Enhanced Independent Vector Analysis for Speech Separation in Room Environments

by

Waqas Rafique

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

February 2017

Communications, Sensors, Signal and Information Processing Research Group (ComS^{2}IP),
School of Electrical and Electronic Engineering,
Newcastle University,
Newcastle upon Tyne, UK, NE1 7RU.

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 and siblings.

Abstract

The human brain has the ability to focus on a desired sound source in the presence of several active sound sources. The machine based method lags behind in mimicking this particular skill of human beings. In the domain of digital signal processing this problem is termed as the cocktail party problem. This thesis thus aims to further the field of acoustic source separation in the frequency domain based on exploiting source independence. The main challenge in such frequency domain algorithms is the permutation problem. Independent vector analysis (IVA) is a frequency domain blind source separation algorithm which can theoretically obviate the permutation problem by preserving the dependency structure within each source vector whilst eliminating the dependency between the frequency bins of different source vectors. This thesis in particular focuses on improving the separation performance of IVA algorithms which are used for frequency domain acoustic source separation in real room environments.

The source prior is crucial to the separation performance of the IVA algorithm as it is used to model the nonlinear dependency structure within the source vectors. An alternative multivariate Student's t distribution source prior is proposed for the IVA algorithm as it is known to be well suited for modelling certain speech signals due to its heavy tail nature. Therefore the nonlinear score function that is derived from the proposed Student's t source prior can better model the dependency structure within the frequency bins and thereby enhance the separation performance and the convergence speed of the IVA and the Fast version of the IVA (FastIVA) algorithms.

A novel energy driven mixed Student's t and the original super Gaussian source prior is also proposed for the IVA algorithms. As speech signals can be composed of many high and low amplitude data points, therefore the Student's t distribution in the mixed source prior can account for the high amplitude data points whereas the original super Gaussian distribution can cater for the other information in the speech signals. Furthermore, the weight of both distributions in the mixed source prior can be adjusted according to the energy of the observed mixtures. Therefore the mixed source prior adapts the measured signals and further enhances the performance of the IVA algorithm.

A common approach within the IVA algorithm is to model different speech sources with an identical source prior, however this does not account for the unique characteristics of each speech signal. Therefore dependency modelling for different speech sources can be improved by modelling different speech sources with different source priors. Hence, the Student's t mixture model (SMM) is introduced as a source prior for the IVA algorithm. This new source prior can adapt according to the nature of different speech signals and the parameters for the proposed SMM source prior are estimated by deriving an efficient expectation maximization (EM) algorithm. As a result of this study, a novel EM framework for the IVA algorithm with the SMM as a source prior is proposed which is capable of separating the sources in an efficient manner.

The proposed algorithms are tested in various realistic reverberant room environments with real speech signals. All the experiments and evaluation demonstrate the robustness and enhanced separation performance of the proposed algorithms.

Contents

Statement of Originality

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

In Chapter 4, the dependency structure within frequency domain speech signals is exploited and a new multivariate Student's t source prior is proposed for the FastIVA method. The proposed source prior can better model the nonlinear dependency structure due to the heavy tailed nature of the Student's t distribution. Therefore, this new source prior improves the separation performance and convergence speed of the FastIVA method in the real room environments. The work was published in:

1. W. Rafique, S. M. Naqvi, P. J. B. Jackson and J. A. Chambers, 'IVA algorithms using a multivariate Student's t source prior for speech source separation in real room environments', International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, pp. 474-478, 2015.

2. W. Rafique, S. Erateb, S. M. Naqvi and J. A. Chambers, 'Evaluation of Source Separation Algorithms, including the IVA algorithm with various source priors, using Binaural Room Impulse Responses', 10th International IMA Conference, Birmingham, UK, 2014.

In Chapter 5, a mixture of the multivariate Student's t and the original super Gaussian distribution is adopted as the source prior for the IVA and the FastIVA algorithms. In the mixed source prior, the Student's t distribution due to its heavy tailed nature can better model the high amplitude information in the speech signals and the super Gaussian distribution can be used to model the rest of the information. Firstly, equal weights were assigned to both distributions in the mixed source prior. Then in order to adapt the mixed source prior according to different speech signals, the weight of both distributions in the mixed source prior was adjusted according to the energy of the observed signals and therefore this energy driven mixed source prior can adapt according to various observed mixture signals. The results are published in:

3. W. Rafique, S. M. Naqvi and J. A. Chambers, 'Speech source separation using the IVA algorithm with multivariate mixed super gaussian student's t source prior in real room environment', in Proc. IET Intelligent Signal Processing (ISP) Conference, London, UK, 2015.

4. W. Rafique, S. M. Naqvi and J. A. Chambers, 'Mixed source prior for the fast independent vector analysis algorithm,' IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM), pp. 1-5, Rio de Janeiro, Brazil, 2016.

5. W. Rafique, S. Erateb, S. M. Naqvi, S. S. Dlay and J. A. Chambers, 'Independent vector analysis for source separation using an energy driven mixed Student's t and super Gaussian source prior', European Signal Processing Conference (EUSIPCO), pp. 858-862, Budapest, Hungry, 2016.

In Chapter 6, an efficient expectation maximization framework is derived for the IVA algorithm and in order to adapt the independent vector analysis algorithm with different speech signals, instead of a conventional single distribution source prior, a new Student's t mixture model is proposed as a source prior for the IVA method. The parameters of the mixture model for different speech signals were estimated with the expectation maximization method. A journal article is in preparation.

6. W. Rafique , S. M. Naqvi and J. A. Chambers, 'An efficient expectation maximization framework for independent vector analysis with a Student's t mixture model source prior for frequency domain blind source separation', journal article in preparation, 2016.

Acknowledgements

I owe a great deal of gratitude to a number of people who have helped me throughout my journey to complete this thesis. First and foremost, I deeply and sincerely thank my supervisor Professor Jonathon Chambers for his invaluable guidance, generous support and huge amounts of encouragement and advice over the past three years, from Loughborough to Surrey and to Newcastle. I have benefited tremendously from his remarkable knowledge, his exceptional enthusiasm and prompt feedback on my papers and reports. He was always available and returned my emails swiftly even in out-of-office hours and on weekends. It is my great privilege and exclusive honour to have been one of his research students.

I would also like to thank Dr. Mohsen Naqvi, for his guidance and assistance during my PhD, specifically at the beginning of my PhD studies. He helped me to enter my research field and was the source of constant support and advice for me.

I would also like to give my appreciation to my colleagues and friends, Atta-ur-Rehman, Yangfeng Liang, Adnan Akbar, Syed Sameed, Ali Sunny, Pengming Feng, Ali Alameer, Yang Sun, Zeyu Fu and Anam Khan for their constant support and help during the duration of my PhD.

Lastly, but most importantly I express my deepest indebtedness to my parents and siblings, who raised me, cared for me and frequently put me before themselves in many circumstance so I can always get the best. I really can not find appropriate words to express my heartfelt gratitude to them for their constant encouragement, attention and prayers throughout my educational career. They are the most important people in my life and I would like to dedicate this thesis to them.

Waqas Rafique

November, 2016

Nomenclature

BRIR	Binaural Room Impulse Response
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
CPP	Cocktail Party Problem
DFT	Discrete Fourier Transform
ECG	ElectroCardioGraphy
EEG	ElectroEncephaloGraphy
EMG	ElectroMyoGraphy
EM	Expectation Maximization
FastIVA	Fast fixed point Independent Vector Analysis
GMM	Gaussian Mixture Model
HOS	Higher Order Statistics
ICA	Independent Component Analysis
IVA	Independent Vector Analysis

NG	Natural Gradient
PCA	Principal Component Analysis
pdf	probability density function
PESQ	Perceptual Evaluation of Speech Quality
PI	Permutation Index
RIR	Room Impulse Response
SDR	Signal-to-Distortion Ratio
SMM	Student’s t Mixture Model
STFT	Short-Time Fourier Transform

List of Symbols

Some of the frequently used notations in this thesis are as follows:

$|.|$ Absolute value

$||.||_2$ Euclidean norm

$(.)^T$ Transpose operator

$(.)^\dagger$ Hermitian transpose operator

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

$\otimes$ convolution operator

$(.)^*$ Complex conjugate operator

$det(.)$ Matrix determinant operator

$E(.)$ Statistical expectation operator

$F(.)$ Nonlinear function for FastIVA

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

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

$\phi(.)$ Non-linear score function

$J(.)$ Cost function

I Identity matrix

k Frequency bin index

K Number of frequency bins

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

$\mathbf{\Sigma}$ Covariance matrix

$\mathbf{\Lambda}$ Precision matrix

G Overall system matrix

A Mixing matrix

W Estimated unmixing matrix

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Human beings process audio information during various activities in their daily life. On numerous occasions, human beings have to focus on a particular sound of interest in the presence of many unwanted and distracting sounds. A person with healthy hearing capability is capable of hearing and identify a particular sound of interest even in a crowded environment. This ability of the human hearing system to identify different acoustic sources and perform difficult acoustic tasks is yet to be fully understood. Numerous efforts in the past decades have been dedicated to understand the capabilities of humans and mapping these qualities to machines but it remains a difficult task. This problem of separating speech signals getting disturbed by surrounding voices is known as the cocktail party problem [1, 2]. The solution of the cocktail party problem is to build a method which can separate the desired speech source while suppressing all the background sound sources [3, 4].

Plenty of research has been conducted during the past few decades in the signal processing community to try and mimic the human hearing abilities in the machines. This research includes various techniques such as source localisation, source detection, source tracking and source separation. Computational auditory scene analysis (CASA) is the result of this research. CASA is known mainly in the computer science community and it is the research which approximates the human ability to localise and isolate the desired acoustic source in a crowded environment [5,6]. In the signal processing community, attempts to solve the machine cocktail party problem are known as blind source separation (BSS). In the past, attempts have been made to solve the cocktail party problem by using the combination of BSS and CASA [7]. The focus of this thesis is mainly on signal processing techniques such as blind source separation.

Figure 1.1: The cocktail party problem (Image courtesy: *Telegraph.co.uk*)

1.1 Blind Source Separation

Blind source separation (BSS) is a statistical technique that refers to the separation of speech sources when there is no a prior information available about the mixing process [8,9]. The BSS methods can generally separate different sound sources by exploiting their statistical properties.

In recent years, various approaches have been proposed to solve the blind source separation problem. One of the state-of-the-art topics in blind source separation is independent component analysis (ICA) which was introduced by Herault and Jutten in 1985 [10,11]. The ICA method maximizes the mutual independence between the source signals however it assumes the unknown mixing process to be instantaneous, which means the source signals are transmitted directly to the listeners without any time delays or reflections [12]. However, in realistic environments, there are reverberations and signals reaching the listeners contain time delayed versions of the original sources. Therefore the instantaneous model is not an appropriate solution for the cocktail party problem as it is an over simplification of the complex real room environment.

In realistic environments, instead of the single direct path, signals mostly take multiple paths to the sensors, therefore the convolutive mixing model is a more appropriate representation of practical scenarios. Generally, there are two types of convolutive model which are anechoic and echoic mixing models. The anechoic mixing model only considers the time delays between the source and sensors whereas the echoic model considers the time delays caused not only by the direct path, but also by the early reflections and late reverberations. The main focus of this thesis is the echoic model, since this model is a more accurate model of the real room environment. As the convolutive mixing model for BSS involves time delays and reverberation, each element of the mixing model is basically a linear filter in the time domain which has sufficient support to describe the multipath between sources and sensors.

The convolutive BSS problem is usually tackled in the frequency domain, mainly because in the time domain the room impulse responses are often on the order of thousands of samples which makes time domain methods computationally complex for convolutive BSS [14]. Since frequency domain methods convert the convolutive time domain problem into a simple multiplication in the frequency domain, computational cost is significantly reduced. In the frequency domain, the convolutive model can be converted into bin-wise instantaneous mixtures at each frequency bin, and the ICA algorithm can be implemented to separate the sources from the mixtures. When the ICA method is implemented bin-wise in the frequency domain, the final separation performance is generally influenced by the permutation problem and the scaling problem.

In the frequency domain ICA algorithm, the scaling problem can be mitigated by using matrix normalization [12, 13]. On the other hand, the permutation ambiguity is the major problem, and causes potential misalignment of sources at different frequency bins. Therefore, when the separated sources are reconstructed in the time domain, it affects the actual separation performance of the algorithm, which is generally poor. Various techniques have been proposed in the signal processing community to mitigate the permutation problem in frequency domain blind source separation [14]. Some of the techniques use localization information to improve the separation performance [15–17]. Other solutions include the use of both audio and video information to solve the permutation problem [18–20]. The main problem with these techniques is that they need pre or post processing which generally add extra complexity and latency in the system.

In order to solve the permutation problem, independent vector analysis (IVA) was proposed by Kim et al [21, 22]. The IVA method can theoretically avoid the permutation problem by maximizing the dependence within the frequency components of each source vector while maximizing the independence between different source vectors [21]. The IVA method exploits a multivariate dependency model to retain the dependency between frequency bins, instead of a univariate model that was used for the ICA method. Therefore it solves the permutation problem within the algorithm convergence process without the need of any pre or post processing techniques.

The main idea of the IVA algorithm is to preserve the dependency structure and the choice of the particular multivariate score function is crucial to its performance [21]. This multivariate score function is derived from the multivariate source prior; therefore selecting an appropriate source prior is very important to improve the performance of the IVA algorithm. The original IVA method uses a multivariate Laplacian distribution as the source prior for the IVA method in an attempt to solve permutation problem. However the improvement in the separation performance of the resulting IVA method is still needed. In order to improve the separation in the IVA method, a chain-like overlapped model has been adopted for modelling the dependency structure, since the dependencies between different frequency bins could be different [23]. Also, an auxiliary function based technique has been used within the IVA method to improve its convergence speed [24] and some other methods based on understanding and exploiting the source structure have been used to improve the separation performance of the IVA method [25–28, 45]. These methods are based on exploiting the harmonic frequency and power variations within the source signals but further improvement is needed in these techniques.

1.2 Applications of Blind Source Separation

Blind source separation has been used in several fields in recent years as it can extract the desired signals from the observed signal mixtures [8]. It can have potential applications in several different fields.

In the field of biomedical signal processing, BSS techniques can be used to process electroencephalography (EEG), electrocardiography (ECG) and electromyography (EMG)signals [30]. During the measurement of ECG, EEG and EMG signals from different body parts and desired measurements can get mixed, so BSS techniques can be used to separate the desired signals [31, 32]. BSS techniques are also used in high quality hearing aids as they can improve performance to reduce the listening effort for the users [33, 34].

BSS techniques can also be used in acoustic surveillance as they can be used to separate the mixed information and help the security agencies in intelligence and spying operations [35, 36]. Also, BSS methods have been used in underwater acoustic systems as BSS techniques can help in detection and separation of underwater acoustic signals which is helpful in understanding of the underwater environment, tracking ship movements and detecting any underwater oil or gas leakages [37].

In the field of image restoration, BSS techniques have been used for deconvolution [38,39]. Also, speech recognition systems must operate in difficult cluttered environments and therefore performance is generally affected in the presence of near by interfering sound sources. For example, the Siri voice recognition system that has been developed by Apple company takes voice commands from users to perform different task on mobile phones [40] and its performance is affected if there are interfering sources in close proximity. BSS methods can potentially be used in such scenarios to suppress the interfering sound sources and enhance the performance of speech recognition systems.

1.3 Aim and Objectives of the Thesis

The purpose of this thesis is to further the research on source separation by investigating and improving the IVA algorithms to achieve an efficient source separation system. The particular objectives of this thesis are:

- Objective 1: to exploit the multivariate Student's t distribution as the source prior in various forms of the IVA algorithm to improve the convergence and separation performance.

Chapter 4 deals with the Objective 1 of the thesis. Since the choice of source prior is critical to the performance of IVA algorithms, the Student's t source prior is adopted as a source prior for the IVA and the fast version of the IVA algorithm as it can better model speech signals thereby improving the separation performance of both algorithms achieving a faster convergence speed in terms of iteration numbers.

- Objective 2: to utilise the statistical property of the mixture signals to select automatically the mixing parameter of a combined distribution source prior and to achieve improved separation performance.

In Chapter 5, a mixture of the Student's t distribution and the original super Gaussian distribution is proposed as a source prior for the IVA algorithms. The weight of both distributions in the mixed source prior is adapted according to the energy of the observed mixture signals. Moreover, the overlapped chain type structure is used to model the dependency within the frequency bins to achieve a robust and improved separation performance with the IVA algorithms.

- Objective 3: to derive and evaluate the expectation maximization (EM) framework to obtain a new form of IVA algorithm which explicitly adapts the source prior according to the measured signal properties.

Chapter 6 addresses this objective by exploiting the EM framework for the original IVA algorithm. The complete EM framework based IVA is derived and this new framework for the IVA algorithm adapts the source prior according to the properties of the measured signals and therefore it enhances the robustness and the separation performance of the IVA algorithm.

- Objective 4: to perform extensive evaluation studies with real speech and room impulse response measurements to confirm the performance gains of the methods proposed in objectives 1, 2 and 3.

In Chapters 4, 5 and 6 the IVA algorithms are evaluated with real speech signals and by using the real room impulse responses which depict the separation performance of the proposed algorithms in the room environments.

1.4 Thesis Outline

The thesis is organised as follows:

Chapter 2 provides an introduction of the frequency domain blind source separation problem. A synopsis of independent component analysis is included and its advantages and limitations are discussed in the context of frequency domain

BSS problem. The natural gradient independent vector analysis algorithm is introduced in order to solve the permutation problem of the frequency domain ICA algorithm. Finally, the fast version of the IVA algorithm is discussed, which improves the convergence speed of the original IVA method.

Chapter 3 illustrates the experimental settings that are used to evaluate different algorithms throughout this thesis. The real room impulses responses are discussed along with the details of real room settings that are used in the experimentations. Furthermore the performance measures that are used to quantify the separation performance of the algorithms are discussed.

Chapter 4 studies a new multivariate Student' t source prior for the different versions of the IVA algorithm. The source prior is used to derive the nonlinear score function for the IVA method, therefore it is critical to the performance of the algorithm. The multivariate Student's t distribution is proposed as a source prior for both the IVA and the FastIVA method and the separation performance and convergence speed of the IVA and the FastIVA algorithms with the new source prior is compared with the original super Gaussian source prior.

Chapter 5 introduces a mixed source prior for the IVA algorithm. A convex combination of the Student's t and the original super Gaussian source prior is adopted as a source prior for the IVA and the FastIVA method, to better model the speech signals. Furthermore, an energy driven version of the mixed source prior is proposed which can adapt the weight of both distributions in the mixed source prior according to the energy of the observed mixture signals. Moreover, an overlapped clique (block) structure was adopted to model the dependency within the frequency bins. This new energy driven source prior was used for both the IVA and the FastIVA methods and the separation performance of both algorithms is tested in different reverberant room environments and it consistently improves the separation performance of both versions of the IVA algorithm.

Chapter 6 describes an efficient implementation of the expectation maximization (EM) framework for the IVA algorithm. Instead of a single distribution source prior for the IVA algorithm, the Student's t mixture model (SMM) is adopted as a source prior for the IVA method. It enables the source prior for the IVA algorithm to adapt according to different mixture signals and therefore the proposed source prior can properly model the nonlinear dependency structure within speech signals. An efficient EM algorithm was derived to estimate the parameters of the SMM source prior. The proposed method was tested with real room impulse responses and the experimental results confirm the advantage of the proposed method.

Chapter 7 concludes the thesis and discusses the directions for future work.

Chapter 2

BACKGROUND AND RELEVANT LITERATURE REVIEW

2.1 Introduction

The process of automated separation of acoustic sources from the measured mixtures is known as acoustic blind source separation (BSS). The typical application of blind source separation is the cocktail party problem. The process of focusing on one particular acoustic source of interest in the presence of multiple sound sources is known as the cocktail party problem [1]. Human beings can easily pay attention to one of the speakers in the presence of multiple active speakers; however, it is much more difficult to replicate the same ability in machines. In the past few decades, plenty of research has been conducted to study different aspects of the cocktail party problem. This research includes the study of the geometry of the microphone array [42], room impulse response identification [43], localisation of speech sources [17] and statistical estimation of speech sources. Independent Component Analysis (ICA) is one of the fundamental techniques to solve the cocktail party problem. The ICA algorithm was proposed by Herault and Jutten [10, 11] and it will be reviewed in this chapter. Independent vector analysis (IVA) is an extension of the ICA algorithm which was proposed to theoretically mitigate the permutation problem of ICA which is inherent to most of the BSS algorithms [22]. Two types of IVA algorithm will be reviewed in detail in this chapter. The first of the IVA algorithms is the original natural gradient IVA algorithm, it makes use of the gradient descent method to optimize the objective function. The fast fixed point IVA is the fast version of the IVA algorithm as it uses the Newton method to minimize the objective function. Mixing models for the BSS problem are discussed in the next section.

2.2 Mixing Models

One of the difficulties of BSS is that this problem particularly relies on the manner in which different source signals are mixed together in the physical environment. The simplest way of signals mixing together deals with an instantaneous mixing model, in which source signals include no delayed versions. This is the ideal scenario for the mixing process of different signals and the initial research and algorithms in the field of BSS were based on this model but such algorithms have limited practical application in the realistic scenarios. In real world multipath and reverberant environments, the signals measured at the acoustic sensors are convolutive mixtures of the sources [44]. Both instantaneous and convolutive mixing models will be discussed in detail in the following sections.

2.2.1 The Instantaneous Mixing Model

When a mixture of signals can be expressed as a linear combination of the original sources at every time instant, it is described as an instantaneous mixing model [12]. This model assumes that different sensors only receive the scaled version of the sound sources from the direct path between microphones and speakers. An instantaneous mixture for the two microphones and two sources case can be mathematically defined as follows:

$$\begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} s_1(t) \\ s_2(t) \end{pmatrix} + \begin{pmatrix} \zeta_1(t) \\ \zeta_2(t) \end{pmatrix} \tag{2.1}$$

where $s_1(t)$ and $s_2(t)$ represent the source signals and $x_1(t)$ and $x_2(t)$ represent the mixture signals. The parameter a_{12} represents the time-invariant transfer function coefficient between the source s_2 and x_1 and it represents the simple acoustic path between the source and the microphone. Likewise, matrix elements a_{mn} holds the other transfer function coefficients between sources and microphones. The signals $\zeta_1(t)$ and $\zeta_2(t)$ represents the additive noise. The noise term $\zeta(t)$ can be considered as extra spatially separated sources, however it is dropped from the reminder of thesis for brevity as in many other works in the field [21]. Equation (2.1) can be written generally for m microphones and n sources without noise as follows:

$$\mathbf{x}(t) = \mathbf{A}(t)\mathbf{s}(t) \tag{2.2}$$

where $\mathbf{x}(t)$ is a received mixture vector, $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_m(t)]^T$ whereas, the source signal is a vector $\mathbf{s}(t) = [s_1(t), s_2(t), ..., s_n(t)]^T$, and $\mathbf{A}$ is a $m \times n$ mixing matrix and the index t denotes the discrete time index.

This is the simplest form of modelling a mixture of signals. This is the case for an ideal surrounding environment where no reverberation or multipath exists. Therefore the instantaneous model only considers the source signals being amplitude modulated and not time delayed or echoed [41]. An ideal instantaneous model is shown in Figure 2.1. There are several applications in signal processing where the instantaneous mixture model is applicable. It can be used in brain science as the BSS algorithms help to determine hidden components of various brain activity from sequences of brain action as recorded by an electroencephalogram (EEG) [12]. An instantaneous mixing model can also be used in image processing applications, in which the separation of independent elements in an image is required to improve the image quality [38]. However, a realistic approach for speech separation must take the convolutive mixing of the acoustic paths into consideration, as in the real reverberant environment multipath effects cannot be ignored.

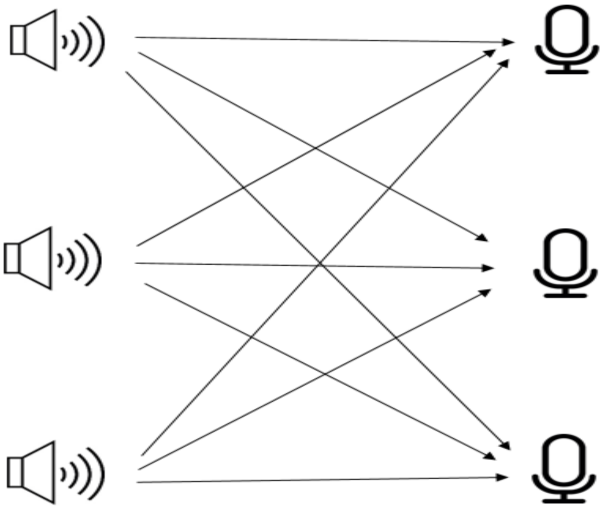

Figure 2.1: Instantaneous mixing model with three sources and microphones

In the real room environment, different speech sources mixed together to form convolutive mixtures and details of this mixing process are discussed in the next section.

2.3 Convolutive Mixing Model

The instantaneous mixture model is not sufficient to represent various problems of the real world environment as it only considers the scaled version of sound signals and does not consider delayed versions of sources from reverberations. Moreover, in the real environment speech signals are non-stationary and are temporally spread by the reverberant environment i.e. they arrive at different times and are delayed as they travel towards the sensors [56]. It makes the problem of source separation more challenging to solve and the complexity of the problem increases with the reverberation time [14]. Therefore in practical scenarios, source separation algorithms consider the convolutive mixing model. For the two sources and two microphones case, the convolutive model can be represented as follows [44]:

$$x_1(t) = [a_{11}(t) \otimes s_1(t) + a_{12}(t) \otimes s_2(t)] + \zeta_1(t) \tag{2.3}$$

$$x_2(t) = [a_{21}(t) \otimes s_1(t) + a_{22}(t) \otimes s_2(t)] + \zeta_2(t) \tag{2.4}$$

where $\otimes$ represents the convolution operator. The general form of the convolutive mixing model can be represented as:

$$x(t) = [A(t) \otimes s(t)] + \zeta(t) \tag{2.5}$$

For the general case the mixing filter is an $m \times n$ polynomial matrix A, where m denotes the number of the sources and n represents the number of sensors. Given the convolutive mixing model in Equation (2.5), the problem then is determining $m \times n$ coefficients of the polynomial $\mathbf{A}$, and thereby, to estimate the source signals. This is a complicated process as matrix $\mathbf{A}$ is a matrix of filters, instead of a matrix of scalars, which is the case for the instantaneous mixture model. In order to overcome the computational complexity, the convolutive mixing model can be moved to the frequency domain, since the convolution in the time domain is equivalent to multiplication in the frequency domain [52]. Hence, the time domain convolutive mixing model without noise can be written in the frequency domain as follows.

$$\mathbf{x}(k) = \mathbf{A}(k)\mathbf{s}(k) \tag{2.6}$$

where $\mathbf{x}(k) = [x_1(k), x_2(k) \cdots x_m(k)]^T$ and $\mathbf{s}(k) = [s_1(k), s_2(k) \cdots s_n(k)]^T$ are the observed mixture signal vector and the source signal vector both in the frequency domain, respectively, and $(.)^T$ denotes vector transpose. $\mathbf{A}(k)$ is the mixing matrix. The dimension for $\mathbf{A}(k)$ is m x n and the index k denotes the k-th frequency bin of this multivariate model. In this thesis, focus is on the exactly determined case, that means there is an equal number of speakers and microphones, therefore the mixing matrix $\mathbf{A}(k)$ is assumed to be a square matrix at all frequency bins. The convolutive mixing model is shown in Figure 2.2. For the reminder of thesis all mixtures are assumed to be convolutive in the time domain and the frequency domain conversion is considered to cut down the computational complexity for the BSS algorithms.

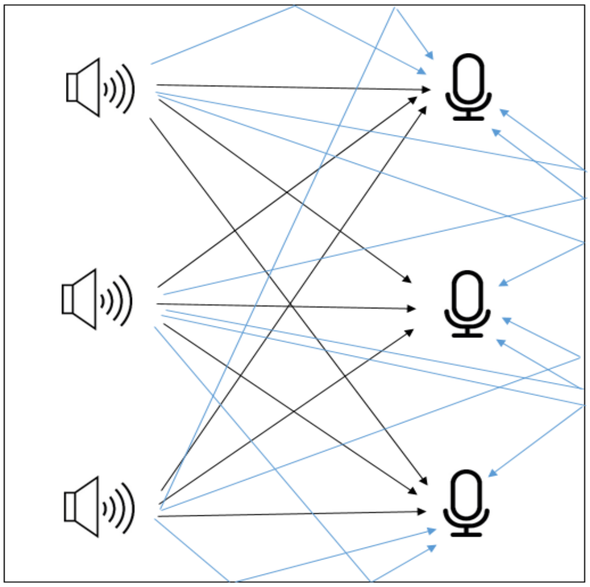

Figure 2.2: Convolutive mixing model with three sources and microphones.

2.4 Statistical Solution to Convolutive BSS Problem

During the past few decades, various research studies have been conducted to find an appropriate solution for the BSS problem [14]. In the start, the main focus of the research was to determine a solution for the time domain BSS problem. However, in realistic environments, the impulses response are very complex as their length is on the order of thousands of samples. Therefore, the time domain methods for source separation are generally computationally high [46,47]. In order to overcome the computational complexity of the time domain BSS methods, a frequency domain solution for the BSS problem was introduced by Parra and Spence [44]. Since the frequency domain methods reduce the computational cost of the algorithms, various frequency domain methods have been proposed for the solution of BSS problem [14, 48–50]. Independent component analysis is a well known method for the source separation problem and it is discussed in the next section.

2.4.1 Independent Component Analysis

Independent component analysis (ICA) is one of the main tools to solve the BSS problem and it was clearly formalised by Comon [11] whereas the initial concept was introduced by Herault and Jutten [10]. ICA is a well-known statistical technique for disclosing hidden variables and factors from sets of observed random measurements. The ICA method in its simplest form was initially introduced to solve the problem of BSS in the instantaneous mixing model. However, in realistic scenarios, different speech sources mixed together according to the convolutive mixing model and therefore the ICA method has to separate the sources from convolutive mixtures. Since the implementation of the ICA method for convolutive mixtures in the time domain is computationally complex therefore in order to overcome the computational complexity, the ICA method is implemented in the frequency domain [52,53]. In the frequency domain implementation, the ICA algorithm for the instantaneous mixing model can be implemented at each frequency bin. This necessarily is the parallel execution of the ICA algorithm for the instantaneous model at each frequency bin for the solution of convolutive BSS. The ICA method is also considered as an improvement over principal component analysis (PCA) as the ICA method is capable of revealing the underlaying sources which in most of the cases the classic techniques like PCA have failed to accomplish [51]. In order to implement the ICA method for the BSS problem, certain assumptions are needed [54], which are as follows.

- The independent components are assumed to be statistically independent.

  The concept of statistical independence can be elaborated by considering two random variables s_1 and s_2. The random variables s_1 and s_2 are said to be independent when the information on the value of s_1 does not provide any information on the value of s_2, and vice versa. Mathematically, two variables can be independent only if the joint probability density function (pdf) is factorisable to the product of marginal distribution as follows:

  $$p(\mathbf{s}) = \prod_{i=1}^{N} p(s_i)$$

  For example, when two source components are considered, in that case the pdf can be factorised as:

  $$p(s_1, s_2) = p(s_1)p(s_2)$$

- The independent components must generally have non-Gaussian distributions.

  In order to perform the independent component analysis the consideration of higher order statistics (HOS) is generally vital. The Gaussian distribution has certain HOS which are zero, therefore the independent component with non-Gaussian distribution can be investigated.

- It is assumed that the unknown mixing matrices are invertible.

  This means that the size of the independent components is equal to the size of the observed mixture or it can be considered as the number of sources that is either smaller or equal to the number of observed mixtures, which is also known as the over-determined or exactly determined problem, respectively.

In order to implement the ICA algorithm for the BSS problem, pre-processing the data can help to reduce the noise, accelerate the convergence speed and reduce computation. This includes e.g. the removal of the sample mean or decorrelation of the mixtures. One typical pre-processing technique is the spatial whitening of the data [53]. Before ICA, the standard PCA algorithm can be applied on the data as it produces uncorrelated signals.
The noise-free ICA model can be written as follows:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{2.7}$$

For simplicity the time index is discarded from the above mentioned equation. The parameters $\mathbf{x}$ and $\mathbf{s}$ can be considered as random vectors. The mixing matrix $\mathbf{A}$ is assumed constant for all observations. The independent components for each source can be estimated as follows:

$$\mathbf{s} = \mathbf{W}\mathbf{x} \tag{2.8}$$

where $\mathbf{W}$ is the unmixing matrix. Estimating an independent component can now be considered as the search for the linear combination that can be represented as $\mathbf{s} = \mathbf{w}^{\dagger}\mathbf{x}$, where $\mathbf{w}$ is the unmixing vector for one source to be determined and $(.)^{\dagger}$ denotes Hermitian transpose. The goal of ICA is to find the $\mathbf{w}$ such that $\mathbf{s}$ is one of the sources. The basic idea in ICA is that the sum of independent variables due to mixing tends to be more Gaussian than the original variable. Therefore the goal for unmixing the sources is to find the unmixing vector $\mathbf{w}$ that can maximize the non-Gaussianity of $\mathbf{s} = \mathbf{w}^{\dagger}\mathbf{x}$.

The nonlinear contrast function for which its extreme values coincide with the independent components is needed for the ICA algorithm and it is given as [52]:

$$J_G(\mathbf{w}) = E\{G|\mathbf{w}^{\dagger}\mathbf{x}|^2\} \tag{2.9}$$

Finding the extrema is only possible if the function G is real. For this reason the suggested contrast function will operate only on the absolute values rather then complex values. With this contrast function, the optimisation problem can be considered as to maximize

$\sum_{n=1}^{N} J_G(\mathbf{w}_i)$ with respect to $\mathbf{w}_n$ where $i = 1, ..., n$

under the constraint $E\{\mathbf{w}_m^{\dagger}\mathbf{x})(\mathbf{w}_n^{H}\mathbf{x})\} = \delta_{nm}$, where $\delta_{nm} = 1$ for $n = m$ and $\delta_{nm} = 0$ otherwise.

Now a nonlinear function G that grows slowly is needed, since the slower the growth of G the more robust the estimator will be to outliers. There are different choices for the selection of G as mentioned in [52], some of them are

$$G_1(y) = \sqrt{(a_1 + y)} \tag{2.10}$$

$$G_2(y) = log(a_2 + y) \tag{2.11}$$

$$G_3(y) = (1/2)y^2 \tag{2.12}$$

where a_1 and a_2 are arbitrary constants and y represents a complex random variable. From the three above mentioned functions G_1 and G_2 are the better choices, as they grow more slowly than G_3.

Independent components can either be determined by the maximization or the minimization of the optimization function, when

$$E(g|s(k)|^2) + |s(k)|^2 g'(|s(k)|^2) - |s_(k)|^2 g(|s(k)|^2) < 0 \tag{2.13}$$

where g(.) is the derivative of G(.) and $g'(.)$ is the derivative of g(.), represents the maximization of the function J_G, whereas for the minimization case it will be

$$E(g|s(k)|^2) + |s(k)|^2 g'(|s(k)|^2) - |s(k)|^2 g(|s(k)|^2) > 0 \tag{2.14}$$

The fixed point algorithm searches for the extrema of the cost function $E\{G(|\mathbf{w}^\dagger \mathbf{x}|^2)\}$. Here G is an even and symmetrical function and the expectation operator will be estimated by the sample mean over the whitened vectors. This can be accomplished by the use of principal component analysis. The fixed point algorithm for one unit is

$$\mathbf{w}^+ = E\{\mathbf{x}(\mathbf{w}^\dagger \mathbf{x}) * g(|\mathbf{w}^\dagger \mathbf{x}|^2)\} - E\{(g|\mathbf{w}^\dagger \mathbf{x}|^2) + |\mathbf{w}^\dagger \mathbf{x}|^2 * g'(|\mathbf{w}^\dagger \mathbf{x}|^2)\}\mathbf{w} \tag{2.15}$$

$$\mathbf{w}_{new} = \mathbf{w}^+ / ||\mathbf{w}^+|| \tag{2.16}$$

When only one-unit case is considered, only one of the rows of $\mathbf{W}$ will be considered and the orthogonalisation in this case will be changed to just normalisation of the vector of a unit length after every single iteration step. This algorithm for one unit can be developed to determine the complete ICA statistical transformation. To obtain the full matrix $\mathbf{W}$, the algorithm will be run for one step for n times and the vectors must be orthonormalised again as they were reduced to just normalisation after one unit iteration. The symmetric orthonormalisation can be used to obtain the complete **W**. By using the symmetric orthonormalisation, the independent components can be computed in a parallel fashion and also it avoids the chance of accumulating estimation error from the first vector to the subsequent ones. Symmetric orthonormalisation is given as:

$$\mathbf{W} = \mathbf{W}(\mathbf{W}^{\dagger}\mathbf{W})^{-1/2} \tag{2.17}$$

Symmetric orthonormalisation is firstly used for performing the iterative step of the one unit fixed point algorithm on all vectors in parallel and after that, making them orthogonal to obtain the desired unmixing matrix **W** [52].

