e. 0°, which means a 5° look direction error. Fig.7.1 demonstrates the performance of various algorithms versus SNR (signal to white noise ratio).

Assuming the number of sources is known accurately. The FastICA algorithm uses PCA as a preprocessing to make the mixing matrix square. When the FastICA algorithm use a random initialization, the performance is bad. This is caused by the different permutation in different frequency bins. To give an example, we observe 60 wrong permutations in a total of 128 frequency bins in one sample experiment. When the FastICA algorithm uses the delay-and-sum beamformer weight vector as the initialization, the performance improves but it is still not good enough as some frequency bins may still be subject to the permutation problem. The CICA use PCA preprocessing as well. It shows better performance than FastICA_DSinit. When the PCA preprocessing is not done properly with the right source number, the CICA algorithm was found to totally fail. The OICA corrects the permutation problem manually and can be taken as a baseline for the ICA algorithms. The proposed QCICA use all 8 channels in the optimization and the delay-and-sum beamformer weight vector is used for initialization. It does not use PCA preprocessing. Its performance is close to OICA's performance. We observe that the proposed ICA_qcpostproc has almost the same performance as the OICA algorithm.



(a) Cepstral distance vs. SNR (b) Mean performance factor vs. SNR

Figure 7.1: The performance of various ICA algorithms vs. SNR (2 sources)

Example 2:

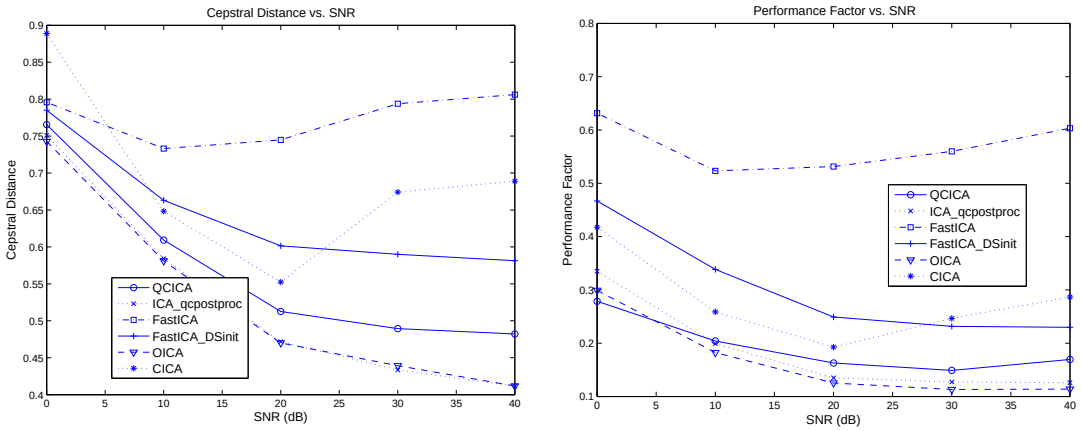
In this example, the scenario is the same as in Example 1 except that the assumed source number is 3. In other words, the number of sources is overestimated. Consequently, PCA preprocessing will employ 3 dimensions. Fig.7.2 demonstrates the performance versus SNR. The experiment results illustrate the performances of QCICA and ICA_qcpostproc are not affected much by the wrong information about the number of sources while CICA fails under such scenario.

Example 3:

Fig.7.3 demonstrates the performances versus SNR when two interference signals exist. The second interference signal comes from direction $-60°$. The number of sources is known accurately. All the other settings are the same as those in Example 1. Similar to Example 2, if the assumed source number is wrong (in this case, we assume 4 sources), the experiment results are shown in Fig.7.4. The experiment results are consistent with the those shown in Example 1 and 2.

Acknowledgement

The text of this chapter, in full, is a reprint of the material which has appeared in Wenyi Zhang, and Bhaskar D. Rao, "Combining Independent Component Analysis with Geometric Information and its Application to Speech Processing", *Acoustics, Speech,*

