Enhanced Independent Vector Analysis for Audio Separation in a Room Environment

by

Yanfeng Liang

A doctoral thesis submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy (PhD), from Loughborough University.

September 2013

Advanced Signal Processing Group,
School of Electronic, Electrical and Systems Engineering,
Loughborough University, Loughborough
Leicestershire, UK, LE11 3TU.

CERTIFICATE OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this thesis, that the original work is my own except as specified in acknowledgements or in footnotes, and that neither the thesis nor the original work contained therein has been submitted to this or any other institution for a degree.

..... (Signed)

..... (candidate)

I dedicate this thesis to my loving parents.

Abstract

Independent vector analysis (IVA) is studied as a frequency domain blind source separation method, which can theoretically avoid the permutation problem by retaining the dependency between different frequency bins of the same source vector while removing the dependency between different source vectors. This thesis focuses upon improving the performance of independent vector analysis when it is used to solve the audio separation problem in a room environment.

A specific stability problem of IVA, i.e. the block permutation problem, is identified and analyzed. Then a robust IVA method is proposed to solve this problem by exploiting the phase continuity of the unmixing matrix. Moreover, an auxiliary function based IVA algorithm with an overlapped chain type source prior is proposed as well to mitigate this problem.

Then an informed IVA scheme is proposed which combines the geometric information of the sources from video to solve the problem by providing an intelligent initialization for optimal convergence. The proposed informed IVA algorithm can also achieve a faster convergence in terms of iteration numbers and better separation performance. A pitch based evaluation method is defined to judge the separation performance objectively when the information describing the mixing matrix and sources is missing.

In order to improve the separation performance of IVA, an appropriate multivariate source prior is needed to better preserve the dependency structure within the source vectors. A particular multivariate generalized

Gaussian distribution is adopted as the source prior. The nonlinear score function derived from this proposed source prior contains the fourth order relationships between different frequency bins, which provides a more informative and stronger dependency structure compared with the original IVA algorithm and thereby improves the separation performance.

Copula theory is a central tool to model the nonlinear dependency structure. The t copula is proposed to describe the dependency structure within the frequency domain speech signals due to its tail dependency property, which means if one variable has an extreme value, other variables are expected to have extreme values. A multivariate student's t distribution constructed by using a t copula with the univariate student's t marginal distribution is proposed as the source prior. Then the IVA algorithm with the proposed source prior is derived.

The proposed algorithms are tested with real speech signals in different reverberant room environments both using modelled room impulse response and real room recordings. State-of-the-art criteria are used to evaluate the separation performance, and the experimental results confirm the advantage of the proposed algorithms.

Contents

1 INTRODUCTION **25**

1.1 Cocktail Party Problem 25

1.2 Blind Source Separation 27

1.3 Aim and Objectives 32

1.4 Thesis Outline 33

2 FUNDAMENTALS OF INDEPENDENT VECTOR ANALYSIS **36**

2.1 Introduction 36

2.2 Convolutive Blind Source Separation 37

2.3 Second Order Statistic Solution to CBSS 39

2.4 Independent Component Analysis 41

2.5 Independent Vector Analysis 45

2.5.1 Natural Gradient IVA 47

2.5.2 Adaptive Step Size Natural Gradient IVA 52

2.5.3 Fast fixed-point IVA 60

2.5.4 Auxiliary Function Based IVA 62

2.5.5 Summary 67

3 BLOCK PERMUTATION PROBLEM OF INDEPENDENT VECTOR ANALYSIS **69**

3.1 Introduction 69

3.2 Block Permutation of IVA 70
3.3 Robust IVA Exploiting Phase Continuity of the Unmixing Matrix 73
3.3.1 Phase Continuity of the Unmixing Matrix 73
3.3.2 Robust IVA Based on Phase Continuity of the Unmixing Matrix 74
3.3.3 Experimental Results 76
3.4 Overcoming Block Permutation by Using an Improved Dependency Model 83
3.4.1 Block Permutation for AuxIVA 84
3.4.2 Overlapped Chain Type Dependency Model for AuxIVA 86
3.4.3 Experimental Results 87
3.5 Summary 89

4 INFORMED INDEPENDENT VECTOR ANALYSIS 91
4.1 Introduction 91
4.2 Block Permutation Problem of FastIVA 93
4.3 Audio Video Based FastIVA 96
4.4 Pitch Based Evaluation For Real Recordings 98
4.5 Experiments and Results 101
4.5.1 Experimental Demonstration of the Block Permutation Problem 101
4.5.2 Experiments in Noisy and Reverberant Room Environment 105
4.5.3 Experiments by Using the Real Room Recordings 109
4.6 Summary 112

5 IVA WITH MULTIVARIATE GENERALIZED GAUSSIAN SOURCE PRIOR 115
5.1 Introduction 115

5.2 Multivariate Generalized Gaussian Source Prior 117
5.3 NG-IVA with the Proposed Source Prior 119
5.4 FastIVA with the Proposed Source Prior 123
5.5 AuxIVA with the Proposed Source Prior 124
5.6 Experimental Results 125
5.7 Summary 132

6 COPULA BASED INDEPENDENT VECTOR ANALYSIS WITH THE MULTIVARIATE STUDENT'S T SOURCE PRIOR 134

6.1 Introduction 134
6.2 Copula Introduction 135
6.3 Dependency within Frequency Domain Speech Signals 140
6.4 IVA with the Multivariate Student's t Source Prior 143
6.5 Experimental Results 148
6.5.1 Experiment in Low Reverberation Room Environment 149
6.5.2 Experiment in Different Reverberant Room Environments 150
6.5.3 Experiment by Using the Real Room Recordings 157
6.6 Summary 158

7 CONCLUSIONS AND FUTURE WORK 160

7.1 Conclusions 160
7.2 Future Work 162

Statement of Originality

The contributions of this thesis are mainly on the improvement of independent vector analysis (IVA) for speech separation in a real room environment. The novelty of the contributions is supported by the following international journal and conference papers.

In Chapter 2, an adaptive step size IVA algorithm is proposed, which can tune the step size adaptively to achieve a faster convergence in terms of iteration number. This has been published in:

1. Y. Liang, S. M. Naqvi and J. A. Chambers, 'Adaptive step size independent vector analysis for blind source separation', IEEE DSP 2011, Corfu, Greece, pp. 1-6, 2011.

In Chapter 3, a specific permutation problem is highlighted and analyzed, i.e. the block permutation problem. Then two solutions for this problem are proposed. The first solution is exploiting the phase continuity of the unmixing matrix to adjust the misalignments to retain the permutation consistently across all the frequency bins. Another solution is adopting the overlapped chain type source prior for IVA. The results of these two solutions have been published in:

2. Y. Liang, S. M. Naqvi and J. A. Chambers, 'Robust independent vector analysis on exploiting phase continuity of the unmixing matrix', in Proc. IEEE ISSPA 2012, Montreal, Canada, pp. 321-326, 2012.

3. Y. Liang, S. M. Naqvi and J. A. Chambers, 'Overcoming block permutation problem in frequency domain blind source separation when using AuxIVA algorithm', Electronics Letters, Vol. 48, No. 8, pp. 460-462, 2012.

In Chapter 4, an informed IVA scheme is proposed, which exploits prior information, i.e. the geometric information captured from the video, to set

an intelligent initialization for the fast fixed-point IVA algorithm to achieve a faster and better separation performance. A new pitch based objective evaluation criterion is also proposed to judge the separation performance when the mixing matrix and source signals are both impossible to obtain. The results of this scheme are presented in:

4. Y. Liang, S. M. Naqvi and J. A. Chambers, 'Audio video based fast fixed-point independent vector analysis for multisource separation in a room environment', EURASIP Journal on Advances in Signal Processing, 2012:183, pp. 1-16, 2012.

5. Y. Liang and J. A. Chambers, 'An audio-video based IVA algorithm for source separation and evaluation on the AV16.3 Corpus', LVA/ICA 2012, Tel-Aviv, Israel, pp. 330-337, 2012.

In Chapter 5, a particular generalized Gaussian distribution is adopted as the source prior, which can exploit the fourth order information between different frequency bins to provide a more informative and stronger dependency structure. Therefore, IVA algorithms with this proposed source prior are shown to achieve an improved separation performance. The results are published in:

6. Y. Liang, S. M. Naqvi, G. Chen and J. A. Chambers, 'Independent vector analysis with a generalized multivariate Gaussian source prior for frequency domain blind source separation', resubmitted after review, Signal Processing.

7. Y. Liang, S. M. Naqvi and J. A. Chambers, 'Independent vector analysis with a multivariate generalized Gaussian source prior for frequency domain blind source separation', IEEE ICASSP 2013, Vancouver, Canada, pp. 6088-6092, 2013.

8. Y. Liang, J. Harris, G. Chen, S. M. Naqvi, C. Jutten and J. A. Chambers, 'Auxiliary function based IVA using a source prior exploiting fourth order relationships', EUSIPCO 2013, Marrakech, Morocco, 2013.

In Chapter 6, the dependency structure within the frequency domain speech signals is exploited, and described by using the t copula. The multivariate student's t distribution is constructed by using a t copula and adopted as the new source prior for IVA. The proposed source prior can better describe the nonlinear dependency structure within the frequency domain speech signals due to the tail dependency of the t copula. Therefore, IVA algorithms with the proposed source prior can improve the separation performance. The results are presented in:

9. Y. Liang, G. Chen, S. M. Naqvi and J. A. Chambers, 'Independent vector analysis with a multivariate student's t distribution source prior for speech separation', accepted for publication in Electronics Letters.

10. Y. Liang, G. Chen, S. M. Naqvi and J. A. Chambers, 'Copula based independent vector analysis with multivariate student's t source prior for frequency domain speech separation', submitted to IEEE Transactions on Audio, Speech and Language Processing.

Acknowledgements

I sincerely and deeply thank my supervisor Professor Jonathon Chambers for his generous support, consistent instruction and great advice throughout the past three years. He kindly offered me the amazing chance for pursuing the PhD degree in Loughborough University. I have benefited tremendously from his wide knowledge and great enthusiasm with students. Without his tireless instruction and encouragement, this thesis would have never been accomplished. It's my exclusive honor to be one of his research students.

I would like to give my great thanks to Dr. Syed Mohsen Naqvi, who helped me to enter my research field and kept giving me advice during my research. I would also like to sincerely thank Dr. Gaojie Chen, who not only helped me in research but also in life.

I would also like to give my appreciation to my friends and colleagues Muhammad Salman Khan, Ata Ur-Rehman, Dr. Miao Yu, Dr. Jie Tang, Lu Ge, Juncheng Hu, Ziming Zhu, Yu Wu and Zhao Tian for making my stay at Loughborough pleasant. I also wish to acknowledge many other friends from the Loughborough Chinese Students and Scholars Association, who helped me and support me to be the president of this society. I would also thank my friend Ying Wang to help me draw the model figure.

I wish to take this opportunity to thank Dr. Yonggang Zhang, who introduced me to Professor Jonathon Chambers. I also want to express my appreciation to my previous supervisor Professor Hao Dong and Professor Yanling Hao in Harbin Engineering University. They helped me to build a solid academic foundation and encouraged me to pursue the PhD degree.

Lastly, but most importantly, I wish to express my deepest gratitude and love to my loving parents. They are always supporting me and encouraging me to pursue knowledge. Moreover, I am extremely thankful to the China Scholarships Council for providing me financial support during the last three years.

List of Acronyms

ASS-IVA	Adaptive Step Size Independent Vector Analysis
AuxIVA	Auxiliary funtion based Independent Vector Analysis
AVIVA	Audio Video based Independent Vector Analysis
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
CBSS	Convolutive Blind Source Separation
cdf	cumulative distribution funtion
CPP	Cocktail Party Problem
DOA	Direction Of Arrival
DFT	Discrete Fourier Transform
ECG	Electrocardiography
EEG	Electroencephalography
EMG	Electromyography
FastIVA	Fast Fixed Point Independent Vector Analysis
FD-BSS	Frequency Domain-Blind Source Separation
GMM	Gaussian Mixture Model

HOS	Higher-Order Statistics
ICA	Independent Component Analysis
IVA	Independent Vector Analysis
IAuxIVA	Improved dependency model AuxIVA
MEG	Magnetoencephalography
MIMO	Multi-Input-Multi-Output
MOS	Mean Opinion Score
NG-IVA	Nature Gradient Independent Vector Analysis
PCA	Principal Component Analysis
pdf	probability density function
PI	Performance Index
PM	Permutation Measurement
RIVA	Robust Independent Vector Analysis
SDR	Signal-to-Distortion Ratio
SIR	Signal-to-Interference Ratio
SOS	Second-Order Statistics
SSS	Spherical Symetric Sparse source prior
STFT	Short Time Fourier Transform
SWIPE	Sawtooth Waveform Inspired Pitch Estimation

List of Symbols

Scalar variables are denoted by plain lower-case letters, (e.g., x), vectors by bold-face lower-case letters, (e.g., $\mathbf{x}$), matrices by upper-case letters, (e.g., X), and a group of matrices by bold-face upper-case letters (e.g.,$\mathbf{X}$). Some frequently used notations are as follows:

$|.|$ Absolute value

$||.||_2$ Euclidean norm

$||.||_p$ l_p norm

$(.)^T$ Transpose operator

$(.)^\dagger$ Hermitian transpose operator

$(.)^{-1}$ Inverse operator

$(.)^*$ Conjugate operator

$det(.)$ Matrix determinant operator

$diag(\mathbf{d})$ Diagonal matrix with vector $\boldsymbol{d}$ on its main diagonal

$E(.)$ Statistical expectation operator

I Identity matrix

k Frequency bin index

K Number of frequency bins

T STFT length

N Number of sources

M Number of mixtures

$\mathbf{s}$ Source signal vector

$\hat{\mathbf{s}}$ Estimated source signal vector

$\mathbf{x}$ Mixture signal vector

$\boldsymbol{\mu}$ Mean vector

Σ Covariance matrix

Λ Diagonal matrix

A Permutation adjustment matrix

G Overall system matrix

H Mixing matrix

W Estimated unmixing matrix

Q_w Whitening matrix

$c(.)$ Copula density function

$C(.)$ Copula

$F(.)$ Nonlinear function for FastIVA

$F(.)'$ First derivative of nonlinear function for FastIVA

$F(.)''$ Second derivative of nonlinear function for FastIVA

$g(.)$ Contrast function for AuxIVA

$J(.)$ Cost function

$Q(.)$ Auxiliary function for AuxIVA

List of Figures

1.1 Machine cocktail party problem: to build an intelligent machine which can duplicate some aspects of the human auditory system to solve the cocktail party problem through microphones and video measurements. 27

2.1 Diagram of instantaneous mixing with three sources and three measurements. 38

2.2 Diagram of convolutive mixing with three sources and three measurements. 38

2.3 The mixture model of independent vector analysis. Independent component analysis is extended to a formulation with multidimensional variables, where the mixing process is constrained to the sources on the same horizontal layer. 46

2.4 The two dimensional pdf for the independent Laplacian source prior. 49

2.5 The two dimensional pdf for the multivariate super-Gaussian source prior adopted by original IVA. 50

2.6 Separation performance of original FastICA Method: performance index at each frequency bin for the original FastICA method at the top and evaluation of permutation at the bottom. 55

2.7 Separation performance of Parra's Method: performance index at each frequency bin for Parra's method at the top and evaluation of permutation at the bottom. 56

2.8 Separation performance of IVA method: performance index at each frequency bin for IVA method at the top and evaluation of permutation at the bottom. 57

2.9 Separation performance of adaptive step size IVA method: performance index at each frequency bin for ASS-IVA method at the top and evaluation of permutation at the bottom. 58

2.10 Convergence comparison between IVA and ASS-IVA. The solid line is the convergence of IVA, and the asterisk line is the convergence of ASS-IVA. 59

3.1 Example of the block permutation problem of IVA. 72

3.2 Separation performance by using original IVA. 78

3.3 Phase difference by using orginal IVA. 78

3.4 Separation performance by using robust IVA. 79

3.5 Phase difference by using robust IVA. 79

3.6 Separation performance by using original IVA. 81

3.7 Phase difference by using orginal IVA. 81

3.8 Separation performance by using robust IVA. 82

3.9 Phase difference by using robust IVA. 82

3.10 Average SDR and SIR comparison. 83

3.11 The permutation measurement showing the block permutation problem by using AuxIVA. 88

3.12 The permutation measurement without the block permutation problem by using AuxIVA with the proposed dependency model. 89

4.1 Block diagram of the audio video based fast fixed-point independent vector analysis. Video localization is based on face and head detection. The visual location of each speaker is approximated after processing the 2-D image information and obtained from at least two synchronized colour video cameras through calibration parameters and an optimization method. The position of the microphone array and the output of the visual localizer are used to calculate the direction of arrival information of each speaker. Based on this information, a smart initialization is set for the FastIVA algorithm. 97

4.2 The pitch tracks of two mixture signals. 99

4.3 The pitch tracks of two separated signals. 100

4.4 Separation performance of FastIVA. The upper part is the performance index figure, and the bottom part is the permutation measurement figure. 102

4.5 Separation performance of AVIVA. The upper part is the performance index figure, and the bottom part is the permutation measurement figure. 103

4.6 One permutation measurement of the separation result for three sources when using FastIVA. 104

4.7 One permutation measurement of the separation result for three sources when using AVIVA. 105

4.8 Separation performance of FastIVA in the noisy environment. The upper part is the performance index figure, and the bottom part is the permutation measurement figure. 107

4.9 Separation performance of AVIVA in the noisy environment. The upper part is the performance index figure, and the bottom part is the permutation measurement figure. 108

4.10 SDR comparison in different reverberant environments. 109

4.11 SIR comparisons in different reverberant environments. 110

4.12 Room environment for one of the AV16.3 corpus recordings. A single video frame from camera 1. 111

4.13 Room environment for one of the AV16.3 corpus recordings. A single video frame from camera 2. 112

4.14 The pitch tracks of the mixed signals. 113

4.15 The pitch tracks of the separated signals by FastIVA. 113

4.16 The pitch tracks of the separated signals by AVIVA. 114

5.1 Second order inter-frequency relationships for the speech signal “si1010.wav”, x and y dimensions correspond to frequency bins 1 to 128 of 512. 121

5.2 Fourth order inter-frequency relationships for the speech signal “si1010.wav”, x and y dimensions correspond to frequency bins 1 to 128 of 512. 122

5.3 Plan view of the experiment setting in the room environment with two microphones and two sources 126

5.4 SDR comparison between original and proposed IVA algorithms as a function of reverberation time. 129

5.5 SIR comparison between original and proposed IVA algorithms as a function of reverberation time. 129

5.6 SDR comparison between original and proposed FastIVA algorithms as a function of reverberation time. 130

5.7 SIR comparison between original and proposed FastIVA algorithms as a function of reverberation time. 131

5.8 SDR comparison between original and proposed AuxIVA algorithms as a function of reverberation time. 132

5.9 SIR comparison between original and proposed AuxIVA algorithms as a function of reverberation time. 133

6.1 The copula density of a t copula with 4 degrees of freedom and correlation coefficient $\rho = 0.7$ 138

6.2 The copula density of a t copula with 4 degrees of freedom and correlation coefficient $\rho = -0.6$ 139

6.3 The copula density of a t copula with 4 degrees of freedom and correlation coefficient $\rho = 0$ 139

6.4 The scatter plot of two independent random variables. 141

6.5 The Chi-plot of two independent random variables. 142

6.6 The scatter plot of two random variables with a t copula, $\rho = -0.6$ and $v = 4$ 143

6.7 The Chi-plot of two random variables with a t copula, $\rho = -0.6$ and $v = 4$ 144

6.8 The scatter plot of two random variables with a t copula, $\rho = 0$ and $v = 2$ 145

6.9 The Chi-plot of two random variables with a t copula, $\rho = 0$ and $v = 2$ 146

6.10 The Chi-plot of two frequency bins of a speech signal "sa1.wav" from TIMIT dataset (a) 50th and 51th frequency bins (b) 50th and 55th frequency bins (c) 50th and 60th frequency bins (d) 50th and 100th frequency bins (e) 50th and 200th frequency bins (b) 50th and 500th frequency bins 147

6.11 The probability density function of a multivariate student's t distribution 148

6.12 The separation performance in different reverberant environment for mixtures 1 (a) SDR (b) SIR 152

6.13 The separation performance in different reverberant environment for mixtures 2 (a) SDR (b) SIR 153

6.14 The separation performance in different reverberant environment for mixtures 3 (a) SDR (b) SIR 154

6.15 The separation performance in different reverberant environment for mixtures 4 (a) SDR (b) SIR 155

6.16 The separation performance in different reverberant environment for mixtures 5 (a) SDR (b) SIR 156

6.17 The time-varying pitch tracks of the mixtures 158

6.18 The time-varying pitch tracks of the separated signals by using IVA algorithm with proposed source prior 159

List of Tables

2.1 SIR comparison 59

3.1 SDR and SIR comparison of the first experiment. 80

3.2 SDR and SIR comparison of the second experiment. 83

3.3 Separation performance comparison when there is no block permutation problem. 89

4.1 Separation performance comparison when block permutation problem happens 105

4.2 Separation performance comparison in noisy environment. 107

4.3 Separation performance for the real room recordings. 111

5.1 Separation performance comparison in SDR(dB) 127

5.2 Separation performance comparison in SIR(dB) 127

5.3 Separation performance comparison in terms of SDR and SIR measures in dB 130

5.4 Separation performance comparison in terms of SDR and SIR measures in dB 131

6.1 Separation performance comparison in SDR 150

6.2 Separation performance comparison in SIR 150

6.3 Separation rate comparison when using real room recordings 158

Chapter 1

INTRODUCTION

1.1 Cocktail Party Problem

"One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others. This is such a common experience that we may take it for granted; we may call it "the cocktail party problem". No machine has been constructed to do just this, to filter out one conversation from a number jumbled together" - Colin Cherry [1].

The cocktail party problem (CPP) was first proposed by Colin Cherry in 1953 [1], and further researched in [2]. The problem describes the situation that there are several people talking simultaneously in a room environment, and the target is to focus on one of them. For human beings, it is easy to focus with increased attention. However, for a machine, it is much more difficult to achieve this goal. The solution for the cocktail party problem is to design a method to focus on the desired speech while suppress or ignore all the other competing speech sounds.

During the past decades, much effort has been put on solving the cocktail party problem. The target is to design a machine which can imitate the auditory capability of humans. However, this target hasn't been realized because a complete understanding of the cocktail party phenomenon is still missing, and the human auditory perception capability is not fully understood. Actually, it is not necessary to duplicate the whole human auditory system to achieve this target. But it is still useful and helpful to better

understand the processing used by a human [3].

To address the cocktail party problem, three neural processes have been identified: analysis, recognition and synthesis.

The analysis process mainly involves segmentation or segregation, which means that it can segregate an incoming auditory signal to individual channels. The spatial location information is used by the listener to segregate the signals. If the sounds are coming from the same location, they are grouped together. While if they are originated from different directions, they are segregated.

The recognition process means analyzing the statistical characteristic contained in a sound stream, which is very useful in recognizing the sound patterns. The target of recognition is to establish the neurobiological mechanisms which are used by humans to identify a segregated sound from multiple streams.

The synthesis process indicates the reconstruction of individual sound waveforms from the separated sound streams. This process plays an important role in the human auditory system [4]. Thus the synthesis problem is highly related to the machine cocktail party problem.

From the description of the three processes, it is clear that the recognition doesn't need a perfect analysis process, meanwhile the synthesis doesn't need perfect analysis or recognition either. The synthesis process can be considered as the inverse process of the combination of the analysis and recognition process. It can deal with the received convolved mixtures and extract the desired speech. This process needs further research to be fully understood.

During the last two decades, the increase in computing power has motivated researchers to attempt to produce a real time solution such as [5]. Meanwhile, video information is also combined to help to solve the cocktail party problem as represented in Figure 1.1 [6].

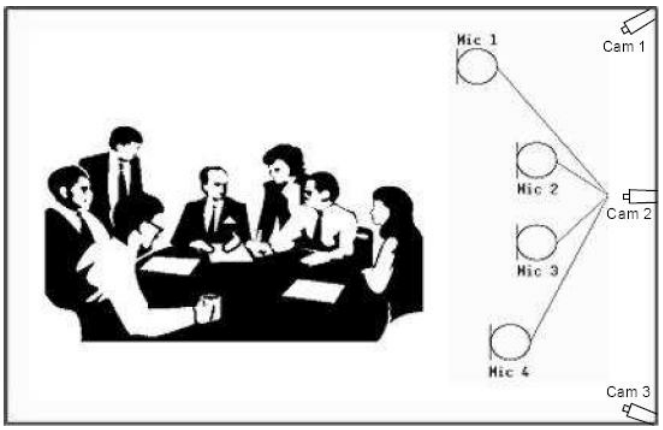


Figure 1.1. Machine cocktail party problem: to build an intelligent machine which can duplicate some aspects of the human auditory system to solve the cocktail party problem through microphones and video measurements.

1.2 Blind Source Separation

Attempts to solve the machine cocktail party problem have come from the signal processing community in the form of blind source separation (BSS) and generally from the computer science community in the form of computational auditory scene analysis (CASA) [7,8]. CASA is motivated by understanding human auditory scene analysis. Recently, the combination of BSS and CASA has also been proposed to improve separation performance [9,10]. The focus of this thesis is on signal processing based approaches such as blind source separation.

Blind source separation (BSS) has been proposed for various fields in recent years [11]. It is used to extract individual signals from observed mixed signals. It can be potentially used in communication systems, biomedical signal processing and image restoration. In the communication field, it is a promising tool for the design of multi-input multi-output (MIMO) equalizers for suppression of intersymbol interference, cochannel and adjacent channel interference and multi-access interference. In biomedical signal processing,

BSS can be used to process electrocardiography (ECG), electroencephalography (EEG), electromyography (EMG) and magnetoencephalograph (MEG) signals. In the image signal processing field, it can be used for image restoration and understanding [12] [13]. While this thesis concentrates upon its use in the cocktail party problem, namely solving the speech separation problem in a real room environment. It can also help to improve the performance of speech recognition by suppressing other competitive sounds and thereby enhance human-computer interface systems, such as the Siri system developed by the Apple company [14].

To address the BSS problem, many methods have been proposed. Herault and Jutten seem to have been the first who addressed the problem of blind source separation in 1985 [15]. The mixtures are assumed to be instantaneous in the standard BSS problem, which means that the sound is transmitted directly from the sources to the microphones without any delay. Comon established an instantaneous linear mixing model and clearly defined the term independent component analysis (ICA) in 1994 [16].

However, the instantaneous model is not suitable for solving the real environment cocktail party problem, because the instantaneous model is too simple to describe the complicated real room environment. For a real room environment, the acoustic sources take multiple paths to the microphone sensors instead of the direct path. Thus, the convolutive model is used to represent the practical situation. There are two types of mixing model which exist in the convolutive case, namely anechoic and echoic. The anechoic mixing model simply describes the transmission delays between the sources and sensors, while the echoic mixing model pays more attention to the delays and also the reverberations of the sources. This thesis is mainly concerned with the echoic mixture model due to its use for representing a real room environment and it also includes the anechoic model as a special case. In convolutive BSS, each element of the mixing filter is in fact a linear filter to

describe the multipaths from sources to sensors.

In a real room environment, the length of the room impulse response is typically on the order of thousands of samples. Thus, time domain methods are generally not suitable for the CBSS problem due to the computational complexity [17]. In order to reduce the computational cost, frequency domain methods have been proposed [18]. The convolution operation in the time domain becomes multiplication in the frequency domain, so the computational cost reduces significantly [17]. When the mixtures are transferred into the frequency domain by using the discrete Fourier transform (DFT), the ICA method can be used in each frequency bin to separate the mixtures.

Transformation into the frequency domain reduces the computational cost, but there are two indeterminacies which are inherent to ICA, namely the scaling and permutation ambiguities. The scaling ambiguities across frequencies can be managed by matrix normalization [12, 19–22]. On the other hand, the permutation ambiguity inherent to ICA is magnified due to the potential misalignment of the separated source at different frequency bins. In this case, should the separated results be transformed back to the time domain, the separation performance will be poor. Therefore, different methods to mitigate the permutation problem have been proposed [17]. In [18] the permutation problem is addressed by imposing a smoothness constraint on the mixing filter. The smoothing essentially forces the estimated sources in the frequency bins to align and is achieved by constraining the filter length in the time domain to be less than the block length of the DFT. Another kind of method exploits the special spectral structure contained in speech signals. Murata et al. used this kind of method to eliminate the cross-correlation of the reconstructed signals [23]. Localization information is added to help to constrain the signals in [24]. Sawada et al. [25] proposed a method to solve the permutation problem by integrating earlier approaches, direction of arrival and inter-frequency correlation of signal envelopes. Naqvi et al. [6] used

a multimodal approach to solve the permutation problem, which uses both audio and video information. Most of them needs pre or post processing by using extra information. For instance, source geometry is estimated first to help to solve the permutation problem in [24]. In [23] [26], the source structure is further exploited after separation stage to address the permutation problem. However, both pre and post processing generally introduce system delay and additional complexity.

Independent vector analysis (IVA) was proposed by Kim et al. around 2007 [27, 28]. It can theoretically avoid the permutation problem by retaining the inter-frequency dependency within each source vector while removing the dependency between different source vectors [5, 27]. Thus, IVA can mitigate the permutation problem during the convergence process without any requirement for additional information such as geometrical information. For ICA, the nonlinear score function is a univariate function. However, IVA adopts a multivariate score function to preserve the dependency between different frequency bins. Thus, IVA can use the data across all the frequency bins to separate the mixture in each individual frequency bin. There are three main types of IVA methods. The first one is the original NG-IVA, which adopts the Kullback-Leibler divergence between the joint probability density function and the product of marginal probability density functions of the individual source vectors as the cost function. The natural gradient method is used to minimize the cost function [27]. The second type is the fast fixed-point IVA (FastIVA), which adopts Newton's method to optimize the cost function to achieve a fast convergence in terms of the iteration numbers [29]. The third one is auxiliary function based IVA (AuxIVA), which is also a fast form of IVA. AuxIVA can converge quickly without introducing tuning parameters and can guarantee that the objective function decreases monotonically by using the auxiliary function technique [30]. Some other IVA methods based on these frameworks are proposed to exploit the source

activity and dynamic structure to achieve better separation performance such as [31] [32].

The core idea of IVA is to preserve the dependency within each source vector, and the nature of the multivariate nonlinear score function plays an important role in this process [27]. The multivariate nonlinear score function is derived from the multivariate source prior, therefore an appropriate multivariate source prior is needed to improve the separation performance. Original IVA algorithms adopt a multivariate Laplacian distribution as the source prior, which is a spherically symmetric distribution and implies the dependency between different frequency bins is all the same. However, the dependencies between frequency bins could be variable. In order to describe the dependency structure better, a chain-like overlapped source prior has been adopted [33]. Similarly, a harmonic structure dependency model has been proposed recently [34]. The Gaussian mixture model (GMM) is also adopted as the source prior, which can model different kinds of signals and make IVA more applicable for different signals [35] [36]. All the above source priors assume the covariance matrix is a diagonal matrix due to the orthogonal Fourier basis, which implies that there is no correlation between different frequency bins. More recently, in the context of time domain signals, the correlations across datasets are introduced to improve the separation performance. An IVA algorithm based upon a multivariate Gaussian source prior has been proposed to introduce second order correlations in the time domain, which is suitable for an application with large second order correlations such as in functional magnetic resonance imaging studies [37]. For the frequency domain IVA algorithms, other correlation information should be exploited.

1.3 Aim and Objectives

For original IVA algorithm, there are several weaknesses. First of all, the step size is fixed which constrains the convergence speed. Secondly, IVA is not stable due to the block permutation problem. Thirdly, the dependency model could be improved to achieve a better separation performance. The overall aim of the study is to overcome these weaknesses and enhance the separation performance of IVA. The particular objectives are:-

- Objective 1: improve the convergence speed of natural gradient IVA

In Chapter 2, the adaptive step size technique is applied to propose an adaptive step size IVA, which can select the optimal step size automatically to achieve a faster convergence in terms of iteration number compared with the original natural gradient IVA.

- Objective 2: improve the robustness of IVA in terms of the block permutation problem

In Chapter 3, the block permutation problem of IVA is highlighted, then the phase continuity of the unmixing matrix is exploited to propose a robust IVA algorithm. Moreover, the overlapped chain type source prior is applied to auxiliary function based IVA to obtain a robust separation performance.

- Objective 3: improve the separation performance by designing a novel source prior motivated by the nonlinear coupling in frequency domain speech signals

The source prior is important for IVA, because the nonlinear function derived from the source prior is used to preserve the dependency within each source vector. In Chapter 4, a particular multivariate generalized Gaussian distribution is adopted as the source prior, and the resultant nonlinear score function contains fourth order cross items to exploit the relationships

between different frequency bins. In Chapter 5, a multivariate student's t distribution constructed by a t copula is proposed as the source prior, which can better model the nonlinear dependency structure within the frequency domain speech signals and thereby improve the separation performance.

In the early chapters, Sawada's dataset is used to test the proposed algorithms, they are from "http://www.kecl.ntt.co.jp/icl/signal/sawada". In the later chapters, the advantage of the proposed algorithms are further confirmed by introducing the widely used TIMIT dataset [38] and real room recordings from the AV16.3 corpus [39].

1.4 Thesis Outline

The outline of this thesis is as follows:

Chapter 2 firstly provides the fundamentals of frequency domain blind source separation. There are two kinds of solution scheme. The first one is based on using second order statistics (SOS) due to the nonstationarity of the speech signals. The second kind of solution is based on exploiting the higher order statistics (HOS) of the signals. The typical second kind of solution is ICA. ICA is introduced followed by three other basic IVA algorithms, i.e. the natural gradient IVA, the fast fixed-point IVA and auxiliary function based IVA. In the section related to natural gradient IVA, the adaptive step size natural gradient IVA is also proposed and compared with the original natural gradient IVA.

Chapter 3 focuses on the stability of IVA algorithms. The specific stability problem of IVA, i.e. "the block permutation problem", is described, and the reason for it is also analyzed. Two robust IVA solutions are proposed in this chapter. The first solution exploits the continuity of the unmixing matrix and adjusts the misalignment to obtain a robust separation. The second solution adopts an overlapped chain like source prior to mitigate this

problem. The first solution is tested when using natural gradient IVA, and the second solution is tested when using the auxiliary function based IVA. Both of them can provide a robust separation performance.

In Chapter 4, the informed IVA is proposed, which combines prior information, i.e. the geometric information of the sources captured by video, with the FastIVA algorithm. One advantage of the informed IVA is using smart initialization to overcome the problems in convergence due to the nature of the cost function, such as the presence of local minima. It can also achieve faster convergence and improved separation performance. The informed IVA is tested by using real room recordings, and a pitch based objective evaluation method is also proposed to judge the separation performance when the information describing the mixing matrix and sources is missing.

In Chapter 5, a particular multivariate generalized Gaussian distribution is proposed to be the source prior for IVA. The nonlinear score function derived from this proposed source prior contains the fourth order relationships between different frequency bins, which can provide a more informative and stronger dependency structure, and thereby improve the separation performance. The proposed source prior can fit into all three IVA frameworks, and the experimental results confirm the advantage of this proposed source prior.

In Chapter 6, copula theory is introduced to model the dependency structure of the frequency domain speech signals. Then a multivariate student's t distribution is constructed by using a t copula with the univariate student's t marginal distribution. The proposed source prior can properly describe the nonlinear dependency structure within frequency domain speech signals. The natural gradient IVA algorithm with the multivariate student's t distribution is proposed and tested not only by different simulated reverberant room environments but also by real room recordings. The separation performance can be improved by using the IVA algorithm with the proposed

source prior.

Finally, conclusions are drawn, and future work is then discussed in Chapter 7.

Chapter 2

FUNDAMENTALS OF INDEPENDENT VECTOR ANALYSIS

2.1 Introduction

The speech separation problem in a real room environment is a convolutive blind source separation (CBSS) problem, which is often addressed in the frequency domain. Two kinds of approaches to solve this problem will be reviewed, namely the second order statistic (SOS) method and higher order statistic (HOS) method. For the second order statistic methods, the statistical non-stationarity of the speech signals is exploited to separate the mixed speech signals. On the other hand, for the higher order statistic methods, the non-Gaussianity of the speech signals is used to address this problem. Independent component analysis is the typical higher order statistic method, which will also be reviewed in this chapter. Independent vector analysis is an extension of independent component analysis to avoid theoretically the permutation problem inherent to ICA by exploiting the dependency within each source vector. There are three main types of IVA algorithms which will be reviewed in this chapter. The first one is the natural gradient IVA, which adopts the natural gradient method to minimize the objective func-

tion. An adaptive step size natural gradient IVA algorithm is proposed to satisfy the first objective of this thesis, i.e. improving the convergence speed in terms of iteration numbers compared with natural gradient IVA. The second one is fast fixed-point IVA which adopts Newton's method to optimize the objective function. The last one is the auxiliary function based IVA, which uses the auxiliary function technique to achieve a fast form of the IVA method in terms of the convergence iterations. Next, convolutive blind source separation is introduced.

2.2 Convolutive Blind Source Separation

The basic blind source separation model assumes the mixing matrix is an instantaneous case, which means the signals are mixed instantaneously, i.e the microphones pick up only scaled mixtures of the original sources. Figure 2.1 shows an example of instantaneous mixing. For the case of three sources and three microphones:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \times \begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} \tag{2.2.1}$$

where x_1, x_2 and x_3 are the three measured mixtures captured by the microphones; s_1, s_2 and s_3 are the three sources, any time dependency is omitted in this equation. The elements h_{ij} of the mixing matrix are scalar values representing a change in amplitude only.

Practically, perfectly instantaneous mixtures of sounds are seldom encountered. For the practical cocktail party problem in a real room environment, the observed signals received by the microphones are convolutive mixtures of source signals because of the reverberant environment. Therefore, it becomes the convolutive blind source separation problem. Figure 2.2 shows an example of convolutive mixing.

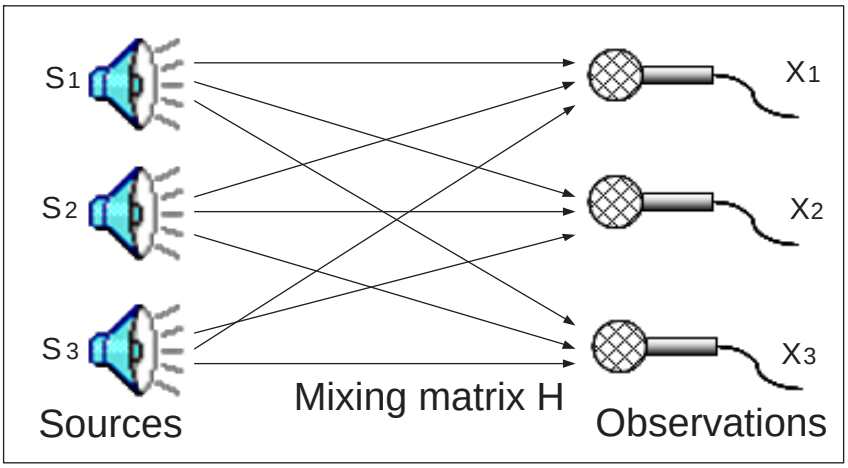


Figure 2.1. Diagram of instantaneous mixing with three sources and three measurements.

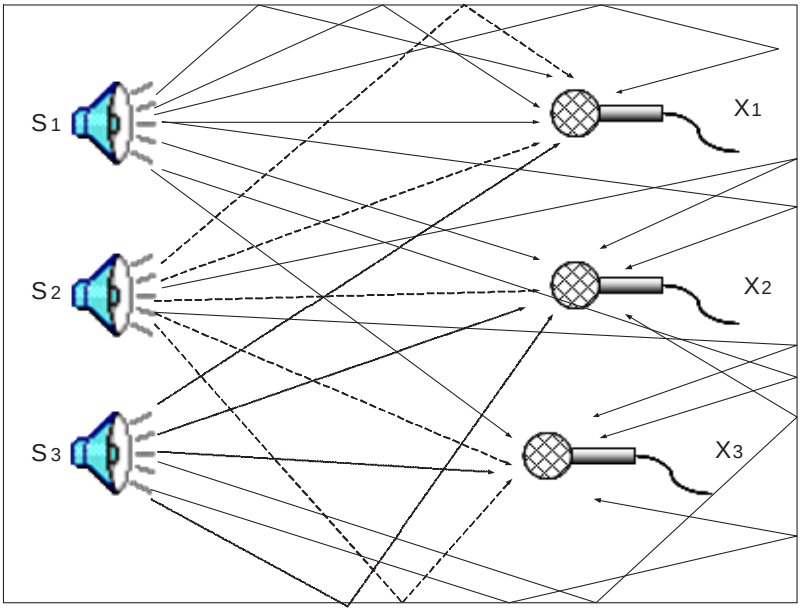


Figure 2.2. Diagram of convolutive mixing with three sources and three measurements.