The ICA algorithm has a good separation performance for the speech signals as they are non-stationary and ICA makes use of the non-Gaussianity for the separation of the sources [54]. Another version of the ICA algorithm is Fast Independent Component Analysis (FastICA), which is a fast converging version of the ICA algorithm. Using FastICA has certain advantages over the other techniques for ICA, step size parameters are not needed, which makes it easy to use. Also in ICA independent components can be calculated one-by-one which is really helpful for exploratory data analysis and it gives more information about the problem. Overall, FastICA is a good algorithm for BSS and is one of the fundamental algorithms that exploits the independence between the sources for BSS. But as discussed earlier, it has scale ambiguity and suffers from the permutation problem [53].

- The first problem is that, the original energies of independent components cannot be determined. As both $\mathbf{s}$ and $\mathbf{A}$ are unknown, any scalar components can cancel the effect of each other. However, the magnitude can be fixed by making the assumption that each component has unit variance. But it still leaves an ambiguity in the sign. But fortunately this problem is generally insignificant in the case of BSS.

- The fundamental problem with the ICA algorithm is that the sequence of the independent components cannot be determined. This problem also occurs because both $\mathbf{A}$ and $\mathbf{s}$ are unknown and can easily switch the order of the terms in the case of the sum and consider any of them the first one. This problem badly affects the performance of frequency domain ICA. At some frequency bins it undoes the work done by the ICA algorithm, so an improved solution for better separation of the source signals from the blind mixture is needed.

In the past decades, various techniques have been introduced to overcome the permutation problem of the ICA method [59–63] . The idea of independent vector analysis was introduced in [21] and it was modelled theoretically to avoid the permutation ambugity by preserving the dependency within each source vector and eliminating the dependency among different vector sources. The Independent Vector Analysis (IVA) algorithm is discussed in detail in the next section.

2.5 Independent Vector Analysis

Independent vector analysis (IVA) is an extension of the ICA algorithm. It is based on a dependency model which retains interfrequency dependencies within each source vector and is represented in Figure 2.3 and it shows the dependency structure of the IVA method when two sources and two measurements are considered. The individual layers of the mixtures in the ICA method can be mapped to the instantaneous mixture at each frequency bin; also the dependent sources that are arranged together as a multivariate variable to correspond to the frequency domain components of a time domain signal [21].

When the IVA algorithm is compared with the ICA algorithm, the inter-frequency dependencies within each source depend on the modified prior of the source signal.

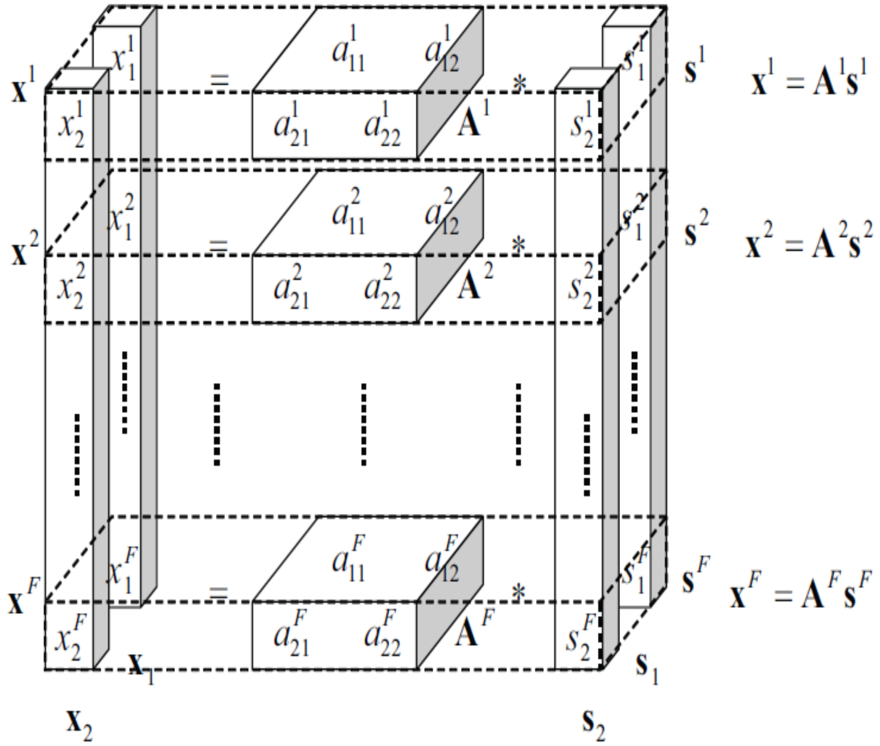

Figure 2.3: The independent vector analysis model for two sources and two measurements case [21]

In the ICA algorithm, independence for each frequency component is measured separately at each frequency bin. The IVA method rather makes use of the higher order dependencies across frequencies and each source prior is described as a multivariate super-Gaussian distribution. Therefore it measures the independence across the whole multivariate source and it can retain the higher order inter-frequency dependencies and structure of frequency components. Furthermore, the permutation ambiguity that is inherent to the ICA method, can be avoided and the separation performance can be improved.

In order to implement the IVA algorithm for the convolutive BSS, the short time Fourier transfer (STFT) is used to convert the problem from the time domain to the frequency domain as it eases the computational complexity of the time domain method. The basic noise free BSS model for the IVA method in the frequency domain can be defined as follows:

$$\mathbf{x}(k) = \mathbf{A}(k)\mathbf{s}(k) \tag{2.18}$$

where $\mathbf{A}$ is a mixing matrix of dimensions $m \times n$. The index k represents the $k - th$ frequency of this multivariate method. In order to separate the source signals from the observed mixtures, an unmixing matrix must be estimated to retrieve the estimate of original sources, as

$$\hat{\mathbf{s}}(k) = \mathbf{W}(k)\mathbf{x}(k) \tag{2.19}$$

where $\hat{\mathbf{s}}(k)$ is the estimated source signal, $\hat{\mathbf{s}}(k) = [\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2(k) \cdots \hat{\mathbf{s}}_n(k)]^T$, $\mathbf{W}(k)$ is the unmixing matrix of dimensions $n \times m$. In this thesis, focus is on the exactly determined case, so the number of sources is considered equal to the number of microphones i.e. $n = m$.

In order to model the independence between sources, the IVA method uses the Kullback-Leilber divergence. So a cost function can be derived as follows:

$$J_{IVA} = \mathcal{KL}(p(\hat{\mathbf{s}})||\prod q(\hat{\mathbf{s}})) \tag{2.20}$$

$$= \int p(\hat{s}_1 \cdots \hat{s}_n) log \frac{p(\hat{s}_1 \cdots \hat{s}_n)}{\prod q(\hat{\mathbf{s}})} d\hat{s}_1 \cdots d\hat{s}_n \tag{2.21}$$

$$= \text{const} - \sum_{k=1}^{K} \text{log}|\text{det}(W^{(k)})| - \sum_{i=1}^{n} E[\text{log}q(\hat{s}_i)] \tag{2.22}$$

where det(.) represents the matrix determinant and $E(.)$ shows the expectation operator. All the sources in the cost function of the IVA algorithm are multivariate and the cost function will be minimised when different vector sources will become independent of each other and the dependency within each source vector is retained. Hence this cost function can be used to eliminate the dependency between the vector sources and preserve the frequency dependency within each vector source.

2.5.1 Natural Gradient IVA Method

In order to minimise the cost function J_{IVA} mentioned in Equation (2.22) for the IVA method, the natural gradient method can be implemented [21, 65]. The Natural Gradient IVA method is also referred to as the original IVA method throughout this thesis. The partial derivative of the cost function with respect to the coefficients of the separating matrices $w_{ij}(k)$ is calculated as follows [21]:

$$\Delta w_{ij}(k) = -\frac{\partial J}{\partial w_{ij}(k)} = (w_{ij}(k))^{-\dagger} - E[\varphi(k)(\hat{s}_i(1)\cdots\hat{s}_i(k))]x_j^*(k) \tag{2.23}$$

where $(.)^*$ denotes the conjugate operator. The natural gradient algorithm can be obtained by multiplying through by $W(k)^\dagger W(k)$:

$$\Delta w_{ij}(k) = \sum_{l=1}^{n} \Big(I_{il} - E\varphi(k)(\hat{s}_i(1)\cdots\hat{s}_i(k))\hat{s}_l{*}(k) \Big) w_{lj}(k) \tag{2.24}$$

where I is an identity matrix only when $i = l$ and 0 otherwise. The update rule from Equation(2.24) can be written as:

$$w_{ij}(k)_{new} = w_{ij}(k)_{old} + \eta\Delta w_{ij}(k) \tag{2.25}$$

where η is learning rate. The nonlinear function $\varphi(k)$ is the multivariate score function and it is based on a super Gaussian distribution source prior. The nonlinear function $\varphi(k)$ is defined as:

$$\varphi(k)(\hat{s}_i(1)\cdots\hat{s}_i(k)) = -\frac{\partial \mathrm{log} q(\hat{s}_i(1)\cdots\hat{s}_i(k))}{\partial \hat{s}_i(k)} \tag{2.26}$$

This multivariate score function preserves the dependency across all the frequency bins. This score function is based on a multivariate super Gaussian source prior and the choice of this source prior is crucial to the performance of the IVA algorithm. In [21], a particular super Gaussian distribution is adopted as a source prior for the IVA algorithm. The source prior for this particular super Gaussian distribution is given as:

$$q(s_i) \propto \exp\left(- \sqrt{(\mathbf{s}_i - \mu_i)^{\dagger} \Sigma_i^{-1} (\mathbf{s}_i - \mu_i)} \right) \tag{2.27}$$

By setting the mean value to zero and covariance matrix to an identity matrix, the non-linear score function for the IVA method can be obtained as follows [21]:

$$\begin{aligned} \varphi^{(k)}(\hat{s}_i(1) \dots \hat{s}_i(K)) &= -\frac{\partial \log\left(p(\hat{s}_i(1) \dots \hat{s}_i(K)\right)}{\partial \hat{s}_i(k)} \\ = \frac{\hat{s}_i(k)}{\sqrt{\sum_{k'=1}^{K} |\hat{s}_i(k')|^2}} \end{aligned} \tag{2.28}$$

It is a multivariate score function now and it is used to represent inter-frequency dependency between the sources. However, this score function is not unique and it depends on the types of sources. It can be adjusted according to the nature of source signals and it is crucial to the performance of the IVA method. The choice of the source prior and the multivariate score function is discussed in more detail in Chapter 4. The original super Gaussian source prior used in [21] is shown in Figure 2.4. The convergence speed of the IVA method can be improved by using the Newton's method in the update and it is discussed in detail in the next section.

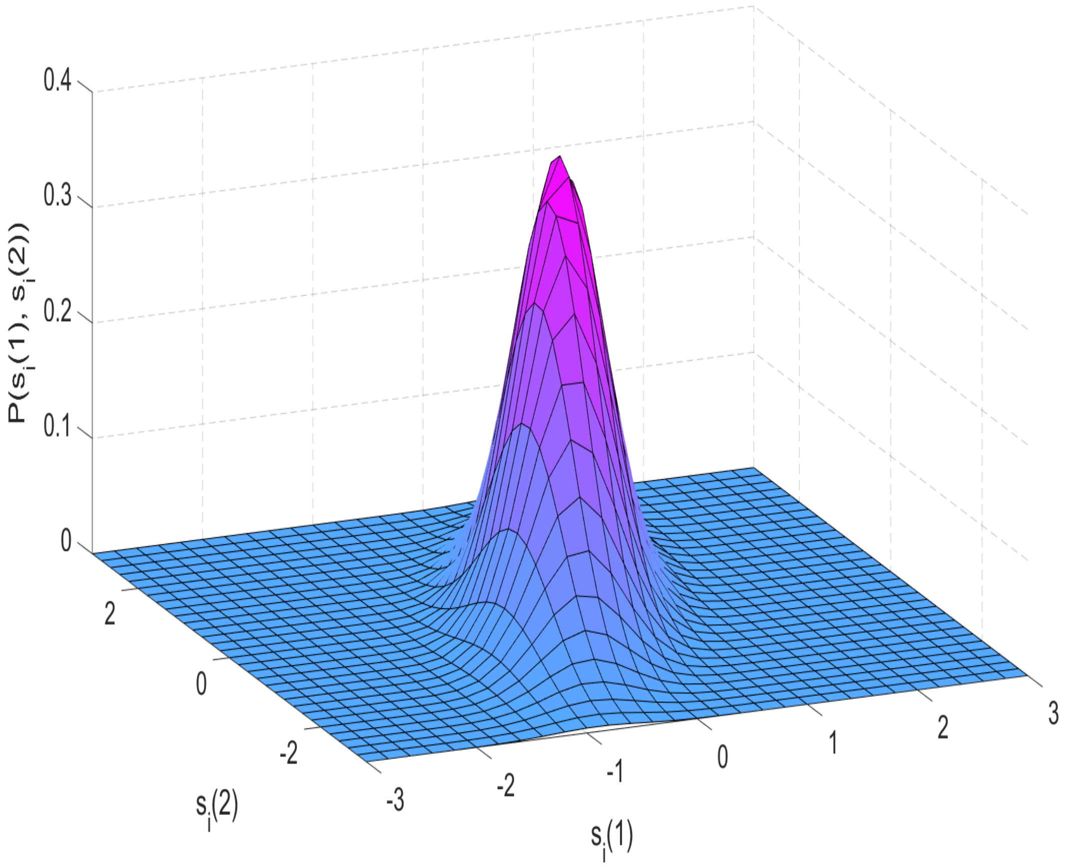

Figure 2.4: Bivariate version of a multivariate super Gaussian distribution used as a source prior in original IVA

2.6 Fast fixed-point IVA algorithm

A new learning algorithm that uses the Newton method is introduced in this section. Fast fixed point IVA is a fast converging version of the IVA method, as it adopt the Newton's method during the update process [58]. The Newton's method is a second order learning algorithm and it can converge quadratically and it is free from the selection of an appropriate learning rate [57]. In order to model the independence between sources, an objective function is needed for the

FastIVA method which is given as [58]:

$$J_{FastIVA} = \sum_{i=1}^{N} \Big[E[F(\sum_{k=1}^{K} |\hat{s}_i(k)|^2)] - \sum_{k=1}^{K} \lambda_i(k)(\mathbf{w}_i(k)^\dagger \mathbf{w}_i(k) - 1) \Big] \tag{2.29}$$

where λ_i denotes the Langrange multiplier and $\mathbf{w}_i^\dagger$ represents the i-th row of the complete unmixing matrix $\mathbf{W}$. This objective function is a multivariate function therefore it can retain the dependency within source vectors and it can be used to make sources independent of each other. The quadratic Taylor series polynomial is adopted for the implementation of the Newton's method in the FastIVA method [58]. It will be introduced in complex variable notation to be used in the contrast function.

$$\begin{aligned} f(\mathbf{w}) =& f(\mathbf{w}_o) + \frac{\partial f(\mathbf{w}_o)}{\partial \mathbf{w}^T}(\mathbf{w} - \mathbf{w}_o) + \frac{\partial f(\mathbf{w}_o)}{\partial \mathbf{w}^\dagger}(\mathbf{w} - \mathbf{w}_o)^* \\ &+ \frac{1}{2}(\mathbf{w} - \mathbf{w}_o)^T \frac{\partial^2 f(\mathbf{w}_o)}{\partial \mathbf{w} \partial \mathbf{w}^T}(\mathbf{w} - \mathbf{w}_o) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_o)^\dagger \frac{\partial^2 f(\mathbf{w}_o)}{\partial \mathbf{w}^* \partial \mathbf{w}^\dagger}(\mathbf{w} - \mathbf{w}_o)^* \\ &+ (\mathbf{w} - \mathbf{w}_o)^\dagger \frac{\partial^2 f(\mathbf{w}_o)}{\partial \mathbf{w}^* \partial \mathbf{w}^T}(\mathbf{w} - \mathbf{w}_o) \end{aligned} \tag{2.30}$$

In order to simplify the objective function, $\mathbf{w}_i(k)$ is replaced with $\mathbf{w}$ and the summation term in Equation (2.29) is consider as $f(\mathbf{w}_i(k))$, thus

$$f(\mathbf{w}_i(k)) = E\Big[F(\sum_{k'=1}^{K} |\hat{s}_i(k')|^2)] - \sum_{k'=1}^{K} \lambda_i(k')(\mathbf{w}_i(k')^\dagger \mathbf{w}_i(k') - 1)\Big] \tag{2.31}$$

In order to optimise $f(\mathbf{w}_i(k))$ taking the first derivative $f(\mathbf{w}_i(k))$ and setting it to zero

$$\begin{aligned}\frac{\partial f(\mathbf{w}_i(k))}{\partial \mathbf{w}_i(k)^*} \approx & \frac{\partial f(\mathbf{w}_{i,o}(k))}{\partial \mathbf{w}_i(k)^*} + \frac{\partial^2 f(\mathbf{w}_{i,o})}{\partial(\mathbf{w}_i(k))^* \partial(\mathbf{w}_i(k))^T}(\mathbf{w}_i(k) - \mathbf{w}_{i,o}(k)) \\ & + \frac{\partial^2 f(\mathbf{w}_{i,o})}{\partial(\mathbf{w}_i(k))^* \partial(\mathbf{w}_i(k))^\dagger}(\mathbf{w}_i(k) - \mathbf{w}_{i,o}(k))^* \approx \mathbf{0}\end{aligned} \tag{2.32}$$

The derivative term in the above mentioned equation can be written as:

$$\frac{\partial f(\mathbf{w}_{i,o}(k))}{\partial \mathbf{w}_i(k)^*} = E\Big[\hat{s}_{i,o}(k)^* F'(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2)\Big] - \lambda_i(k)\mathbf{w}_{i,o}(k) \tag{2.33}$$

The second derivative of the above mentioned equation can be given as:

$$\begin{aligned}&\frac{\partial^2 f(\mathbf{w}_{i,o})}{\partial(\mathbf{w}_i(k))^* \partial(\mathbf{w}_i(k))^T} = E\Big[(F'(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2) + |\hat{s}_{i,o}(k)|^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2))\mathbf{x}(k)\mathbf{x}(k)^*\Big] \\ & - \lambda_i(k) I \approx E\Big[(F'(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2) + |\hat{s}_{i,o}(k)|^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k)|^2))] E[\mathbf{x}(k)\mathbf{x}(k)^*\Big] - \lambda_i(k) I \\ & = \Big(E\Big[(F'(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2) + |\hat{s}_{i,o}(k)|^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2))\Big] - \lambda_i(k)\Big) I\end{aligned} \tag{2.34}$$

where the first and second derivative terms are denoted by $F(.)'$ and $F(.)''$, respectively. By simplifying and also because of the whitening process, and making the assumption of $E[\mathbf{x}(k)\mathbf{x}(k)^*] = I$ in Equation (2.34):

$$\begin{aligned}&\frac{\partial^2 f(\mathbf{w}_{i,o})}{\partial(\mathbf{w}_i(k))^* \partial(\mathbf{w}_i(k))^\dagger} = E\Big[(\hat{s}_{i,o}(k)^*)^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2))\mathbf{x}(k)\mathbf{x}(k)^T\Big] \\ & \approx E\Big[(\hat{s}_{i,o}(k)^*)^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2))\Big] E\Big[\mathbf{x}(k)\mathbf{x}(k)^T\Big] \\ & = \mathbf{0}\end{aligned} \tag{2.35}$$

In order to simplify Equation (2.35) the assumption of $E[\mathbf{x}(k)\mathbf{x}(k)^T] = \mathbf{0}$ due to complex circularity has been considered [67]. From Equation (2.32) and (2.34), the objective function of the FastIVA method is reduced to

$$\mathbf{w}_i(k) - \mathbf{w}_{i,o}(k) = \frac{-1}{c(\mathbf{w}_{i,o})}.\frac{\partial(\mathbf{w}_{i,o})}{\partial(\mathbf{w}_i(k))^*} \tag{2.36}$$

where $c(\mathbf{w}_{i,o})$ is a constant term. By substitution, the iterative algorithm is given as:

$$\mathbf{w}_i(k) \leftarrow \mathbf{w}_{i,o}(k) - \frac{E\Big[\hat{s}_{i,o}(k)^* F'(\sum_{k'}|\hat{s}_{i,o}(k')|^2)\Big] - \lambda_i(k)\mathbf{w}_{i,o}(k)}{E\Big[(F'(\sum_{k'}|\hat{s}_{i,o}(k')|^2) + |\hat{s}_{i,o}(k)|^2 F''(\sum_{k'}|\hat{s}_{i,o}(k')|^2))\Big] - \lambda_i(k)} \tag{2.37}$$

The Lagrange multiplier $\lambda_i(k)$ in the above mentioned equation is given as:

$$\lambda_i(k) = E\Big[|\hat{s}_{i,o}(k)|^2 F'(\sum_{k'=1}^{K}|\hat{s}_{i,o}(k')|^2)\Big] \tag{2.38}$$

and $\lambda_i(k)$ can be removed by multiplying the numerator in Equation (2.37) on both sides of the equation. Then by using normalisation, the learning rule is given as:

$$\begin{aligned}\mathbf{w}_i(k) \leftarrow & E\Big[F^{'}(\sum_{k'=1}^{K}|\hat{s}_{i,o}(k')|^2) + |\hat{s}_{i,o}(k)|^2 F^{''}(\sum_{k'=1}^{K}|\hat{s}_{i}(k')|^2))\Big]\mathbf{w}_i(k) \\ & - E\Big[(\hat{s}_{i,o}(k))^* F^{'}(\sum_{k'=1}^{K}|\hat{s}_{i,o}(k')|^2)\mathbf{x}(k)\Big]\end{aligned} \tag{2.39}$$

In order to obtain the unmixing matrix $\mathbf{W}(k)$, the above mentioned equation is implemented for all the sources. This unmixing matrix is decorrelated by the symmetric decorrlation scheme as follows:

$$\mathbf{W}(k) \leftarrow (\mathbf{W}(k)(\mathbf{W}(k))^{\dagger})^{-1/2}\mathbf{W}(k). \tag{2.40}$$

So an unmixing matrix can be obtained, which can be used to separate the source signals from the observed mixtures. The source prior is used to derive the non linear function for the FastIVA method. Therefore it is important to choose an appropriate source prior for the better separation performance of the algorithm. In [58], a super Gaussian distribution is adopted as a source prior for the FastIVA method, it is the same original source prior that was used for the original IVA method as well. When the mean is assumed to be zero and the variance is considered as unity, the original source prior for the FastIVA method is given as follows:

$$F(\sum_{k'=1}^{K}|\hat{s}_i(k')|^2) = \sqrt{(\sum_{k'=1}^{K}|\hat{s}_i(k')|^2)} \tag{2.41}$$

A detail discussion on the choice of source prior and its effect on the separation performance of the FastIVA is included in Chapter 4 of this thesis.

2.7 Summary

In this chapter, different mixing models were discussed. Then the basic techniques in the field of source separation were outlined in this chapter. Then the ICA algorithm was introduced for convolutive blind source separation and its limitations were examined. Also, the gradient decent IVA method was introduced to solve the permutation problem of the frequency domain ICA algorithm. At the end, a preliminary study of the FastIVA algorithms, which is the fast version of the IVA algorithm, was included.

In order to evaluate the BSS algorithms, various performance measures can be used and details of these performance measures along with datasets used to generate room impulse responses are discussed in the next section.

Chapter 3

ROOM ACOUSTICS AND PERFORMANCE MEASURES

This chapter provides a discussion of different room models and the performance measures that are used to evaluate the separation performance of the BSS algorithms. Firstly, to evaluate the BSS algorithms, performance measurement criteria will be discussed. Then the simulation environment will be described, that will be used to analyse the behaviour of different algorithms, by exploiting artificial and real room impulse responses.

3.1 Performance Measure

The separation performance of the BSS algorithms is calculated by different performance measures. There are subjective as well as objective measures. In this thesis the following performance measures will be used to evaluate the separation performance of BSS algorithms.

3.1.1 Performance Index

The first criterion that will be used to measure the separation performance is called the performance index (PI) [12]. 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 $\mathbf{G} = \mathbf{W}^{\dagger}\mathbf{A}$, where matrix $\mathbf{W}$ is the separating matrix obtained by the BSS algorithm. Mathematically it can be written as:

$$\begin{aligned} PI(G) =& [\frac{1}{n}\sum_{i=1}^{n}(\sum_{k=1}^{m}\frac{|G_{ik}|}{max_k|G_{ik}|}-1)]+ \\ & [\frac{1}{m}\sum_{k=1}^{m}(\sum_{i=1}^{n}\frac{|G_{ik}|}{max_i|G_{ik}|}-1)] \end{aligned} \tag{3.1}$$

Although PI can show separation performance in each frequency bin, it can not show the permutation directly [89]. Thus, for a two-input two-output model, the criterion $[abs(\mathbf{G}_{11}\mathbf{G}_{22}) - abs(\mathbf{G}_{12}\mathbf{G}_{21})]$ will be used to measure the permutation.

3.1.2 Signal to Distortion Ratio

Secondly, the SiSec toolbox [77] will be used to measure the separation performance of BSS algorithms. This is a toolbox that can give reliable results in the form of source to interference ratio (SIR) and source to distortion ratio (SDR). To define the SiSec toolbox, first consider an original source s_i and after separation, an estimate of the original source $\hat{s}_i$. The estimated source $\hat{s}_i$ can be decomposed as

$$\hat{s}_i = s_{target} + e_{interf} + e_{noise} + e_{artif} \tag{3.2}$$

where s_{target} is a modified version of the original source s_i by an allowed distortion and $e_{interf}, e_{noise}, e_{artif}$ respectively, are the interference, noise, and artifact error terms. These four terms should represent the part of $\hat{s}_i$ perceived as coming from the wanted source s_i, and from other unwanted sources i.e. from sensor noises and interference. Now the SDR and SIR terms that can compute the energy ratios of estimated sources in decibels (dB) are defined as :

$$\begin{aligned} \mathrm{SDR} =& 10\log_{10}\frac{||s_{target}||^2}{||e_{interf}+e_{noise}+e_{artif}||^2} \\ \mathrm{SIR} =& 10\log_{10}\frac{||s_{target}||^2}{||e_{interf}||^2} \end{aligned} \tag{3.3}$$

The SIR measure only considers the interfering sources that influence the separated source, whereas the SDR measure, in addition to the interfering sources also considers the additive noise in the separated source. These SDR and SIR measures provide reliable information about the fidelity of the recovered signal [77]. In this thesis, the value of SDR and SIR is assumed to be 0 dB at the microphone as the sources are considered to have similar variance at the microphones.

3.1.3 Perceptual Evaluation of Speech Quality

The third criterion that will be used to measure the separation performance for BSS algorithm is known as perceptual evaluation of speech quality (PESQ). It provides a subjective measure to evaluate the separation performance of the BSS algorithm. Objective performance measures can provide good comparison for the separation performance of different algorithms, whereas a subjective measure can portray the true quality of the separated speech signals. Therefore the subjective measure of PESQ is used to evaluate the separation performance of the BSS algorithm and it is basically an approach designed to predict the subjective opinion scores of a degraded audio sample. It compares the original speech signal and the estimated speech signals as shown in Figure 3.1. After comparison it provides results in the form of mean opinion score for the speech quality, with values from 0-4.5, where 0 denotes a very poor separation performance and 4.5 an excellent separation performance [78].

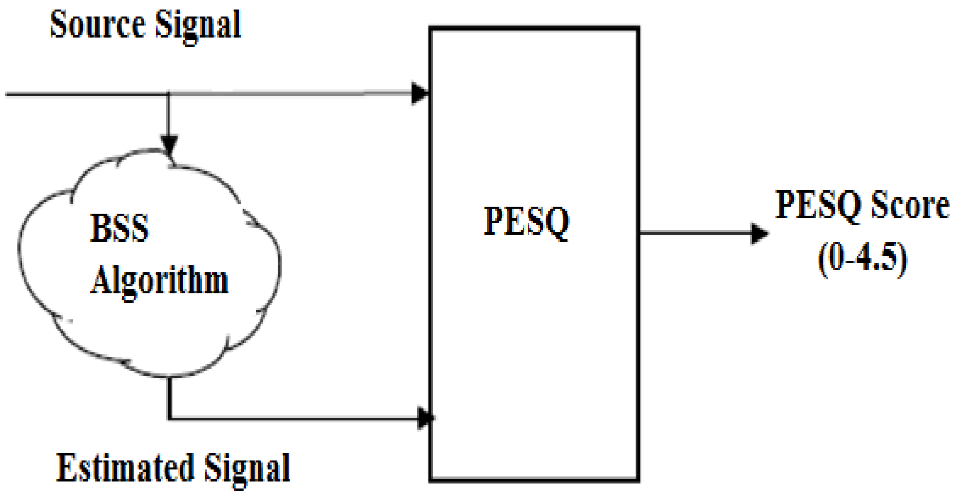

Figure 3.1: Measurement of PESQ score for the separated source signal

Different simulated and real room impulse responses are used to analyse and evaluate the separation performance of the BSS algorithms, which are discussed in the next section.

3.1.4 TIMIT Acoustic-Phonetic Continuous Speech Corpus

In order to evaluate BSS algorithms, the TIMIT dataset is extensively used throughout the experiments [79]. The TIMIT corpus is specifically designed to evaluate different automatic speech recognition systems and it has recordings of 6300 sentences by 630 native American speakers with eight different American English accents. These recordings were recorded by using a Sennheiser close talking microphone at the rate of 16kHz with 16 bit sample resolution. The length of the recordings varies roughly between 3 second to 7 seconds. The TIMIT dataset provides a universal speech library and is widely used to evaluate speech separation algorithms.

3.2 Room Impulse Responses

In the physical world, the observed mixtures are convolutive, which means that the observed mixtures have contributions from the delayed and weighted versions of the original signal. This is because of the reverberant physical environment as the signals can take several different paths and can suffer different levels of attenuations [80]. The reverberation time (RT_{60}) of a room is defined as the time period in which the energy of an impulse response is dropped below a certain threshold, which is usually considered as 60dB [81, 82]. The reverberant room environment can be modelled with room impulse responses (RIRs) [83]. These RIRs are used to model the acoustic pathways of the sound waves within an enclosed physical environment. In this thesis, three different types of RIRs have been used for experimentation and their details are as follows:

3.2.1 Image Method

The first technique for room impulse response generation is known as the image method. This method produces impulse responses which are artificially generated for an anechoic environment and can be used to simulate listening to a loudspeaker in an anechoic environment [84]. But this method lacks room related properties and produces very artificial RIRs; as in real life there are echoes and reverberations form the reflecting surfaces and walls of the room. However, the impulse responses generated by the image method can still be used to compare different algorithms and for proof of concept. Uncertainties in the measurement of the RIRs are further discussed in [85, 86]. RIRs that can better model the real room impulses and provides a better evaluation of separation performance of BSS algorithms in realistic scenarios are also considered in the experimentation.

3.2.2 Binaural Room Impulse Responses

In order to measure the separation performance of a BSS algorithm, a more natural and real room impulse model will be used, that is known as the binaural room impulse responses (BRIRs). It is an important tool for high-quality 3-D audio rendering. The BRIRs that are generated by Shinn are used [87]. These BRIRs are recorded in a real room environment by using a KEMAR (Knowles Electronics Manikin for Acoustic Research) dummy head to mimic the effect of a human head. This method produces a comprehensive database of real room recordings at different source location azimuths of $(15°, 30°, 45°, 60°, 75°, 90°)$ at distances of (0.15m, 0.40m and 1m) with reference to KEMAR dummy head. These BRIRs are recorded in a real classroom which roughly has dimensions of 5 x 9 x $3.5m^3$. The floor of the room was carpeted and the walls on three sides of the room are made of hard concrete while on the other side there is 9 m long sound absorptive partition. It also considers the distance between the speaker and microphone, so it models the 3-D acoustic space in a better way. The measurements for the BRIRs are taken at four different listener locations (back, ear, corner and center) and the distance between the floor and ears was approximately 1.50m. In this thesis only the center location is used and the RT_{60} at the center location for the classroom was 565ms. All the measurements for these BRIRs are repeated at three different occasions by taking down the equipment and reassembled, which improves the reliability of the measurements. The schematic of the room is shown in Figure 3.2.

These RIRs provide good evaluation of the BSS algorithm in a highly reverberant real room environment. In order to improve the credibility of the results another set of RIRs that have varying RT_{60} over four different rooms is also used for experimentation and details of these RIRs are discussed in the next section.

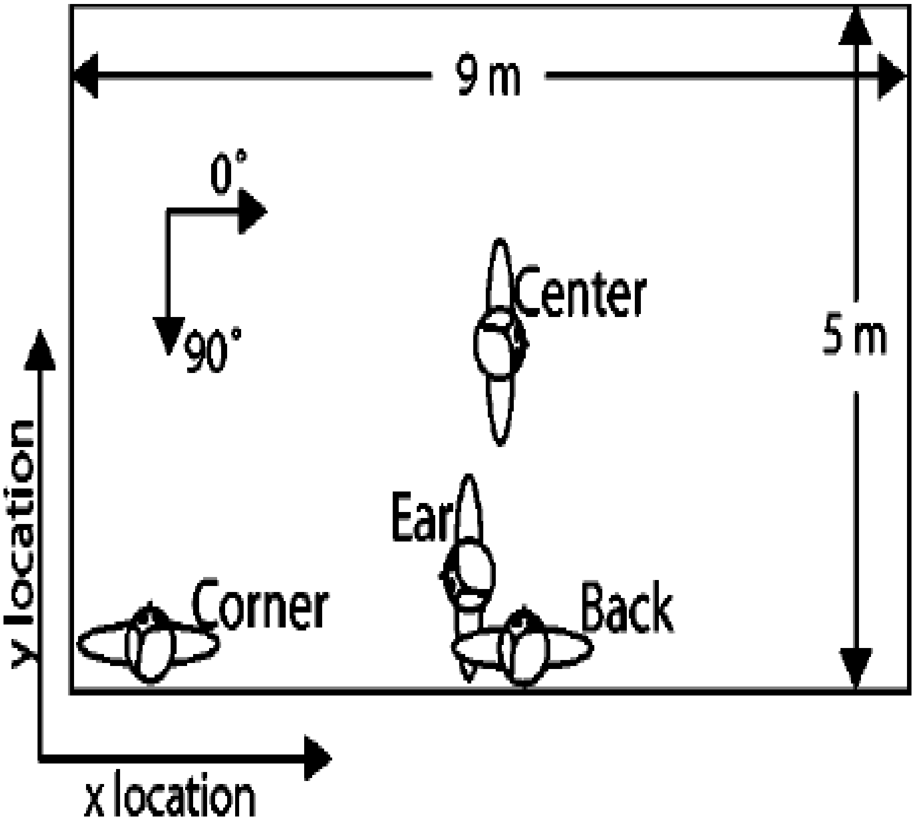

Figure 3.2: Rough orientation and layout for the classroom [87]. Measurements were taken at four different listener locations. Only center location is used in the experiments.

