# Enhanced IVA for Audio Separation in Highly Reverberant Environments

by

Suleiman Erateb

A thesis submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy (PhD)

June 2018

Signal Processing and Networks Research Group,
Wolfson School of Mechanical, Electrical and Manufacturing Engineering,
Loughborough University, Loughborough,
Leicestershire, UK, LE11 3TU

## CERTIFICATE OF ORIGINALITY

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

................................ (Signed)

................................ (candidate)

*To my late grandmother, my parents and my family.*

# Abstract

Blind Audio Source Separation (BASS), inspired by the “cocktail-party problem”, has been a leading research application for blind source separation (BSS). This thesis concerns the enhancement of frequency domain convolutive blind source separation (FDCBSS) techniques for audio separation in highly reverberant room environments.

Independent component analysis (ICA) is a higher order statistics (HOS) approach commonly used in the BSS framework. When applied to audio FDCBSS, ICA based methods suffer from the permutation problem across the frequency bins of each source. Independent vector analysis (IVA) is an FD-BSS algorithm that theoretically solves the permutation problem by using a multivariate source prior, where the sources are considered to be random vectors. The algorithm allows independence between multivariate source signals, and retains dependency between the source signals within each source vector. The source prior adopted to model the nonlinear dependency structure within the source vectors is crucial to the separation performance of the IVA algorithm. The focus of this thesis is on improving the separation performance of the IVA algorithm in the application of BASS.

An alternative multivariate Student’s t distribution is proposed as the source prior for the batch IVA algorithm. A Student’s t probability density function can better model certain frequency domain speech signals due to its tail dependency property. Then, the nonlinear score function, for the IVA, is derived from the proposed source prior.

A novel energy driven mixed super Gaussian and Student's t source prior is proposed for the IVA and FastIVA algorithms. The Student's t distribution, in the mixed source prior, can model the high amplitude data points whereas the super Gaussian distribution can model the lower amplitude information in the speech signals. The ratio of both distributions can be adjusted according to the energy of the observed mixtures to adapt for different types of speech signals.

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

An adaptive learning scheme is developed to improve the performance of the online IVA algorithm. The scheme adjusts the learning rate as a function of proximity to the target solutions. The scheme is also accompanied with a novel switched source prior technique taking the best performance properties of the super Gaussian source prior and the generalized Gaussian source prior as the algorithm converges.

The methods and techniques, proposed in this thesis, are evaluated with real speech source signals in different simulated and real reverberant acoustic environments. A variety of measures are used within the evaluation criteria of the various algorithms. The experimental results demonstrate improved performance of the proposed methods and their robustness in a wide range of situations.

# Acknowledgements

I would like to express my very great appreciation to my supervisor Prof. Jonathon Chambers. Without his guidance and encouragement, this thesis would have never been accomplished. He has taught me more than I could ever give him credit for here. He has provided me extensive personal and professional guidance and taught me a great deal about both scientific research and life in general. He has shown me, by his example, what a good scientist and person should be. I have benefited tremendously from his wide knowledge and great enthusiasm. I have had his full support, even after he had left Loughborough University. He has always been available to meet me in person, talk to me on the phone or via video conferencing and he replied to my emails promptly any time of the day, any day of the week and anywhere in the world. His advice and feedback on my work and thesis have been invaluable.

I would also like to extend my deepest gratitude to Prof. Sangarapillai Lambotharan, my second supervisor and Head of Signal Processing and Networks Research Group. He supported me greatly and was always willing to help me. He provided me amazing opportunities to attend international conferences and meet some wonderful researchers in the area of signal processing. I am grateful to all of those with whom I have had the pleasure to work during my study within the research group. I would especially like to thank Dr. Jack Harris for his assistance.

I would also like to thank the staff at Wolfson School of Mechanical, Electrical and Manufacturing Engineering, the research office and around the University. They have all facilitated my study and contributed to the wonderful experience at Loughborough. I wish to acknowledge the help provided by my managers and colleagues within the Engineering team at CU Coventry, in particular during the writing up.

Nobody has been more important to me in the pursuit of this PhD than the members of my family. Most importantly, I would like to thank my parents, whose love and guidance are with me in whatever I pursue. I wish to thank all members of my family for their support, patience and endless inspiration throughout the PhD journey.

# Statement of Originality

The contributions of this thesis are mainly related to the enhancement of the independent vector analysis (IVA) algorithm for speech separation in real room reverberant environments. The novelty of the contributions is supported by the following international journal and conference papers.

In Chapter 4, a new multivariate Student's t source prior is proposed for the IVA method to model the dependency structure of frequency domain speech signals. The heavy tails of the proposed multivariate source prior can be advantageous in modelling the spectrum of frequency domain non-stationary speech signals. Therefore, it improves the separation performance and convergence speed of the IVA method in the real room environments. The work was published in:

**1.** S. Erateb, W. Rafique, M. Naqvi and J.A. Chambers, "Evaluation of Source Separation Algorithms, including the IVA algorithm with various source priors, using Binaural Room Impulse Responses," *International Conference on Mathematics in Signal Processing, 10th IMA*, Birmingham, UK, 2014.

In Chapter 5, a new multivariate source prior for the IVA algorithm is introduced as a mixture of two distributions to better model speech signals. The source prior is constructed by mixing the original multivariate super Gaussian distribution and the multivariate Student's t distribution with a certain ratio. The Student's t distribution can better model the high amplitude information in the speech signals due to its heavy tailed nature and the super Gaussian distribution is used to model the rest of the information. The mixed source prior is empowered with an energy driven scheme that adjusts the weight of each distribution according to the energy of the observed mixtures so it can adapt to different types of speech mixture signals. The model is further strengthened by an overlapped clique based dependency model for frequency spectrum structure of the sources. The IVA and the FastIVA methods are evaluated using the new techniques in different real room environments. The findings were published in:

**2.** W. Rafique, S. Erateb, M. Naqvi, 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," *Signal Processing Conference (EUSIPCO), 24th European*, pp. 858 - 862, Budapest, Hungary, 2016.

In Chapter 6, a new robust adaptive learning based scheme is proposed to improve the separation performance of the online IVA algorithm. The scheme controls the learning rate by exploiting a gear-shifting technique to start with high value learning rate and reduces it as the algorithm converges. In addition, a generalized Gaussian distribution is introduced as a multivariate source prior for the online IVA algorithm. The scheme was tested using the super Gaussian source and the new source prior and compared with the original online IVA algorithm in real room environments and real speech signals. The results demonstrated improved convergence speed and steady state performance. The results led to a new switched source prior that switches between the original super Gaussian and the generalized Gaussian source priors controlled by the adaptive learning scheme to acquire the best properties of each distribution. The results are published in:

**3.** S. Erateb, W. Rafique, J. Harris, M. Naqvi and J.A. Chambers, "Enhanced Online Independent Vector Analysis Algorithm for Speech Separation using Adaptive Learning," *International Conference on Mathematics in Signal Processing, 11th IMA*, Birmingham, UK, 2016.

**4.** S. Erateb, M. Naqvi and J.A. Chambers, "Online IVA with Adaptive Learning for Speech Separation using Various Source Priors," *Sensor Signal Processing for Defence Conference (SSPD)*, pp. 74 - 78, London, UK, 2017.

**5.** S. Erateb, M. Naqvi and J.A. Chambers, "Enhanced Online IVA with Adaptive Learning and Switched Source Prior for Speech Separation," *Article in preparation for IET Proc. Signal Processing*, 2018.

# Contents

# List of Acronyms

| | |
|---|---|
| **ASA** | Auditory Scene Analysis |
| **BASS** | Blind Audio Source Separation |
| **BRIR** | Binaural Room Impulse Response |
| **BSS** | Blind Source Separation |
| **CASA** | Computational Auditory Scene Analysis |
| **CBSS** | Convolutive Blind Source Separation |
| **CPP** | Cocktail Party Problem |
| **DFT** | Discrete Fourier Transform |
| **DNN** | Deep Neural Networks |
| **DOA** | Directions of Arrival |
| **DRR** | Direct-to-Reverberant Ratio |
| **EVD** | Eigen-Value Decomposition |
| **FastICA** | Fast Fixed Point Independent Component Analysis |
| **FastIVA** | Fast Fixed Point Independent Vector Analysis |
| **FD-BSS** | Frequency Domain Blind Source Separation |
| **FD-CBSS** | Frequency Domain Convolutive Blind Source Separation |
| **FFT** | Fast Fourier Transform |
| **GD** | Gradient Descent |
| **HOS** | Higher Order Statistics |
| **ICA** | Independent Component Analysis |
| **IR** | Impulse Response |
| **ISM** | Image Source Method |

| | |
|---|---|
| **ITDG** | Initial Time Delay Gap |
| **IVA** | Independent Vector Analysis |
| **KEMAR** | Knowles Electronics Manikin for Acoustic Research |
| **KL** | Kullback-Leibler |
| **LS** | Least Squares |
| **MOS** | Mean Opinion Score |
| **MSS** | Mean-Squared Sum |
| **NG** | Natural Gradient |
| **NG-ICA** | Natural Gradient ICA |
| **NG-IVA** | Natural Gradient IVA |
| **NMF** | Non-negative Matrix Factorisation |
| **PCA** | Principal Component Analysis |
| **PDF** | Probability Density Function |
| **PESQ** | Perceptual Evaluation of Speech Quality |
| **PI** | Permutation Index |
| **PM** | Permutation Measurement |
| **RIR** | Room Impulse Response |
| **RMS** | Root Mean Square |
| **RT** | Reverberation Time |
| $\mathbf{RT_{60}}$ | Reverberation Time (60dB) |
| **SDR** | Signal-to-Distortion Ratio |
| **SIR** | Signal-to-Interference Ratio |
| **SOS** | Second Order Statistics |
| **SSL** | Spherically Symmetric Laplace |
| **STFT** | Short-Time Fourier Transform |
| **TIMIT** | Texas Instruments and Massachusetts Institute of Technology |

# List of Symbols

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

| | |
|---|---|
| $\lvert . \rvert$ | Absolute value |
| $\lvert\lvert . \rvert\rvert_2$ | Euclidean norm |
| $\lvert\lvert . \rvert\rvert_F$ | Frobenius norm |
| $(.)^T$ | Transpose operator |
| $(.)^H$ | Hermitian transpose operator |
| $(.)^{-1}$ | Inverse operator |
| $\otimes$ | Convolution operator |
| $(.)^*$ | Complex conjugate operator |
| $det(.)$ | Matrix determinant operator |
| $E(.)$ | Statistical expectation operator |
| $\Pi(.)$ | Matrix projection |
| $F(.)$ | 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 |
| $\mathbf{I}$ | Identity matrix |

| | |
|---|---|
| $T$ | STFT length |
| $k$ | 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 |
| $\eta$ | Learning rate |
| $\mu$ | Mean value |
| $\boldsymbol{\mu}$ | Mean vector |
| $\sigma$ | Standard deviation |
| $\beta$ | Smoothing factor |
| $\lambda_d$ | Weighting factor |
| $\mathbf{\Sigma}$ | Covariance matrix |
| $\mathbf{\Lambda}$ | Diagonal matrix |
| $\Re$ | Scored correlation matrix |
| $\mathbf{G}$ | Overall system matrix |
| $\mathbf{H}$ | Mixing matrix |
| $\mathbf{W}$ | Estimated unmixing matrix |
| $\mathbf{P}$ | Permutation matrix |
| $\mathbf{Q}$ | Whitening matrix |
| $\mathbf{U}$ | Rotation matrix |
| $n$ | Time block index of STFT |
| $\nu$ | Degrees of freedom parameter |

# List of Figures

# List of Tables

Chapter 1

# INTRODUCTION

## 1.1 Introduction

In recent years, blind source separation (BSS) has attracted much research attention from the signal processing community [1]. BSS is a statistical signal processing approach that aims to separate individual sources from measurements, containing mixtures of the sources, observed at multiple sensors [2]. The estimation is performed blindly, i.e., without possessing information about the sources and the mixing process. The sources are commonly recovered by exploiting the assumption of mutual independence between the sources [3]. BSS can be used to recover all sources from the recorded mixtures, or to segregate a particular source from the mixtures. It may also be useful, in some situations, to reveal the mixing process itself to identify the mixing system.

BSS has been proposed for various fields in recent years [4]. The technique is applicable to a wide variety of signal processing applications including communication systems, biomedical signal processing, image restoration, radar antenna systems, and image and acoustic signal processing systems [5, 6]. A well-recognised BSS application is the separation of audio sources that have been mixed and then captured by multiple microphones in a real room environment, known as the cocktail party problem (CPP).