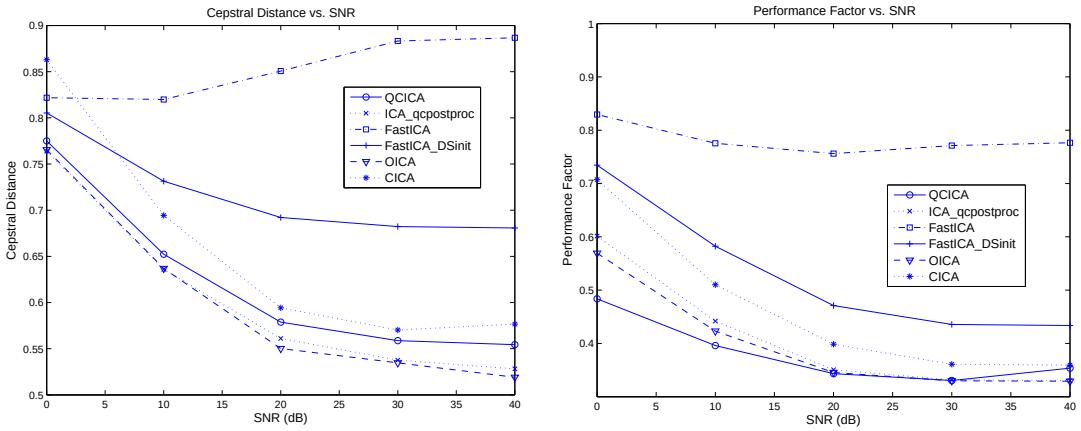


(a) Cepstral distance vs. SNR

(b) Mean performance factor vs. SNR

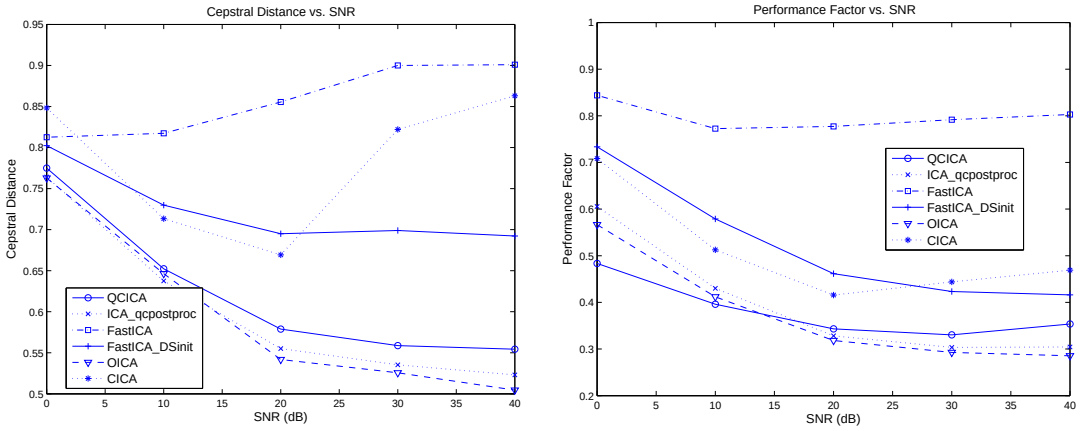
Figure 7.2: The performance of various ICA algorithms vs. SNR (2 sources exist, assumed source number is 3)

and Signal Processing, 2009. ICASSP 09. IEEE International Conference on, pp. 3065-3068, Apr. 2009.



(a) Cepstral distance vs. SNR (b) Mean performance factor vs. SNR

Figure 7.3: The performance of various ICA algorithms vs. SNR (3 sources)



(a) Cepstral distance vs. SNR (b) Mean performance factor vs. SNR

Figure 7.4: The performance of various ICA algorithms vs. SNR (3 sources exist, assumed source number is 4)

8 Conclusion

Compared with single channel speech processing, multi-microphone based speech processing methods are capable of high interference suppression in noisy environments because of their spatial filtering capability. Despite the many years of development of antenna array processing techniques in radar or sonar applications, microphone array speech processing brings up new challenges because of its specific characteristics. Therefore, new array processing algorithms for speech are desired. This dissertation develops novel microphone array speech processing methods in a variety of configurations and also analyzes and provides insights into existing popular techniques.

Although speech's special characteristics poses new challenges in microphone array processing, it may also provide specific opportunities which are not available in general signal processing. In this work, we consider DOA estimation for multiple speech signals based on only two microphones. Utilizing speech's specific properties such as the sparsity in the time-frequency domain and the sinusoidal modeling, we propose a two microphone based DOA estimation technique for multiple speech sources using the generalized mixture decomposition clustering algorithm. Voiced speech is sparse in the frequency domain and can be represented by sinusoidal tracks via sinusoidal modeling which provides high local SNR. By utilizing the inter-channel phase differences (IPD) between the dual channels on the sinusoidal tracks, the source localization of the mixed multiple speech sources is turned into a clustering problem on the IPD vs. frequency plot. The generalized mixture decomposition algorithm (GMDA) is used to cluster the groups of points corresponding to multiple sources and thus estimate the DOA of the sources. Experiments illustrate the proposed GMDA algorithm with the Laplacian noise model can estimate the number of sources accurately and exhibits smaller DOA estimation error than the baseline histogram based DOA estimation algorithm in various

scenarios including reverberant and additive white noise environments. Experiments suggest that appropriate power thresholding can be a simple and good approximation to the sinusoidal modeling, for the purpose of selecting time-frequency points with high local SNR, with slight loss in performance.

Data dependent adaptive beamformers have high resolution and interference rejection capability when the array steering vector is accurately known. However, the performance of the traditional adaptive beamformer can degrade severely in practice when there exist SOI steering vector errors. This problem should be addressed before the adaptive beamforming algorithms can be applied to process real world speech data. We compare and analyze recent developments in the academic world of adaptive beamforming. Experiments illustrate the performance of the Frost beamformer (optimum solution) can degrade severely when the steering vector errors exist, however, the Frost LMS algorithm is robust to the look direction error when the adaptation does not go through too many iterations. Motivated by these observations, we analyze the effect that the signal incidence angle has on the convergence rate of the Frost LMS beamforming weight vector. Our analysis confirms that the Frost LMS algorithm is robust to the look direction error when the adaptation does not go through too many iterations.

We develop a robust broadband adaptive beamforming algorithm which combines the robustness of the DS beamforming in the look direction with the high interference rejection capability of the conventional adaptive beamforming algorithm. A quadratic constraint is employed to deal with the uncertainty in the look direction. In order to address the ill-conditioning associated with the constraint matrix, a diagonal loading (DL) is added to the constraint matrix thereby ensuring a robust solution to the quadratic constraint beamforming problem. The advantage of adding DL to constraint matrix is that the constraint matrix is only determined by the geometry of the array thereby allowing the DL level to be chosen offline. This is superior to adding DL to the signal covariance matrix where the DL level has to be chosen online. It is shown that the diagonal loading is equivalent to an additional norm constraint without introducing it explicitly. We also develop an iterative algorithm (and corresponding adaptive algorithm) to solve for the robust beamformer coefficients. The developed algorithm is applied to the problem of beamforming using microphone arrays for speech recognition

and shown to be superior to existing algorithms.