For a CBSS problem, the noise free relationship between the sources and observations in the time domain is defined as:

$$x_j(t) = \sum_{i=1}^{N} \sum_{\tau=0}^{l-1} h_{ij}(\tau) s_i(t-\tau) \tag{2.2.2}$$

where $s_i(t)$ is the i-th source of all N sources, and $x_j(t)$ denotes the j-th mixture of all M mixtures; $h_{ij}(t)$ is the room impulse response between them

which has l length in time, and t denotes the time sample index.

2.3 Second Order Statistic Solution to CBSS

During the past decades, there has been considerable research performed in the field of CBSS [17]. Initially, research was aimed at solutions based in the time domain. In a real room environment, however, where the impulse response is on the order of thousands of samples in length, the time domain algorithm would be computationally very expensive to separate the sources. To overcome this problem, a solution in the frequency domain was proposed by Parra and Spence [18]. As convolution in the time domain corresponds to multiplication in the frequency domain provided the block length of the transform is substantially larger than the length of the time domain filter, the transformation into the frequency domain converts the convolutive mixing problem to that of independent complex instantaneous mixing operations at each frequency bin. Time domain signals $x_i(t)$ are converted into the frequency domain time series signals $x_i(k,m)$ by a T-point window discrete Fourier transform, where k denotes the frequency index and m denotes the time block index. Thus the frequency domain blind source separation (FD-BSS) method reduces the computational cost greatly.

The noise free model for frequency domain convolutive blind source separation problem can be described as:

$$\mathbf{x}(k,m) = H(k)\mathbf{s}(k,m) \tag{2.3.1}$$

where $\mathbf{s}(k,m) = [s_1(k,m),\ldots,s_N(k,m)]^T$ is the source vector for the k-th frequency bin and $\mathbf{x}(k,m) = [x_1(k,m),\ldots,x_M(k,m)]^T$ is the mixture vector for the k-th frequency bin, and $[\cdot]^T$ denotes the transpose operator; $H(k)$ is the mixing matrix in the k-th frequency bin. The target is to find the unmixing matrix $W(k)$ and unmixed signals $\hat{\mathbf{s}}(k,m) = [\hat{s}_1(k,m),\ldots,\hat{s}_N(k,m)]^T$ at

each frequency bin, which are calculated as

$$\hat{\mathbf{s}}(k,m) = W(k)\mathbf{x}(k,m) \tag{2.3.2}$$

Parra and Spence proposed a method which utilizes second order statistic by exploiting the non-stationarity of speech in 2000 [18]. This method works in the frequency domain and adopts a joint signalization approach. Then least squares optimization is used to estimate the unmixing matrix as well as the signal power. The gradient descent algorithm is used to jointly diagonalize the unmixing matrix $W(k)$ for all the frequency bins by minimizing the sum-squared error (as the sum of off diagonal elements of the covariance matrix of the estimated sources). The unmixing matrix $W(k)$ is found across all the frequency bins from

$$\begin{aligned} R_{\hat{s}}(k,t_d) &= W(k)R_x(k,t_d)W^{\dagger}(k) \\ &= W(k)H(k)\Lambda_s(k,t_d)H^{\dagger}(k)W^{\dagger}(k) \end{aligned} \tag{2.3.3}$$

where $[\cdot]^{\dagger}$ denotes the Hermitian transpose. $\Lambda_s(k,t_d)$ is a diagonal covariance matrix describing the source signals and is assumed to be a distinct diagonal matrix for each time block t_d[1] , and $R_x(k,t_d)$ is the covariance matrix of $x(k,t_d)$. The covariance matrices are estimated using an averaged cross-power spectrum

$$\hat{R}_x(k,t_d) = \frac{1}{L}\sum_{n=1}^{L-1}\mathbf{x}(w,t_d+nT)\mathbf{x}^{\dagger}(k,t_d+nT) \tag{2.3.4}$$

The cost function J_m based on the off-diagonal elements of $R_{\hat{\mathbf{s}}}(k,t_d)$ estimated at $t_d = dTL$, $d = 1,2,\ldots,D$, with D being the number of matrices to diagonalize, is

[1]The index m is not used in order that the calculation of the time block index can be described.

$$J_m = \sum_{k=1}^{T} \sum_{d=1}^{D} \|E(k, t_d)]\|_F^2 \tag{2.3.5}$$

where $E(k, t_d) = W(k)\hat{R}_x(k, t_d)W^\dagger(k) - \Lambda_s(k, t_d)$, and $\|.\|_F^2$ is the squared Frobenius norm. To avoid the trivial $W(k) = \mathbf{0}\ \forall k$ solutions, the constant $diag(W(k)) = I\ \forall k$ is applied, where $diag(\cdot)$ denotes taking the diagonal elements to construct a diagonal matrix. To minimize (2.3.5) the method of steepest descent is applied to yield

$$\frac{\partial J_m}{\partial W^*(k)} = 2 \sum_{d=1}^{D} E(k, t_d)W(k)\hat{R}_x(k, t_d) \tag{2.3.6}$$

where $(\cdot)^*$ denotes the conjugate operators, and the update equation for $W(k)$ becomes

$$W_{i+1}(k) = W_i(k) - \eta \sum_{d=1}^{D} E(k, t_d)W_i(k)R_x(k, t_d) \tag{2.3.7}$$

where i and η are the iteration index and learning rate respectively. The unmixing filter matrix $W(k)$ is updated for all the frequency bins. The source covariance matrix can be estimated at each iteration by $\hat{\Lambda}_s(k, t_d) = diag(W(k)R_x(k, t_d)W^\dagger(k))$. A higher order statistic approach is next considered.

2.4 Independent Component Analysis

"ICA (independent component analysis) is a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements, or signals. ICA defines a generative model for the observed multivariate data, which is typically given as a large database of samples. In the model, the data variables are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown. The latent variables are assumed non-Gaussian and mutually independent, and

they are called the independent components of the observed data. These independent components, also called sources or factors, can be found by ICA." pp. xvii, [19]. ICA is superficially related to principal component analysis and factor analysis. ICA is a much more powerful technique, however, capable of finding the underlying factors or sources when these classical methods fail completely. In order to make ICA work, some assumptions are necessary [40].

- The source components are assumed to be statistically independent of each other.

According to the mathematical definition of independence, variables are independent if and only if the joint probability density function (pdf) is factorizable in the following way:

$$p(s_1, \cdots, s_N) = \prod_{i=1}^{N} p(s_i)$$

- At most one source has *Gaussian* distribution.

ICA works by exploiting the higher order statistic of the signals. However, the higher order cumulants of a Gaussian distribution are zero. Thus, ICA is essentially impossible if all the sources are Gaussian signals.

- The unknown mixing matrix is assumed to be invertible.

In other words, it assumes the number of sources is equal to or smaller than the number of mixtures, i.e. an exactly determined or over-determined problem. However, this assumption can be relaxed if other information such as the time frequency representation of the sources is exploited [41].

There are several kinds of independent component analysis methods. The first one is ICA by maximization of non-Gaussianity. The measurement of non-Gaussianity can be the kurtosis or the negentropy which is introduced

from information theory [19]. The second one is ICA by maximum likelihood estimation. The last one is ICA by minimization of mutual information [19]. The motivation of this approach is that it may not be very realistic in many cases to assume that the data follows the ICA model. Therefore, an approach that does not assume anything about the data is needed. A general-purpose measure of the dependence of the components of a random vector is used, then ICA can be defined as a linear decomposition that minimizes that dependence measure. Such an approach can be developed using mutual information, which is a well-motivated information-theoretic measure of statistical dependence.

However, ICA does have two ambiguities. The first of which is called the scaling ambiguity, namely the variances (energy) of the independent components are not necessarily matched to the original sources. The reason is that, both $\mathbf{s}$ and H being unknown, any scalar multiplier applied to one of the sources s_i could always be canceled by dividing the corresponding column $\mathbf{h}_i$ of H by the same non zero scalar:

$$\mathbf{x} = \sum_i (\frac{1}{\alpha}\mathbf{h}_i)(\alpha s_i)$$

In order to address this problem, the most usual way is to standardize the independent components to have unit variance.

The second ambiguity is the permutation ambiguity, which means the order of the independent components can not be determined. The reason is again as both $\mathbf{s}$ and H are unknown, the order of the terms can be exchanged without losing the restored independence. When ICA based methods are used to address the CBSS problem in the frequency domain to reduce the computational load of the time domain solution, the permutation ambiguity is magnified because the alignment is different for each individual frequency bin. Many solutions have been proposed to solve the permutation problem

as discussed in Chapter 1.

The statistical independence implies the uncorrelated sources, but the reverse is not necessarily true. Most ICA algorithms decorrelate the mixtures via spatial whitening, before optimizing their separating objective contrast or cost functions. This spatial whitening is achieved by employing principal component analysis (PCA).

In the context of BSS, principal component analysis seeks to remove the cross-correlation between the observed signals, and ensures that they have unit variance [19]. PCA operates by finding the projections of the mixture data on orthogonal directions of maximum variance. A zero mean vector $\mathbf{z}$ containing observations from spatially distinct locations is said to be spatially white if

$$E[\mathbf{z}\mathbf{z}^T] = I \tag{2.4.1}$$

where $E[\cdot]$ is the statistical expectation operator, and I is the identity matrix. The unmixing matrix W can be decomposed into two components as:

$$W = U_w Q_w \tag{2.4.2}$$

where Q_w denotes the whitening matrix and U_w is the rotation matrix [19]. The whitening matrix Q_w can be formulated as:

$$Q_w = D_x^{-\frac{1}{2}} E_x^T \tag{2.4.3}$$

where E_x is the matrix whose columns are the unit-norm eigenvectors of the spatial covariance matrix $C_x = E_x D_x E_x^T$ and D_x is the diagonal matrix of the eigenvalues of C_x. The matrix $D_x^{-\frac{1}{2}}$ plays an important role in $E[\mathbf{z}\mathbf{z}^T] = I$ and it is also important to note that the whitening matrix Q_w is not unique because it can be pre-multiplied by an orthogonal matrix to obtain another version of Q_w. In order to overcome the permutation problem

algorithmically, a new algorithm is considered.

2.5 Independent Vector Analysis

For traditional frequency domain methods, ICA is applied to separate the mixture at each frequency bin individually. However, the alignment of the separated signals is not consistent across all the frequency bins. Independent vector analysis was proposed as a frequency domain blind source separation method around 2007 [27, 28]. It is designed theoretically to avoid the permutation problem inherent to ICA by retaining the dependency within each individual source vector, while removing the dependency between different source vectors.

Independent vector analysis is based on the ICA method with some modifications. It exploits a dependency model capturing interfrequency dependencies, which is represented diagrammatically in Fig 2.3 for the case of two sources and two measurements. In this diagram, each horizontal slice corresponds to a single discrete frequency and the vertical shaded regions represent the interfrequency dependencies between the sources, whilst the horizontal shaded regions are the intra-frequency dependencies introduced by the mixing process.

Compared with ICA methods, the interfrequency dependencies depend on a modified model for the source signal prior. In the conventional ICA based algorithms, the source signal prior is defined independently at each frequency, while the IVA method uses higher order dependencies across frequencies. The IVA method defines each source prior as a multivariate super-Gaussian distribution. Thus, it can be used to preserve the higher order interfrequency dependencies and structures of frequency components. Moreover, the permutation problem can be avoided and leads to an improved separation performance [27].

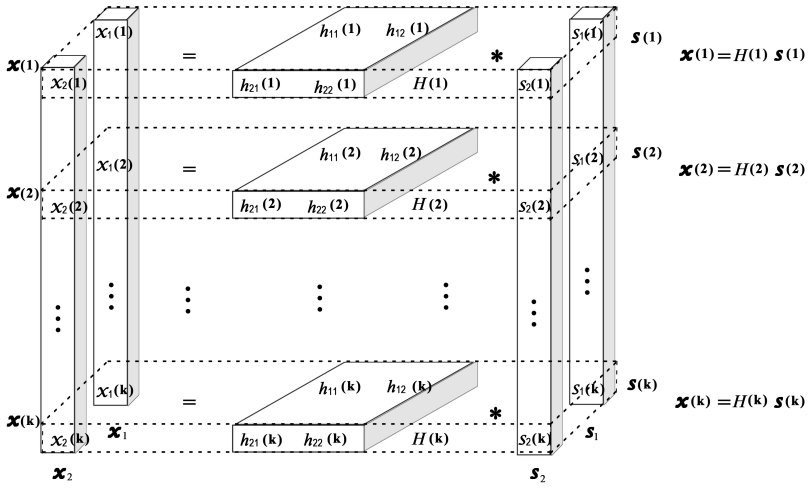


Figure 2.3. The mixture model of independent vector analysis. Independent component analysis is extended to a formulation with multi-dimensional variables, where the mixing process is constrained to the sources on the same horizontal layer.

In order to solve the CBSS problem by IVA, the short time Fourier transform (STFT) is applied to transfer the problem into the frequency domain to avoid the heavy computational load of time domain operation. The basic noise free frequency domain BSS model has been described in equation (2.3.1), and the time index is omitted here for convenience:

$$\mathbf{x}(k) = H(k)\mathbf{s}(k) \tag{2.5.1}$$

The index $k = 1, 2, \ldots, K$ denotes the k-th frequency bin, and K is the number of frequency bins. $H(k)$ is the mixing matrix at k-th frequency bin, whose dimension is $M \times N$.

In order to solve this problem and recover the source signals from the mixtures, an unmixng matrix must be determined to obtain the estimated sources

$$\hat{\mathbf{s}}(k) = W(k)\mathbf{x}(k) \tag{2.5.2}$$

where $\hat{\mathbf{s}}(k) = [\hat{s}_1(k), \hat{s}_2(k), \ldots, \hat{s}_n(k)]^T$ is the estimated signal vector in the frequency domain, and $W(k)$ is the unmixing matrix at the k-th frequency bin, whose dimension is $N \times M$. In this thesis, most of the time it is assumed that the number of sources is the same as the number of microphones, i.e. $M = N$.

The cost function for IVA is the Kullback-Leibler divergence between the joint probability density function $p(\hat{\mathbf{s}}_1 \ldots \hat{\mathbf{s}}_N)$ and the product of marginal probability density functions of the individual source vectors $\prod p(\hat{\mathbf{s}}_i)$ [27].

$$\begin{aligned} J &= \mathcal{KL}\big(p(\hat{\mathbf{s}}_1 \ldots \hat{\mathbf{s}}_N) || \prod p(\hat{\mathbf{s}}_i)\big) \\ &= \int p(\hat{\mathbf{s}}_1 \cdots \hat{\mathbf{s}}_N) \log \frac{p(\hat{\mathbf{s}}_1 \cdots \hat{\mathbf{s}}_N)}{\prod p(\hat{\mathbf{s}}_i)} d\hat{\mathbf{s}}_1 \cdots d\hat{\mathbf{s}}_N \\ &= const - \sum_{k=1}^{K} \log |det(W(k))| - \sum_{i=1}^{N} E[\log p(\hat{\mathbf{s}}_i)] \end{aligned} \tag{2.5.3}$$

where $det(\cdot)$ is the matrix determinant operator and $|\cdot|$ denotes the absolute value. The source prior $p(\hat{\mathbf{s}}_i)$ is different from traditional ICA methods because it is a vector across all frequency bins instead of the product of source prior of each frequency bin $\prod_{k=1}^{K} p(s_i(k))$. Therefore, when the cost function is minimized, the dependency between different source vectors would be removed but the dependency between the components within each vector is preserved.

2.5.1 Natural Gradient IVA

The cost function for this optimization problem has been defined in (2.5.3), from which the natural gradient IVA is derived straightforwardly by applying the natural gradient method to minimize the cost function and update the unmixing matrix as:

$$W(k)_{new} = W(k)_{old} + \eta \Delta W(k) \tag{2.5.4}$$

where η is the step size, and ΔW is derived from the cost function, and then multiplied by $W(k)^{\dagger}W(k)$ to use the natural gradient [27].

$$\Delta W(k) = \big(I - E[\varphi^{(k)}(\hat{\boldsymbol{s}})\hat{\boldsymbol{s}}^{*}(k)]\big)W(k) \tag{2.5.5}$$

The term $\varphi^{(k)}(\hat{\boldsymbol{s}})$ is the nonlinear score function vector

$$\varphi^{(k)}(\hat{\boldsymbol{s}}) = [\varphi^{(k)}(\hat{\boldsymbol{s}}_{\mathbf{1}}), \cdots, \varphi^{(k)}(\hat{\boldsymbol{s}}_{\boldsymbol{N}})]^{T} \tag{2.5.6}$$

and

$$\varphi^{(k)}(\hat{\boldsymbol{s}}_{\boldsymbol{i}}) = -\frac{\partial \log p(\hat{\boldsymbol{s}}_{\boldsymbol{i}})}{\partial \hat{s}_i(k)} \tag{2.5.7}$$

which is a multivariate function and is used to retain dependency across the frequency bins. Because it is derived from the source prior, it is important to establish an appropriate multivariate source prior to preserve the dependency structure and achieve a good separation performance.

In traditional BSS approaches, the univariate Laplacian distribution is often adopted as the source prior. Suppose that the source prior of a vector is an independent Laplacian source prior in each frequency, thus, the source prior vector can be written as

$$p(\mathbf{s}_i) = \prod_{k=1}^{K} p(s_i(k)) \propto \prod_{k=1}^{K} \exp\Big(- \frac{|s_i(k) - \mu_i(k)|}{\sigma_i(k)}\Big) \tag{2.5.8}$$

where $\mu_i(k)$ and $\sigma_i(k)$ are the mean and standard deviation of the i-th signal at the k-th frequency bin respectively. The two dimensional pdf for this source prior is shown in Fig 2.4.

Assuming zero mean and unit variance, the nonlinear score function can

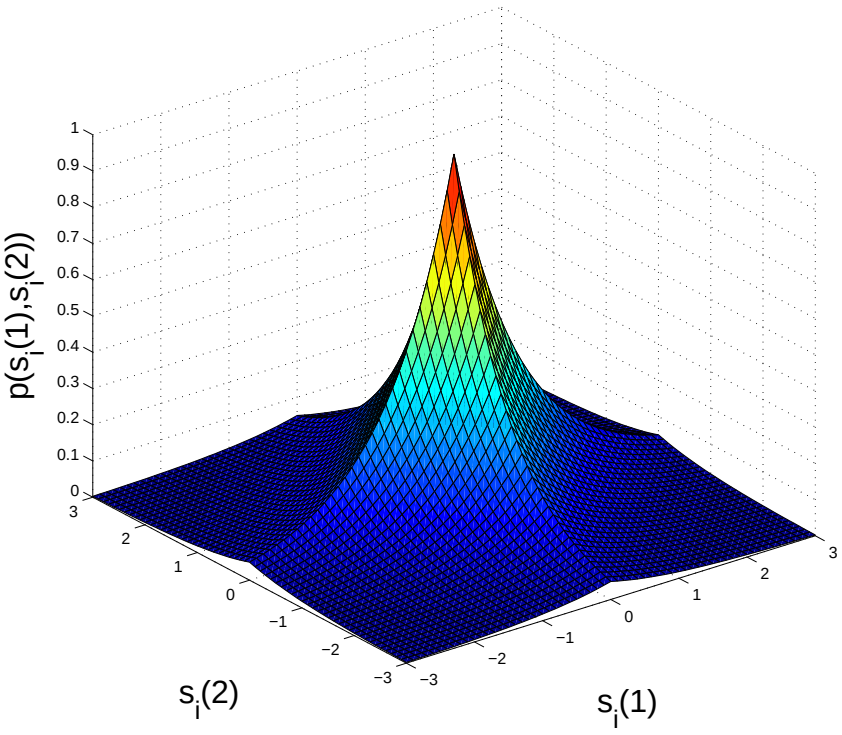


Figure 2.4. The two dimensional pdf for the independent Laplacian source prior.

be derived according to equation (2.5.7)

$$\varphi^{(k)}(\hat{s}_i(1)\cdots\hat{s}_i(K)) = \frac{\hat{s}_i(k)}{|\hat{s}_i(k)|} \tag{2.5.9}$$

which is a univariate function, because it only contains a single variable $\hat{s}_i(k)$, which can not keep the dependency within the source vector. Thus a dependent source prior is needed, and the elements of which are modelled as dependent with each other.

For the original IVA algorithm, a dependent multivariate super-Gaussian distribution is adopted as the dependent source prior, which takes the form

$$p(\mathbf{s}_i) \propto exp\Big(- \sqrt{(\mathbf{s}_i - \boldsymbol{\mu}_i)^{\dagger}\Sigma_i^{-1}(\mathbf{s}_i - \boldsymbol{\mu}_i)}\Big) \tag{2.5.10}$$

where $\boldsymbol{\mu}_i$ and Σ_i^{-1} are respectively the mean vector and inverse covariance

matrix of the i-th source. The two dimensional pdf for this source prior is shown in Fig 2.5. The product of the marginal probability density functions is not equal to the joint probability density function [27], which indicates the elements in the source vector are dependent with each other.

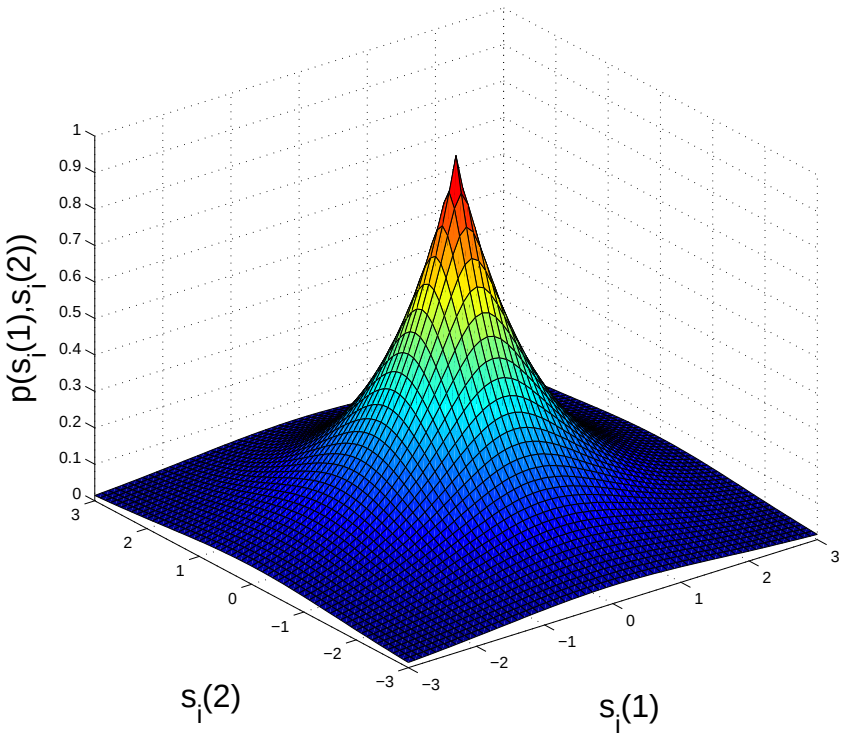


Figure 2.5. The two dimensional pdf for the multivariate super-Gaussian source prior adopted by original IVA.

This distribution can be related to a multivariate Gaussian with a fixed mean and a variable variance

$$\mathbf{s}_i = \gamma^{1/2}\boldsymbol{\xi}_i + \boldsymbol{\mu}_i \tag{2.5.11}$$

where γ is a scalar random variable, and $\boldsymbol{\xi}_i$ is a K-dimensional random variable which has Gaussian distribution with zero mean and covariance matrix Σ_i

$$p(\boldsymbol{\xi}_i) \propto \exp\big(-\frac{\boldsymbol{\xi}_i^{\dagger}\Sigma_i^{-1}\boldsymbol{\xi}_i}{2}\big) \tag{2.5.12}$$

Suppose that γ obeys a Gamma distribution as follows:

$$p(\gamma) \propto \gamma^{1/2}\exp\big(-\frac{\gamma}{2}\big) \tag{2.5.13}$$

Thus the original source prior can be achieved by integrating joint distribution of $\mathbf{s}_i$ and γ over γ as follows:

$$\begin{aligned} p(\mathbf{s}_i) &= \int_0^{\infty} p(\mathbf{s}_i|\gamma)p(\gamma)d\gamma \\ &= \alpha_1 \int_0^{\infty} \gamma^{1/2}\exp\big(-\frac{1}{2}\big(\frac{(\mathbf{s}_i-\boldsymbol{\mu}_i)^{\dagger}\Sigma_i^{-1}(\mathbf{s}_i-\boldsymbol{\mu}_i)}{\gamma}+\gamma\big)\big)d\gamma \\ &= \alpha_2\exp\big(-\sqrt{(\mathbf{s}_i-\boldsymbol{\mu}_i)^{\dagger}\Sigma_i^{-1}(\mathbf{s}_i-\boldsymbol{\mu}_i)}\big) \end{aligned} \tag{2.5.14}$$

where α_1 and α_2 are normalization terms. This indicates that there is variance dependency generated by γ.

The nonlinear score function can be derived according to the source prior. It is assumed that the mean vector is a zero vector and the covariance matrix is a diagonal matrix due to the orthogonality of the Fourier bases, which implies that each frequency bin sample is uncorrelated with the others. As such, the nonlinear score function to extract the i-th source at the k-th frequency bin can be obtained as:

$$\begin{aligned} \varphi^{(k)}(\hat{s}_i(1)\dots\hat{s}_i(K)) &= -\frac{\partial\log\big(p(\hat{s}_i(1)\dots\hat{s}_i(K)\big)}{\partial\hat{s}_i(k)} \\ &= \frac{\partial\sqrt{\sum_{k'=1}^{K}\left|\frac{\hat{s}_i(k')}{\sigma_i(k')}\right|^2}}{\partial\hat{s}_i(k)} = \frac{\hat{s}_i(k)}{\sigma_i(k)\sqrt{\sum_{k'=1}^{K}\left|\frac{\hat{s}_i(k')}{\sigma_i(k')}\right|^2}} \end{aligned} \tag{2.5.15}$$

This is a multivariate function, and the dependency between the frequency bins is thereby accounted for in learning.

As for the scaling problem, original IVA uses the minimal distortion

principle method to adjust the learned unmixing matrix [42], because natural signals are generally dynamic and nonstationary and the variance isn't known. Having designed the learning algorithm, the unmixing matrix is an arbitrary version of the exact one, which is given by

$$W(k) = A(k)H^{-1}(k) \tag{2.5.16}$$

where $A(k)$ is an arbitrary diagonal matrix. Therefore, the unmixing matrix can be updated as

$$W(k) \leftarrow diag\big(W^{-1}(k)\big)W(k). \tag{2.5.17}$$

Improving the convergence of the algorithm is next considered.

2.5.2 Adaptive Step Size Natural Gradient IVA

It is highlighted that the original IVA method above uses a fixed step to update the separating matrix. However, a fixed step update has its own shortcomings, such as relative slow convergence speed, poor tracking ability and relatively poor separation performance. Thus, an adaptive step size algorithm for the IVA method is derived here. The frequency index k is omitted for convenience. The update rule for each frequency is as follows:

$$W(t'+1) = W(t') + \eta(t')\Delta W(t') \tag{2.5.18}$$

where t' is the iteration index. The step size variation should be correlated with the change in the estimated cost function. When the change is large, which means the algorithm is in the early stage of learning, and high convergence speed is needed, the step size should be relatively large. When the change is small, which means the algorithm is approaching steady state, and accuracy should be considered more, so the step size should be reduced.

Based on the discussion above, the update for the step size [43] is:

$$\eta(t') = \eta(t'-1) - \theta \nabla_\eta J(t')|_{\eta=\eta(t'-1)} \tag{2.5.19}$$

where θ is a small positive constant, and $J(t)$ is the instantaneous estimate of the cost function from which IVA is derived. To proceed, an inner product of matrices is defined in [44] as:

$$\langle D_1, D_2 \rangle = tr(D_1^T D_2) \tag{2.5.20}$$

where $\langle \cdot \rangle$ denotes the inner product, $tr(\cdot)$ represents the trace operator. D_1 and D_2 are two matrices. Thus, the gradient term of the step size update is derived as follows:

$$\begin{aligned} \nabla_\eta J(t')|_{\eta=\eta(t'-1)} &= \Big\langle \frac{\partial J(t')}{\partial W(t')}, \frac{\partial W(t')}{\partial \eta(t'-1)} \Big\rangle \\ &= tr\Big(\frac{\partial J(t')}{\partial W(t')}^T \frac{\partial W(t')}{\partial \eta(t'-1)} \Big) \end{aligned} \tag{2.5.21}$$

where $\partial J(t')/\partial W(t') = -\Delta W(t')$ is simply the update of the separating matrix. Due to the separating matrix update equation (2.5.18),

$$\frac{\partial W(t')}{\partial \eta(t'-1)} = \Delta W(t'-1) \tag{2.5.22}$$

Thus, the gradient term of the step size update is obtained:

$$\nabla_\eta J(t')|_{\eta=\eta(t'-1)} = -tr(\Delta W(t')^T \Delta W(t'-1)) \tag{2.5.23}$$

Finally, the adaptive step size IVA algorithm is described as follows:

$$W(t'+1) = W(t') + \eta(t') \Delta W(t') \tag{2.5.24}$$

$$\eta(t') = \eta(t'-1) + \theta tr(\Delta W(t')^T \Delta W(t'-1)) \tag{2.5.25}$$

An illustrative experiment is shown to indicate that the IVA algorithm can mitigate the permutation problem, and the separation performance is compared with second order statistic method, i.e. Parra's method. Then the adaptive step size natural gradient IVA is used to separate the mixtures and compared with the original natural gradient IVA algorithm to show the advantage.

In the experiments, two real recorded speech signals are used as the data set, and a two-input two-output system model is established. The dimension of the simulated room is $5m \times 5m \times 5m$, and the reverberation time RT60 is 130ms. The sources are assumed to be positioned at [2, 3.1, 1.5] and [3.25, 3.25, 1.5], and the microphones at [2.48, 4.5, 1.5] and [2.52, 4.5, 1.5], relative to the reference of the room, which is the corner. The length of the short time Fourier transform $T = 1024$ samples. The sampling frequency is 8kHZ, thus the length of STFT is 128ms. The initial step size $\eta = 0.1$. The initial value of the separation matrix W is an identity matrix, and θ is chosen as 2×10^{-7}.

The separation performance is evaluated in terms of two aspects: the separation evaluation and the convergence performance. In the BSS field, two typical criteria for the separation performance are the signal-to-interference ratio (SIR) and performance index (PI).

The SIR criterion is commonly used in the signal processing field, which indicates the purity of a signal. The calculation of SIR used here is [21]:

$$SIR = 10 \log_{10} \frac{\sum_k \sum_i |H_{ii}(k)|^2 \langle |s_i(k)|^2 \rangle}{\sum_k \sum_{i \neq j} |H_{ij}(k)|^2 \langle |s_j(k)|^2 \rangle} \tag{2.5.26}$$

where H_{ii} and H_{ij} denote the diagonal and off-diagonal elements of the frequency domain mixing filter, and s_i is the frequency domain representation of the source of interest.

The PI criterion is widely used in the blind source separation field. It is

calculated at each frequency bin and is based on the overall system matrix $G = WH$, where matrix W is the obtained unmixing matrix (the k index is dropped here for convenience in presentation). The PI is described as follows [12]:

$$PI(G) = [\frac{1}{N}\sum_{i=1}^{N}(\sum_{j=1}^{M}\frac{|G_{ij}|}{max_j|G_{ij}|}-1)]+[\frac{1}{M}\sum_{j=1}^{M}(\sum_{i=1}^{N}\frac{|G_{ij}|}{max_i|G_{ij}|}-1)] \quad (2.5.27)$$

where G_{ij} is the ij-th element of matrix G. Although PI can show the separation performance in each frequency bin, it can not show the permutation directly. Thus, for a two-input two-output model, a criterion $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})]$ is used to measure the permutation [21], it is called permutation measurement (PM) in this thesis.

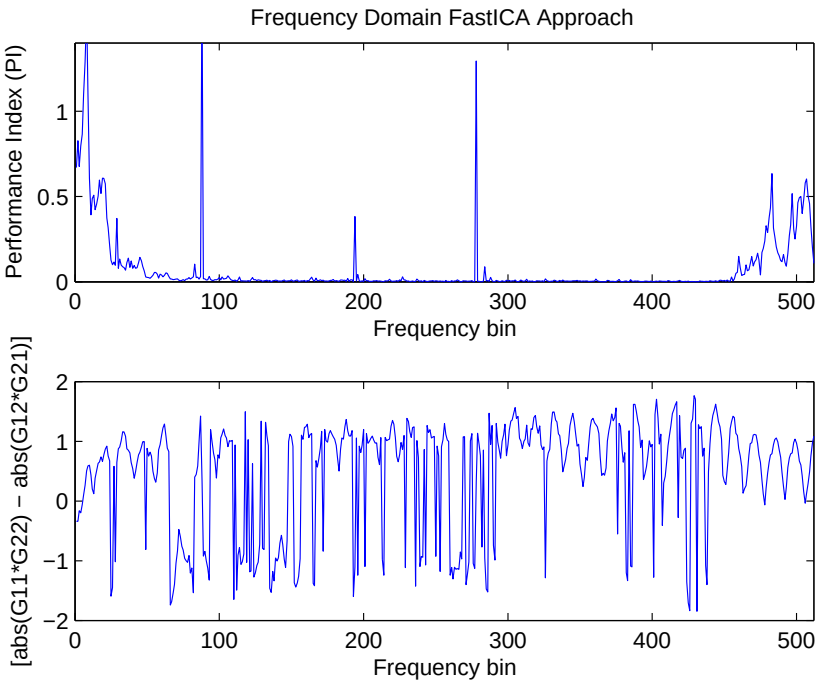


Figure 2.6. Separation performance of original FastICA Method: performance index at each frequency bin for the original FastICA method at the top and evaluation of permutation at the bottom.

In the first simulation, the mixed signals are separated by the original FastICA method with random initialization [45]. The PI and evaluation of

permutation are shown in Fig 2.6, the frequency bin range [0,512] corresponds to [0, 4000]Hz as the sampling frequency is 8kHz. The PI is approximately zero in many frequency bins, which shows that the FastICA method can separate the mixed signals well at most frequency bins. The poor behavior at low frequencies can be explained by the inter-microphone spacing which is 4cm, whereas at high frequencies it is due to low energy in the speech signals, and these effects are common for all algorithms. However, the PI is insensitive to the permutation problem; PM is not always greater than zero, which indicates there is a permutation problem, and the separation is degenerate. This simulation is used to show that the separation performance can be poor due to the permutation problem, even when the performance index is small in the majority of frequencies.

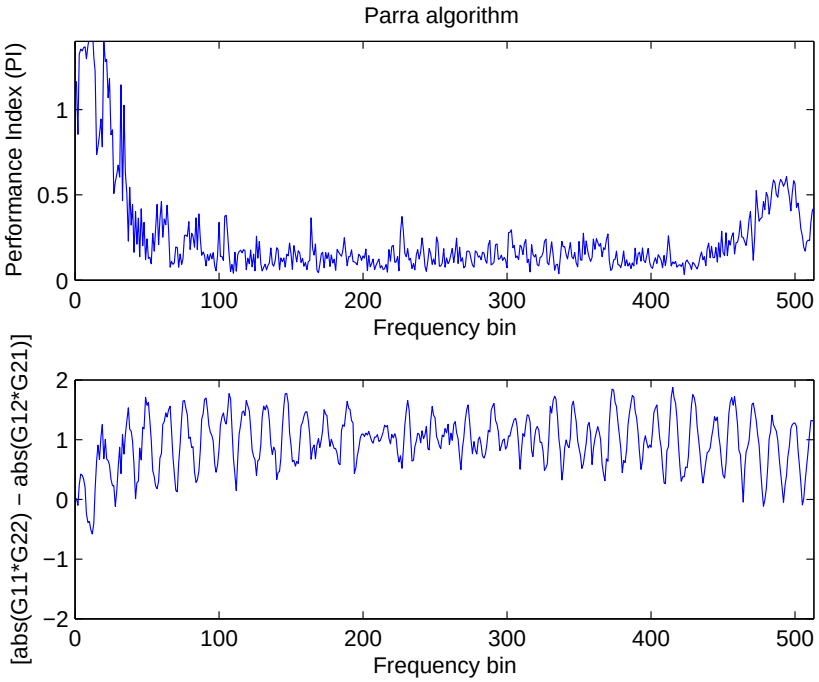


Figure 2.7. Separation performance of Parra's Method: performance index at each frequency bin for Parra's method at the top and evaluation of permutation at the bottom.

In the second simulation, the same mixed signals are separated by Parra's method [18]. The PI and permutation evaluation by using this method are

shown in Fig 2.7. Although it can separate the signals, the performance is not very good, because the PI in each frequency bin is not close to zero, which means the mixed signal is not separated very well; PM are all greater than zero except at some low frequencies and some high frequencies, which means it can mitigate the permutation problem but not in every frequency bin.

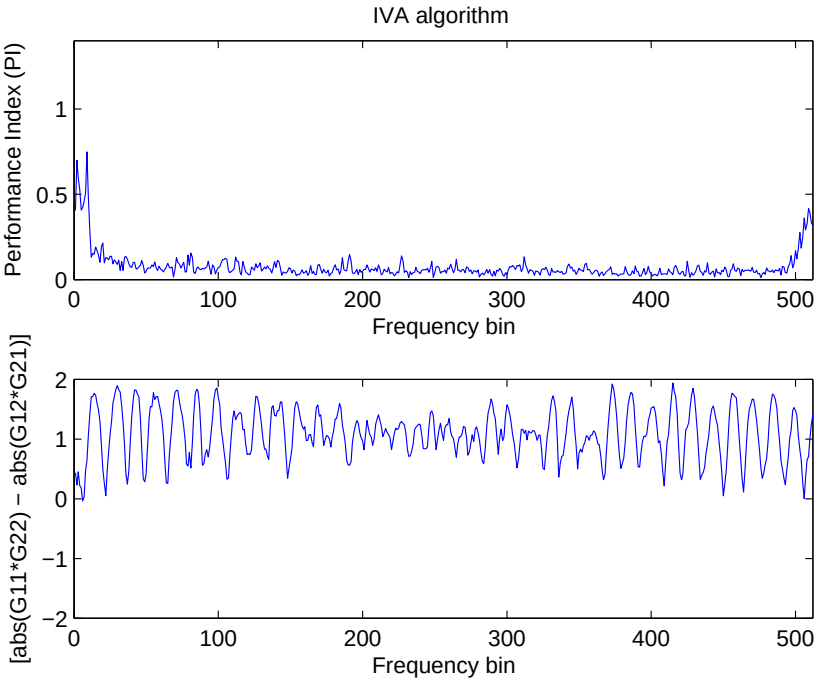


Figure 2.8. Separation performance of IVA method: performance index at each frequency bin for IVA method at the top and evaluation of permutation at the bottom.

In the third simulation, the same mixed signals are separated by the original natural gradient IVA method [27]. The PI and permutation evaluation are shown in Fig 2.8. PI is approximately zero in almost every frequency, which means it can separate the mixed signals in almost every frequency. It shows that the proposed method can separate the mixed signals very well; PM are all greater than zero in each frequency bin, which indicates it can solve the permutation problem better than Parra's method. Thus, it is obvious that the IVA method performs better than Parra's method.

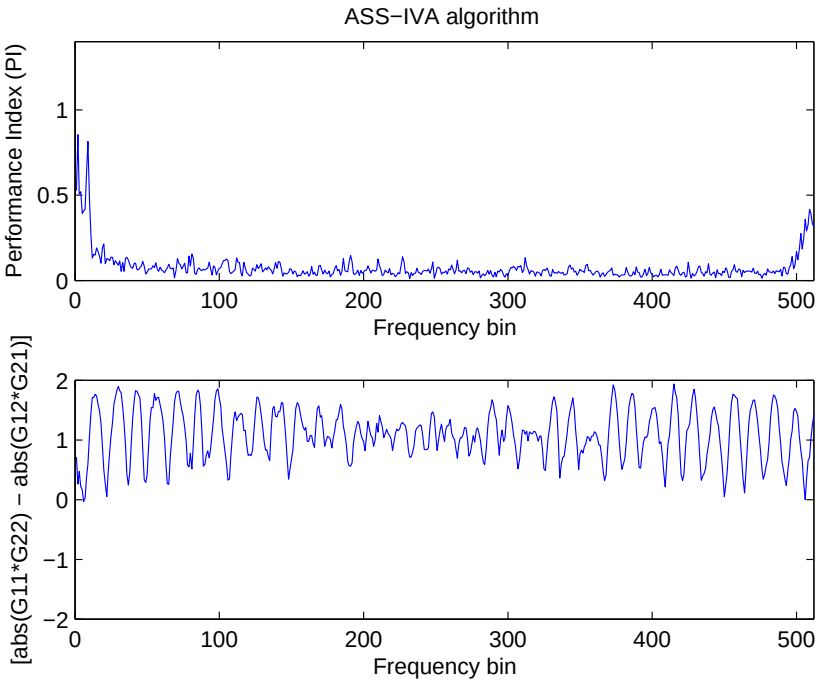


Figure 2.9. Separation performance of adaptive step size IVA method: performance index at each frequency bin for ASS-IVA method at the top and evaluation of permutation at the bottom.