**Figure 1.1.** The cocktail party environment (Image from *Telegraph.co.uk*).

## 1.2 Cocktail Party Problem

The cocktail party effect is the ability of humans to focus their auditory attention on a particular stimulus while filtering out other stimuli in a complex auditory setting where a number of people are simultaneously participating in a conversation as illustrated in Figure 1.1. In such a setting, competing speech sounds or a variety of noises that are often assumed to be independent of each other may produce hearing interferences. However, this effect allows the majority of people to focus on a single voice reducing interferences [7].

The human auditory system has a great ability to distinguish between sounds from different sources in a cocktail party environment. However, for a machine, it is a much more challenging task to accomplish. This is what scientists term as the cocktail party problem (CPP). The term was first introduced by Colin Cherry [8] and further explored in [9]. The cocktail party problem refers to a machine's task of re-

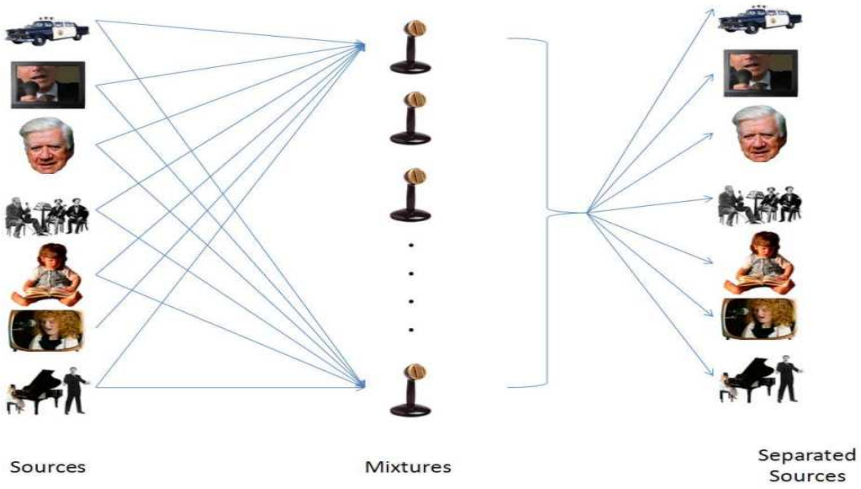


**Figure 1.2.** The cocktail party problem (Image from *onionesquereality.wordpress.com*).

covering speech in a room of simultaneous independent speakers. This may require the machine to imitate the complex cognitive processes of humans to achieve this goal.

The increase in computing power has motivated researchers to attempt to develop solutions to the cocktail party problem using microphone sensors. The solution for the cocktail party problem is to design a method to separate speech sources from their mixture or to focus on a desired speech source while suppressing all the other competing speech sources including noise. The problem is illustrated in Figure 1.2.

Using advanced computing and signal-processing technologies, the long term research aim of scientists working on the cocktail party problem is to build an intelligent machine which can mimic the ability of the human auditory system to solve the cocktail party problem. However, this aim has not been met because a complete understanding of the cocktail party phenomenon is still missing, and the human auditory perception capability is not fully understood [7].

To imitate the human performance with a machine, computational neuroscientists, computer scientists, and engineers have attempted to view and simplify this complex perceptual task as a learning problem, for which a computational solution is sought. Following the early pioneering works of several scholars [10–14] numerous efforts have been dedicated to address the CPP problem in diverse fields: physiology, neurobiology, psychophysiology, cognitive psychology, biophysics, computer science, and engineering.

In the computer science community, the topic is known as computational auditory scene analysis (CASA). CASA is driven by understanding and imitating the capabilities of human auditory scene analysis [11, 15]. In the signal processing community, it is known as blind source separation (BSS). In addition, approaches that combine CASA and BSS have also been proposed [16, 17]. Several methods to address BSS problems have emerged over the years such as non-negative matrix factorisation (NMF) [18] and deep learning for neural networks (DNN) [19]. This thesis focuses on statistical signal processing based BSS approaches exploiting, mainly, higher order statistics (HOS), in particular independent vector analysis (IVA) [20, 21].

## 1.3 Blind Source Separation

Blind source separation (BSS) is a technique for estimating individual source components from their observed mixtures. The observed signals are obtained at a set of sensors, each receiving a different linear combination of the source signals. The term blind refers to the fact that only the mixtures are available and both the sources and the mixing process are unknown [13]. Separation may be achieved in different ways

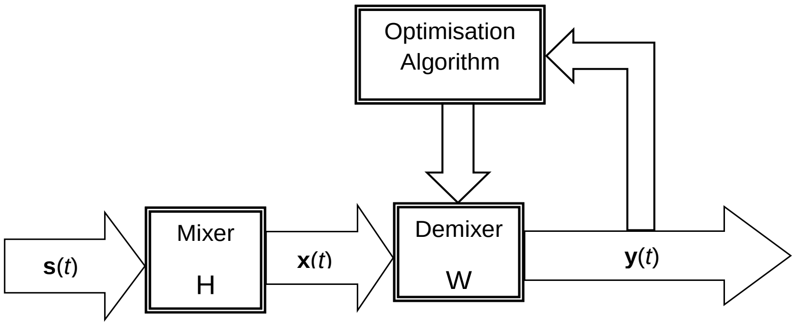


**Figure 1.3.** Block diagram of the mixing and BSS processes

according to the amount of prior information available [14].

The BSS problem can be stated as the estimation of $N$ source signals from $M$ observed mixture signals that are unknown function of the sources. The basic BSS model is shown in Figure 1.3. In the figure, the source data $\mathbf{s}(t)$ are mixed by a mixing matrix $\mathbf{H}$ to produce the sensor data $\mathbf{x}(t)$, where $t$ is the discrete time index. Optimisation algorithms act on $\mathbf{x}(t)$ to produce a separating matrix $\mathbf{W}$ that has the capability to extract the original sources $\mathbf{y}(t)$, an estimation of $\mathbf{s}(t)$, from the mixed sources [14]. The vectors of the model can be expressed as the linear transformations by:

$$\mathbf{x}(t) = \mathbf{H}\mathbf{s}(t) \tag{1.3.1}$$

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) \tag{1.3.2}$$

where [1]$\mathbf{x}$ is the observation vector $(M \times 1)$ components,
$\mathbf{s}$ is the source vector $(N \times 1)$ components,
$\mathbf{y}$ is the estimated (output) vector $(N \times 1)$ components,
$\mathbf{H}$ is the mixing matrix $(M \times N)$, and
$\mathbf{W}$ is the unmixing matrix $(N \times M)$.

[1] The time index (t) is dropped for notational convenience.

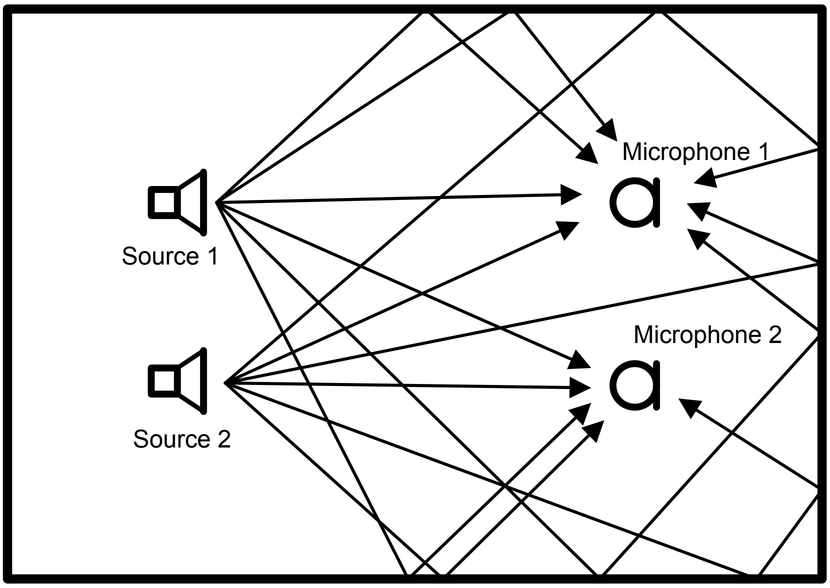


**Figure 1.4.** A schematic diagram of convolutive mixing environment between two sources and two microphones.

When $M = N$ it is an exactly determined system;

$M > N$ an over determined system; and

$M < N$ an under determined system.

If the source signals can travel directly without delay or filtering, they will arrive simultaneously[2] at the sensors. This leads to the basic instantaneous mixing model, which simplifies the solution of the problem. However, when BSS is applied to solve the cocktail party problem, the mixtures of audio sources, in a real reverberant environment, are convolutive rather than instantaneous due to propagation time delays and sound reverberation in the room, as illustrated in Figure 1.4. Thus the mixture can be expressed as:

[2] In practice, any real signal, such as speech, will take a finite time to travel from the source to the sensor array.

$$\mathbf{x}(t) = \sum_{\tau} \mathbf{H}(\tau)\mathbf{s}(t-\tau) \tag{1.3.3}$$

where $\tau$ is time delay and $\mathbf{H}(t)$ is the room impulse response (transfer function). Time domain methods have been proposed to separate convolutive mixture. However, in a real room environment, the length of the room impulse response is typically on the order of thousands of samples even if sampled with the 8 kHz sampling rate (Nyquist rate) for speech bandwidth. This makes solving the convolutive BSS (CBSS) problem computationally expensive in the time domain [1]. In order to improve the computational efficiency of the CBSS algorithms, frequency domain BSS (FD-BSS) methods have been proposed to tackle the problem. The convolution operation in the time domain can be approximated by multiplication in the frequency domain, which reduces the computational cost significantly [1]. Ideally, in FD-BSS methods, an instantaneous mixture is obtained at each frequency component. In each frequency bin, $k$, the mixing and separation can be denoted as:

$$\mathbf{x}^{(k)} = \mathbf{H}^{(k)}\mathbf{s}^{(k)} \tag{1.3.4}$$

$$\mathbf{y}^{(k)} = \mathbf{W}^{(k)}\mathbf{x}^{(k)} \tag{1.3.5}$$

BSS algorithms are, generally, designed to exploit the statistical independence of different sources in an acoustic environment. These algorithms attempt to maximise the independence between the estimated output signals. Most FD-BSS algorithms are based on extracting second order statistics (SOS) or higher order statistics (HOS) from the recorded data. BSS algorithms exploiting source independence under the instantaneous mixing model, lead to solutions for independent component analysis (ICA) [22]. For convolutive mixtures, FD-BSS can employ the ICA based techniques to separate the instantaneous mixtures in each frequency bin independently [5]. However, the component-wise separation approach exhibits some limitations [23–26]. The main ambiguity is the permutation problem, where the source permutation may appear independently in each frequency bin. To mitigate the permutation problem, extra measures have to be taken after the separation process to repair this internal permutation ambiguity. The extra processing stage, generally, makes the approach computationally expensive and may not lead to successful solutions.

Independent vector analysis (IVA) proposed by Kim et al. [20,21] is an approach that has proved successful in solving the permutation problem of FD-BSS for speech separation. It is based on an improved model of the ICA method exploiting higher order frequency dependencies to capture inherent interfrequency dependencies of the speech signals. The proposed method introduces the concept of multivariate components by extending the ICA formulation of univariate source signals to multivariate source signals. The algorithm allows independence between multivariate source signals and retains dependency between the source signals within each source vector. Hence, the IVA approach proposed a new cost function that measures the independence among multivariate signals with multivariate probability density functions (PDFs), instead of the univariate PDF adopted by the ICA method. The multivariate score function can be obtained from the multivariate source prior. The IVA mixing and separating model is illustrated in Figure 1.5, where both sources and observations are multivariate [20].