Motivated by recent work in robust MVDR beamforming (RCB&DCRCB) [26], we develop variants of the constrained robust MVDR beamformer that attempt to limit the search in the underlying optimization problem to a feasible set of steering vectors thereby achieving improved performance. The robustness against steering vector error is provided through a spherical uncertainty set constraint, while a set of magnitude constraints is enforced on each element of the steering vector to better constrain the search in the space of feasible steering vectors. By appropriately changing the variables, the optimization problem is modified such that the need for the magnitude constraints are avoided. We also develop adaptive algorithms for the RCB and the time delay based robust MVDR beamformer. The adaptive algorithms have two updating steps. The first step updates the steering vector estimation or the time delay estimation; the second step updates the beamformer's weight vector given an estimated steering vector. The developed algorithms are tested in the context of speech enhancement using a microphone array.

Another class of promising multi-channel signal separation algorithms besides beamforming is independent component analysis (ICA). For separating convolutively mixed source signals, the frequency domain ICA approach is often used because it simplifies the time domain convolutive mixing problem into the instantaneous mixing ICA problem in each frequency bin. We examine and provide insights into the frequency domain ICA methods for source separation in reverberant environments. For the modeling of the source signals, we develop the concept of a dynamic random process to model the source signals. It formalizes the concept of signals that are stationary in a frame but exhibit dynamics at the frame level. Frame dynamics is an important characteristics of these signals and prove important to the success of the ICA methods. With suitable assumptions, the dynamic random process is stationary in the ensemble sense while a given realization may in an engineering sense exhibit 'non-stationarity'. We show for dynamic random processes, the unconditional distribution of the source signal in each frequency bin is a Gaussian scale mixture (GSM). The non-Gaussianity, which is critical to ICA, of the source signal in each frequency bin is shown to be a direct consequence of the frame dynamics. Furthermore, the independence between the unconditional dis-

tributions of the source signals in each frequency bin is related to the independence of the frame dynamics of the mixed time domain source signals. The GSM mathematical modeling is extended to the vector random processes formed by stacking the different frequency components of a source. This provides insights into the mathematical models suitable for the frequency domain independent vector analysis (IVA) type approaches. A special case of the distribution turns out to be the 'spherical distribution' employed in IVA source modeling providing support to their use in source separation. Concentrating on the bin-wise ICA methods, a significant contribution of the work is to show that signals modeled using GSM density can be separated using ICA even though they might be dependent on each other as long as the the frame dynamics of the source signals are different almost surely. We also analyze the stability conditions of the complex maximum likelihood ICA/IVA.

Lastly, in an attempt to make the best of ICA and beamforming methods, we propose two approaches for combining DOA information of the desired source with ICA algorithm to solve the permutation problem. The first approach is a new blind extraction algorithm with a soft quadratic geometric constraint. The quadratic constraint restricts the weighted square error between the desired and actual response of the processor over a small spatial uncertainty region chosen to deal with look direction uncertainty. Thereby the desired source is guaranteed to be conveyed to the output with little distortion and the negentropy maximization criterion is used to ensure that the other sources get suppressed at the output. The second approach employs a quadratic geometric test as a post-processing step to pickup the desired source after ICA processing. In every frequency bin, the ICA algorithm separates instantaneously mixed source signals, then the quadratic geometric test will pick up the desired source. An advantage of the proposed two approaches is that they do not require accurate knowledge of the number of sources in the mixtures to recover the desired source, in contrast, other geometric ICA approaches usually fail if the number of sources is not known accurately.

Bibliography

[1] H. L. V. Trees, *Optimum Array Processing (Detection, Estimation, and Modulation Theory, Part IV)*. Wiley-Interscience, 2002.

[2] Y. A. Huang and J. Benesty, *Adaptive Multichannel Time Delay Estimation Based on Blind System Identification for Acoustic Source Localization*. Springer, 2003, ch. 8, pp. 227–248.

[3] G. Jacovitti and G. Scarano, "Discrete time techniques for time delay estimation," *Signal Processing, IEEE Transactions on*, vol. 41, no. 2, pp. 525–533, Feb. 1993.

[4] F. Reed, P. Feintuch, and N. Bershad, "Time delay estimation using the lms adaptive filter – static behavior," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 29, no. 3, pp. 561–571, Jun. 1981.

[5] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *Acoustical Society of America Journal*, vol. 107, pp. 384–391, 2000.

[6] Y. Huang, J. Benesty, and G. Elko, "Adaptive eigenvalue decomposition algorithm for real time acoustic source localization system," in *Acoustics, Speech, and Signal Processing, 1999. ICASSP '99. Proceedings., 1999 IEEE International Conference on*, vol. 2, Mar. 1999, pp. 937–940.

[7] Y. Chan, R. Hattin, and J. Plant, "The least squares estimation of time delay and its use in signal detection," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 3, pp. 217–222, Jun. 1978.

[8] M. Brandstein and H. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., Proceedings., 1997 IEEE International Conference on*, vol. 1, Apr. 1997, pp. 375–378.

[9] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *Signal Processing, IEEE Transactions on*, vol. 51, no. 1, pp. 11–24, Jan. 2003.

[10] G. Carter, "Coherence and time delay estimation," *Proceedings of the IEEE*, vol. 75, no. 2, pp. 236–255, Feb. 1987.

[11] D. Hertz, "Time delay estimation by combining efficient algorithms and generalized cross-correlation methods," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 34, no. 1, pp. 1–7, Feb. 1986.