3.2.3 Real RIRs

The second set of real RIRs consists of four different room types and these real RIRs are obtained from [88]. These RIRs were generated by using a Cortex Instruments Mk.2 Head and Torso simulator (HATS) in real room environments. This method produces a complete database of real room recording in five different room types. The first of the rooms is an anechoic room and it is not used for experimentations in these thesis. The four other rooms are named as Room A, Room B, Room C and Room D and the details of RIRs generated in these rooms are as follows:

Room A

A medium-sized office was used to record the real RIRs at different source location azimuths that vary from ($-90°$ to $90°$) with reference to HATS. This room has relatively smaller RT_{60} of 320ms. In the experimentations source location azimuths of ($15°, 30°, 45°, 60°, 75°, 90°$) are considered for this room. A complete layout of this room along with its dimensions is shown in Figure 3.3

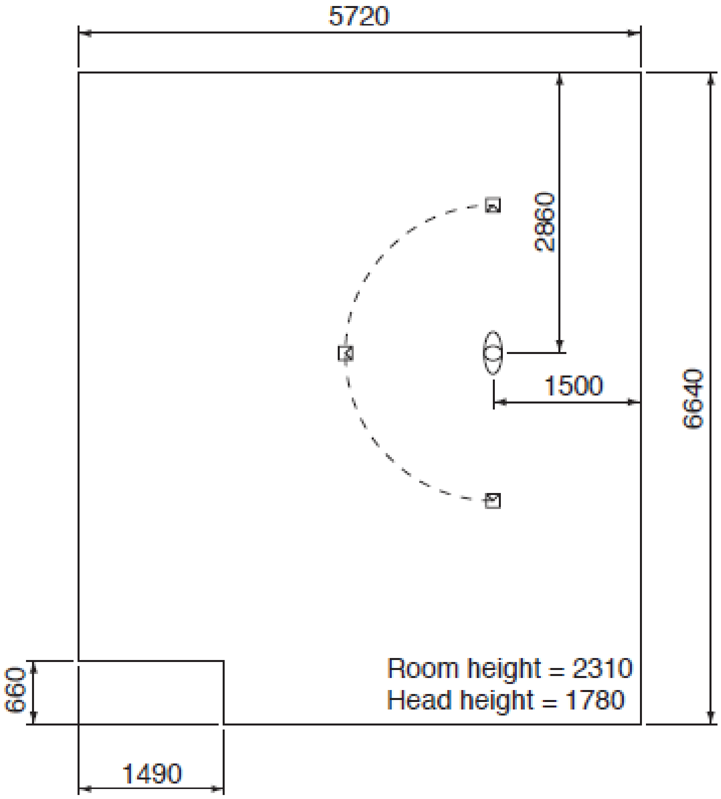

Figure 3.3: Layout and dimensions for Room-A (RT_{60} = 320ms) [88] .

Room B

A medium size classroom was used to record the real RIRs. For this room, $RT_{60} =$ 470ms, which is relatively high and it is therefore used to test the performance of the BSS algorithms in a reverberant real room environment. Source location azimuths of $(15°, 30°, 45°, 60°, 75°, 90°)$ are used in experiments. The layout of Room B is presented in Figure 3.4.

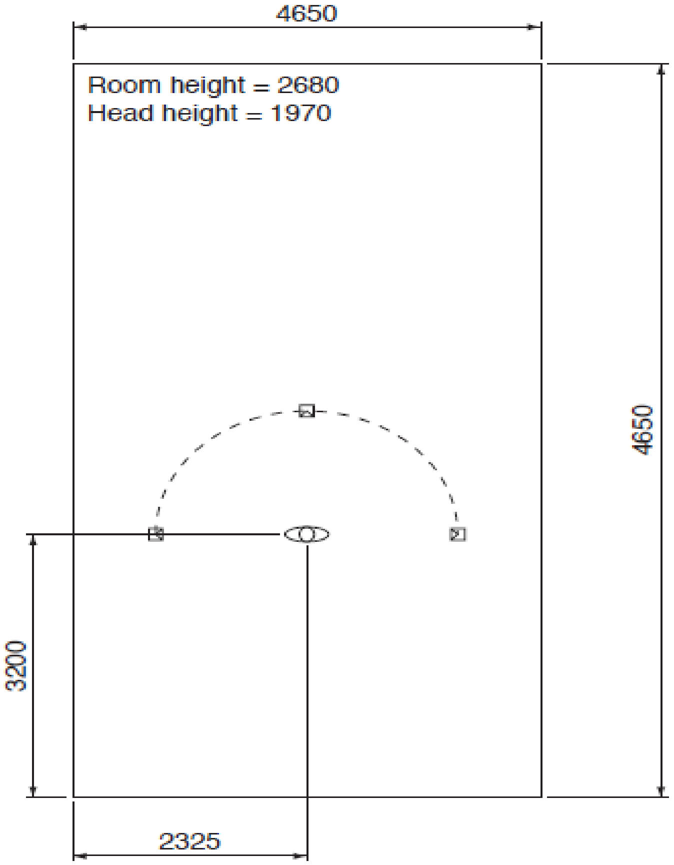

Figure 3.4: Plan elevation and dimensions for HATS and Room-B ($RT_{60} = 470$ms) [88].

Room C

A large lecture room with 418 seats was used to record the real RIRs and the RT_{60} for this particular lecture theatre is 680ms. A comprehensive layout for Room C is shown in Figure 3.5. In the experimentations, source location azimuths of (15° to 90°) are used in steps of 15°.

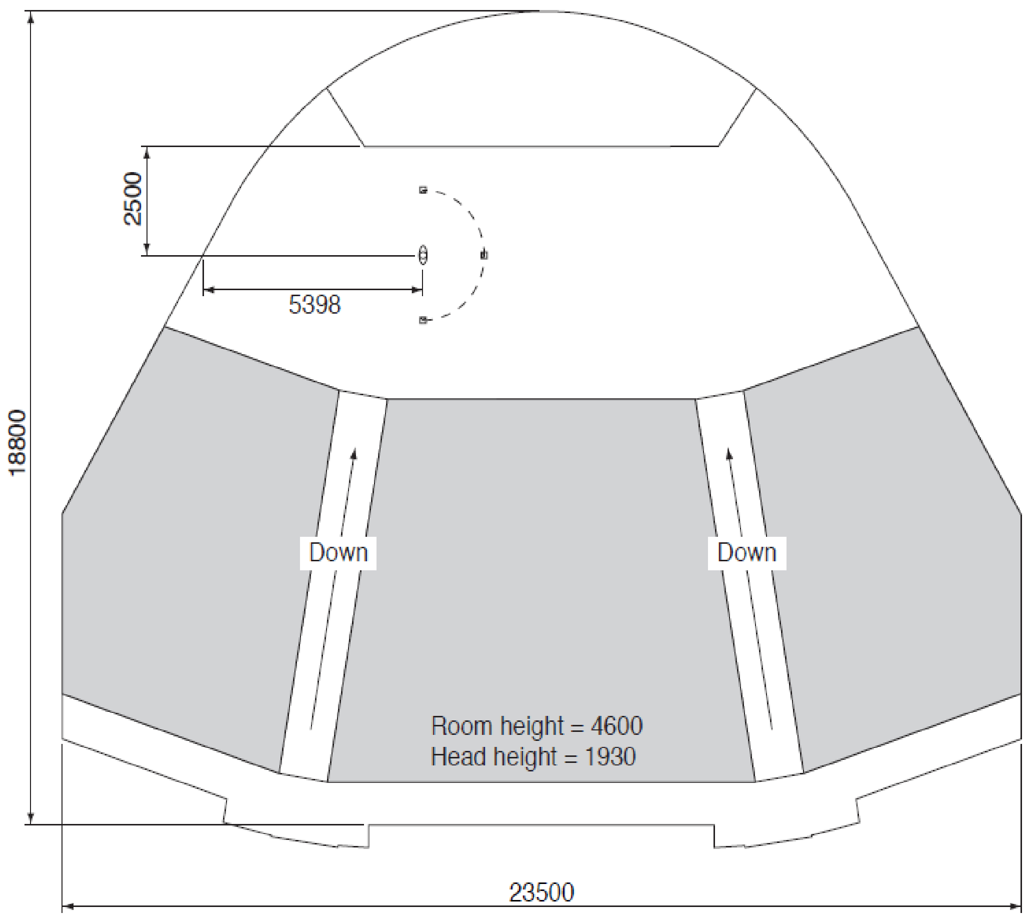

Figure 3.5: Layout and dimensions for Room-C (RT_{60} = 680ms). Shaded area represents the seating [88] .

Room D

The final set of RIRs is recorded in a typical large sized seminar hall which has high ceiling. This seminar hall has extremely high RT_{60} of 890ms. Therefore when different algorithms are evaluated by using the RIRs generated in this room, it provides accurate information about the algorithm performance in highly reverberant realistic environments. The layout for this seminar hall is shown in Figure 3.6.

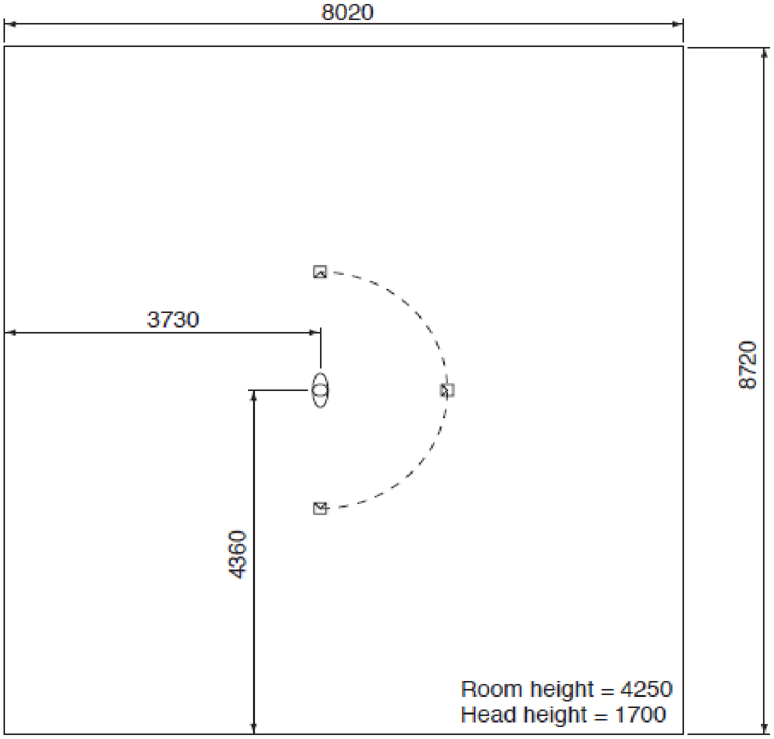

Figure 3.6: Layout and dimensions for Room-D (RT_{60} = 890ms). Shaded area represents the seating [88] .

Since the RIRs are generated in four different rooms with different RT_{60}, these RIRs are used for experimentation as they can provide the evaluation of algorithms in varying realistic conditions. An overview of the acoustic properties for all four rooms is included in Table 3.1

Table 3.1: Room types with the respective RT_{60} and direct-to-reverberant ratio (DRR).

Room	Type	RT_{60} (ms)	DRR (dB)
A	Medium office	320	6.09
B	Small class room	470	5.31
C	Large lecture room	680	8.82
D	Large seminar theatre	890	6.12

3.3 Summary

This chapter introduces different separation performance measures and the TIMIT dataset that are widely used in this thesis to evaluate the BSS algorithms. The chapter also discusses the concept of the reverberation time and its effect on the separation performance of BSS algorithms. This chapter also includes the different approaches by which RIRs are generated for the experimentation. These datasets included the simulated RIRs and also the real room recordings which can be used to test the separation performance of the algorithms in realistic scenarios. The source prior for the IVA and the FastIVA algorithms is crucial to the separation performance of the algorithm. Therefore a novel source prior for the IVA algorithms is presented in the next chapter.

Chapter 4

INDEPENDENT VECTOR ANALYSIS WITH A MULTIVARIATE STUDENT'S T SOURCE PRIOR

4.1 Introduction

Independent vector analysis (IVA) attempts to mitigate the permutation problem of the frequency domain ICA method. The IVA method exploits the higher order dependencies across the frequency bins and instead of the univariate distributions used by traditional frequency domain BSS methods, it describes each source prior vector with a dependent multivariate super Gaussian distribution. Such modelling preserves the higher order intra-vector source dependencies, namely the structural dependency between frequency bins of each source vector, while imposing the inter-vector source independence. Since the performance of the IVA method relies heavily on the source prior, a befitting source prior for speech signals is required for the IVA method [21].

In the past, various efforts have been made to improve the dependency structure of the IVA method. In [92], a chain type overlapped source prior is introduced to preserve the dependency structure. Whereas in [93], an online version of the IVA method is described which also uses a multivariate super Gaussian distribution to model the dependency structure. Another implementation of a multivariate super Gaussian source prior based IVA method in the time domain is proposed in [94]. Recently in [95], a multivariate generalized Gaussian source prior was adopted to model the dependency structure. However, in order to model the speech signals, a certain set of parameters is used to shape the source prior.

In this chapter, a multivariate Student's t source prior is proposed for the IVA and the FastIVA algorithms. The multivariate Student's t distribution is a super Gaussian distribution and it has heavier tails as compared with the multivariate Laplacian distribution, that is used as the source prior in the original IVA method introduced in [21]. Due to the heavy tail nature of the Student's t distribution, it is well suited to model certain types of speech signals [70, 72]. Since speech signals are highly non-stationary in nature. many useful samples can be of high amplitudes. Therefore the Student's t distribution with heavier tails can have a certain advantage in modelling the dependency between frequency bins of speech signals. The proposed Student's t source prior is implemented within the original IVA and the Fast-IVA method and the performance of both IVA methods is tested in real room environments. The detailed simulation studies will confirm the consistently improved separation performance of the proposed multivariate Student's t source prior.

4.2 Method

As discussed earlier, in practical situations due to reverberations, convolutive BSS methods are more appropriate and is discussed earlier they are generally implemented in the frequency domain. Hence, the noise free model in the frequency domain is described as:

$$\mathbf{x}(k) = \mathbf{A}(k)\mathbf{s}(k) \tag{4.1}$$

$$\hat{\mathbf{s}}(k) = \mathbf{W}(k)\mathbf{x}(k) \tag{4.2}$$

where $\mathbf{x}(k) = [x_1(k), x_2(k) \cdots x_m(k)]^T$ and $\hat{\mathbf{s}}(k) = [\hat{s}_1(k), \hat{s}_2(k) \cdots \hat{s}_n(k)]^T$ are the observed mixture signal vector and estimated signal vector both in the frequency domain, respectively, and $(.)^T$ denotes vector transpose. $\mathbf{A}(k)$ and $\mathbf{W}(k)$ are the mixing matrix and the unmixing matrix respectively. The dimension for $\mathbf{A}(k)$ is m x n and $\mathbf{W}(k)$ is n x m and the index k denotes the k-th frequency bin of this multivariate model. In this work focus is on the equally determined case, namely the 2 x 2 problem, i.e. two sources and two microphones.

In order to separate the multivariate observations, a multivariate cost function is needed. The Kullback-Leibler divergence between the joint pdf $p(\hat{s}_1 \cdots \hat{s}_n)$ and the product of pdf of the individual source vectors $\prod q(\hat{\mathbf{s}})$ is used for IVA [21].

$$\begin{aligned} J &= KL(p(\hat{s}_1 \cdots \hat{s}_n))||\prod q(\hat{\mathbf{s}})) \\ &= \int p(\hat{s}_1 \cdots \hat{s}_n)\log\frac{p(\hat{s}_1 \cdots \hat{s}_n)}{\prod q(\hat{\mathbf{s}})}d\hat{s}_1 \cdots d\hat{s}_n \\ &= \text{const} - \sum_{k=1}^{K} \log|det(W^{(k)})| - \sum_{i=1}^{n} E[\log q(\hat{s}_i)] \end{aligned} \tag{4.3}$$

where $E[\cdot]$ represents the statistical expectation operator, and $det(\cdot)$ is the matrix determinant operator. In the cost function, all sources are multivariate and it would be minimised when the vector sources are independent whereas the dependency between the components of each vector is still preserved. Therefore, the cost function can be used to remove the dependency between the vector sources and preserve the frequency dependency within each vector source.

4.3 Learning Algorithm: Gradient Descent Method

To minimise the above mentioned KL divergence in the cost function J, the gradient descent method is used. By differentiating the cost function J with respect to the coefficients of the separating matrices $w_{ij}(k)$, the gradients for the coefficients can be obtained as follows [21]:

$$\Delta w_{ij}(k) = -\frac{\partial J}{\partial w_{ij}(k)} = (w_{ij}(k))^{-H} - E[\varphi(k)(\hat{s}_i(1)\cdots\hat{s}_i(k))]x_j^*(k) \tag{4.4}$$

where $(\cdot)^*$ and $(\cdot)^H$ denote the conjugate and Hermitian transpose respectively. By multiplying both sides of the gradient matrices with $W(k)^H W(k)$ the natural gradient algorithm can be obtained.

$$\Delta w_{ij}(k) = \sum_{l=1}^{n} \Big(I_{il} - E\varphi(k)(\hat{s}_i(1)\cdots\hat{s}_i(k))\hat{s}_l^*(k) \Big) w_{lj}(k) \tag{4.5}$$

where I is an identity matrix with unity elements only when $i = l$ and 0 otherwise. The update rule is

$$w_{ij}(k)(t+1) = w_{ij}(k)(t) + \eta \Delta w_{ij}(k)(t) \tag{4.6}$$

where η is the learning rate. The nonlinear score function $\varphi(k)$ is:

$$\varphi(k)(\hat{s}_i(1)\cdots\hat{s}_i(k)) = -\frac{\partial \text{log} q(\hat{s}_i(1)\cdots\hat{s}_i(k))}{\partial \hat{s}_i(k)} \tag{4.7}$$

where $\varphi(k)(\hat{s}_i(1)\cdots\hat{s}_i(k))$ is a multivariate function. This score function preserves the dependency across the frequency bins. Since it is determined from the source prior, it is vital to establish a befitting multivariate source prior to retain the dependency structure as it can lead to good separation resutls. In the original IVA method, the source prior is defined by a multivariate Laplacian distribution and a simple and effective form of non-linear function can be obtained.

Since the score function is a multivariate function, it is therefore used to represent inter-frequency dependency between the sources. However, this score function is not unique and it depends on the types of sources. It can be adjusted according to the nature of the sources and finding the appropriate score function is crucial to the performance of the IVA method. Therefore the following section introduces a new multivariate Student's t source prior for the IVA algorithm.

4.3.1 Multivariate Student's t Source Prior

A multivariate Student's t distribution is proposed as a source prior for the IVA method, instead of the original super Gaussian distribution used in [21]. As human speech is complex and highly non-stationary in nature, it can have many high and low amplitude data points [68, 69]. This new Student's t source prior is well suited to model the dependency among the high amplitude information in frequency domain speech signals. The Student's t distribution due to its heavy tail nature can better model the information in the outliers [70, 71]. Therefore, when it is adopted as a source prior for the IVA method, it can better model the high amplitude information in the speech signals than the original super-Gaussian source prior.

The heavy tail nature of the Student's t distribution is confirmed in Figure 4.1. The Student's t distribution with different values of the degrees of freedom parameter is plotted and compared with the original super Gaussian distribution. It is also evident from Figure 4.1 that when the value of the degrees of freedom parameter in the Student's t distribution is decreased, the tails of distribution becomes heavier. For all the values of the degrees of freedom parameter, the Student's t distribution has heavier tails than the super Gaussian distribution. Also an example of the multivariate version of the Student's t distribution (two-dimensional) is shown in Figure 4.2.

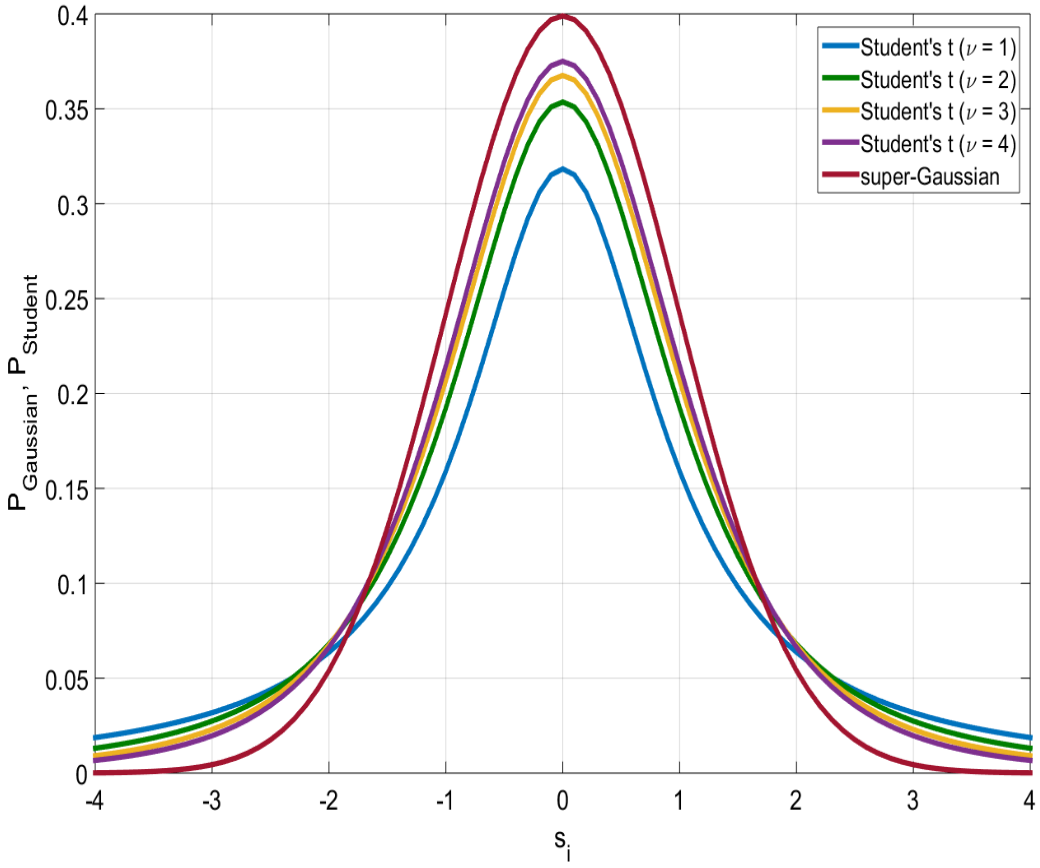

Figure 4.1: Comparison between the Student's t distribution and the super-Gaussian distribution as a function of the degrees of freedom parameter(ν). The Student's t distribution has heavier tails as compared to the super-Gaussian distribution.

The Student's t distribution is defined as follows. A K-dimensional random source vector $\mathbf{s} = (s_1, \ldots, s_K)^T$ is said to have a K-variate t distribution with degrees of freedom ν, mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{\Lambda}$, if its joint probability density function is given by [70]:

$$St(\mathbf{s}|\boldsymbol{\mu}, \boldsymbol{\Lambda}, \nu) = \frac{\Gamma(\frac{\nu+K}{2})|\boldsymbol{\Lambda}|^{1/2}}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\Big(1 + \frac{(\mathbf{s}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\mathbf{s}-\boldsymbol{\mu})}{\nu}\Big)^{-\frac{\nu+K}{2}} \tag{4.8}$$

where $\Gamma(.)$ is the Gamma function. The variance and the leptokurtic nature of the Student's t distribution can be tuned by varying the degrees of freedom

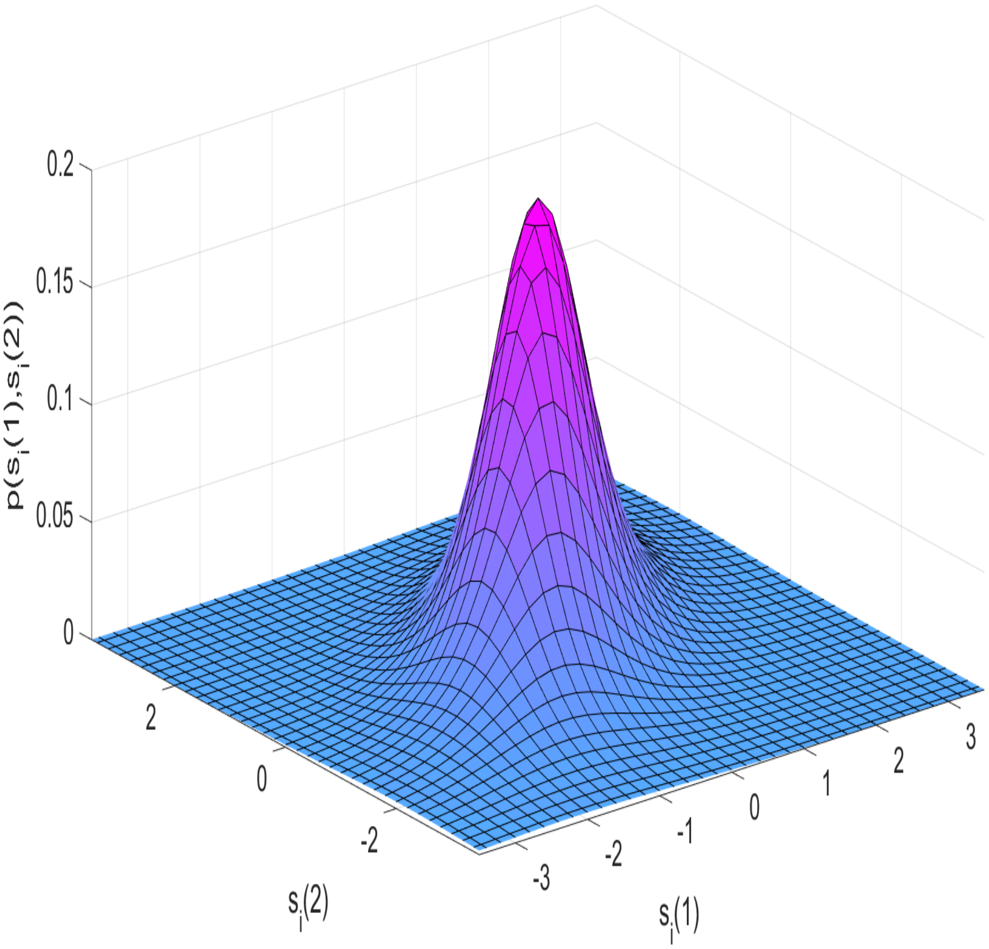

Figure 4.2: Bivariate version of Student's t distribution with degrees of freedom (ν) set to four.

parameter ν [72,74]. When degrees of freedom parameter is tuned to lower value, the tails of the distribution become heavier and if ν is increased to infinity, the Student's t distribution tends to a Gaussian distribution [75]. By ignoring the constant terms, the multivariate Student's t source prior for the IVA method can be represented as follows:

$$p(\mathbf{s}_i) \propto \left(1 + \frac{(\mathbf{s}_i - \mu_i)^T \mathbf{\Lambda}(\mathbf{s}_i - \mu_i)}{\nu}\right)^{-\frac{\nu+K}{2}} \tag{4.9}$$

The joint density function shown in Equation (4.9) is different from the product of the marginal distributions, which indicates that the multivariate Student's t has different variables which are dependent. Therefore, similar to the original super Gaussian multivariate source prior, the multivariate Student's t distribution can also retain the dependence among the frequency bins and can be used as source prior for the IVA method. In Equation (4.9), due to orthogonal Fourier bases, the covariance matrix is set to the identity matrix and zero mean is assumed, so it takes the following form:

$$p(\mathbf{s}_i) \propto \left(1 + \frac{\sum_{k=1}^{K} |s_i(k)|^2}{\nu}\right)^{-\frac{\nu+K}{2}} \tag{4.10}$$

The score function for the multivariate Student's t distribution can be obtained by replacing the source prior in the score function for the IVA algorithm in Equation (4.7) and with appropriate normalisation:

$$\varphi(k)(\hat{s}_i(1) \cdots \hat{s}_i(k)) = \frac{\hat{s}_i(k)}{1 + (\frac{1}{\nu}) \sum_{k=1}^{K} |\hat{s}_i(k)|^2} \tag{4.11}$$

The choice of the degrees of freedom in the new score function and the performance of the new source prior for the IVA method will be discussed in the next section.

4.3.2 Experimental Results

The performance of the new multivariate Student's t source prior with the original IVA method is tested in this section. The original IVA method with the proposed Student's t source prior will be evaluated in both the simulated and the real room environments and the results will be compared with the IVA method with the original super Gaussian source prior.

Experiments with Image Method

Firstly, room impulse responses (RIRs) were generated by the image method [84]. The speech signals for the experiments were randomly chosen from the whole TIMIT dataset [79] and the length of each signal was approximately four seconds. The size of the room was 7 x 5 x $3m^3$ and the reverberation time (RT_{60}) was 200ms. A 2x2 case was considered and the locations of microphones were [3.42 2.60 1.50]m and [3.48 2.60 1.50]m respectively. The STFT length was 1024 and the sampling frequency of 8 kHz was used. The separation performance of the multivariate source prior for the original IVA method was tested by using the objective measure of SDR [77]. The value of the degrees of freedom parameter for the Student's t source prior was chosen to be four, which is empirically found to be an appropriate choice. The mixtures were then separated by using the original IVA method with both the original super Gaussian source prior and the new Student's t source prior. The separation performance for both source priors is shown in Table 4.1 for ten different sets of mixtures.

Table 4.1: SDR (dB) values for both source priors with image room impulse response [84]. The Student's t source prior enhances the separation performance of the IVA algorithm for all mixtures.

	Original	Student's t	Improvement (dB)
Set-1	9.85	11.02	1.17
Set-2	9.98	10.90	0.92
Set-3	12.26	13.13	0.87
Set-4	11.02	11.91	0.89
Set-5	10.93	11.44	0.51
Set-6	11.08	11.62	0.54
Set-7	13.41	14.08	0.67
Set-8	10.22	10.97	0.75
Set-9	9.67	10.21	0.54
Set-10	12.08	12.81	0.73

Since RIRs were generated by the image method, which is a simulated method with $RT_{60} = 200$ms, the SDR values for both source priors are generally high. In comparison the Student's t source prior improves the separation performance of the original IVA method for all ten sets of mixtures. The average performance improvement by using the Student's t source prior for the original IVA method is approximately 0.7 dB, as shown in Table 4.1.

Furthermore, the separation performance of the IVA method with the new Student's t source prior will be tested in real room environments. The image room impulse response is a good simulated environment to compare the separation performance of different methods but it can not provide accurate evaluation of BSS methods in realistic scenarios. Therefore, in the next section, the IVA method with both source priors will be tested in a real room environment.

Experiments with Real Room Impulse Responses

Table 4.2: Parameter settings used in experiments.

Sampling rate	8kHz
STFT frame length	1024
Velocity of sound	343 m/s
Reverberation time	565 ms (BRIRs)
Room dimensions	9 m x 5 m x 3.5 m
Source signal duration	3.5 s (TIMIT)

In this section, separation performance of the IVA method is evaluated with binaural room impulse responses (BRIRs) which were recorded in a real classroom by Shinn-Cunningham [87]. The speech signals were again randomly chosen from the whole of the TIMIT data, which has length of approximately 4 seconds. The size of the classroom is 9 x 5 x 3.5 m^3. As mentioned earlier the measurements are taken at four different positions in the classroom (back, ear, corner and center).

For this particular set of experiments, measurements were considered from the center of the classroom and at this position $RT_{60} = 565$ms.
As these experiments are performed in a classroom with RT_{60} of 565ms, they examine the achieved performance of the algorithm in a difficult and highly reverberant real room environment. In these experiments six different source location azimuths relative to the second source were consider which vary over $(15°, 30°, 45°, 60°, 75°, 90°)$, this provides good evaluation of the separation performance with the changing positions in the real room environment. The summary of different parameters used for the experiments is given in Table 4.2.

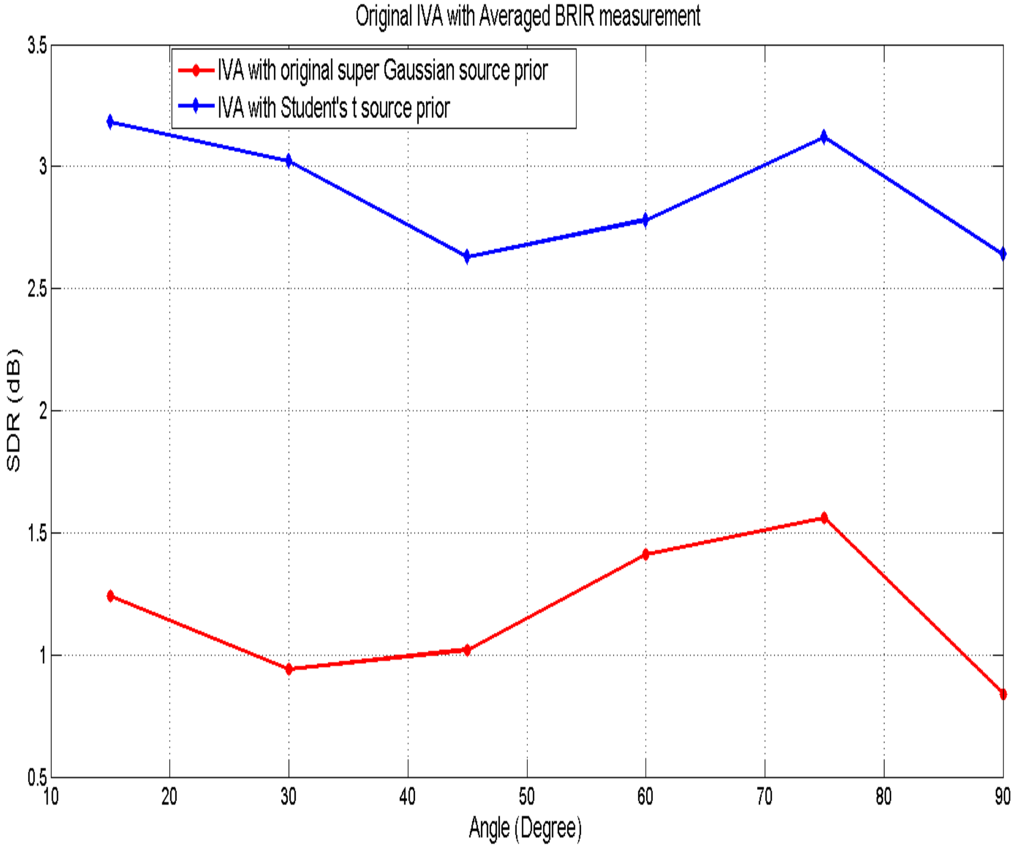

Figure 4.3: The graph indicates the separation performance with real BRIRs. The position of the listener was changed from 15 ° to 90 ° in steps of 15 °. The Student's t source prior enhance the separation performance of the IVA algorithm at all separation angles.

After the mixtures are created by using the BRIRs with high RT_{60} of 565ms, they are separated by using the original IVA method with both multivariate Student's t and the original super Gaussian source prior. The mixtures are separated at all six different azimuth angles available in the classroom. The degrees of freedom parameter is set to 4 as it is found empirically to be an appropriate choice. Figure 4.3 shows the separation performance of the IVA method with Student's t source prior and the comparison is made with the original super Gaussian source prior. All the experiments are repeated three times for reliability and at all six azimuth angles available in the room, SDR values are averaged for both speech signals. Figure 4.3 confirms the average improvement of 1.4 dB in the separation performance of the original gradient descent IVA algorithm with the multivariate Student's t source prior for this particular set of speech mixtures.