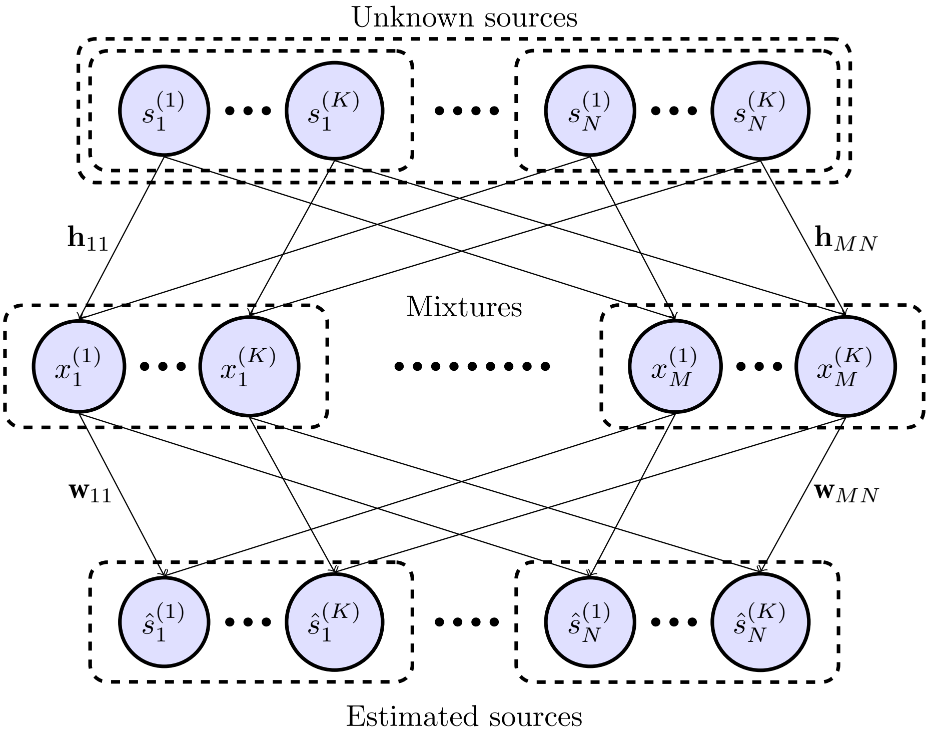


**Figure 1.5.** Mixing and separating model for the frequency domain independent vector analysis (FD-IVA).

The performance of the IVA algorithms relies on the statistical model employed as a source prior. The original IVA method employs a multivariate super-Gaussian (Laplacian) distribution as source prior. The performance of the IVA method can be potentially improved by exploiting alternative models and techniques.

The batch and online modes of the IVA algorithm are considered in this thesis. The batch algorithm requires all (or sufficient) signal data to be available before processing is performed. While the online algorithm is performed iteratively as signal data arrives.

## 1.4 Applications of Blind Audio Source Separation (BASS)

BSS has many potential audio applications. Given the observations, in some applications only one source signal is of interest and other applications require recovering all the source signals. The application of BSS for the separation of simultaneous speech sources in reverberating environment, such as in a room, is the focus of this thesis.

Speech enhancement by removing noise or other unwanted signal components is a major application area of BASS [27]. Enhancement of voice quality in mobile phones is one important application, particularly in noisy surroundings [28]. Voice dialling or speech recognition in general in a cocktail party environment is another application [29, 30]. Hearing aids are also another lucrative application for BSS speech enhancement [31, 32]. Other applications whereby the interest might be in picking up one target signal include spying, intelligence or forensic applications [33, 34].

BASS is useful in teleconferencing setup and speakerphones, where it is desirable to acquire speech signal free from reverberation, noise, acoustical echoes and mixed other speakers [35]. BASS techniques can also be applied in the detection and separation of acoustic signals in underwater systems. The application is utilised in understanding the underwater environment, ship tracking and detecting any underwater substance leakages [36, 37]. Another applications include high quality separation of musical sources [38] and source localization for auditory scene analysis.

## 1.5 Aim and Objectives

The main aim of the thesis is to analyse, develop and evaluate novel techniques to enhance the separation performance of speech signals acquired in reverberant environments through improved statistical modelling of the source dependencies. The focus is on enhancing the separation performance of the independent vector analysis (IVA) technique. The particular objectives of this thesis are:

- Objective 1: to examine alternative statistical dependency models for the IVA algorithm to improve the convergence and separation performance of the algorithm.

In Chapter 4, the Student's t distribution is adopted as a source prior for the batch IVA. The PDF can better model certain speech signals due to its tail dependency property, thereby improving the separation performance of the algorithm achieving a faster convergence speed in lower number of iteration. In Chapter 6, a generalized Gaussian distribution is adopted as a source prior for the online IVA algorithm. It improves the convergence time of the algorithm.

- Objective 2: to exploit the statistical property of the speech mixture signals to produce a combined distribution source prior that adapts to different types of speech signals and hence achieves improved separation performance.

In Chapter 5, a mixture source prior of the original super Gaussian distribution and the Student's t distribution is proposed for the batch IVA algorithm and the fast version of the algorithm. The weight of each distribution in the mixed source prior is adapted automatically according to the energy of the observed mixture signals.

- Objective 3: to investigate the frequency spectrum dependency structure within each source to avoid the permutation problem and enhance the separation performance of the IVA algorithm.

In Chapter 5, a clique based overlapped chain type dependency model is used to model the dependency structure within the frequency bins of each source to achieve a robust and improved separation performance for the IVA algorithms.

- Objective 4: to develop new techniques to enhance the convergence and separation performance of the online IVA.

Chapter 6 introduces a new adaptive learning scheme as a function of proximity to the target solution to improve the performance of the online IVA in terms of convergence time and steady state separation value and accuracy. In addition, the scheme is enhanced by a switched source prior technique between the super Gaussian and generalized Gaussian distributions gaining the advantages of each distribution at different stages of the learning algorithm.

- Objective 5: to apply and evaluate the different proposed techniques and methods for speech separation in a variety of real room environments and settings using robust criteria.

In all contribution Chapters 4, 5 and 6, all the different forms of the IVA algorithm are evaluated using different reverberant real room settings and room impulse responses. The evaluation criteria, including a package of performance measures is devised in Chapter 3.

## 1.6 Thesis Outline

The thesis is organised as follows:

- Chapter 2 introduces the background theory on BSS in the time and frequency domains related to the material of the thesis. The

BSS mixing models and the major techniques for solving the BSS problem in the frequency domain are discussed along with their ambiguities. The Parra-Spence frequency domain BSS algorithm based on SOS is reviewed. An overview of independent component analysis (ICA) is provided with its limitations in FD-BASS applications. Independent vector analysis (IVA) and how it addresses the permutation problem, inherent to ICA, is reviewed in detail along with its fast version (FastIVA).

- Chapter 3 outlines the datasets, techniques and experimental setups required for the implementation and evaluation of the separation of convolutive speech mixtures. The speech sources library is described. The different room models, deployed, along with the room settings and various parameters involved in the mixing process of speech sources are discussed. Furthermore, the separation performance criteria are detailed including the performance measures adopted for the purpose.

- Chapter 4, firstly, provides a comparison between separation performance of the ICA and IVA algorithms to demonstrate the effect of the permutation problem in CBSS and how it is addressed by the IVA method. Then, the multivariate Student's t distribution is adopted as a source prior for the IVA method. The source prior is used to derive the nonlinear score function for the algorithms. The separation performance of the IVA algorithm with the new source prior is compared with the original super Gaussian source prior in real room environments.

- In Chapter 5, a new multivariate source prior for the IVA algorithm is introduced as a mixture of two distributions to better model speech signals. The source prior is constructed by mixing the original multivariate super Gaussian distribution and the multivariate Student's t distribution with a certain ratio. The mixed source prior is empowered with an energy driven scheme that adjusts the weight of each distribution according to the energy of the observed mixtures so it can adapt to different types of speech signals. Moreover, an overlapped clique based dependency model is adopted for frequency spectrum structure of the sources. The IVA and the FastIVA methods are evaluated using the new techniques in different real room environments.

- Chapter 6 proposes a new robust adaptive learning based scheme to improve the separation performance of the online IVA algorithm. The scheme controls the learning rate by exploiting gear-shifting to address the trade-off between the high and small values of the learning rate. A generalized Gaussian source prior is introduced to the online IVA algorithm with the proposed scheme. The scheme was tested and compared with the original online IVA algorithm in real room environments and real recordings. Then, a new switched source prior technique is added to the adaptive learning scheme. The technique acquires the best aspect of the original super Gaussian and the generalized Gaussian source priors.

- Finally, conclusions are drawn from the thesis in Chapter 7. It summarises the contributions of the thesis and discussions including suggestions for future work.

Chapter 2

# BACKGROUND AND RELATED LITERATURE REVIEW

## 2.1 Introduction

Blind source separation (BSS) is a technique for recovering individual source signals from observed mixture signals at multiple sensors. The estimation technique is performed without having information about the original source signals or the mixing process. The separation of speech signals from their mixtures, known as the cocktail party problem [8], is an application of BSS. The BSS task strongly depends on the way in which the original source signals are mixed within the physical environment. The simplest mixing model is the instantaneous mixing. However, speech signal mixtures in a real reverberant environment are generally convolutive mixtures.

The separation of convolutive mixtures can be addressed in the time domain. However, time domain BSS methods are generally not suitable for the convolutive BSS (CBSS) problem due to the computational complexity [1]. In order to reduce the computational cost, CBSS is usually addressed in the frequency domain termed FD-BSS [1].

This chapter introduces the background theory of BSS in the time and frequency domains. The major techniques for solving the BSS problem in the frequency domain are discussed along with their ambiguities. Most FD-BSS algorithms are based on extracting second order statistics (SOS) or higher order statistics (HOS) from the observations data. SOS methods exploit the statistical non-stationarity of the speech signals and HOS methods use the non-Gaussianity of the speech signals to separate the mixed speech signals. The Parra-Spence algorithm based on SOS in the frequency domain [39] will be reviewed.

Independent component analysis (ICA), a well-known BSS algorithmic technique [40,41], will be reviewed along with its limitations in audio BSS applications. It separates sources from the observed mixtures by maximising statistical independence among source signals. Independent vector analysis (IVA), is an extension of ICA from univariate to multivariate components to avoid theoretically the permutation problem inherent to ICA [21]. It utilises the statistical independence among multivariate signal sources and the statistical inter-dependency of each multivariate source signal.

Enhancing the performance of the IVA algorithm is the focus of this thesis and will be reviewed in detail in this chapter. The original natural gradient IVA (NG-IVA) algorithm [20] uses the gradient descent method [42] to optimise the contrast function. The fast fixed point IVA (FastIVA) algorithm [43] is a fast version of the IVA algorithm and it uses the Newton method [44] to minimise the contrast function.

## 2.2 Mixture and Separation Models

The blind source separation task depends on the way in which the signals are mixed within a physical environment. Many methods have been proposed to attempt to solve the BSS problem. In 1985, Herault and Jutten [45] were the first to address the problem of blind source separation.

### 2.2.1 Instantaneous Mixing

The basic mixing model is called *instantaneous mixing*. In this model, it is assumed that the mixtures are instantaneous. It is assumed that the signals arrive at the sensors (microphones) at the same time without any delay[1] or filtering. Figure 2.1 illustrates the *instantaneous mixing* process for the case of three sources and three microphones.

In instantaneous mixing, $N$ unknown source signals are combined to yield the $M$ measured sensor signals.

The noise free instantaneous mixing model is defined in the time domain as [46]:

$$x_j(t) = \sum_{i=1}^{N} h_{ji} s_i(t) \qquad j = 1, \cdots, M \tag{2.2.1}$$

where $x_j(t)$ is the $j$th element of the mixture vector, $s_i(t)$ is the $i$th element of the source vector and $h_{ji}$ is the $j$th row and $i$th column element of the mixing matrix $\mathbf{H}$. In matrix form:

$$\mathbf{x}(t) = \mathbf{H}\mathbf{s}(t) \tag{2.2.2}$$

[1] In practice, speech signals will take a finite time to travel from the source to the sensors.

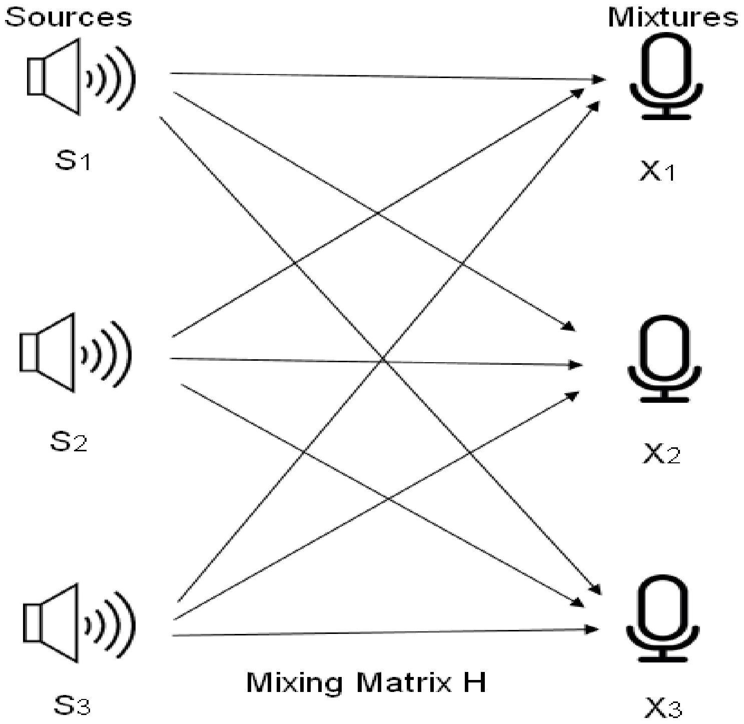


**Figure 2.1.** Instantaneous mixing of three sources and three microphone observations.

Assuming the unmixing matrix $\mathbf{W}$ is known then source are estimated as:

$$y_i(t) = \sum_{j=1}^{M} \mathrm{w}_{ij} x_j(t) \qquad i = 1, \cdots, N \tag{2.2.3}$$

where $y_i(t)$ is the $i$th element of the estimated source vector and $\mathrm{w}_{ij}$ is the $i$th row and $j$th column element of the unmixing matrix $\mathbf{W}$. In matrix form:

$$\mathbf{y}(t) = \mathbf{W}\mathbf{x}(t) \tag{2.2.4}$$

Many algorithms have been developed to solve the instantaneous mixing case namely to find the unmixing matrix $\mathbf{W}$ from the observations $\mathbf{x}(t)$ [22, 47]. Although useful for theoretical derivations, such algorithms do not offer practical solutions for speech source separation. The instantaneous model does not generally represent real-world room environments. For a real room environment, the acoustic signals take multiple paths to the microphone sensors instead of only the direct path. Thus, the convolutive model is used to represent the practical situation.