[12] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[13] G. Carter, A. Nuttall, and P. Cable, "The smoothed coherence transform," *Proceedings of the IEEE*, vol. 61, no. 10, pp. 1497–1498, Oct. 1973.

[14] T. Gustafsson, B. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 6, pp. 791–803, Nov. 2003.

[15] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: an overview," *EURASIP Journal on Applied Signal Processing*, vol. 2006, no. 1, pp. 170–170, 2006.

[16] P. Aarabi, "Self-localizing dynamic microphone arrays," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 32, no. 4, pp. 474–484, Nov. 2002.

[17] C. Zhang, D. Florencio, and Z. Zhang, "Why does phat work well in lownoise, reverberative environments?" in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-April 4 2008, pp. 2565–2568.

[18] J. Blauert, *Spatial Hearing, the Psychophysics of Human Sound Localization*. Cambridge, MA: MIT Press, 1983.

[19] I. Frost, O.L., "An algorithm for linearly constrained adaptive array processing," *Proceedings of the IEEE*, vol. 60, no. 8, pp. 926–935, Aug. 1972.

[20] J. Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.

[21] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 35, no. 10, pp. 1365–1376, Oct 1987.

[22] K. Buckley and L. Griffiths, "An adaptive generalized sidelobe canceller with derivative constraints," *Antennas and Propagation, IEEE Transactions on*, vol. 34, no. 3, pp. 311–319, Mar. 1986.

[23] M. Er, "Adaptive antenna array under directional and spatial derivative constraints," *Microwaves, Antennas and Propagation, IEE Proceedings H*, vol. l35, no. 6, pp. 414–419, Dec 1988.

[24] K. Bell, Y. Ephraim, and H. Van Trees, "A bayesian approach to robust adaptive beamforming," *Signal Processing, IEEE Transactions on*, vol. 48, no. 2, pp. 386–398, Feb 2000.

[25] S. Vorobyov, A. Gershman, and Z.-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: a solution to the signal mismatch problem," *Signal Processing, IEEE Transactions on*, vol. 51, no. 2, pp. 313–324, Feb 2003.

[26] J. Li, P. Stoica, and Z. Wang, "Doubly constrained robust capon beamformer," *Signal Processing, IEEE Transactions on*, vol. 52, no. 9, pp. 2407–2423, Sept. 2004.

[27] Y. Selén, R. Abrahamsson, and P. Stoica, "Automatic robust adaptive beamforming via ridge regression," *Signal Processing*, vol. 88, no. 1, pp. 33–49, 2008.

[28] P. Stoica, J. Li, X. Zhu, and J. Guerci, "On using a priori knowledge in space-time adaptive processing," *Signal Processing, IEEE Transactions on*, vol. 56, no. 6, pp. 2598–2602, Jun. 2008.

[29] J. Yang, X. Ma, C. Hou, and Y. Liu, "Automatic generalized loading for robust adaptive beamforming," *Signal Processing Letters, IEEE*, vol. 16, no. 3, pp. 219–222, Mar. 2009.

[30] W. Zhang and B. Rao, "Robust adaptive beamformer with feasibility constraint on the steering vector," in *European Signal Processing Conference*, Sep. 2006.

[31] J. Guerci, "Theory and application of covariance matrix tapers for robust adaptive beamforming," *Signal Processing, IEEE Transactions on*, vol. 47, no. 4, pp. 977–985, Apr 1999.

[32] A. Gershman, U. Nickel, and J. Bohme, "Adaptive beamforming algorithms with robustness against jammer motion," *Signal Processing, IEEE Transactions on*, vol. 45, no. 7, pp. 1878–1885, Jul 1997.

[33] C.-Y. Tseng, "Minimum variance beamforming with phase-independent derivative constraints," *Antennas and Propagation, IEEE Transactions on*, vol. 40, no. 3, pp. 285–294, Mar 1992.

[34] O. Besson, A. Monakov, and C. Chalus, "Signal waveform estimation in the presence of uncertainties about the steering vector," in *Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2, Nov. 2003, pp. 1873–1877 Vol.2.

[35] W.-K. Ma, P.-C. Ching, and B.-N. Vo, "Crosstalk resilient interference cancellation in microphone arrays using capon beamforming," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 5, pp. 468–477, Sept. 2004.

[36] P. Aarabi and G. Shi, "Phase-based dual-microphone robust speech enhancement," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, no. 4, pp. 1763–1773, Aug. 2004.

[37] M. Seltzer, B. Raj, and R. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 5, pp. 489–498, Sept. 2004.

[38] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function gsc and postfiltering," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 6, pp. 561–571, Nov. 2004.

[39] J.-H. Lee and Y.-H. Lee, "Two-dimensional adaptive array beamforming with multiple beam constraints using a generalized sidelobe canceller," *Signal Processing, IEEE Transactions on*, vol. 53, no. 9, pp. 3517–3529, Sept. 2005.

[40] M. Er and A. Cantoni, "A unified approach to the design of robust narrow-band antenna array processors," *Antennas and Propagation, IEEE Transactions on*, vol. 38, no. 1, pp. 17–23, Jan 1990.

[41] S. Applebaum and D. Chapman, "Adaptive arrays with main beam constraints," *Antennas and Propagation, IEEE Transactions on*, vol. 24, no. 5, pp. 650–662, Sep 1976.

[42] K. Buckley, "Spatial/spectral filtering with linearly constrained minimum variance beamformers," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 35, no. 3, pp. 249–266, Mar 1987.

[43] J. Hudson, *Adaptive Array Principles*. New York: Peter Peregrinus Ltd., 1981.