In order to establish the robustness of the multivariate Student's t source prior, further results for the separation performance of the IVA method for five different randomly chosen sets of speech signals are shown in Table 4.3. Again, the signals are separated at six source location azimuths varying from $(15°, 30°, 45°, 60°, 75°, 90°)$ with both source priors for the IVA algorithm. All the SDR values shown in Table 4.3 are averaged over six different angles. Table 4.3 confirms that the IVA method with multivariate Student's t source prior improves the average separation performance by approximately 0.8dB for all sets in highly reverberant real room environments.

Table 4.3: Separation results of the IVA algorithm with different source priors. All the SDR (dB) values shown are averaged over six different source locations. Student's t source prior for the IVA algorithms yields improvement for all the speech mixture.

	Original	Student's t	Improvement
Set-1	3.35	4.11	0.76
Set-2	4.03	4.85	0.82
Set-3	2.64	3.37	0.73
Set-4	3.05	4.13	1.08
Set-5	3.22	4.09	0.87

To gain further improvement in separation performance a fast converging algorithm is considered, which is discussed in the next section.

4.4 Learning Algorithm: Newton Method-FastIVA

A fast version of the IVA method (FastIVA) with the Newton's method as the learning algorithm will be discussed in this section. It is a rapidly converging form of IVA algorithm and is also known as fast fixed-point IVA(FastIVA). Newton's method converges quadratically and it is free from selecting an efficient learning rate.

The objective function used by FastIVA is as follows [58]:

$$J_{FastIVA} = \sum_{i=1}^{N} \Big[E[F(\sum_{k=1}^{K} |\hat{s}_i(k)|^2)] - \sum_{k=1}^{K} \lambda_i(k)(\mathbf{w}_i(k)^{\dagger}\mathbf{w}_i(k) - 1) \Big] \tag{4.12}$$

where $(.)^{\dagger}$ denotes a Hermitian transpose and $\mathbf{w}_i^{\dagger}$ is the i-th row of the unmixing matrix $\mathbf{W}$, and λ is the i-th Langrange multiplier. Equation (4.12) shows a multivariate function for the FastIVA algorithm and it is sum of the desired signals in all frequency bins.

The quadratic Taylor series polynomial will be used to implement the Newton's method in the update rule [22]. It will be introduced in complex variable notation to be used in the contrast function. The corresponding Taylor series expansion becomes

$$\begin{aligned} f(\mathbf{w}) \approx & f(\mathbf{w}_o) + \frac{\partial f(\mathbf{w}_o)}{\partial \mathbf{w}^T}(\mathbf{w}-\mathbf{w}_o) + \frac{\partial f(\mathbf{w}_o)}{\partial \mathbf{w}^\dagger}(\mathbf{w}-\mathbf{w}_o)^* \\ & + \frac{1}{2}(\mathbf{w}-\mathbf{w}_o)^T \frac{\partial^2 f(\mathbf{w}_o)}{\partial \mathbf{w} \partial \mathbf{w}^T}(\mathbf{w}-\mathbf{w}_o) \\ & + \frac{1}{2}(\mathbf{w}-\mathbf{w}_o)^\dagger \frac{\partial^2 f(\mathbf{w}_o)}{\partial \mathbf{w}^* \partial \mathbf{w}^\dagger}(\mathbf{w}-\mathbf{w}_o)^* + (\mathbf{w}-\mathbf{w}_o)^\dagger \frac{\partial^2 f(\mathbf{w}_o)}{\partial \mathbf{w}^* \partial \mathbf{w}^T}(\mathbf{w}-\mathbf{w}_o) \end{aligned} \tag{4.13}$$

To simplify the objective function, replace $\mathbf{w}_i(k)$ with $\mathbf{w}$ and consider $f(\mathbf{w}_i(k))$ to be the summation term of the Equation (4.12), thus

$$f(\mathbf{w}_i(k)) = E\Big[F(\sum_{k'=1}^{K}|\hat{s}_i(k')|^2)] - \sum_{k'=1}^{K}\lambda_i(k')(\mathbf{w}_i(k')^\dagger \mathbf{w}_i(k') - 1)\Big] \tag{4.14}$$

To optimise $f(\mathbf{w}_i(k))$ take the first derivative of $f(\mathbf{w}_i(k))$ and set it to zero, i.e.

$$\begin{aligned} \frac{\partial f(\mathbf{w}_i(k))}{\partial \mathbf{w}_i(k)^*} \approx & \frac{\partial f(\mathbf{w}_{i,o}(k))}{\partial \mathbf{w}_i(k)^*} + \frac{\partial^2 f(\mathbf{w}_{i,o})}{\partial(\mathbf{w}_i(k))^* \partial(\mathbf{w}_i(k))^T}(\mathbf{w}_i(k) - \mathbf{w}_{i,o}(k)) \\ & + \frac{\partial^2 f(\mathbf{w}_{i,o})}{\partial(\mathbf{w}_i(k))^* \partial(\mathbf{w}_i(k))^\dagger}(\mathbf{w}_i(k) - \mathbf{w}_{i,o}(k))^* = \mathbf{0} \end{aligned} \tag{4.15}$$

Then the derivative term in the above equation will become:

$$\frac{\partial f(\mathbf{w}_{i,o}(k))}{\partial \mathbf{w}_i(k)^*} = E\Big[\hat{s}_{i,o}(k)^* F'(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2)\Big] - \lambda_i(k)\mathbf{w}_{i,o}(k) \tag{4.16}$$

The second derivative of the above equation can be written as:

$$\begin{aligned}
\frac{\partial^2 f(\mathbf{w}_{i,o})}{\partial(\mathbf{w}_i(k))^* \partial(\mathbf{w}_i(k))^T} &= E\Big[(F'(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2) \\
&+ |\hat{s}_{i,o}(k)|^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2))\mathbf{x}(k)\mathbf{x}(k)^*\Big] - \lambda_i(k)I \\
&\approx E\Big[(F'(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2) + |\hat{s}_{i,o}(k)|^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k)|^2))]E[\mathbf{x}(k)\mathbf{x}(k)^*\Big] - \lambda_i(k)I \\
&= \Big(E\Big[(F'(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2) + |\hat{s}_{i,o}(k)|^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2))\Big] - \lambda_i(k)\Big)I
\end{aligned} \tag{4.17}$$

where $F'(\cdot)$ and $F''(\cdot)$ denote the first derivative and second derivative of $F(\cdot)$ respectively. By simplifying and due to the whitening process making the assumption of $E[\mathbf{x}(k)\mathbf{x}(k)^*] = I$ in Equation (4.17), it can be written as:

$$\begin{aligned}
\frac{\partial^2 f(\mathbf{w}_{i,o})}{\partial(\mathbf{w}_i(k))^* \partial(\mathbf{w}_i(k))^\dagger} &= E\Big[(\hat{s}_{i,o}(k)^*)^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2))\mathbf{x}(k)\mathbf{x}(k)^T\Big] \\
&\approx E\Big[(\hat{s}_{i,o}(k)^*)^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2))\Big] E\Big[\mathbf{x}(k)\mathbf{x}(k)^T\Big] \\
&= \mathbf{0}
\end{aligned} \tag{4.18}$$

where $E[\mathbf{x}(k)\mathbf{x}(k)^T] = 0$ because of the assumption of complex circularity. Now by using appropriate normalisation and simplification, the learning rule can be written as:

$$\begin{aligned}\mathbf{w}_i(k) \leftarrow & E\Big[F^{'}(\sum_{k'=1}^{K}|\hat{s}_{i,o}(k')|^2) + |\hat{s}_{i,o}(k)|^2 F^{''}(\sum_{k'=1}^{K}|\hat{s}_i(k')|^2))\Big]\mathbf{w}_i(k) \\ & - E\Big[(\hat{s}_{i,o}(k))^{*} F^{'}(\sum_{k'=1}^{K}|\hat{s}_{i,o}(k')|^2)\mathbf{x}(k)\Big]\end{aligned} \tag{4.19}$$

An unmixing matrix $\mathbf{W}(k)$ can be formed if the above equation is implemented for each source [22]. The unmixing matrix $\mathbf{W}(k)$ can be decorrelated by the symmetric decorrelation scheme as follows:

$$\mathbf{W}(k) \leftarrow (\mathbf{W}(k)(\mathbf{W}(k))^{\dagger})^{-1/2}\mathbf{W}(k). \tag{4.20}$$

4.4.1 FastIVA with Student's t Source Prior

As discussed earlier, the source prior is crucial to the performance of the IVA algorithm, therefore by choosing an appropriate source prior can ehnhance the separation performance of the IVA method. Also from earlier discussion, it is known that speech signals are highly non-stationary in nature and can have many useful samples with high amplitudes.

The Student's t distribution due to its heavy tail nature can better model the high amplitude data points in speech signals. So, the multivariate Student's t distribution is adopted as the source prior for the FastIVA method [76]. By using the multivariate Student's t distribution, the dependency within the source vectors can be preserved and because of the heavy tail nature of the Student's t distribution, it can improve the modelling of high amplitude information in different speech sources and thereby improve the separation performance of the FastIVA method.

For the FastIVA method, the non linear function F(.) is derived from the source prior. When the multivariate Student's t distribution as described in Equation (4.8) is adopted as the source prior in the FastIVA method, the non-lnear function can be found as follows. The multivariate Student's t distribution is adopted as the source prior for the FastIVA algorithm, namely

$$p(\mathbf{s}_i) \propto \left(1 + \frac{(\mathbf{s}_i - \mu_i)^T \mathbf{\Lambda} (\mathbf{s}_i - \mu_i)}{\nu}\right)^{-\frac{\nu+K}{2}} \tag{4.21}$$

and by using Equation (4.21), the non linear function can be calculated as:

$$F(\sum_{k'=1}^{K} |\hat{s}_i(k')|^2)' = \frac{\nu + K}{\nu} \left(1 + \frac{\sum_{k=1}^{K} |s_i(k)|^2}{\nu}\right)^{-\frac{\nu+K}{2}} \tag{4.22}$$

The leading coefficient $\frac{\nu+K}{\nu}$ can be absorbed in the step size in the update equation, therefore by normalising it to unity and with zero mean and unity variance assumption, Equation (4.22) can be written as:

$$F(\sum_{k'=1}^{K} |\hat{s}_i(k')|^2)'' = \frac{1 - \sum_{k'=1}^{K} |\hat{s}_i(k')|^2}{(1 + \sum_{k'=1}^{K} |\hat{s}_i(k')|^2)^2} \tag{4.23}$$

The above mentioned non-linear function is a multivariate function. Hence, this non-linear function can retain the interfrequency dependency as all the frequency bins are accounted for during the learning process. Also by changing the value of the degrees of freedom parameter ν, the tails of the distribution become heavier and therefore it can enhance the modelling of the information in the high amplitude data points in speech measurements. The separation performance of the proposed FastIVA algorithm will be evaluated in simulated and real room environments and the results will be compared with the FastIVA method using the original super Gaussian source prior in the next section.

4.4.2 Experimental Results

The separation performance of the FastIVA method with new Student's t source prior is evaluated in this section. Firstly, the proposed FastIVA algorithm is tested with the image method and then to evaluate the performance of the proposed method in the realistic scenarios, it is tested with the real room impulse responses.

Evaluation of FastIVA with Image Method

The proposed FastIVA method with multivariate Student's t source prior is firstly tested in a simulated environment generated with the image method [84]. Mostly, the experimental settings are similar to those in the case of experiments with the original IVA method. A 2 x 2 case is considered and the speech signals for the experiments are randomly selected from the whole of the TIMIT dataset [79]. The length of each speech signal is approximately four seconds. The STFT length is 1024 and the sampling frequency is 8kHz. The size of the room is 7 x 5 x 3 m^3 and the location of microphones are [3.42 2.50 1.50]m and [3.48 2.50 1.50]m respectively. The RT_{60} for these experimental settings is 200ms. The mixed signals are separated by using the FastIVA method with both the proposed Student's t source prior and the original super Gaussian source prior. The separation performance is measured in SDR and the results are shown in Table 4.4.

Table 4.4 shows the separation performance of the FastIVA method with both Student's t and the original super Gaussian source prior for ten different set of mixtures. All the SDR values shown in Table 4.4 are the average of two separated signals. In comparison the Student's t source prior based algorithm performs

Table 4.4: SDR (dB) values for FastIVA method with both source priors. Student's t source prior for the FastIVA method improves the separation performance for all the mixtures.

	Original (dB)	Student's t (dB)	Improvement (dB)
Set-1	10.88	12.02	1.14
Set-2	10.49	11.31	0.82
Set-3	12.76	13.73	0.97
Set-4	13.02	13.93	0.91
Set-5	11.84	12.59	0.75
Set-6	13.38	14.42	1.04
Set-7	13.47	14.28	0.81
Set-8	12.15	12.97	0.82
Set-9	10.66	11.42	0.76
Set-10	11.38	12.44	1.06

better than that using the original super Gaussian source prior for all set of mixtures, which is evident from the table. The average performance improvement in the SDR for the multivariate Student's t source prior is approximately 0.9 dB. The room impulse responses generated by the image method are helpful in comparing different methods but they can't evaluate the separation performance of BSS methods in the realistic scenarios. Therefore the separation performance of the proposed FastIVA method with the multivariate Student's t source prior in realistic scenarios is discussed in the next section.

Evaluation of FastIVA with the Real Room Impulse Responses

In this section the proposed FastIVA algorithm is evaluated in a real classroom environment by using the binaural room impulse responses (BRIR) generated by Shinn-Cunningham [87]. Experimental settings are kept the same as for the case of the original IVA method. Again, the centre location of the room is considered for these experiments and the RT60 = 565ms. As the RT_{60} is really high for this particular set of experiments therefore it provides good evaluation of the proposed algorithm in highly reverberant real room environment. Speech signals are randomly chosen from the whole of the TIMIT dataset [79]. A 2x2 case is consider and in order to consider the changing position of sources in real room environment, six different source location $(15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 90^\circ)$ azimuths relative to second source were considered. Furthermore, to improve the reliability of results, all the simulations are repeated three times. The summary of different parameters used for this set of experiments is given in Table 4.5

Table 4.5: Summary of parameters used in experiments.

Sampling rate	8kHz
STFT frame length	1024
Velocity of sound	343 m/s
Reverberation time	565 ms (BRIRs)
Room dimensions	9 m x 5 m x 3.5 m
Source signal duration	4 s (TIMIT)

In the first set of experiments, mixtures are created from the speech signals from the TIMIT dataset and by the impulse responses generated by BRIRs with RT_{60} of 565ms. These mixtures are then separated by using the proposed FastIVA method with the multivariate Student's t source prior and its separation performance is measured in SDR (dB) and the results are then compared with the FastIVA method with the original super Gaussian source prior. As benchmarks the basic FastICA [57] and intelligently initialised FastICA [89] are also included in comparisons and the results are shown in Figure 4.4. It is evident from Figure 4.4 that the FastIVA algorithm with proposed multivariate Student's t source prior performs better then the original super Gaussian source prior at all the separation angles. The FastICA and the intelligently initialised FastICA used for the separation of mixtures have poor separation performance in these experiments because of the permutation problem and also there is no pre or post processing used for these methods, which is generally needed in FastICA methods. All the SDR values shown in Figure 4.4 are averaged over eighteen random speech mixtures at all the separation angles that established the improved separation performance of the proposed multivariate Student's t source prior for the FastIVA method. Overall, the proposed source prior improves the separation performance of the FastIVA method by approximately 0.9dB.

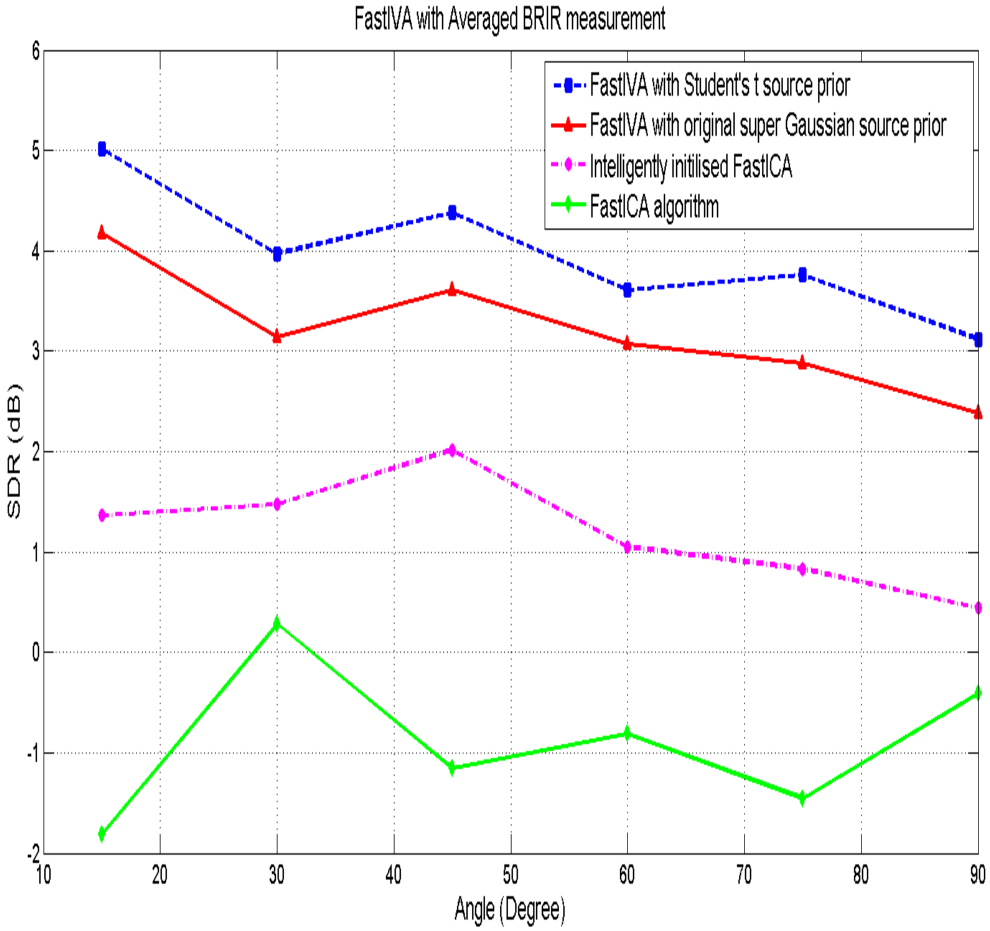

Figure 4.4: The graph indicates the separation performance of the FastIVA and FastICA algorithms. All the SDR (dB) values are averaged over eighteen random speech mixtures. The Student's t source prior enhance the separation performance of the FastIVA algorithm at all separation angles.

Generally, objective evaluations for real mixtures such as SDR are very useful in order to compare the performance of different methods but they can not portray the true quality of separated speech signals. Therefore in addition to the objective evaluation, the separation performance of the proposed Student's t source prior for the FastIVA method is also evaluated by using the subjective measure of perceptual evaluation of speech quality (PESQ) [78]. The same experimental settings were used as before and the mixtures were created by using the speech signals from TIMIT dataset in BRIRs. Then mixtures were separated by using the FastIVA method with the original super Gaussian source prior and the proposed Student's t source prior. Then the PESQ score is calculated for the separated signals from both methods by comparing the separated speech signals with the original speech signals. PESQ score is generally between 0 to 4.5, with 0 being poor score and the score of 4.5 is assigned to signal, that are almost identical. PESQ score for both source priors for five different set of mixtures is shown in Table 4.6 and for each set, PESQ score is averaged over six different locations in the room which are source azimuth angles varying over $(15°, 30°, 45°, 60°, 75°, 90°)$ relative to the second source which improves the reliability of the results. Table 4.6 indicates that the proposed multivariate Student's t source prior even in highly reverberant real room environment can consistently achieve better PESQ score than the original super Gaussian source prior for the FastIVA algorithm.

Table 4.6: PESQ score for the Student's t source prior and the Original Super Gaussian source prior for the FastIVA algorithm. All PESQ values are averaged for six different source locations and for all sets of mixtures, the Student's t source prior has better PESQ score.

	Original super Gaussian Source Prior	Student's t Source Prior
Set-1	1.65	1.81
Set-2	2.03	2.25
Set-3	2.14	2.29
Set-4	1.92	2.09
Set-5	2.05	2.16

Furthermore, the convergence speed of the FastIVA method was measured with the new Student's t source prior. Since the main purpose of introducing FastIVA method was to improve the convergence speed of the original IVA method, therefore it is vital to test the convergence speed for the proposed source prior for the FastIVA method. The same set of experiments was repeated in order to measure the convergence speed of the proposed method. The convergence speed was measured by counting the number of iterations that the FastIVA method was needed to converge as measured by changing likelihood of the algorithm. The convergence of the algorithm is calculated when the change of the norm of the weight matrix is less then 10^{-6} and it was measured for the FastIVA with both the new Student's t and the original super Gaussian source prior and the results are shown in Figure 4.5.

It is clear from Figure 4.5 that the FastIVA with new Student's t source prior converges swiftly as compared with the original super Gaussian source prior based FastIVA algorithm. For most of the angles the new Student's t based FastIVA method only needs almost half the number of iterations that were needed for the original super Gaussian based IVA method. The main purpose of the FastIVA method was to make the algorithm converge faster and the new Student's t source prior further improves the convergence speed of the FastIVA method, which is vital when using the algorithm in real time applications.

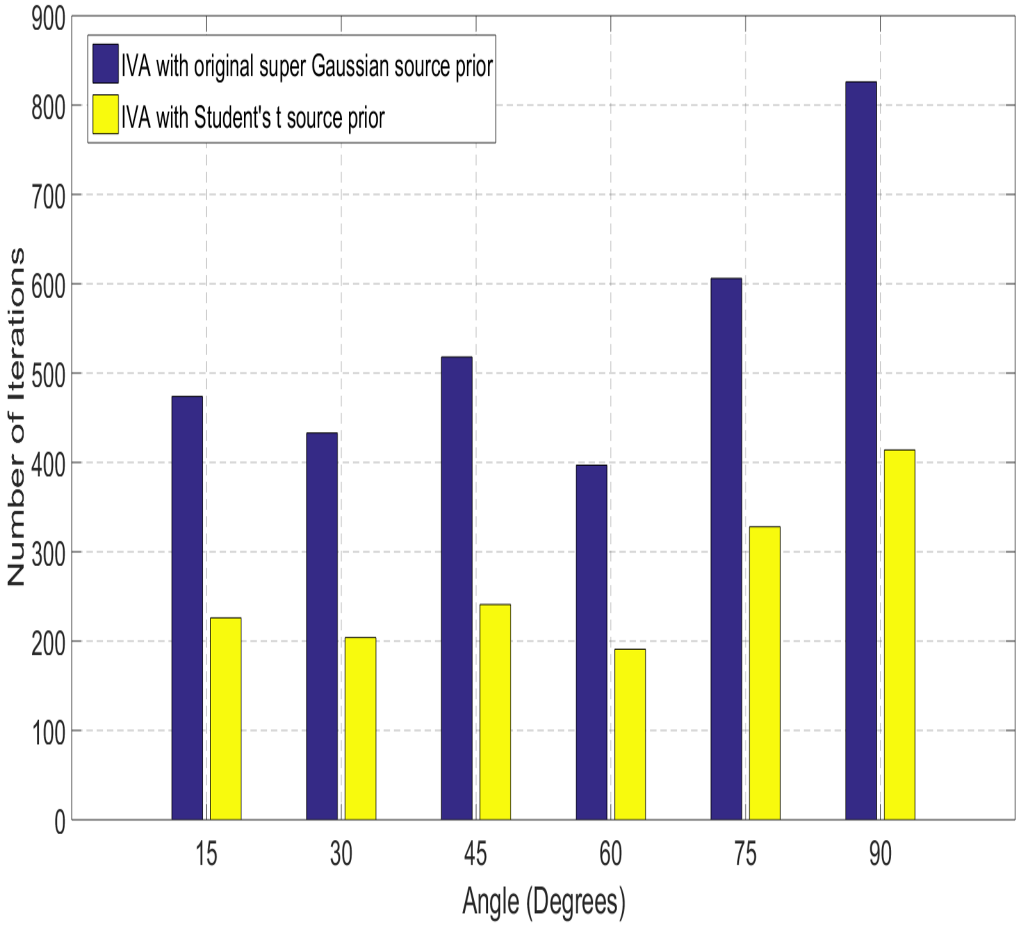

Figure 4.5: The number of iterations needed for the FastIVA algorithm to converge using both the original super Gaussian [21] and Student's t source priors in realistic RIRs is shown. The Student's t source prior at most angles need almost half the number of iterations.

4.5 Summary

In this chapter, a new multivariate Student's t source prior was introduced for the IVA and the FastIVA algorithm. The source prior for the IVA method is crucial to the performance of the algorithm as the non-linear score function is used to retain the inter-frequency dependency derived based on the PDF of the source. The multivariate Student's t distribution that belongs to the family of multivariate super Gaussian distributions is used in this work to model the high amplitude data points in speech signals. The multivariate Student's t distribution has heavier tails, thereby it can make use of the information lying in high amplitudes. Speech signals can have significant high amplitude data points such as voice sounds, therefore the multivariate Student's distribution is well suited to model the speech signals. Also, highly reverberant mixtures were used to evaluate the performance of the proposed source prior, which were more challenging to separate as compared to previous studies. The new experimental results in the highly reverberant real room environments, confirms that the proposed Student's t source prior consistently improves the separation performance of both the IVA and the FastIVA algorithm.

Chapter 5

ENERGY DRIVEN MIXED SOURCE PRIOR FOR THE INDEPENDENT VECTOR ANALYSIS ALGORITHM

The independent vector analysis algorithm preserves the dependency within each source vector to solve the permutation problem. Statistical models that can improve the dependency structure within each source vector are still needed to further improve the separation performance of the IVA method. As discussed in Chapter 4, in the past various statistical models have been proposed to improve the statistical dependence within the IVA method [94–96]. The multivariate source prior is important in all versions of the IVA algorithm, since it is used to derive the nonlinear score function and retain the dependency between different frequency bins [22].

In this chapter, a new enhanced multivariate source prior is introduced for the IVA algorithm. Instead of a conventional single distribution source prior, the proposed source prior is a mixture of the original multivariate super Gaussian distribution as in [21] and the multivariate Student's t distribution. The Student's t distribution is a super Gaussian distribution which has heavier tails and it can have a certain advantage when modelling speech signals. It is also stated in [72], that the Student's t distribution is well suited to model certain types of speech signals. Human speech is highly random in nature and can have many high and low amplitude components [12]. Therefore, the Student's t distribution due to its heavy tail nature can capture and model the information in high amplitude components in an efficient manner [72] and at the same time, the original super Gaussian distribution can be used to model the other data points in the speech signal. The contribution in this chapter is that the weight of both distributions can also be adjusted in the mixed source prior, which enables the source prior to adapt to different types of speech signals. The ratio of both distributions in the mixed source prior is adjusted according to the energy of the observed mixtures. Importantly, this method is found to be successful only when the observed mixtures are available and not the original sources. Moreover, to further enhance the separation performance of the proposed IVA algorithm, the fully connected frequency bin structure is decomposed into smaller groups as the neighbouring frequency bins generally have much stronger dependency as compared to distant frequency bins where the dependency is generally much weaker [23, 92]. Therefore, the strong dependency between neighbouring frequency bins is exploited by dividing them into smaller cliques whilst retaining considerable overlap between adjacent cliques. Furthermore, the new energy driven mixed source prior with clique based dependency structure is evaluated in real room environments and it consistently improves the separation performance of the IVA algorithm.

5.1 Source Prior for the IVA method

A new multivariate source prior that can better preserve the dependency structure within different frequency bins is needed to improve the separation performance of the IVA algorithm. Instead of a single distribution source prior, a mixture of original multivariate super Gaussian source prior and the multivariate Student's t distribution is found to be a suitable source prior for the IVA method.

The cost function for the original IVA algorithm is only minimised when the vector sources are independent while the dependency within the components of each vector is still preserved. Thus the cost function retains the inherent frequency dependency within each source vector, whilst removing the dependency among the sources [22]. When the cost function for the IVA method is minimised by using the gradient descent algorithm, the nonlinear function $\varphi(k)$ for source $\hat{s}_i$ is given as [21]:

$$\varphi(k)(\hat{s}_i(1)\cdots\hat{s}_i(k)\cdots\hat{s}_i(K)) = -\frac{\partial \text{log} q(\hat{s}_i(1)\cdots\hat{s}_i(k)\cdots\hat{s}_i(K))}{\partial\hat{s}_i(k)} \tag{5.1}$$

where $\varphi(k)(\hat{s}_i(1)\cdots\hat{s}_i(k)\cdots\hat{s}_i(K))$ is a multivariate score function and is used to preserve dependency across the frequency bins, denoted by index k. This nonlinearity represents the core idea of the IVA algorithm, as it is a multivariate function so it can preserve the dependency between different frequency bins. Since this multivariate score function is obtained from the source prior, it is vital to choose an appropriate multivariate source prior to retain the dependency structure for a better separation performance of the IVA algorithm.

In the original IVA method [21], the source prior representing the inter-frequency dependencies is a dependent multivariate super-Gaussian distribution and it can be derived as follows: Suppose a K dimensional random variable is explained by:

$$\mathbf{s}_i = \sqrt{v}.\mathbf{z}_i + \mu_i \tag{5.2}$$

where v is a scalar random variable, μ_i is a K-dimensional deterministic variable and $\mathbf{z}_i$ is a K-dimensional random vector. This random vector is assumed to have a Gaussian distribution with covariance matrix Σ_i and zero mean, that is

$$p(\mathbf{z}_i) = \alpha_z \exp\big(-\frac{\mathbf{z}_i^{\dagger}\Sigma_i^{-1}\mathbf{z}_i}{2}\big) \tag{5.3}$$

where $(.)^{\dagger}$ denotes a Hermitian transpose and α_z is a normalization term. Suppose that v has a Gamma distribution, that is:

$$p(v) = \alpha_v v^{\frac{K-1}{2}} \exp\big(-\frac{v}{2}\big) \tag{5.4}$$

where α_v is also a normalization term and the conditional distribution $p(s_i|v)$ is a Gaussian with mean μ and covariance σ_i. Therefore the original source prior can be obtained [21]:

$$\begin{aligned} p(\mathbf{s}_i) &= \int_0^{\infty} p(\mathbf{s}_i|v)p(v)dv \\ &= \alpha_1 \int_0^{\infty} \sqrt{v}\exp\Bigg(-\frac{1}{2}\Bigg(\frac{(\mathbf{s}_i-\mu_i)^{\dagger}\Sigma_i^{-1}(\mathbf{s}_i-\mu_i)}{v}+v\Bigg)\Bigg)dv \\ &= \alpha_2 \exp\Bigg(-\sqrt{(\mathbf{s}_i-\mu_i)^{\dagger}\Sigma_i^{-1}(\mathbf{s}_i-\mu_i)}\Bigg) \end{aligned} \tag{5.5}$$

Equation (5.5) shows there is a variance dependency between the frequency bins, which means when the variance of one frequency component is large then the variance for other frequency components will be large as well. Assumption of zero mean vector is taken and also the covariance matrix is assumed to be a diagonal matrix, since the frequency outputs are obtained by the orthogonal Fourier bases. So Equation (5.5) can be written as:

$$p(s_i) = \alpha \exp\Bigg(-\sqrt{\sum_{k=1}^{K}\Bigg|\frac{\hat{s}_i(k)}{\sigma_i(k)}\Bigg|}\Bigg) \tag{5.6}$$

where $\sigma_i(k)$ is the standard deviation of the *ith* source at the *kth* frequency bin.

By setting $\sigma_i(k)$ to unity, the multivariate score function can be written as [21]:

$$\varphi^{(k)}(\hat{s}_i(1)\ldots\hat{s}_i(K)) = -\frac{\partial \log\left(p(\hat{s}_i(1)\ldots\hat{s}_i(K)\right)}{\partial \hat{s}_i(k)} \\ = \frac{\partial\sqrt{\sum_{k'=1}^{K}\left|\hat{s}_i(k')\right|^2}}{\partial \hat{s}_i(k)} = \frac{\hat{s}_i(k)}{\sqrt{\sum_{k'=1}^{K}|\hat{s}_i(k')|^2}} \tag{5.7}$$

Equation (5.7) shows the multivariate score function for the original IVA method with the original super Gaussian multivariate source prior and it is used to represent inter-frequency dependency between sources. However, this score function is not unique and it depends on the types of sources. Therefore, a new source prior that can adapt according to different speech sources is needed and one is proposed in detail in the next section.

5.1.1 The Student's t Source Prior for the IVA method

The multivariate Student's t distribution is well suited to model certain types of speech signals [70]. As discussed in the Chapter 3, when the multivariate Student's t distribution is adopted as a source prior for the IVA algorithm, it takes the following form

$$p(\mathbf{s}_i) \propto \left(1 + \frac{(\mathbf{s}_i - \mu_i)^T \mathbf{\Lambda}(\mathbf{s}_i - \mu_i)}{\nu}\right)^{-\frac{\nu+K}{2}} \tag{5.8}$$

where μ and $\mathbf{\Lambda}$ are the mean and the precision matrix, respectively. The precision matrix is defined as the inverse of the covariance matrix $\mathbf{\Lambda} = \sum_i^{-1}$ and ν is the degrees of freedom parameter, which can tune the variance and the leptokurtic nature of the Student's t distribution [72]. The tails of the distribution becomes heavier when the degrees of freedom parameter ν approaches to zero which makes it more suitable for certain types of speech signals [70].

A score function for the original IVA method can be obtained by using Equation (5.8) and the multivariate Student's t distribution. Due to the orthogonal Fourier bases, the covariance matrix is set to the identity matrix and when zero mean is assumed, the source prior for the IVA method using the multivariate Student's t source prior can be obtained as follows:

$$\varphi(k)(\hat{s}_i(1)\cdots\hat{s}_i(K)) = \frac{\hat{s}_i(k)}{1+(\frac{1}{\nu})\sum_{k=1}^{K}|\hat{s}_i(k)|^2} \tag{5.9}$$

The separation performance of the IVA method can potentially be improved by using the combination of distributions as a source prior instead of a conventional single distribution source prior for the IVA method. Therefore a mixed source prior is proposed for the IVA method in the next section.

5.1.2 Mixed Source Prior for the IVA Method

Different speech source signals can have different statistical properties, therefore instead of modelling all speech sources by a single distribution, a mixed Student's t and original super Gaussian distribution source prior is proposed. The Student's t distribution because of its heavy tail nature can improve the modelling for the high amplitude information in the speech sources and the rest of the information can be better modelled with the original super Gaussian multivariate distribution [99].

The nonlinear score function for the IVA algorithm with new mixed multivariate Student's t and multivariate super Gaussian source prior takes following form

$$p(\mathbf{s}_i) = (\lambda_d).f_{St} + (1-\lambda_d).f_G \tag{5.10}$$

where f_G and f_{St} are the original multivariate super Gaussian distributions and the multivariate Student's t distribution, respectively; $\lambda_d \epsilon [0,1]$ is a weighting parameter and determines the ratio of each distribution in the mixed source prior. In the above equation, when the multivariate Student's t is replaced by using Equation (5.8) and the original multivariate super Gaussian is replaced by using Equation (5.5), it takes the following form.

$$\begin{aligned} p(\mathbf{s}_i) =& (\lambda_d)\left(1+\frac{(\mathbf{s}_i-\mu_i)^T\mathbf{\Lambda}(\mathbf{s}_i-\mu_i)}{\nu}\right)^{-\frac{\nu+K}{2}}+ \\ & (1-\lambda_d)\exp\left(-\sqrt{(\mathbf{s}_i-\mu_i)^\dagger\Sigma_i^{-1}(\mathbf{s}_i-\mu_i)}\right) \end{aligned} \tag{5.11}$$

The nonlinear score function for the IVA algorithm with the mixed Student's t and super Gaussian source prior can be obtained by Equation (5.11). Considering the zero mean case and due to Fourier bases, with the assumption the covariance matrix is set to identity and also by using appropriate normalisation, the overall non linear score function for source $\hat{\mathbf{s}}_i$ can be written as:

$$\begin{aligned} \varphi(k)(\hat{s}_i(1)\cdots\hat{s}_i(K)) \propto & (\lambda_d)\left(\frac{\hat{s}_i(k)}{1+\frac{1}{\nu}\sum_{k=1}^{K}|\hat{s}_i(k)|^2}\right) \\ & +(1-\lambda_d)\left(\frac{\hat{s}_i(k)}{\sqrt{\sum_{k=1}^{K}|\hat{s}_i(k)|^2}}\right) \end{aligned} \tag{5.12}$$

The nonlinear score function for the IVA method by using a mixed source prior is shown in Equation (5.12) and it is a multivariate function. Therefore, this score function can retain the inter-frequency dependency as all the frequency bins are accounted for during the learning process. The weights of both distribution in the source prior can be adjusted according to the speech signals by changing the value of $\lambda_d \epsilon [0, 1]$. When $\lambda_d = 1$, the mixed source prior only has the Student's t distribution and when the $\lambda_d = 0$, it only provides the original super Gaussian distribution as a source prior. The separation performance of this new mixed source prior within the original IVA method is evaluated in the next section.