### 2.2.2 Convolutive Mixing

Real-world acoustical paths in a reverberant room environment lead to the *convolutive mixing* of the sources when measured at the acoustic sensors. *Convolutive mixing* occurs due to the time delays resulting from sound propagation over space and the *multipath* generated by reflections of sound off the walls and different objects in the room [48]. CBSS has driven much recent research work in the field of BSS. Figure 2.2 illustrates the *convolutive mixing* process for the case of three sources and three microphones.

In CBSS the sources are assumed to be convolved with a linear model [49]. CBSS introduces the following noise free relation between the $j$th mixed signal and the original source signals:

$$\mathbf{x}_j(t) = \sum_{i=1}^{N} \sum_{p=0}^{P-1} h_{ji}(p)\mathbf{s}_i(t-p) \qquad j = 1, \cdots, M \tag{2.2.5}$$

The mixed signal $\mathbf{x}_j(t)$ is a linear mixture of filtered versions of each of the source signals $\mathbf{s}_i(t)$, where $h_{ji}(p)$ represent the corresponding mixture filter coefficient from source $i$ to microphone $j$ and $P$ is the mixing filter length in time. The room impulse response can be represented in the form of a multichannel FIR filter $\mathbf{H}(p)$, $p = 0, \ldots, P-1$ to produce $M$ sensor signals, where:

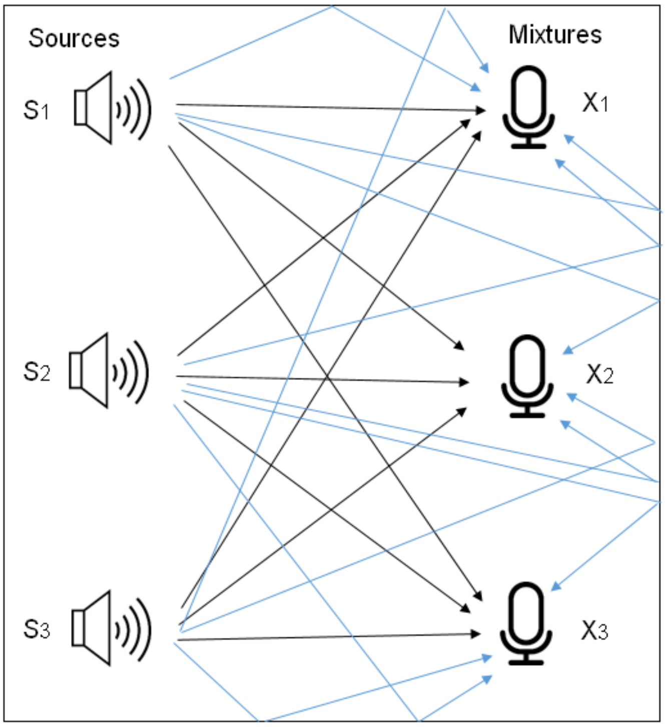


**Figure 2.2.** Convolutive mixing of three sources and three microphone observations.

$$\mathbf{H}(p) = \begin{bmatrix} h_{11}(p) & \dots & h_{1N}(p) \\ \vdots & \ddots & \vdots \\ h_{M1}(p) & \dots & h_{MN}(p) \end{bmatrix} \tag{2.2.6}$$

In time domain CBSS, the sources are estimated using a set of inverse FIR filter matrices $\mathbf{W}(q)$, $q = 0, \dots, Q-1$ such that:

$$\mathbf{y}_i(t) = \sum_{j=1}^{M} \sum_{q=0}^{Q-1} \mathrm{w}_{ij}(q)\mathbf{x}_j(t-q) \qquad i = 1, \cdots, N \tag{2.2.7}$$

The estimated signal $\mathbf{y}_i(t)$ is a linear mixture of filtered versions of each of the mixture signals $\mathbf{x}_j(t)$, where $\mathrm{w}_{ij}(q)$ represent the corresponding separating filter coefficient from mixture $j$ to output source

$i$ and $Q$ is the unmixing filter length in time. The $q$th slice of the unmixing filter $\mathbf{W}(q)$ is:

$$\mathbf{W}(q) = \begin{bmatrix} \mathrm{w}_{11}(p) & \dots & \mathrm{w}_{1M}(p) \\ \vdots & \ddots & \vdots \\ \mathrm{w}_{N1}(p) & \dots & \mathrm{w}_{NM}(p) \end{bmatrix} \tag{2.2.8}$$

There is advantage to perform CBSS in the frequency domain which is next considered.

## 2.3 Frequency Domain CBSS

When CBSS is applied to speech mixtures, it involves relatively long multichannel FIR filters to achieve separation even with moderate room reverberation. Although time-domain algorithms can be developed to perform the task, they can be hard to implement due to the multichannel convolution operations involved. The CBSS process can be simplified by transforming the task in the frequency domain, as convolution in the time domain theoretically becomes multiplication in the frequency domain. Ideally, each frequency component of the mixture signal contains an instantaneous mixture of the corresponding frequency components of the underlying source signals.

Transformation of time-domain signals into the frequency-domain is usually performed via the discrete Fourier transform (DFT) or the short-time Fourier transform (STFT). Using a T-point windowed discrete Fourier transformation (DFT), the time domain signals $\mathbf{x}_j(t)$, $j = 1, \dots, M$, can be converted into time-frequency domain signals $\mathbf{x}_j(\omega, t_k)$ where $\omega$ is a normalised frequency index and $t_k$ is a discrete time block index, $k = 1, \ldots, K$ ($K$ represents the total number of data blocks). The transform is given as:

$$\mathbf{x}(\omega, t_k) = \sum_{\tau=0}^{T-1} \mathbf{x}(t_k, \tau) e^{-j\omega\tau/T} \tag{2.3.1}$$

The linear convolution in the time domain can be written in the frequency domain as separate multiplications for each frequency bin as:

$$\mathbf{x}(\omega, t_k) = \mathbf{H}(\omega)\mathbf{s}(\omega, t_k) \tag{2.3.2}$$

where $\mathbf{x}(\omega, t_k)$ denotes the DFT of the current mixed vector frame at discrete time block instant $t_k$ and frequency $\omega$ and $\mathbf{s}(\omega, t_k)$ denotes the corresponding DFT of the source vector frame. The matrix $\mathbf{H}(\omega)$ is the frequency representation for the mixing impulse response $h_{ji}(p)$. It is an $M \times N$ time invariant mixing matrix at frequency $\omega$ and can be represented as:

$$\mathbf{H}(\omega) = \begin{bmatrix} h_{11}(\omega) & \ldots & h_{1N}(\omega) \\ \vdots & \ddots & \vdots \\ h_{M1}(\omega) & \ldots & h_{MN}(\omega) \end{bmatrix} \tag{2.3.3}$$

For a particular frequency bin $\omega$, the system represents an instantaneous mixing system. The block diagram of a generic frequency-domain CBSS procedure is shown in Figure 2.3. The separation system at frequency $\omega$ and discrete time block $t_k$ is represented as:

$$\mathbf{y}(\omega, t_k) = \mathbf{W}(\omega)\mathbf{x}(\omega, t_k) \tag{2.3.4}$$

where $\mathbf{y}(\omega, t_k)$ denotes the DFT of the estimated source signals at dis-

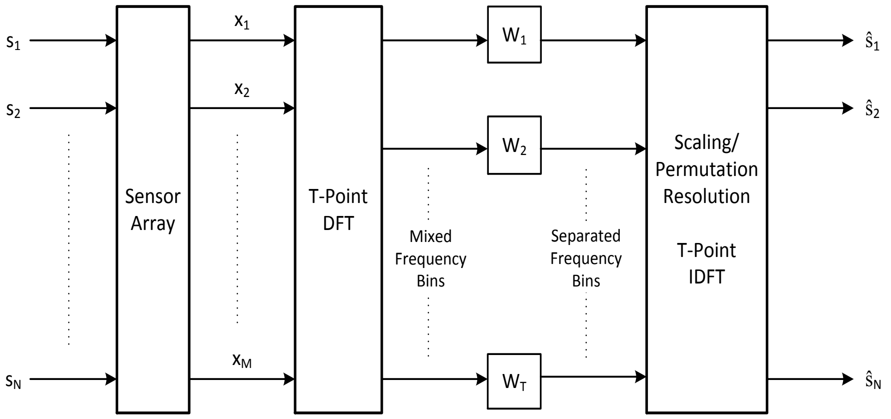


**Figure 2.3.** Block diagram of frequency domain BSS (FD-BSS).

crete time block instant $t_k$ and frequency $\omega$. $\mathbf{W}(\omega)$ is the $N \times M$ frequency representation of the unmixing matrix. $\mathbf{W}(\omega)$ is determined so that $\hat{\mathbf{s}}_i(\omega, t_k) = \mathbf{y}_i(\omega, t_k)$, $i = 1, \ldots, N$, become as independent as possible. $\mathbf{W}(\omega)$ can be represented as:

$$\mathbf{W}(\omega) = \begin{bmatrix} \mathrm{w}_{11}(\omega) & \ldots & \mathrm{w}_{1M}(\omega) \\ \vdots & \ddots & \vdots \\ \mathrm{w}_{N1}(\omega) & \ldots & \mathrm{w}_{NM}(\omega) \end{bmatrix} \tag{2.3.5}$$

The time domain separated signals $\hat{\mathbf{s}}(t) = \mathbf{y}_i(t)$ can then be obtained by using an inverse DFT (IDFT) operation. The signals that are reconstructed via the IDFT after the separation step may have spectral components corresponding to multiple sources, and they may suffer from distortions due to spectral errors [49]. Thus, to obtain good performance from FD-BSS methods, it is necessary to solve the ambiguities in FD-BSS. Once ambiguities are mitigated, the final time-domain source estimates can be reconstructed using overlap-add methods [50]. Ambiguities in FD-BSS are discussed next.

### 2.3.1 Ambiguities in Frequency Domain BSS

In BSS algorithms, the source signals and mixing matrix $\mathbf{H}$ are assumed to be unknown and they generally target at restoring independence [2]. This may lead to some ambiguities in the possible solutions across different frequency bins obtained in FD-BSS. Therefore, there is no guarantee that the separated signals obtained by such procedures will have the same scaling and permutation properties for different values of $\omega$. Ideally, in FD-BSS, the separation system is adapted such that:

$$\mathbf{W}(\omega)\mathbf{H}(\omega) = \mathbf{I} \tag{2.3.6}$$

Due to the ***scaling*** and ***permutation*** ambiguities it is not necessarily the case that the separating matrix $\mathbf{W}$ corresponds exactly to the inverse of the mixing matrix $\mathbf{H}$. However, it is the case that:

$$\mathbf{W}(\omega)\mathbf{H}(\omega) = \mathbf{P}(\omega)\mathbf{D}(\omega) \tag{2.3.7}$$

where $\mathbf{P}(\omega)$ and $\mathbf{D}(\omega)$ are frequency-dependent permutation and diagonal scaling matrices, respectively.

#### 2.3.1.1 Scaling

Due to the lack of prior information about the sources and the mixing matrix $\mathbf{H}$, the energies of the source signals cannot be normally estimated. Inevitability, the energies of the independent estimated source signals are not equivalent to those of the original source signals. If one of the sources $s_i$ is multiplied by a non-zero scalar $\alpha$, dividing the corresponding column $h_i$ of $\mathbf{H}$ by the same non-zero scalar could cancel its effect:

$$x = \sum_i (\frac{1}{\alpha} h_i)(s_i \alpha) \tag{2.3.8}$$

This demonstrates that the sources can be estimated only up to a scaling constant. Scaling ambiguities can result in unequal scaling of the spectral components before reconstruction. Scaling ambiguities are most often resolved by some form of normalisation of each separation matrix at each frequency bin or by assuming a unit variance for the independent estimated zero mean source components. The sign ambiguity normally has minor effect on speech signals in BSS.