[44] J. Hiemstra, M. Weippert, J. Goldstein, and T. Pratt, "Application of the l-curve technique to loading level determination in adaptive beamforming," *Signals, Systems and Computers, 2002. Conference Record of the Thirty-Sixth Asilomar Conference on*, vol. 2, pp. 1261–1266, Nov. 2002.

[45] S. Shahbazpanahi, A. Gershman, Z.-Q. Luo, and K. M. Wong, "Robust adaptive beamforming for general-rank signal models," *Signal Processing, IEEE Transactions on*, vol. 51, no. 9, pp. 2257–2269, Sept. 2003.

[46] R. Lorenz and S. Boyd, "Robust minimum variance beamforming," *Signal Processing, IEEE Transactions on*, vol. 53, no. 5, pp. 1684–1696, May 2005.

[47] J. Li, P. Stoica, and Z. Wang, "On robust capon beamforming and diagonal loading," *Signal Processing, IEEE Transactions on*, vol. 51, no. 7, pp. 1702–1715, July 2003.

[48] J. Li, L. Du, and P. Stoica, "Fully automatic computation of diagonal loading levels for robust adaptive beamforming," *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 2325–2328, Apr. 2008.

[49] J. Ward, H. Cox, and S. Kogon, "A comparison of robust adaptive beamforming algorithms," *Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on*, vol. 2, pp. 1340–1344, Nov. 2003.

[50] B. Widrow, K. Duvall, R. Gooch, and W. Newman, "Signal cancellation phenomena in adaptive antennas: Causes and cures," *Antennas and Propagation, IEEE Transactions on*, vol. 30, no. 3, pp. 469–478, May 1982.

[51] T.-J. Shan and T. Kailath, "Adaptive beamforming for coherent signals and interference," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 3, pp. 527–536, Jun 1985.

[52] A. Luthra, "A solution to the adaptive nulling problem with a look-direction constraint in the presence of coherent jammers," *Antennas and Propagation, IEEE Transactions on*, vol. 34, no. 5, pp. 702–710, May 1986.

[53] M. Agrawal and S. Prasad, "Robust adaptive beamforming for wide-band, moving, and coherent jammers via uniform linear arrays," *Antennas and Propagation, IEEE Transactions on*, vol. 47, no. 8, pp. 1267–1275, Aug 1999.

[54] L. Tong and S. Perreau, "Multichannel blind identification: from subspace to maximum likelihood methods," *Proceedings of the IEEE*, vol. 86, no. 10, pp. 1951–1968, Oct 1998.

[55] J. Tugnait and B. Huang, "Multistep linear predictors-based blind identification and equalization of multiple-input multiple-output channels," *Signal Processing, IEEE Transactions on*, vol. 48, no. 1, pp. 26–38, Jan 2000.

[56] X.-L. Li and X.-D. Zhang, "A family of generalized constant modulus algorithms for blind equalization," *Communications, IEEE Transactions on*, vol. 54, no. 11, pp. 1913–1917, Nov. 2006.

[57] P. Comon, "Independent component analysis, a new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, 1994.

[58] A. Hyvrinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[59] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *Neural Networks, IEEE Transactions on*, vol. 10, no. 3, pp. 626–634, May 1999.

[60] E. Bingham and A. Hyvarinen, "A fast fixed point algorithm for independent component analysis of complex valued signals," *Int. J. Neural Networks*, vol. 10, no. 1, pp. 1–8, 2000.

[61] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[62] D. T. Pham and P. Garat, "Blind separation of mixture of independent sources through a quasi-maximum likelihood approach," *Signal Processing, IEEE Transactions on*, vol. 45, no. 7, pp. 1712–1725, Jul. 1997.

[63] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, pp. 251–276, 1998.

[64] K. J. Pope and R. E. Bogner, "Blind signal separation i: Linear, instantaneous combinations," *Digital Signal Processing*, vol. 6, no. 1, pp. 5–16, 1996.

[65] ——, "Blind signal separation ii: Linear, convolutive combinations," *Digital Signal Processing*, vol. 6, no. 1, pp. 17–28, 1996.

[66] S. Choi, A. Cichocki, and A. Beloucharni, "Second order nonstationary source separation," *J. VLSI Signal Process. Syst.*, vol. 32, no. 1-2, pp. 93–104, 2002.

[67] L. Parra and C. Spence, "Separation of nonstationary natural signals," in *Independent Components Analysis: Principles and Practice*, C. Roberts, Ed. Cambridge Univ. Press, 2000.

[68] S. ichi Amari, T. ping Chen, and A. Cichocki, "Stability analysis of learning algorithms for blind source separation," *Neural Networks*, vol. 10, no. 8, pp. 1345 – 1351, 1997.

[69] R. Everson and S. Roberts, "Non-stationary independent component analysis," in *The 8th International Conference on Artificial Neural Networks*, Berlin: Springer-Verlag, 1999, pp. 503–508.

[70] S. Choi and A. Cichocki, "Blind separation of nonstationary and temporally correlated sources from noisy mixtures," *Neural Networks for Signal Processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop*, vol. 1, pp. 405–414, 2000.

[71] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *Signal Processing, IEEE Transactions on*, vol. 44, no. 12, pp. 3017–3030, Dec 1996.

[72] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, Oct 2001.

[73] T.-W. Lee, M. S. Lewicki, and T. J. Sejnowski, "ICA mixture models for unsupervised classification of non-gaussian classes and automatic context switching in blind signal separation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 10, pp. 1078–1089, 2000.

[74] T.-W. Lee and M. Lewicki, "Unsupervised image classification, segmentation, and enhancement using ICA mixture models," *Image Processing, IEEE Transactions on*, vol. 11, no. 3, pp. 270–279, Mar 2002.