The fourth simulation is using the proposed adaptive step size IVA (ASS-IVA) method to separate the mixed signals. Fig 2.9 shows the separation performance of the ASS-IVA method. The PIs are all approximately zero and the PM values are all greater than zero in almost all frequency bins. It indicates that the adaptive step size natural gradient IVA method separates the mixed signals as well as the IVA method, which means the proposed method still has a very good separation performance without permutation problem.

The convergence performance comparison between the IVA method and ASS-IVA method is shown in Fig 2.10. The convergence performance corresponds to the performance achieved in Fig 2.8 and Fig 2.9. It is clear that the convergence speed of the adaptive step size IVA is faster than the IVA method in terms of iteration numbers. The iteration times for IVA to reach the steady is approximately 100, while the iteration times for adaptive IVA

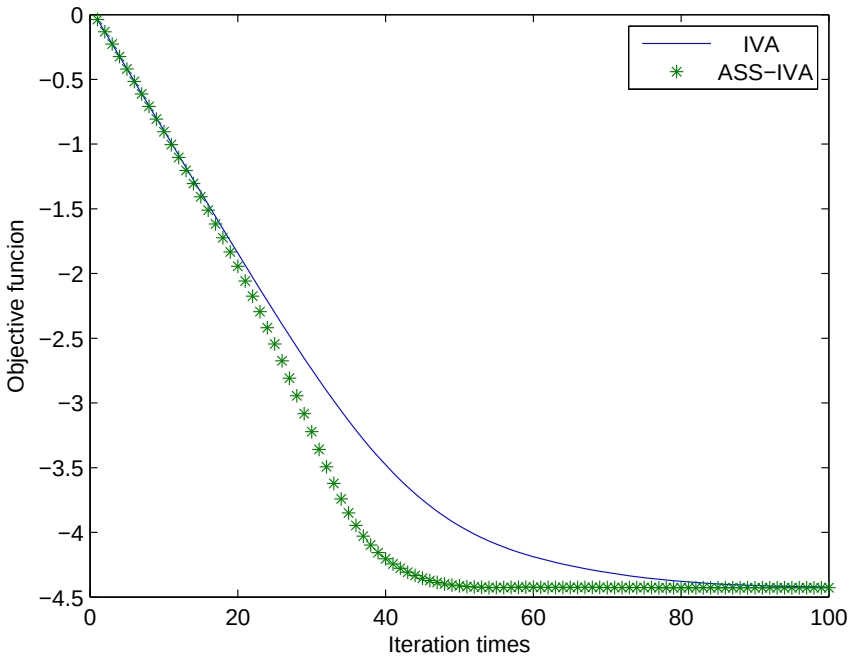


Figure 2.10. Convergence comparison between IVA and ASS-IVA. The solid line is the convergence of IVA, and the asterisk line is the convergence of ASS-IVA.

is approximately 50. It saves almost half of iteration times compared with IVA.

Moreover the separation results are also evaluated objectively by using the SIR criterion. Table 2.1 is the SIR result of different methods. The SIR improvement by using the adaptive step size IVA method is approximately 4dB. The SIR result indicates that the adaptive step size IVA has the best performance among these methods. To further improve the convergence of IVA, another more rapidly converging algorithm is considered.

Table 2.1. SIR comparison

method	Parra's	IVA	ASS-IVA
SIR	19.84	23.12	23.72

2.5.3 Fast fixed-point IVA

Fast fixed-point independent vector analysis is a rapidly converging form of IVA algorithm. Newton's method is adopted in the update, which converges quadratically and is free from selecting an efficient learning rate [29].

The objective function used by FastIVA is as follows:

$$J_{FastIVA} = \sum_{i=1}^{N} \Big(E[F(\sum_{k=1}^{K} |\hat{s}_i(k)|^2)] - \sum_{k=1}^{K} \lambda_i(k)(\mathbf{w}_i(k)^\dagger \mathbf{w}_i(k) - 1) \Big) \quad (2.5.28)$$

where $\mathbf{w}_i^\dagger$ is the i-th row of the unmixing matrix W, and λ_i is the i-th Lagrange multiplier. $F(\cdot)$ is the nonlinear function, which can take on several different forms as discussed in [29]. It is a multivariate function of the summation of the desired signals in all frequency bins.

In order to apply Newton's method in the update rules, a quadratic Taylor series polynomial approximation is introduced in the notations of complex variables as follows, which can be used for a contrast function.

$$\begin{aligned} f(\mathbf{w}) \approx & f(\mathbf{w}_o) + \frac{\partial f(\mathbf{w}_o)}{\partial \mathbf{w}^T}(\mathbf{w} - \mathbf{w}_o) + \frac{\partial f(\mathbf{w}_o)}{\partial \mathbf{w}^\dagger}(\mathbf{w} - \mathbf{w}_o)^* \\ & + \frac{1}{2}(\mathbf{w} - \mathbf{w}_o)^T \frac{\partial^2 f(\mathbf{w}_o)}{\partial \mathbf{w} \partial \mathbf{w}^T}(\mathbf{w} - \mathbf{w}_o) \\ & + \frac{1}{2}(\mathbf{w} - \mathbf{w}_o)^\dagger \frac{\partial^2 f(\mathbf{w}_o)}{\partial \mathbf{w}^* \partial \mathbf{w}^\dagger}(\mathbf{w} - \mathbf{w}_o)^* \\ & + (\mathbf{w} - \mathbf{w}_o)^\dagger \frac{\partial^2 f(\mathbf{w}_o)}{\partial \mathbf{w}^* \partial \mathbf{w}^T}(\mathbf{w} - \mathbf{w}_o) \end{aligned} \quad (2.5.29)$$

Let $\mathbf{w}_i(k)$ take place of $\mathbf{w}$, and set $f(\mathbf{w}_i(k))$ to be the summation term of the objective function

$$f(\mathbf{w}_i(k)) = E[F(\sum_{k'=1}^{K} |\hat{s}_i(k')|^2)] - \sum_{k'=1}^{K} \lambda_i(k')(\mathbf{w}_i(k')^\dagger \mathbf{w}_i(k') - 1) \quad (2.5.30)$$

The $\mathbf{w}_i(k)$ that optimizes $f(\mathbf{w}_i(k))$ will set the first order derivative

$\partial f(\mathbf{w}_i(k))/\partial \mathbf{w}_i(k)^*$ to be zero.

$$\begin{aligned}\frac{\partial f(\mathbf{w}_i(k))}{\partial \mathbf{w}_i(k)^*} \approx & \frac{\partial f(\mathbf{w}_{i,o}(k))}{\partial \mathbf{w}_i(k)^*} + \frac{\partial^2 f(\mathbf{w}_{i,o})}{\partial(\mathbf{w}_i(k))^*\partial(\mathbf{w}_i(k))^T}(\mathbf{w}_i(k) - \mathbf{w}_{i,o}(k)) \\ & + \frac{\partial^2 f(\mathbf{w}_{i,o})}{\partial(\mathbf{w}_i(k))^*\partial(\mathbf{w}_i(k))^\dagger}(\mathbf{w}_i(k) - \mathbf{w}_{i,o}(k))^* \equiv \mathbf{0}\end{aligned} \tag{2.5.31}$$

The derivative terms contained in equation (2.5.31) become:

$$\frac{\partial f(\mathbf{w}_{i,o}(k))}{\partial \mathbf{w}_i(k)^*} = E[\hat{s}_{i,o}(k)^* F'(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2)] - \lambda_i(k)\mathbf{w}_{i,o}(k) \tag{2.5.32}$$

$$\begin{aligned}&\frac{\partial^2 f(\mathbf{w}_{i,o})}{\partial(\mathbf{w}_i(k))^*\partial(\mathbf{w}_i(k))^T} \\ &= E[(F'(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2) + |\hat{s}_{i,o}(k)|^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2))\mathbf{x}(k)\mathbf{x}(k)^*] - \lambda_i(k)I \\ &\approx E[(F'(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2) + |\hat{s}_{i,o}(k)|^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k)|^2))]E[\mathbf{x}(k)\mathbf{x}(k)^*] - \lambda_i(k)I \\ &= \Big(E[(F'(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2) + |\hat{s}_{i,o}(k)|^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2))] - \lambda_i(k)\Big)I\end{aligned} \tag{2.5.33}$$

$$\begin{aligned}&\frac{\partial^2 f(\mathbf{w}_{i,o})}{\partial(\mathbf{w}_i(k))^*\partial(\mathbf{w}_i(k))^\dagger} \\ &= E[(\hat{s}_{i,o}(k)^*)^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2))\mathbf{x}(k)\mathbf{x}(k)^T] \\ &\approx E[(\hat{s}_{i,o}(k)^*)^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2))]E[\mathbf{x}(k)\mathbf{x}(k)^T] \\ &= \mathbf{0}\end{aligned} \tag{2.5.34}$$

where $F'(\cdot)$ and $F''(\cdot)$ denote the derivative and second derivative of $F(\cdot)$ respectively. The assumption in equation (2.5.33) is $E[\mathbf{x}(k)\mathbf{x}(k)^*] = I$, which is due to the whitening processing, and the assumption in equation (2.5.34) is $E[\mathbf{x}(k)\mathbf{x}(k)^T] = \mathbf{0}$, which is the complex circularity assumption.

By substitution, the iterative algorithm becomes as follows:

$$\mathbf{w}_i(k) \leftarrow \mathbf{w}_{i,o}(k) - \frac{E[\hat{s}_{i,o}(k)^* F'(\sum_{k'} |\hat{s}_{i,o}(k')|^2)] - \lambda_i(k)\mathbf{w}_{i,o}(k)}{E[(F'(\sum_{k'} |\hat{s}_{i,o}(k')|^2) + |\hat{s}_{i,o}(k)|^2 F''(\sum_{k'} |\hat{s}_{i,o}(k')|^2))] - \lambda_i(k)} \tag{2.5.35}$$

where the Lagrange multiplier $\lambda_i(k)$ is

$$\lambda_i(k) = E[|\hat{s}_{i,o}(k)|^2 F'(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2)] \tag{2.5.36}$$

Then with normalization, the learning rule is:

$$\begin{aligned} \mathbf{w}_i(k) \leftarrow & E[F^{'}(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2) + |\hat{s}_{i,o}(k)|^2 F^{''}(\sum_{k'=1}^{K} |\hat{s}_i(k')|^2))]\mathbf{w}_i(k) \\ & - E[(\hat{s}_{i,o}(k))^* F^{'}(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2)\mathbf{x}(k)] \end{aligned} \tag{2.5.37}$$

And if this is used for all sources, an unmixing matrix $W(k)$ can be constructed which needs to be decorrelated with

$$W(k) \leftarrow (W(k)(W(k))^{\dagger})^{-1/2} W(k). \tag{2.5.38}$$

The nonlinear function is derived from the source prior. When the super-Gaussian distribution used by the original natural gradient IVA algorithm is used as the source prior for the FastIVA algorithm, with the zero mean and unity variance assumptions, it takes the form

$$F(\sum_{k'=1}^{K} |\hat{s}_i(k')|^2) = \big(\sum_{k'=1}^{K} |\hat{s}_i(k')|^2\big)^{\frac{1}{2}} \tag{2.5.39}$$

2.5.4 Auxiliary Function Based IVA

In order to avoid step size tuning and derive effective iterative update rules, the auxiliary function technique is introduced, which is an extension of the expectation-maximization algorithm [46]. In the auxiliary function tech-

nique, an auxiliary function is designed for optimization. Instead of minimizing the cost function, the auxiliary function is minimized in terms of auxiliary variables. The auxiliary function technique can guarantee monotonic decrease of the cost function, and therefore provide effective iterative update rules [30]. Thus the design of the auxiliary function is the central problem.

For a general optimization problem, the target is to find a parameter vector $\Theta = \bar{\Theta}$ satisfying

$$\bar{\Theta} = argmin_{\Theta} J(\Theta) \tag{2.5.40}$$

where $J(\Theta)$ is an objective function.

In the auxiliary function technique, an auxiliary function $Q(\Theta, \tilde{\Theta})$ is designed to satisfy

$$J(\Theta) = min_{\tilde{\Theta}} Q(\Theta, \tilde{\Theta}) \tag{2.5.41}$$

where $\tilde{\Theta}$ is a vector of auxiliary variables. Then, the auxiliary function instead of the objective function is minimized. The variables being iteratively updated as

$$\tilde{\Theta}(i+1) = argmin_{\tilde{\Theta}} Q(\Theta(i), \tilde{\Theta}) \tag{2.5.42}$$

$$\Theta(i+1) = argmin_{\Theta} Q(\Theta, \tilde{\Theta}(i+1)) \tag{2.5.43}$$

where i is the iteration index. When both equations (2.5.42) and (2.5.43) are written in closed forms, the auxiliary function technique gives an efficient iterative update rule [30].

As in [30] and [46], in order to determine the proper auxiliary function for the IVA cost function, a definition is needed at first.

Definition 1 A set of real-valued functions of a vector random variable $\mathbf{z}$, S_g is defined as

$$S_g = \{g(\mathbf{z})|g(\mathbf{z}) = g_R(\|\mathbf{z}\|_2)\} \tag{2.5.44}$$

where $g_R(r)$ is a continuous and differentiable function of a real variable r and $g'_R(r)/r$ is continuous everywhere together with monotonically decreasing in $r > 0$.

The function $g_R(r)$ is derived from the source prior [46]. For the original IVA algorithm,

$$g(\mathbf{z}) = r \tag{2.5.45}$$

where $r = \|\mathbf{z}\|_2$.

Based on this definition, two theorems are introduced to design the auxiliary function [30].

Theorem 1 For any $g(\mathbf{z}) = g_R(||\mathbf{z}||_2) \in S_g$

$$g(\mathbf{z}) \leq \frac{g'_R(r_0)}{2r_0}r^2 + (g_R(r_0) - \frac{r_0 g'_R(r_0)}{2}) \tag{2.5.46}$$

where $r = ||\mathbf{z}||_2$, holds for any $\mathbf{z}$ and r_0. The equality sign is satisfied if and only if $r_0 = r = ||\mathbf{z}||_2$.

Proof: Construct the function :

$$f(r) = \frac{g'_R(r_0)}{2r_0}r^2 + (g_R(r_0) - \frac{r_0 g'_R(r_0)}{2}) - g_R(r) \tag{2.5.47}$$

and differentiating,

$$f'(r) = \frac{g'_R(r_0)}{r_0}r - g'_R(r) = r(\frac{g'_R(r_0)}{r_0} - \frac{g'_R(r)}{r}) \tag{2.5.48}$$

According to the Definition 1, $g'_R(r)/r^2$ monotonically decreases in $r > 0$. It is also evident that $f'(r_0) = 0$. Then, $f(r)$ has a unique minimum value at $r = r_0$, because $f(r)$ is continuous everywhere and $f(r_0) = 0$. Therefore,

it follows that Theorem 1 is proved.

Theorem 2 For any $g(\mathbf{z}) = g_R(\|\mathbf{z}\|_2)) \in S_g$, let

$$Q(\mathbf{W}, \mathbf{V}) = \sum_{k=1}^{K} Q_k(W(k), \mathbf{V}(k)) \tag{2.5.49}$$

where

$$Q_k(W(k), \mathbf{V}(k)) = \frac{1}{2} \sum_{i=1}^{n} \mathbf{w}_i^{\dagger}(k) V_i(k) \mathbf{w}_i(k) - log|detW(k)| + R \tag{2.5.50}$$

and

$$V_i(k) = E[\frac{g_R'(r_i)}{r_i} \mathbf{x}(k)\mathbf{x}(k)^{\dagger}] \tag{2.5.51}$$

where r_i is a positive random variable, $\mathbf{V}(k)$ represents a set of $V_i(k)$ for any i, $\mathbf{V}$ represents a set of $V_i(k)$ for any i and k, and R is a constant term independent of W. Then,

$$J(\mathbf{W}) \leq Q(\mathbf{W}, \mathbf{V}) \tag{2.5.52}$$

holds for any $\mathbf{W}$ and any $\mathbf{V}$. The equality sign holds if and only if

$$r_i = \sqrt{\sum_{k=1}^{K} |\mathbf{w}_i^{\dagger}(k)\mathbf{x}(k)|^2} \tag{2.5.53}$$

and $J(\mathbf{W})$ is the IVA cost function:

$$J(\mathbf{W}) = \sum_{i=1}^{n} E[g(\mathbf{y}_i)] - \sum_{k=1}^{K} log|detW(k)| \tag{2.5.54}$$

Proof: Apply Theorem 1 to $E[g(\mathbf{y}_i)]$

$$\begin{aligned} E[g(\mathbf{y}_i)] &\leq E[\frac{g'_R(r_i)}{2r_i}\sum_{k=1}^{K}|\mathbf{w}_i^\dagger(k)\mathbf{x}(k)|^2] + R_i \\ &= \sum_{k=1}^{K}\mathbf{w}_i^\dagger(k)E[\frac{g'_R(r_i)}{2r_i}\mathbf{x}(k)\mathbf{x}(k)^h]\mathbf{w}_i(k) + R_i \\ &= \frac{1}{2}\sum_{k=1}^{K}\mathbf{w}_i^\dagger(k)V_i(k)\mathbf{w}_i(k) + R_i \end{aligned} \tag{2.5.55}$$

where $V_i(k)$ is defined as in equation (2.5.51) and R_i is a constant term independent of $\mathbf{w}_i(k)$. The equality sign holds if and only if equation (2.5.53) is satisfied. Summing up equation (2.5.55) over all i, proves Theorem 2.

Theorem 2 shows that $Q(\mathbf{W}, \mathbf{V})$ is the proper auxiliary function for the IVA cost function. So the update rules can be derived according to this auxiliary function.

The update rules to minimize the auxiliary function $Q(\mathbf{W}, \mathbf{V})$ are derived dependent on the auxiliary variables update, namely $\mathbf{W}$ and $\mathbf{V}$ in this case. The minimization of $\mathbf{V}$ can be achieved by using equations (2.5.51) and (2.5.53) according to Theorem 2. The update of $\mathbf{W}$ can be obtained

$$\frac{\partial Q(\mathbf{W}, \mathbf{V})}{\partial \mathbf{w}_i^*(k)} = \frac{1}{2}V_i(k)\mathbf{w}_i(k) - \frac{\partial log|detW(k)|}{\partial \mathbf{w}_i^*(k)} = \mathbf{0} \tag{2.5.56}$$

Instead of simultaneously updating all $\mathbf{w}_i(k)$, only the update of one $\mathbf{w}_i(k)$ is focused upon [30]. It is deduced that

$$\mathbf{w}_i^\dagger(k)V_i(k)\mathbf{w}_i(k) = 1 \tag{2.5.57}$$

and

$$\mathbf{w}_i^\dagger(k)V_i(k)\mathbf{w}_j(k) = 0 \qquad (i \neq j) \tag{2.5.58}$$

According to the method described in [46], the updates of $\mathbf{w}_i(k)$ can be

achieved by

$$\mathbf{w}_i(k) = (W(k)V_i(k))^{-1}\mathbf{e}_i \tag{2.5.59}$$

where $\mathbf{e}_i$ is a unity vector, the i-th element of which is unity. Finally, the normalization process is needed to satisfy equation (2.5.57)

$$\mathbf{w}_i(k) = \frac{\mathbf{w}_i(k)}{\sqrt{\mathbf{w}_i^{\dagger}(k)V_i(k)\mathbf{w}_i(k)}} \tag{2.5.60}$$

In summary, the overall update rules are the following:

(1) $r_i = \sqrt{\sum_{k=1}^{K} |\mathbf{w}_i^{\dagger}(k)\mathbf{x}(k)|^2}$

(2) $V_i(k) = E[\frac{g_R'(r_i)}{r_i}\mathbf{x}(k)\mathbf{x}(k)^{\dagger}]$

(3) $\mathbf{w}_i(k) = (W(k)V_i(k))^{-1}\mathbf{e}_i$

(4) $\mathbf{w}_i(k) = \frac{\mathbf{w}_i(k)}{\sqrt{\mathbf{w}_i^{\dagger}(k)V_i(k)\mathbf{w}_i(k)}}$

2.5.5 Summary

In this chapter, background knowledge related to convolutive blind source separation problem was firstly introduced. The second order statistic methods and higher order statistic methods were discussed. Then, the original natural gradient independent vector analysis algorithm was introduced, followed by the proposed adaptive step size natural gradient IVA. The illustrative experimental results confirm that the IVA algorithm can mitigate the permutation problem, and the proposed adaptive step size natural gradient IVA can achieve faster convergence in terms of the iteration numbers compared with the original natural gradient IVA algorithm. At the end of this chapter, two fast form IVA algorithms, i.e. the FastIVA algorithm and Aux-

IVA algorithm, were briefly introduced. However, the IVA algorithms are not always robust, a particular permutation problem sometimes happens. This specific problem and the corresponding solutions will be discussed in the next chapter.

Chapter 3

BLOCK PERMUTATION PROBLEM OF INDEPENDENT VECTOR ANALYSIS

3.1 Introduction

Independent vector analysis is proposed to theoretically avoid the classical permutation problem inherent to ICA method. However, a specific problem, i.e. the block permutation problem, sometimes happens when using IVA. In recent research work [47], a similar problem with the convergence of IVA is termed as "partial permutation", but without analysis about why this problem can occur. In this chapter, this problem is discussed by analyzing the cost function, and two kinds of solutions are proposed. The first solution exploits the phase continuity of the unmixing matrix to adjust the misalignments and thereby retain consistent permutation across all frequency bins. The second solution adopts an improved overlapped chain type dependency model to mitigate this problem. The first scheme is shown when the original natural gradient IVA is used, and the second one is illustrated by taking AuxIVA as an example. However, both of these two strategies can be adapted

to all kinds of IVA algorithm for addressing the block permutation problem. This chapter is targeted at satisfying the second objective of this thesis.

3.2 Block Permutation of IVA

The independent vector analysis (IVA) algorithm is essentially based on the ICA algorithm with certain modifications. It exploits a dependency model which captures inter-frequency dependencies. However, when employing the IVA algorithm in practice, it has been found that alignment is generally achieved between frequency bins which are closely spaced, but it is difficult to guarantee the alignment of all frequency bins because distant spectral components may not be highly dependent. Moreover, IVA can converge well when the temporal activity between the sources is strongly uncorrelated, otherwise the IVA method will exhibit poor convergence with an accompanying block permutation problem. Additionally, this problem can also be understood by examining the cost function. When the block permutation problem happens, there is a block of frequency bins which can be denoted $[k_b, k_e]$, whose alignment is different from the other frequency bins. This indicates that the corresponding rows of the unmixing matrix $W(k)$ are exchanged. This modified unmixing matrix is denoted as $\overline{W}(k)$.

This problem can be discussed by analyzing the cost function of IVA. Thus, the cost function as introduced in Chapter 2 is repeated here. The cost function for IVA is the Kullback-Leibler divergence between the joint probability density function $p(\hat{\mathbf{s}})$ and the product of probability density functions of the individual source vectors $\prod p(\hat{\mathbf{s}}_i)$ [27],

$$\begin{aligned} J &= KL(p(\hat{\mathbf{s}})||\prod p(\hat{\mathbf{s}}_i)) \\ &= \int p(\hat{s}_1 \cdots \hat{s}_n) \log \frac{p(\hat{s}_1 \cdots \hat{s}_n)}{\prod p(\hat{\mathbf{s}}_i)} d\hat{s}_1 \cdots d\hat{s}_n \\ &= const - \sum_{k=1}^{K} \log |det(W(k))| - \sum_{i=1}^{n} E[\log p(\hat{s}_i)] \end{aligned} \tag{3.2.1}$$

A 2×2 case is used to illustrate this problem. Because the first term of the cost function is a constant and

$$\log|det(W(k))| = \log|det(\overline{W}(k))|, \tag{3.2.2}$$

the last term of the cost function is the only thing changed when the block permutation problem happens. Due to the assumption that the multivariate Laplacian distribution is used as the source prior for the original IVA algorithm, the last term is defined as

$$J_1 = E\Big[-\sqrt{\sum_k |\hat{s}_1(k)|^2}\Big] + E[-\sqrt{\sum_k |\hat{s}_2(k)|^2}\Big]. \tag{3.2.3}$$

When the block permutation happens, the last term becomes

$$\begin{aligned} J_2 = E\Big[-\sqrt{\sum_1^{k_b-1} |\hat{s}_1(k)|^2 + \sum_{k_b}^{k_e} |\hat{s}_2(k)|^2 + \sum_{k_e+1}^{K} |\hat{s}_1(k)|^2}\Big] \\ + E\Big[-\sqrt{\sum_1^{k_b-1} |\hat{s}_2(k)|^2 + \sum_{k_b}^{k_e} |\hat{s}_1(k)|^2 + \sum_{k_e+1}^{K} |\hat{s}_2(k)|^2}\Big]. \end{aligned} \tag{3.2.4}$$

So that when

$$\sum_{k_b}^{k_e} |\hat{s}_1(k)|^2 = \sum_{k_b}^{k_e} |\hat{s}_2(k)|^2 \tag{3.2.5}$$

it follows that $J_1 = J_2$ which implies the cost function achieves the same value as there is no block permutation. This analysis indicates that there is no penalty for IVA converging to a block permutation solution. This explains why the block permutation problem happens, and it is emphasized that this problem is different from the conventional permutation problem encountered in frequency domain ICA where such block permutation problems are generally not observed. An example of this problem is shown in Fig 3.1.

Fig. 3.1 demonstrates the block permutation problem in IVA for an ex-

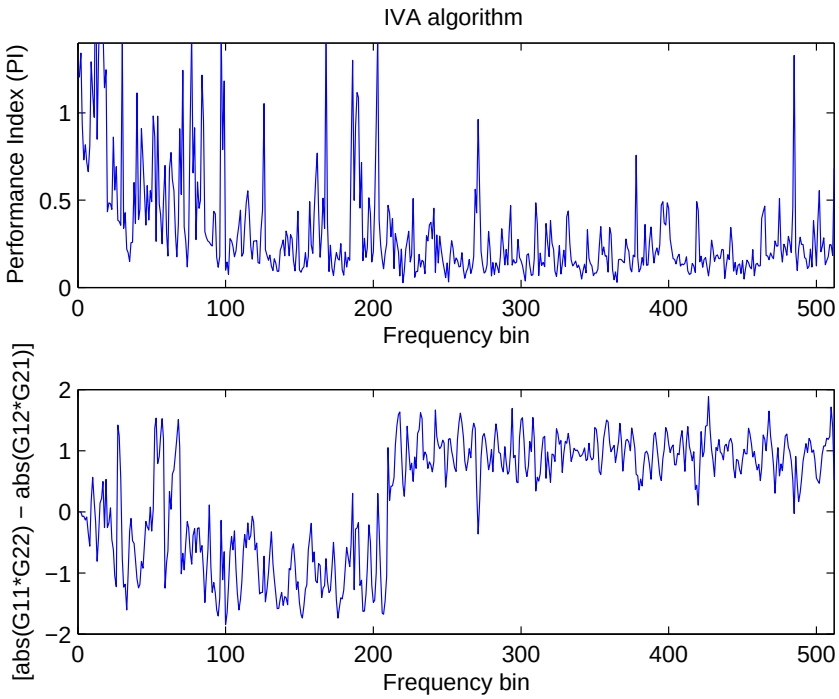


Figure 3.1. Example of the block permutation problem of IVA.

actly determined 2×2 case. The upper part of the figure is the performance index as a function of frequency [12], as given by equation (2.5.27). Note that the larger values of PI outside of the axis range represent poor separation, and therefore can be chopped. The lower part is the permutation measurement. When the permutation measurement is equal to 1, it means the overall system matrix G is an identity matrix; while if it equals to -1, it means G is [0 1;1 0]. Both of the mixtures are well separated, however, they indicate two alignments across frequencies. It is clear to see that there is a block permutation in the permutation measurement, which means the alignments of the separated signals are different. Thus, when the separated signals are transformed back to the time domain, they will be still mixed without full separation.

3.3 Robust IVA Exploiting Phase Continuity of the Unmixing Matrix

The block permutation problem can be addressed by exploiting phase continuity of the unmixing matrix. The natural gradient IVA is taken as an example to show how to use phase continuity to correct the misalignment and keep the permutation consistent across all the frequency bins.

3.3.1 Phase Continuity of the Unmixing Matrix

In the exactly determined 2×2 anechoic case, each observed mixture can be viewed as a sum of delayed and scaled versions of the original signals according to the position of the sources. In the frequency domain, each observed component could be viewed as a sum of phase-rotated versions of the frequency domain component of the original signals [48]. The mixing matrix at the k-th frequency becomes

$$H(k) = \begin{pmatrix} |h_{11}|e^{-j2\pi k t_{11}} & |h_{12}|e^{-j2\pi k t_{12}} \\ |h_{21}|e^{-j2\pi k t_{21}} & |h_{22}|e^{-j2\pi k t_{22}} \end{pmatrix} \tag{3.3.1}$$

where $|h_{ij}|$ and t_{ij} denote the channel gain and transmit time between the i-th microphone and j-th source respectively, and $|\cdot|$ is the magnitude operator.

Ideally, the unmixing matrix should be the inverse of the mixing matrix.

$$W(k) = \begin{pmatrix} |h_{22}|e^{-j2\pi k t_{22}} & -|h_{12}|e^{-j2\pi k t_{12}} \\ -|h_{21}|e^{-j2\pi k t_{21}} & |h_{11}|e^{-j2\pi k t_{11}} \end{pmatrix} \frac{1}{det(H(k))} \tag{3.3.2}$$

It is clear that the phase difference between any two elements of the unmixing matrix is a linear function according to the frequency, because t_{11}, t_{12}, t_{21} and t_{22} are fixed in this ideal situation. Considering the scaling

problem inherent to ICA, the unmixing matrix becomes:

$$W(k) = \begin{pmatrix} \frac{c1(k)}{det(H(k))} & 0 \\ 0 & \frac{c2(k)}{det(H(k))} \end{pmatrix} \times \begin{pmatrix} |h_{22}|e^{-j2\pi kt_{22}} & -|h_{12}|e^{-j2\pi kt_{12}} \\ -|h_{21}|e^{-j2\pi kt_{21}} & |h_{11}|e^{-j2\pi kt_{11}} \end{pmatrix} \tag{3.3.3}$$

where c1 and c2 are arbitrary and generally different non zero complex scaling factors. Because they are complex, they can affect the phase information of the unmixing matrix. However, the phase difference between the columns will be invariant to the scale factors as observed by the ratio between the elements of the first row

$$r = -\frac{|h_{22}|}{|h_{12}|}e^{-j2\pi k(t_{22}-t_{12})}. \tag{3.3.4}$$

In the exactly determined case with more than two sources, the phase difference between columns will still be invariant [49]. In a real environment, due to the reverberations, the linearity may be distorted. However, this information can still be useful if it is taken as a criterion to judge the separation performance, because after mitigating reverberation problems, the phase difference will be essentially continuous with a trend approaching linearity.

3.3.2 Robust IVA Based on Phase Continuity of the Unmixing Matrix

By observing Fig. 3.1, it is clear the problem can be solved if where the block permutation problem happens is known. However, the permutation measurement needs prior knowledge of the mixing matrix. In practical BSS it is impossible to have access to such prior knowledge. However, the phase continuity of the unmixing matrix can be exploited to determine where this problem happens, because the continuity of the unmixing matrix is likely to

be destroyed when the block permutation happens.

When the block permutation happens, the sign of the permutation measurement $[abs(G_{11}G_{22}) - abs(G_{12}G_{21})]$ in the lower frequency bins is opposite to the higher bins. In order to solve the block permutation problem, a reference sign for the permutation measurement is required. Then an permutation matrix A is used to keep the sign of the permutation measurement consistent with the reference sign. Thus the overall system matrix $G = AWH$, where A is the identity matrix when the sign of the permutation measurement is the same as the reference sign, and A is [0 1;1 0] when the sign of the permutation measurement is opposite to the reference sign.

In this scheme therefore the IVA method is still used initially to separate the mixtures, but after the unmixing matrix is obtained by IVA, the phase information of the unmixing matrix is checked to observe the occurrence of the block permutation problem, where

$$phase(W(k)) = \begin{pmatrix} phase(w_{11}(k)) & phase(w_{12}(k)) \\ phase(w_{21}(k)) & phase(w_{22}(k)) \end{pmatrix} \tag{3.3.5}$$

Then the phase difference between columns is calculated. The first row is taken as the observation target as in equation (3.3.4).

$$\begin{aligned} \Delta phase(W(k)) &= phase(w_{11}(k)) - phase(w_{12}(k)) \\ &= 2\pi k(t_{22} - t_{12}) \end{aligned} \tag{3.3.6}$$

Ideally, the sign of $\Delta phase(W(k))$ should be the same across all frequencies. In order to identify the reference sign, a test block over the high frequencies is set assuming that high frequency range information is reliable, thereby avoiding spatial aliasing problems at lower frequencies, and the mean of $\Delta phase(W(k))$ is calculated. If it is greater than zero, the sign

of all the phase differences should be positive. Otherwise, all of them should be negative. Then the phase difference is checked frequency-by-frequency, and the unmixing matrices are adjusted to make sure that the sign of the permutation measurement is consistent.

The flow of this robust IVA (RIVA) method becomes:

(1) Obtain the unmixing matrix for every frequency by using the IVA method.

(2) Calculate the phase of the unmixing matrices according to equation (3.3.5).

(3) Calculate the phase differences of the first row according to equation (3.3.6).

(4) Form a test block over the high frequencies, and calculate the mean of the phase difference in this block to set the reference sign.

(5) Check the signs of the phase differences across all frequencies, if any are different from the reference sign, interchange the rows of the unmixing matrix otherwise leave them unaltered.

3.3.3 Experimental Results

In the experiments, the proposed robust IVA algorithm is used to solve the block permutation problem. The source signals are from Sawada's website "http://www.kecl.ntt.co.jp/icl/signal/sawada", which is approximately 7 seconds long for each source. The image method is used to generate the room response model [50], and the size of the room is $7 \times 5 \times 3m^3$, the STFT length T = 1024, and RT60 = $150ms$. A 2×2 mixing case is used, for which the microphone positions are [2.36, 2.50, 1.50] and [2.40, 2.50, 1.50] respectively. The sampling frequency Fs is 8kHz. The separation performance is evaluated objectively by the performance index (PI), the signal-to-distortion ratio (SDR) and signal-to-interference ratio (SIR).

The SDR and SIR used here are proposed by Vincent [51]. The estimated

source signal is decomposed as:

$$\hat{s}_j = s_{target} + e_{interf} + e_{noise} + e_{artif} \tag{3.3.7}$$

where s_{target} is a version of s_j modified by an allowed distortion; e_{interf}, e_{noise} and e_{artif} are the interferences, noise and artifacts error terms respectively.

Then the SDR and SIR can be calculated as

$$SDR = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf} + e_{noise} + e_{artif}||^2} \tag{3.3.8}$$

$$SIR = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf}||^2} \tag{3.3.9}$$

where $||\cdot||^2$ denotes the energy of the signal.

In the first experiment the source positions are set as [3.25, 3.8, 1.50] and [1.75, 3.8, 1.50]. Fig. 3.2 shows the separation performance, and Fig. 3.3 shows the phase difference of the unmixing matrix. The frequency bin range [0, 512] corresponds to [0, 4000]Hz as the sampling frequency is 8kHz. The upper part of Fig. 3.2 is the performance index, and the lower part is the permutation measurement. It is clear to see the occurrence of the block permutation problem in the original IVA method [27]. The phase difference in Fig. 3.3 also confirms the block permutation problem. And the objective evaluations of SIR and SDR are all negative for the original IVA method (shown in the Table 3.1). This verifies that there is poor source separation due to the block permutation problem.

Next the robust IVA (RIVA) method is used to separate the same mixtures, and the length of the test block which is used to identify the reference sign is 200 frequency bins from 312 to 512. Fig. 3.4 shows an improved separation performance in comparison with Fig. 3.2. By observing Fig. 3.5, it is evident that the phase difference has a linear trend and there is no block permutation problem. Moreover, the comparison of objective measures is

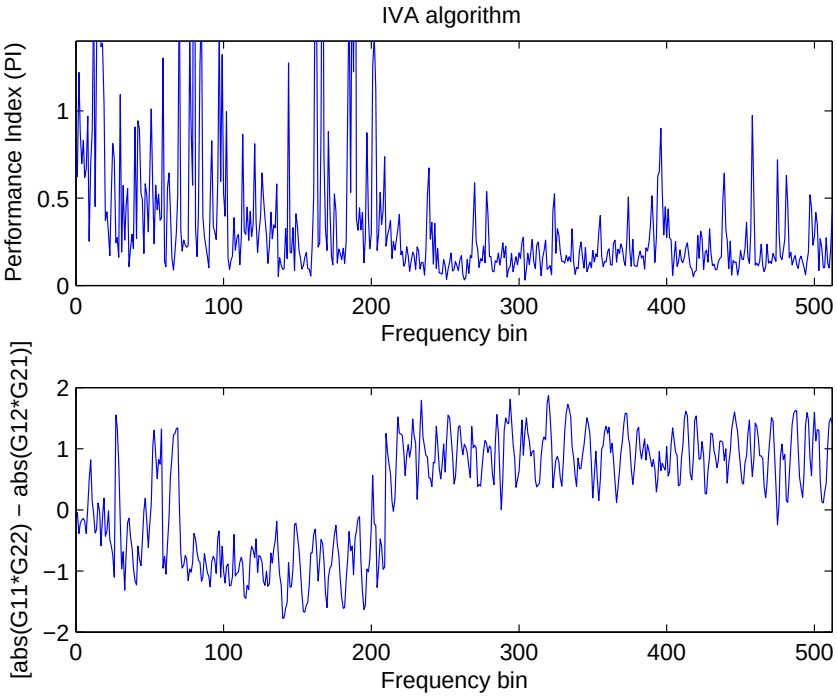


Figure 3.2. Separation performance by using original IVA.

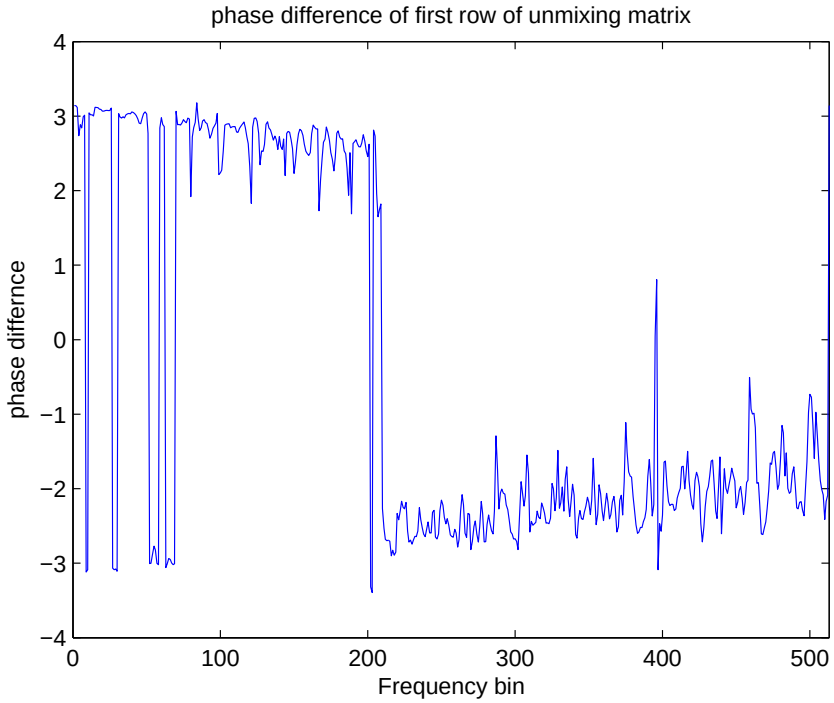


Figure 3.3. Phase difference by using orginal IVA.

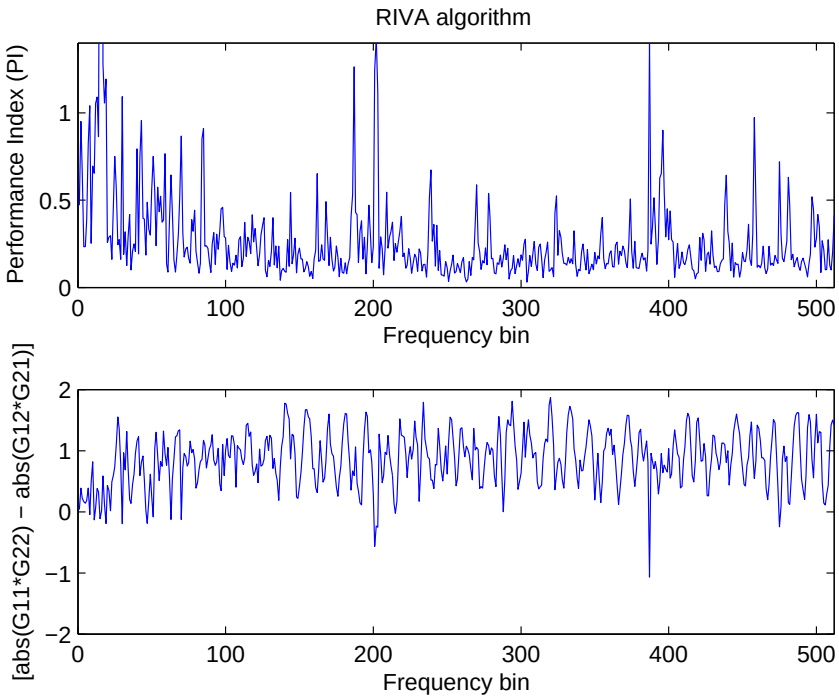


Figure 3.4. Separation performance by using robust IVA.