5.1.3 Experimental Results

In this section, the new mixed source prior for the original IVA method is evaluated with two different room impulse responses. Firstly, it is evaluated with the simulated room impulse responses generated by the Image method [84], which are synthetic and can not provide proper evaluation of the algorithm in the real life context but these evaluations can be useful to compare the performance of different algorithms. Therefore, the proposed mixed source for the IVA method is further evaluated with the real binaural room impulse responses (BRIRs), which were recorded in the real classroom by Shinn, et al. [87]. These BRIRs are real room recordings with very high RT_{60} of 565ms, therefore they provide more accurate evaluation of the algorithm in realistic scenarios. The proposed mixed source prior IVA method is evaluated with both real and synthetic room impulse responses and the results are compared with the original IVA method with the original super Gaussian source prior [21].

Evaluation with Image Method

For the first set of experiments, room impulse response are generated by using the Image method [84]. The size of the room was 7 x 5 x $3m^3$ and the microphone sources were positioned at $[3.48, 2.50, 1.50]m$ and $[3.44, 2.50, 1.50]m$ respectively. The STFT length was set to 1024 and the sampling frequency was 8kHz. In these experiments the RT_{60} was set to 200ms. Two different speech signals of length of approximately four seconds were chosen randomly from the whole of the TIMIT dataset [79] and convolved into two mixtures. These mixtures were then separated by using the IVA algorithm with new mixed Student's t and the original super Gaussian source prior. The separation performance of the method was measured by using the objective measure of signal to distortion ratio (SDR) in decibels (dB). The value of the degrees of freedom in the mixed source prior was chosen to be four, which is empirically found to give the best separation performance. The weighting parameter $\lambda_d = 0.5$ was used in the mixed source prior as it will assign equal weight to both the Student's t and the original super Gaussian distribution in the mixed source prior.

Table 5.1: SDR (dB) values for both source priors for the original IVA method with Image room impulse response [84]. The SDR (dB) values are average for six different positions for each mixture. The mixed Student's t-original super Gaussian source prior shows considerable improvement for all mixtures.

	As in [22]	Mixed Source Prior	Improvement (dB)
Set-1	9.24	10.38	1.04
Set-2	8.33	9.21	0.88
Set-3	9.11	9.94	0.83
Set-4	8.85	9.77	0.92
Set-5	8.48	9.32	0.84

The separation performance of the IVA algorithm with the new mixed source prior is shown in Table 5.1 and the results are compared with the original IVA method [21]. All the values shown are the average separation performance of both source priors at six different positions in the same room. All the SDR values are generally high as the room impulse responses are simulated at relatively lower RT_{60} of 200ms. For the experiments, five different sets of speech signals were randomly chosen from the TIMIT dataset and then convolved into mixtures in the room impulse responses generated with the Image method [84]. For all the five sets of speech signals, the new mixed Student's t and original super

Gaussian source prior based IVA method has better separation performance than the original IVA method. The results in Table 5.1 confirm that the new mixed source prior improves the average separation performance of the IVA method by approximately 0.9 dB.

Evaluation with Real BRIRs

In this section, the IVA method with new mixed source prior is evaluated with BRIRs, which were obtained from [87]. These BRIRs are real room recording, therefore they provide more representative information about the separation performance of the algorithms in realistic scenarios. Two speech signals were randomly chosen from the whole of the TIMIT dataset [79] and convolved into mixtures. The size of the classroom is 9 x 5 x $3.5m^3$ and six different source location azimuths relative to the second source were considered. These source location azimuths vary over $(15°, 30°, 45°, 60°, 75°, 90°)$ which provide good evaluation of separation performance of the algorithm with changing positions in the room, representative of when speakers move around in the room. All the experiments at all the source location azimuths were repeated three times to improve the reliability of the results. The RT_{60} for the classroom is 565ms, which is relatively high and provides a good evaluation of the separation performance of BSS algorithms in highly reverberant realistic scenarios. The value of the degrees of freedom parameter in the mixed source prior was again chosen to be four, which was found empirically to yield best separation. The weighting parameter λ_d was set to 0.5, as it assigns equal weight to both the Student's t and the original super Gaussian distribution in the mixed source prior. The summary of different parameters used in the experiments is provided in Table 5.2.

The speech mixtures were created by using the BRIRs with high RT_{60} of 565ms and then separated by using the IVA algorithm with the new mixed multivariate

Table 5.2: Different parameters in experiments.

Sampling rate	8kHz
STFT frame length	1024
Weighting parameter	0.5
Degrees of freedom	4
Reverberation time	565 ms (BRIRs)
Room dimensions	9 m x 5 m x 3.5 m
Source signal duration	4 s (TIMIT)

Student's t-original super Gaussian source prior and the results were compared with the original IVA method as in [21]. The separation performance in terms of SDR for both the source priors for the IVA method for six different source location azimuths $(15°, 30°, 45°, 60°, 75°, 90°)$ is shown in Figure 5.1. In the bar plot, blue bars represent the separation performance of the original IVA method [21] and yellow bars represent the separation performance of the proposed mixed source prior based IVA method. In order to improve the reliability of results, all the SDR values at each angle, shown in Figure 5.1 were averaged over the separation performance for eighteen speech mixtures. It is evident from Figure 5.1, that for all the separation angles the IVA method with new mixed Student's t and original super Gaussian source prior has better separation performance as compared to the original IVA method. The average improvement of 0.85 dB is recorded at all the separation angles when the new mixed source prior is adopted for the IVA method. For the separation angles of $15°, 30°$ and $90°$ the SDR values for both the methods are lower because of difficult contexts but even with these difficult separation angles the new mixed source prior performs better than the original super Gaussian source prior. The new mixed Student's t and original super Gaussian source prior is also adopted for the fast version of the IVA algorithm and it is discussed in detail in the next section.

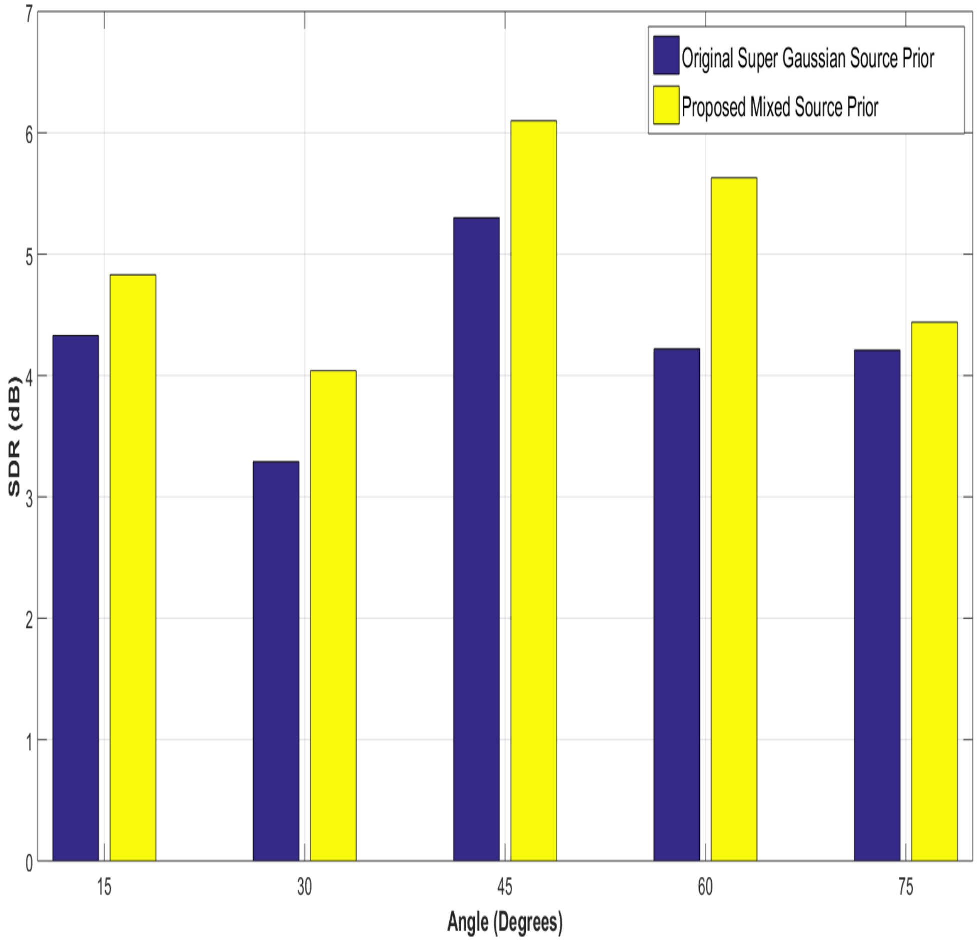

Figure 5.1: The graph shows the SDR (dB) values at six different separation angles. Real BRIRs from [87] were used. Results were averaged over eighteen mixtures. Mixed Student's t-original super Gaussian source prior enhance the separation performance of the IVA algorithm at all separation angles.

5.2 The Mixed Source Prior for the FastIVA algorithm

The proposed mixed Student's t-original super Gaussian source prior is also adopted as a source prior for the FastIVA method. The FastIVA algorithm is the fast converging version of the IVA method and it uses Newton's method in the learning process which can converge quadratically. The objective function used by the FastIVA algorithm is given as [58]:

$$J_{FastIVA} = \sum_{i=1}^{N} \Big[E[F(\sum_{k=1}^{K} |\hat{s}_i(k)|^2)] - \sum_{k=1}^{K} \lambda_i(k)(\mathbf{w}_i(k)^{\dagger}\mathbf{w}_i(k) - 1) \Big] \tag{5.13}$$

where $\mathbf{w}_i^{\dagger}$ is the i-th row of the unmixing matrix $\mathbf{W}$, and λ is the i-th Langrange multiplier. Also in Equation (5.13), $F(\cdot)$ represents the nonlinear function which is the summation of the desired signals in all frequency bins. This nonlinear score function can take several different forms as shown in [58]. By using the appropriate normalisation and the derivation discussed in the Chapter 3, the learning rule for the FastIVA method can be written as:

$$\begin{aligned} \mathbf{w}_i(k) \leftarrow & E\Big[F^{'}(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2) + |\hat{s}_{i,o}(k)|^2 F^{''}(\sum_{k'=1}^{K} |\hat{s}_i(k')|^2))\Big]\mathbf{w}_i(k) \\ & - E\Big[(\hat{s}_{i,o}(k))^* F^{'}(\sum_{k'=1}^{K} |\hat{s}_{i,o}(k')|^2)\mathbf{x}(k)\Big] \end{aligned} \tag{5.14}$$

where $F^{'}(\cdot)$ and $F^{''}(\cdot)$ represent the first and the second derivative of $F(\cdot)$ respectively. When the learning rule is used for all the sources, an unmixing matrix $\mathbf{W(k)}$ can be constructed which needs to be uncorrelated as follows

$$\mathbf{W}(k) \leftarrow (\mathbf{W}(k)(\mathbf{W}(k))^{\dagger})^{-1/2}\mathbf{W}(k). \tag{5.15}$$

The nonlinear score function $F(\cdot)$ is derived from the source prior and it is crucial to the separation performance of the algorithm. It can take several different forms according to the source prior. In [58], a particular super Gaussian distribution is used as a source prior for the FastIVA algorithm. When the variance is assumed to be unity and the mean is consider to be zero, this particular super Gaussian source prior can be represented as:

$$F\left(\sum_{k'=1}^{K}|\hat{s}_i(k)|^2\right) = \sqrt{\sum_{k'=1}^{K}|\hat{s}_i(k')|^2} \tag{5.16}$$

With the appropriate normalisation and by using the super Gaussian source prior mentioned in equation (5.16), the nonlinear score function for the FastIVA method by using the original super Gaussian distribution as a source prior can be derived as follows:

$$F''\left(\sum_{k'=1}^{K}|\hat{s}_i(k)|^2\right) = \left(\frac{1}{\sqrt{\sum_{k'=1}^{K}|\hat{s}_i(k')|^2}}\right)^3 \tag{5.17}$$

The separation performance of the FastIVA method can be further improved by carefully selecting an appropriate source prior.

5.2.1 The multivariate Student's t source prior for the FastIVA method

As shown in the Chapter 3, when the Student's t distribution is adopted as a source prior for the FastIVA method, it takes the following form.

$$p(\mathbf{s}_i) \propto \left(1 + \frac{(\mathbf{s}_i - \mu_i)^T \mathbf{\Lambda}(\mathbf{s}_i - \mu_i)}{\nu}\right)^{-\frac{\nu+K}{2}} \tag{5.18}$$

The nonlinear score function for the FastIVA method with multivariate Student's t distribution as a source prior can be calculated by using Equation (5.18). When the covariance matrix is set to zero due to Fourier bases, the mean is assumed to be zero and with appropriate normalisation, the nonlinear score function for the multivariate Student's t source prior based FastIVA algorithm can be written as:

$$F''(\sum_{k'=1}^{K}|\hat{s}_i(k')|^2) = \frac{1-\sum_{k'=1}^{K}|\hat{s}_i(k')|^2}{(1+\sum_{k'=1}^{K}|\hat{s}_i(k')|^2)^2} \tag{5.19}$$

The nonlinear score function obtained from Equation (5.19) is a multivariate function. Therefore, this nonlinear score function will preserve the inter-frequency dependency as all the frequency bins are accounted for during the learning process. In order to further improve the separation performance of the FastIVA method, more appropriate source prior that can better model all the samples in the speech signals is still needed and instead of using a single distribution source prior, a mixed distribution source prior is again proposed in the next section.

5.2.2 Mixed source prior for the FastIVA Method

A mixture of the multivariate Student's t and original multivariate super Gaussian source prior is also adopted as a source prior for the FastIVA method [91]. The Student's t distribution part in the mixed source prior can account for the high amplitude samples in speech signals and the other samples can be modelled with the original super Gaussian distribution. When this new mixed source prior is adopted for the FastIVA method, the nonlinear score function in general form can be written as:

$$p(\mathbf{s}_i) = (\lambda_d).f_{St} + (1-\lambda_d).f_G \tag{5.20}$$

When f_{St} is replaced by the multivariate Student's t from Equation (5.18) and f_G is replaced with the original super Gaussian from Equation (5.16), the general equation for nonlinear score function takes the following form.

$$
\begin{aligned}
p(\mathbf{s}_i) =&(\lambda_d)\Bigg(1+\frac{(\mathbf{s}_i-\mu_i)^T\mathbf{\Lambda}(\mathbf{s}_i-\mu_i)}{\nu}\Bigg)^{-\frac{\nu+K}{2}}\\
&+(1-\lambda_d)\Bigg(\sqrt{\sum_{k'=1}^{K}|\hat{s}_i(k')|^2}\Bigg)
\end{aligned}
\tag{5.21}
$$

The nonlinear score function for the FastIVA method can be obtained by using the mixed multivariate Student's t and original super Gaussian source prior shown in equation (5.21). By using the appropriate normalisation, the overall score function for the FastIVA method with new mixed source prior for source $\hat{s}_i$ can be obtained as:

$$
\begin{aligned}
F''(\sum_{k'=1}^{K}|\hat{s}_i(k')|^2) =&(\lambda_d)\Bigg(\frac{1-\sum_{k'=1}^{K}|\hat{s}_i(k')|^2}{(1+\sum_{k'=1}^{K}|\hat{s}_i(k')|^2)^2}\Bigg)\\
&+(1-\lambda_d)\Bigg(\frac{1}{\sqrt{\sum_{k'=1}^{K}|\hat{s}_i(k')|^2}}\Bigg)^3
\end{aligned}
\tag{5.22}
$$

Equation (5.22) represents the nonlinear score function for the FastIVA algorithm and it has λ_d as a weighting parameter, which can be used to adjust the ratio of both distributions in the mixed source prior and cater for different types of speech signals. The separation performance of the FastIVA method with this new mixed multivariate Student's t-original super Gaussian source prior is evaluated in the next section.

5.2.3 Experimental Results for the Mixed Source Prior FastIVA algorithm

The separation performance of the FastIVA algorithm with new mixed source prior is evaluated in simulated and realistic scenarios in this section. Again, to compare the performance of different methods, firstly the Image method [84] was used to generate the room impulse responses. In order to test the performance of the algorithm in the realistic environment, BRIRs [87] were also used to generate the room impulse responses, as they are real recordings in a classroom and provide a good evaluation in realistic scenarios.

Evaluation with Image Method

The same set of experiments were repeated for the FastIVA method as in the case of the original IVA method with new mixed source prior. The room impulse responses were again generated by using the Image method [84] and randomly chosen speech signals from TIMIT dataset [79] were convolved into mixtures.

	As in [58]	Mixed Source Prior	Improvement (dB)
Set-1	9.44	10.36	0.92
Set-2	9.75	10.82	1.07
Set-3	10.36	11.32	0.96
Set-4	10.18	11.76	1.58
Set-5	9.82	11.06	1.24

Table 5.3: The table indicates the improvement in separation performance of the FastIVA algorithm in terms of SDR (dB) for five speech mixtures using the Image method [84]. For each set of speech signals the SDR values are averaged for both estimated source signals.

In this set of experiments, five different speech mixtures were separated by using the new mixed source prior FastIVA method and the original FastIVA method [58]. The separation performance for both methods for all five speech mixtures is shown in Table 5.3 and for all the mixtures, the proposed mixed source prior improves the separation performance of the FastIVA method. All the values shown in Table 5.3 are the average SDR value in dB for both separated speech signals. When the new mixed Student's t-original super Gaussian source prior is adopted as a source prior for the FastIVA method, it improves the average separation performance of the FastIVA method by approximately 1.10 dB, as shown in Table 5.3.

Evaluation with Real BRIRs

In this subsection, the FastIVA method with new mixed Student's t-original superGaussian source prior is evaluated in realistic scenarios by generating the room impulse responses by BRIRs, which were obtained from [87]. The speech signals for these experiments were again randomly chosen from the whole of the TIMIT dataset [79] and the length of speech signals was approximately 5 seconds. The same experimental setting was used for these experiments as in the case of the original IVA method with mixed source prior. As these BRIRs were recorded in a real classroom therefore they have relatively high RT_{60} of 565ms. Hence it provides good evaluation of the separation performance of the algorithms in highly reverberant real life scenarios. The speech mixtures created with the BRIRs are then separated by using both the new mixed source prior FastIVA method and the original FastIVA method [58]. The weighting parameter λ_d is set to 0.5 as it will assign equal weight to both distributions in the mixed source prior. The common parameters used in this set of experiments are summarised in Table 5.4.

Table 5.4: Different parameters in experiments.

Sampling rate	8kHz
STFT frame length	1024
Reverberation time	565 ms (BRIRs)
Room dimensions	9 m x 5 m x 3.5 m
Source signal duration	5 s (TIMIT)
Weighting parameter	0.5
Degrees of freedom	4

In order to evaluate the separation performance of the proposed mixed source prior FastIVA method, six speech mixtures were separated at five different source location azimuths varying from 15° to 75° in steps of 15°. All the measurements in this set of experiments were repeated three times in order to improve the reliability of the results. The separation performance of both the methods for five different position is shown in Figure 5.2. All the SDR values shown are in dB and for both the proposed and the original FastIVA method, at all the source location azimuths, all the SDR values are the average separation performance of eighteen random speech mixtures. The better separation performance of the new mixed source prior is evident from Figure 5.2, as for all the different positions in the real room environment with RT_{60} = 565ms, the mixed source prior yields better separation performance. On average the new mixed source prior improves the separation performance of the FastIVA method by 0.90 dB in the realistic scenarios.

The performance of the proposed mixed Student's t-original super Gaussian source prior can be further improved by changing the weight of the distributions in the mixed source prior according to the nature of speech signals. Therefore, a new energy driven source prior that can adapt to different speech mixtures is proposed in the next section.

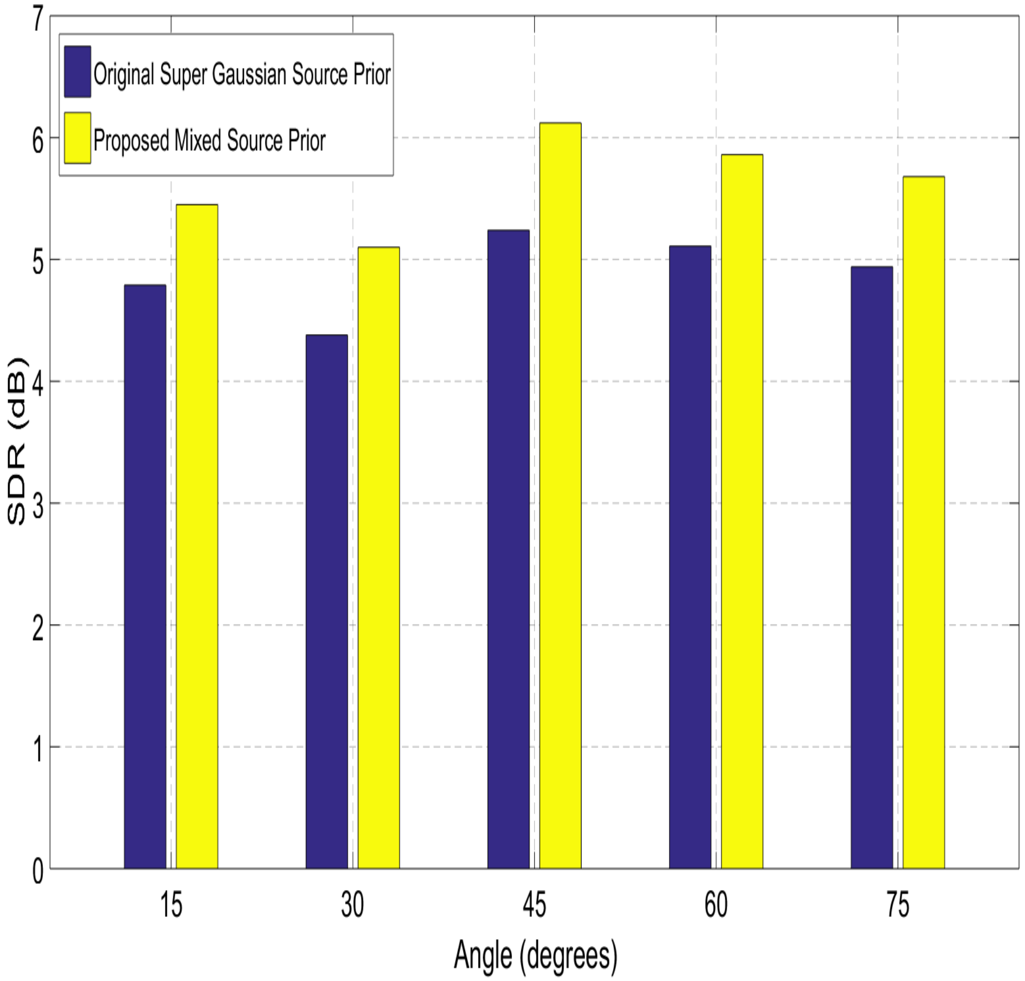

Figure 5.2: The bar graph provides SDR (dB) for the FastIVA method [58] and the proposed mixed source prior FastIVA for five different angles. All the SDR values are averaged over eighteen random mixtures. Real BRIRs from [87] were used. The new mixed source prior enhance the separation performance at all separation angles.

5.3 Energy Driven Mixed Source Prior for the Original IVA Method

Since different speech sources can have different statistical properties, therefore it is important to adapt the mixed Student's t and super Gaussian source prior according to different speech sources. In this mixed source prior for the original IVA method, equal weight was assigned to both the Student's t and the original super Gaussian in the mixed source prior for all speech sources. However, adjusting the weight of both distributions automatically in the mixed source prior according to the variation in the speech sources can potentially further improve the separation performance of the original IVA method. Therefore, in this section the weight of both distributions in the mixed source prior is adapted automatically according to the energy of the measured speech mixture signals. Importantly, this method is found to be successful only with access to mixtures not the original sources [104]. Moreover, to further improve the separation performance of the IVA algorithm, fully connected frequency bins are decomposed into smaller groups because the dependency among the neighbouring frequency bins is generally stronger and much weaker between distant frequency bins [23]. So the strong dependency between neighbouring frequency bins is exploited by dividing them into smaller cliques whilst retaining considerable overlap between adjacent cliques. The clique based approach for frequency bins is adopted for the IVA method and the mixed Student's t and original super Gaussian distribution is used as a source prior and the ratio of distributions in the mixed source prior is automatically adjusted according to the energy of the measured speech mixtures.

5.3.1 Clique based IVA method

In the original IVA method, retaining the inter-frequency dependency is crucial to its performance and it is preserved by using the multivariate source prior. Also, the neighbouring and distant frequency components are assigned the same dependency in the IVA method whereas in real life speech sources, the dependency between the neighbouring frequency components is much stronger than that of distant frequency components [23]. Therefore, in order to enhance the frequency dependency within the IVA method, the single and fully connected statistical model is decomposed into several overlapping cliques of fixed size. Since the dependency is much stronger in the neighbouring frequency bins, by dividing the fully connected statistical model into smaller cliques, the strong dependency between neighbouring frequency bins can be exploited and at the same time weaker dependency between the distant frequency bins can be reduced. When the fully connected statistical dependency model of the IVA method is decomposed into several overlapping cliques of fixed size, the corresponding multivariate probability density function can be written as [23]:

$$p(\mathbf{s}_i) \propto \exp\left(-\sum_{c=1}^{C} \sqrt{\sum_{k=f_c}^{l_c} \left| \frac{\hat{\mathbf{s}}_i(k)}{\sigma_i(k)} \right|^2} \right) \tag{5.23}$$

where C is the number of cliques and f_c and l_c are the first and last indices of the c-th clique, respectively. So this new dependency structure consists of several cliques of fixed and identical size and the centre frequency is increasing for every clique. As an example, in the case of 1024 frequency bins, in order to consider strong dependency within the neighbouring frequency bins, the fully connected statistical model of the IVA method is decomposed into 128 cliques each of fixed size of 256 frequency bins and clique ranges are $[f_1, l_1] = [1, 256]$, $[f_2, l_2] = [17, 272], \ldots, [f_c, l_c] = [769, 1024]$. This fixed size clique based structure improves the dependency structure for the IVA method by making better use of the strong dependency between the neighbouring frequency bins and is likely to improve the separation performance of the IVA method with the new mixed energy driven mixed source prior. The energy calculation of the measured speech signals and tuning of the mixed source prior according to the energy of the measured speech signals is discussed in the next section.

5.3.2 Energy Calculation of Measured Speech Mixtures

As discussed earlier, the mixed source prior for the original IVA method is given as:

$$\begin{aligned} p(\mathbf{s}_i) =&(\lambda_d)\left(1+\frac{(\mathbf{s}_i-\mu_i)^T\mathbf{\Lambda}(\mathbf{s}_i-\mu_i)}{\nu}\right)^{-\frac{\nu+K}{2}}+ \\ &(1-\lambda_d)\exp\left(-\sqrt{(\mathbf{s}_i-\mu_i)^{\dagger}\Sigma_i^{-1}(\mathbf{s}_i-\mu_i)}\right) \end{aligned} \tag{5.24}$$

The nonlinear score function for the original IVA algorithm with the mixed Student's t and super Gaussian source prior can be obtained as follows.

$$\begin{aligned} \varphi(k)(\hat{s}_i(1)\cdots\hat{s}_i(k)) \propto&(\lambda_d)\left(\frac{\hat{s}_i(k)}{1+\frac{1}{\nu}\sum_{k=1}^{K}|\hat{s}_i(k)|^2}\right) \\ &+(1-\lambda_d)\left(\frac{\hat{s}_i(k)}{\sqrt{\sum_{k=1}^{K}|\hat{s}_i(k)|^2}}\right) \end{aligned} \tag{5.25}$$

In Equation (5.25), λ_d is the weighting parameter, which defines the ratio of both the Student's t and the original super Gaussian distribution in the mixed source prior. Since different speech sources can have different properties and their resulting speech mixtures can have different energies, therefore by tuning the value of λ_d, according to the local energy of speech mixtures, the mixed source prior can adapt to different speech mixtures. The weighting parameter is frequency dependent i.e. $\lambda_d(k)$ is estimated according to the energy of the observed speech mixture. It is calculated as the normalised energy of the speech mixtures in the frequency domain blocks. The complete frequency bins are subdivided into smaller non overlapping blocks and then the normalised energy of each block is calculated. The frequency bins are divided into smaller blocks because different frequency ranges can have different energy and then to model a particular block, an appropriate value for $\lambda_d(k)$ can be selected. So the energy of a particular block can be calculated as follows:

$$E_b = \frac{1}{E_t}\left(\sum_{k=f_b}^{l_b} ||\mathbf{x}_p(k)||^2\right) \tag{5.26}$$

where f_b and l_b are the first and last indices of the block, respectively and $\mathbf{x}_p(k)$ denotes the vector of all frequency components k calculated by dividing the entire speech observation into subblocks indexed by p, whereas E_b and E_t are the energy of the particular block (clique) and the total energy of the source mixture, respectively, and $||(\cdot)||$ denotes Euclidean norm. When a particular block has high energy, the value of $\lambda_d(k)$ is tuned so that the ratio of the Student's t distribution in the mixed source prior as the Student's t distribution is high due to its heavy tail nature, it can improve the modelling of the high amplitude information. Similarly, when the energy of a particular block is relatively low, the weighting parameter λ_d is tuned to assign more weight to the original super Gaussian distribution in the mixed Student's t source prior. Since low energy for a particular block generally indicates lack of high amplitude information, therefore in order to appropriately model the speech sources, the mixed source is tuned to have less ratio of Student's t distribution and more of the original super Gaussian source prior. Hence, this energy driven mixed source prior should be able to better model the underlying non-stationary speech signals by adapting to the nature of the measured speech mixtures thereby improving the separation performance of the IVA method. In the next section, the new energy driven mixed source prior based IVA algorithm is evaluated in the two experimental setups.

5.3.3 Experimental Results

The energy driven mixed Student's t-original super Gaussian source prior for the original IVA method is evaluated in this section by using three different types of room impulse responses. Firstly, it is tested with the simulated room impulses responses that were generated by the Image method [84]. Then the separation performance of the new energy driven mixed source prior based IVA method is tested in the real room impulses that were obtained from [88] and [87] which provide the accurate evaluation of the algorithm in the real life context.

Evaluation with Image Method

In this section, the energy driven mixed source prior is tested with room impulse generated by the Image method [84]. The same experimental settings were used as in the case of the original IVA method with fixed mixed source prior as in Section 5.1.3. The RT_{60} for this particular set of experiments was 250ms. The speech signals were again randomly chosen from the whole of the TIMIT dataset [79] and convolved into mixtures. The weighting parameter λ_d in the mixed source prior was tuned according to the energy of the measured speech mixtures.

	As in [22]	Proposed Source Prior	Improvement (dB)
Set-1	8.58	9.01	0.95
Set-2	9.01	9.93	0.92
Set-3	8.61	9.70	1.09
Set-4	7.24	8.12	0.88
Set-5	8.03	9.09	1.06

Table 5.5: SDR (dB) values for the energy driven mixed source prior and the original super Gaussian source prior for the original IVA method with Image room impulse response [84]. For all mixture, the separation results are average separation performance for six different positions.

The separation performance of the energy driven mixed source prior based IVA method with image room impulse responses is presented in Table 5.5. The separation performance was evaluated objectively with SDR (dB) and all the values shown in Table 5.5 are the average of separation performance for each set at six different positions. All the results are compared with the separation performance of the original IVA method with the original super Gaussian source prior and the new energy based mixed source prior consistently improves the separation performance of the IVA method. When the energy driven mixed source prior is used as a source prior for the IVA method, the average improvement of approximately 1dB is recorded in the separation performance of the algorithm.

Evaluation with Real Room Impulse Responses from Hummerstone [88]

In order to evaluate the separation performance of the proposed energy driven mixed source prior in realistic scenarios, real room impulse responses were used, which were obtained from [88]. These room impulse responses are the real room recording in four different rooms with different sizes and geometry. Therefore all four rooms have different reverberation time and provide a different real life environment. Hence when the proposed energy driven mixed source prior is evaluated with these room impulse responses, it provides a good evaluation of the algorithm in the changing real life scenarios. The four room types and their respective RT_{60} is shown in Table 5.6. The RT_{60} of rooms varies from 320 ms to 890 ms, so these room impulse responses provide a good evaluation of the algorithm over a range of reverberation times. Also the source location azimuths ranging from (-90° to 90°) relative to the second source were available for all the room types, which provide a robust evaluation at different positions for moving sources within each room.

For this set of experiments, the 2 x 2 case was considered. The speech signals were randomly selected from the whole of the TIMIT dataset and convolved into

Table 5.6: Room types and the respective RT_{60}.

Room	Type	RT_{60} (ms)
A	Medium office	320
B	Small class room	470
C	Large lecture room	680
D	Large seminar theatre	890

mixtures for all the rooms. The source location azimuths in step of 15 ° was considered from 15 ° to 90 ° in all the rooms for this set of experiments. The separation performance was again measured objectively with SDR in dB. The mixtures were then separated by using the new energy driven mixed Student's t and original super Gaussian source prior and its separation performance is compared with the original IVA method [21]. The separation performance of both methods for all four rooms over the range of RT_{60} is shown in Figure 5.3.