[75] N. Mitianoudis and M. E. Davies, "Audio source separation: Solutions and problems," *International Journal of Adaptive Control and Signal Processing*, vol. 18, pp. 299–314, 2004.

[76] S. Choi, A. Cichocki, and S. Amari, "Equivariant nonstationary source separation," *Neural Networks*, vol. 15, no. 1, pp. 121–130, 2002.

[77] H. Sawada, S. Winter, R. Mukai, S. Araki, and S. Makino, "Estimating the number of sources for frequency-domain blind source separation," in *ICA '04*, ser. Lecture Notes in Computer Science, C. G. Puntonet and A. Prieto, Eds., vol. 3195. Springer, 2004, pp. 610–617.

[78] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 2, pp. 109–116, Mar. 2003.

[79] H. Attias, "Independent factor analysis," *Neural Computation*, vol. 11, pp. 803–851, 1999.

[80] J. A. Palmer, K. Kreutz-Delgado, and S. Makeig, "Super-gaussian mixture source model for ICA," in *Independent Component Analysis and Blind Signal Separation*. Berlin: Springer-Verlag, 2006, pp. 854–861.

[81] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley and Sons, 2001.

[82] A. Cichocki and S.-i. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. New York, NY, USA: John Wiley & Sons, Inc., 2002.

[83] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in *Springer Handbook*

of Speech Processing. Springer Press, Sep. 2007. [Online]. Available: http://www2.imm.dtu.dk/pubdb/p.php?4924

[84] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, pp. 21–34, 1998.

[85] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Evaluation of blind signal separation method using directivity pattern under reverberant conditions," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings., 2000 IEEE International Conference on*, 2000, pp. 3140–3143.

[86] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1135–1146, 2003.

[87] L. Parra and C. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 6, pp. 352–362, Sep. 2002.

[88] M. Knaak, S. Araki, and S. Makino, "Geometrically constrained independent component analysis," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 715–726, Feb. 2007.

[89] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 3, pp. 204–215, May 2003.

[90] W. Zhang and B. Rao, "Combining independent component analysis with geometric information and its application to speech processing," in *Acoustics, Speech, and Signal Processing, 2009. ICASSP '09. Proceedings., 2009 IEEE International Conference on*, Apr. 2009.

[91] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," in *International Conference on Independent Component Analysis and Blind Source Separation*. IEEE, 2003, pp. 505–510.

[92] C. Servière and D.-T. Pham, "A novel method for permutation correction in frequency-domain in blind separation of speech mixtures," in *ICA '04*, 2004, pp. 807–815.

[93] C. Servière and D. T. Pham, "Permutation correction in the frequency domain in blind separation of speech mixtures," *EURASIP Journal on Applied Signal Processing*, vol. 2006, no. 1, pp. 177–177, 2006.

[94] N. Mitianoudis and M. Davies, "Audio source separation of convolutive mixtures," *Audio and Speech Processing, IEEE Transactions on*, vol. 11, pp. 489–497, 2003.

[95] M. Ikram and D. Morgan, "Permutation inconsistency in blind speech separation: investigation and solutions," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 1, pp. 1–13, Jan. 2005.

[96] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," *Lecture Notes in Computer Science*, pp. 601–608, 2006.

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

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

[99] I. Lee and T.-W. Lee, "On the assumption of spherical symmetry and sparseness for the frequency-domain speech model," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1521–1528, Jul. 2007.

[100] I. Lee, J. Hao, and T.-W. Lee, "Adaptive independent vector analysis for the separation of convoluted mixtures using em algorithm," *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 145–148, Apr. 2008.

[101] G.-J. Jang, I. Lee, and T.-W. Lee, "Independent vector analysis using non-spherical joint densities for the separation of speech signals," *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 2, pp. II–629–II–632, Apr. 2007.

[102] A. Masnadi-Shirazi, W. Zhang, and B. D. Rao, "Glimpsing IVA: A framework for overcomplete/complete/undercomplete convolutive source separation," *submitted to Audio, Speech and Language Processing, IEEE Transactions on, special issue on reverberant speech: methodologies and applications*.

[103] I. Lee, T. Kim, and T.-W. Lee, "Complex FastIVA: A robust maximum likelihood approach of MICA for convolutive BSS," *Lecture Notes in Computer Science*, pp. 625–632, 2006.

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

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

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

[107] S. Choi and A. Cichocki, "Blind separation of nonstationary sources in noisy mixtures," *Electronics Letters*, vol. 36, no. 9, pp. 848–849, Apr. 2000.

[108] A. Hyvarinen, "Blind source separation by nonstationarity of variance: a cumulant-based approach," *Neural Networks, IEEE Transactions on*, vol. 12, no. 6, pp. 1471–1474, Nov 2001.

[109] I. Russell, J. Xi, A. Mertins, and J. Chicharo, "Blind source separation of nonstationary convolutively mixed signals in the subband domain," *Acoustics, Speech, and Signal Processing, 2004. ICASSP '04. Proceedings., IEEE International Conference on*, vol. 5, pp. V–481–4 vol.5, May 2004.

[110] Z. Koldovsky, J. Malek, P. Tichavsky, Y. Deville, and S. Hosseini, "Blind separation of piecewise stationary non-gaussian sources," *Signal Processing*, vol. In Press, Corrected Proof, pp. –, 2009.

[111] W. Lindemann, "Extension of a binaural cross-correlation model by contralateral inhibition. i. simulation of lateralization for stationary signals," *Acoustical Society of America Journal*, vol. 80, pp. 1608–1622, 1986.