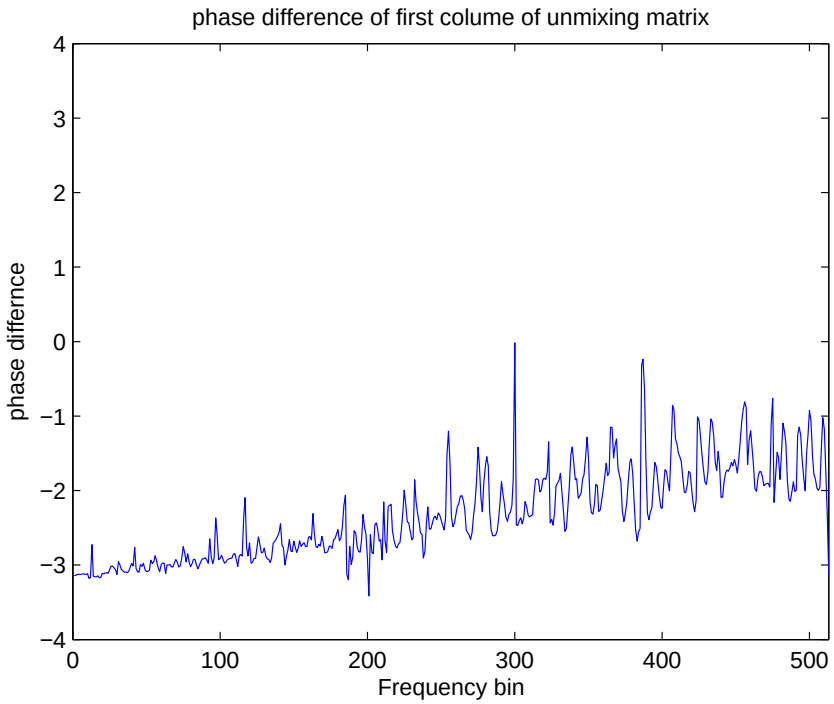


Figure 3.5. Phase difference by using robust IVA.

shown in Table 3.1. The SIR and SDR averaged across the two sources are 8.50dB and 17.89dB respectively, which also verify that the proposed RIVA solves the block permutation problem and separates the sources.

Table 3.1. SDR and SIR comparison of the first experiment.

	IVA			RIVA		
	S1	S2	Mean	S1	S2	Mean
SDR(dB)	-5.13	-4.23	-4.68	8.63	8.36	8.50
SIR(dB)	-3.40	-3.26	-3.33	22.03	13.74	17.89

In the second experiment, the source positions are set as [2.75, 3.8, 1.50] and [1.75, 3.8, 1.50]. Figs. 3.6 and 3.7 show the separation performance and the phase difference when the original IVA is used. The permutation measurement in Fig. 3.6 shows that there is no block permutation problem. Moreover, Fig. 3.7 also confirms this. However, there is a permutation problem at certain frequencies. And the objective measures are shown in Table 3.2.

Then the robust IVA (RIVA) scheme is also used to separate the same mixtures. The results are given in Fig 3.8 and Fig 3.9. The separation performance is improved by mitigating the permutation problem at certain frequency bins. And the comparisons based on objective measures of SDR and SIR are shown in Table 3.2. The results verify that the proposed scheme improves SIR and SDR by 2.8dB and 1.8dB respectively, when the original IVA method has no block permutation problem, which again confirms the robustness of the proposed technique.

Different source signals and different locations are adopted to perform 10 experiments where there is no block permutation problem, and finally the comparison is obtained as shown in Fig 3.10. The average SDR and SIR improve approximately by 1.3dB and 3.0dB respectively.

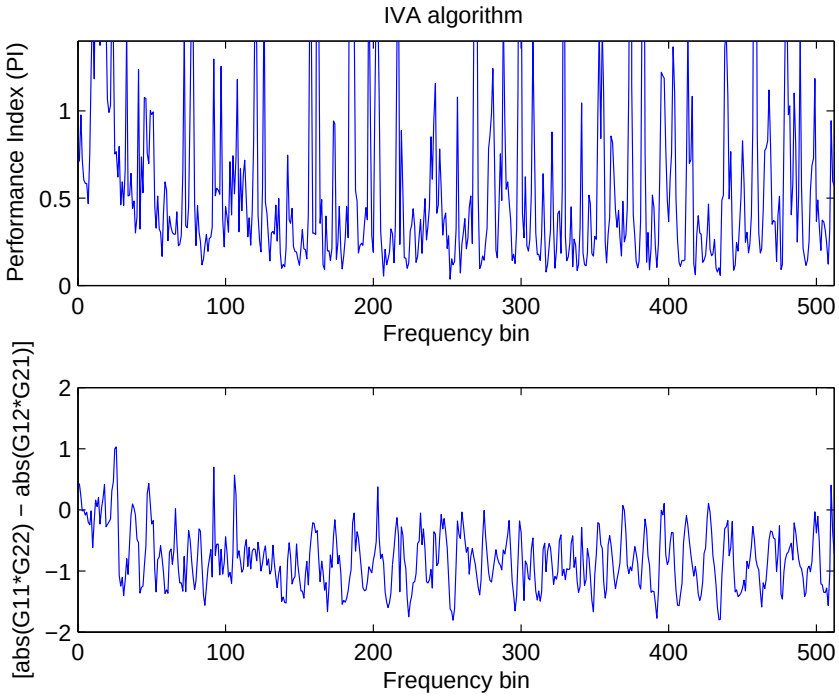


Figure 3.6. Separation performance by using original IVA.

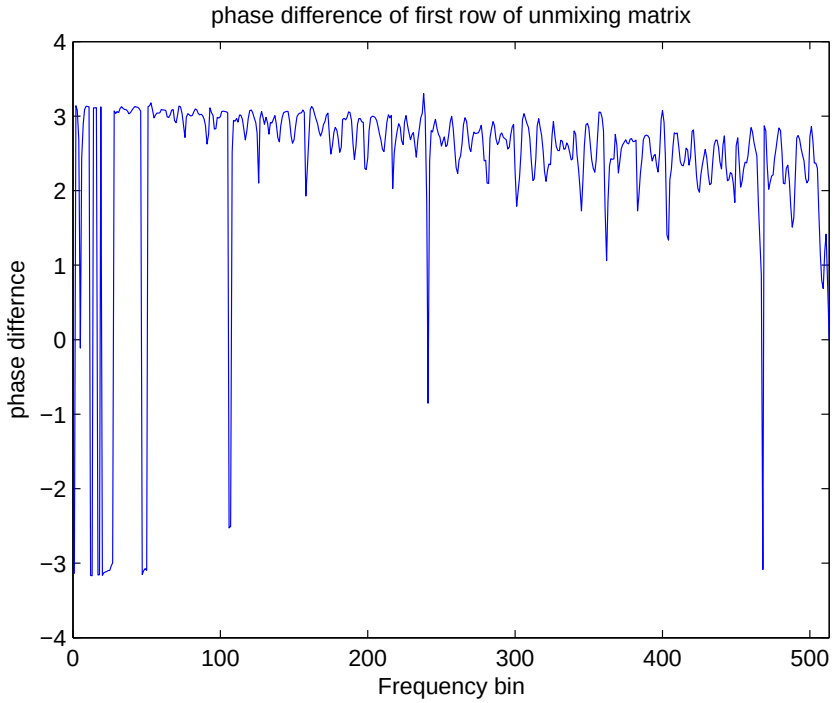


Figure 3.7. Phase difference by using orginal IVA.

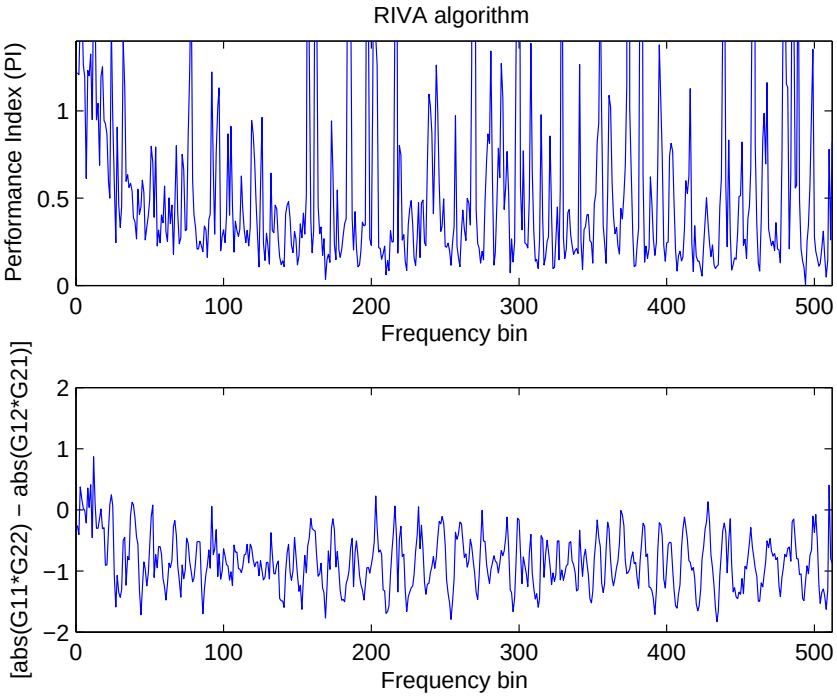


Figure 3.8. Separation performance by using robust IVA.

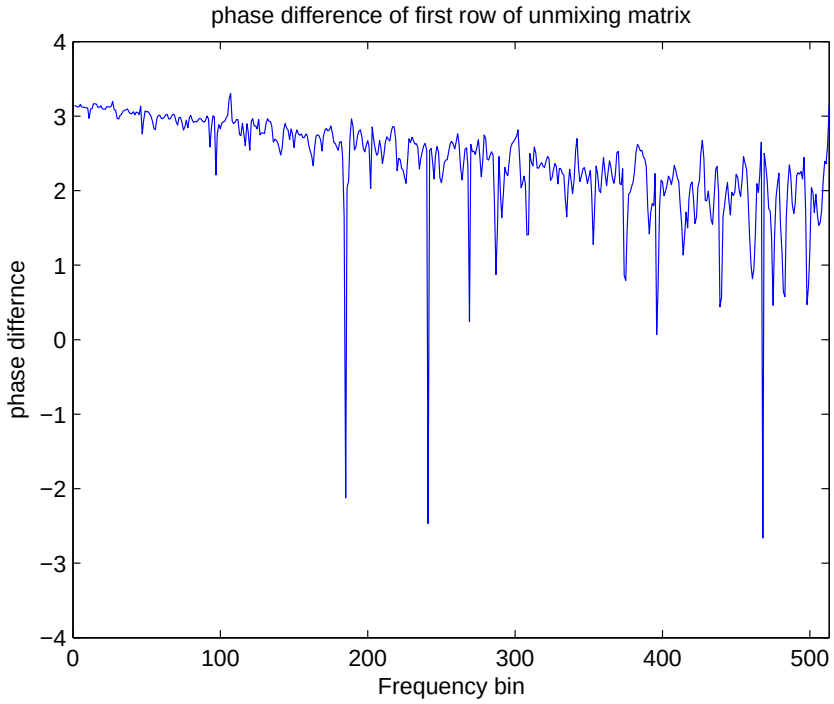


Figure 3.9. Phase difference by using robust IVA.

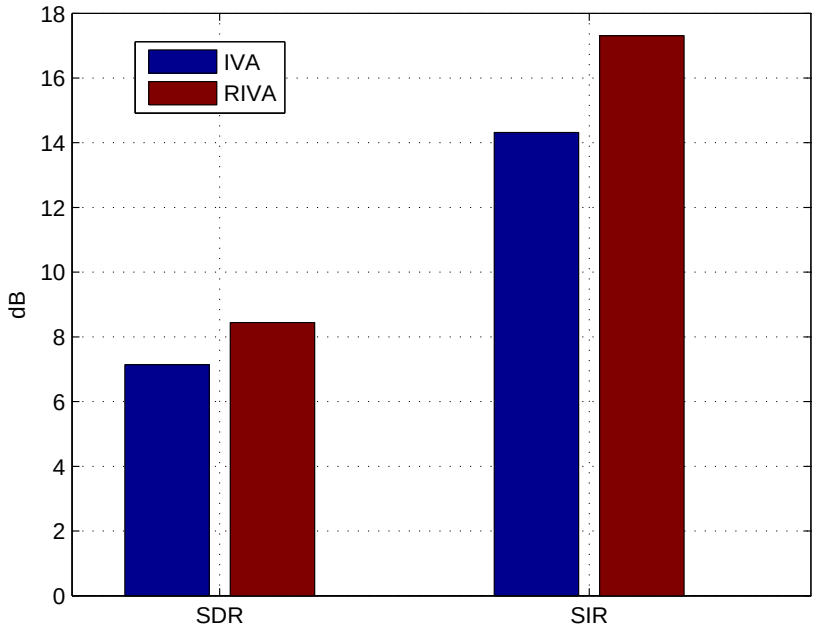


Figure 3.10. Average SDR and SIR comparison.

Table 3.2. SDR and SIR comparison of the second experiment.

	IVA			RIVA		
	S1	S2	Mean	S1	S2	Mean
SDR(dB)	4.38	3.84	4.11	5.42	6.40	5.91
SIR(dB)	8.54	6.25	7.40	9.51	10.92	10.21

3.4 Overcoming Block Permutation by Using an Improved Dependency Model

Another solution for overcoming the block permutation problem is adopting an improved overlapped chain type dependency model. The AuxIVA is taken as the example to show that AuxIVA with the improved dependency model can mitigate the block permutation problem.

3.4.1 Block Permutation for AuxIVA

An auxiliary function based independent vector analysis (AuxIVA) algorithm has been proposed recently [30]. The auxiliary function method is an extension of the expectation-maximization algorithm, which is widely used for statistical inference problems in signal processing. The AuxIVA algorithm can avoid the step size tuning problem in conventional IVA and gives effective iterative update rules which can guarantee the monotonic decrease of the objective function at each update [30]. The original AuxIVA assumes the source probability density function is multivariate super-Gaussian which is overall hyper-spherical or radially symmetric. The radial symmetry assumes that the dependency between all the frequency bins is the same, which is a constraint and can lead to a block permutation problem which results in poorer separation.

The auxiliary function used in the AuxIVA algorithm formulation is a function of unmixing matrices $\mathbf{W}$ and weighted covariance matrices $\mathbf{V}$. $\mathbf{W}$ is a set of unmixing matrices $W(k)$ for all the frequency bins, and $W(k) = (\mathbf{w}_1, \cdots, \mathbf{w}_N)^\dagger$. $\mathbf{V}$ is a set of $V_i(k)$ for any i and k.

$$Q(\mathbf{W}, \mathbf{V}) = \sum_{k=1}^{K} Q(k) = \sum_{k=1}^{K} (\frac{1}{2} \sum_{i=1}^{N} \mathbf{w}_i^\dagger(k) V_i(k) \mathbf{w}_i(k) - \log|det(W(k))|) + R \tag{3.4.1}$$

where R is a scalar constant term, and $V_i(k)$ is the weighted covariance matrix at the k-th frequency bin, which can be calculated as:

$$V_i(k) = E[\frac{g_R'(r_i)}{r_i} \mathbf{x}(k)\mathbf{x}(k)^\dagger] \tag{3.4.2}$$

When the multivariate Laplacian source prior is adopted, the contrast function becomes

$$g_R(r_i) = r_i = \sqrt{\sum_{k=1}^{K} |\hat{s}_i(k)|^2} \tag{3.4.3}$$

For the original AuxIVA algorithm, $g'_R(r_i)/r_i$ is used to retain the dependencies between frequencies. It corresponds to a hyper-spherical model and assumes the dependencies between all frequency bins are all the same. However, it is high likely that the dependencies between frequency bins which are far away from each other are weak. Due to this reason, the block permutation problem can happen as shown in Fig 3.1. The reason for this problem can also be understood by examining the cost function. It is defined that

$$Q' \triangleq \sum_{i=1}^{N} \mathbf{w}_i^{\dagger}(k) V_i(k) \mathbf{w}_i(k) \tag{3.4.4}$$

Assuming there are two sources and two mixtures, by using (3.4.2) and (3.4.3)

$$\begin{aligned} Q'_1 &= \mathbf{w}_1^{\dagger}(k) E\Big[\frac{\mathbf{x}(k)\mathbf{x}(k)^{\dagger}}{\sqrt{\sum_{k=1}^{K} |\hat{s}_1(k)|^2}}\Big] \mathbf{w}_1(k) + \mathbf{w}_2^{\dagger}(k) E\Big[\frac{\mathbf{x}(k)\mathbf{x}(k)^{\dagger}}{\sqrt{\sum_{k=1}^{K} |\hat{s}_2(k)|^2}}\Big] \mathbf{w}_2(k) \\ &= E\Big[\frac{|\hat{s}_1(k)|^2}{\sqrt{\sum_{k=1}^{K} |\hat{s}_1(k)|^2}}\Big] + E\Big[\frac{|\hat{s}_2(k)|^2}{\sqrt{\sum_{k=1}^{K} |\hat{s}_2(k)|^2}}\Big] \\ &\approx \frac{1}{K}\Big(\sqrt{\sum_{k=1}^{K} |\hat{s}_1(k)|^2} + \sqrt{\sum_{k=1}^{K} |\hat{s}_2(k)|^2}\Big) \end{aligned} \tag{3.4.5}$$

Then, when block permutation happens in frequency bin block $[k_b, k_e]$, the corresponding rows of the unmixing matrices $W(k)$ are exchanged, and the modified unmixing matrices are denoted as $\bar{W}(k)$. Because

$$\log|det(W(k))| = \log|det(\bar{W}(k))| \tag{3.4.6}$$

and R in equation (3.4.1) is a scalar constant term which is independent of the unmixing matrix. Thus, Q' is the only term changed in equation (3.4.1),

which becomes

$$Q_2' \approx \frac{1}{K}\Big(\sqrt{\sum_{1}^{k_b-1}|\hat{s}_1(k)|^2+\sum_{k_b}^{k_e}|\hat{s}_2(k)|^2+\sum_{k_e+1}^{K}|\hat{s}_1(k)|^2} + \sqrt{\sum_{1}^{k_b-1}|\hat{s}_2(k)|^2+\sum_{k_b}^{k_e}|\hat{s}_1(k)|^2+\sum_{k_e+1}^{K}|\hat{s}_2(k)|^2}\Big) \tag{3.4.7}$$

When

$$\sum_{k_b}^{k_e}|\hat{s}_1(k)|^2 = \sum_{k_b}^{k_e}|\hat{s}_2(k)|^2 \tag{3.4.8}$$

it follows that $Q_1' = Q_2'$, and therefore there is no penalty for the AuxIVA converging to a block permutation solution. To confirm the problem occurs regularly, different speech signals chosen randomly from Sawada's dataset as mentioned in the previous experiment, are positioned at a variety of different locations in a room environment to generate microphone measurements by using the image method. Then the AuxIVA method was used to separate them. It is found that approximately 29% of them suffer from the block permutation problem which justifies the need to overcome the ill-convergence.

3.4.2 Overlapped Chain Type Dependency Model for AuxIVA

For the original dependency model, it assumes same dependency between any of two frequency bins. However, for the speech signal, it is inappropriate to make the assumption that two far apart frequency bins have the same strong dependency as two neighboring frequency bins. Generally, the dependency between two neighboring frequency bins is much stronger than that of frequency bins far apart. Thus, a chain type dependency model is adopted for AuxIVA. This improved dependency model divides the whole range of frequency bins to several overlapped frequency bin blocks, which is linked as a chain [33]. As such, it can strengthen the neighborhood dependence, while weakens the dependence when frequency bins are far way. Therefore,

$g'_R(r_i)/r_i$ becomes

$$\frac{g'_R(r_i)}{r_i} = \frac{1}{\sqrt{\sum_{k_{b1}}^{k_{e1}} |\hat{s}_i(k)|^2}} + \frac{1}{\sqrt{\sum_{k_{b2}}^{k_{e2}} |\hat{s}_i(k)|^2}} + \cdots + \frac{1}{\sqrt{\sum_{k_{bl}}^{k_{el}} |\hat{s}_i(k)|^2}} \quad (3.4.9)$$

where k_{bl} and k_{el} denote the beginning and end frequency bins of the l-th frequency bin sub-block. The updates of the rows of the unmixing matrix are

$$\mathbf{w}_i(k) = [W(k)V_i(k)]^{-1}\mathbf{e}_i \qquad i = 1, \cdots, N \quad (3.4.10)$$

where $\mathbf{e}_i$ denotes the unit column vector with the i-th element unity. The advantage of this improved source dependency model will be confirmed in experimental evaluations.

3.4.3 Experimental Results

In the simulations, the speech signals used are also from Sawada's dataset, each of them is approximately 7 seconds long. The image method was used to generate the room impulse responses [50], and the size of the room is $7 \times 5 \times 3m^3$. The STFT length is 1024, and RT60 = 200ms. A 2×2 mixing case is used, for which the microphone positions are [3.48, 2.50, 1.50] and [3.52, 2.50, 1.50] respectively. The sampling frequency is 8kHz. The improved source dependency model divides the whole frequency range into three parts with 50% overlap. The separation performance is evaluated objectively by SDR and SIR [51].

An example of solving the block permutation problem is given. Two speech signals are chosen from Sawada's dataset, and placed at the positions [4.8 3.25 1.5] and [2.75 3.8 1.5], whose directions of arrival are respectively 60 and 120 degrees with reference to the center of the microphones. The experimental result shows that AuxIVA can not separate the mixtures due to the block permutation problem which can be observed by the permutation

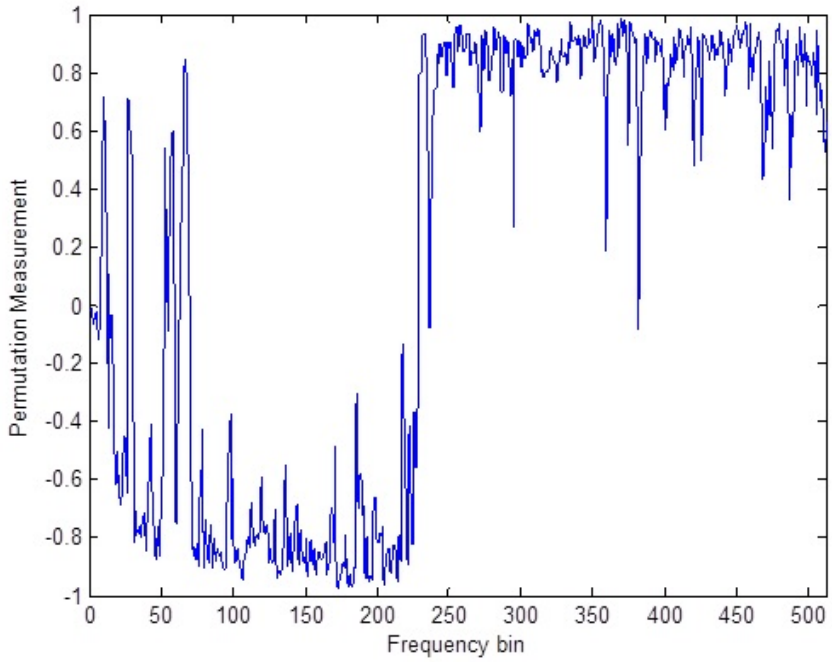


Figure 3.11. The permutation measurement showing the block permutation problem by using AuxIVA.

measurement shown in Fig. 3.11. The SDR is 1.82dB and SIR is 3.00dB, which also confirms that the mixtures are not separated well. By using the improved dependency model AuxIVA (IAuxIVA) algorithm, the block permutation problem can be solved and thereby the mixtures can be separated. The result is shown in Fig. 3.12. The objective evaluation SDR is 5.44dB and SIR is 6.54dB, which indicate that the mixtures are separated.

The separation performance comparison is also shown when there is no block permutation problem. The AuxIVA method is firstly used to separate the speech mixtures which are generated by positioning the other two source speech signals at different locations. Then the IAuxIVA is used to separate the same speech mixtures. The comparison results are shown in Table 3.3. It is clear that the proposed method can also improve the convergence speed to achieve essentially the same separation performance as AuxIVA.

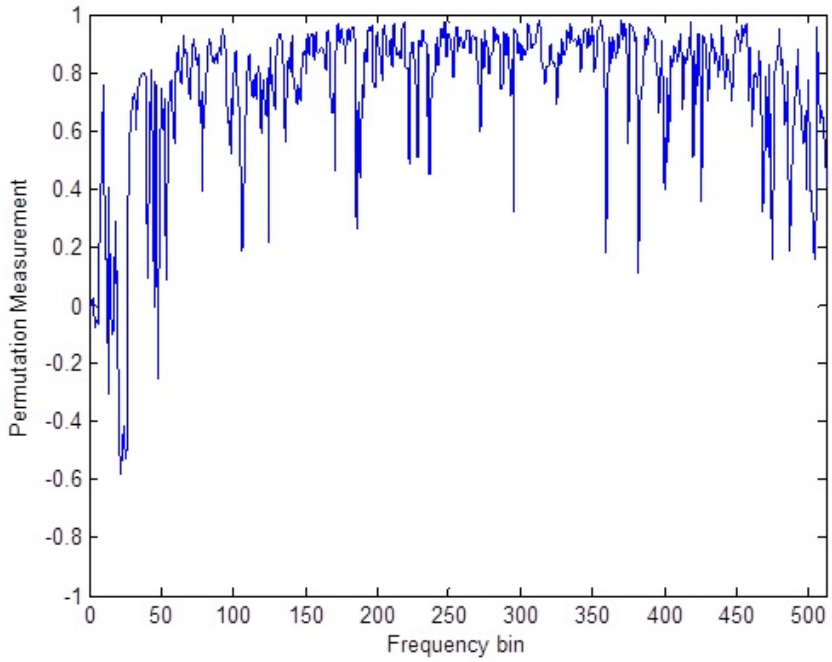


Figure 3.12. The permutation measurement without the block permutation problem by using AuxIVA with the proposed dependency model.

Table 3.3. Separation performance comparison when there is no block permutation problem.

angles	iterations		SDR(dB)		SIR(dB)	
	AuxIVA	IAuxIVA	AuxIVA	IAuxIVA	AuxIVA	IAuxIVA
30,120	32	23	20.24	20.17	22.33	22.27
60,120	34	29	17.98	18.17	19.61	19.70
30,150	34	23	20.91	20.96	22.96	22.87
60,150	38	32	19.50	19.44	21.29	21.17

3.5 Summary

In this chapter, the specific block permutation problem inherent to IVA was highlighted, and discussed by analyzing the cost function of IVA. Then, two kind of solutions were proposed. The first one exploited the phase continuity of the unmixing matrix, and the second adopted the overlapped chain type

source prior to provide improved dependency model. Both of these two schemes can be adapted to all IVA algorithms to solve the block permutation problem. In the next chapter, an informed IVA scheme will be proposed, which can also solve this problem by introducing prior information.

Chapter 4

INFORMED INDEPENDENT VECTOR ANALYSIS

4.1 Introduction

IVA is a blind source separation method; which generally implies that no prior information such as geometric positions of the sources is used to aid performance. However, as people not only use their ears to solve the cocktail party problem, but also their eyes, it is natural to exploit video information within such machine learning algorithms [52] [53] [54] [55] [56] [57] [58]. In this chapter, an informed IVA scheme is proposed, which combines the FastIVA algorithm with prior geometric information of the sources, which is obtained from video, to help the separation.

For FastIVA, although it can achieve fast convergence, sometimes it can still suffer the special block permutation problem. In this chapter, this special problem is highlighted and analytically demonstrated. It is also shown that such ill-convergence can be mitigated by setting a good initialization of the unmixing matrix, which also satisfies the second objective of this thesis.

Initialization is important for the optimization problem because it can improve convergence speed by ensuring a short cut convergence path avoiding local minimum points which yield poor separation. Source position information is important prior knowledge for setting a good initialization, and

it can be obtained by audio localization or video localization. Audio localization for a single active speaker is difficult because human speech is an intermittent signal and contains much of its energy in the low-frequency bins where spatial discrimination is imprecise [59]. Audio localization can also be affected by noise and room environment [60] [61]. Additionally, audio localization is not always effective due to the complexity in the case of multiple concurrent speakers [53]. Therefore, the accuracy of the audio localization would be degraded in a multisource real room environment with noise and reverberations, but video localization is generally robust in such an environment. On the other hand, video localization is not always effective, especially when the face of a human being is not visible to at least two cameras due to some obstacles, for example when the environment is cluttered, camera angle is wide, or illumination conditions are varying. For an audio video combined source separation method, besides the direction of arrival (DOA) information, another type of combination is using audio video coherence for separation [22] [62] [63] [64]. However, for the room environment (as used in the AV16.3 recordings [39]), it is not possible to perform lip reading due to the sources being far from the cameras. Therefore, cameras are only used to capture the locations of the speakers in this chapter. Then the positions can be used to obtain a smart initialization for the convergence of the learning algorithm. Thus, a new audio video based fast fixed-point independent vector analysis (AVIVA) method is proposed, which uses video information to initialize the algorithm.

In order to verify the advantages of AVIVA, datasets containing multiple speech and noise signals are used in its evaluation. Most speech separation evaluations have been undertaken by using artificial recordings. Few of them use real room recordings due to the practical constraints. However, in this chapter, the proposed AVIVA method is tested with real room recordings, i.e. the AV16.3 corpus [39], which not only confirms the advantages of the

proposed method, but also confirms the practical advantage of this work.

For real dataset, the separation performance evaluation becomes a problem. There is no objective evaluation method proposed to evaluate such real room recordings. Traditional evaluations are all based on prior knowledge such as the mixing filters or source signals. For instance, the performance index needs the mixing filters [12], and the signal-to-interference ratio or signal-to-distortion ratio require the original speech signals [51]. However, for a real recorded dataset, the only obtained information is the audio mixtures. Therefore, a new evaluation method is needed without requiring any other prior knowledge. In this chapter, a new evaluation method based on pitch information is employed. It detects the pitches of all the separated signals, and then calculates the pitch differences between them, and thereby provides an objective evaluation between methods. This chapter begins with the block permutation problem of the FastIVA algorithm.

4.2 Block Permutation Problem of FastIVA

The analysis of the block permutation problem is similar to that in Chapter 3. The occurrence of the block permutation problem can be understood by examining the cost function. For a 2×2 case, the cost function for FastIVA takes the form:

$$\begin{aligned} J_{FastIVA} =& E\Big[F(\sum_k |\hat{s}_1(k)|^2)\Big] - \sum_k \lambda_1(k)\Big(\mathbf{w}_1^\dagger(k)\mathbf{w}_1(k) - 1\Big) \\ &+ E\Big[F(\sum_k |\hat{s}_2(k)|^2)\Big] - \sum_k \lambda_2(k)\Big(\mathbf{w}_2^\dagger(k)\mathbf{w}_2(k) - 1\Big) \end{aligned} \tag{4.2.1}$$

as in [29], where the Lagrange multiplier $\lambda_i^{(k)}$ is:

$$\lambda_i^{(k)} = E\Big[|\hat{s}_i^{(k)}|^2 F'(\sum_k |\hat{s}_i^{(k)}|^2)\Big] \tag{4.2.2}$$

The original FastIVA adopts the multivariate Laplacian distribution as the source prior. The corresponding nonlinear function is:

$$F(z) = \sqrt{z} \tag{4.2.3}$$

Thus, the cost function becomes:

$$\begin{aligned} J1_{FastIVA} =& E\Big[\sqrt{\sum_k |\hat{s}_1(k)|^2}\Big] - \sum_k E\Big[\frac{|\hat{s}_1(k)|^2}{2\sqrt{\sum_k |\hat{s}_1(k)|^2}}\Big]\Big(\mathbf{w}_1^\dagger(k)\mathbf{w}_1(k) - 1\Big) \\ &+ E\Big[\sqrt{\sum_k |\hat{s}_2(k)|^2}\Big] - \sum_k E\Big[\frac{|\hat{s}_2(k)|^2}{2\sqrt{\sum_k |\hat{s}_2(k)|^2}}\Big]\Big(\mathbf{w}_2^\dagger(k)\mathbf{w}_2(k) - 1\Big) \end{aligned} \tag{4.2.4}$$

If the block permutation problem happens, there is a frequency bin block over the range $[k_b, k_e]$ with a different separation alignment from other frequency bins, and $\mathbf{w}_1$, $\mathbf{w}_2$ are exchanged. Then, the cost function becomes:

$$
\begin{aligned}
&J2_{FastIVA} = \\
&E\Big[\sqrt{\sum_{k=1}^{k_b-1}|\hat{s}_1(k)|^2+\sum_{k=k_b}^{k_e}|\hat{s}_2(k)|^2+\sum_{k=k_e+1}^{K}|\hat{s}_1(k)|^2}\Big] \\
&+E\Big[\sqrt{\sum_{k=1}^{k_b-1}|\hat{s}_2(k)|^2+\sum_{k=k_b}^{k_e}|\hat{s}_1(k)|^2+\sum_{k=k_e+1}^{K}|\hat{s}_2(k)|^2}\Big] \\
&-\sum_{k=1}^{k_b-1}E\Big[\frac{|\hat{s}_1(k)|^2}{2\sqrt{\sum_{k=1}^{k_b-1}|\hat{s}_1(k)|^2+\sum_{k=k_b}^{k_e}|\hat{s}_2(k)|^2+\sum_{k=k_e+1}^{K}|\hat{s}_1(k)|^2}}\Big]\Big(\mathbf{w}_1^{\dagger}(k)\mathbf{w}_1(k)-1\Big) \\
&-\sum_{k=k_b}^{k_e}E\Big[\frac{|\hat{s}_2(k)|^2}{2\sqrt{\sum_{k=1}^{k_b-1}|\hat{s}_1(k)|^2+\sum_{k=k_b}^{k_e}|\hat{s}_2(k)|^2+\sum_{k=k_e+1}^{K}|\hat{s}_1(k)|^2}}\Big]\Big(\mathbf{w}_2^{\dagger}(k)\mathbf{w}_2(k)-1\Big) \\
&-\sum_{k=k_e+1}^{K}E\Big[\frac{|\hat{s}_1(k)|^2}{2\sqrt{\sum_{k=1}^{k_b-1}|\hat{s}_1(k)|^2+\sum_{k=k_b}^{k_e}|\hat{s}_2(k)|^2+\sum_{k=k_e+1}^{K}|\hat{s}_1(k)|^2}}\Big]\Big(\mathbf{w}_1^{\dagger}(k)\mathbf{w}_1(k)-1\Big) \\
&-\sum_{k=1}^{k_b-1}E\Big[\frac{|\hat{s}_2(k)|^2}{2\sqrt{\sum_{k=1}^{k_b-1}|\hat{s}_2(k)|^2+\sum_{k=k_b}^{k_e}|\hat{s}_1(k)|^2+\sum_{k=k_e+1}^{K}|\hat{s}_2(k)|^2}}\Big]\Big(\mathbf{w}_2^{\dagger}(k)\mathbf{w}_2(k)-1\Big) \\
&-\sum_{k=k_b}^{k_e}E\Big[\frac{|\hat{s}_1(k)|^2}{2\sqrt{\sum_{k=1}^{k_b-1}|\hat{s}_2(k)|^2+\sum_{k=k_b}^{k_e}|\hat{s}_1(k)|^2+\sum_{k=k_e+1}^{K}|\hat{s}_2(k)|^2}}\Big]\Big(\mathbf{w}_1^{\dagger}(k)\mathbf{w}_1(k)-1\Big) \\
&-\sum_{k=k_e+1}^{K}E\Big[\frac{|\hat{s}_2(k)|^2}{2\sqrt{\sum_{k=1}^{k_b-1}|\hat{s}_2(k)|^2+\sum_{k=k_b}^{k_e}|\hat{s}_1(k)|^2+\sum_{k=k_e+1}^{K}|\hat{s}_2(k)|^2}}\Big]\Big(\mathbf{w}_2^{\dagger}(k)\mathbf{w}_2(k)-1\Big)
\end{aligned}
\tag{4.2.5}
$$

It is evident that when

$$
\sum_{k=k_b}^{k_e}|\hat{s}_1(k)|^2=\sum_{k=k_b}^{k_e}|\hat{s}_2(k)|^2 \tag{4.2.6}
$$

is satisfied, the cost function has the same value, i.e. $J1_{FastIVA}=J2_{FastIVA}$. This indicates that there is no penalty for the FastIVA converging to a block permutation solution, which is also a global minimum with the same value

as the correct solution. For the case where there are more sources, a similar analysis can also be used to confirm that the block permutation can happen.

4.3 Audio Video Based FastIVA

Based on the analysis and discussion in the above section, it is evident that it is necessary to set a proper initialization for the FastIVA algorithm to mitigate the block permutation problem. Moreover, a proper initialization can also achieve faster convergence and better performance, which is common for any optimization problem. Additionally, such a video localization based algorithm can improve the separation performance especially when there is background noise and a highly reverberant room environment, because audio localization can be seriously affected by such noise and reverberation [60]. The system configuration is shown in Fig. 4.1, further details of the processing can be found in [6] [65].

Firstly, video localization based on face and head detection is used to obtain the visual location of each speaker which is approximated after processing the 2-D image information and obtained from at least two synchronized colour video cameras through calibration parameters [66] and an optimization method [67].

After estimating the 3-D position of each speaker i, the elevation (θ_i) and azimuth (ϕ_i) angles of arrival to the center of the microphone array are calculated from

$$R_i = \sqrt{(u_{x_i} - u'_{x_c})^2 + (u_{y_i} - u'_{y_c})^2 + (u_{z_i} - u'_{z_c})^2} \tag{4.3.1}$$

$$\theta_i = tan^{-1}(\frac{u_{y_i} - u'_{y_c}}{u_{x_i} - u'_{x_c}}) \tag{4.3.2}$$

$$\phi_i = sin^{-1}(\frac{u_{y_i} - u'_{y_c}}{R_i sin(\theta_i)}) \tag{4.3.3}$$

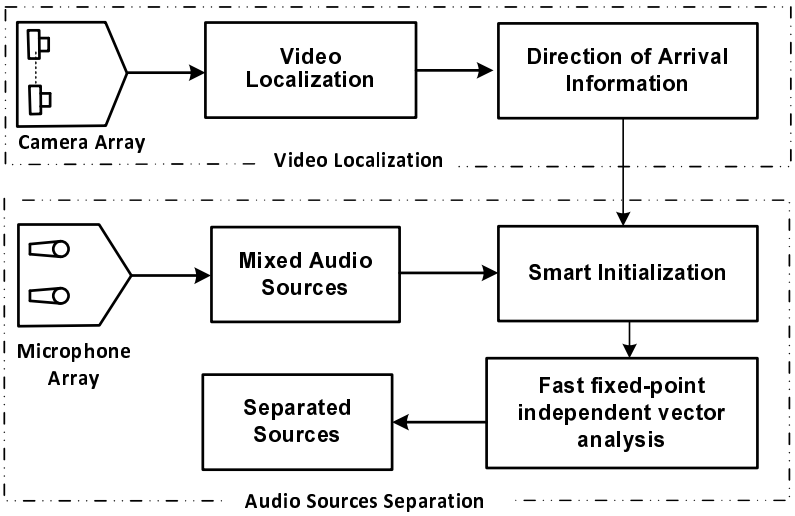


Figure 4.1. Block diagram of the audio video based fast fixed-point independent vector analysis. Video localization is based on face and head detection. The visual location of each speaker is approximated after processing the 2-D image information and obtained from at least two synchronized colour video cameras through calibration parameters and an optimization method. The position of the microphone array and the output of the visual localizer are used to calculate the direction of arrival information of each speaker. Based on this information, a smart initialization is set for the FastIVA algorithm.

where u_{x_i}, u_{y_i} and u_{z_i} are the 3-D positions of the speaker i, while u'_{x_c}, u'_{y_c} and u'_{z_c} are Cartesian coordinates of the center of the microphone array.

Then the mixing matrix can be calculated under the plane wave propagation assumption by using the DOA information.

$$H(k) = [\mathbf{h}^{(k)}(\theta_1, \phi_1) \cdots \mathbf{h}^{(k)}(\theta_n, \phi_n)] \tag{4.3.4}$$

where

$$\mathbf{h}^{(k)}(\theta_i, \phi_i) = \begin{bmatrix} e^{-j\kappa\left(sin(\theta_i)cos(\phi_i)u'_{x_1} + sin(\theta_i)sin(\phi_i)u'_{y_1} + cos(\theta_i)u'_{z_1}\right)} \\ \vdots \\ e^{-j\kappa\left(sin(\theta_i)cos(\phi_i)u'_{x_m} + sin(\theta_i)sin(\phi_i)u'_{y_m} + cos(\theta_i)u'_{z_m}\right)} \end{bmatrix} \tag{4.3.5}$$

and $\kappa = k/c$ where c is the speed of sound in air at room temperature. The

coordinates u'_{x_i}, u'_{y_i} and u'_{z_i} are the 3-D positions of the i-th microphone.