It is evident from the Figure 5.3 that in the room A, the energy based mixed source prior consistently improves the separation results of the algorithm at all the separation angles. Overall SDR values for both the algorithms is generally high as the RT_{60} for room is 320ms. The proposed energy driven mixed source prior tunes the weight of both distribution according to the energy of the measured mixtures and it provides better separation for the IVA method at all the separation angles. All the results shown in Figure 5.3 at all the separation angles are the average of twelve mixtures, which improves the reliability of the results. For room B, the SDR values drop down for both algorithms, as the room type B has the higher reverberation time of 470 ms. It is clear form Figure 5.3 that even at the higher reverberation time, the new mixed source prior still performs better then the original super Gaussian source prior at the separation angles. Likewise for room type C and D, the proposed energy driven IVA method performs better then the original super Gaussian source prior. Also the SDR value drops down

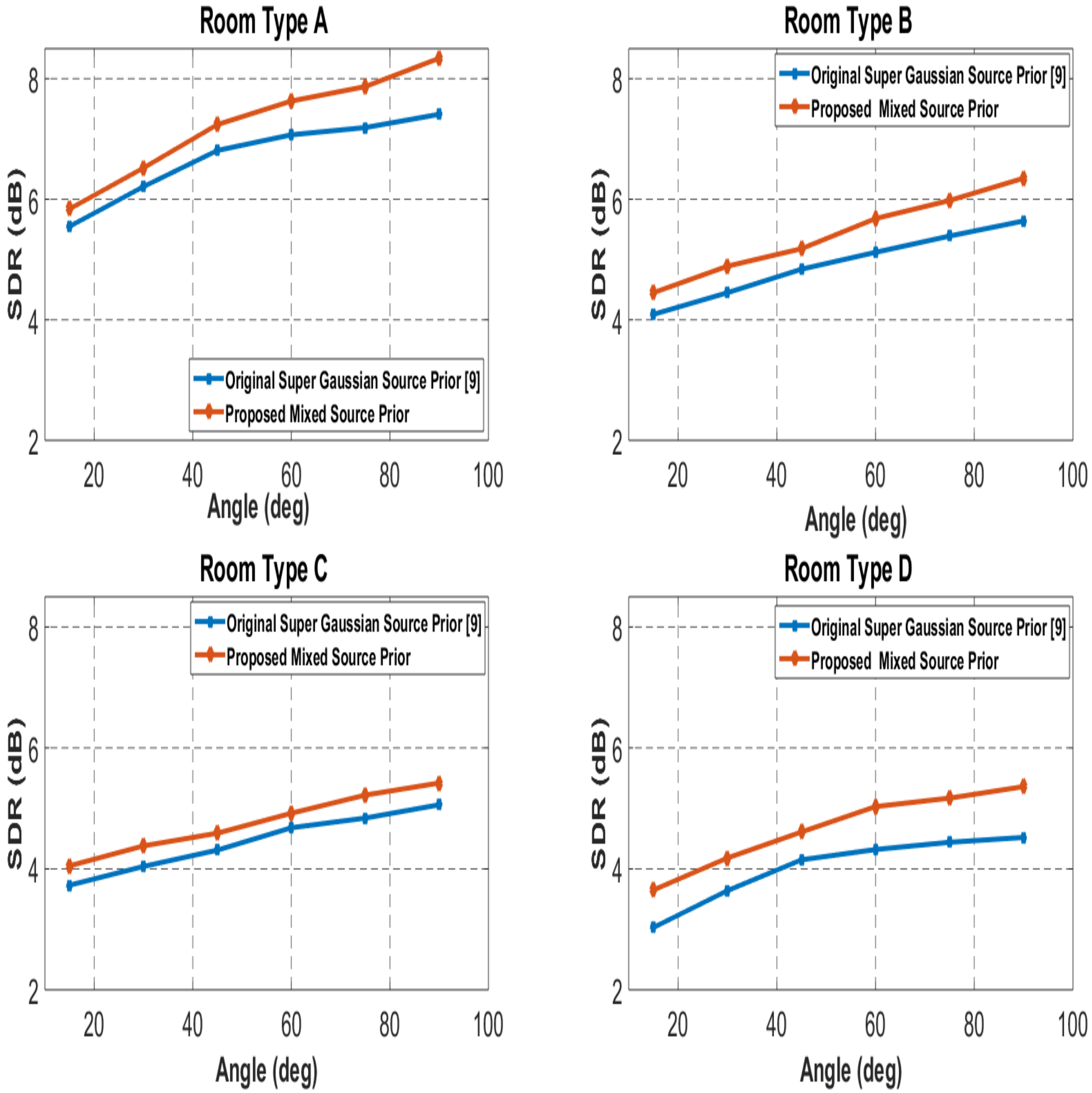

Figure 5.3: The graphs indicate results for different reverberant rooms. All the SDR values (dB) were averaged over twelve mixtures at each separation angle. Energy driven mixed source prior enhance the separation performance of the IVA algorithm in all types of reverberant conditions.

further for room C and is worse for room D for both the algorithms. Since the RT_{60} for room C and room D is 680ms and 890ms, respectively, it shows the separation performance of the algorithm in the extremely difficult and highly reverberant environments. It is evident from Figure 5.3 that even in the highly reverberant room, which provides a good indication of real life context, the new energy driven mixed Student's t and original super Gaussian source prior based has better separation performance as compared to the original IVA method as in [21].

Evaluation with Real Room Impulse Responses from Shinn [87]

Finally, the separation performance of the proposed method is tested with the room impulse response generated by the BRIRs which were obtained from Shinn [87]. Since it is a different room with different settings at high RT_{60} of 565ms, it also provides a good evaluation of the algorithms in the real life context. For these experiments all the speech signals were chosen randomly from the TIMIT dataset and the length of signals was approximately five seconds. Six different source location azimuths varying from 15 ° to 90 ° with a step of 15 ° relative to the second source were used. In order to improve the reliability of the results, all the measurements were recorded on three separate occasions. The degrees of freedom parameter in the mixed source prior was again chosen empirically to be four. The common parameters that were used in the experiments are mentioned in Table 5.7.

Table 5.7: Different parameters used in experiments.

Sampling rate	8kHz
STFT frame length	1024
Reverberation time	565 ms (BRIRs)
Room dimensions	9 m x 5 m x 3.5 m
Source signal duration	5 s (TIMIT)
Degrees of freedom	4

The speech sources were convoluted into mixtures with the room impulse responses generated by BRIRs [87] and then separated by using the new energy driven mixed source prior based IVA method. The separation performance of the proposed method was compared with the original super Gaussian source prior based IVA method [21] at all the separation angles varying form ($15\,^{\circ}$ to $90\,^{\circ}$). The separation performance was measured objectively with SDR in dB and the results are shown in Figure 5.4. For each separation angle, the results shown in Figure 5.4 were averaged over eighteen different mixtures and at all the source location azimuth. The new energy driven mixed source prior provides approximately 1dB average performance improvement for the IVA method. Figure 5.4 also shows the proposed energy driven mixed source prior can consistently achieve better separation performance in a highly reverberant real life scenarios.

In addition to the objective measure of SDR, a subjective measure of perceptual evaluation of speech quality (PESQ) is also used to measure the separation performance of the algorithm. The PESQ is a commonly used measure to check the quality of the separated signal as it compares the estimated signals with the original signals and gives a score between 0-4.5, 0 for very poor separation and 4.5 for excellent separation. The signals were separated from mixtures by using the energy driven mixed source prior based IVA method in the highly reverberant real room environment with RT_{60} of 565ms and the PESQ scores for separated signals were measured as shown in Table 5.8. All the PESQ scores shown in Table 5.8 for each mixture are the average of PESQ scores for six different source location azimuths varying from ($15\,^{\circ}$ to $90\,^{\circ}$). The PESQ score for the proposed energy driven source prior is compared with the PESQ score of the estimated signals that were separated by the original IVA method and this subjective study also confirms the improved separation performance for the IVA method with the energy driven mixed source prior.

The final set of experiments will establish the advantage of the energy calculation of the measure mixtures and automatically adapting the weight of distributions

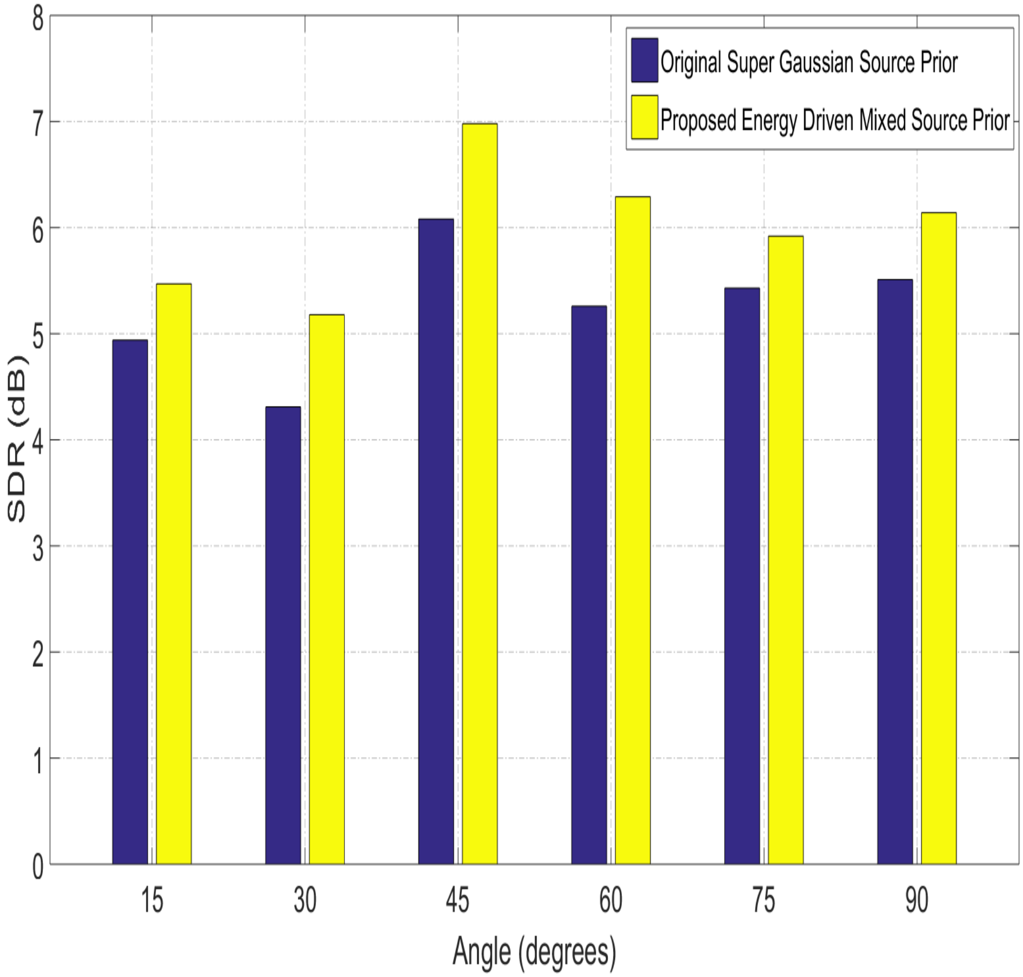

Figure 5.4: Separation performance in terms of SDR (dB) values is shown for the energy based mixed source prior and the original IVA algorithm. All the SDR values are averaged over eighteen mixtures to improve the reliability of results. The energy based mixed source prior m enhance the separation performance of IVA algorithm at all the locations in the room.

in the mixed source prior. The same set of experiments with same experimental setting were repeated for the fixed mixed source prior and the value of the weighting parameter was set to $\lambda_d = 0.5$. The same set of mixtures was also separated by using the IVA method with the Student's t source prior, which was also described in the Chapter 3. For all three methods, the mixtures were created by using the room impulse response generated by the BRIRs with high

	Original Source Prior [22]	Proposed Source Prior
Set-1	1.66	1.97
Set-2	2.04	2.27
Set-3	2.09	2.32
Set-4	1.92	2.11
Set-5	2.02	2.21

Table 5.8: The table shows PESQ values for IVA algorithm with two source priors. PESQ scores are averaged over six different locations in the room.

reverberation time of 565ms. Then all three methods were used to separate the mixtures at six different source location azimuths and the separation performance in terms of SDR is shown in Figure 5.5. The blue and red lines show the separation performance of the Student's t source prior based IVA and the fixed mixed source prior based IVA method, respectively, whereas the green line represents the separation performance of the new energy driven mixed source prior based IVA method. It is evident from Figure 5.5 that for all the separation angles varying from (15 ° to 90 °), the proposed energy driven mixed source prior based IVA method has the improved separation performance. Since the energy driven mixed source prior adapts to the statistical properties of the measured mixtures, it is well suited to model different types of speech sources and therefore it improves the separation performance of the IVA algorithm.

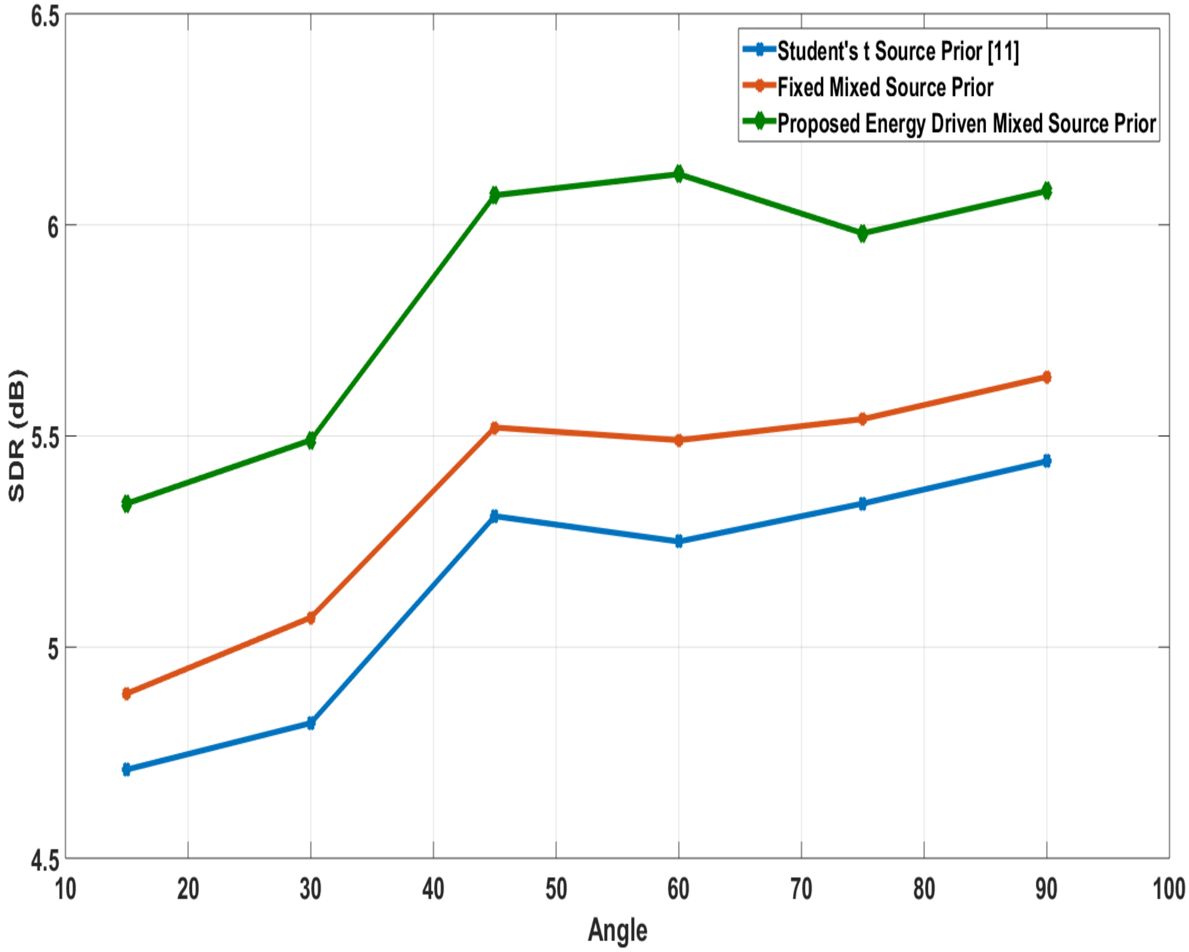

Figure 5.5: The separation performance of the IVA algorithm with three different source prior is shown. RIRs were generated by BRIRs dataset and the SDR values were averaged over twelve mixtures at each separation angle. The energy based mixed source prior yields the improved sepration performance for all room settings.

5.4 Summary

In this chapter, instead of a single distribution source prior for the IVA algorithm, a new mixed multivariate Student's t and original super Gaussian source prior was proposed for the IVA algorithm. This multivariate mixed source prior can improve the modelling of the speech signals as the speech signals are highly random in nature and can have high amplitude information. The Student's t distribution in the mixed source prior was adopted to better model the high amplitude information in the speech signals and at the same time the original super Gaussian distribution was used to model the remaining information. This mixed source prior was adopted for the IVA and the FastIVA algorithm and performance improvement was recorded. The separation performance of the mixed source prior IVA was further enhanced by adopting the ratio of distributions in the mixed source prior according to the energy of the measured mixtures, as different speech sources can have different statistical properties. The detailed experimental studies in the simulated and real room environment with different reverberation times confirmed the improved separation performance of the energy driven mixed source prior. In the next chapter, in order to adapt to different speech sources, a mixture model approach will be exploited and a new EM framework will be proposed for the IVA algorithm.

Chapter 6

AN EXPECTATION MAXIMIZATION FRAMEWORK FOR THE IVA ALGORITHM USING STUDENT'S T MIXTURE MODEL

In this chapter a new general probabilistic framework for the IVA algorithm is introduced. The performance of the IVA algorithm depends on the choice of the source prior to better model the speech signals. Previously, identical source priors were used in different methods, however different speech sources will generally have different statistical properties. In this chapter, a novel IVA algorithm is introduced which can adapt to the statistical properties of different speech sources and efficiently separate different types of speech signals. In order to make the IVA algorithm robust to different speech mixtures, instead of identical source priors, different Student's t mixture models are introduced as source priors. These flexible Student's t mixture models can adapt to the statistical properties of different speech signals and thereby enhance the separation performance of the IVA algorithm. The unknown parameters of the source prior and unmixing matrices are estimated together by deriving an efficient expectation maximization (EM) algorithm. As a result of the proposed EM framework for the IVA method, the algorithm can adjust according to the statistical properties of the speech sources. Useful improvement in the separation performance in realistic scenarios is confirmed by simulation studies on real datasets.

6.1 Introduction

The process of human speech production is really complex [1] and the human speech signal is non-stationary in nature. Human speech is difficult to model because there can be wide variations in human speech [10]. The properties of natural speech varies from person-to-person and depend on which language is being spoken as the pronunciation rates and phonemes can be totally different in different parts of the world. Also recorded speech is dependent on variations in room acoustics and microphone characteristics e.g. different rooms will have different reverberation effects and different microphones will have variable frequency responses [4]. All of these factors can change the observed human speech signal and thereby different speech signals generally have different statistical properties. Therefore it is important that the BSS algorithms adapt their statistical structure according to the characteristics of the observed speech signals.

As discussed earlier in Chapter 2, the IVA algorithm preserves the interfrequency dependency between the individual sources in the frequency domain. The IVA method uses the score function and its form is crucial to the performance of the IVA algorithm [22]. The score function is derived by statistical modelling of the speech sources by selecting an appropriate source prior. Speech signals are often characterized with fixed statistical models. In the original IVA [21] method all the speech sources were modelled by the identical multivariate Laplacian distributions. Different sources can have different statistical properties and modelling all the sources with identical distribution may not be the most appropriate solution. Therefore in this chapter, a Student's t mixture model (SMM) is adopted as a source prior for the IVA algorithm, instead of the conventional identical multivariate distributions. The probability density function (pdf) of the SMM has heavier tails as compared to the Gaussian mixture models and therefore it can model outliers in the data [75]. As human speech is highly random, the spread of samples can be very wide and the Student's t distribution, due to its heavier tails, can model it more accurately. Therefore, the high amplitude information in human speech can generally be modelled more accurately by using the SMMs. The new framework of the expectation maximization (EM) algorithm is implemented efficiently within the proposed IVA algorithm to estimate the unmixing matrices. The EM algorithm is a two steps iterative approach which efficiently estimates the unknown parameters of source prior and unmixing matrices. The EM method overcomes non-analytically solvable problems and it has been commonly used in the field of statistics, pattern recognition, signal processing and machine learning [90]. By using SMMs as the source prior and implementing the new framework of EM, the proposed IVA algorithm shows performance improvement when compared with previous approaches [21]. The detailed EM framework for the IVA algorithm is derived in the following sections.

6.2 Independent Vector Analysis

Previously, in IVA method, speech signals have been modelled with various superGaussian distributions, e.g. the Laplacian distribution [21] or generalized Gaussian distribution [95] but speech signals can have very high and low amplitudes and the Laplacian distribution may not be able to accurately model the

high amplitudes in the speech signals [70]. Therefore, the Student's t distribution is adopted to model the speech signals. The Student's t distribution due to its heavy tail nature can better model the information in the outliers [70]. The multivariate Student's t distribution is introduced in Chapter 4 which is defined as follows:
A K-dimensional random separated source vector $\mathbf{s} = (s(1), \ldots, s(K))^T$ can have a K-variate t distribution with degree of freedom ν, precision matrix $\mathbf{\Lambda}$ and mean $\boldsymbol{\mu}$, if its joint pdf is given by [70]:

$$St(\mathbf{s}|\boldsymbol{\mu}, \mathbf{\Lambda}, \nu) = \frac{\Gamma(\frac{\nu+d}{2})|\mathbf{\Lambda}|^{1/2}}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\Big(1 + \frac{(\mathbf{s}-\boldsymbol{\mu})^T\mathbf{\Lambda}(\mathbf{s}-\boldsymbol{\mu})}{\nu}\Big)^{-\frac{\nu+d}{2}} \tag{6.1}$$

In the joint pdf of the Student's t distribution, the leptokurtic nature and the variance of distribution can be adjusted by tuning the degrees of freedom parameter ν [72]. When the ν is set to a lower value, the tails of the distribution becomes heavier and if ν is increased to infinity, the Student's t distribution tends to a Gaussian distribution [75]. Since different sources can have different statistical properties, so instead of using identical Student's t source prior for all sources, the Student's t mixture model (SMM) is adopted as a source prior in this work. By adopting the SMM as a source prior, different speech sources can be modelled with different SMMs thereby enabling the IVA method to adapt to the statistical properties of different types of signals. Assuming the sources are statistically independent, for a 2 x 2 case, an SMM can be represented as:

$$\begin{aligned} p(\mathbf{s}(1)\cdots\mathbf{s}(K)) &= \prod_{i=1}^{2} p(\mathrm{s}_i(1)\cdots\mathrm{s}_i(K)) \\ p(\mathrm{s}_i(1)\cdots\mathrm{s}_i(K)) &= \sum_{q_i} p(q_i)\prod_{k} S_t(\mathbf{s}_i(k)|\boldsymbol{\mu}_i(k), \mathbf{\Lambda}_i(k)) \end{aligned} \tag{6.2}$$

where q_i is the weight of the mixture component of the SMM source prior for source i and K represents the total number of frequency bins in the multivariate model. The precision matrix $\mathbf{\Lambda}$ is defined as the inverse of the covariance matrix and its $ik-th$ element satisfies $1/\mathbf{\Lambda}_i(k) = E\{\mid \mathbf{s}_i(k) \mid^2 q_i\}$. With appropriate normalisation and zero mean assumption, the Student's t distribution can be written as:

$$St(\mathbf{s}_i(k)|0, \mathbf{\Lambda}_i(k)) = \frac{\mathbf{\Lambda}_i(k)}{\pi}\left(1 + \frac{\mathbf{\Lambda}_i(k)\mid \mathbf{s}_i(k) \mid^2}{\nu}\right)^{-\frac{\nu+d}{2}} \tag{6.3}$$

When the vector of frequency components is considered from the same source i, the interdependency between these frequency components is preserved whereas the vectors that originate from different sources are independent of each other. Therefore by adopting this interfrequency dependency model, the IVA method prevents the permutation problem that is inherent to most BSS methods [12].

In the IVA algorithm, the scaling of mixture signal $\mathbf{x}(k)$ and mixing matrix $\mathbf{A}(k)$ cannot be determined by the separated source signals $\mathbf{s}(k)$, therefore observations are prewhitened. Because of the prewhitening process, both the mixing $\mathbf{A}(k)$ and the unmixing matrix $\mathbf{W}(k)$ are unitary matrices. In this work, the 2 x 2 case has been considered, so the Cayley Klein parameterizations [103] for the unitary matrix $\mathbf{W}(k)$ is as follows:

$$\mathbf{W}(k) = \begin{pmatrix} a_k & b_k \\ -b_k^* & a_k^* \end{pmatrix} \therefore |\mathbf{W}(k)| = a_k a_k^* + b_k b_k^* = 1 \tag{6.4}$$

In the next section, the maximum likelihood estimate is derived for the IVA algorithm.

6.3 Maximum Likelihood Estimation of SMM

The maximum likelihood estimate is a well known method that is usually used to estimate the mixture parameters. Based on the maximum likelihood method, the mixture parameters can be effectively estimated iteratively via the expectation maximization (EM) algorithm [90]. The log likelihood function for t components mixture of Student's t distributions is considered and it is given as:

$$\mathcal{L}(\mathbf{x}, \theta) = \sum_{i=1}^{t} \log \; p(\mathbf{x}_i(1), \cdots, \mathbf{x}_i(K)) = \sum_{i=1}^{t} \log \Bigg(\sum_{q_i} \prod_{k=1}^{K} p(\mathbf{x}_i(k)|q_i) p(q_i) \Bigg) \quad (6.5)$$

where $\theta = \{\mathbf{W}_i, \mathbf{\Lambda}_i, p(q_i)\}$ consists of the model parameters for the log likelihood function; $p(\mathbf{x}_i(1), \cdots, \mathbf{x}_i(K))$ is the PDF of the observed source mixture signals which is a SMM as it is generated by the SMM source priors. The $\mathbf{W}_i$ shows the unmixing matrix, $\mathbf{\Lambda}_i$ represents the precision matrix and $q_i = [q_1, q_2]$ is the collective mixture index of the SMMs for source prior. In the maximum likelihood estimation, the best fitting model helps to estimate parameters that can maximize the log-likelihood function, which is usually performed by using the EM algorithm [90]. Therefore the model parameters set $\theta = \{\mathbf{W}_i, \mathbf{\Lambda}_i, p(q_i)\}$ can be estimated by training the SMM and maximizing the log likelihood function by using an EM algorithm. The detailed method for estimating the model parameters by the EM algorithm is explained in the next section.

6.4 The Expectation Maximization Algorithm

The EM algorithm is well-matched to finding latent parameters in probabilistic models by using an iterative optimization technique [90]. The EM algorithm is implemented by introducing discrete random variables $z(q_i)$ which are dependent on the observations $(\mathbf{x}_i(1), \cdots, \mathbf{x}_i(K))$ and the model parameter set θ. The log likelihood function with these variables is given by

$$\begin{aligned}\mathcal{L}(\mathbf{x},\theta) &= \sum_{i=1}^{t}\log\bigg(\sum_{x_i}\prod_{k=1}^{K}p(\mathbf{x}_i(k)|\mathbf{q}_i)p(q_i)\bigg)\\ &= \sum_{i=1}^{t}\log\bigg(\sum_{q_i}\frac{z(q_i)\prod_{k=1}^{K}p(\mathbf{x}_i(k)|q_i)p(q_i)}{z(q_i)}\bigg)\end{aligned} \tag{6.6}$$

and can be used to optimise the model parameters. In the case of an increasing log likelihood function, a lower bound is formed on the increasing log likelihood for the observations $(\mathbf{x}_i(1),\cdots,\mathbf{x}_i(K))$. So the new parameters that increase the log likelihood function of the complete data with respect to current parameters, can be found. Hence there is an increase in the expected log likelihood of the complete data with respect to current parameters and it is produced by the updated parameters. Therefore an auxiliary function can be used to represent the expected log likelihood function. There will be a definite increase in the log likelihood function when the auxiliary function will be optimised but it doesn't necessarily yield a maximum likelihood solution [90]. Therefore it is important to iteratively calculate and maximize the auxiliary function until convergence. Hence a local approximation is made which is the lower bound to the objective function. By using the Jensen's inequality [102], the lower bound for the log likelihood function in Equation (6.6) can be calculated as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{x},\theta) \geq \sum_{i=1}^{t}z(q_i)\log\bigg(\frac{\prod_{k=1}^{K}p(\mathbf{x}_i(k)|q_i)p(q_i)}{z(q_i)}\bigg)\\ = \mathfrak{F}(z,\theta)\end{aligned} \tag{6.7}$$

The EM algorithm will run until convergence and it will iteratively maximize $\mathcal{L}(\mathbf{s},\theta)$ in two steps. The first step is the expectation step in which the posterior probability of the hidden variable $\mathfrak{F}(z,\theta)$ is calculated over $z(x_i)$ and in the second step, the θ is updated.

6.4.1 The Expectation Step

In the expectation step, θ is fixed and $\mathfrak{F}(z,\theta)$ is maximised over $z(q_i)$. In order to maximise $\mathfrak{F}(z,\theta)$, the derivative of log-likelihood equation with respect to $z(\mathbf{q}_i)$ is calculated as follows:

$$\frac{\partial}{\partial z(q_i)}(\mathcal{L}(\mathbf{x},\theta)) = \frac{\partial}{\partial z_i}\left(\sum_{i=1}^{t} z(q_i)\log\frac{\prod_{k=1}^{K} p(\mathbf{x}_i(k)|\mathbf{q}_i)p(q_i)}{z(q_i)}\right) \tag{6.8}$$

$$= 1.\log\left(\frac{\prod_{k=1}^{K} p(\mathbf{x}_i(k)|\mathbf{q}_i)p(q_i)}{z(s_i)}\right)+z(q_i)\left(\frac{z(q_i)}{\prod_{k=1}^{K} p(\mathbf{x}_i(k)|\mathbf{q}_i)p(q_i)}\right)\left(\frac{-\prod_{k=1}^{K} p(\mathbf{x}_i(k)|\mathbf{q}_i)p(q_i)}{z(q_i)}\right) \tag{6.9}$$

$$= \log\left(\frac{\prod_{k=1}^{K} p(\mathbf{x}_i(k)|\mathbf{q}_i)p(q_i)}{z(q_i)}\right) - z(q_i) \tag{6.10}$$

In order to maximize θ for fixed $\mathfrak{F}(z,\theta)$, equating the above equation equal to zero and with appropriate normalization,

$$z(q_i) = \frac{\prod_{k=1}^{K} p(\mathbf{x}_i(k)|q_i)p(q_i)}{p(\mathbf{x}_i(1),\cdots,\mathbf{x}_i(K))} \tag{6.11}$$

Now by using $\mathbf{x}(k) = \mathbf{A}(k)\mathbf{s}(k)$,

$$p(\mathbf{s}_i(k)|\mathbf{q}_i) = St(\mathbf{s}_i(k)|0,\boldsymbol{\Lambda}_i(k)) \tag{6.12}$$

and the precision matrix for the 2x2 case can be written as:

$$\mathbf{\Lambda}_{ik} = \mathbf{W}(k)^{\dagger}\mathbf{\Phi}_i(k)\mathbf{W}(k); \mathbf{\Phi}_i(k) = \begin{pmatrix} v_1(k) & 0 \\ 0 & v_2(k) \end{pmatrix} \tag{6.13}$$

As $\mathbf{W}(k)$ is a unitary matrix, therefore $\det(\mathbf{\Lambda}_i(k)) = v_1(k)v_2(k)$ and from Equation (6.5), the function $f(x_i)$ can be defined as:

$$f(q_i) = \sum_k \text{log}p(\mathbf{x}_i(k)|\mathbf{q}_i) + \text{log}p(q_i) \tag{6.14}$$

By using Equation (6.11), function $f(x_i)$ can be rewritten as $z(x_i) \propto e^{f(x_i)}$, therefore:

$$\begin{aligned} j_i &= \sum_{s_i} e^f(x_i); \\ z(x_i) &= \frac{1}{j_i} e^f(x_i) \end{aligned} \tag{6.15}$$

Next, the maximization step is considered.

6.4.2 The Maximization Step

The maximization step (M-step) the parameters $\theta = \{\mathbf{W}_i, \mathbf{\Lambda}_i, p(q_i)\}$ can be estimated by maximising the cost function. In this step, each parameter is estimated separately. In the first step, the maximisation of $\mathbf{W}_i$ over the unitary constraint is considered. In order to maximize the $\mathbf{W}_i$, the precision matrix for 2x2 can take the following form:

$$\mathbf{\Phi}_{ik} = \begin{pmatrix} v_1(k) - v_2(k) & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} v_2(k) & 0 \\ 0 & v_2(k) \end{pmatrix} \tag{6.16}$$

Now Equation (6.12) can be rearranged as:

$$p(\mathbf{x}_i(k)|\mathbf{q}_i) = S_t(\mathbf{x}_i(k)|0, \mathbf{\Lambda}_i(k)) = \frac{\mathbf{\Lambda}_i(k)}{\pi}\left(1 + \frac{\mathbf{\Lambda}_i(k)|\ \mathbf{x}_i(k)\ |^2}{\nu}\right)^{-\nu/2-d/2} \tag{6.17}$$

When $p(\mathbf{x}_i(k)|\mathbf{q}_i)$ is replaced in the log likelihood Equation (6.7), it will take the following form:

$$\mathcal{L}(\mathbf{x}, \theta) = \sum_{i=1}^{t} z(q_i)\log\left(\frac{\prod_{k=1}^{K} \frac{\mathbf{\Lambda}_i(k)}{\pi}\left(1 + \frac{\mathbf{\Lambda}_i(k)|\mathbf{x}_i(k)|^2}{\nu}\right)^{-\nu/2-d/2} p(q_i)}{z(q_i)}\right) \tag{6.18}$$

$$= \sum_{i=1}^{t} z(q_i)\left\{\log(\frac{\mathbf{\Lambda}_i(k)}{\pi})(p(q_i))(z(q_i))\right\}\left\{\log\left(1 + \frac{\mathbf{\Lambda}_i(k)|\ \mathbf{x}_{ik}\ |^2}{\nu}\right)^{-\nu/2-d/2}\right\} \tag{6.19}$$

$$= \sum_{i=1}^{t} z(q_i)\left\{\lambda\right\}\left\{(-\nu/2 - d/2)\log\left(1 + \frac{\mathbf{\Lambda}_i(k)|\ \mathbf{x}_i(k)\ |^2}{\nu}\right)\right\} \tag{6.20}$$

By using the log approximation $\log(1+a) \approx a$, where a is a small value, the above mentioned equation can take the following form, wherein equality is assumed for convenience.

$$= \sum_{i=1}^{t} z(q_i)\left\{\lambda\right\}\left\{(-\nu/2 - d/2)\left(\frac{\mathbf{\Lambda}_i(k)|\ \mathbf{x}_i(k)\ |^2}{\nu}\right)\right\} \tag{6.21}$$

Now by replacing the value of the precision $\mathbf{\Lambda}_i(k) = \mathbf{W}(k)^\dagger\mathbf{\Phi}_i(k)\mathbf{W}(k)$ in the above equation

$$= \sum_{i=1}^{t} z(q_i)\,\{\lambda\}\left\{(-\nu/2 - d/2)\left(\frac{\mathbf{x}_i(k)^{\dagger}\mathbf{W}(k)^{\dagger}\mathbf{\Phi}_{ik}\mathbf{W}(k)\mathbf{x}_i(k)}{\nu}\right)\right\} \tag{6.22}$$

The above equation can be rewritten by replacing the value of Φ_{ik} from the Equation (6.16) as follows:

$$\begin{gathered} = -\sum_{i=1}^{t} z(q_i)\,\{\lambda\} \\ \left\{\frac{(\nu/2 + d/2)}{\nu}\left(\mathbf{x}_i(k)^{\dagger}\mathbf{W}(k)^{\dagger}\begin{pmatrix} v_1(k) - v_2(k) & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} v_2(k) & 0 \\ 0 & v_2(k) \end{pmatrix}\mathbf{W}(k)\mathbf{x}_i(k)\right)\right\} \end{gathered} \tag{6.23}$$

After appropriate manipulation and ignoring the constant terms, Equation (6.23) takes the following form:

$$= -\sum_{i=1}^{t} z(q_i)\,\{\lambda\}\left\{\frac{(\nu/2 + d/2)}{\nu}(v_1(k) - v_2(k))\left(\mathbf{x}_i(k)^{\dagger}\mathbf{W}(k)^{\dagger}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\mathbf{W}(k)\mathbf{x}_i(k)\right)\right\} \tag{6.24}$$

$$\begin{aligned} &= -\sum_{i=1}^{t} z(q_i)\,\{\lambda\} \\ &\quad \left\{\frac{(\nu/2 + d/2)}{\nu}(v_1(k) - v_2(k))\left(\mathbf{x}_i(k)^{\dagger}\mathbf{W}(k)^{\dagger}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}\mathbf{W}(k)\mathbf{x}_i(k)\right)\right\} \\ &\quad + \beta_k(a_k a_k^* + b_k b_k^* - 1) \end{aligned} \tag{6.25}$$

Now by replacing the value of $\mathbf{x}_i(k)$ and $\mathbf{W}(k)$ for the 2 x 2 case:

$$\begin{aligned}&= -\sum_{i=1}^{t} z(q_i)\,\{\lambda\}\,\frac{(\nu/2+d/2)}{\nu}\Big(v_1(k)-v_2(k)\Big)\\ &\left\{\begin{pmatrix}\mathbf{x}_1(k) & \mathbf{x}_2(k)\end{pmatrix}\begin{pmatrix}a_k & -b_k^*\\ b_k & a_k^*\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}\begin{pmatrix}a_k & b_k\\ -b_k^* & a_k^*\end{pmatrix}\begin{pmatrix}\mathbf{x}_1(k)\\ \mathbf{x}_2(k)\end{pmatrix}\right\}\\ &+\beta_k(a_k a_k^* + b_k b_k^* - 1)\end{aligned} \tag{6.26}$$