[112] M. Bodden, "Modeling human sound source localization and the cocktail-party-effect," *Acta Acust.*, vol. 1, pp. 43–55, 1993.

[113] D. Banks, "Localisation and separation of simultaneous voices with two microphones," *Communications, Speech and Vision, IEE Proceedings I*, vol. 140, no. 4, pp. 229–234, Aug. 1993.

[114] T. Nakatani, M. Goto, and H. G. Okuno, "Localization by harmonic structure and its application to harmonic sound stream segregation," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP '96. Proceedings., 1996 IEEE International Conference on*, Washington, DC, USA, May 1996, pp. 653–656.

[115] T. Nakatani and H. G. Okuno, "Harmonic sound stream segregation using localization and its application to speech stream segregation," *Speech Communication*, vol. 27, no. 3-4, pp. 209–222, 1999.

[116] C. Liu, B. C. Wheeler, W. D. O'Brien, Jr., R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *Acoustical Society of America Journal*, vol. 108, pp. 1888–1905, 2000.

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

[118] S. Araki, H. Sawada, R. Mukai, and S. Makino, "DOA estimation for multiple sparse sources with normalized observation vector clustering," *Acoustics, Speech and Signal Processing, 2006. ICASSP '06. Proceedings. 2006 IEEE International Conference on*, vol. 5, pp. V33–V36, May 2006.

[119] K. Suyama, K. Takahashi, and R. Hirabayashi, "A robust technique for sound source localization in consideration of room capacity," *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pp. 63–66, 2001.

[120] M. I. Mandel, D. P. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds. Cambridge, MA: MIT Press, 2007, pp. 953–960.

[121] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[122] U. Madhow, *Fundamentals of Digital Communication*. Cambridge University Press, 2008.

[123] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*. Berlin: Springer-Verlag, 1972.

[124] S. T. Roweis, "One microphone source separation," in *Advances in Neural Information Processing Systems*, 2000, pp. 793–799.

[125] A. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 6, pp. 1766–1776, 2007.

[126] B. C. J. MOORE, *An Introduction to the Psychology of Hearing*, 3rd ed. Academic Press, 1994.

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

[128] D. Ellis, "http://www.ee.columbia.edu/ dpwe/resources/matlab/sinemodel/."

[129] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Third Edition*. Orlando, FL, USA: Academic Press, Inc., 2006.

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

[131] Defense, *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT)*, speech disc cd1-1.1 ed., DARPA-ISTO, 1990.

[132] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, Aug. 2006.

[133] B. Champagne, S. Bedard, and A. Stephenne, "Performance of time-delay estimation in the presence of room reverberation," *Speech and Audio Processing, IEEE Transactions on*, vol. 4, no. 2, pp. 148–152, Mar. 1996.

[134] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Acoustical Society of America Journal*, vol. 65, pp. 943–950, Apr. 1979.

[135] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall PTR, Apr. 1993.

[136] "Multichannel numbers corpus distribution. http://www.idiap.ch/speech/."

[137] M. Er and A. Cantoni, "A new set of linear constraints for broad-band time domain element space processors," *Antennas and Propagation, IEEE Transactions on*, vol. 34, no. 3, pp. 320–329, Mar 1986.

[138] ——, "A new approach to the design of broad-band element space antenna array processors," *Oceanic Engineering, IEEE Journal of*, vol. 10, no. 3, pp. 231–240, Jul. 1985.

[139] Q. Zou, Z. L. Yu, and Z. Lin, "A robust algorithm for linearly constrained adaptive beamforming," *Signal Processing Letters, IEEE*, vol. 11, no. 1, pp. 26–29, Jan. 2004.

[140] A. El-Keyi, T. Kirubarajan, and A. Gershman, "Wideband robust beamforming based on worst-case performance optimization," *Statistical Signal Processing, 2005 IEEE/SP 13th Workshop on*, pp. 265–270, July 2005.

[141] Y. Zheng, R. Goubran, and M. El-Tanany, "Robust near-field adaptive beamforming with distance discrimination," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 5, pp. 478–488, Sep. 2004.

[142] W. Zhang and B. Rao, "Robust broadband beamformer with diagonally loaded constraint matrix and its application to speech recognition," in *Acoustics, Speech, and Signal Processing, 2006. ICASSP '06. Proceedings., 2006 IEEE International Conference on*, May 2006, pp. 785–788.

[143] C.-Y. Tseng and L. J. Griffiths, "Adaptive broadband beamforming using phase-independent derivative constraint techniques," in *ICASSP '93: Proceedings of the Acoustics, Speech, and Signal Processing, 1993. ICASSP-93 Vol 4., 1993 IEEE International Conference on*. Washington, DC, USA: IEEE Computer Society, 1993, pp. 536–539.

[144] S. Simanapalli and M. Kaveh, "Broadband focusing for partially adaptive beamforming," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 30, no. 1, pp. 68–80, Jan 1994.

[145] A. Krieger and E. Masry, "Constrained adaptive filtering algorithms: asymptotic convergence properties for dependent data," *Information Theory, IEEE Transactions on*, vol. 35, no. 6, pp. 1166–1176, Nov 1989.

[146] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *IEEE Workshop on Neural Networks for Signal Processing, Kyoto, Japan*, Sep. 1996, pp. 423–432.

[147] T.-W. Lee, A. Bell, and R. Orglmeister, "Blind source separation of real world signals," *Neural Networks,1997., International Conference on*, vol. 4, pp. 2129–2134 vol.4, Jun 1997.