Thus, the initialization of the unmixing matrix can be obtained by following the approach in [21]

$$W^{\dagger}(k) = Q_w(k)H(k) \tag{4.3.6}$$

where Q_w is the whitening matrix. The above estimation is biased because it only takes the DOA information to construct the mixing matrix. However, this biased estimation can be used as the initialization of the unmixing matrix of FastIVA rather than an identity matrix or random matrix. The real room recordings will be used to test this proposed method, and an evaluation criterion for real room recording will be presented in the following section.

4.4 Pitch Based Evaluation For Real Recordings

In this chapter, the real datasets with multiple signals are used to test the algorithm. Thus how to evaluate the separation performance becomes an issue. For real room recordings, the only measurements obtained are the mixed signals captured by the microphone array. It is impossible to access either the mixing matrix or the pure source signals. Thus, it is impossible to evaluate the separation performance by traditional methods, such as performance index [12] which is based on the prior knowledge of the mixing matrix, or the SIR or SDR [51] which require prior knowledge about the source signals. It is a challenging problem to evaluate objectively real recording separation performance. People can listen to the separated speech signals, but it is just a form of subjective evaluation, such as mean opinion score (MOS). In order to evaluate the results objectively, the features of the separated signals should be used. Pitch information is one of the features which can help to evaluate the separation performance, because different

speech sections at different time slots have different pitches [68] provided that the original sources do not have substantially overlapping pitch characteristics. The sawtooth waveform inspired pitch estimator (SWIPE) method is adopted [69], which has better performance compared with traditional pitch estimators [70] [71] [72] [73].

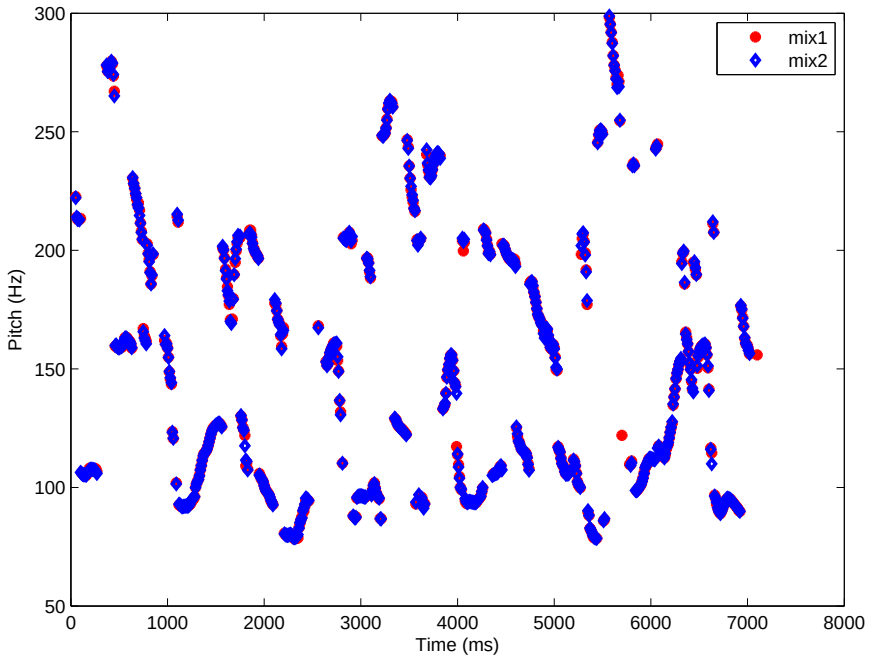


Figure 4.2. The pitch tracks of two mixture signals.

Fig. 4.2 shows that the pitches of the mixed signals are still mixed, while the pitches of the source signals in Fig. 4.3 are well separated. It is obvious that good separated pitches can indicate good separation performance provided that the original sources do not have substantially overlapping pitch characteristics. In order to evaluate performance objectively, the pitch differences are calculated:

$$p_{diff}(t) = \sqrt{\sum_{i \neq j} (p_i(t) - p_j(t))^2} \quad i, j = 1, \cdots, m \text{ and } t = 1, \cdots, T_L \quad (4.4.1)$$

where T_L is the number of time slots. Then a threshold p_{thr} is set, if the

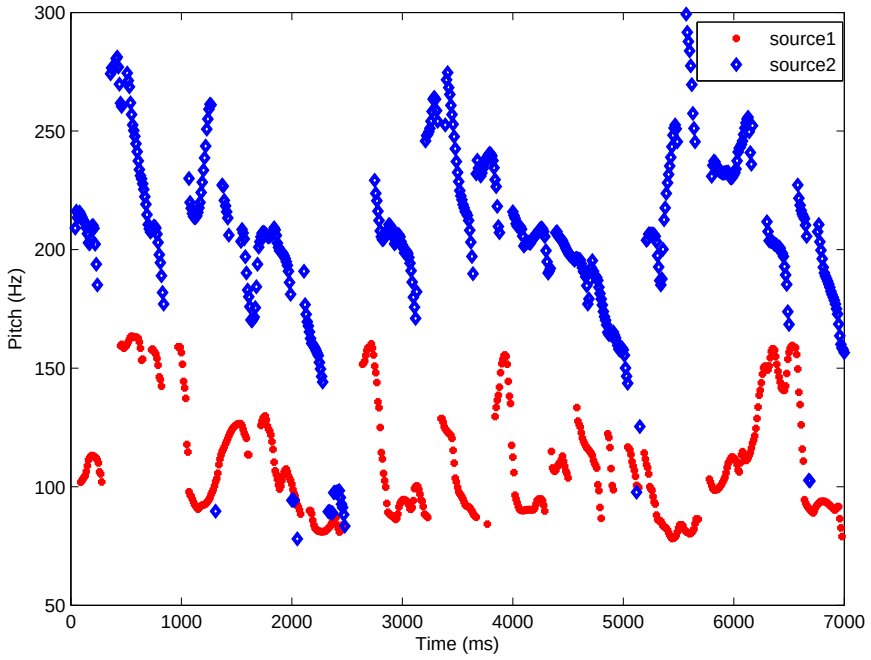


Figure 4.3. The pitch tracks of two separated signals.

pitch difference is greater than the threshold at a certain time slot, it can be considered that the mixed signals are separated at that time slot and set the separation status equal to 1, otherwise 0, as defined by

$$sep\_status(t) = \begin{cases} 1 & if \quad p_{diff}(t) > p_{thr} \\ 0 & otherwise \end{cases} \tag{4.4.2}$$

Finally, a separation rate can be calculated to evaluate the separation performance.

$$sep\_rate = \frac{\sum_t sep\_status(t)}{T_L} \tag{4.4.3}$$

The separation performance improves as the separation rate increases. It is highlighted here that it can not evaluate the absolute quality of the separated speech signal, but it can be used for comparing the separation performance when using different separation methods.

4.5 Experiments and Results

In this section, three different kinds of experimental results are shown by using different multisource datasets to show the advantage of the proposed AVIVA algorithm. The first experiment will show that the proposed AVIVA algorithm can successfully avoid the block permutation problem. The second experiment will demonstrate the advantage of AVIVA in the aspect of convergence speed and separation performance improvement in a noisy environment and in a highly reverberant environment. The positions of the source speech signal are assumed known in these two experiments, and the initialization is based on these positions. The last experiment shows the proposed method used in a real application by using the real multisource dataset. The 3-D video localizer is used to capture the source positions.

4.5.1 Experimental Demonstration of the Block Permutation Problem

For the real room recordings, it is impossible to obtain the mixing filters, therefore the block permutation can not be observed visually. In the first simulation, it assumes that the source signals and mixing filters are known to experimentally demonstrate the block permutation problem. The speech signals are also from Sawada's dataset as previous chapters. Each speech signal is approximately 7 seconds long. The image method is used to generate the room impulse responses [50], and the size of the room is [7,5,3], which represents the length, the width and the height respectively, and the measure unit is meter. The STFT length is 1024 samples, and reverberation time RT60 = 200ms. A 2×2 mixing case is used, for which the microphone positions are [3.48, 2.50, 1.50] and [3.52, 2.50, 1.50] respectively in Cartesian coordinates. The sampling frequency is 8kHz. The separation performance is evaluated objectively by performance index (PI) [12], the signal-to-distortion

ratio (SDR) and signal-to-interference ratio (SIR) [51].

Two speech signals are chosen, and placed at positions [4.8, 3.25, 1.5] and [2.75, 3.8, 1.5], whose azimuth angles are respectively 60 and -30 degrees with reference to the normal to the microphones. Then the FastIVA method is used to separate the mixtures. The result is shown in Fig. 4.4. The frequency bin range [0,512] corresponds to [0, 4000]Hz as the sampling frequency is 8kHz. The upper part of the figure is the performance index, the closer it is to zero, the better the separation performance. And the bottom part is the permutation measurement. It is clear that there is a block permutation problem. Thus the mixtures can not be properly separated by FastIVA. The objective measurements are shown in Table 4.1. SDR is 2.81dB and SIR is 4.12dB, which also confirms that it is still mixed.

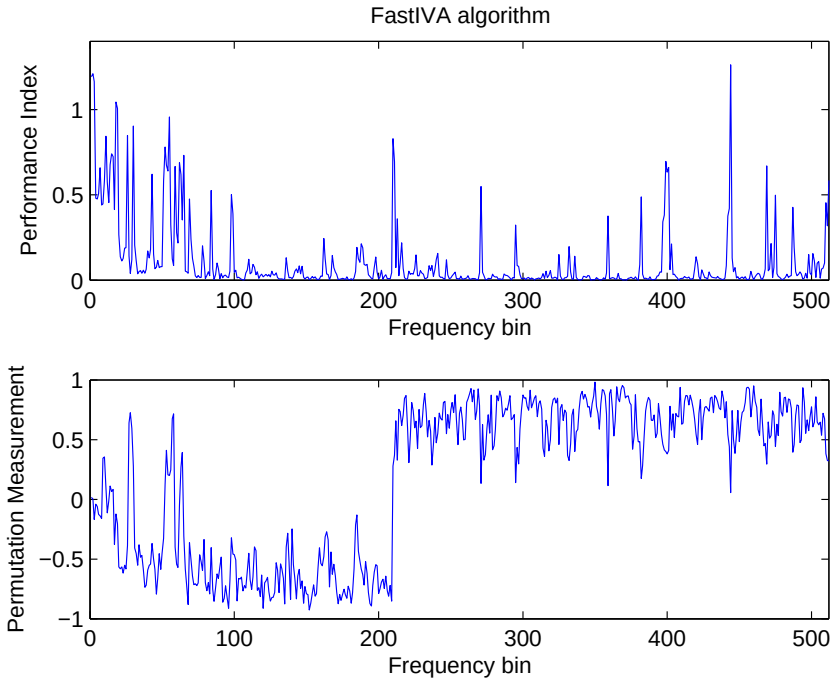


Figure 4.4. Separation performance of FastIVA. The upper part is the performance index figure, and the bottom part is the permutation measurement figure.

Then, the proposed AVIVA method is used to separate the mixtures. The result is shown in Fig. 4.5. It confirms that the block permutation

problem has been solved. As such, the performance is improved, which can be verified by the performance index figure in Fig. 4.5. Moreover, the objective measurement SDR is 6.11dB and SIR is 7.35dB, which confirms the mixtures are better separated.

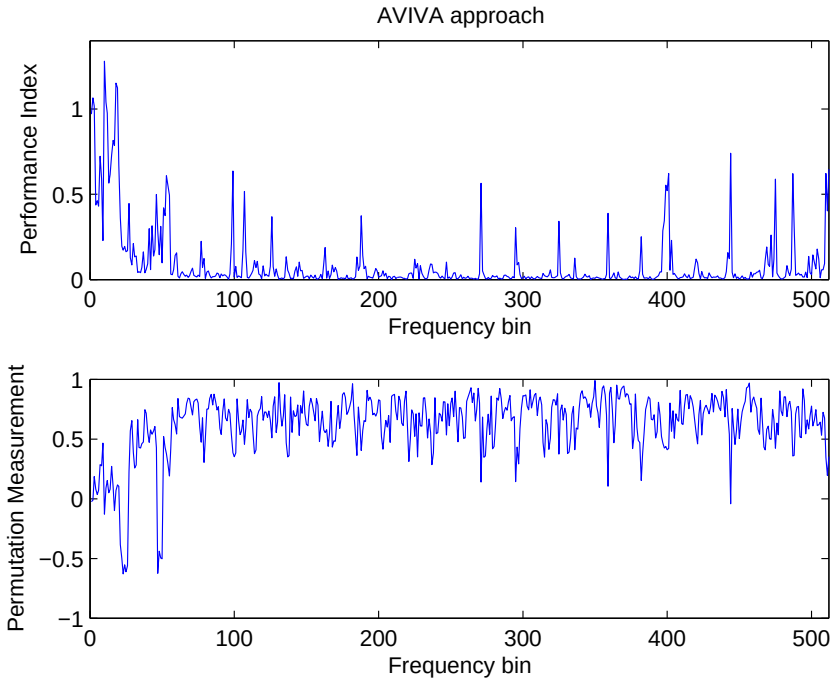


Figure 4.5. Separation performance of AVIVA. The upper part is the performance index figure, and the bottom part is the permutation measurement figure.

Then a 3×3 mixing case is used to confirm the block permutation problem, for which the microphone positions are [3.46, 2.50, 1.50], [3.50, 2.50, 1.50] and [3.54, 2.50, 1.50] respectively. Three speech signals are chosen, and placed at positions [4.80, 3.25, 1.5], [3.50, 4.00, 1.50] and [2.75, 3.8, 1.5], whose azimuth angles are respectively 60, 0 and -30 degrees with reference to the normal to the microphones. For the 3×3 case, it is hard to use the permutation measurement directly, and the permutation measurement of each 2×2 sub matrix in the 3×3 matrix needs to be calculated. The FastIVA algorithm is first used to separate the mixtures, and Fig. 4.6 shows one permutation measurement which has the block permutation problem

between source 1 and source 3. For the frequencies above frequency bin 220, the mean value of the permutation measurement is negative, whereas for the other frequencies, the mean value is positive, which shows the block permutation problem. And the objective results shown in Table 4.1 confirm the bad separation, the SDR is 0.12dB and SIR is 1.06dB.

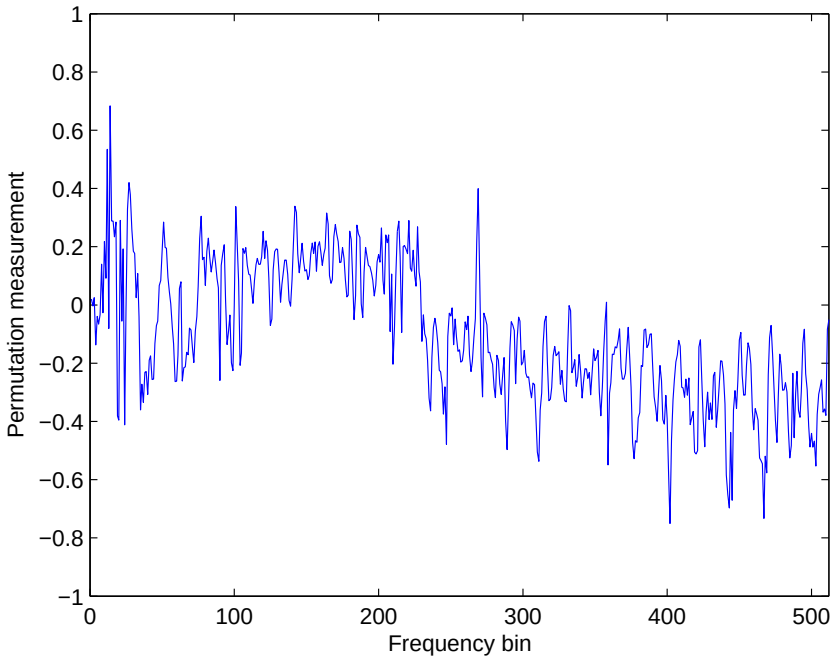


Figure 4.6. One permutation measurement of the separation result for three sources when using FastIVA.

Then the AVIVA approach is used to separate the mixtures. The permutation measurement is shown in Fig. 4.7. Combining with the objective measurements SDR and SIR, which are 6.63dB and 8.42dB respectively, it confirms that the block permutation problem has been solved.

These simulations have confirmed that the block permutation problem can happen, and the experimental results verify that the AVIVA algorithm can avoid the block permutation problem successfully by using a proper initialization.

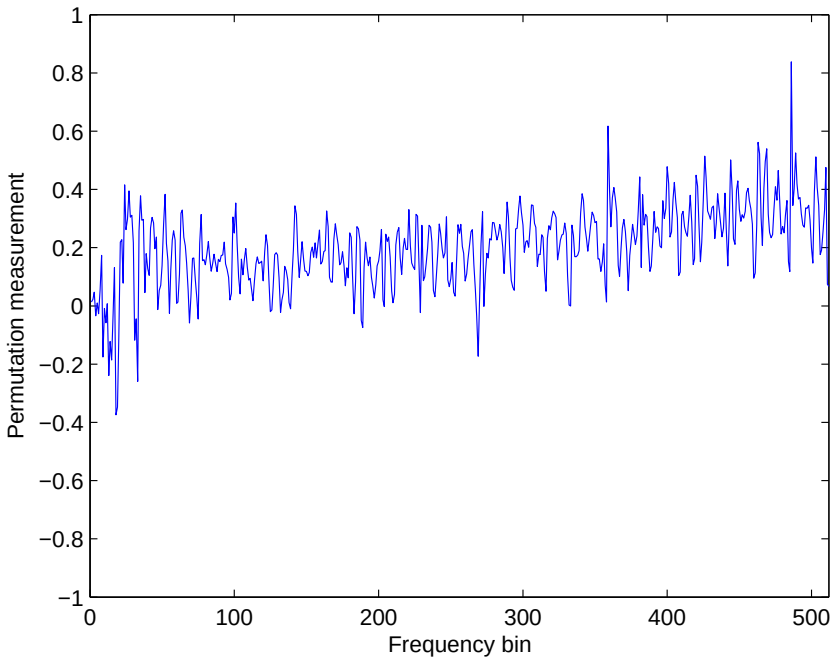


Figure 4.7. One permutation measurement of the separation result for three sources when using AVIVA.

Table 4.1. Separation performance comparison when block permutation problem happens

	FastIVA SDR(dB)	FastIVA SIR(dB)	AVIVA SDR(dB)	AVIVA SIR(dB)
two sources	2.81	4.12	6.11	7.35
three sources	0.12	1.06	6.63	8.42

4.5.2 Experiments in Noisy and Reverberant Room Environment

In the second simulation, the separation performance of the AVIVA approach in a noisy environment is shown for a multisource case. Moreover, it also shows that the AVIVA approach can achieve better separation performance in a highly reverberant environment. The positions of the sources and microphones are assumed known to generate different reverberant environments by changing the absorption coefficients of the image method. A 2×2 mix-

ing case is used, for which the microphone positions are [3.48, 2.50, 1.50] and [3.52, 2.50, 1.50] respectively. The noise is assumed to be Gaussian distributed and its standard deviation is selected to be 2.5% of the maximum magnitude of the speech signal. Different speech signals are chosen from the TIMIT dataset [38]. This simulation is used to show the AVIVA algorithm is suitable for different kinds of mixtures, and can achieve a better separation performance with faster convergence in a noisy environment. All the experiment parameters are the same as the 2×2 case in experiment 4.5.1. The separation performance is also evaluated by SDR and SIR.

First of all, it shows that the block permutation problem still can happen when using the TIMIT dataset in a noisy environment. Two speech signals are chosen from the TIMIT dataset, placed at positions [4.8, 3.25, 1.5] and [2.75, 3.8, 1.5], whose azimuth angles are respectively 60 and -30 degrees with reference to the normal to the microphones. FastIVA and AVIVA are used to separate the mixtures respectively. The results are shown in Fig. 4.8 and Fig. 4.9. When using the FastIVA algorithm, the objective separation performance measures SDR and SIR are 0.19dB and 2.45dB, which confirms the limited separation performance. When using the AVIVA approach, the block permutation problem is solved and the SDR and SIR are 6.43dB and 14.90dB which indicate a good separation performance.

Two different speech signals are chosen randomly from the TIMIT dataset and these are convolved into two mixtures. Then FastIVA and AVIVA are used to separate the mixtures respectively. Next, the source positions are changed to repeat the simulation. For every pair of speech signals, three different azimuth angles for the sources relative to the normal to the microphone array are set for testing, these angles are selected from 30, 45, 60 and -30 degrees. After that, another pair of speech signals is chosen to repeat the above simulations. In total, five different pairs of speech signals are used (including combinations with one male speech signal and one female speech

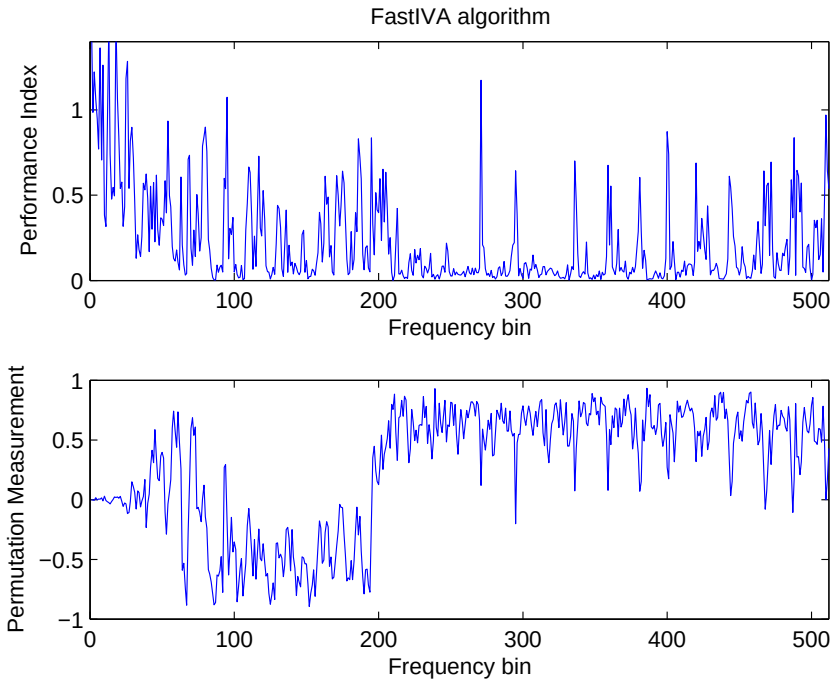


Figure 4.8. Separation performance of FastIVA in the noisy environment. The upper part is the performance index figure, and the bottom part is the permutation measurement figure.

signal, combinations with two male speech signals and combinations with two female speech signals), and the simulation are repeated for 15 times at different positions. Table 4.2 shows the average separation performance for each pair of speech signals. The convergence advantage of the AVIVA approach is also considered.

Table 4.2. Separation performance comparison in noisy environment.

	FastIVA			AVIVA		
	iter	SDR(dB)	SIR(dB)	iter	SDR(dB)	SIR(dB)
mixtures 1	30	3.68	7.00	21	6.34	10.70
mixtures 2	28	6.60	10.68	25	7.01	11.47
mixtures 3	23	7.51	14.16	14	7.61	14.41
mixtures 4	26	6.33	11.27	11	6.76	12.79
mixtures 5	22	6.24	12.38	15	6.45	13.30

The results shown in Table 4.2 confirm the advantage of the proposed

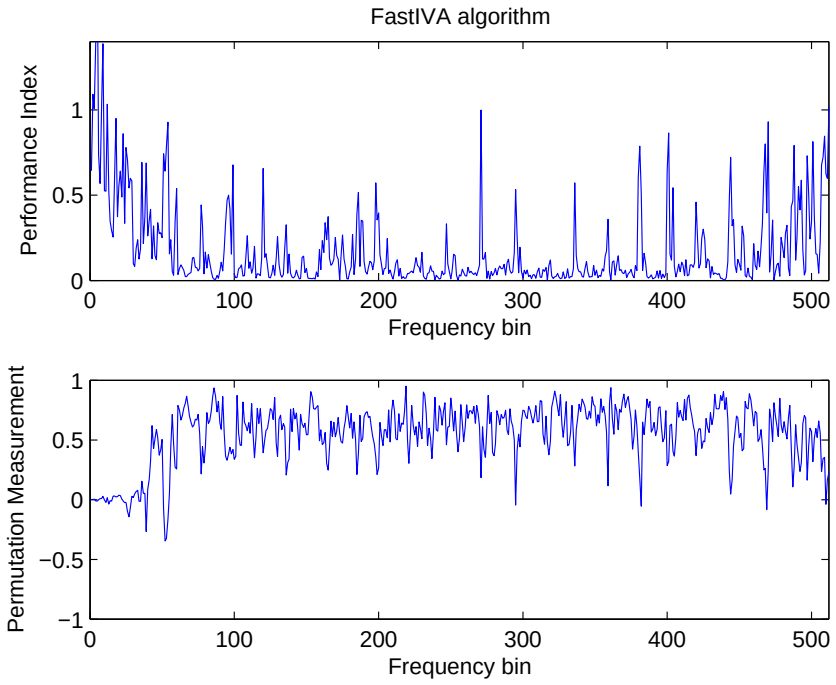


Figure 4.9. Separation performance of AVIVA in the noisy environment. The upper part is the performance index figure, and the bottom part is the permutation measurement figure.

AVIVA algorithm in that it can achieve a faster convergence and better separation performance in a noisy environment. The FastIVA is already a fast form algorithm, however, the AVIVA can improve the convergence speed approximately by 60%. Meanwhile, the separation performances are also improved generally. Comparing with the FastIVA algorithm, the average further improvement in SDR is approximately 0.75dB, and the average further improvement in SIR is approximately 1.4dB.

Then, a pair of speech signals are chosen randomly from the TIMIT dataset and placed at the positions whose azimuth angles are 60 and -30 relative to the normal to the microphone array. The room reverberation RT60 changed from 200ms to 700ms to test the separation ability of FastIVA and AVIVA algorithms in a highly reverberant environment. The results are shown in Fig. 4.10 and Fig. 4.11. The experimental results indicate that the AVIVA approach can consistently achieve better separation performance

than FastIVA algorithm in different reverberant environments. Comparing with the FastIVA algorithm, the average further improvements in SDR and SIR are 2.4dB and 2.9dB respectively.

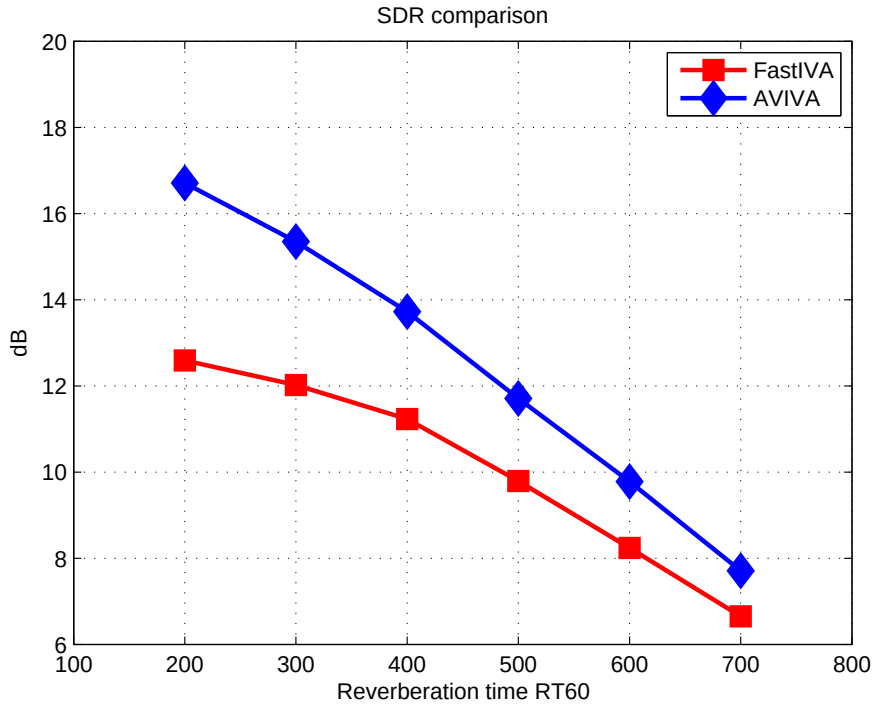


Figure 4.10. SDR comparison in different reverberant environments.

4.5.3 Experiments by Using the Real Room Recordings

In the last simulation, the real room recordings AV16.3 corpus are used to test the proposed AVIVA algorithm [39]. "16.3" stands for 16 microphones and 3 cameras, recorded in a fully synchronized manner. The "seq37-3p-0001" recording is used to perform the experiment, which contains three speakers. Fig. 4.12 and Fig. 4.13 show the room environment, the positions of microphone arrays and the positions of the three speakers. There are two microphone arrays, three microphones (mic3, mic5 and mic7) from microphone array 1 are chosen which is in the red circle. The audio sampling frequency of the recording is 16kHz. The RT60 is approximately 700ms,

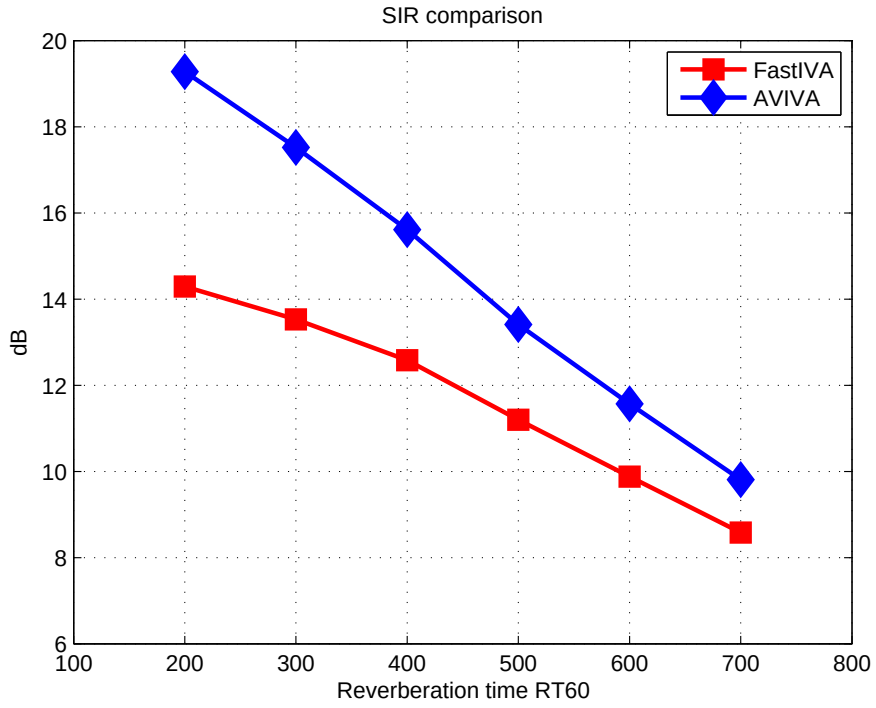


Figure 4.11. SIR comparisons in different reverberant environments.

which means that it is a highly reverberant environment and the accuracy of audio localization will be seriously affected. For this simulation, the proposed pitch based evaluation method is used, and the pitch threshold is set to 5, which has been found empirically.

The recorded speech is extracted from 200s to 205s, during which three speakers are speaking simultaneously. Then, the positions of the speakers are obtained by using the video information. After that, FastIVA and AVIVA are applied respectively. The experimental results are shown in Fig. 4.14, Fig. 4.15, Fig. 4.16 and Table 4.3. The convergence advantage of the AVIVA approach is considered.

Fig. 4.14 shows that the pitches of the mixed signals are all mixed. Fig. 4.15 is the separation result by using FastIVA. Although the pitches are separated to some extent, there are still many mixed pitches. Fig. 4.16 is the separation results by using AVIVA. It shows that the pitches are better

Figure 4.12. Room environment for one of the AV16.3 corpus recordings. A single video frame from camera 1.

Table 4.3. Separation performance for the real room recordings.

Time slot	FastIVA iterations	FastIVA separation rate	AVIVA iterations	AVIVA separation rate
200s-205s	70	0.03	49	0.14
220s-225s	192	0.05	54	0.06
240s-245s	58	0.20	56	0.23
200s-220s	77	0.14	71	0.16

separated compared with the result of FastIVA. The objective evaluation separation rate is shown in Table 4.3. Then different time slots are chosen to repeat the simulation, and the results are also shown in Table 4.3. It is highlighted that all the three speakers in this experiment are all male, and the proposed pitch based evaluation method still works well. The experimental results indicate that the proposed AVIVA algorithm can be used in a real multisource room environment successfully with faster convergence and better separation performance than FastIVA.

Figure 4.13. Room environment for one of the AV16.3 corpus recordings. A single video frame from camera 2.

4.6 Summary

In this chapter, firstly, the block permutation problem of FastIVA was analyzed. Then an audio video based FastIVA algorithm was proposed, which can use the geometric information obtained from video to set a proper initialization. The proposed algorithm can avoid the block permutation problem of independent vector analysis methods. Moreover, it can also achieve a faster and better separation performance in a noisy environment and a highly reverberant environment when compared with FastIVA. Meanwhile, a pitch based evaluation method was also proposed for the real multisource dataset, which doesn't need any prior information such as the mixing filters and source signals. The experimental results confirmed the advantages of the proposed AVIVA algorithm, and also verified that the proposed pitch based evaluation method can be used for comparing the separation performance. In the next chapter, a new source prior will be proposed to improve the separation performance of IVA algorithms.

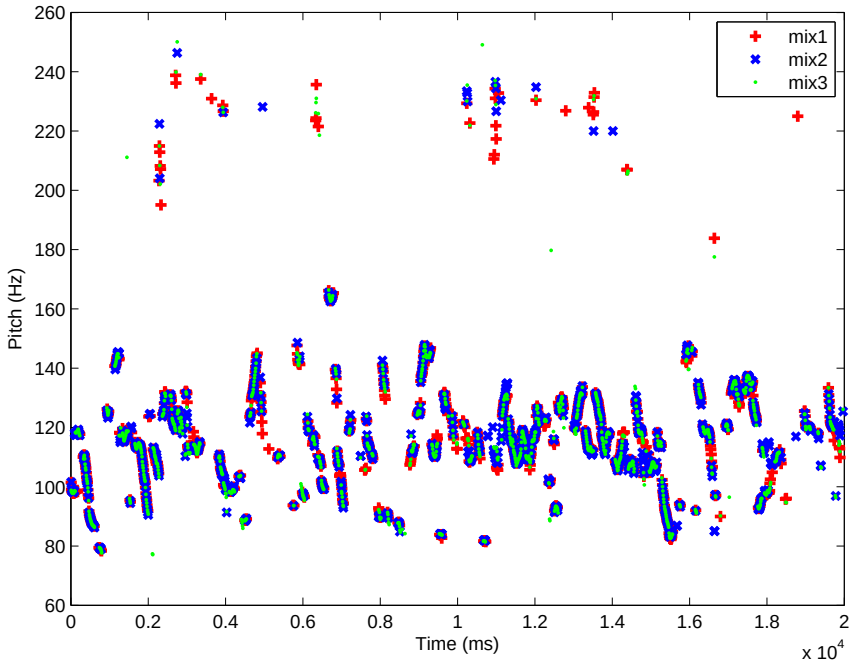


Figure 4.14. The pitch tracks of the mixed signals.

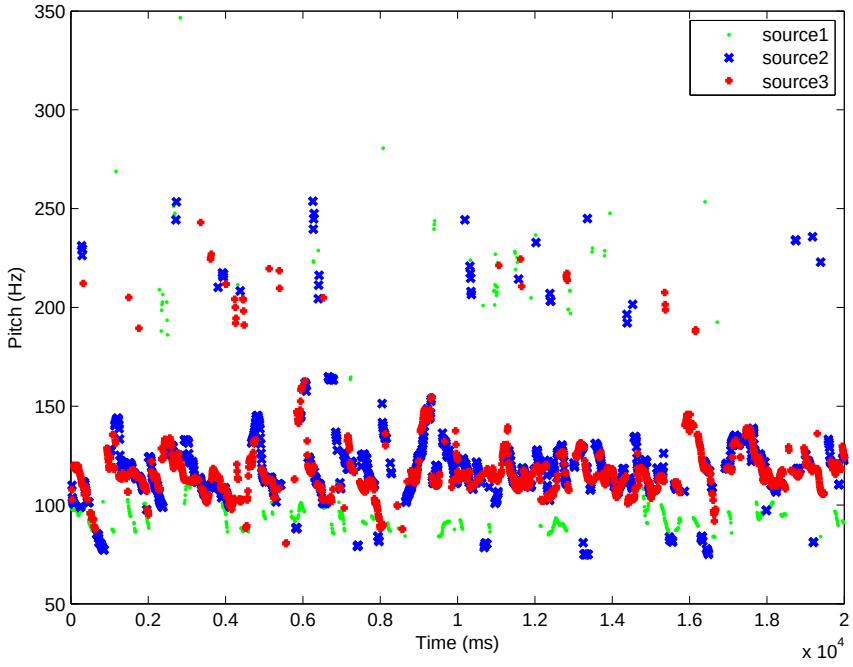


Figure 4.15. The pitch tracks of the separated signals by FastIVA.

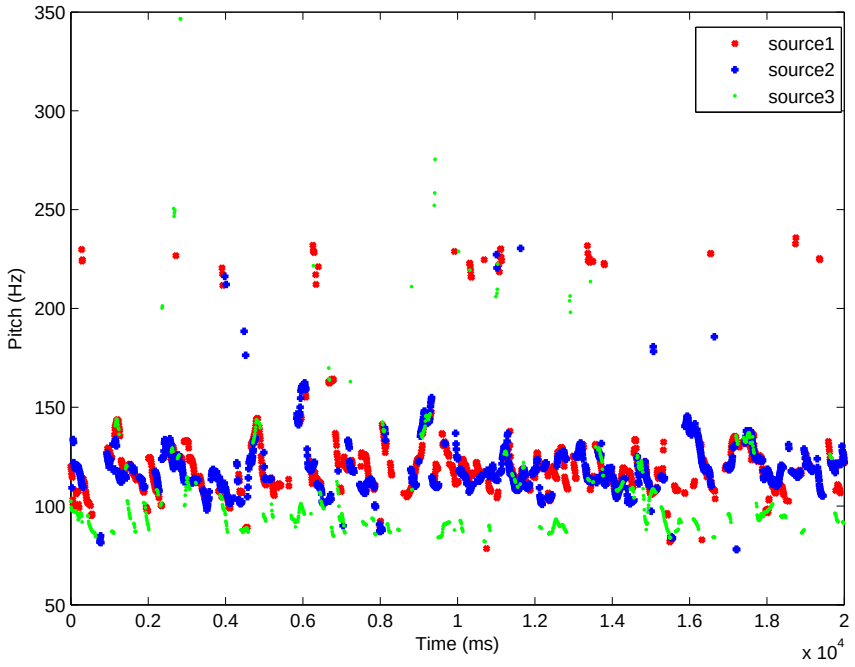


Figure 4.16. The pitch tracks of the separated signals by AVIVA.

Chapter 5

IVA WITH MULTIVARIATE GENERALIZED GAUSSIAN SOURCE PRIOR

5.1 Introduction

The core idea of IVA algorithms applied to frequency domain BSS is preserving inter-frequency dependencies for individual sources. The nature of the score function used in the algorithm derivation is crucial in this process [27]. The nonlinear score function is derived from the source prior, therefore an appropriate source prior is needed. For the original IVA algorithms, a spherically symmetric distribution is adopted as the source prior, which implies the dependencies between different frequency bins are all the same. However, the dependencies between frequency bins should be variable. In order to describe the dependency structure better, a chain type overlapped source prior has been proposed [33]. More recently, a harmonic structure dependency model has been proposed [34]. Another possible source prior is the Gaussian mixture model, whose advantage is that it enables the IVA algorithms to separate a wider class of signals [35] [36]. However, for all of these source priors, the covariance matrix of each source vector is an identity matrix because the Fourier basis is an orthogonal basis. This implies that there

is no correlation between different frequency bins. Moreover, the higher order correlations between the elements of the source vectors are ignored when separating the mixtures with IVA algorithms. Recently, an IVA algorithm based upon a multivariate Gaussian source prior has been proposed to introduce the second order correlations in the time domain [37]. However, it is used in applications which have large second order correlations such as functional magnetic resonance imaging studies. For the frequency domain IVA algorithms, other correlation information should be exploited.