#### 2.3.1.2 Permutation

The order in which the components is recovered may not be determined correctly. Permutation ambiguities can result in the spectral mixing of sources upon reconstruction. With instantaneous mixing models permutation does not usually affect the solution of the BSS algorithm. However, in the case of convolutive mixing models, the permutation ambiguities becomes a major problem. If the FD-BSS problem is solved independently at each frequency bin, the order of the estimated source signals at each frequency bin will most likely be inconsistent across all frequency bins. The problem is depicted in Figure 2.4. In the diagram the order of the estimated signals $(\hat{s}_1^{(k)}, \hat{s}_2^{(k)}, \cdots, \hat{s}_N^{(k)})$ is different from the order of the original source signals $(s_1^{(k)}, s_2^{(k)}, \cdots, s_N^{(k)})$ at each frequency bin.

Source permutation is a much more challenging problem and has received considerable attention by several researchers. Most methods for resolving frequency-dependent permutation fall into one of three categories [2], namely methods exploiting:

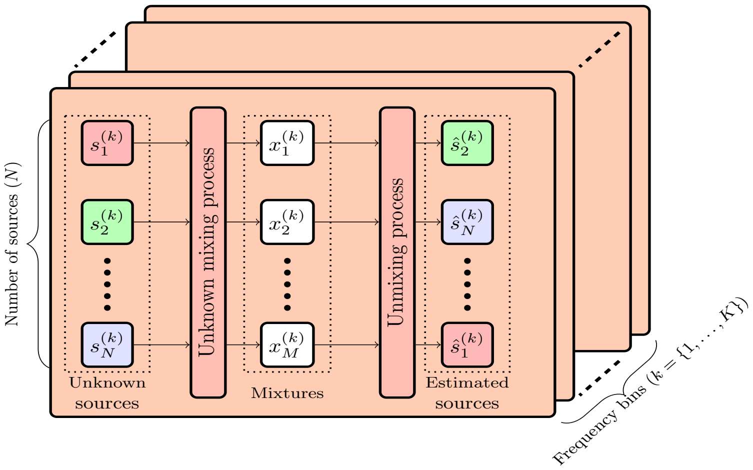


**Figure 2.4.** A graphical representation of the permutation problem in FD-BSS [51].

- Signal properties of the DFT.
- Properties of speech.
- Geometric properties of the sensor array, such as directions of arrival (DOAs).

All three classes of methods require additional information about the measurement setup or the signals being separated. The major FD-BSS techniques are presented in the next section.

## 2.4 BSS Techniques

There are various techniques in the literature to address the BSS problem. BSS algorithms are based on different assumptions on the sources and the mixing and separation model or system. The sources are usually assumed to be independent or decorrelated. The algorithms can be divided according to the separation criterion into methods based on second order statistics (SOS), and methods based on higher order statistics (HOS). In CBSS it is also assumed that sensors receive $N$ linearly independent versions of the source signals and there are at least as many sensors $M$ in an exactly determined system, i.e., $M \geq N$. Different algorithms make different assumptions on the statistics of the sources. These methods are motivated by the present understanding on the grouping principles of auditory perception commonly referred to as “Auditory Scene Analysis” (ASA) [1].

### 2.4.1 Second Order Statistics BSS

BSS algorithms based on second order statistics (SOS) separate the sources depending on decorrelation instead of the stronger condition of independence between the sources. SOS conditions alone are not adequate for separation of sources. Hence, these methods require additional conditions for separation [1]. They work on assumptions such as the statistical non-stationarity of the sources [39] or a minimum phase mixing system [52]. The main advantage of SOS is that they are less sensitive to noise and outliers, as a result they do not require a huge amount of data for the estimation process [53]. A separation algorithm based on SOS in the frequency domain which was proposed by Parra and Spence [39] is discussed next.

#### 2.4.1.1 Parra-Spence Algoirthm

The algorithm exploits the non-stationarity of speech which can be considered statistically non-stationary for time scales beyond 10ms [54,55]. It employs cross-correlation at multiple times to provide sufficient separation conditions and uses least squares (LS) optimisation to estimate the unmixing matrix $\mathbf{W}(\omega)$. The separation matrix is estimated by decorrelating the cross-correlation matrices at different lags by searching for $\mathbf{W}(\omega)$ that diagonalise simultaneously the cross-correlation matrices of the estimated sources at $K$ different times:

The gradient descent algorithm is used to diagonalise the unmixing matrix for all the frequency bins by minimising the sum-squared error (as the sum of off-diagonal elements of the covariance matrix of the estimated sources).

$$\mathbf{y}(\omega, t_k) = \mathbf{W}\mathbf{x}(\omega, t_k) \tag{2.4.1}$$

SOS in the frequency domain is captured by the cross-power spectrum:

$$\begin{aligned}\mathbf{R}_y(\omega, t_k) &= \mathbf{W}(\omega)\mathbf{R}_x(\omega, t_k)\mathbf{W}^H(\omega) \\ &= \mathbf{W}(\omega)\mathbf{H}(\omega)\mathbf{\Lambda}_s(\omega, t_k)\mathbf{H}^H(\omega)\mathbf{W}^H(\omega) \end{aligned} \tag{2.4.2}$$

where $\mathbf{\Lambda}_s(\omega, t_k)$ is a diagonal covariance matrix describing the source signals for each discrete time block $t_k$, $\mathbf{R}_x(\omega, t_k)$ is the covariance matrix of $\mathbf{x}(\omega, t_k)$ and $(.)^H$ denotes Hermitian transpose.

The aim is to minimise the cross-powers on the off-diagonal of the matrix $\mathbf{R}_y(\omega, t_k)$. The covariance matrices are estimated using an averaged cross-power spectrum:

$$\hat{\mathbf{R}}_x(\omega, t_k) = \frac{1}{L}\sum_{l=0}^{L-1} \mathbf{x}(\omega, t_k + lT)\mathbf{x}^H(\omega, t_k + lT) \tag{2.4.3}$$

where $T$ is the block length of the FFT.

The cost function $J_m$ based on the off-diagonal elements of $\mathbf{R}_y(\omega, t_k)$ estimated at $t_k = kTL, k = 1, \cdots, K$, with $K$ being the number of matrices to diagonalise, is given by:

$$J_m = \sum_{\omega=1}^{T}\sum_{k=1}^{K} ||\mathbf{E}(\omega, t_k)||_F^2 \tag{2.4.4}$$

where $\mathbf{E}(\omega, t_k) = \mathbf{W}(\omega)\hat{\mathbf{R}}_x(\omega, t_k)\mathbf{W}^H(\omega) - \mathbf{\Lambda}_s(\omega, t_k)$ and $\|.\|_F^2$ is the squared Frobenius norm.

To minimise $J_m$ the method of steepest descent is used giving:

$$\frac{\partial J_m}{\partial \mathbf{W}^*(\omega)} = 2\sum_{k=1}^{K} \mathbf{E}(\omega, t_k)\mathbf{W}(\omega)\hat{\mathbf{R}}_x(\omega, t_k) \tag{2.4.5}$$

where $(.)^*$ denotes the conjugate operator, and the update equation for $\mathbf{W}(\omega)$ becomes:

$$\mathbf{W}_{j+1}(\omega) = \mathbf{W}_j(\omega) - \eta\sum_{k=1}^{K} \mathbf{E}(\omega, t_k)\mathbf{W}(\omega)\hat{\mathbf{R}}_x(\omega, t_k) \tag{2.4.6}$$

where $j$ is the iteration index and $\eta$ is the learning rate.

The unmixing matrix $\mathbf{W}(\omega)$ is updated for all the frequency bins. The source covariance matrix can be estimated at each iteration by:

$$\hat{\mathbf{\Lambda}}_s(\omega, t_k) = diag\{\mathbf{W}(\omega)\mathbf{R}_x(\omega, t_k)\mathbf{W}^H(\omega)\} \tag{2.4.7}$$

The arbitrary permutation of the coordinates for each frequency will lead to the same error $\mathbf{E}(\omega, t_k)$. Therefore, choosing a different permutation of the solutions for each frequency bin will not change the total cost.

Since accurate reconstruction of the sources requires consistent permutation for all frequencies, the Parra-Spence algorithm suffers from the permutation problem. Parra and Spence proposed a possible solution to the permutation problem [39] by imposing a smoothness constraint on the separating filters to produce better alignment of the frequency bins. This can be achieved by constraining the filter length Q to be much less than the size of the DFT ($\mathbf{W}(\tau) = 0$ for $\tau > Q$ and

$Q \ll T$). BSS based on higher order statistics (HOS) is explained in the next section.

### 2.4.2 Higher Order Statistics BSS

BSS algorithms based on higher order statistics HOS separate the sources on the assumption that they are statistically independent. The statistical independence implies uncorrelated sources, but the reverse is not necessarily true. Many algorithms are based on minimising second and fourth order dependence between the source signals [1]. To successfully separate the sources by means of higher order moments, it is necessary for the sources to be non-Gaussian (with the exception of one at the most), given the fact that Gaussian sources have zero higher cumulants [5]. The two major HOS BSS techniques, independent component analysis ICA and independent vector analysis IVA, are described next.

## 2.5 Independent Component Analysis (ICA)

ICA is a statistical and computational efficient technique that reveals hidden factors that are contained within sets of random variables, measurements, or signals. It defines a generative model that expresses the data variables as a linear combination of some unknown latent variables with an unknown mixing system [47]. The latent variables, known as the independent components of the observed data, are assumed to be non-Gaussian and mutually independent which can be found by ICA. Although ICA is superficially related to principal component analysis and factor analysis, it is a more powerful technique capable of finding the underlying factors or sources. Certain fundamental assumptions are necessary for ICA to work [6, 47, 56]:

- The source signals are assumed to be statistically independent. Statistical independence between the source signals is expressed in terms of the probability density functions (PDF). If the model sources are independent, the joint probability density function can be written as:

  $$p(s_1, \ldots, s_N) = \prod_{i=1}^{N} p(s_i) \tag{2.5.1}$$

  where $p(s_i)$ is the marginal distribution of the i$th$ source.

  This is equivalent to stating that model sources $s_i$ do not carry mutual information.

- With the exception of one, all other sources must be non-Gaussian signals. It is not possible to use HOS ICA if all the sources are Gaussian because the higher order cumulants of a Gaussian distribution are zero.

- The system is exactly determined when the number of sources is equal to the number of mixtures. This means the mixing matrix $\mathbf{H}$ is assumed to be square ($N = M$) and invertible.

Generally, ICA algorithms are carried out in two stages. First, the mixtures are decorrelated via spatial whitening and then the estimation process is performed by optimising their separating objective contrast or cost functions. This spatial whitening is accomplished by employing principal component analysis (PCA).

### 2.5.1 Principle Component Analysis (PCA)

In ICA BSS, PCA is used to whiten the observed signals by removing the cross-correlation between them, and ensuring that they have unit variance [5]. PCA operates by finding the projections of the mixture data in orthogonal directions of maximum variance. The whitening process is usually done after the data were centred by subtracting the mean from the observed data. A zero mean vector $\mathbf{z}$ containing observations from spatially distinct locations is said to be spatially white if:

$$E(\mathbf{z}\mathbf{z}^T) = \mathbf{I} \tag{2.5.2}$$

where $E(.)$ is the statistical expectation operator, $(.)^T$ is the transpose operator and $\mathbf{I}$ is the identity matrix. The unmixing matrix, $\mathbf{W}$, can be decomposed into two components as:

$$\mathbf{W} = \mathbf{U}\mathbf{Q} \tag{2.5.3}$$

where $\mathbf{Q}$ denotes the whitening matrix and $\mathbf{U}$ is the rotation matrix [11].

PCA might be done using eigen-value decomposition (EVD) of the covariance matrix $C_x$:

$$\mathbf{C}_x = \mathbf{E}\mathbf{D}\mathbf{E}^T \tag{2.5.4}$$

The whitening matrix $\mathbf{Q}$ can be formulated as:

$$\mathbf{Q} = \mathbf{D}^{-1/2}\mathbf{E}^T \tag{2.5.5}$$

where $\mathbf{E}$ is the matrix of eigenvectors of the covariance matrix $\mathbf{C}_x$ and $\mathbf{D}$ is the diagonal matrix of the eigenvalues of $\mathbf{C}_x(\mathbf{D} = diag(d_1, \cdots, d_n))$.