[148] T.-W. Lee, A. J. Bell, and R. H. Lambert, "Blind separation of delayed and convolved sources," in *Advances in Neural Information Processing Systems*, 1997.

[149] W. Feller, *An Introduction to Probability Theory and Its Applications*. Wiley, Jan. 1968, vol. 1.

[150] D. Andrews and C. Mallows, "Scale mixtures of normal distributions," *Journal of the Royal Statistical Society*, vol. 36, pp. 99–102, 1974.

[151] T. Eltoft, T. Kim, and T.-W. Lee, "Multivariate scale mixture of gaussians modeling," in *Independent Component Analysis and Blind Signal Separation*. Berlin: Springer-Verlag, 2006, pp. 799–806.

[152] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," in *Advances Neural Information Processing Systems*. MIT Press, 2000, pp. 855–861.

[153] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of gaussians in the wavelet domain," *Image Process, IEEE Transactions on*, vol. 12, pp. 1338–1351, 2003.

[154] H. Brehm and W. Stammler, "Description and generation of spherically invariant speech-model signals," *Signal Process.*, vol. 12, no. 2, pp. 119–141, 1987.

[155] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in modeling and analyzing natural images," *Applied and Computational Harmonic Analysis*, vol. 11, no. 1, pp. 89–123, Jul. 2001.

[156] J. A. Palmer, K. Kreutz-Delgado, and S. Makeig, "Probabilistic formulation of independent vector analysis using complex gaussian scale mixtures," in *ICA '09: Proceedings of the 8th International Conference on Independent Component Analysis and Signal Separation*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 90–97.

[157] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. New York: McGraw-Hill, 1991.

[158] A. Hyvarinen and P. Hoyer, "Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces," *Neural Computation*, vol. 12, no. 7, pp. 1705–1720, 2000.

[159] A. Hyvarinen, P. O. Hoyer, and M. Inki, "Topographic independent component analysis," *Neural Computation*, vol. 13, no. 7, pp. 1527–1558, 2001.

[160] R. Buccigrossi and E. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *Image Processing, IEEE Transactions on*, vol. 8, no. 12, pp. 1688–1701, Dec. 1999.

[161] Y. Karklin and M. S. Lewicki, "A hierarchical bayesian model for learning nonlinear statistical regularities in nonstationary natural signals," *Neural Comput.*, vol. 17, no. 2, pp. 397–423, 2005.

[162] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, March 2004.

[163] K. Matsuoka, M. Ohya, and M. Kawamoto, "A neural net for blind separation of nonstationary signals," *Neural Networks*, vol. 8, no. 3, pp. 411–419, 1995.

[164] L. Parra and C. Spence, "On-line convolutive blind source separation of non-stationary signals," *Journal of VLSI Signal Processing Systems*, vol. 26, no. 1-2, pp. 39–46, 2000.

[165] B. Krongold and D. Jones, "Blind source separation of nonstationary convolutively mixed signals," *Statistical Signal and Array Processing, 2000. Proceedings of the Tenth IEEE Workshop on*, pp. 53–57, 2000.

[166] T. ping Chen and A. Cichocki, "Stability analysis of adaptive blind source separation," *Neural Networks*, vol. 10, pp. 1345–1351, 1997.

[167] J.-F. Cardoso, "On the stability of source separation algorithms," *J. VLSI Signal Process. Syst.*, vol. 26, no. 1-2, pp. 7–14, 2000.

[168] S. Choi, A. Cichocki, and S. Amari, "Local stability analysis of flexible independent component analysis algorithm," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, vol. 6, 2000, pp. 3426–3429.

[169] H. Li and T. Adali, "Stability analysis of complex maximum likelihood ICA using wirtinger calculus," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-April 4 2008, pp. 1801–1804.

[170] D. H. Brandwood, "A complex gradient operator and its application in adaptive array theory," *IEE Proceedings F: Communications Radar and Signal Processing*, vol. 130, pp. 11–16, Feb. 1983.

[171] A. van den Bos, "Complex gradient and hessian," *Vision, Image and Signal Processing, IEE Proceedings* -, vol. 141, no. 6, pp. 380–383, Dec. 1994.

[172] K. Kreutz-Delgado and Y. Isukapalli, "Use of the newton method for blind adaptive equalization based on the constant modulus algorithm," *Signal Processing, IEEE Transactions on*, vol. 56, no. 8, pp. 3983–3995, Aug. 2008.

[173] W. Zhang and B. Rao, "Two microphone based direction of arrival estimation for multiple speech sources using spectral properties of speech," in *Acoustics, Speech, and Signal Processing, 2009. ICASSP '09. Proceedings., 2009 IEEE International Conference on*, Apr. 2009, pp. 2193–2196.

[174] W. Zhang and B. D. Rao, "A two microphone based approach for source localization of multiple speech sources," *Audio, Speech and Language Processing, IEEE Transactions on*, accepted for publication.

[175] ——, "Combining independent component analysis with geometric information and its application to speech processing," in *Acoustics, Speech, and Signal Processing, 2009. ICASSP 09. IEEE International Conference on*, 2009, pp. 3065–3068.

[176] W. Zhang, A. Masnadi-Shirazi, and B. D. Rao, "Insights into the frequency domain ICA/IVA approach," *submitted to Audio, Speech and Language Processing, IEEE Transactions on, special issue on reverberant speech: methodologies and applications*.

[177] A. Masnadi-Shirazi, W. Zhang, and B. D. Rao, "Glimpsing independent vector analysis: Separating more sources than sensors using active and inactive states," in *Acoustics, Speech, and Signal Processing, 2010. ICASSP 2010. IEEE International Conference on*, 2010.