In [74], a family of l_p-norm-invariant sparse probability density functions is used as the source prior; then the separation performance of NG-IVA algorithms is compared. The experimental results indicate that the spherical symmetry pdf, i.e. $p = 2$, is suitable for modeling speech. The sparseness parameter is also discussed, and it is claimed that the best separation can be obtained when the sparseness parameter is around 7.

In this chapter, in order to satisfy the third objective of this thesis, a particular multivariate generalized Gaussian distribution is adopted as the source prior, which also belongs to the family of l_2-norm-invariant sparse probability density functions, and the sparseness parameter is chosen to be $\frac{3}{2}$. This proposed source prior has heavier tails compared with the original multivariate Laplacian distribution. It can preserve the dependency across different frequency bins in a similar way as when the original multivariate Laplace distribution is used to derive the IVA algorithm. Moreover, the nonlinear score functions which are derived based on the proposed source prior additionally contain fourth order relationships between the elements of each source vector, thus they contain more information describing the dependency structure which can thereby better preserve the inter-frequency dependency to achieve an improved separation performance, as suggested by Hyvärinen [75]. The experimental results show that using the new source prior can consistently achieve improved separation performance. The IVA

algorithm with the source prior whose sparseness parameter is around 7 is found not to be robust in that it not always increases the separation performance.

5.2 Multivariate Generalized Gaussian Source Prior

In order to improve the separation performance, a new multivariate source prior which can better retain the dependency between different frequency bins is needed. The multivariate generalized Gaussian distribution is found to be suitable for being the source prior for IVA.

The univariate Laplace distribution is a special case of the univariate generalized Gaussian distribution which takes the form

$$p(s_i) \propto exp\Big(- \Big(\frac{|s_i - \mu|}{\alpha} \Big)^{\beta} \Big) \tag{5.2.1}$$

where $\alpha, \beta \in \mathcal{R}^+$ and are respectively scale and shape parameters. If α is chosen properly, it becomes the Gaussian distribution when $\beta = 2$, and it is the Laplace distribution when $\beta = 1$. Moreover, as β reduces the heavier the tails become.

On the other hand, the family of multivariate generalized Gaussian distributions has the form

$$p(\mathbf{s}_i) \propto exp\Big(- \Big(\frac{1}{\alpha} \sqrt{(\mathbf{s}_i - \boldsymbol{\mu}_i)^{\dagger} \Sigma_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i)} \Big)^{\beta} \Big) \tag{5.2.2}$$

when $\alpha = 1$ and $\beta = 1$, it is the multivariate Laplace distribution adopted by the original IVA algorithm [27].

To derive a new nonlinear score function, $\beta = \frac{2}{3}$ and $\alpha = 1$ are set to yield

$$p(\mathbf{s}_i) \propto exp\Big(- \sqrt[3]{(\mathbf{s}_i - \boldsymbol{\mu}_i)^{\dagger} \Sigma_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i)} \Big) \tag{5.2.3}$$

the target multivariate generalized Gaussian source prior. This proposed source prior has a heavier tail than the original one, which can have advan-

tage in separating speech-like non-stationary signals [76].

This proposed source prior can also preserve the inter-frequency dependencies within each source vector following the approach in [27].

It begins with the definition of a K-dimensional random variable

$$\mathbf{s}_i = \gamma^{\frac{3}{4}} \boldsymbol{\xi}_i + \boldsymbol{\mu}_i \tag{5.2.4}$$

where γ is a scalar random variable, and $\boldsymbol{\xi}_i$ obeys a generalized Gaussian distribution which has the form:

$$p(\boldsymbol{\xi}_i) \propto exp\big(- \big(\frac{\boldsymbol{\xi}_i^{\dagger} \Sigma_i^{-1} \boldsymbol{\xi}_i}{2\sqrt{2}}\big)^{\frac{2}{3}}\big). \tag{5.2.5}$$

If γ has a Gamma distribution of the form:

$$p(\gamma) \propto \gamma^{\frac{1}{2}} exp\big(- \frac{\gamma}{2}\big) \tag{5.2.6}$$

then the proposed source prior can be achieved by integrating the joint distribution of $\mathbf{s}_i$ and γ over γ as follows:

$$\begin{aligned} p(\mathbf{s}_i) &= \int_0^{\infty} q(\mathbf{s}_i|\gamma) p(\gamma) d\gamma \\ &= \alpha_1 \int_0^{\infty} \gamma^{\frac{1}{2}} exp\big(- \frac{1}{2}\big(\frac{((\mathbf{s}_i - \boldsymbol{\mu}_i)^{\dagger} \Sigma_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i))^{\frac{2}{3}}}{\gamma} + \gamma\big)\big) d\gamma \\ &= \alpha_2 exp\big(- \sqrt[3]{(\mathbf{s}_i - \boldsymbol{\mu}_i)^{\dagger} \Sigma_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i)}\big) \end{aligned} \tag{5.2.7}$$

where α_1 and α_2 are both normalization terms. Therefore, equation (5.2.7) confirms that the proposed source prior has the dependency generated by γ.

In [74] Lee discusses the source priors suitable for IVA, which are termed as the spherical symmetric sparse (SSS) source priors. A general form of this source prior is described as:

$$p(\mathbf{s}_i) \propto exp(-(\|\mathbf{s}_i\|_p)^{\frac{1}{L_s}}) = exp\Big(- \Big(\sum_k |s_i(k)|^p \Big)^{\frac{1}{pL_s}} \Big) \tag{5.2.8}$$

where $\|\cdot\|_p$ denotes the l_p norm, and L_s is termed as the sparseness parameter. Lee suggested that the spherical symmetry assumption is suitable for modeling the frequency components of speech, i.e. $p = 2$, and through certain experimental studies found that the best separation performance can be achieved when L_s is around 7.

The proposed source prior also belongs to this family. If choosing $p = 2$ to make it spherically symmetric, and choosing $L_s = \frac{3}{2}$, the proposed source prior can be obtained. In the detailed experimental results in the experimental section, inconsistent improvement in separation performance will be shown when L_s is around 7, as proposed in [74], while the NG-IVA which adopts the proposed source prior can consistently achieve improved separation performance.

5.3 NG-IVA with the Proposed Source Prior

Applying this proposed source prior to derive the nonlinear score function with the assumption that the mean vector of the sources is zero and the covariance matrix is a diagonal matrix, the nonlinear function becomes

$$\varphi^{(k)}(\hat{s}_i(1)\dots\hat{s}_i(K)) = \frac{2\frac{\hat{s}_i(k)}{\sigma_i(k)}}{3\sqrt[3]{\left(\sum_{k'=1}^{K}\left|\frac{\hat{s}_i(k')}{\sigma_i(k')}\right|^2\right)^2}}. \tag{5.3.1}$$

If the equation under the cubic root is expanded, it can be written as:

$$\left(\sum_{k'=1}^{K}\left|\frac{\hat{s}_i(k')}{\sigma_i(k')}\right|^2\right)^2 = \sum_{k'=1}^{K}\left|\frac{\hat{s}_i(k')}{\sigma_i(k')}\right|^4 + \sum_{a\neq b} c_{ab}|\hat{s}_i(a)|^2|\hat{s}_i(b)|^2 \tag{5.3.2}$$

which contains cross items $\sum_{a\neq b} c_{ab}|\hat{s}_i(a)|^2|\hat{s}_i(b)|^2$, and c_{ab} is a scalar constant between the a-th and b-th frequency bins. These terms are related to the fourth order relationships between different components for each source vector, and capture the level of interdependency between different frequency

bins. Thus, this new multivariate nonlinear function includes important information describing the dependency structure [75].

The use of such fourth order relationships of speech signals was not previously highlighted in frequency domain BSS based on IVA. An example of the second order relationships and the fourth order relationships inherent to a particular speech signal "si1010.wav" from the TIMIT database [38], with 8kHz sampling frequency and 1024 STFT length, will be shown. Fig. 5.1 is part of the image display of elements of the covariance matrix formed by sample interrelationships between the elements of the signal vector, which is correspondent to the low frequency bins. It is hard to observe any information correspondent to the high frequency bins due to the limited energy. Therefore, only part of the image is shown. It shows that only the diagonal has significant second order relationships information. This is because of the orthogonal Fourier basis.

Now a similar fourth order matrix is constructed to exploit the fourth order relationships, which is structured as

$$\begin{pmatrix} E[|s_i(1)|^2|s_i(1)|^2] & \cdots & E[|s_i(1)|^2|s_i(K)|^2] \\ \vdots & \ddots & \vdots \\ E[|s_i(K)|^2|s_i(1)|^2] & \cdots & E[|s_i(K)|^2|s_i(K)|^2] \end{pmatrix}. \tag{5.3.3}$$

Fig. 5.2 is part of this fourth order matrix, which is also correspondent to the same low frequency bins as Fig. 5.1. It is evident that there are fourth order relationships throughout the matrix not only on the diagonal. Thus, such fourth order relationships should be exploited to help separation.

Next it will be shown that the proposed source prior is the best choice to introduce the fourth order relationship as shown in equation (5.3.2). According to equation (5.2.2), if assuming $\alpha = 1$, the source prior can have the

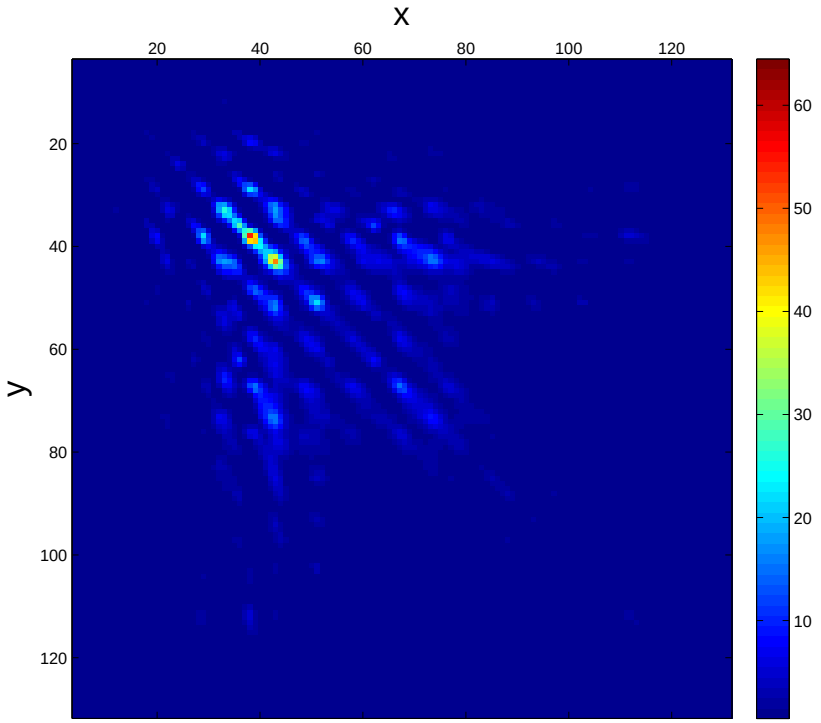


Figure 5.1. Second order inter-frequency relationships for the speech signal "si1010.wav", x and y dimensions correspond to frequency bins 1 to 128 of 512.

general form

$$p(\mathbf{s}_i) \propto exp\Big(- \Big(\sum_{k=1}^{K} |s_i(k)|^2 \Big)^{\beta} \Big) \tag{5.3.4}$$

And the nonlinear score function derived from this source prior is

$$\varphi^{(k)}(\hat{s}_i(1) \dots \hat{s}_i(K)) = \frac{2\beta \hat{s}_i(k)}{(\sum_{k'=1}^{K} |\hat{s}_i(k')|^2)^{1-\beta}} \tag{5.3.5}$$

In order to preserve the fourth order relationship as shown in equation (5.3.2), the root should be an odd number. Thus the following condition must be satisfied

$$1 - \beta = \frac{2}{2I' + 1} \tag{5.3.6}$$

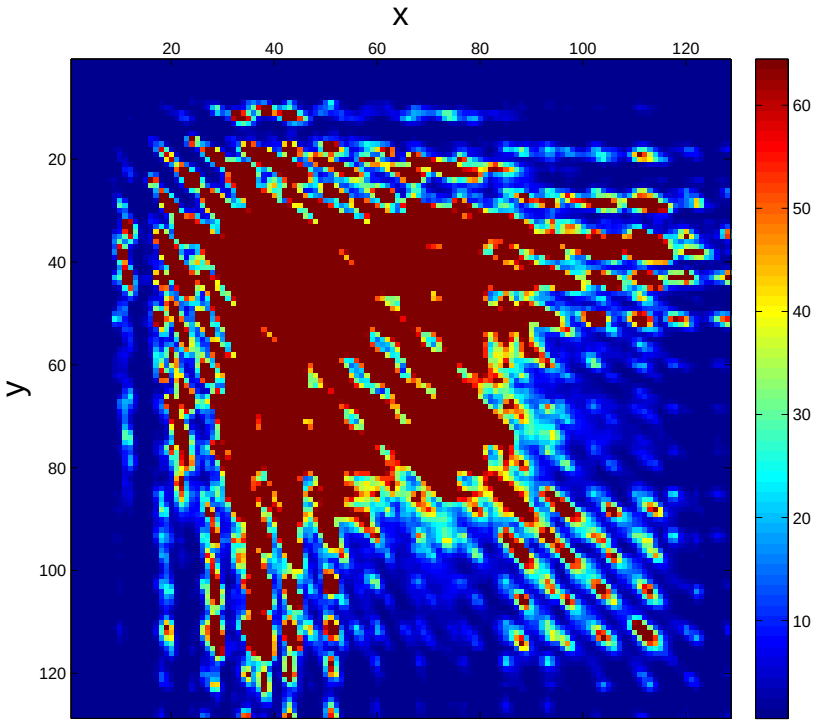


Figure 5.2. Fourth order inter-frequency relationships for the speech signal "si1010.wav", x and y dimensions correspond to frequency bins 1 to 128 of 512.

where I' is a positive integer. Then the condition for β can be obtained.

$$\beta = \frac{2I' - 1}{2I' + 1} \tag{5.3.7}$$

On the other hand, β is the shape parameter of the generalized multivariate Gaussian distribution. In order to make the proposed source prior heavier tail and more robust when separating the statistically non-stationary signals compared with the original source prior whose β is 1/2, β should be less than the 1/2. Thus

$$\frac{2I' - 1}{2I' + 1} < \frac{1}{2} \tag{5.3.8}$$

Finally, $I' = 1$ is the only solution, and the correspondent β is 1/3, as proposed in this chapter.

5.4 FastIVA with the Proposed Source Prior

Fast fixed-point independent vector analysis is a rapidly converging form of IVA algorithm, which has been introduced in Chapter 2. Newton's method is adopted in the update, which converges quadratically and is free from selecting an efficient learning rate. In order to apply Newton's method in the update rules, a quadratic Taylor series polynomial approximation is introduced in the notations of complex variables which can be used for a contrast function [29]. The cost function used by FastIVA is as follows:

$$J_{FastIVA} = \sum_{i=1}^{N} \Big(E\Big[F(\sum_{k=1}^{K} |\hat{s}_i(k)|^2)\Big] - \sum_{k=1}^{K} \lambda_i^{(k)} (\mathbf{w}_i(k)^\dagger \mathbf{w}_i(k) - 1)\Big) \quad (5.4.1)$$

where $F(\cdot)$ is the nonlinear function, which can take on several different forms as discussed in [29]. It is a multivariate function of the summation of the desired signals in all frequency bins. With normalization, the learning rule is:

$$\begin{aligned} \mathbf{w}_i(k) \leftarrow & E\Big[F^{'}(\sum_{k'=1}^{K} |\hat{s}_i(k')|^2) + |\hat{s}_i(k)|^2 F^{''}(\sum_{k'=1}^{K} |\hat{s}_i(k')|^2))\Big]\mathbf{w}_i(k) \\ & - E\Big[(\hat{s}_i(k))^* F^{'}(\sum_{k'=1}^{K} |\hat{s}_i(k')|^2)\mathbf{x}(k)\Big] \end{aligned} \quad (5.4.2)$$

If this is used for all sources, an unmixing matrix $W(k)$ can be constructed which needs to be decorrelated with

$$W(k) \leftarrow (W(k)(W(k))^\dagger)^{-1/2} W(k). \quad (5.4.3)$$

When the multivariate Laplacian distribution is used as the source prior for the FastIVA algorithm, with the zero mean and unity variance assumptions, the nonlinear function takes the form

$$F(\sum_{k'=1}^{K} |\hat{s}_i(k')|^2) = \big(\sum_{k'=1}^{K} |\hat{s}_i(k')|^2\big)^{\frac{1}{2}}. \quad (5.4.4)$$

When the proposed multivariate generalized Gaussian distribution is used as the source prior, with the same assumptions, the nonlinear function becomes:

$$F(\sum_{k'=1}^{K}|\hat{s}_i(k')|^2) = \big(\sum_{k'=1}^{K}|\hat{s}_i(k')|^2\big)^{\frac{1}{3}}. \tag{5.4.5}$$

Therefore, the first derivative becomes:

$$F'(\sum_{k'=1}^{K}|\hat{s}_i(k')|^2) = \frac{2}{3\sqrt[3]{(\sum_{k'=1}^{K}|\hat{s}_i(k')|^2)^2}}. \tag{5.4.6}$$

It is very similar to equation (5.3.1), and it also contains cross terms which can exploit the fourth order relationships between different frequency bins. Thus, the FastIVA algorithm with the proposed source prior is likely to help improve the separation performance.

5.5 AuxIVA with the Proposed Source Prior

As introduced in Chapter 2, the update rules for AuxIVA contains two parts, i.e. the auxiliary variable updates and unmixing matrix updates. In summary, the update rules are as follows:

$$r_i = \sqrt{\sum_{k=1}^{K}|\mathbf{w}_i^{\dagger}(k)\mathbf{x}(k)|^2} \tag{5.5.1}$$

$$V_i(k) = E[\frac{g_R'(r_i)}{r_i}\mathbf{x}(k)\mathbf{x}(k)^{\dagger}] \tag{5.5.2}$$

$$\mathbf{w}_i(k) = (W(k)V_i(k))^{-1}\mathbf{e}_i \tag{5.5.3}$$

$$\mathbf{w}_i(k) = \frac{\mathbf{w}_i(k)}{\sqrt{\mathbf{w}_i^{\dagger}(k)V_i(k)\mathbf{w}_i(k)}}. \tag{5.5.4}$$

The contrast function $g(\mathbf{z})$ is derived from the source prior [46]. For the

original AuxIVA algorithm,

$$g(\mathbf{z}) = r \tag{5.5.5}$$

where $r = \|\mathbf{z}\|_2$.

By using the proposed source prior, a new contrast function can be obtained

$$g(\mathbf{z}) = r^{\frac{2}{3}} \tag{5.5.6}$$

both the original and new proposed contrast function belong to S_g.

During the update process of the auxiliary variable $V_i(k)$ as (5.5.2), it is noticed that $\frac{g'_R(r_i)}{r_i}$ is used to retain the dependency between different frequency bins for source i. In this chapter, as defined previously, $g_R(r) = r^{\frac{2}{3}}$. Therefore

$$\frac{g'_R(r_i)}{r_i} = \frac{2}{3r_i^{\frac{4}{3}}} = \frac{2}{3\sqrt[3]{(\sum_{k'=1}^{K} |\hat{s}_i(k')|^2)^2}} \tag{5.5.7}$$

which has the same form as equation (5.4.6). The update rules also contain the terms to exploit the fourth order relationships within the speech signal vectors and should thereby help to achieve a better separation performance, which will be assessed by simulation study.

5.6 Experimental Results

In this section, it will be shown that all three types of IVA algorithm with the proposed multivariate generalized Gaussian source prior can improve the separation performance consistently when measurements are taken in a reverberant room environment. In these experiments, the TIMIT dataset [38] is used. Each speech signal is approximately seven seconds long. The image method is used to generate the room impulse responses [50], and the size of the room is $7 \times 5 \times 3\text{m}^3$. The STFT length is 1024, and the reverberation time RT60 = 200ms. A 2×2 mixing case is used, for which the microphone positions are [3.48, 2.50, 1.50]m and [3.52, 2.50, 1.50]m respectively. The

sampling frequency is 8kHz. The separation performance is evaluated objectively by SDR and SIR [51]. Fig. 5.3 is the plan view of the experimental setting.

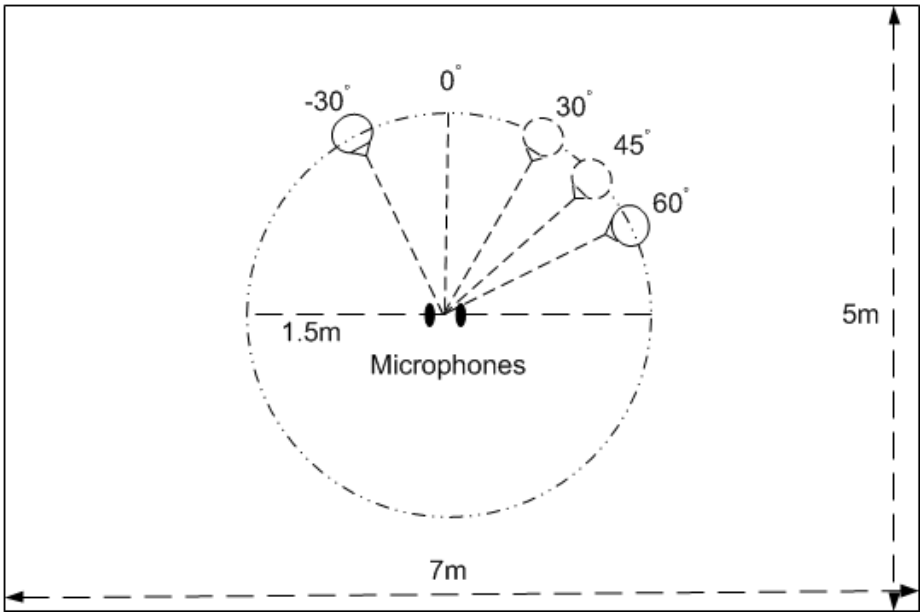


Figure 5.3. Plan view of the experiment setting in the room environment with two microphones and two sources

In the first experiment, two different speech signals are chosen randomly from the TIMIT dataset and convolved into two mixtures. Then the original NG-IVA, the NG-IVA method with the proposed source prior and the NG-IVA with Lee's SSS source prior where the sparseness parameter $L_s = 6, 7$ and 8, around the value suggested in [74], are all used to separate the mixtures respectively. Then the source positions are changed to repeat the simulation. For every pair of speech signals, three different azimuth angles for the sources relative to the normal to the microphone array are set for testing, these angles are selected from 30, 45, 60 and -30 degrees. After that, another pair of speech signals is chosen to repeat the above simulations. In total, ten different pairs of speech signals are used, and the simulation is repeated 30 times at different positions. Tables 5.1 and 5.2 show the average

separation performance for each pair of speech signals in terms of SDR and SIR in dB.

Table 5.1. Separation performance comparison in SDR(dB)

	original	proposed	SSS L_s=6	SSS L_s=7	SSS L_s=8
mixture 1	12.27	12.90	12.62	4.74	5.88
mixture 2	18.13	18.47	18.39	18.34	18.27
mixture 3	8.88	11.83	11.44	11.41	7.84
mixture 4	15.57	16.92	15.48	5.95	6.29
mixture 5	18.10	18.69	15.78	15.44	19.44
mixture 6	18.81	19.58	5.04	3.71	5.41
mixture 7	15.94	16.59	15.35	8.63	8.82
mixture 8	15.29	15.75	16.05	16.03	16.01
mixture 9	18.58	19.05	19.21	17.35	10.05
mixture 10	18.80	19.31	0.76	0.78	0.79

Table 5.2. Separation performance comparison in SIR(dB)

	original	proposed	SSS L_s=6	SSS L_s=7	SSS L_s=8
mixture 1	14.08	14.84	14.82	5.62	7.07
mixture 2	19.57	19.86	19.86	19.81	19.75
mixture 3	10.72	13.74	13.22	13.19	9.14
mixture 4	16.98	18.46	16.89	7.16	7.55
mixture 5	20.14	20.47	17.32	16.94	20.75
mixture 6	20.30	20.98	5.92	4.35	6.33
mixture 7	17.88	18.40	16.39	10.73	9.93
mixture 8	19.88	20.41	20.65	20.61	20.56
mixture 9	20.75	20.89	20.85	18.80	11.00
mixture 10	20.28	20.60	1.45	1.48	1.51

The experimental results show clearly that IVA with the proposed source prior can consistently improve the separation performance. However, for the IVA with SSS source prior, the separation improvement is not consistent. For example, when $L_s = 7$, in some cases there is essentially no separation such as mixtures 1, 6 and 10. Even though it can achieve better separation than the original IVA, it is still no better than the proposed method. Only for

mixture 8, does it achieve the best separation performance. Therefore, the IVA with the proposed source prior can generally achieve better separation performance. Based on totally 30 tests, the average SDR improvement and SIR improvement are approximately 0.9dB and 0.8dB, respectively.

Then the performance of the IVA with the proposed source prior in different reverberant room environments is tested. Two speech signals from the TIMIT dataset are selected randomly and convolved into two mixtures. The azimuth angles for the sources relative to the normal to the microphone array are set as 60 and -30 degrees. Both the original IVA and the proposed method are used to separate the mixtures. The results are shown in Fig. 5.4 and Fig. 5.5, which show the separation performance comparisons in different reverberant environments. Fig. 5.4 and Fig. 5.5 show the SDR and SIR comparison respectively. They indicate that the proposed algorithm can consistently improve the separation performance in different reverberant environments, up to a reverberation time of 450ms. The advantage reduces with increasing RT60 due to the greater challenge in extracting the individual source vectors.

In the second experiment, all the experimental settings and the processes are all the same as the first experiment. Here five pairs of speech signals from the TIMIT dataset are selected and convolved into mixtures. The original FastIVA algorithm and the FastIVA algorithm with the proposed source prior are used to separate the speech mixtures. Then the source positions are changed to repeat the experiment, the average separation performance comparison is shown in Table 5.3. It shows that the separation performance can be improved by adopting the proposed source prior. The average SDR improvement and SIR improvement both are approximately 0.6dB.

The separation performance of these two algorithms are also compared in different reverberant room environments as in the first experiment. The SDR and SIR comparisons are shown in Fig. 5.6 and Fig. 5.7. They show

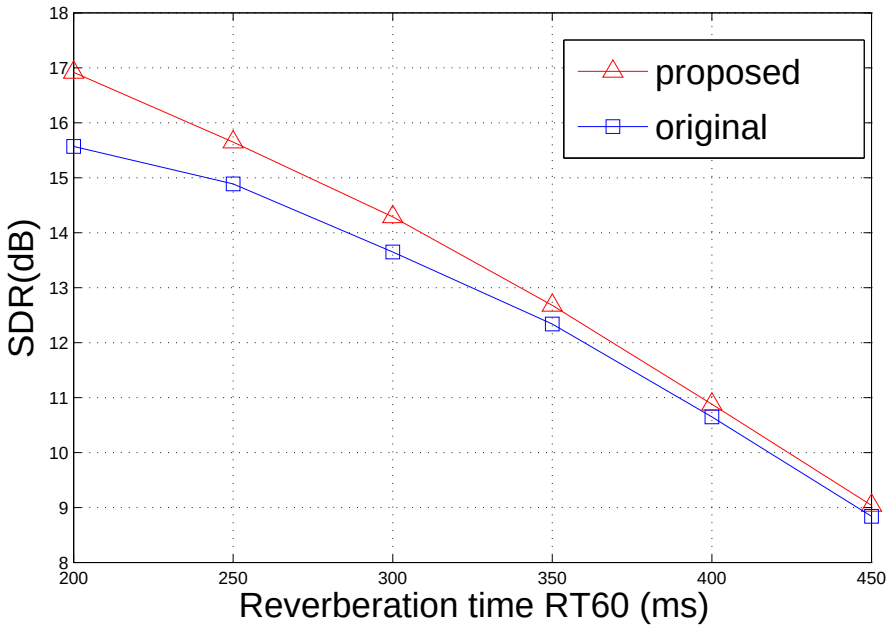


Figure 5.4. SDR comparison between original and proposed IVA algorithms as a function of reverberation time.

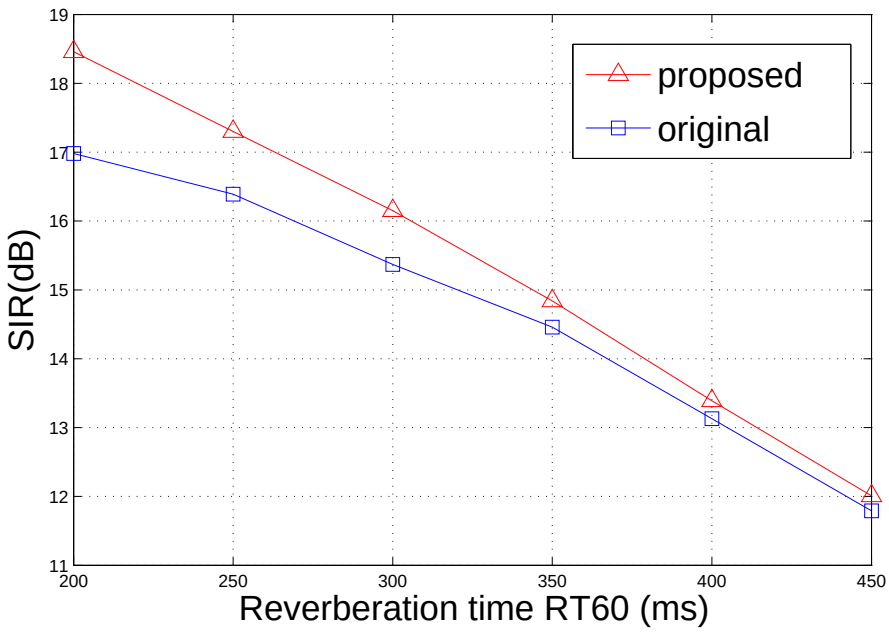


Figure 5.5. SIR comparison between original and proposed IVA algorithms as a function of reverberation time.

Table 5.3. Separation performance comparison in terms of SDR and SIR measures in dB

Mixtures	mix1	mix2	mix3	mix4	mix5
Original FastIVA (SDR)	17.77	19.48	14.75	18.12	16.79
Proposed FastIVA (SDR)	18.04	20.63	15.08	18.88	17.31
Original FastIVA (SIR)	19.32	21.01	17.04	19.80	19.18
Proposed FastIVA (SIR)	19.59	22.04	17.31	20.51	19.74

the SDR and SIR comparison respectively. The results indicate that the FastIVA algorithm with the proposed source prior can improve the separation performance, but again the advantage is reduced with increasing RT60.

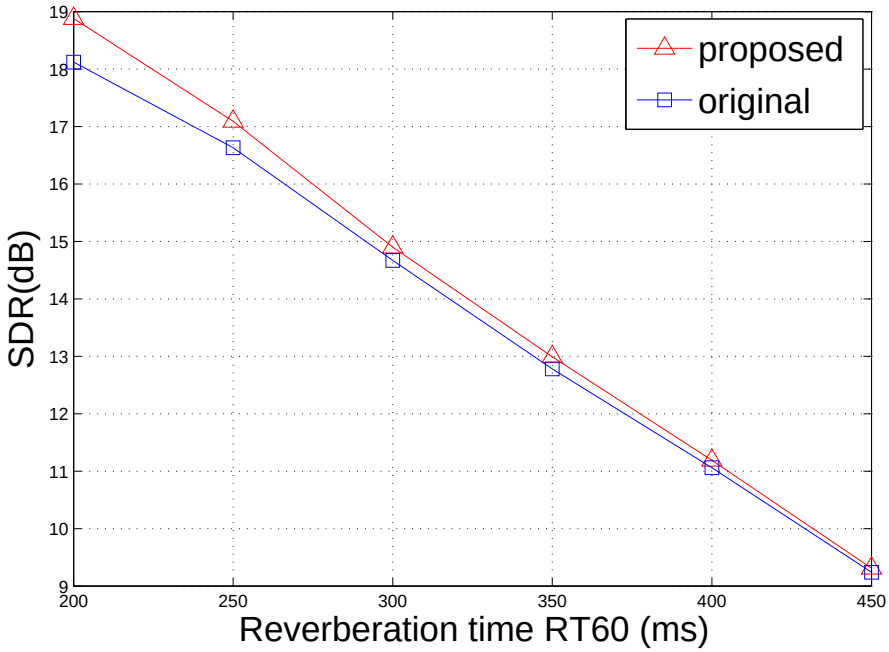


Figure 5.6. SDR comparison between original and proposed FastIVA algorithms as a function of reverberation time.

In the third experiment, the second experiment is repeated by using the original AuxIVA and AuxIVA with the proposed source prior. Five different pairs of speech signals are used, and the simulation is repeated 15 times at different positions. Table 5.4 shows the average separation performance for each pair of speech signals in terms of SDR and SIR. The average SDR and

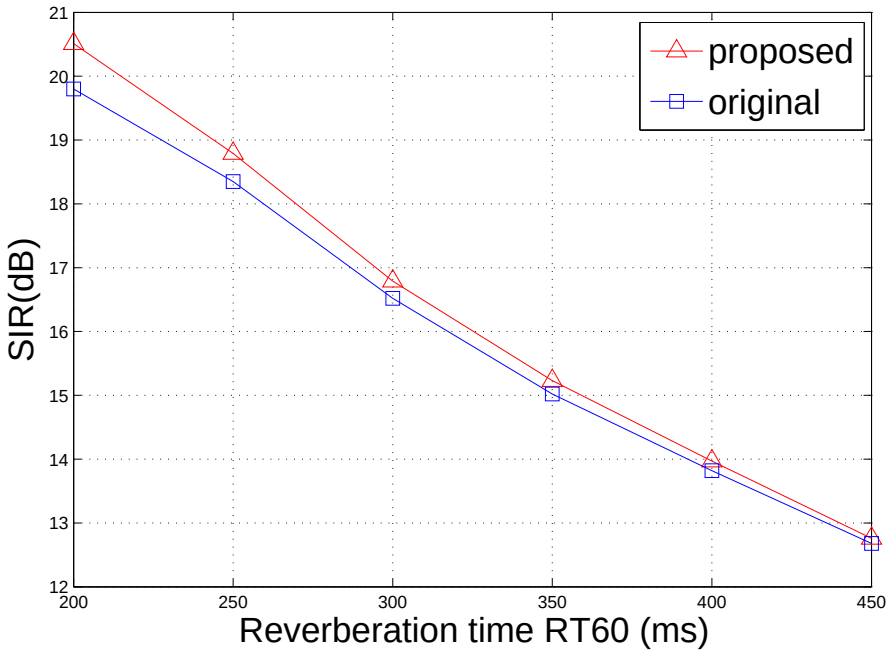


Figure 5.7. SIR comparison between original and proposed FastIVA algorithms as a function of reverberation time.

SIR improvements are approximately 1.7dB and 1.9dB respectively. The results confirm the advantage of the proposed AuxIVA method which can better preserve the dependency between different frequency bins of each source and thereby achieve a better separation performance.

Table 5.4. Separation performance comparison in terms of SDR and SIR measures in dB

Mixtures	mix1	mix2	mix3	mix4	mix5
Original AuxIVA (SDR)	12.13	14.62	9.86	19.23	18.64
Proposed AuxIVA (SDR)	14.82	16.30	12.45	19.92	19.50
Original AuxIVA (SIR)	14.06	16.72	11.59	20.54	20.12
Proposed AuxIVA (SIR)	17.26	18.42	14.58	21.20	20.90

Then the performance of the proposed AuxIVA method in different reverberant room environments is also tested. The experimental settings are all the same as the first experiment. The results are shown in Fig. 5.8 and Fig. 5.9, which show the separation performance comparison in different

reverberant environments. Fig. 5.8 and Fig. 5.9 show the SDR and SIR comparison respectively. It indicates that the AuxIVA algorithm with the proposed source prior can consistently improve the separation performance in different reverberant environments.

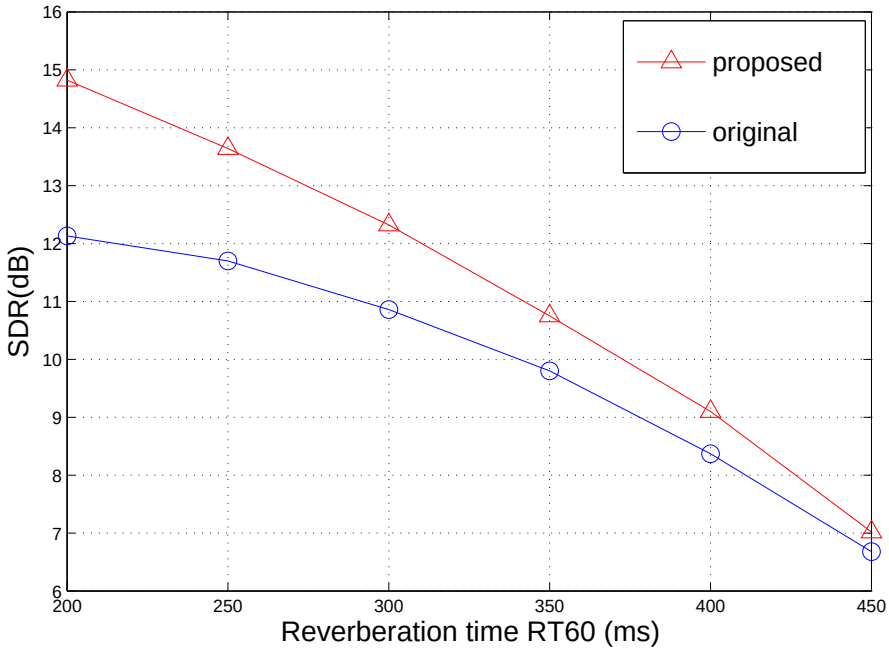


Figure 5.8. SDR comparison between original and proposed AuxIVA algorithms as a function of reverberation time.

Examing the results for the three algorithms, the proposed source prior offers the maximum improvement in the AuxIVA algorithm. It is difficult however to make a general recommendation which is the best algorithm due to the variability of performance with different speech signals and mixing environments.

5.7 Summary

In this chapter, a particular multivariate generalized Gaussian source prior was proposed to adopt in independent vector analysis. This particular source prior can better preserve the inter-frequency dependencies as compared to

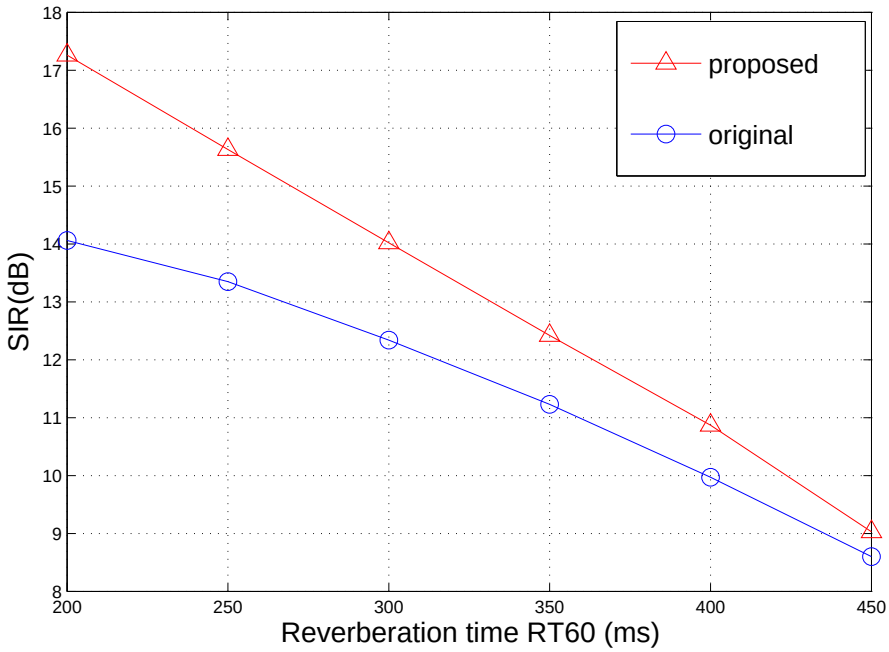


Figure 5.9. SIR comparison between original and proposed AuxIVA algorithms as a function of reverberation time.

the original multivariate Laplace source prior. When used in IVA algorithms, it introduced fourth order relationships commonly found in speech signals to improve the learning process and enhance separation. Three basic forms of IVA algorithms with the proposed source prior, i.e. NG-IVA, FastIVA and AuxIVA, were all analyzed with non-stationary source signals, and the experimental results confirm the advantage of adopting the proposed source prior particular for smaller reverberation time. In the next chapter, the dependency within the frequency domain speech signals is further exploited, and another new source prior is proposed for IVA algorithms.