After the matrix multiplication, the previous equation takes the following form:

$$\begin{aligned}= -\sum_{i=1}^{t} z(q_i)\,\{\lambda\}\left\{\frac{(\nu/2+d/2)}{\nu}(v_1(k)-v_2(k))(a_k\mathbf{x}_1(k)+b_k\mathbf{x}_2(k))(a_k\mathbf{x}_1(k)+b_k\mathbf{x}_2(k))\right\}\\ +\beta_k(a_k a_k^* + b_k b_k^* - 1)\end{aligned} \tag{6.27}$$

$$\begin{aligned}&= -\sum_{i=1}^{t} z(q_i)\,\{\lambda\}\left\{\frac{(\nu/2+d/2)}{\nu}(v_1(k)-v_2(k))(a_k\mathbf{x}_1(k)+b_k\mathbf{x}_2(k))^2\right\}\\ &+\beta_k(a_k a_k^* + b_k b_k^* - 1)\end{aligned} \tag{6.28}$$

Now by taking the derivative of above mentioned equation with respect to a_k and equating it equals to zero.

$$\begin{aligned}&=\sum_{i=1}^{t} z(q_i)\,\{\lambda\}\left\{\frac{(\nu/2+d/2)}{\nu}(v_1(k)-v_2(k))(2(a_k\mathbf{x}_1(k)+b_k\mathbf{x}_2(k))\mathbf{x}_i(k)^\dagger)\right\}\\ &+a_k^*\beta_k = 0\end{aligned} \tag{6.29}$$

$$\sum_{i=1}^{t} z(q_i)\left\{\lambda\right\}\left\{\frac{(\nu/2+d/2)}{\nu}(v_1(k)-v_2(k))\begin{pmatrix}\mathbf{x}_i(k) & \mathbf{x}_i(k)\end{pmatrix}\begin{pmatrix}a_k \\ b_k\end{pmatrix}\right\} = a_k^*\beta_k \tag{6.30}$$

Likewise, taking the derivative with respect to b_k and equating it to zero

$$\sum_{i=1}^{t} z(q_i)\left\{\lambda\right\}\left\{\frac{(\nu/2+d/2)}{\nu}(v_1(k)-v_2(k))\begin{pmatrix}\mathbf{x}_i(k) & \mathbf{x}_i(k)\end{pmatrix}\begin{pmatrix}a_k \\ b_k\end{pmatrix}\right\} = b_k^*\beta_k \tag{6.31}$$

Assuming $\mathbf{M_{ik}} = \sum_{i=1}^{t} z(s_i)\frac{(\nu/2+d/2)}{\nu}(v_1(k)-v_2(k))\mathbf{x}_i(k)\mathbf{x}_i(k)^{\dagger}$ and by using Equations (6.30) and (6.31):

$$\mathbf{M_{ik}}\begin{pmatrix}a_k^* \\ b_k^*\end{pmatrix} = \beta_k\begin{pmatrix}a_k^* \\ b_k^*\end{pmatrix} \tag{6.32}$$

where vector $(a_k, b_k)^{\dagger}$ is the eigenvector of $\mathbf{M_{ik}}$ with the smaller eigenvalue. This can be found by replacing $\mathbf{M_{ik}}$ in Equation (6.25) and taking trace of the equation:

$$-Tr\left\{\mathbf{M_{ik}}\begin{pmatrix}a_k^* \\ b_k^*\end{pmatrix}\begin{pmatrix}a_k & b_k\end{pmatrix}^{\dagger}\right\} = \beta_k \tag{6.33}$$

where $Tr(x)$ denotes the trace of the matrix. Whereas the eigenvectors associated with the smaller eigenvalues will give the higher value of the cost function. Therefore $(a_k, b_k)^{\dagger}$ is the eigenvector of $\mathbf{M}_{ik}$ with the smaller eigenvalue. In order to calculate the eigenvalues associated with the $\mathbf{M}_{ik}$ for the 2 x 2 case, $\mathbf{M}_{ik}$ can be written as:

$$\mathbf{M}_{ik} = \begin{pmatrix}M_{11} & M_{12} \\ M_{21} & M_{22}\end{pmatrix} \tag{6.34}$$

where M_{11}, M_{22} are real and $M_{21} = M_{12}^*$, because $\mathbf{M}_{ik}$ is Hermitian. Eigenvalues in this case can be calculated as:

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0 \tag{6.35}$$

$$det \left\{ \begin{pmatrix} M_{11} - \lambda & M_{12} \\ M_{21} & M_{22} - \lambda \end{pmatrix} \right\} = 0 \tag{6.36}$$

$$(M_{11} - \lambda)(M_{22} - \lambda) - (M_{12})(M_{21}) = 0 \tag{6.37}$$

Since the above equation is a quadratic equation, therefore the quadratic formula can be used to find the eigenvalues which are $\frac{M_{11}+M_{22}}{2} \pm \sqrt{\frac{(M_{11}-M22^2}{4} + |M_{12}|^2}$, so the smaller eigenvalue can be written as:

$$\beta_k = \frac{M_{11} + M_{22}}{2} - \sqrt{\frac{(M_{11} - M22^2}{4} + |M_{12}|^2} \tag{6.38}$$

From the eigen value the corresponding eigenvector can be calculated, which is as follows:

$$\begin{pmatrix} a_k^* \\ b_k^* \end{pmatrix} = \frac{1}{\sqrt{1 + (\frac{\beta_k - M_{11}}{M_{12}})^2}} \begin{pmatrix} 1 \\ \frac{\beta_k - M_{11}}{M_{12}} \end{pmatrix} \tag{6.39}$$

Since the unmixing matrix $\mathbf{W_i(k)} = \begin{pmatrix} a_k & b_k \\ -b_k^* & a_k^* \end{pmatrix}$, so it can be estimated by using the above mentioned analytical solution. It is an efficient method to estimate the unmixing matrix as it avoids the matrix calculations.

The model parameters $\theta = \{\mathbf{W}(k), \mathbf{\Lambda}(k), p(q_i)\}$ are estimated by maximizing the log likelihood function. Therefore, now $\mathfrak{F}(z, \theta)$ will be maximized over $\mathbf{\Lambda}_{iK}$. In order to estimate $\mathbf{\Lambda}_i(k)$, $p(\mathbf{x}_i(k)|\mathbf{q}_i)$ can be replaced in Equation (6.7) as follows:

$$\mathcal{L}(\mathbf{x}, \theta) = \sum_{i=1}^{t} z(q_i)\log\left(\frac{\prod_{k=1}^{K} \frac{|\mathbf{\Lambda}_i(k)|}{\pi}\left(1+\frac{\mathbf{x}_i(k)^{\dagger}\mathbf{\Lambda}_i(k)\mathbf{x}_i(k)}{\nu}\right)^{-\nu/2-d/2} p(q_i)}{z(q_i)}\right) \tag{6.40}$$

After appropriate manipulation and ignoring the constant terms, the above equation will take the following form.

$$= \sum_{i=1}^{t} z(q_i)\left\{\log(\frac{|\mathbf{\Lambda}_i(k)|}{\pi}) + \left\{\log\left(1+\frac{\mathbf{x}_i(k)^{\dagger}\mathbf{\Lambda}_i(k)\mathbf{x}_i(k)}{\nu}\right)^{-\nu/2-d/2}\right\}\right\} \tag{6.41}$$

$$= \sum_{i=1}^{t} z(q_i)\left\{\log(\frac{|\mathbf{\Lambda}_i(k)|}{\pi}) + \left\{(-\nu/2-d/2)\log\left(1+\frac{\mathbf{x}_i(k)^{\dagger}\mathbf{\Lambda}_i(k)\mathbf{x}_i(k)}{\nu}\right)\right\}\right\} \tag{6.42}$$

Again, by using the log approximation $\log(1+a) \approx a$, where a is a small value, the above equation can take the following form, wherein equality is consider for convenience.

$$= \sum_{i=1}^{t} z(q_i)\left\{\log\left(\frac{|\mathbf{\Lambda}_i(k)|}{\pi}\right) + \left\{(-\nu/2-d/2)\left(\frac{\mathbf{x}_i(k)^{\dagger}\mathbf{\Lambda}_i(k)\mathbf{x}_i(k)}{\nu}\right)\right\}\right\} \tag{6.43}$$

By replacing the value of $\mathbf{\Lambda}_i(k)$ in the Equation (6.43), it will take the following form:

$$=\sum_{i=1}^{t} z(q_i)$$
$$\left\{\log\left(\frac{|\mathbf{W}(k)^{\dagger}\mathbf{\Phi}_i(k)\mathbf{W}(k)|}{\pi}\right)+\left\{(-\nu/2-d/2)\left(\frac{\mathbf{x}_i(k)^{\dagger}\mathbf{W}(k)^{\dagger}\mathbf{\Phi}_i(k)\mathbf{W}(k)\mathbf{x}_i(k)}{\nu}\right)\right\}\right\} \tag{6.44}$$

Now in order to maximize it over the precision $\mathbf{\Lambda_{ik}}$, the derivative of the Equation(6.44) with respect to v_{k1} is taken to yield

$$=\sum_{i=1}^{t} z(q_i)\left\{\left(\frac{1}{v_1(k)}\right)-\left\{\frac{(-\nu/2-d/2)}{\nu}\left(\mathbf{x}_i(k)^{\dagger}\mathbf{W}(k)^{\dagger}\mathbf{W}(k)\mathbf{x}_i(k)\right)\right\}\right\} \tag{6.45}$$

Therefore,

$$\frac{1}{v_{ik_{j=r}}}=\left(\frac{-\nu/2-d/2}{\nu}\right)\frac{\left[\sum_{i=1}^{t} z(q_{ij=r})(\mathbf{x}_i(k)^{\dagger}\mathbf{W}(k)^{\dagger}\mathbf{W}(k)\mathbf{x}_i(k))\right]_{jj}}{\sum_{i=1}^{t} z(q_{i=jr})} \tag{6.46}$$

where $[.]_{jj}$ denotes the (j,j) element of the matrix. So $\mathfrak{F}(\mathbf{z},\theta)$ over $\mathbf{\Lambda}_i$ using the above mentioned solution.

Now, maximisation of $\mathfrak{F}(\mathbf{z}, \theta)$ over $p(q_i)$ is performed. The lower bound of the log likelihood equation is:

$$\mathfrak{F}(z, \theta) = \sum_{i=1}^{t} z(q_i) \log\left(\frac{\prod_{k=1}^{K} p(\mathbf{x}_i(k)|q_i) p(q_i)}{z(q_i)} \right) \tag{6.47}$$

If $\mathbf{q}_i$ can take s possible states, $p(q_t) = r$ has to satisfy $\sum_r p(\mathbf{q}_i = r) = 1$. So $p(q_t) = r$ doesn't have s degrees of freedom instead it has $q - 1$ free parameters. So, the Lagrange multiplier is used in this case. Therefore the cost function can be described as:

$$\sum_{i=1}^{t} z(q_i) \log\left(\frac{p(\mathbf{x}_i(k)|x_i) p(q_i)}{z(q_i)} \right) + \beta \left(1 - \sum_{i=1}^{t} \left(p(q_i = r) \right) \right) \tag{6.48}$$

Now taking the derivative of the above mentioned equation with respect to $p(q_i = r)$ and equating it to zero

$$\sum_{i=1}^{t} z(q_i = r) \left(\frac{1}{p(q_i = r)} \right) - \beta = 0 \tag{6.49}$$

$$p(q_i = r) = \frac{\sum_{i=1}^{t} z(q_i = r)}{\beta} \tag{6.50}$$

Now $p(q_i = r) = 1$ and $\sum_{i=1}^{t} z(q_i = r) = 1$, therefore the above equation can be rewritten as:

$$1 = \frac{\sum_{i=1}^{t}}{\beta} \Rightarrow 1 = \frac{t}{\beta} \Rightarrow \beta = t \tag{6.51}$$

Hence,

$$p(q_i = r) = \frac{\sum_{i=1}^{t} z(q_i = r)}{t} \tag{6.52}$$

Hence the weighting parameter can be calculated by using the above mentioned equation. It can be seen that the EM algorithm effectively estimates all the model parameters $\theta = \{\mathbf{W}_i, \mathbf{\Lambda}_i, p(q_i)\}$. The E-step updates the $z(q_i)$, while the M-steps effectively estimates the model parameters. In the EM algorithm the degrees of freedom parameter ν is fixed in advance for all the sources, then the M-step exists in the close form. The value for degrees of freedom can be estimated empirically for different source signals. The complete EM algorithm for the IVA algorithm by using the Student's t mixtures model is summarized as follows.

Require: Given a Student's t mixture model, the aim is to maximize the log likelihood function with respect to the parameters $\theta = \{\mathbf{W}_i, \mathbf{\Lambda}_i, p(q_i)\}$.

1: Initialize the model parameters, the unmixing matrix $\mathbf{W}_i$, the precision $\mathbf{\Lambda}_i$ and the weight coefficients $p(q_i)$ and evaluate the initial value of the log likelihood.

2: **Expectation Step**: Evaluate the probabilities using the current parameter values

$$z(x_i) = \frac{\prod_{k=1}^{K} p(\mathbf{x}_i(k)|q_i)p(q_i)}{p(x_i(1), \cdots, x_i(K))} \tag{6.53}$$

3: **Maximization Step**: Using the current probabilities, re-estimate the parameters

- Coefficient of the unmixing matrices $\mathbf{W}_i$ are estimated by:

$$\begin{pmatrix} a_k^* \\ b_k^* \end{pmatrix} = \frac{1}{\sqrt{1 + (\frac{\beta_k - M_{11}}{M_{12}})^2}} \begin{pmatrix} 1 \\ \frac{\beta_k - M_{11}}{M_{12}} \end{pmatrix} \tag{6.54}$$

- Coefficients of the precision matrix $\mathbf{\Lambda}_i$ are estimated by

$$\frac{1}{v_{ik_{j=r}}} = \left(\frac{-\nu/2 - d/2}{\nu}\right) \frac{\left[\sum_{i=1}^{t} z(x_{ij=r})(\mathbf{x}_i(k)^\dagger \mathbf{W}(k)^\dagger \mathbf{W}(k) \mathbf{x}_i(k)\right]_{jj}}{\sum_{i=1}^{t} z(x_{i=jr})} \tag{6.55}$$

- The weighting coefficients can be estimated as

$$p(q_i = r) = \frac{\sum_{i=1}^{t} z(q_i = r)}{t} \tag{6.56}$$

4: Evaluate the log likelihood

$$\mathcal{L}(\mathbf{x}, \theta) = \sum_{i=1}^{t} \log\left(\sum_{x_i} \prod_{k=1}^{K} p(\mathbf{x}_i(k)|q_i)p(q_i)\right) \tag{6.57}$$

and check for convergence of the log likelihood function, if the criterion for convergence is not fulfilled, return to step 2.

Algorithm 1: EM algorithm for Student's t Mixtures

The separation performance of this EM framework for the IVA method will be evaluated in the next section.

6.5 Experiments and Results

In this section, the separation performance of the EM framework for the IVA algorithm will be tested in three different experimental setups. Firstly the new framework for the IVA algorithm will be tested in simulated environment and then its separation performance will also be tested in real RIRs, which can depict the performance of the proposed method in realistic scenarios. The results from all three sets of experiment for the proposed algorithm will be compared with the original IVA algorithm with different source priors.

6.5.1 Simulations with the Image Method

Table 6.1: Summary of parameters used in experiments.

Sampling rate	8kHz
STFT frame length	1024
Reverberation time	200 ms
Room dimensions	7 m x 5 m x 3 m
Source signal duration	4 s (TIMIT)
Room impulse responses	Image method
Objective measure	Signal to Distortion Ratio (SDR)

Firstly the proposed method will be tested with RIRs that were generated by using the image method. The speech signals were selected randomly form the whole of TIMIT dataset [79] and the length of speech signals were approximately 4 seconds. A 2 x 2 case was considered and the room has the RT_{60} = 200ms and it provides a good setup for comparing the evaluation the separation performance of different algorithms. The position of microphones in the room were set to [3.44, 2.50, 1.50] and [3.48, 2.50, 1.50] with azimuth angles of 60° and 30°, respectively with reference to the normal of the microphone position. The STFT length is 1024 and sampling frequency is 8kHz. The separation performance of the algorithm was evaluated with the objective measure of SDR [77]. The common parameters used in these experiments are summarised in Table 6.1.

Table 6.2: SDR (dB) values for different source priors for the IVA method with image room impulse response [84]. The SMM source prior shows improvement for all mixtures.

	Original super Gaussian [21]	Student's t [76]	SMM source prior
Set-1	9.09	9.84	10.27
Set-2	8.98	9.72	10.24
Set-3	9.26	10.11	10.87
Set-4	9.02	9.95	10.49
Set-5	9.53	10.21	10.62
Set-6	9.51	10.14	10.74
Set-7	8.91	9.67	10.09
Set-8	9.86	10.48	11.05
Set-9	9.94	10.66	11.24
Set-10	10.02	10.56	10.97

The speech signals were convolved into mixtures in the above mentioned room settings. These speech mixtures were then separated by using the SMM source prior based IVA method and the separation results for different mixtures were compared with the separation performance of the original IVA method with the original super Gaussian source prior [21] and also with the IVA method with Student's t source prior [76] and the results are shown in Table 6.2 and all the values shown for SDR are in dB. For each mixture SDR performance shown in the Table 6.2 is the average of two speech signals. It is evident from the Table 6.2 that the SMM is adopted as a source prior, the average SDR improvement is approximately 1.1 dB for all the mixtures as compared to the original super Gaussian source prior for the IVA method.

Table 6.3: Comparison between SMM and GMM source prior for the EM framework IVA algorithm with Imaging method [84]. Proposed SMM source prior for the EM framework IVA has better separation performance for all the mixtures.

	GMM source prior [100]	SMM source prior
Set-1	9.91	10.34
Set-2	9.28	9.62
Set-3	10.32	10.71
Set-4	10.04	10.45
Set-5	9.93	10.27
Set-6	9.42	9.84
Set-7	10.19	10.57
Set-8	9.56	10.05
Set-9	9.84	10.26
Set-10	10.02	10.39

It is evident from Table 6.2 that the SMM source prior based IVA algorithm enhance the separation performance of the IVA method with single distribution source prior such as the Student's t distribution and also the original super Gaussian. Therefore to further investigate the separation performance of the SMM source prior for the IVA algorithm, its separation results are compared with the other mixture model source prior such as Gaussian mixture model [100]. The same experimental settings were used for this experiment. Same room of size 7 x 5 x $3m^3$ with RT_{60} of 200ms was used. The two speech sources were positioned at [4.6, 3.25, 1.5] and [2.7, 3.8, 1.5] respectively. Then the new IVA algorithm with SMM as source prior and the IVA algorithm with GMM as source prior was implemented to separate the speech mixtures. The separation performance of both source priors is shown in Table 6.3. Again, all the SDR (dB) values shown in table are the average of SDR for two separated signals. It is clear from Table 6.3 that when SMM is used as source prior for the IVA technique, it consistently shows the better separation performance, as it improves the separation performance by approximately 0.4 dB when compared with with other mixture models i.e. GMM, as a source prior for the IVA method.

6.5.2 Simulations with Real RIRs

In the second set of experiments, the proposed framework of EM algorithm for the IVA method is tested with real RIRs. These real RIRs were obtained from [88] and these are recorded in different rooms with different acoustic properties. Three different room types (A, B,D) have been used with RT_{60} of 320ms, 470ms and 890ms, respectively. By using these RIRs the proposed method can be tested with the range of reverberation time. Therefore, these simulations show the performance of the proposed algorithm in real life scenarios as the RT_{60} can vary drastically in realistic environments. There are source location azimuth angles available which are ranging from (15° to 90°) relative to the second source.

Firstly, the proposed algorithm is tested in the Room A, which is a typical medium sized office and it has the RT_{60} of 320ms, which is relatively small for a medium size office. In the experiments two speech signals are randomly chosen from the whole of TIMIT dataset and the source location azimuth angles are set to be from (15° to 90° with a step of 15°). The mixed sources are separated by using the proposed IVA method with SMM source prior and the separation performance in terms of SDR is compared with the IVA using the identical Student's t source prior [76] and also with the original super Gaussian source prior based IVA method [21]. The separation performance for both methods is evaluated for six different angles varying from (15° to 90° with a step of 15°). At all the angles separation performance is averaged over six different speech mixture and the results are presented in Figure 6.1. It is evident from Figure 6.1 that when proposed algorithm is used to separate the mixtures and the performance is compared with identical distribution source prior for the IVA, it consistently has a better separation performance at all the selected azimuths angles and approximately 1.1dB of improvement in SDR values is recorded at all the angles as compared to the

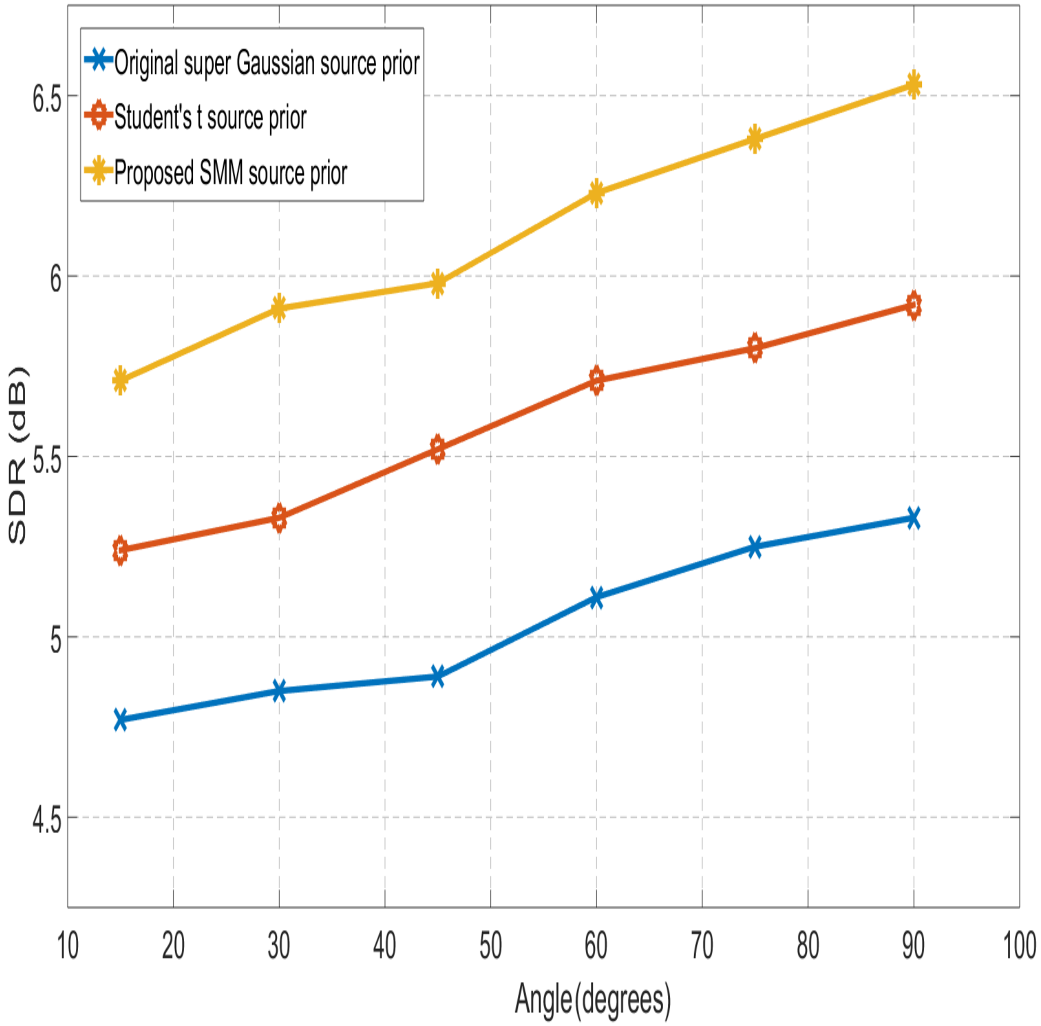

Figure 6.1: Comparison between original IVA with original super Gaussian source prior, Student's t source prior and EM framework IVA with SMM source prior for Room-A (RT_{60} = 320ms). The separation performance at each angle is averaged over six different speech mixtures. The proposed mixture model IVA perform better then single Student's t distribution at all the separation angles

original IVA method [21].

In order to evaluate the separation performance of the proposed algorithm in changing realistic scenarios, it is further tested in the Room B. It is a medium size class room which has RT_{60} of 470ms, which is a highly reverberant room environment and therefore, it presents a good estimate of the separation performance of the algorithm in realistic environment. Again, all the speech signals are randomly chosen from whole of the TIMIT dataset. In this room, same experimental settings were used as in case of Room A and the speech sources were separated at six different azimuth angles varying from (15° to 90° with a step of 15°).

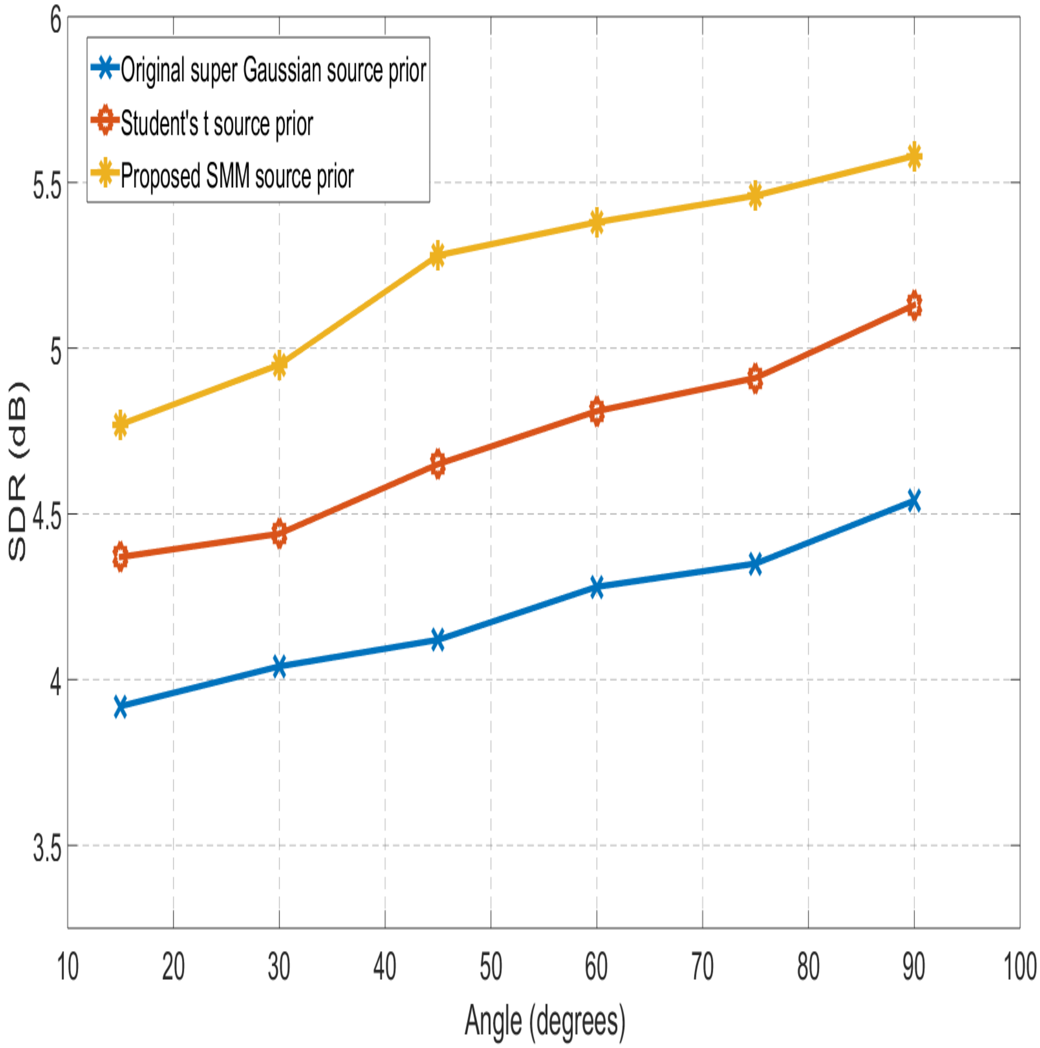

Figure 6.2: Comparison between original IVA with Student's t source prior and EM framework IVA with SMM source prior for Room-B (RT_{60} = 470ms). The separation performance at each angle is averaged over six different speech mixtures. The proposed mixture model IVA perform better then single Student's t distribution at all the separation angles.

The separation performance in terms of SDR of both methods for six different azimuth angles is showed in Figure 6.2. In order to enhance the reliability of the results, at all the angles separation performance shown is the average of six different speech mixtures for all methods. The value of SDR increases as the angle between the sources is increased from 15° to 90°. In comparison with Room A, the overall SDR values are decreased for all the angles because of the high RT_{60} of Room B. From Figure 6.2, it is evident that the EM framework IVA with SMM source prior perform better than the identical source priors for the original IVA method at all separation angles in highly reverberant real room environment.

Finally, the separation performance of the proposed EM framework for the IVA method is evaluated in a highly reverberant realistic environment that can depict the performance of the algorithm in the real life scenarios. For the highly reverberant environment, Room D was used which is a medium size seminar and presentation hall with a very high ceiling. The RT_{60} for this seminar hall is 890ms, which is high reverberation time and therefore it provides a good insight into algorithm's performance in a extremely difficult real life situations.

The experimental setup in this highly reverberant room D is similar to previous two rooms. Again, two speech signals were randomly chosen from the whole of TIMIT database each time and they were convolved in room D with high RT_{60} of 890ms. Experiments were performed by varying the azimuth angle of the source location relative to microphone location by 15° from 15° to 90°. The mixtures were separated with the IVA method with different source priors and the separation performance in terms of SDR for all methods is shown in Figure 6.3 for all six angles varying from 15° to 90°. As the angle between the sources increased, the separation performance is improved. The SDR values in room D is lower in comparison SDR values for Room-A and Room-B, it is mainly because the RT_{60} for Room D is really high as compared to the other two rooms. Also, it is evident from the Figure 6.3 that even in highly reverberant environment the IVA method with SMM source prior performs better than the identical distribution source

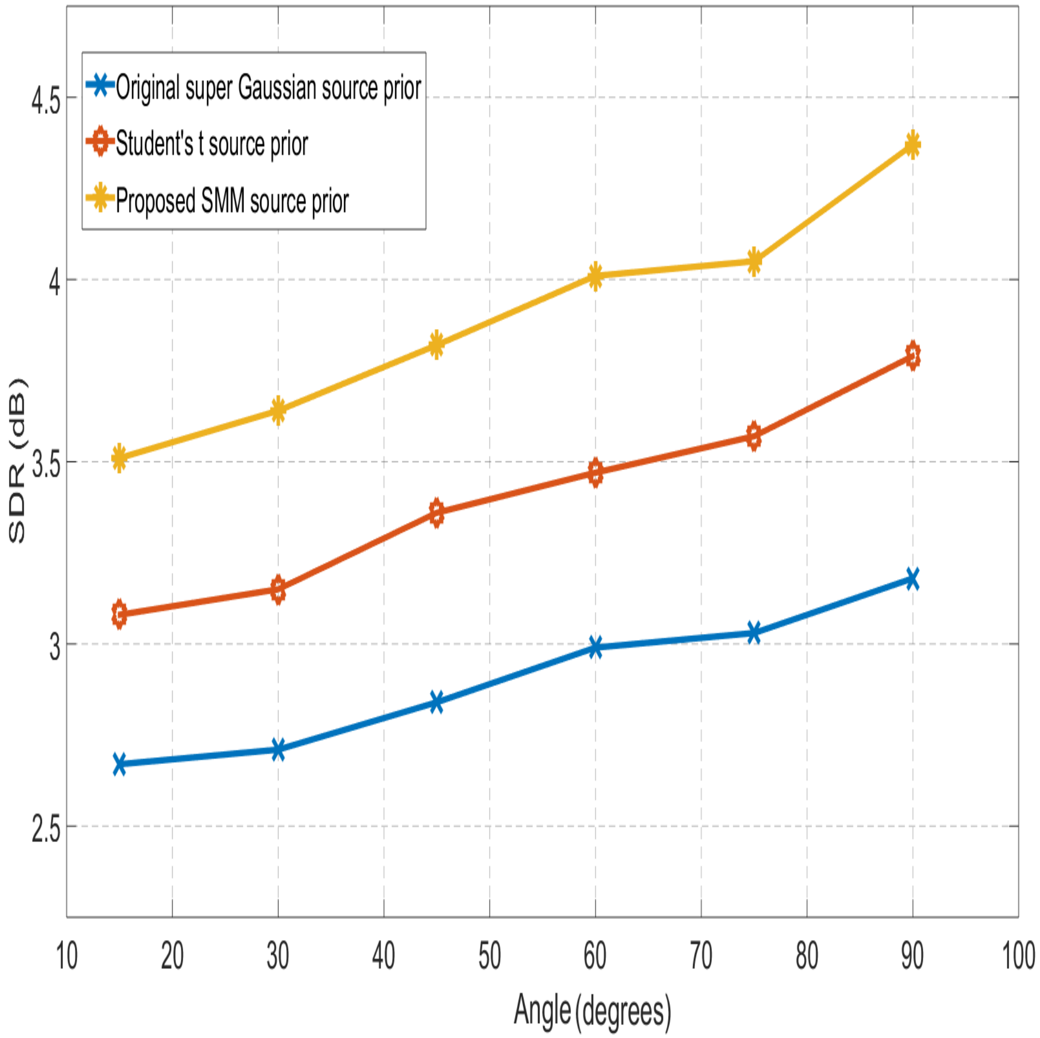

Figure 6.3: Comparison between Original IVA with Student's t source prior and EM framework IVA with SMM source prior for Room-D (RT_{60} = 890ms). The separation performance at each angle is averaged over six different speech mixtures. The proposed mixture model IVA perform better then single Student's t distribution at all the separation angles..

priors for the original IVA with Student's t source prior at all the separation angles.

The separation performance of the proposed EM framework for the IVA algorithm with SMMs as a source prior is also compared with the IVA algorithm with GMM as a source prior. Since the mixture model is adapted as a source prior for the IVA algorithm, the comparison with other mixture models i.e. GMM can provide the better understanding of the separation performance of the proposed source prior. Therefore the same experimental settings for Room A, B and D are used as in previous case and speech signals are randomly chosen from TIMT dataset. Firstly, experiments are performed in room A, which has RT_{60} of 320ms and it is repeated for six different source location varying from 15° to 90°. Similarly the same experimental setup is used for room B with RT_{60} of 470ms and for room D with RT_{60} of 890ms. In all the rooms mixtures are separated by using EM framework IVA with both SMM and GMM source priors and the separation performance in terms of SDR is compared with the proposed method at six different source azimuth angles varying from 15° to 90°. All the SDR values at all the angles are the average of separation performance of six different mixtures. The separation performance of both methods for all three rooms with the range of RT_{60} is shown is Figure 6.4 and it is evident that the IVA method with SMM as a source prior has better separation performance then IVA with GMM as a source prior.