It is important to note that the whitening matrix $\mathbf{Q}$ is not unique because it can be pre-multiplied by an orthogonal matrix to obtain another version of $\mathbf{Q}$. $\mathbf{C}_x$ can be estimated as a time average using samples of the observed vector $\mathbf{x}(1), \cdots, \mathbf{x}(T)$. The whitened vector $\mathbf{z}$ is obtained as follows:

$$\mathbf{z} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{E}^T\mathbf{x} = \tilde{\mathbf{A}}s \tag{2.5.6}$$

where $\mathbf{D}^{-1/2} = diag(d_1, \cdots, d_n)$.

PCA requires the $n$ diagonal elements of the whitened data covariance matrix $\mathbf{C}_z$ to be unity. Due to the symmetry of $\mathbf{C}_z$, $(n^2 - n)/2$ of its off-diagonal elements can be set to zero. This means spatial whiteness imposes $n(n+1)/2$ constraints. Therefore the whitening process reduces the number of unknown parameters to $n(n-1)/2$ instead of the $n^2$ originally required.

### 2.5.2 Learning Algorithm: Natural Gradient ICA

The fundamental idea of ICA is to minimise the dependency among the output components. The independence is measured by the average mutual information (MI) of the estimated sources [5]. The Kullback-Leibler divergence between the joint distribution $p(\hat{\mathbf{s}})$ and the product of the marginal distributions of the source outputs $\prod_{1=1}^{N} q(\hat{s}_i)$:

$$C_{ICA} = KL\Big(p(\hat{\mathbf{s}}) \| \prod_{1=1}^{N} q(\hat{s}_i)\Big) \tag{2.5.7}$$

$$= \int p(\hat{\mathbf{s}}) \log \frac{p(\hat{\mathbf{s}})}{\prod_{1}^{N} q(\hat{s}_i)} d\hat{\mathbf{s}} \tag{2.5.8}$$

$$= \int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} - \log |det(\mathbf{W})| - \sum_{i=1}^{N} \int p(\mathbf{y}_i) \log q(\hat{s}) d\hat{\mathbf{s}}_i \quad (2.5.9)$$

$$= const. - \sum_{k=1}^{K} \log |det(\mathbf{W})| - \sum_{i=1}^{N} E[\log q(\hat{s}_i)] \quad (2.5.10)$$

By differentiating Equation (2.5.10) with respect to the separating matrix $\mathbf{W}$, the gradient of the cost function can be calculated as follows:

$$\Delta \mathbf{W} = -\frac{\partial C_{ICA}}{\partial \mathbf{W}} = \mathbf{W}^{-1} E[\varphi_{ICA}(\hat{\mathbf{s}})] \mathbf{x}_j^T \quad (2.5.11)$$

where $(.)^{-1}$ denotes the inverse of a matrix and the nonlinear score function for ICA in its general form:

$$\varphi_{ICA}(\hat{\mathbf{s}}) = \frac{\partial \log q(\hat{s}_i)}{\partial \mathbf{s}_i} \quad (2.5.12)$$

The natural gradient [42] can be calculated by multiplying through by $\mathbf{W}^T\mathbf{W}$:

$$\Delta \mathbf{W} \propto (\mathbf{I} - E[\varphi_{ICA}(\hat{\mathbf{s}})\hat{\mathbf{s}}^T])\mathbf{W} \quad (2.5.13)$$

Then the update rule for natural gradient ICA (NG-ICA):

$$\mathbf{W}(l+1) = \mathbf{W}(l) + \eta \Delta \mathbf{W} \mathbf{W}(l) \quad (2.5.14)$$

$$\mathbf{W}(l+1) = \mathbf{W}(l) + \eta(\mathbf{I} - E[\varphi_{ICA}(\hat{\mathbf{s}})\hat{\mathbf{s}}^T])\mathbf{W}(l) \quad (2.5.15)$$

where $\eta$ is a learning rate, and $l$ is the iteration index.

The exact non-linear score function is based on the distribution (PDF) used to model the statistics of the original sources. For example, a Laplacian source prior in the form:

$$q(s_i) \propto \exp\left(\frac{|s_i - \mu_i|}{\sigma_i}\right) \quad (2.5.16)$$

where $\sigma_i$ is the standard deviation of each source. The non-linear score function becomes:

$$\varphi_{ICA}(\hat{\mathbf{s}}) = \frac{\partial \log q(\hat{s}_i)}{\partial \mathbf{s}_i} = \frac{\hat{s}_i}{|\hat{s}_i|} \tag{2.5.17}$$

Various source priors can be deployed to model the speech signals [22, 57]. The separation performance depends on the source prior selected.

Different forms of the ICA algorithm have been introduced, including the popular FastICA algorithm described in [58] which will be explained in Chapter 4. In the ICA method, the alignment of the separated signals is not consistent across all the frequency bins. Therefore, it is necessary to correct the permutations of separating matrices at each frequency to achieve an accurate reconstruction of the separated signal in the time domain. A widely used approach is to impose a smoothness constraint of the source that translates into smoothing of the separating filter. This approach has been recognised by several techniques such as averaging separating matrices with adjacent frequencies [39], limiting the filter length in the time domain [59]. Direction of arrival estimation has also been exploited [23,60] and video tracking of sources was suggested [61]. Although these methods perform well under certain conditions, they may not provide good perform in general conditions.

The independent vector analysis (IVA) algorithm is a recent FD-BSS algorithm introduced to solve the permutation problem algorithmically [20]. It models independence between source signal vectors and dependency between frequency bins within each source vector. The IVA algorithm is discussed in detail in the next section.

## 2.6 Independent Vector Analysis (IVA)

In the ICA approach, the source signal prior is defined independently at each frequency bin. ICA methods suffer from the unknown permutation of the output signals over different frequency bins due to the indeterminacy of permutation inherent in the ICA algorithm.

To mitigate the permutation problem, Kim et al. proposed the independent vector analysis (IVA) approach [20, 21]. It is a frequency domain BSS method based on an improved model of the ICA method. It assumes that dependencies exist between frequency bins instead of defining independence for each frequency bin. The proposed method introduces the concept of multivariate components by extending the ICA formulation of univariate source signals to multivariate source signals. It exploits a dependency model capturing inherent interfrequency dependencies of the speech signals.

In IVA, the sources are considered to be multidimensional random vectors, not just single variables as in the ICA. Since the elements of a random vector are related to each other, elements within a source vector are dependent as well as correlated [21]. The algorithm allows independence between multivariate source signals represented as random vectors, and retains dependency between the source signals within each source vector. It also considers that multivariate signals have a multidimensional mixing linear model [20].

Compared with ICA methods, the interfrequency relationships depend on a modified model for the source signal prior. The method uses higher order dependencies across frequencies. The IVA method defines each source prior as a multivariate super-Gaussian distribution which is a simple extension of the independent Laplacian distribution. Thus, it can preserve the higher order interfrequency dependencies and structures of frequency components. It therefore mitigates the permutation problem and improves the separation performance of sources [20]. In addition to the dependency model which captures interfrequency dependencies in data, the IVA approach proposes a new cost function that measures the independence among multivariate signals with multivariate probability density functions (PDFs). It is an extension of mutual information between multivariate random variables. The learning algorithm for the parameters of the separating filters is derived by minimising the cost function [20].

### 2.6.1 The IVA Model

In the IVA method, the time domain signal is converted to the frequency domain signal using the short time Fourier transform (STFT). This is then processed assuming the IVA mixing and separating model where both sources and observations are multivariate [20]. The noise free frequency domain BSS model is described as:

$$\mathbf{x}^{(k)} = \mathbf{H}^{(k)}\mathbf{s}^{(k)} \tag{2.6.1}$$

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 signal vector and the source signal vector in the frequency domain at the $k$th frequency bin respectively. $\mathbf{H}^{(k)}$ is the $M \times N$ mixing matrix at k-th frequency bin, $k = 1, 2, ..., K$, and $K$ is the number of frequency bins.

The separation model of the source signal is given as:

$$\hat{\mathbf{s}}^{(k)} = \mathbf{W}^{(k)}\mathbf{x}^{(k)} \tag{2.6.2}$$

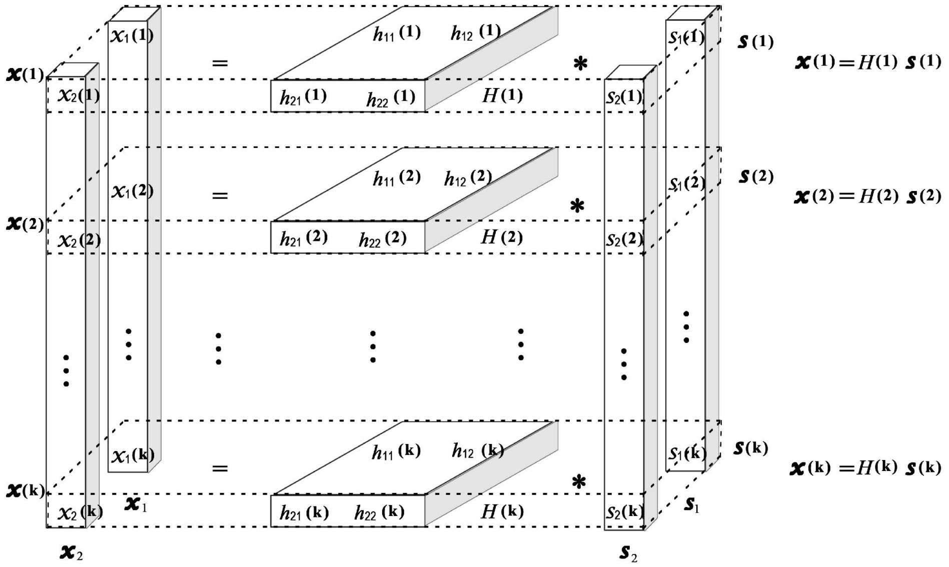


**Figure 2.5.** 3D independent vector analysis model for the case of two sources and two sensors case. IVA groups the dependent sources as a multivariate vector and learns each group as a whole. In the ICA, the mixing process is restricted to the source components on the same horizontal layer [62].

where $\hat{\mathbf{s}}^{(k)} = [\hat{s}_1^{(k)}, \hat{s}_2^{(k)}, \cdots, \hat{s}_N^{(k)}]^T$ is the estimated signal vector in the frequency domain, and $\mathbf{W}^{(k)}$is the $N \times M$ unmixing matrix at the $k$th frequency bin. Figure 2.5 shows the three-dimensional (3D) dependency structure of the IVA algorithm for the case of two sources and two sensors. The IVA model consists of a set of standard ICA models where the univariate sources across different layers are dependent such that they can be aligned and grouped together as a multivariate vector and each group is learned as a whole.

### 2.6.2 Cost Function

Separating multivariate sources from multivariate observations requires a cost function for multivariate random variables. The Kullback-Leibler

(KL) divergence between two functions is chosen as the measure of independence. In IVA, the two functions are the exact joint probability density function of the estimated sources $p(\hat{\mathbf{s}}_1, \cdots, \hat{\mathbf{s}}_N)$ and the product of marginal probability density functions of the individual source vectors $\prod_{1=1}^{N} q(\hat{\mathbf{s}}_i)$ [20]:

$$C = KL\Big(p(\hat{\mathbf{s}}_1, ..., \hat{\mathbf{s}}_N) \| \prod_{1=1}^{N} q(\hat{\mathbf{s}}_i)\Big)$$

$$= \int p(\hat{\mathbf{s}}_1, \cdots, \hat{\mathbf{s}}_N) \log \frac{p(\hat{\mathbf{s}}_1, \cdots, \hat{\mathbf{s}}_N)}{\prod_1^N q(\hat{\mathbf{s}}_i)} d\hat{\mathbf{s}}_1, \cdots, d\hat{\mathbf{s}}_N$$

$$= \int p(\mathbf{x}_1, \cdots, \mathbf{x}_M) \log p(\mathbf{x}_1, \cdots, \mathbf{x}_M) d\mathbf{x}_1, \cdots, d\mathbf{x}_M$$
$$- \sum_{k=1}^{K} \log |det(\mathbf{W}^{(k)})| - \sum_{i=1}^{N} \int p(\mathbf{y}_i) \log q(\hat{\mathbf{s}}) d\hat{\mathbf{s}}_i$$

$$= const - \sum_{k=1}^{K} \log |det(\mathbf{W}^{(k)})| - \sum_{i=1}^{N} E[\log q(\hat{\mathbf{s}})] \qquad (2.6.3)$$

where $\int p(\mathbf{x}_1, \cdots, \mathbf{x}_M) \log p(\mathbf{x}_1, \cdots, \mathbf{x}_M) d\mathbf{x}_1, \cdots, d\mathbf{x}_M$ is the entropy of the given observations, which is a constant, $det(.)$ is the matrix determinant operator and $|.|$ denotes the absolute value. The random variables are multivariate.