Chapter 6

COPULA BASED INDEPENDENT VECTOR ANALYSIS WITH THE MULTIVARIATE STUDENT'S T SOURCE PRIOR

6.1 Introduction

For IVA, new statistical models which can better preserve the dependency within the source vector still need to be exploited to satisfy the third objective of this study. The dependency between different frequency bins can be nonlinear, and the simple linear dependency, i.e the Pearson correlation, can not always describe it accurately. The copula concept is widely used for modeling nonlinear dependency, and was first widely used in the field of finance [77]. Nowadays, it has been adopted in various engineering fields. For example, a copula is used for modeling stochastic dependence in power system uncertainty analysis in [78]. Moreover, the Gaussian copula is adopted to model texture in image in [79]. However, few works introduce the copula in the speech separation field, especially for IVA. As the copula is a cen-

tral tool for modeling nonlinear dependency, the copula should be exploited in IVA. Thus, the IVA algorithm can thereby achieve improved separation performance with a better dependency structure to retain the dependency within the source vector.

In this chapter, the t copula is used to construct a multivariate student's t distribution with univariate student's t marginal distributions as the source prior for IVA. Recently, the student's t distribution has been popular for modeling speech signals [80, 81]. The student's t distribution is a super Gaussian distribution, which has heavier tails than the Gaussian distribution and is suitable for modeling certain speech signals. The t copula is chosen because it has tail dependence, which means if one variable has an extreme value, other variables are expected to have extreme values [82]. Due to the heavy tail property of certain speech signals, many useful samples can be in the tails. Thus the tail dependence can be an advantage when modeling the dependency between different frequency bins of a speech signal. This will be shown in this chapter. Moreover, it will be shown that the multivariate student's t distribution constructed by the t copula with univariate marginal student's t distributions can retain the dependency within each source vector. The NG-IVA algorithm with the proposed source prior will be tested in different room environments, and the experimental results will confirm the advantage of the proposed multivariate student's t source prior.

6.2 Copula Introduction

Copulas are used to model dependence of several random variables, and have been widely developed in the finance field. A copula is a joint cumulative distribution function (cdf). The dependency structure is entirely described by itself, and is not related to the marginal distribution. The definition of copula is [82]:

Definition 1: A d-dimensional copula $C : [0,1]^d :\to [0,1]$ is a function which is a cumulative distribution function with uniform marginals.

$$C(\boldsymbol{u}) = C(u_1, \cdots, u_d) \tag{6.2.1}$$

The basic and most important theorem for a copula was proposed by Sklar in 1959 [82].

Sklar's theorem: Consider a d-dimensional cdf F with marginals $F_1, \cdots, F_d$. There exists a copula C, such that

$$F(z_1, \cdots, z_d) = C(F_1(z_1), \cdots, F_d(z_d)) \tag{6.2.2}$$

for all z_i in $[-\infty, \infty]$, $i = 1, \cdots, d$. If F_i is continuous for all $i = 1, \cdots, d$ then C is unique; otherwise C is uniquely determined only on $RanF_1 \times \cdots \times RanF_d$, where $RanF_i$ denotes the range of the cdf F_i. It is also noticed that $u_i = F_i(z_i)$.

If C is continuous and differentiable, the copula density function c can by achieved by taking the d-th order partial derivative of C.

$$c(\boldsymbol{u}) = \frac{\partial^d C(u_1, \cdots, u_d)}{\partial u_1 \cdots u_d} \tag{6.2.3}$$

Therefore, the multivariate joint probability density function can be derived according to equations (6.2.2) and (6.2.3). The joint probability density function is obtained by taking the d-th order partial derivative of equation (6.2.2).

$$\begin{aligned} p(z_1, \cdots, z_d) &= \frac{\partial^d F(z_1, \cdots, z_d)}{\partial z_1 \cdots z_d} \\ &= \frac{\partial^d C(F_1, \cdots, F_d)}{\partial F_1 \cdots F_d} \frac{\partial F_1}{\partial z_1} \cdots \frac{\partial F_d}{\partial z_d} \\ &= c(F_1, \cdots, F_d) \prod_{i=1}^{d} p_i(z_i) \end{aligned} \tag{6.2.4}$$

By observing this multivariate joint pdf, it is evident that marginal pdfs are independent, but the copula density function c is used to describe the dependency structure among all the marginal pdfs. When the copula density c equals to unity, it is the independent case.

There are several copula families which have been proposed, such as the Gaussian copula, Archimedean copulas and t copula [77]. In this chapter, the t copula is the focus, since an algebraic update equation for IVA can be derived.

The t copula can be used to represent the dependency structure implicit in a multivariate student's t distribution [83], which has the form

$$C(\boldsymbol{u}) = \int_{-\infty}^{F_1^{-1}} \cdots \int_{-\infty}^{F_d^{-1}} \frac{\Gamma(\frac{v+d}{2})}{\Gamma(\frac{v}{2})(\sqrt{\pi v|\Sigma|})} \Big(1 + \frac{\boldsymbol{z}^{\dagger}\Sigma^{-1}\boldsymbol{z}}{v}\Big)^{-\frac{v+d}{2}} d\boldsymbol{z} \qquad (6.2.5)$$

where F_i^{-1} is the quantile function [84] of a standard univariate student's t distribution with v degrees of freedom; Σ is a positive definite matrix; $\Gamma(\cdot)$ is the Gamma function.

The t copula density function takes the form [85]:

$$c(u_1, \cdots, u_d) = \frac{\Gamma(\frac{v+d}{2})\Gamma(\frac{v}{2})^{d-1}}{|\Sigma|^{\frac{1}{2}}\Gamma(\frac{v+1}{2})^d} \frac{\prod_{i=1}^{d}(1 + \frac{|z_i|^2}{v})^{\frac{v+1}{2}}}{(1 + \frac{\boldsymbol{z}^{\dagger}\Sigma^{-1}\boldsymbol{z}}{v})^{\frac{v+d}{2}}} \qquad (6.2.6)$$

Fig. 6.1 is the bivariate case of a t copula density, and it is used to show the property of the t copula. For a bivariate case,

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

where ρ is the correlation coefficient, and $\rho = 0.7$ in Fig 6.1.

One of the attractive properties of a t copula is the tail dependence. Fig. 6.1 shows that the t copula has tail dependence, which means that if one variable has an extreme value, another variable is most likely to have

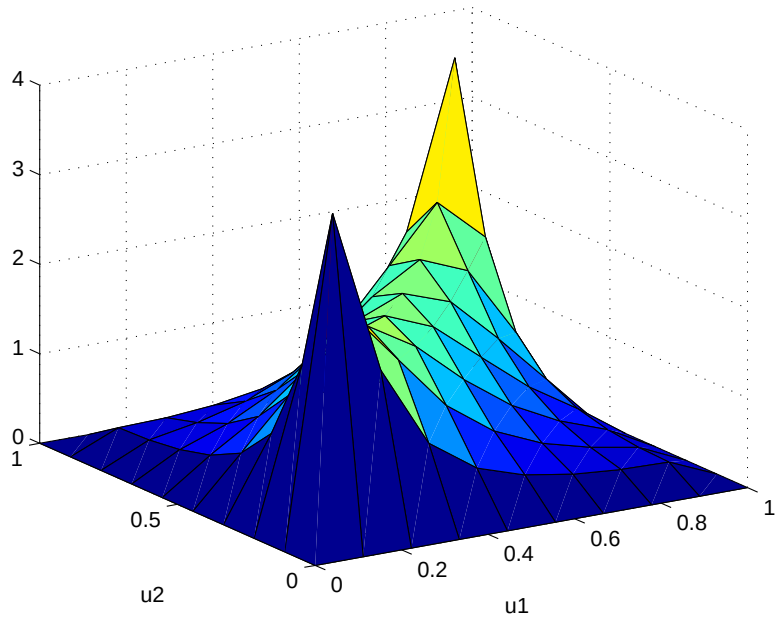


Figure 6.1. The copula density of a t copula with 4 degrees of freedom and correlation coefficient $\rho = 0.7$

an extreme value as well. The two peaks in Fig. 6.1 correspond to the copula density for the tails of two variables. It has a relative large value, which means the dependency is strong. Fig 6.2 is another bivariate case of t copula density with 4 degrees of freedom and $\rho = -0.6$.

For the t copula, even for the zero correlation, i.e $\rho = 0$, it still shows tail dependency [82]. Fig 6.3 confirms this by showing that the copula is not always unity.

Tail dependence has great advantage for modeling frequency domain speech signals. The distribution for a frequency domain speech signal is commonly a heavy tail distribution, which means most useful information is likely to be in the tails. The tail dependence captures the dependency between tails, thus it can emphasize the dependency among useful samples.

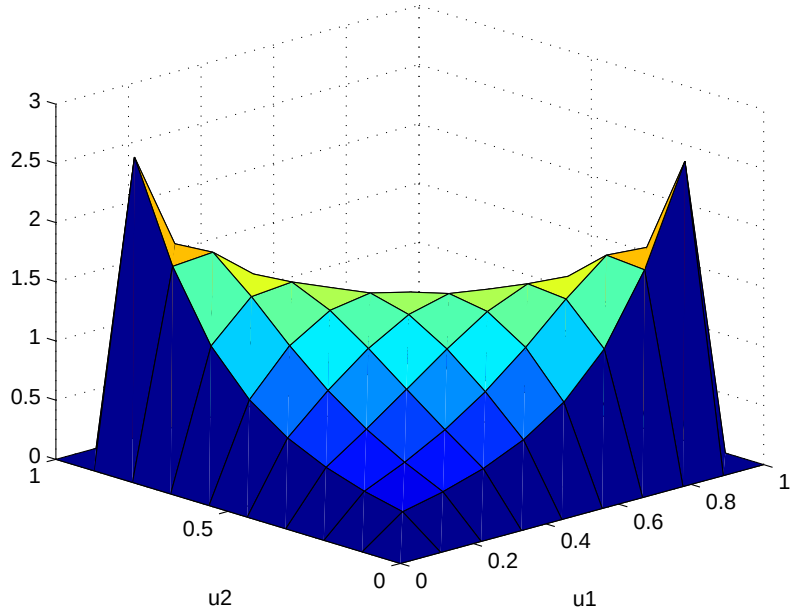


Figure 6.2. The copula density of a t copula with 4 degrees of freedom and correlation coefficient $\rho = -0.6$

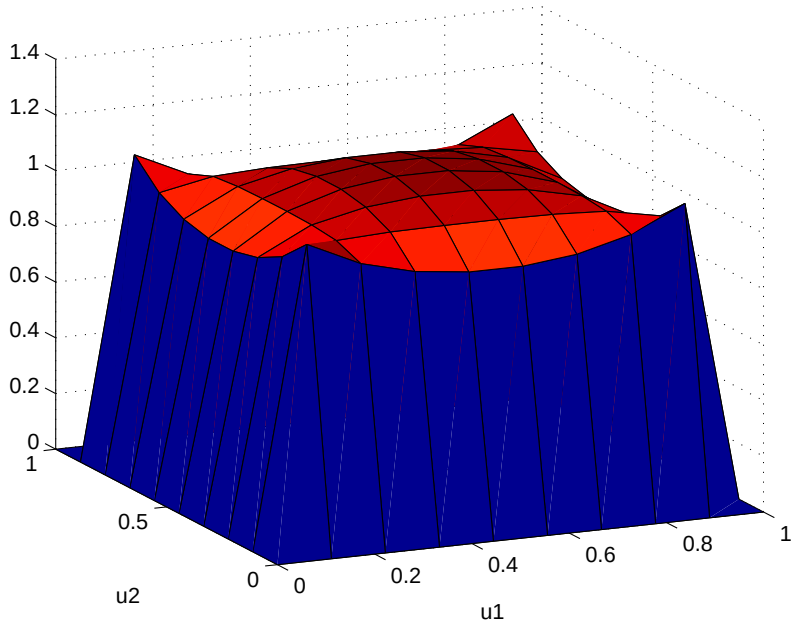


Figure 6.3. The copula density of a t copula with 4 degrees of freedom and correlation coefficient $\rho = 0$

6.3 Dependency within Frequency Domain Speech Signals

Now it will be shown that the t copula can be used to describe the dependency structure within frequency domain speech signals. The Chi-plot is used to observe the dependency [86]. It can be considered as an extension of the scatter plot to illustrate the dependency. The scatter plot usually has certain pattern when there are certain types of dependency. However, sometimes it is difficult to judge the pattern for some characteristic by only using a scatter plot. Then the Chi-plot is proposed to provide a graph illustrating the characteristic patterns, which is easier for observation.

Fig. 6.4 and Fig. 6.5 show the scatter plot and Chi-plot of two independent random variables z_1 and z_2 respectively. In the scatter plot, there is no pattern for the points. Meanwhile, in the correspondent Chi-plot, almost all the points are in the tolerance band, the band between two straight lines [87], which corresponds to approximate 95% probability region. As for the Chi-plot, the deviation from the tolerance band indicates a dependency structure. The x axis of the Chi-plot is the measure of the distance from the center of the dataset, which is denoted by *lamda*. The y axis is the correlation coefficients between dichotomized values, which is denoted by Chi [86]. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a random sample from $\tilde{H}$, the joint distribution function for a pair of random variables (X, Y), and let $I(.)$ be the indicator function. The calculations of Chi and $lamda$ are as follows:

$$\tilde{H}_i = \sum_{j \neq i} I(x_j \leq x_i, y_j \leq y_i)/(n-1) \tag{6.3.1}$$

$$\tilde{F}_i = \sum_{j \neq i} I(x_j \leq x_i)/(n-1) \tag{6.3.2}$$

$$\tilde{G}_i = \sum_{j \neq i} I(y_j \leq y_i)/(n-1) \tag{6.3.3}$$

$$\tilde{S}_i = sign[(\tilde{F}_i - 0.5)(\tilde{G}_i - 0.5)] \tag{6.3.4}$$

$$Chi_i = (\tilde{H}_i - \tilde{F}_i\tilde{G}_i)/\sqrt{\tilde{F}_i(1-\tilde{F}_i)\tilde{G}_i(1-\tilde{G}_i)} \tag{6.3.5}$$

$$lamda_i = 4S_i max((\tilde{F}_i - 0.5)^2, (\tilde{G}_i - 0.5)^2) \tag{6.3.6}$$

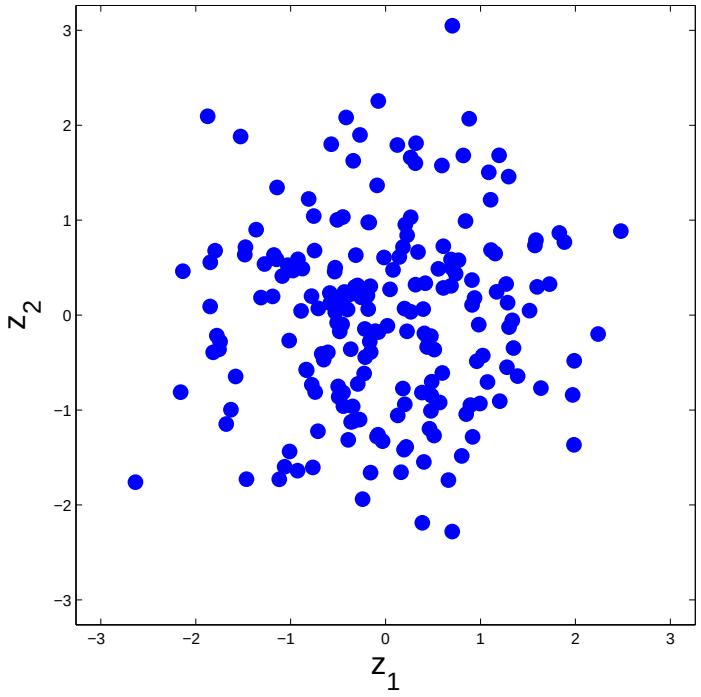


Figure 6.4. The scatter plot of two independent random variables.

Two random variables z_1 and z_2 with a bivariate t copula are generated, whose correlation coefficient $\rho = -0.6$ and degrees of freedom v is 4. Fig. 6.6 and Fig. 6.7 show the scatter plot and Chi-plot of these two random variables respectively. It clearly shows there is a pattern in the scatter plot, and an obvious deviation from the tolerance band, which shows a strong dependency structure. Thus, the t copula can generate a dependency structure.

Now it will be shown that even when the correlation coefficient is zero, a dependency structure still exists. Two random variables with a bivariate t copula with zero correlation and two degrees of freedom are generated. The scatter plot and Chi-plot are shown in Fig. 6.8 and Fig. 6.9 respectively. By observing the Chi-plot, which is different from the independent case, since

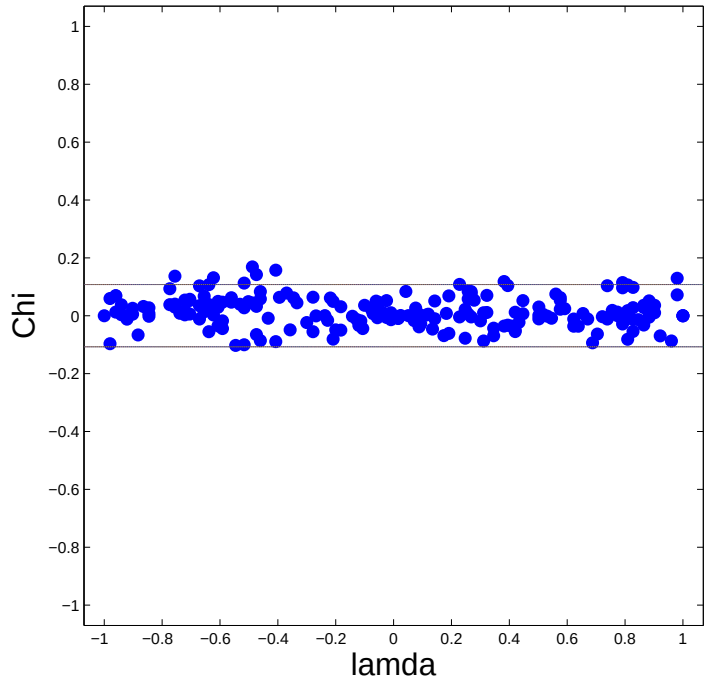


Figure 6.5. The Chi-plot of two independent random variables.

some points deviate from the tolerance band. This indicates that a t copula with zero correlation still has dependency structure.

For a frequency domain speech signal, it's impossible to draw a figure to show the dependency structure across all the frequency bins. Instead, the dependency structure between two different frequency bins is observed by showing the Chi-plot. A speech signal from the TIMIT database [38] is randomly selected, which is "sa1.wav", then the 1024 length STFT is applied to transform it into frequency domain. By observing Fig. 6.10, it can be seen that all the Chi-plots are similar to the t copula Chi-plot as shown in Fig. 6.7 and Fig. 6.9. Fig. 6.10(a) indicates that the dependency between adjacent frequency bins, i.e 50th and 51th frequency bins, is strong. Fig. 6.10(b), Fig. 6.10(c), Fig. 6.10(d), Fig. 6.10(e), Fig. 6.10(f) shows the dependency between 50th and 55th, 50th and 60th, 50th and 100th, 50th and 200th, 50th and 500th frequency bins respectively. These figures

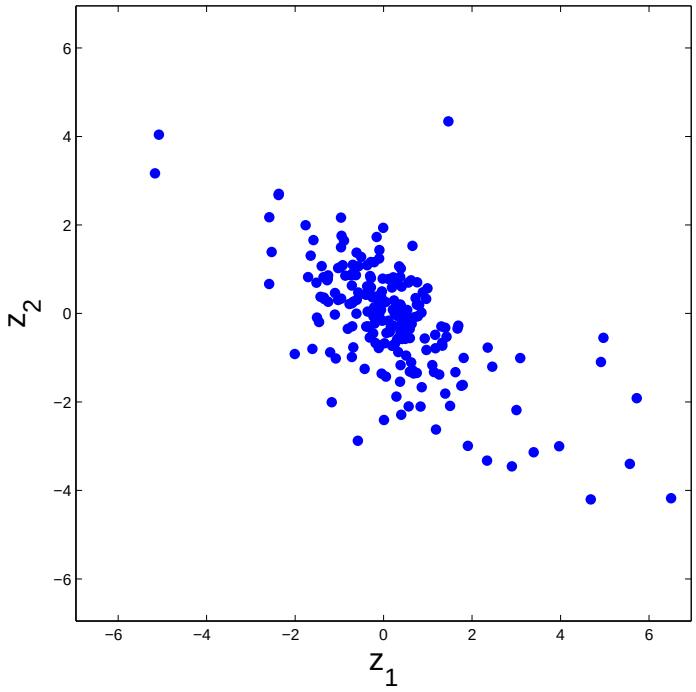


Figure 6.6. The scatter plot of two random variables with a t copula, $\rho = -0.6$ and $v = 4$

illustrate that the dependency becomes weak when the two frequency bins are far away. However, there is still dependency which can be described by a t copula. Therefore, the t copula is appropriate to describe the dependency structure within the frequency domain speech signals. In the next section, a multivariate source prior will be constructed by using a t copula to be the source prior for IVA.

6.4 IVA with the Multivariate Student's t Source Prior

It has been found that t copula is suitable for modeling the dependence structure for frequency domain speech signals. Thus a multivariate source prior will be constructed by using a t copula in this section.

According to equation (6.2.4), the marginal density function must be determined to construct the multivariate source prior. The marginal density

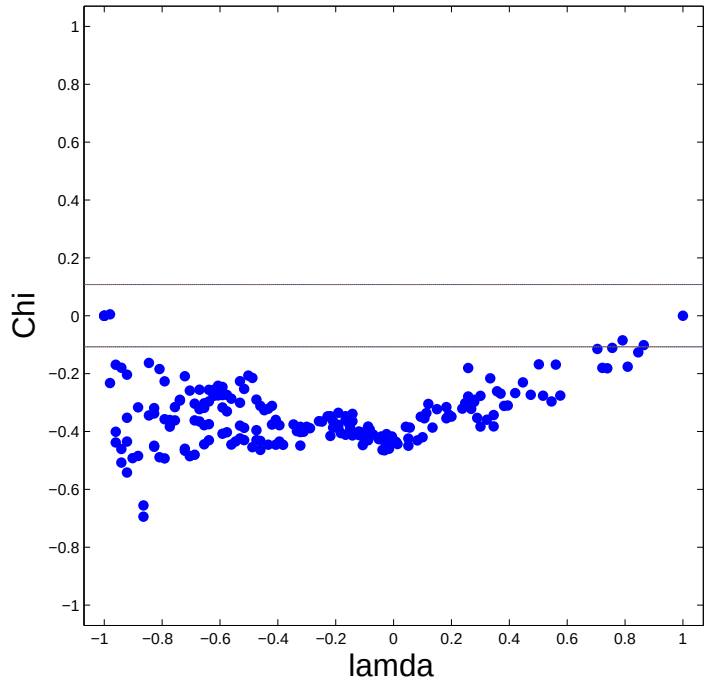


Figure 6.7. The Chi-plot of two random variables with a t copula, $\rho = -0.6$ and $v = 4$

function is used to describe the distribution of each frequency bin. The univariate student's t distribution is proposed as the marginal density function, which takes the form:

$$p(s_i(k)) = \frac{\Gamma(\frac{v+K}{2})}{\sqrt{v\pi}\Gamma(\frac{v}{2})}\left(1 + \frac{|s_i(k)|^2}{v}\right)^{-\frac{v+1}{2}} \tag{6.4.1}$$

The student's t distribution has a heavier tail than the Gaussian distribution, thus it can be suitable for modeling the spectrum of a speech signal. The degrees of freedom parameter v can tune the variance and leptokurtic nature of the distribution. With decreasing v, the tail of the distribution becomes heavier.

According to equations (6.2.4), (6.2.6) and (6.4.1), the multivariate source

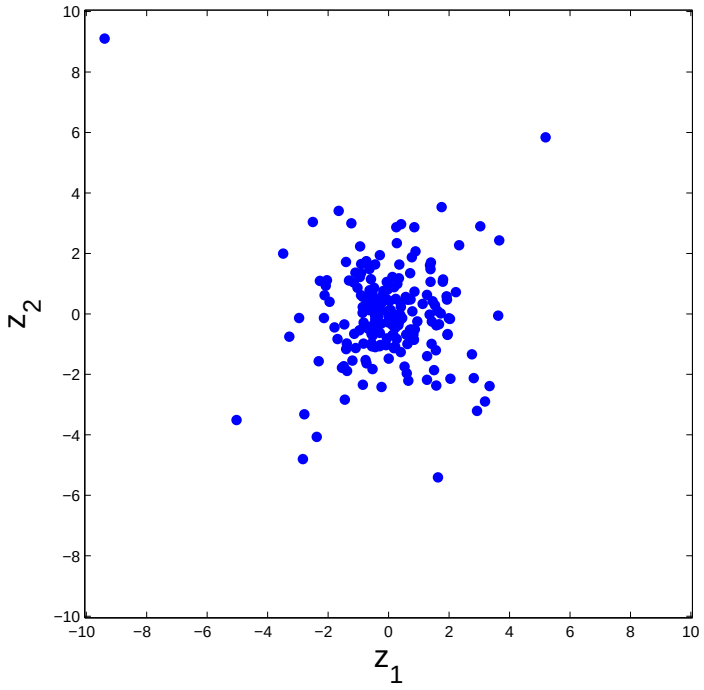


Figure 6.8. The scatter plot of two random variables with a t copula, $\rho = 0$ and $v = 2$

prior can be obtained

$$p(\mathbf{s}_i) \propto \left(1 + \frac{\mathbf{s}_i^{\dagger}\Sigma_i^{-1}\mathbf{s}_i}{v}\right)^{-\frac{v+K}{2}} \tag{6.4.2}$$

which is a K-dimensional student's t distribution. Fig. 6.11 is the probability density function for a two dimensional student's t distribution. The marginal probability density function is a univariate student's t distribution. However, the joint density function takes the form of equation (6.4.2) with $K = 2$, which is different from the product of marginal probability density functions. This indicates that different variables of the multivariate student's t distribution are dependent. Therefore, the multivariate student's t distribution can be used as a source prior to retain the dependence across the frequency bins.

Due to the orthogonal Fourier basis, theoretically there is no correlation

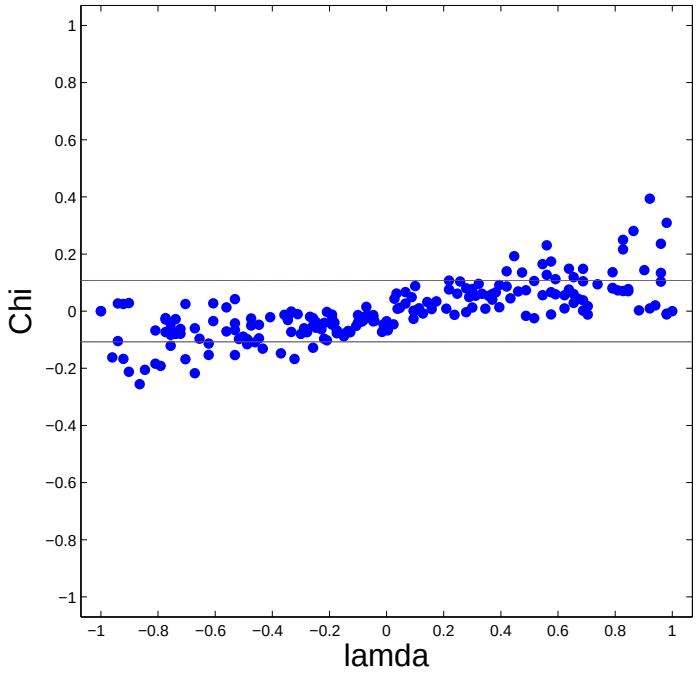


Figure 6.9. The Chi-plot of two random variables with a t copula, $\rho = 0$ and $v = 2$

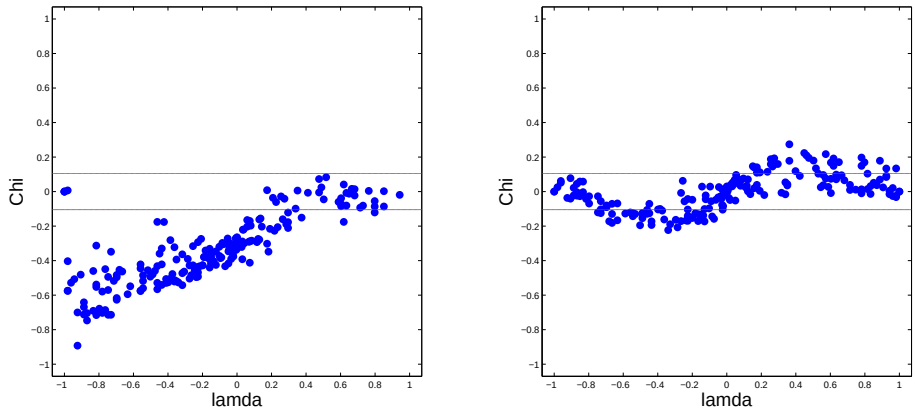
between different frequency bins. Thus Σ_i should be a diagonal matrix. Figs. 6.10(b), 6.10(c), 6.10(d), 6.10(e), 6.10(f) show that the dependency within frequency domain speech signals is similar to the dependency described by the t copula without correlation as shown in Fig. 6.9. As discussed in the last section, the dependency still exists even without correlation. It is assumed that Σ_i is an identity matrix, and equation (6.4.2) becomes

$$p(\mathbf{s}_i) \propto \left(1 + \frac{\sum_{k=1}^{K} s_i(k)}{v}\right)^{-\frac{v+K}{2}} \tag{6.4.3}$$

When equation (6.4.3) is used to derive the score function of IVA, the new nonlinear score function can be achieved

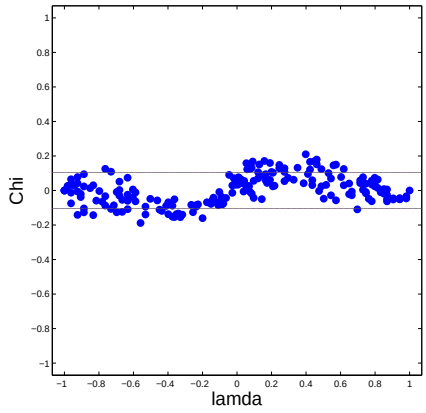
$$\varphi^{(k)}(\hat{s}_i(1) \cdots \hat{s}_i(k)) = \frac{v+K}{v} \frac{\hat{s}_i(k)}{1 + \frac{1}{v}\sum |\hat{s}_i(k)|^2} \tag{6.4.4}$$

The coefficient $\frac{v+K}{v}$ can be absorbed by the step size η in the update

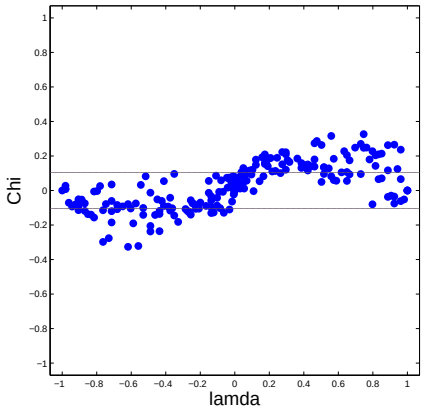


(a)

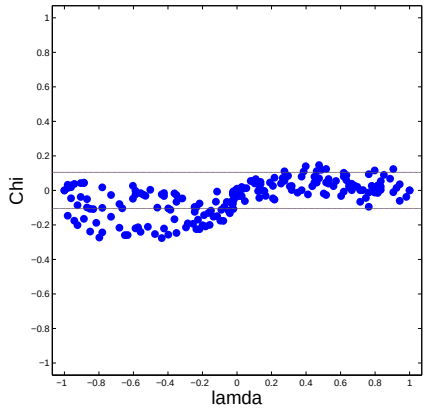
(b)



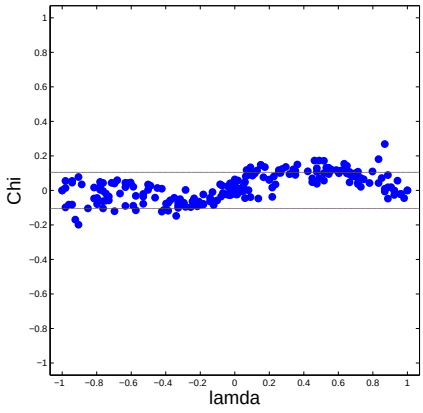
(c)



(d)



(e)



(f)

Figure 6.10. The Chi-plot of two frequency bins of a speech signal "sa1.wav" from TIMIT dataset (a) 50th and 51th frequency bins (b) 50th and 55th frequency bins (c) 50th and 60th frequency bins (d) 50th and 100th frequency bins (e) 50th and 200th frequency bins (b) 50th and 500th frequency bins

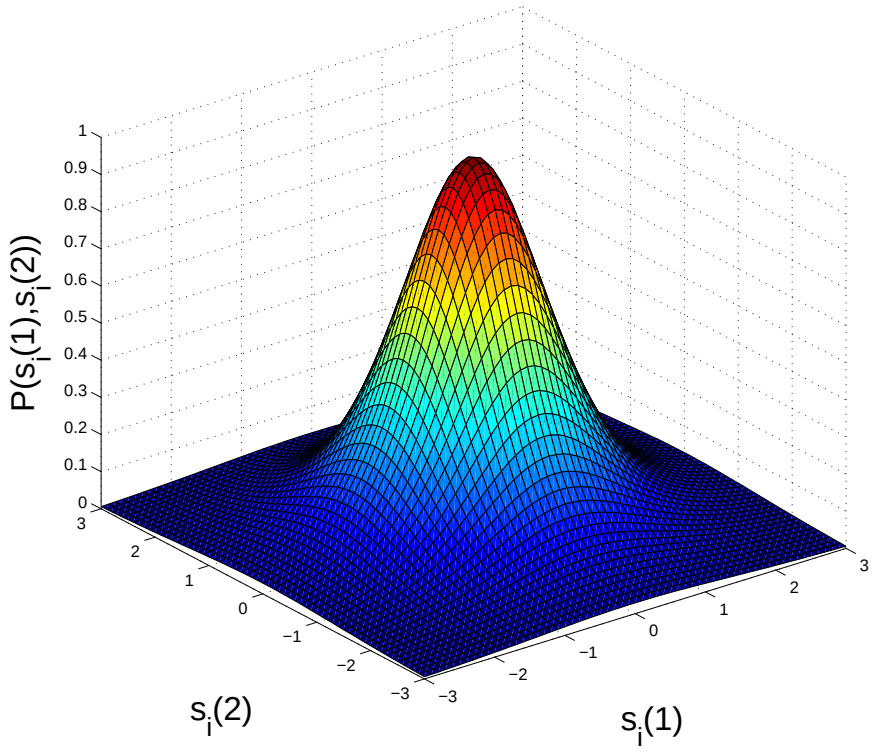


Figure 6.11. The probability density function of a multivariate student's t distribution

equation. Thus it can be normalized to unity. The performance of the IVA algorithm with the proposed source prior will be tested in the next section.

6.5 Experimental Results

In this section, the NG-IVA algorithm with the proposed multivariate student's t distribution in different environments will be tested and the results will show that it can achieve an improved separation performance. The speech signals are selected from the TIMIT dataset [38]. Each of the speech signals is approximately seven seconds long, and the sampling frequency is 8kHz. The image method is used to generate the room impulse responses [50], and the size of the room is $7 \times 5 \times 3\text{m}^3$. A 2×2 case is used, for which the microphone positions are [3.48, 2.50, 1.50]m and [3.52, 2.50,

1.50]m respectively. The STFT length is set to be 1024. The separation performance is evaluated objectively by SIR and SDR [51].

As for the selection of the degrees of freedom v, it is a very difficult problem because what can be served is the speech mixtures instead of individual clean speech signal. With v increasing, the tails of the distribution will become lighter. The Gaussian distribution is the limiting case of the student's t distribution as $v \to \infty$. Thus, v should not be a large value. The separation performance of algorithms with different small v values is tested, the performance is essentially the same. In this section, v is set to be 4 for all the experiments.

6.5.1 Experiment in Low Reverberation Room Environment

In the first experiment, the separation performance of NG-IVA algorithm with the proposed source prior in a low reverberation room environment is tested. The reverberation time RT60 is set to be 200ms. Two different speech signals are chosen randomly from the TIMIT dataset and convolved into two mixtures. Then the orignal NG-IVA algorithm and the NG-IVA algorithm with the new source prior are used to separate the mixtures respectively. Then the source positions are changed to repeat the simulation. For every pair of speech signals, three different azimuth angles for the sources relative to the normal to the microphone array are set for testing, these angles are selected from 30, 45, 60 and -30 degrees. After that, another pair of speech signals is chosen to repeat the above simulations. The separation performance for each pair of speech signals is calculated by averaging the performance in different positions. Table 6.1 and Table 6.2 show the separation performance for ten different pairs of speech signals in terms of SDR and SIR respectively.

50 different mixtures are also formed in total from the TIMIT database to test the separation performance, and the average SDR and SIR improve-

Table 6.1. Separation performance comparison in SDR

mixtures	original(dB)	proposed(dB)	improvement(dB)
mixture1	18.81	20.12	1.31
mixture2	15.94	17.26	1.32
mixture3	9.97	11.73	1.76
mixture4	11.68	12.40	0.72
mixture5	18.80	19.91	1.11
mixture6	12.27	18.74	6.47
mixture7	8.88	11.10	2.22
mixture8	15.57	17.09	1.52
mixture9	18.10	19.50	1.4
mixture10	16.84	19.65	2.81

Table 6.2. Separation performance comparison in SIR

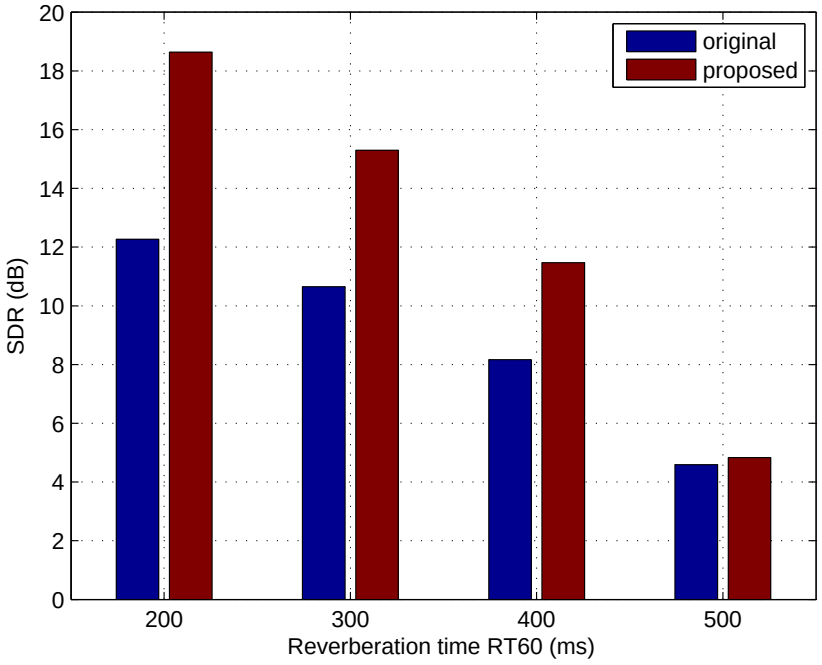
mixtures	original(dB)	proposed(dB)	improvement(dB)
mixture1	20.30	21.43	1.13
mixture2	17.88	19.00	1.12
mixture3	12.08	12.77	0.69
mixture4	14.42	14.97	0.55
mixture5	20.28	20.95	0.67
mixture6	14.08	20.94	6.86
mixture7	10.72	12.57	1.85
mixture8	16.98	18.77	1.79
mixture9	20.14	20.80	0.66
mixture10	19.53	21.54	2.01

ments are 1.3dB and 1.1dB respectively. These improvements confirm the advantage of the IVA algorithm with the proposed source prior in a low reverberation room environment.

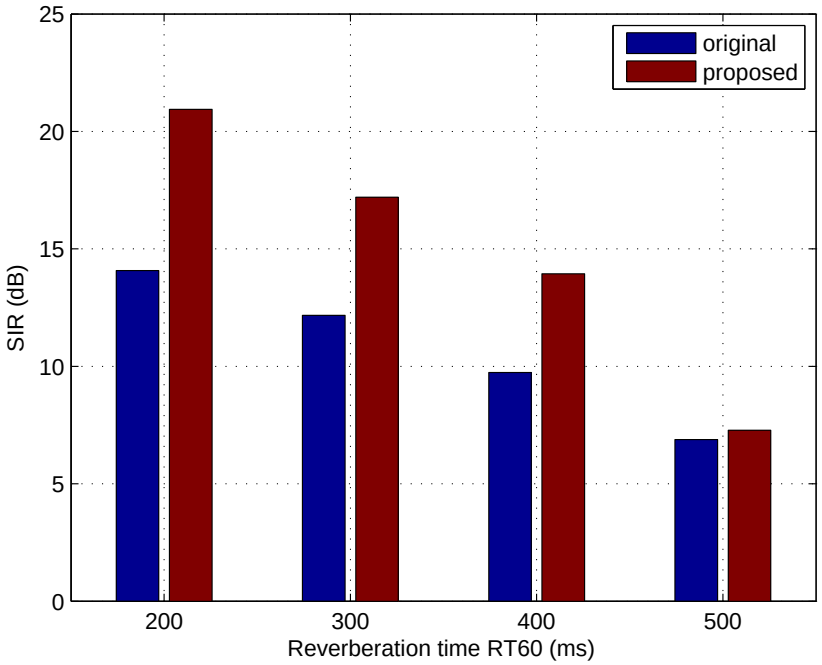
6.5.2 Experiment in Different Reverberant Room Environments

Then the separation performance of the NG-IVA algorithm with the proposed source prior in different reverberant room environments are tested. The reverberation time RT60 is set to be 200, 300, 400 and 500ms. All the

experimental settings and procedures are the same as the first experiment. Again the speech signals are selected from the TIMIT dataset to generate mixtures. Five different pairs of mixtures are selected randomly to do the experiments, and separation results in terms of SDR and SIR are shown from Fig. 6.12 to Fig. 6.16. The rad bar represents the separation performance of the proposed algorithm, and the blue bar represents the separation performance of the original NG-IVA algorithm. The x axis is the reverberation time RT60, and the y axis is SDR or SIR. It is shown that the red bar is always higher than the blue bar in different room environments and by using different mixtures, which means a better separation performance. The figures confirm that the IVA algorithm with the proposed multivariate source prior can consistently improve the separation performance in different reverberant room environments. However, the improvement reduces with increasing reverberation time due to the greater challenge in extracting the individual source vectors.