6.5.3 Simulations with Binaural Room Impulse Responses

The proposed algorithm is further tested with binaural room impulse response (BRIRs) obtained from [87]. These BRIRs are recorded in a real classroom which roughly has dimensions of 5 x 9 x 3.5m^3. The six source location azimuths $(15°, 30°, 45°, 60°, 75°, 90°)$ to the right of listener were used for the experimentation. Also distance between the source were changed three times $(0.15, 0.40$ and 1 m). The measurements for the BRIRs are taken at four different listener locations (back, ear, corner and center) and the distance between the floor and ears was approximately 1.50m. In these experiments only center location is used and the RT_{60} at the center location for the classroom was 565ms. All the measurements are repeated at three different occasions by taking down the equipment and re-

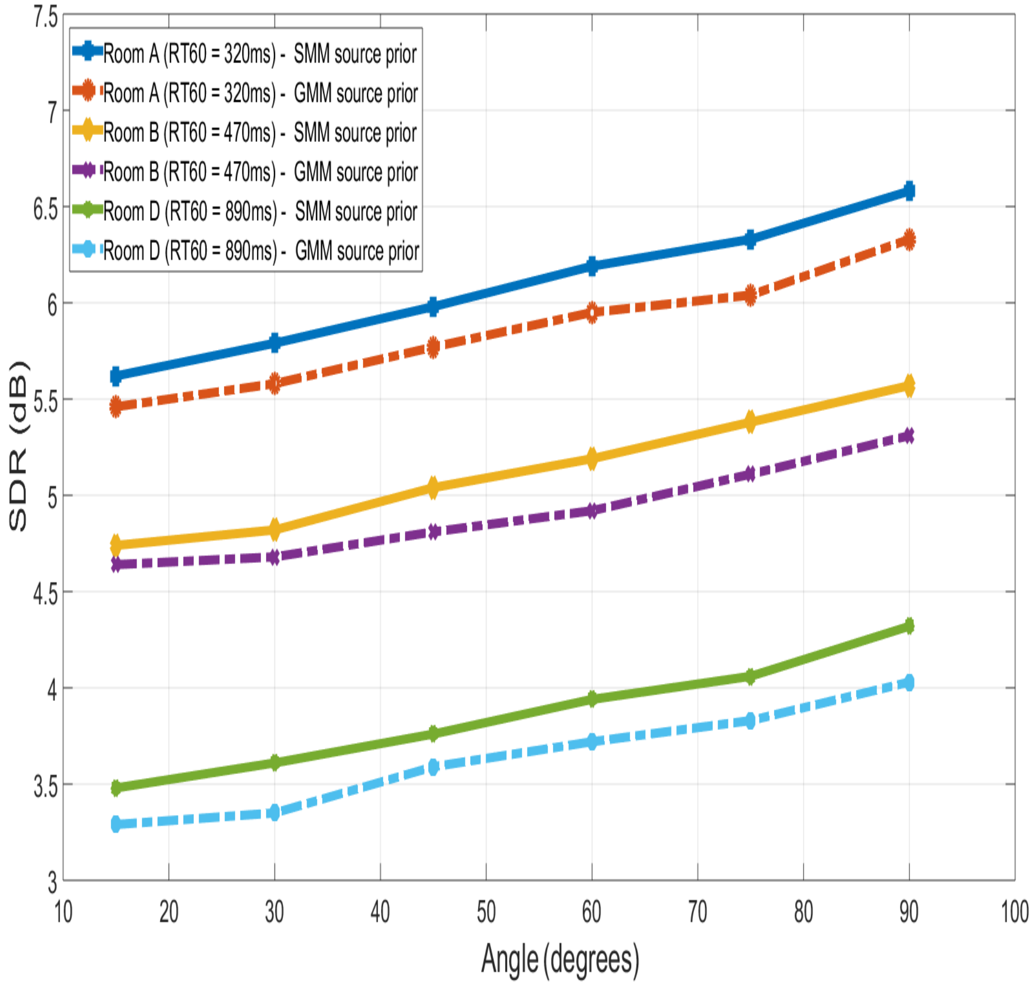

Figure 6.4: Comparison between EM framework IVA with SMM and GMM source prior for three different rooms (Room-A, Room-B, Room-D). The separation performance at each angle is averaged over six different speech mixtures. The EM framework IVA algorithm with proposed SMM source prior perform better then GMM source prior at all the separation angles.

assembled which improves the reliability of the measurements. Therefore these BRIRs has been used in the experiments as they are reliable and also provide the accurate estimate of the separation performance of the BSS algorithms in the highly reverberant room environment. A summary of different parameters used in this set of experiments is given in Table 6.4.

The 2 x 2 case was considered for the experiments and speech signals were ran-

Table 6.4: Summary of parameters used in experiments.

Sampling rate	8kHz
STFT frame length	1024
Velocity of sound	343 m/s
Reverberation time	565 ms (BRIRs)
Room dimensions	9 m x 5 m x 3.5 m
Source signal duration	3.5 s (TIMIT)

domly chosen from the whole TIMIT dataset and mixtures were created by using BRIRs. The length of the speech signals were approximately four seconds. The speech signals were then separated from the mixtures by using the proposed EM framework for the IVA algorithm with SMM as source prior. The separation performance of the proposed algorithm is compared with the separation performance of the IVA with GMM as source prior for the IVA algorithm. It provides a good estimate for the separation performance of the proposed algorithm and source prior as comparison is drawn with mixture model source priors. The separation performance in terms of SDR is shown in Figure 6.5 for the six different source location $(15°, 30°, 45°, 60°, 75°, 90°)$. All the experiments are repeated three times and at each source location six different speech mixtures are separated. In order to improve the reliability of results shown in Figure 6.5, all the SDR values are the averaged of separation performance of the algorithms over eighteen different speech mixture.

From Figure 6.5 it is evident that when SMM is used as a source prior for the IVA algorithm it performs better as compared with the GMM as a source prior. Since speech signals are highly non-stationary in nature and there can be many useful samples in outliers which might not be properly modelled with the Gaussian mixtures but Student's t mixtures because of its heavy tails can model the outliers information and therefore enhance the separation results of the IVA method.

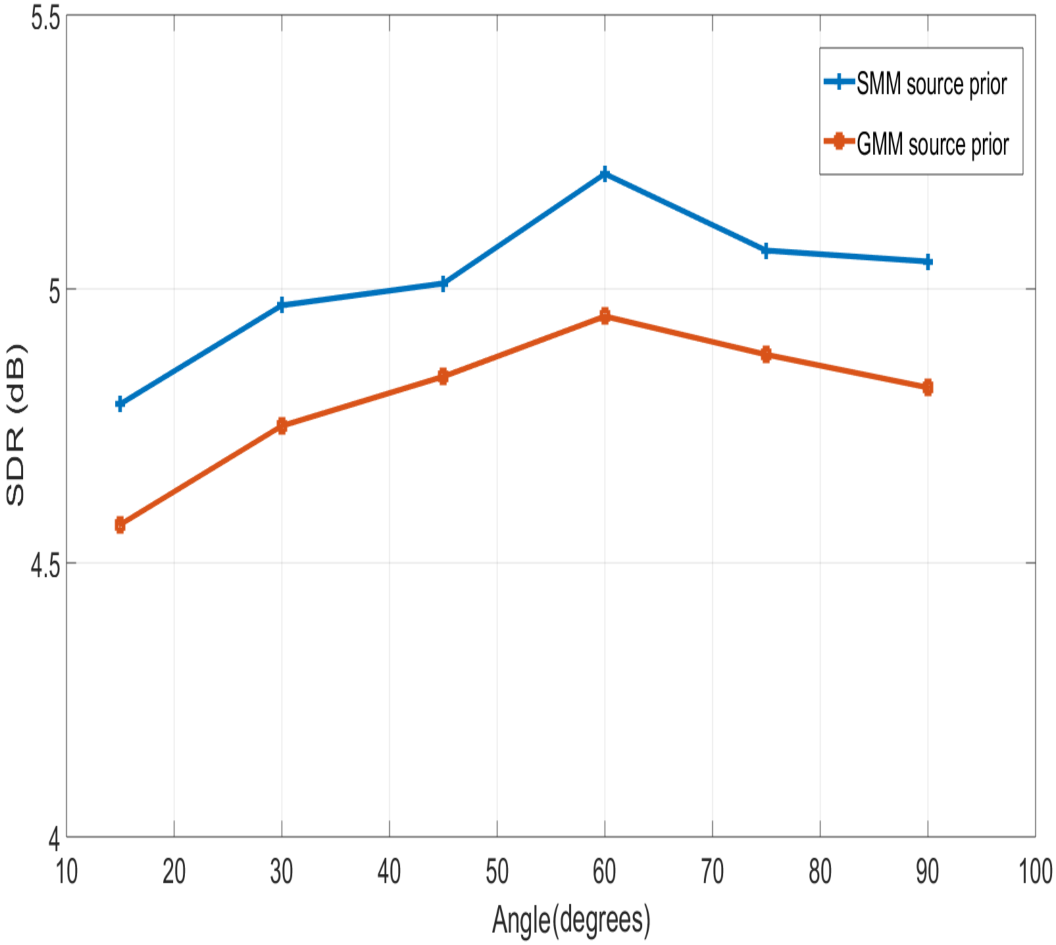

Figure 6.5: Comparison between SMM source prior and GMM source prior for the EM framework IVA algorithm with BRIRs ($RT_{60} = 565$ms). The separation performance at each angle is averaged over eighteen different speech mixtures. The IVA algorithm with proposed mixture model Student's t source prior perform better at all the separation angles then the GMM source prior.

When SMM is adopted as a source prior for the IVA method, at all the source location azimuths it improves the average separation performance for the IVA method by approximately by 0.4dB, as shown in Figure 6.5.

Furthermore,the separation performance is evaluated with the subjective measure of PESQ. This subjective measure compare original signals and separated signals and gives a score from 0 to 4.5, 0 for the poor separation performance and 4.5 being the excellent separation performance. This measure therefore provides a good estimate about the similarity between the original and separated sources. So the speech mixtures made with BRIRs are separated with the proposed SMM source prior for the EM framework IVA and also with the GMM source prior IVA and the PESQ score is calculated for both the methods. The PESQ score for the IVA method with both source priors is shown in Table 6.5 and the IVA method with SMM source prior consistently has the better PESQ score as compared with the GMM source prior for the IVA algorithm. Therefore it is evident from the table that when SMM is adapted as a source prior, it improves the separation performance for the IVA method.

Table 6.5: PESQ score for GMM and SMM source prior for the IVA algorithm. PESQ score shown is the average over six different locations in the room. SMM source prior for the IVA algorithm provides the better estimate of source signals.

	GMM Source Prior	SMM Source Prior
Set-1	1.85	2.02
Set-2	1.98	2.11
Set-3	1.96	2.13
Set-4	2.02	2.19
Set-5	1.93	2.14
Set-6	2.08	2.21

Finally, the separation performance of the proposed EM framework for the IVA method with SMM as source prior is compared with original IVA with identical source priors. BRIRs with RT_{60} of 565ms are used to evaluate the algorithms in highly reverberant environment that can depict the performance of the algorithms in the realistic scenarios. Same experimental settings are used as in first experiments and the source location is varied six times from $(15°, 30°, 45°, 60°, 75°, 90°)$. All the measurements are repeated three times and six different speech mixtures are separated at each angle by using IVA method with SMM as source prior and the results are compared with the separation performance of IVA method with multivariate Student's t distribution as source prior, the IVA method with original multivariate super Gaussian source prior and also with IVA method with the mixed Student's t and original super Gaussian source prior. This provides an overall comparison of the separation performance of different source prior and the framework for the IVA method. The results in terms of SDR (dB) for six different source location are shown in figure 6.6.

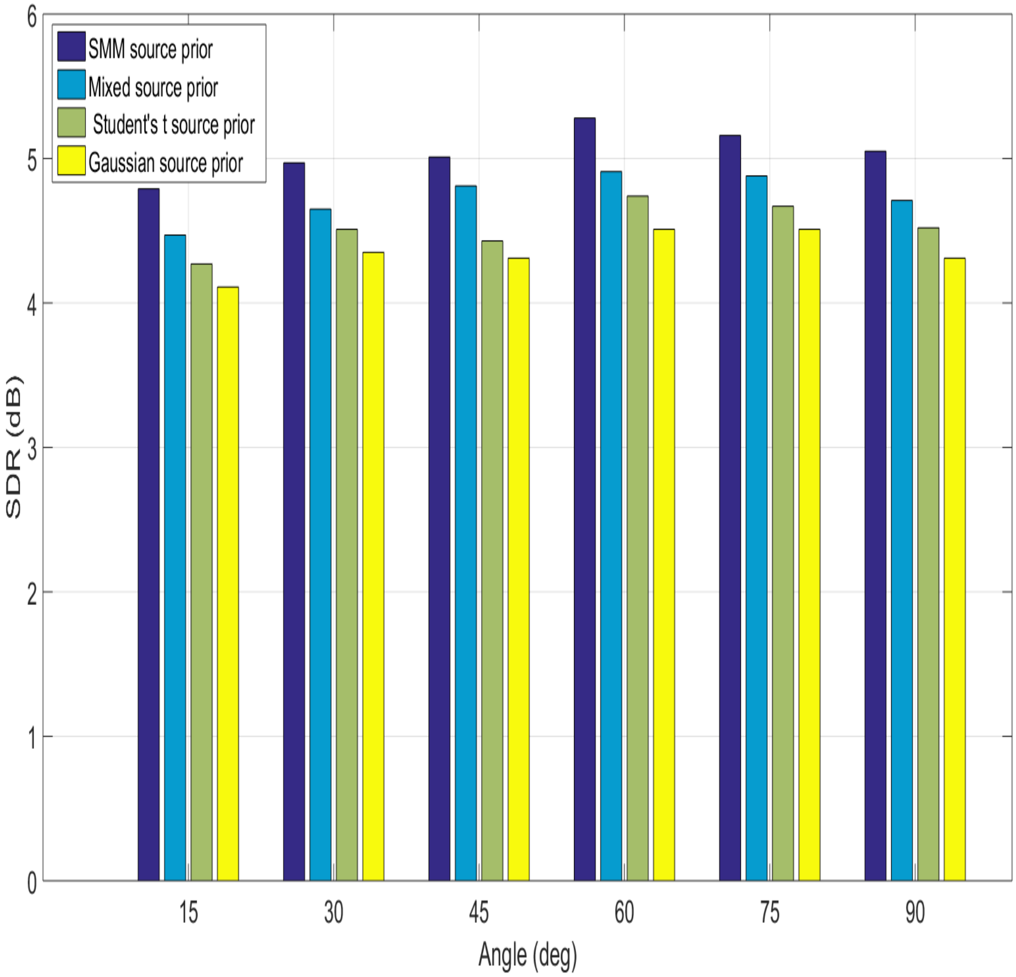

Figure 6.6: Comparison between different source priors for the IVA algorithm for BRIRs (RT_{60} = 565ms). The separation performance at each angle is averaged over eighteen different speech mixtures. The IVA algorithm with proposed mixture model Student's t source prior perform better at all the separation angles in comparison to identical source prior for all the sources.

It is evident from Figure 6.6 that the mixture model source prior performs better then the identical distribution source prior at all the source locations. Since different speech sources can have different statistical properties and the mixture model such as SMM source prior can model different sources with different Student's t distribution in the mixture model while identical source prior model all the sources with the identical distribution and therefore there separation performance suffers as compared to the mixture model source priors.

6.6 Summary

This chapter presented the EM framework for the IVA method that uses the mixture of Student's t distribution as a source prior in order to better model the different statistical properties in different speech sources. The mixture of Student's t source prior made use of the heavy tails nature of the Student's t distribution to effectively model the high amplitude information in the speech signal. The complete EM framework was derived efficiently to estimate the model parameters for the IVA method. The separation performance for the proposed method was tested with image room impulse method and it confirms the advantage of using the proposed framework for the IVA method. Further experiments were conducted in real room environments with different reverberation times. All the experiments with real room recordings confirmed that the proposed EM framework for the IVA algorithm that make use of the SMM source prior improves the separation performance even in highly reverberant real room environments.

Chapter 7

CONCLUSIONS AND FUTURE WORK

This study focused upon enhancing the performance of the independent vector analysis algorithm for separating multiple speech sources from their reverberant mixtures in a real room environment. Humans are proficient at selectively focusing on sound of interest in the presence of multiple sound sources. In contrast, machines usually struggle to mimic this particular human ability. The performance of current source separation techniques is limited as well. Therefore the work in this thesis was aimed at improving the separation performance of the independent vector analysis algorithm in real room environments.

7.1 Conclusions

The contributions of this work satisfy the three research objectives specified in the introduction chapter. The first contribution is to improve the separation performance and the convergence speed of the IVA algorithm by exploiting a new multivariate Student's t source prior to preserve the inter-frequency dependency within the frequency domain signals. The second contribution is using the combined distribution model to improve the source prior of the IVA method and utilise the energy of the mixture signals to automatically adapt the mixing parameter of the combined source prior. The third contribution is deriving the expectation-maximization framework for the IVA algorithm which can explicitly adapt according to the measured speech signal and improve the separation performance of the IVA algorithm. The details of the contributions and background information are as follows:

Fundamental information and background for convolutive BSS is introduced in Chapter 2. This chapter also discussed the need to conduct the processing in the frequency domain, which initiates the bin-wise permutation problem that is an implicit problem of the frequency domain ICA algorithm. Moreover, Chapter 2 also examined the natural gradient IVA algorithm and the fast fixed point IVA algorithm. The complete derivation of both IVA algorithms was included and the choice of the source prior was discussed for the original IVA algorithm and the Fast IVA algorithm.

In Chapter 3, different experimental setups were discussed in the detail. The data sets for the generation of real room impulse responses were introduced. Moreover, different objective and subjective performance measures were introduced to evaluate the separation performance of different algorithms.

In Chapter 4, a new multivariate Student's t source prior was proposed for the original IVA and the FastIVA algorithm. The source prior for the IVA method is imperative to the performance of the algorithm as the non-linear score function is used to retain the inter-frequency dependency derived based on the PDF of the source. Speech signals are highly non stationary in nature and many useful samples in speech signals can be of high amplitude. The Student's t distribution was adopted to model the speech signals, since the Student's t distribution has heavier tails and it can better model the information in high amplitude data points. Therefore the Student's t source prior better models the dependency structure in the frequency domain speech signals and improves the separation performance and the convergence speed of the IVA and the FastIVA algorithm. The separation performance of the IVA and the FastIVA algorithm with the Student's t source prior was tested in both the simulated and the real reverberant room environments and the improved averaged separation performance of 0.90 dB was recorded. Also, the faster convergence speed was confirmed for the FastIVA algorithm when the results were compared with the original FastIVA algorithm.

Chapter 5 introduced a mixed multivariate Student's t and original multivariate super Gaussian source prior for the IVA algorithms. In the mixed multivariate source prior, the Student's t distribution was adopted to better model the high amplitude information in the speech signals and at the same time the original super Gaussian distribution was used to model the remaining information. In the mixed source prior equal weightage was assigned to both the distributions and experiments were performed with real room impulses; and the performance improvement was recorded when compared with the original IVA and the FastIVA method. The average improvement in the separation performance was approximately 1 dB. The separation performance of the mixed source prior was further enhanced by adopting the ratio of distributions in the mixed source prior according to the normalised energy of the measured mixtures in the frequency domain blocks, as different speech sources can have different statistical properties. The complete frequency bins were divided into smaller non overlapping blocks and then the normalised energy was calculated for each block as different frequency ranges can have different energy. The weightage of distributions in the mixed source prior was then adapted according to the energy of a particular block. The new energy driven mixed source prior for the IVA algorithm was evaluated in different reverberant environments and it further improved the separation performance by 1.2 dB when compared with the original IVA algorithm.

A new expectation maximization framework for the IVA algorithm was efficiently derived in Chapter 6. Instead of a conventional identical multivariate distribution, a new multivariate Student's t mixture model was adopted as a source prior for the IVA method. The SMM was able to better model the different speech sources as different sources can have distinctive statistical properties and the SMMs can adapt according to the statistical properties of different sources. Also, by using the SMM as the source prior for the IVA algorithm had the advantage of modelling the high amplitude data points more efficiently. In order to estimate the unmixing matrix, an efficient EM algorithm was implemented for the IVA algorithm and then the new framework was tested in different real reverberant room environments. The separation performance of the EM based IVA algorithm with SMM as the source prior was compared with the original IVA algorithm with different source prior and the proposed method improved the separation performance of algorithm for all the reverberant room environments.

All algorithms described in the main body of this thesis deal with real reverberant room environments. Therefore in this respect, this thesis can serve as a stepping stone for future researchers to expend on the ideas and improve the solution of the machine cocktail party problem.

7.2 Future Work

The techniques proposed in this thesis could be expanded in a number of potential ways and different directions can be explored.

In this thesis, the number of frequency bins are consider to be similar to the length of the room impulse response, unless otherwise stated (1024). This number for frequency bins is chosen for the IVA algorithm to cultivate a good SDR performance. In future work, it is possible to explore the methods to reduce the number of frequency bins for the IVA algorithm, whilst maintaining a reasonable SDR separation performance. As by reducing the number of frequency bins for the IVA algorithm, computational complexity can be cut down and convergence speed of the algorithm can be increased. Chapter 4 improves the convergence speed of the IVA algorithm, however convergence speed of the algorithm needs to increase significantly in order to implement the IVA algorithm in real time applications.

Since the separation performance of the IVA algorithm deteriorates with the increase in the reverberation time of any particular room environment. So one of the directions for further work is to investigate different dereverberation methods that can be used in conjunction with the IVA algorithm to alleviate the problem. One of the methods used for dereverberation is beamforming, which is widely used in speech processing. It can be used as a pre-processing step for the IVA method in order to dereverberate the speech signals. Some other methods used for dereverberation of speech signals include the linear prediction technique. Numerous other methods have been proposed as a potential solution for this problem [106–110]. Most of these methods are developed for the one source case, while in the cocktail party problem, the minimum number of sources is two, which makes the implementation of these methods difficult. So further studies can be devoted to achieve a combined model which can use the dereverberation methods as a pre-processing stage for the IVA algorithm and can potentially improve the separation performance of the IVA algorithm in a highly reverberant environment.

Finally, future work can be focused on investigating the speech signals and choosing the dependency structures that can further improve the modelling of the speech signals. As the IVA method relies heavily on choosing an appropriate source prior so future research can be conducted on further improving the dependency model which can potentially improve the separation performance of the IVA algorithm.

References

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

[2] E. Cherry and W. Taylor, “Some further experiments on the recognition of speech, with one and with two ears,” *Journal of the Acoustical Society of America*, vol. 26, pp. 554-559, 1954.

[3] J. H. McDermott, “The cocktail party problem,” *Current Biology*, vol. 19, R1024-R1027, 2009.

[4] S. Haykin and Z. Chen, “The cocktail party problem,” *Neural Computation*, vol. 17, pp. 1875-1902, 2005.

[5] 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.

[6] D. Wang and G. Brown, *Fundamentals of computational auditory scene analysis, in computational auditory scene analysis: Principles, algorithms and applications,* Hoboken, NJ: John Wiley and Sons, 144, 2006.

[7] M. I. Mandel, R. J. Weiss, and D. Ellis, “Model-based expectation maximization source separation and localization,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 382-394, 2010.

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

[9] J. F. Cardoso, “Blind signal separation: statistical principles,” *Proceedings of the IEEE*, vol. 86, pp. 2009-2025, 1998.

[10] C. Jutten and J. Herault, “Blind Seperation of sources, part I: An adaptive algorithm based on neuromimetic architecture,”*Signal Processing*, vol. 24, pp. 1-10, 1991.

[11] C. Jutten and P. Comon. *Handbook of Blind Source Separation: Independent Component Analysis and Applications.* Academic Press, 2010.

[12] A. Cichocki and S. Amari, *Adaptive Blind Signal and Image Processing*, John Wiley, 2002.

[13] A. Cichocki, R. Zdunek, A. H. Phan, and S. I. Amari, *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation.* Wiley, 2009.

[14] 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*, vol. 8, pp. 1-34, 2007.

[15] 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

[16] 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.

[17] N. Madhu and J. Wouters, "Localisation-based, situation-adaptive mask generation for source separation," *International Symposium on Communications, Control and Signal Processing (ISCCSP)*, pp 1-6, 2010.

[18] 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 Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 96-108, 2007.

[19] 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.

[20] B. Rivet, W. Wang, S. M. Naqvi, and J. A. Chambers, "Audiovisual speech source separation: An overview of key methodologies," *IEEE Signal processing magazine*, vol. 31, pp. 125-134, 2014.

[21] 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.

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

[23] I. Lee, G. J. Jang and T. W. Lee, "Indpendent 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.

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

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

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

[27] 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.

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

[29] T. Barker, and T. Virtanen, "Blind Separation of Audio Mixtures Through Nonnegative Tensor Factorization of Modulation Spectrograms," *IEEETransactions on Audio, Speech, and Language Processing*, vol. 50, pp. 2377-2389, 2016.

[30] L. De Lathauwer, B. De Moor, and J. Vandewalle, "Electrocardiogram extraction by blind source subspace separation," *IEEE Transactions on Biomedical Engineering,* vol. 47, pp. 567-572, 2000.

[31] T.-P. Jung, S. Makeig, C. Humphries, T.-W. Lee, M. J. Mckeown, V. Iragui, and T. J. Sejnowski, "Removing electroencephalographic artifacts by blind source separation," *Psychophysiology*, vol. 37, pp. 163-178, 2000.

[32] Z. Zhang, H. Li and D. Mandic, "Blind source separation and artefact cancellation for single channel bioelectrical signal," *IEEE International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pp. 177-182, 2016.

[33] K. Reindl, Y. Zheng, and W. Kellermann, “Speech enhancement for binaural hearing aids based on blind source separation,” *IEEE International Symposium on Communications, Control and Signal Processing (ISCCSP)*, pp. 1-6, 2010.

[34] T. Zhang, F. Mustiere and C. Micheyl, “Intelligent hearing aids: The next revolution” *International Conference of the Engineering in Medicine and Biology Society (EMBC)*, pp. 72-76, 2016.

[35] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, “A realtime blind source separation scheme and its application to reverberant and noisy acoustic environments,” *Signal Processing*, vol. 86, pp. 1260-1277, 2006.

[36] S. F. Wu, *An autonomous surveillance system for blind source localization and separation*, http://www.freepatentsonline.com/y2016/0061929.html, Accessed: 10-11-2016, 2016.

[37] A. Mansour, N. Benchekroun, and C. Gervaise, “Blind separation of underwater acoustic signals,” *International Communication Association (ICA)*, pp. 181-188, 2006.

[38] B. Marhaba, M. Zribi, W. Khodar, “Image restoration using a combination of blind and non-blind deconvolution technqiues,” *International Journal of Engineering Research and Science*, vol. 2, pp. 225-239, 2016.

[39] P. Ravichandran, “Literature Survey on Image Deblurring Techniques,” *International Journal of Science and Research*, vol. 5, pp. 1670-1674, 2016.

[40] Apple Corporation Inc., *About Siri*, https://support.apple.com/en-gb/HT204389, Accessed: 11-11-2016, 2016.

[41] M. G. Jafari, “Novel sequential algorithms for blind source separation of instantaneous mixtures,” *PhD thesis, Kings College London*, 2002.

[42] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, 114: 2236-2252, 2003.

[43] T. Adali, M. Anderson, and F. Geng-Shen, "Diversity in independent component and vector analyses:Identiability, algorithms, and applications in medical imaging," *IEEE Signal Processing Magazine*, vol. 31, pp. 18-33, 2014.

[44] 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.

[45] T. Barker, and T. Virtanen, "Blind Separation of Audio Mixtures Through Nonnegative Tensor Factorization of Modulation Spectrograms," *IEEETransactions on Audio, Speech, and Language Processing*, vol. 50, pp. 2377-2389, 2016.

[46] M. Davies, "Audio source separation," *Institute of Mathematics and its applications conference series*, vol. 71, pp. 57-68, 2002.

[47] P. Smaragdis, "Efficient blind separation of convolved sound mixtures," *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 1997.

[48] W. Wang, S. Sanei and J. A. Chambers, "Penalty function based joint diagonalization approach for convolutive blind separation of nonstationary sources," *IEEE Transcations on Signal Processing*, vol. 53, pp. 1654-1669, May 2005.

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

[50] A. Alinaghi, P. Jackson, Q. Liu and W. Wang, " Joint mixing vector and binaural model based stereo source separation," *IEEE/ACM Transcations on Audio, Speech and Language Processing*, vol. 22, pp. 1434-1448, 2014.

[51] I. Jolliffe, *Principal component analysis*, Wiley Online Library, 2002.

[52] E. Bingham and A. Hyvarinen, "A fast fixed point algorithm for independent component analysis of complex valued signals," *International Journal Neural Networks*, vol. 10, pp. 1-8, 2000.

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

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

[55] A. Hyvrinen, "Fast and robust fixed-point algorithms for independent component analysis," IEEE Transactions on Neural Networks, vol. 10, pp. 626-634, 1999.

[56] N. Mitianoudis and M. E. Davies. "Audio source separation of convolutive mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 489-497, 2003.

[57] A. Hyvärinen, "Independent component analysis: recent advances," *Philosophical transactions Series A, Mathematical, physical, and engineering sciences*, 371, 2013.

[58] 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.

[59] J. Xi and J. Reilly, “Blind separation and restoration of signals mixed in convolutive environment,” *ICASSP*, pp. 1327-1330, Munich, Germany, 1997.

[60] S. F. Minhas and P. Gaydecki, “A hybrid algorithm for blind source separation of a convolutive mixture of three speech sources,” *EURASIP Journal on Advances in Signal Processing*, vol. 2014, pp. 92, 2014.

[61] S. Araki, H. Sawada, R. Mukai, and S. Makino, “Blind sparse source separation with spatially smoothed time frequency masking,” *Proc. IWAENC*, Paris, 2006.

[62] M. Z. Ikram and D. R. Morgan, “Exploring permutation inconsistency in blind separation of speech signals in a reverberant environment,” *Proc. IEEE International Conference on Acoustics Speech and Signal Processing*, 2000.

[63] W. Wang, J. A. Chambers and S. Sanei, ” A novel hybrid approach to the permutation problem of frequency domain blind source separation,” *Proc. International Conference on Independent Component Analysis and Blind Source Separation*, pp. 530-537, 2004.

[64] M. S. Pedersen, D. Liang, J. Larsen, and U. Kjems, “Two-microphone separation of speech mixtures,” *IEEE Transactions on Neural Networks*, vol. 19, pp. 475-492, 2008.

[65] S. I. Amari, “Natural gradient works effciently in learning,” *Neural computation*, vol. 10, pp. 251-276, 1998.

[66] N. Ono, “Blind source separation on iPhone in real environment,” *European Signal Processing Conference,(EUSIPCO)*, 2013.

[67] Y. Liang in *Enhanced Independent Vector Analysis for Audio Separation in a Room Environment* , (PhD thesis, Loughborough University, UK), 2013.

[68] S. Gazor and Z. Wei, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, pp. 204-207, 2003.

[69] A. Aroudi, H. Veisi, H. Sameti, Hossein, Z. Mafakheri, "Speech signal modeling using multivariate distributions," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, pp. 35, 2015.

[70] D. Peel and G. J. McLachlan, "Robust mixture modelling using the t distribution", *Satistics and Computing*, vol 10, pp. 339-348, 2000.

[71] W. G. Gilchrist, *Statistical Modelling with Quantile Functions.* Chapman and Hall, 2000.

[72] I. Cohen, "Speech enhancement using super-Gaussian speech models and noncausal a priori SNR estimation," *Speech Communication*, vol. 47, pp. 336-350, 2005.

[73] Z. Y. Zohny, S. M. Naqvi, and J. A. Chambers, "Enhancing MESSL algorithm with robust clustering based on student's t-distribution" *Electronic Letters*, vol. 50, pp. 552-554, 2014.

[74] 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.

[75] 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.

[76] W. Rafique, S. M. Naqvi, P. J. B. Jackson and J. A. Chambers, "IVA algorithms using a multivariate Student's t source prior for speech source separation in real room environments," *IEEE ICASSP*, South Brisbane, QLD, pp. 474-478, 2015.

[77] 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.

[78] Y. Hu and P.C. Loizou,"Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol.16, pp. 229-238, 2008

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

[80] K. Matsuoka and S. Nakashima, "Minimal distrotion principle for blind source separation," *SICE conference*, vol. 4 pp.2138-2143, 2001

[81] C. Brown, Matlab central, t60.m, http://de.mathworks.com/matlabcentral/fileexchange/1212-t60-m, Accessed: 11-11-2016, 2002.

[82] E. ISO, "3382-2: 2008," Acoustics. Measurements of room acoustics parameters, Part 2.

[83] H. Kuttruff, "Room acoustics," *Spon Press, Oxon*, 2009.

[84] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *Journal of the Acoustical Society of America*, vol. 65, pp. 943-950, 1979.

[85] A. Lundeby, T. E. Vigran, H. Bietz, and M. Vorlander, "Uncertainties of measurements in room acoustics," *Acta Acustica united with Acustica,* vol. 81, pp. 344-355, 1995.

[86] J. Mourjopoulos. "On the variation and invertibility of room impulse response functions", *Journal of Sound and Vibration*, vol. 102, pp. 217-228, 1985.

[87] 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.

[88] C. Hummersone, "A psychopsychoacoustic engineering approach to machine sound source separation in reverberant environments," *Ph.D. dissertation*, University of Surrey, 2011.

[89] S. M. R. Naqvi in *Multimodal methods for blind source separation of audio sources*, (PhD thesis, Loughborough University), UK, 2009.

[90] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[91] W. Rafique, S. M. Naqvi and J. A. Chambers, "Mixed source prior for the fast independent vector analysis algorithm," *IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, Rio de Janerio, pp. 1-5, 2016.

[92] I. Lee and G. J. Jang, "Independent vector analysis based on overlapped cliques of variable width for frequency-domain blind signal separation,"*EURASIP Journal on Advances in Signal Processing*, vol. 2012, pp. 113, 2012.

[93] J. Harris, B. Rivet, S.M. Naqvi, J.A. Chambers and C. Jutten, "Real-time independent vector analysis with Student's t source prior for convolutive speech mixtures," *IEEE International Conference on Acoustics, Speech and Signal Processing,* pp.1856-1860, 2015.

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

[95] Y. Liang, J. Harris, 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," *Signal Processing*, vol. 105, pp. 175-184, 2014.

[96] 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.

[97] 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.

[98] Y. Liang, G. Chen, S.M.R. Naqvi and J.A Chambers, "Independent vector analysis with multivariate Student's t distribution source prior for speech separation," *Electronics Letters*, vol. 49, pp. 1035-1036, 2013

[99] W. Rafique, S.M. Naqvi and J. A. Chambers, "Speech source separation using the IVA algorithm with multivariate mixed super Gaussian Student's t source prior in real room environment,"*IET Conference Proceedings*, 2015.

[100] J. Hao, I. Lee, T. W. Lee and T. J. Sejnowski, "Independent Vector Analysis for Source Separation Using a Mixture of Gaussians Prior," *Neural computation.*, vol. 22, pp. 1646-1673, 2010

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

[102] S. S. Dragmor and C. J. Goh, "Some counterpart inequalities in for a functional associated with Jensen's inequality," *Jounal of Inequalities and Applications*, vol. 1, pp. 311-325, 1997.

[103] O. A. Bauchau, L. Trainelli, "The vectorial parametrization of rotation," *Journal of Nonlinear Dynamics*, vol. 32, pp. 7192, 2003.

[104] W. Rafique, S. Erateb, S. M. Naqvi, S. S. Dlay, J. A. Chambers, " Independent vector analysis for source separation using an energy driven mixed Student's t and super Gaussian source prior" *Proc. of Eusipco*, 2016.

[105] H. Wang and A. Zhang, "Underwater acoustic signals blind separation based on time-frequency analysis,"*Int. J. Computational Intelligence Research.*, vol 2. pp. 91-94, 2006.

[106] T. Nakatani, T. Yoshioka, K. Kinoshita, and M. Miyoshi, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation" *Proc. ICASSP*, Las Vegas, U.S.A, 2008.

[107] 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.

[108] P. A. Naylor and N. D. G. (Eds.), *Speech Dereverberation, Signals and Communication Technology.* Springer, 1st Edition, 2010.

[109] 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.

[110] N. Mohammadiha and S. Doclo, "Speech Dereverberation Using Non-Negative Convolutive Transfer Function and Spectro-Temporal Modeling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 276-289, 2016.