The source prior $q(\hat{\mathbf{s}})$ in the cost function is a vector across all frequency bins. Each source is multivariate and the cost would be minimised when the dependency between the source vectors is removed but the dependency between the components of each vector can be retained. Therefore, the cost function removes dependency between the vector sources and preserves the inherent frequency dependency within each source vector.

### 2.6.3 Learning Algorithm: A Gradient Descent Method (Natural Gradient IVA)

The learning algorithm for the parameters of the separating filters is derived by minimising the KL cost function using a gradient descent method. By differentiating the cost function with respect to the coefficients of the separating matrices ($\mathrm{w}_{ij}^{(k)}$), the gradients for the coefficients ($\Delta \mathrm{w}_{ij}^{(k)}$) can be obtained as follows:

$$\Delta \mathrm{w}_{ij}^{(k)} = -\frac{\partial C}{\partial \mathrm{w}_{ij}^{(k)}} \tag{2.6.4}$$

$$= \mathrm{w}_{ij}^{-H(k)} - E\Big[\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}, \cdots, \hat{\mathbf{s}}_i^{(k)})\hat{\mathbf{x}}_j^{*(k)}\Big] \tag{2.6.5}$$

where $(\mathbf{W}^{(k)^{-1}})^H \equiv \mathrm{w}_{ij}^{-H(k)}$

The natural gradient is a fast convergence method [42] which can be obtained by multiplying scaling matrices $\mathbf{W}^{(k)H}\mathbf{W}^{(k)}$ to the gradient matrices $\Delta\mathbf{W}^{(k)} \equiv \{\Delta \mathrm{w}_{ij}^{(k)}\}$:

$$\Delta \mathrm{w}_{ij}^{(k)} = \sum_{l=1}^{N}(I_{il} - E\Big[\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}, \cdots, \hat{\mathbf{s}}_i^{(k)})\hat{\mathbf{s}}_l^{*(k)}\Big]\mathrm{w}_{ij}^{(k)} \tag{2.6.6}$$

where $\mathbf{I}$ is the identity matrix ($I_{il} = 1$ when $i = l$) and ($I_{il} = 0$ when $i \neq l$). The nonlinear score function vector $\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}, \cdots, \hat{\mathbf{s}}_i^{(k)})$ is defined as:

$$\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}, \cdots, \hat{\mathbf{s}}_i^{(k)}) = -\frac{\partial \log q(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)})}{\partial \hat{\mathbf{s}}_i^{(k)}} \tag{2.6.7}$$

The coefficients of the separating matrices can be updated by the batch update rule as [20]:

$$\mathrm{w}_{ij}^{(k)new} = \mathrm{w}_{ij}^{(k)old} + \eta \Delta \mathrm{w}_{ij}^{(k)} \tag{2.6.8}$$

where $\eta$ is the learning rate.

$\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}, \cdots, \hat{\mathbf{s}}_i^{(k)})$ is a multivariate score function which if defined as a univariate score function $\varphi^{(k)}(\hat{s}_i^{(k)})$ , the algorithm becomes conventional ICA. IVA is used to preserve the dependency structure across the frequency bins and to achieve a good separation performance.

The multivariate score function is strongly related to a source prior, because the cost function includes $q(\hat{s}_i)$, which is an approximated probability density function of a source vector, that is, $q(s_i) \approx p(s_i)$. Thus, the multivariate score function can be obtained by differentiating the log prior with respect to each element of a source vector.

In BSS approaches, when the sources have super-Gaussian distribution, a Laplacian distribution is widely used as a source prior. The source prior of a vector as an independent Laplacian source prior in each frequency can be expressed as:

$$p(\mathbf{s}_i) = \prod_{k=1}^{K} p(s_i^{(k)}) = \alpha \prod_{k=1}^{K} \exp(\frac{|s_i^{(k)} - \mu_i^{(k)}|}{\sigma_i^{(k)}}) \tag{2.6.9}$$

where $\alpha$ is a normalization factor, $\mu_i^{(k)}$ and $\sigma_i^{(k)}$ are respectively the mean and standard deviation of the $i$th source signal at the $k$th frequency bin. Figure 2.6 shows the two-dimensional PDF for this independent Laplacian distribution source prior [20].

Assuming zero mean and unit variance, the non-linear score function is given as:

$$\varphi^{(k)}(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)}) = \frac{\partial \sum_{k=1}^{K} |\hat{s}_i^{(k)}|}{\partial \hat{s}_i^{(k)}} = \frac{\hat{s}_i^{(k)}}{|\hat{s}_i^{(k)}|} \tag{2.6.10}$$

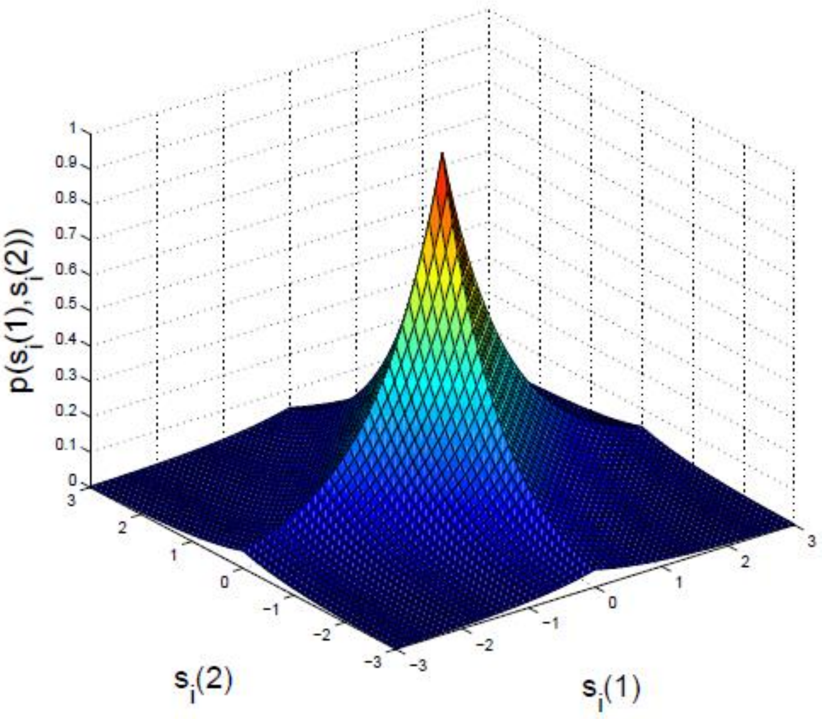


**Figure 2.6.** Two-dimensional PDF for independent Laplacian source prior. $\mathbf{s}_i(1)$ and $\mathbf{s}_i(2)$ can be considered as either real or imaginary parts.

As the above score function depends only on a single variable $\hat{s}_i^{(k)}$, it is a univariate function which is not capable of maintaining the dependency within the source vector. Therefore, a new source prior, which is greatly dependent on the other elements of a source vector, is required. The IVA proposed in [20] defines the source prior as a dependent multivariate super-Gaussian distribution in the form:

$$p(\mathbf{s}_i) = \alpha \exp\left( - \sqrt{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i)} \right) \qquad (2.6.11)$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean vector and covariance matrix of the $i$th source signal, respectively. Figure 2.7 shows the two-dimensional

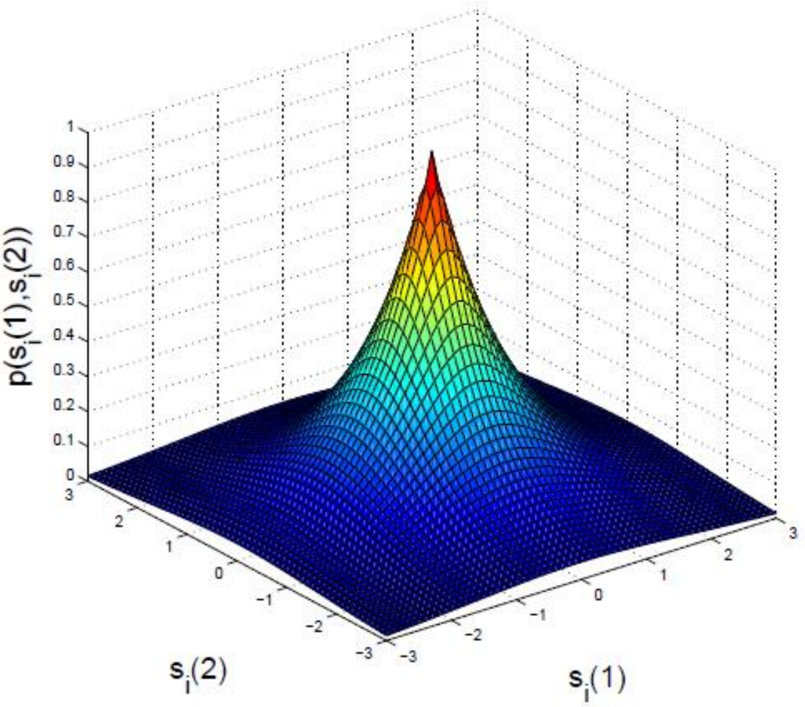


**Figure 2.7.** Two-dimensional PDF for the dependent multivariate super-Gaussian distribution source prior. $\mathbf{s}_i(1)$ and $\mathbf{s}_i(2)$ can be considered as either real or imaginary parts.

PDF for this dependent multivariate super-Gaussian distribution source prior [20]. As can be seen from the figure, the joint distribution of $p(s_i^{(1)}, s_i^{(2)})$ does not exhibit any directionality which means $s_i^{(1)}$ and $s_i^{(2)}$ are uncorrelated. However, the marginal distribution of $s_i^{(1)}$is different from the joint distribution of $s_i^{(1)}$ given $s_i^{(2)}$, that is, $s_i^{(1)}$ and $s_i^{(2)}$ are highly dependent.

The distribution shown in Figure 2.7 can be derived by a scaled mixture of Gaussians with a fixed mean and a variable variance, as follows:

Suppose that there is a $K$-dimensional random variable, which is defined by [63]:

$$\mathbf{s}_i = \sqrt{\nu}.\mathbf{z}_i + \boldsymbol{\mu}_i \tag{2.6.12}$$

where $\nu$ is a scalar random variable, $\boldsymbol{\mu}_i$ is a $K$-dimensional deterministic variable and $\mathbf{z}_i$ is a $K$-dimensional random variable that has Gaussian distribution with zero mean and $\boldsymbol{\Sigma}_i$ covariance matrix

$$p(\mathbf{z}_i) = \alpha_z \exp\bigg( - \frac{\mathbf{z}_i{}^H \boldsymbol{\Sigma}_i^{-1} \mathbf{z}_i}{2} \bigg) \tag{2.6.13}$$

where $\alpha_z$ is a normalization factor.

Assume $\nu$ has a Gamma distribution defined by:

$$p(\nu) = \alpha_\nu \nu^{\frac{K-1}{2}} \exp(-\frac{\nu}{2}) \tag{2.6.14}$$

where $\alpha_\nu$ is a normalization factor.

Then, the random variable $\mathbf{s}_i$ given $\nu$ has joint Gaussian distribution $p(\mathbf{s}_i|\nu)$ with mean $\boldsymbol{\mu}_i$ and covariance $\nu\boldsymbol{\Sigma}_i$. The original source prior can be obtained by integrating joint distribution $p(\mathbf{s}_i|\nu)$ with respect to $\nu$ as follows:

$$\begin{aligned} p(\mathbf{s}_i) &= \int_0^\infty p(\mathbf{s}_i|\nu)p(\nu)d\nu \\ = \hat{\alpha} \int_0^\infty \sqrt{\nu} \exp\bigg( -\frac{1}{2}\bigg(\frac{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i)}{\nu}\bigg)\bigg) d\nu \\ &= \alpha \exp\bigg( - \sqrt{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i)} \bigg) \end{aligned} \tag{2.6.15}$$

This indicates that each component of $\mathbf{s}_i$ is correlated to others and there is variance dependency generated by $\nu$. Even if the covariance matrix $\boldsymbol{\Sigma}_i$ is assumed to be identity, which means that each component of $\mathbf{s}_i$ is uncorrelated, the components are still dependent on each other.