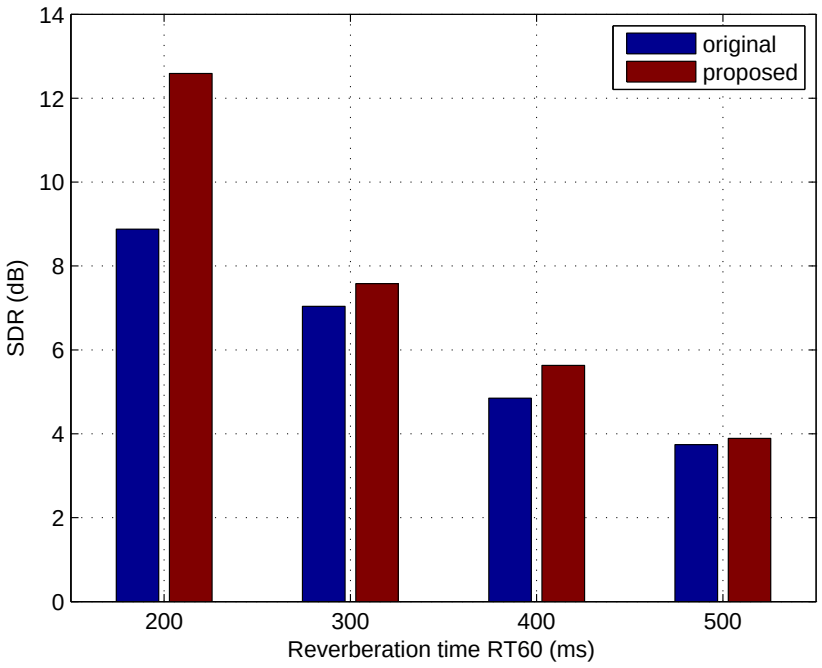


(a)

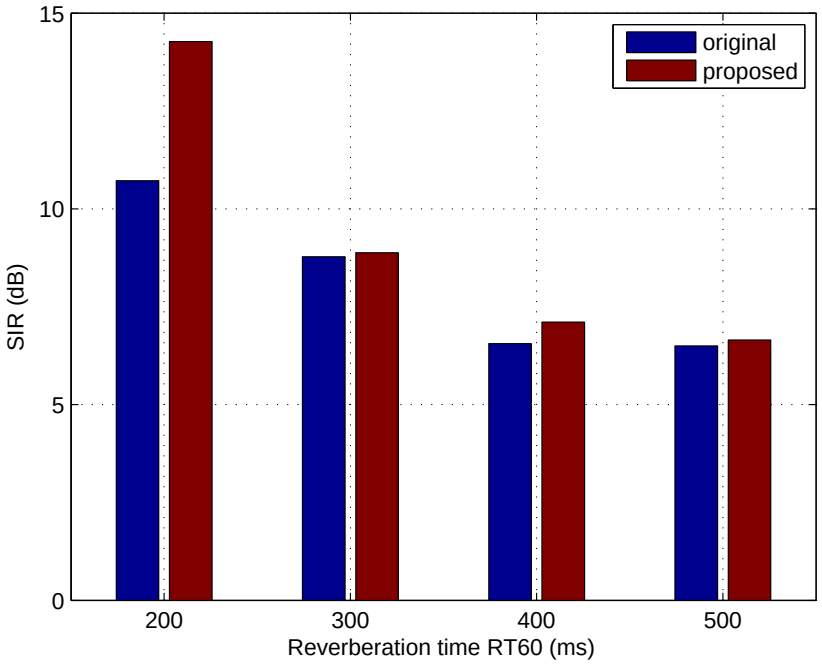


(b)

Figure 6.12. The separation performance in different reverberant environment for mixtures 1 (a) SDR (b) SIR

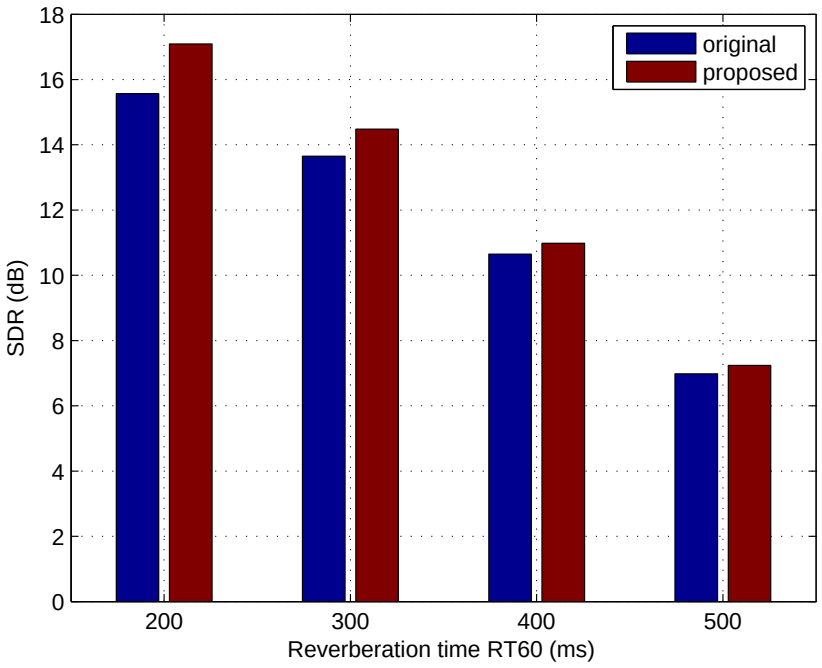


(a)

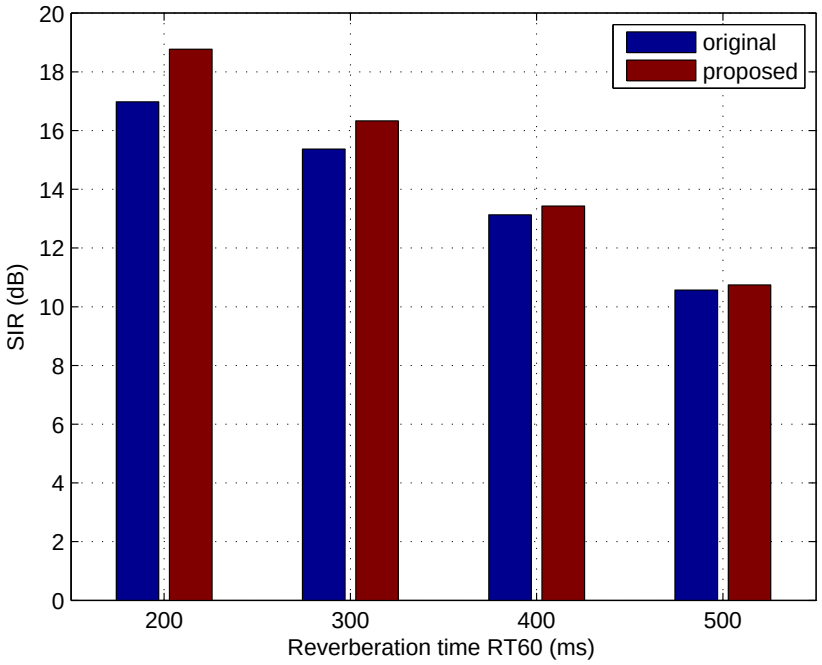


(b)

Figure 6.13. The separation performance in different reverberant environment for mixtures 2 (a) SDR (b) SIR

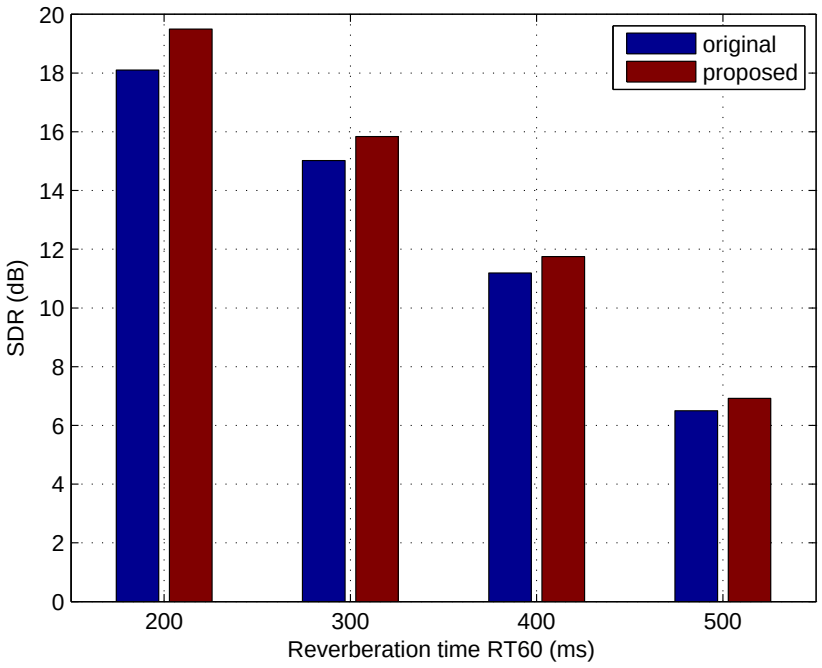


(a)

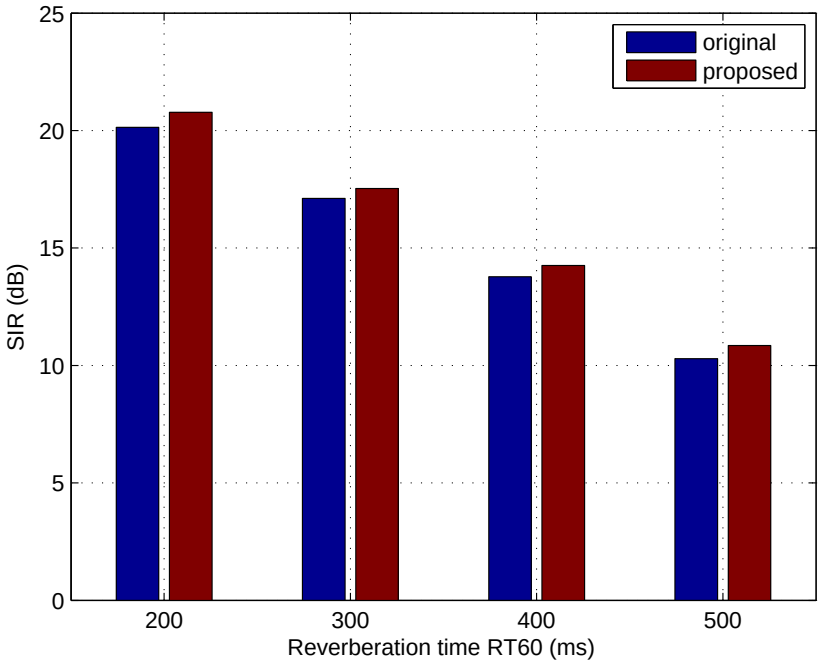


(b)

Figure 6.14. The separation performance in different reverberant environment for mixtures 3 (a) SDR (b) SIR

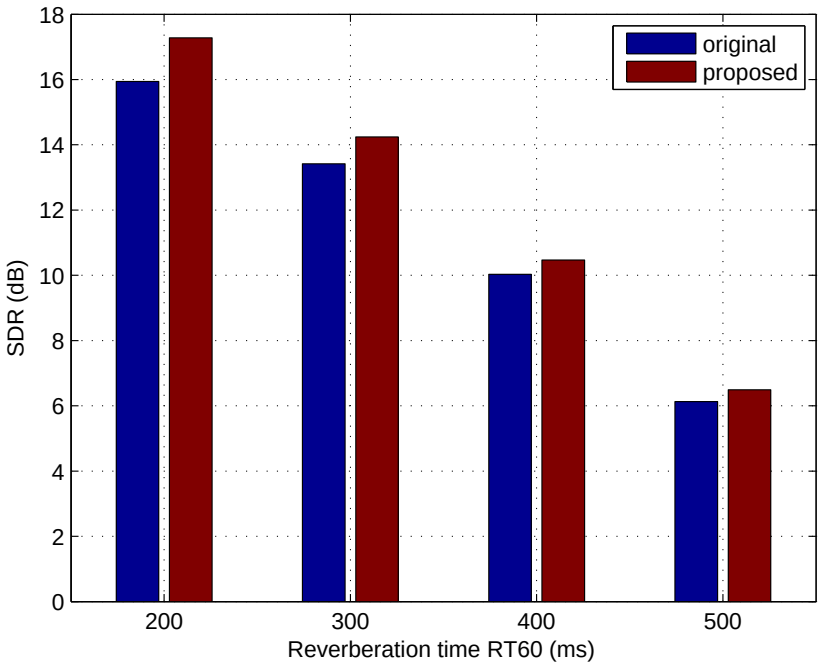


(a)

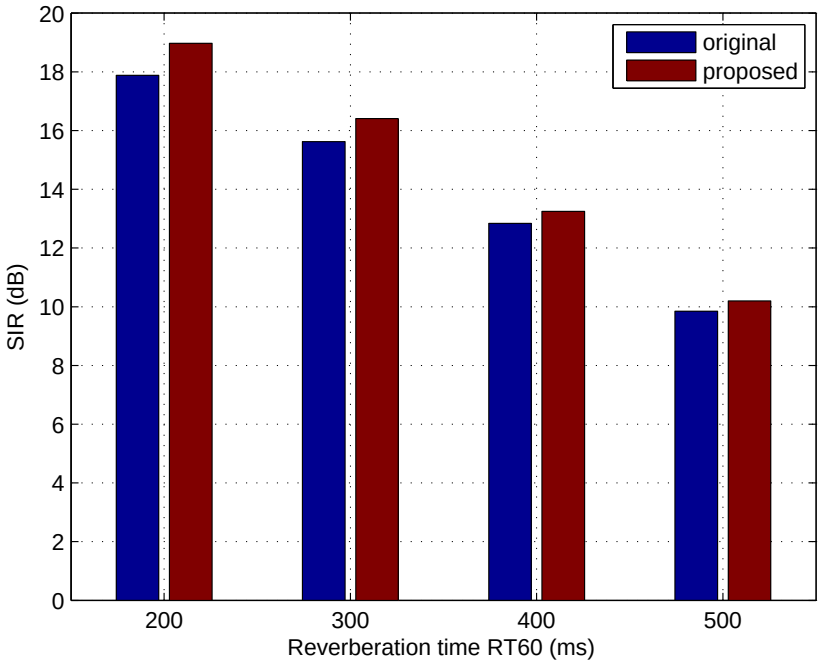


(b)

Figure 6.15. The separation performance in different reverberant environment for mixtures 4 (a) SDR (b) SIR



(a)



(b)

Figure 6.16. The separation performance in different reverberant environment for mixtures 5 (a) SDR (b) SIR

6.5.3 Experiment by Using the Real Room Recordings

In the last experiment, the real room recordings AV16.3 [39] corpus is used to compare the separation performance of original NG-IVA and NG-IVA with proposed source prior. The "seq37-3p-0001" recording is used to perform the experiment, which contains three speakers. The room environment has already been shown in the experiment section of Chapter 4. Three microphones (mic3, mic5 and mic7) from microphone array 1 are chosen to collect the mixtures. The audio sampling frequency of the recording is 16kHz. The RT60 is approximately 700ms, which means that it is a high reverberant environment.

The recorded speech is extracted from 210s to 215s, during which three speakers are speaking simultaneously. This multi-speaker speech separation problem is tried to be solved by using the original NG-IVA and NG-IVA with the proposed source prior. As for the performance evaluation, the information about the mixing matrix and sources are both missing, thus it is impossible to use the traditional SDR and SIR criteria. Thus the pitch based evaluation method proposed in Chapter 4 is adopted. If the speech signals are mixed, the pitches are also mixed as shown in Fig. 6.17. If the mixtures are separated, the pitches are separated as well as shown in Fig. 6.18. The pitch based evaluation method can also provide an objective evaluation criterion, i.e. the separation rate, which can be used to compare the separation performance of different algorithms. The bigger the separation rate, the more pitches are separated, which indicates a better separation. Fig. 6.17 and Fig. 6.18 show the pitch tracks of the mixtures and the pitches of separated signals by using IVA with the proposed source prior. It's hard to observe the difference when using two NG-IVA algorithms by comparing the pitches of the separated signals. The pitch track figure of the separated signals is omitted when using original IVA, and the separation rate is used to compare the separation performance when using different IVA algorithms

as shown in Table 6.3. The experimental results show that the proposed method can also achieve improvement by using real room recordings to solve the multi-speaker speech separation problem.

Table 6.3. Separation rate comparison when using real room recordings

	mixtures	original	proposed
separation rate	0.0379	0.2515	0.2794

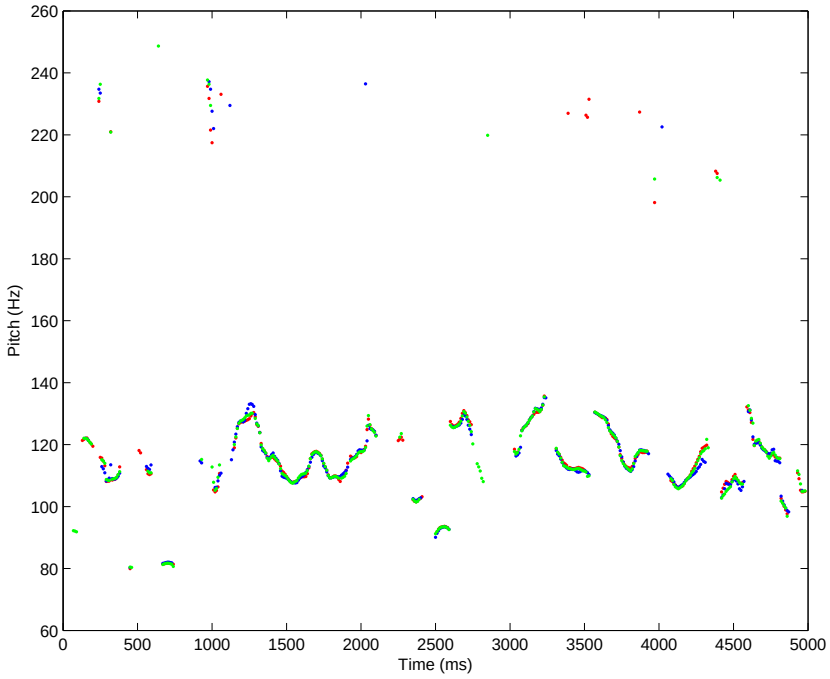


Figure 6.17. The time-varying pitch tracks of the mixtures

6.6 Summary

In this chapter, the dependency structure within the frequency domain speech signals was further exploited by introducing copula theory. The t copula was found suitable to model the inter-frequency dependency, which

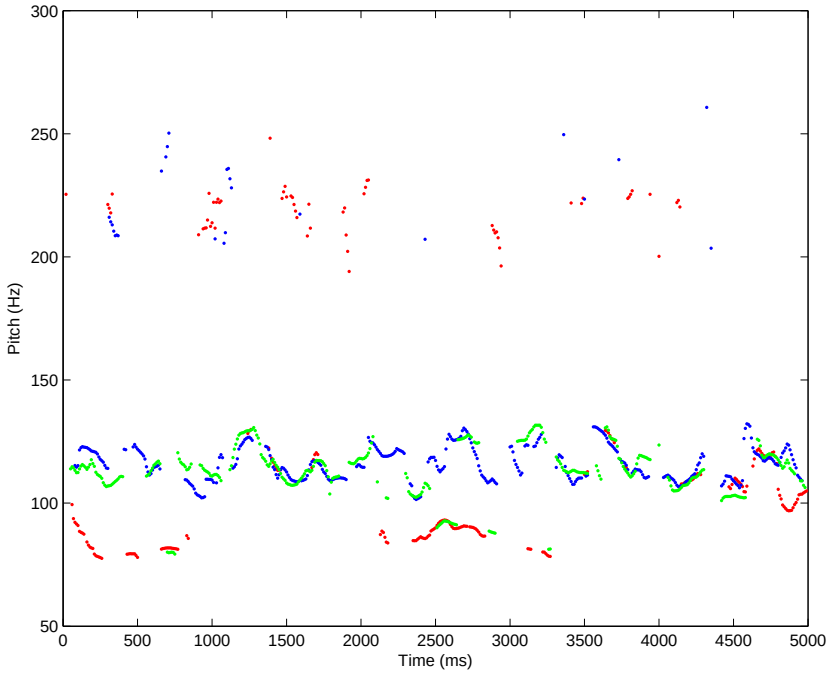


Figure 6.18. The time-varying pitch tracks of the separated signals by using IVA algorithm with proposed source prior

was also confirmed by observing the Chi-plot between two frequency bins of a real speech signal. Then, a multivariate student's t distribution was constructed by using the t copula density function and univariate student's t marginal distribution, which was adopted as the new source prior for the NG-IVA algorithm. The separation performance was tested in different reverberant room environments and also by using real room recordings. All the experimental results confirmed the advantage of this proposed source prior.

Chapter 7

CONCLUSIONS AND FUTURE WORK

The contributions of this thesis are summarized below, followed by a discussion on future works.

7.1 Conclusions

This study has provided enhanced independent vector analysis algorithms for audio separation in a room environment. The contributions of this study satisfy the three objectives mentioned in the introduction chapter. The first contribution is improving the convergence speed of the natural gradient IVA algorithm. The second contribution is highlighting the specific block permutation problem and proposing corresponding robust solutions. The third contribution is improving the separation performance by adopting a new source prior to preserve the inter-frequency dependency within the frequency domain speech signals. The details of the contributions are as follows:

In Chapter 2, besides the introduction of the fundamental knowledge of CBSS and ICA, the original natural gradient IVA algorithm, fast fixed-point IVA and auxiliary function based IVA were all discussed. Moreover, an adaptive step size natural gradient IVA algorithm was proposed in this chapter, which can automatically tune the step size to achieve a faster convergence compared with IVA. The proposed algorithm can save almost half of the

iteration numbers to converge compared with original NG-IVA algorithm.

In Chapter 3, the specific block permutation problem is discussed by analyzing the cost function. Then two kinds of solutions were proposed to solve this problem. Firstly, a robust IVA algorithm was proposed to address this problem by exploiting the phase continuity of the unmixing matrix to adjust the misalignments and thereby keep the permutation consistent across all the frequency bins. Then, an overlapped chain type dependency structure was also proposed to mitigate this problem. The experimental results confirmed that when the block permutation happened, the separation performance was poor, the SDR and SIR were negative values or small positive values. When the proposed methods were used, the block permutation problem was mitigate and a good separation performance can be achieved. Moreover, even when there was no block permutation problem, the separation performance can still be improved about 1.3dB and 3.0dB respectively by using the robust IVA algorithm. When the chain type dependency structure was applied to AuxIVA, the iteration numbers can be saved approximately 20% compared with the original AuxIVA algorithm.

In Chapter 4, the informed IVA scheme was proposed, which introduced the geometric information captured from video to combine with the FastIVA algorithm. The geometric information was used to set a smart initialization for optimization problem. The proposed scheme can not only make FastIVA more robust in terms of avoiding the block permutation problem, but also improve convergence speed and separation performance. The experimental results indicated that the improvement in noisy and reverberant room environment were approximately 0.75dB and 1.4dB in terms of SDR and SIR respectively compared with the original FastIVA algorithm. Moreover, a pitch based objective evaluation method was also proposed for evaluating the separation performance when using real room recordings.

In Chapter 5, a particular multivariate generalized Gaussian distribution

was proposed to be the source prior for IVA. The nonlinear score function derived from this particular source prior contained fourth order cross terms, which introduced extra relationships between different frequency bins and improved the dependency structure to achieve a better separation performance. The proposed source prior was applied to NG-IVA, FastIVA and AuxIVA, and experimental results confirmed the advantage of the proposed source prior. When the new source prior was applied to the NG-IVA algorithm, the SDR and SIR improvement were approximately 0.9dB and 0.8dB respectively. When it was applied to the FastIVA algorithm, both the SDR and SIR improvement were approximately 0.6dB. Finally, when AuxIVA adopted the new source prior, the improvement was 1.7dB and 1.9dB in terms of SDR and SIR respectively.

In Chapter 6, the dependency structure within the frequency domain speech signals was researched, and the t copula was found to be suitable to describe this dependency structure. Then, the multivariate student's t distribution is constructed by using a t copula with univariate student's t marginal distribution, and the NG-IVA with this multivariate student's t source prior was derived. The experimental results showed that the proposed method can consistently achieve improved separation performance in different reverberant room environments. The average improvement in terms of SDR and SIR were 1.3dB and 1.1dB respectively compared with the original NG-IVA algorithm.

7.2 Future Work

In order to further improve this study, there are several topics which could be further researched.

Firstly, in order to improve IVA, other dependency structure needs to be exploited. With a stronger dependency structure, the block permutation

problem may be mitigated, which will make the separation performance of IVA more robust. Moreover, an improved dependency structure will also potentially help to improve the separation performance of IVA.

Secondly, the theoretical link between the dependency structure and the separation performance is still missing. It will be helpful to exploit how the dependency structure affects the separation performance. This knowledge could be exploited in future algorithm design.

Thirdly, as for the multivariate student's t source prior, the selection of the degrees of freedom parameter should be studied. Although there are several methods to estimate this parameter for a pure speech signal, such as the tail index estimation method [88], it is difficult to estimate this parameter by using only the mixtures. A potential solution for this problem is to roughly separate the mixtures at the first stage, then estimate the degrees of freedom for each source.

Fourthly, the combination of IVA algorithms and CASA based methods such as the time frequency masking technique can potentially further improve the separation performance. Recently, several such combinations have been proposed such as [9] [10]. Moreover, this combination scheme can also be used to solve the underdetermined case by using the time frequency representations to exploit the number of sources and the direction of arrival information.

Finally, IVA still suffers from the challenging problem of a high reverberant room environment [89]. There are several dereverberation methods that can be considered to combine with the IVA algorithm to mitigate this problem. Beamforming is widely used for dereverberation in the field of speech processing. Thus, using beamforming as a pre-processing stage to dereverberate the speech signal seems to be a potential solution to deal with this problem [90]. Linear prediction is another popular dereverberation technique. Many related methods have been published [91] [92] [93]. However,

few of them have been used for the blind source separation field, because they all focus on the situations that there is only one source in the measurements. For the cocktail party problem, the number of sources is at least two, which makes it is difficult to use the linear prediction method to dereverberate the speech mixtures. As for the combination, the pre-processing will affect the room impulse response and change the original speech mixtures, sometimes it will make the speech mixtures are unsuitable for IVA algorithms at the second stage. Thus, how to design a joint optimization algorithm to combine the dereverberation stage with IVA to achieve a good performance in highly reverberant environment is open to future study [94].

References

[1] C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of The Acoustical Society of America*, vol. 25, pp. 975–979, 1953.

[2] C. Cherry and W. Taylor, "Some further experiments upon the recognition of speech, with one and with two ears," *The Journal of The Acoustical Society of America*, vol. 26, pp. 554–559, 1954.

[3] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, pp. 1875–1902, 2005.

[4] A. S. Bregman, *Auditory scene analysis: the perceptual organization of sound*. Cambridge: MIT Press, 1990.

[5] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *IEEE Transactions on Circuits and Systems I*, vol. 57, pp. 1431–1438, 2010.

[6] S. M. Naqvi, M. Yu, and J. A. Chambers, "A multimodal approach to blind source separation of moving sources," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, pp. 895–910, 2010.

[7] M. Cooke and D. Ellis, "The auditory orgnization of speech and other sources in listeners and computational models," *Speech Communication*, vol. 35, pp. 141–177, 2001.

[8] D. Wang, "Time frequency masking for speech separation and its potential for hearing aid design," *Trends in Amplification*, vol. 12, pp. 332–353, 2008.

[9] T. Jan, W. Wang, and D. Wang, "A multistage approach to blind separation of convolutive speech mixtures," *Speech Communication*, vol. 53, pp. 524–539, 2011.

[10] F. S. P. Clark, M. R. Petraglia, and D. B. Haddad, "A new initialization method for frequency-domain blind source separation algorithms," *IEEE Signal Processing Letters*, vol. 18, pp. 343–346, 2011.

[11] S. Haykin et al., *Unsupervised Adaptive Filtering (Volume I: Blind Source Separation)*. Wiley, 2000.

[12] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing: learning algorithms and applications*. Wiley, 2003.

[13] S. Haykin et al., *New directions in Statistical Signal Processing: From Systems to Brain*. Cambridge, MA: MIT Press, 2007.

[14] "iPhone user guide," *Apple Inc*, pp. 36–42, 2013.

[15] J. Herault, C. Jutten, and B. Ans, "Detection de grandeurs primitives dans un message composite par une architecture de calcul neuromimetique en apprentissage non supervis," in *Proc. GRETSI*, (Nice, France), 1985.

[16] P. Comon, "Independent component analysis - a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[17] M. S. Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," *Springer Handbook on Speech Processing and Speech Communication*, 2007.

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

[19] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis.* Wiley, 2001.

[20] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications.* Academic Press, 2009.

[21] S. M. Naqvi, Y. Zhang, T. Tsalaile, S. Sanei, and J. A. Chambers, "A multimodal approach for frequency domain independent component analysis with geometrically-based initialization," in *Proc. EUSIPCO*, (Lausanne, Switzerland), 2008.

[22] B. Rivet, L. Girin, and C. Jutten, "Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, pp. 96–108, 2007.

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

[24] L. Parra and C. Alvino, "Geometric source separation: merging convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 352–362, 2002.

[25] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, pp. 530–538, 2004.

[26] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," in *Proc. IWIANN*, (University of Laguna, Tenerife, Spain), 1998.

[27] T. Kim, H. Attias, S. Lee, and T. Lee, "Blind source separation exploiting

higher-order frequency dependencies," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 70–79, 2007.

[28] T. Kim, I. Lee, and T.-W. Lee, "Independent vector analysis: definition and algorithms," in *Fortieth Asilomar Conference on Signals, Systems and Computers 2006*, (Asilomar, USA), 2006.

[29] I. Lee, T. Kim, and T.-W. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Processing*, vol. 87, pp. 1859–1871, 2007.

[30] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *2011 IEEE WASPAA*, (New Paltz, USA), 2011.

[31] A. Masnadi-Shirazi, W. Zhang, and B. D. Rao, "Glimpsing IVA: A framework for overcomplete/complete/undercomplete convolutive source separation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, pp. 1841–1855, 2010.

[32] T. Ono, N. Ono, and S. Sagayama, "User-guided independent vector analysis with source activity tuning," in *ICASSP 2012*, (Kyoto, Japan), 2012.

[33] I. Lee, G.-J. Jang, and T.-W. Lee, "Independent vector analysis using densities represented by chain-like overlapped cliques in graphical models for separation of convolutedly mixed signals," *Electronic Letters*, vol. 45, pp. 710–711, 2009.

[34] C. H. Choi, W. Chang, and S. Y. Lee, "Blind source separation of speech and music signals using harmonic frequency dependent independent vector analysis," *Electronic Letters*, vol. 48, pp. 124–125, 2012.

[35] I. Lee, J. Hao, and T. W. Lee, "Adaptive independent vector analysis for the separation of convoluted mixtures using EM algorithm," in *ICASSP 2008*, (Las Vegas, U.S.A.), 2008.

[36] J. Hao, I. Lee, and T. W. Lee, "Independent vector analysis for source separation using a mixture of Gaussian prior," *Neural Computation*, vol. 22, pp. 1646–1673, 2010.

[37] M. Anderson, T. Adali, and X.-L. Li, "Joint blind source separation with multivariate Gaussian model: algorithms and performance analysis," *IEEE Trans. on Signal Processing*, vol. 60, pp. 1672–1682, 2012.

[38] J. S. Garofolo et al., "TIMIT acoustic-phonetic continuous speech corpus," in *Linguistic Data Consortium*, (Philadelphia), 1993.

[39] G. Lathoud, J. Odobez, and D. Gatica-Perez, "AV16.3: an audio-visual corpus for speaker localization and tracking.," in *Proceedings of the MLMI'04 Workshop*, 2004.

[40] T. W. Lee, *Independent Component Analysis: Theory and Applications*. Kluwer Academic, 2000.

[41] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 516–527, 2011.

[42] K. Matsuoka and S. Nakashima, "Minimal distrotion principle for blind source separation," pp. 722–727, 2001.

[43] L. Yuan, W. Wang, and J. Chambers, "Variable step-size sign natural gradient algorithm for sequential blind source separation," *IEEE Signal Processing Letters*, vol. 12, pp. 589–592, 2005.

[44] J. Chambers, M. Jafari, and S. McLaughlin, "Variable step-size EASI algorithm for sequential blind source separation," *Electronics Letters*, vol. 40, pp. 393–394, 2004.

[45] E. Bingham and A. Hyvarinen, "A fast fixed point algorithm for independent component analysis of complex valued signals," *Int. J. Neural Netw.*, vol. 10, pp. 1–8, 2000.

[46] N. Ono and S. Miyabe, "Auxiliary-function-based independent component analysis for super-Gaussian source," in *LVA/IVA 2010*, (St. Malo, France), 2010.

[47] T. Itahashi and K. Matsuoka, "Stability of independent vector analysis," *Signal Processing*, vol. 93, pp. 1809–1820, 2012.

[48] F. Nesta, P. Svaizer, and M. Omologo, "A BSS method for short utterances by a recursive solution to the permutation problem," in *Proc. SAM*, (Darmstadt, Germany), 2008.

[49] F. Nesta, P. Svaizer, and M. Omologo, "Convolutive BSS of short mixtures by ICA recursively regularized across frequencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 624–639, 2011.

[50] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[51] E. Vincent, C. Fevotte, and R. Gribonval, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1462–1469, 2006.

[52] J. L. Schwartz, F. Berthommier, and C. Savariaux, "Seeing to hear better: evidence for early audio-visual interactions in speech identification," *Cognition*, vol. 93, pp. 69–78, 2004.

[53] H. K. Maganti, D. Gatica-Perez, and I. McCowan, "Speech enhancement and recognition in meetings with an audio-visual sensor array," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2257–2269, 2007.

[54] K. W. Grant and P. Seitz, "The use of visible speech cues for improving auditory detection of spoken sentences," *Journal of the Acoustical Society of America*, vol. 108, pp. 1197–1208, 2000.

[55] K. Wilson, N. Checka, D. Demirdjian, and T. Darrell, "Audio-video array source separation for perceptual user interfaces," *in Proc. Workshop on Perceptive User Interfaces, Orlando, Florida*, pp. 1–7, 2001.

[56] W. H. Sumby and I. Pollack, "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, vol. 26, no. 2, pp. 212–215, 1954.

[57] D. Sodoyer, J. Schwartz, L. Girin, J. Klinkisch, and C. Jutten, "Separation of audio-visual speech sources: a new approach exploiting the audio-visual coherence of speech stimuli," *EURASIP J. Appl. Signal Process.*, vol. 2002, pp. 1165–1173, 2002.

[58] D. Sodoyer, L. Girin, C. Jutten, and J. Schwartz, "Developing an audio-visual speech source separation algorithm," *Speech Communication*, vol. 44, no. 1-4, pp. 113–125, 2004.

[59] B. Shinn-Cunningham, N. Kopco, and T. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *Journal of the Acoustical Society of America*, vol. 117, pp. 3100–3115, 2005.

[60] T. Gustafsson, B. Rao, and M. Trivedi, "Source localization in reverberant environments: modeling and statistical analysis," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 791–803, 2003.

[61] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, no. 114, pp. 2236–2252, 2003.

[62] W. Wang, D. Cosker, Y. Hicks, S. Sanei, and J. A. Chambers, "Video assisted speech source separation," in *Proc. ICASSP*, (Philadelphia, U.S.A), 2005.

[63] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[64] E. Petajan, B. Bischoff, D. Bodoff, and N. M. Brooke, "An improved automatic lipreading system to enhance speech recognition," *in Proc. SIGCHI conference on Human factors in computing systems, Washington, D.C., United States*, pp. 19–25, 1988.

[65] S. M. R. Naqvi in *Multimodal Methods for Blind Source Separation of Audio Sources*, (PhD thesis, Loughborough University, UK), 2009.

[66] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE Journal of Robotics and Automation*, vol. RA-3, pp. 323–344, 1987.

[67] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2001.

[68] H. Shabani and M. H. Kahaei, "Missing feature mask generation in BSS outputs using pitch frequency," in *17th International Conference on Digital Signal Processing*, (Corfu, Greece), 2011.

[69] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of The Acoustical Society of America*, vol. 124, pp. 1638–1652, 2008.

[70] A. Camacho and J. G. Harris, "A pitch estimation algorithm based on the smooth harmonic average peak-to-valley envelope," *in Proc. the International Symposium on Circuits and Systems*, pp. 3940–3943, 2007.

[71] M. M. Sondhi, "New methoeds of pitch extraction," *IEEE Transactions on Audio Electroacoustics*, vol. AU-16, pp. 262–266, 1968.

[72] D. J. Hermes, "Measurement of pitch by subharmonic summation," *Journal of the Acoustical Society of America*, vol. 83, pp. 257–264, 1988.

[73] H. Duifhuis, L. F. Willems, and R. J. Sluyter, "Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception," *Journal of the Acoustical Society of America*, vol. 71, pp. 1568–1580, 1982.

[74] I. Lee and T. W. Lee, "On the assumption of spherical symmetry and sparseness for the frequency-domain speech model," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, pp. 1521–1528, 2007.

[75] A. Hyvärinen, "Independent component analysis: recent advances," *Philos Transact A Math Phys Eng Sci*, vol. 371(1984), pp. 1–19, 2012.

[76] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[77] R. B. Nelsen, *An Introduction to Copulas*. USA: Springer, 2006.

[78] G. Papaefthymiou and D. Kurowicka, "Using copulas for modeling stochastic dependence in power system uncertinty analysis," *IEEE Transactions on Power Systems*, vol. 24, pp. 40–49, 2009.

[79] Y. Stitou, N. Lasmar, and Y. Berthoumieu, "Copulas based multivariate gamma modeling for texture classification," in *Proc. ICASSP 2009*, (Taipei, China), 2009.

[80] C. Fevotte and J. Godsil, "A Bayesian approach for blind separation of sparse sources," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 2174–2188, 2006.

[81] H. Sundar, C. S. Seelamantula, and T. Sreenivas, "A mixture model approach for formant tracking and the robustness of student's t distribution," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, pp. 2626–2636, 2012.

[82] T. Schmidt, *Coping with Copulas*. USA: Springer, 2006.

[83] S. Demarta and A. J. McNeil, "The t copula and related copulas," *International Statistical Review*, vol. 73, pp. 111–129, 2005.

[84] W. G. Gilchrist, *Statistical Modelling with Quantile Functions*. Chapman and Hall, 2000.

[85] S. Daul, E. D. Giorgi, F. Lindskog, and A. McNeil, "The grouped t-copula with an application to credit risk," *RISK*, vol. 16, pp. 73–76, 2003.

[86] N. I. Fisher and P. Switzer, "Chi-plots for assessing dependence," *Biometrika*, vol. 72, pp. 253–265, 1985.

[87] N. I. Fisher and P. Switzer, "Graphical assessment of dependence: is a picture worth 100 tests?," *The American Statistician*, vol. 55, pp. 233–239, 2001.

[88] R. Huisman, K. G. Koedijk, J. M. C. Kool, and F. Palm, "Tail-index estimate in small samples," *Journal of Business and Economic Statistics*, vol. 19, pp. 208–216, 2001.

[89] 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," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 109–116, 2003.

[90] L. Wang, H. Ding, and F. Yin, "Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–13, 2010.

[91] T. Nakatani, T. Yoshioka, K. Kinoshita, and M. Miyoshi, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *Proc. ICASSP*, (Las Vegas, U.S.A), 2008.

[92] M. Delcroix, T. Hikichi, and M. Miyoshi, "Precise dereverberation using multichannel linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 430–440, 2007.

[93] M. Delcroix, T. Hikichi, and M. Miyoshi, "Dereverberation and denosing using multichannel linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1791–1801, 2007.

[94] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 69–84, 2011.