Speech signals have inherent dependencies between frequency bins such as variance dependency. That is, the variances of the different frequency components are directly proportional to each other. However, since the Fourier transform has orthogonal bases and its outputs have zero means, the mean vector $\boldsymbol{\mu}_i$ can be set to zero and the covariance matrix $\boldsymbol{\Sigma}_i$ becomes a diagonal matrix. This implies that each frequency bin is uncorrelated with the others. Therefore, Equation (4.3.6) can be rewritten as:

$$p(s_i) = \alpha \exp\left( - \sqrt{\sum_{k=1}^{K} \left| \frac{\hat{s}_i{}^{(k)}}{\sigma_i{}^{(k)}} \right|^2} \right) \tag{2.6.16}$$

where $\sigma_i^{(k)}$ is the standard deviation of the $i$th source at the $k$th frequency bin which determines the scale of each element of a source vector. The parameter $\sigma_i^{(k)}$ is set to unity to adjust the scale after learning the separating filters. Thus:

$$p(\mathbf{s}_i) = \alpha \exp\left( - \sqrt{\sum_{k=1}^{K} |\ \hat{s}_i{}^{(k)}|^2} \right) \tag{2.6.17}$$

Accordingly, the multivariate nonlinear score function used in the algorithm to extract the $i$th source at the $k$th frequency is obtained as:

$$\varphi^{(k)}(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)}) = -\frac{\partial \log q(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)})}{\partial \hat{s}_i^{(k)}} \tag{2.6.18}$$

$$\varphi^{(k)}(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)}) = -\frac{\partial \log\left( - \exp\left( \sqrt{\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2} \right) \right)}{\partial \hat{s}_i^{(k)}} \tag{2.6.19}$$

$$\varphi^{(k)}(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)}) = \frac{\partial \sqrt{\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2}}{\partial \hat{s}_i^{(k)}} \tag{2.6.20}$$

$$\varphi^{(k)}(\hat{s}_i^{(1)}, \cdots, .\hat{s}_i^{(k)}) = \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2}} \tag{2.6.21}$$

This is a multivariate function which takes into account the dependency between the frequency bins in the learning process. This function is the proposed form of a multivariate score function for separating source signals in this algorithm. However, the form of a multivariate score function may vary based on different types of dependency. Designing and devising proper multivariate score functions for various dependency models is a promising area for further research and a focus of this thesis.

### 2.6.4 Scaling

Since natural signal sources are generally dynamic non-stationary signals, and their variances are unknown, the scaling problem is solved using the minimal distortion principle method [64] to adjust the learned separating (unmixing) filter matrix. On completion of the learning algorithm, the learned separating filter matrix is an arbitrary scaled version of the exact one, which is given as:

$$\mathbf{W}^{k} = \mathbf{D}^{k}\mathbf{H}^{-1(k)} \tag{2.6.22}$$

where $\mathbf{D}^{(k)}$ is an arbitrary diagonal matrix.

Therefore, the separating filter matrix can be updated to achieve reasonable scales by replacing $\mathbf{W}^{(k)}$ as:

$$\mathbf{W}^{(k)} = diag(\mathbf{W}^{-1(k)})\mathbf{W}^{(k)} \tag{2.6.23}$$

After solving the scaling problem, finally the separated sources are estimated in the frequency domain. Then, an IDFT is performed and overlap added to reconstruct the time domain signal. A fast fixed-point version of the IVA algorithm that uses the Newton method as a learning algorithm is introduced in the next section.

### 2.6.5 Fast Fixed-Point IVA Algorithm

The fast fixed-point IVA (FastIVA) algorithm is a fast converging version of the IVA method, as it adopts the Newton's method during the learning process [43]. The Newton's method [44] is a second order learning algorithm that converges quadratically and does not require a learning rate [65]. The FastIVA algorithm uses the following contrast function to model the independence between the sources [43]:

$$C_{FastIVA} = \sum_{i=1}^{N} \Big[ E\Big[F\Big(\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2\Big)\Big] - \sum_{k=1}^{K} \lambda_i{}^{(k)} \Big( (\mathbf{w}_i{}^{(k)})^H \mathbf{w}_i{}^{(k)} - 1 \Big) \Big] \tag{2.6.24}$$

where $\lambda_i$ denotes the Langrange multiplier, $\hat{s}_i^{(k)} = (\mathbf{w}_i{}^{(k)})^H \mathbf{x}^{(k)}$ and $\mathbf{w}_i^H$ denotes the $i$th row of the complete unmixing matrix $\mathbf{W}^{(k)}$. $F(.)$ is the nonlinear function, which can take different forms [43]. The above contrast function is a multivariate function which can retain the dependency within the source vectors and it can minimise independence between the sources vectors in all frequency bins. In this algorithm, the Newton's method is applied to the contrast function using the quadratic Taylor polynomial around $\mathbf{w}_o$ in the complex variable notation [43]:

$$\begin{aligned} f(\mathbf{w}) \approx & f(\mathbf{w}_o) + \frac{\partial f(\mathbf{w}_o)}{\partial \mathbf{w}^T}(\mathbf{w} - \mathbf{w}_o) \\ & + \frac{\partial f(\mathbf{w}_o)}{\partial \mathbf{w}^H}(\mathbf{w} - \mathbf{w}_o)^* \\ & + \frac{1}{2}(\mathbf{w} - \mathbf{w}_o)^T \frac{\partial^2 f(\mathbf{w}_o)}{\partial \mathbf{w} \partial \mathbf{w}^T}(\mathbf{w} - \mathbf{w}_o) \\ & + \frac{1}{2}(\mathbf{w} - \mathbf{w}_o)^H \frac{\partial^2 f(\mathbf{w}_o)}{\partial \mathbf{w}^* \partial \mathbf{w}^H}(\mathbf{w} - \mathbf{w}_o)^* \\ & + (\mathbf{w} - \mathbf{w}_o)^H \frac{\partial^2 f(\mathbf{w}_o)}{\partial \mathbf{w}^* \partial \mathbf{w}^T}(\mathbf{w} - \mathbf{w}_o) \end{aligned} \tag{2.6.25}$$

The term $\mathbf{w}$ in Equation (2.6.25) is replaced with $\mathbf{w}_i^{(k)}$ and $f(\mathbf{w}_i^{(k)})$ is set to the summation term of the contrast function in Equation (2.6.24):

$$f(\mathbf{w}_i^{(k)}) = E\Big[F\Big(\sum_{k'=1}^{K} |\hat{s}_i^{(k')}|^2\Big)\Big] - \sum_{k'=1}^{K} \lambda_i^{(k')}\Big((\mathbf{w}_i^{(k')})^H \mathbf{w}_i^{(k')} - 1\Big) \tag{2.6.26}$$

where $\hat{s}_i^{(k')} = ({\mathbf{w}_i}^{(k')})^H \mathbf{x}^{(k)}$. The function $f(\mathbf{w}_i^{(k)})$ will be optimised when the gradient $\partial f({\mathbf{w}_i}^{(k)})/\partial(\mathbf{w}_i^{(k)})^*$ is set to zero. From Equation (2.6.25):

$$\begin{aligned}\frac{\partial f(\mathbf{w}_i^{(k)})}{\partial(\mathbf{w}_i^{(k)})^*} \approx & \frac{\partial f(\mathbf{w}_{i,o}^{(k)})}{\partial(\mathbf{w}_i^{(k)})^*} + \frac{\partial^2 f(\mathbf{w}_{i,o})}{\partial({\mathbf{w}_i}^{(k)})^* \partial(\mathbf{w}_i^{(k)})^T}(\mathbf{w}_i^{(k)} - \mathbf{w}_{i,o}^{(k)}) \\ & + \frac{\partial^2 f(\mathbf{w}_{i,o})}{\partial(\mathbf{w}_i^{(k)})^* \partial(\mathbf{w}_i^{(k)})^H}\big({\mathbf{w}_i}^{(k)} - \mathbf{w}_{i,o}^{(k)}\big)^* \equiv \mathbf{0}\end{aligned} \tag{2.6.27}$$

The derivative terms in Equation (2.6.27) are written as:

$$\frac{\partial f(\mathbf{w}_{i,o}^{(k)})}{\partial(\mathbf{w}_i^{(k)})^*} = E\bigg[\big(\hat{s}_{i,o}^{(k)}\big)^* F'\bigg(\sum_{k'=1}^{K} |\hat{s}_{i,o}^{(k')}|^2\bigg)\mathbf{x}^{(k)}\bigg] - \lambda_i^{(k)} \mathbf{w}_{i,o}^{(k)} \tag{2.6.28}$$

By keeping the observations $\mathbf{x}^{(k)}$ to be zero mean and white such that $(E[\mathbf{x}^{(k)}\big(\mathbf{x}^{(k)}\big)^H] = \mathbf{I})$ and assuming complex circular symmetry in the source vectors such that $(E[\mathbf{x}^{(k)}\big(\mathbf{x}^{(k)}\big)^T] = \mathbf{0})$, the second derivatives can be obtained:

$$\begin{aligned}&\frac{\partial^2 f(\mathbf{w}_{i,o}^{(k)})}{\partial(\mathbf{w}_i^{(k)})^* \partial(\mathbf{w}_i^{(k)})^T} = \\ &E\bigg[\bigg(F'\bigg(\sum_{k'=1}^{K} |\hat{s}_{i,o}^{(k')}|^2\bigg) + |\hat{s}_{i,o}^{(k)}|^2 F''\bigg(\sum_{k'=1}^{K} |\hat{s}_{i,o}^{(k')}|^2\bigg)\bigg)\mathbf{x}^{(k)}\big(\mathbf{x}^{(k)}\big)^H\bigg] - \lambda_i^{(k)}\mathbf{I} \\ &\approx E\bigg[F'\bigg(\sum_{k'=1}^{K} |\hat{s}_{i,o}^{(k')}|^2\bigg) + |\hat{s}_{i,o}^{(k)}|^2 F''\bigg(\sum_{k'=1}^{K} |\hat{s}_{i,o}^{(k)}|^2\bigg)\bigg] E[\mathbf{x}^{(k)}\big(\mathbf{x}^{(k)}\big)^H] - \lambda_i^{(k)}\mathbf{I} \\ &= \bigg(E\bigg[F'\bigg(\sum_{k'=1}^{K} |\hat{s}_{i,o}^{(k')}|^2\bigg) + |\hat{s}_{i,o}^{(k)}|^2 F''\bigg(\sum_{k'=1}^{K} |\hat{s}_{i,o}^{(k')}|^2\bigg)\bigg] - \lambda_i^{(k)}\bigg)\mathbf{I}\end{aligned} \tag{2.6.29}$$

$$\begin{aligned}
\frac{\partial^2 f(\mathbf{w}_{i,o}^{(k)})}{\partial(\mathbf{w}_i^{(k)})^* \partial(\mathbf{w}_i^{(k)})^H} &= E\Big[((\hat{s}_{i,o}^{(k)})^*)^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}^{(k')}|^2))\mathbf{x}^{(k)}(\mathbf{x}^{(k)})^T\Big] \\
&\approx E\Big[((\hat{s}_{i,o}^{(k)})^*)^2 F''(\sum_{k'=1}^{K} |\hat{s}_{i,o}^{(k')}|^2))\Big] E[\mathbf{x}^{(k)}(\mathbf{x}^{(k)})^T] \qquad (2.6.30) \\
&= \mathbf{0}
\end{aligned}$$

where $\hat{s}_{i,o}^{(k)} = (\mathbf{w}_{i,o}^{(k)})^H \mathbf{x}^{(k)}$, $F(.)'$ and $F(.)''$ are the first and second derivatives of $F(.)$, respectively. Approximation by separation of expectations [58] were applied to Equations (2.6.29) and (2.6.30).

From Equations (2.6.29) and (2.6.30), the Newton step Equation (2.6.27) is reduced to:

$$\mathbf{w}_i^{(k)} - \mathbf{w}_{i,o}^{(k)} = \frac{-1}{c(\mathbf{w}_{i,o})} . \frac{\partial(\mathbf{w}_{i,o})}{\partial(\mathbf{w}_i^{(k)})^*} \qquad (2.6.31)$$

where $c(\mathbf{w}_{i,o})$ is the constant term multiplied to matrix $\mathbf{I}$ in Equation (2.6.29). By substitution, the corresponding iterative algorithm becomes:

$$\mathbf{w}_i^{(k)} \leftarrow \mathbf{w}_{i,o}^{(k)} - \frac{E\Big[(\hat{s}_{i,o}^{(k)})^* F'(\sum_k |\hat{s}_{i,o}^{(k)}|^2)\Big] - \lambda_i^{(k)} \mathbf{w}_{i,o}^{(k)}}{E\Big[(F'(\sum_k |\hat{s}_{i,o}^{(k)}|^2) + |\hat{s}_{i,o}^{(k)}|^2 F''(\sum_k |\hat{s}_{i,o}^{(k)}|^2))\Big] - \lambda_i^{(k)}} \qquad (2.6.32)$$

where the Lagrange multiplier $\lambda_i^{(k)}$ is given by:

$$\lambda_i^{(k)} = E\Big[|\hat{s}_{i,o}^{(k)}|^2 F'(\sum_{k=1}^{K} |\hat{s}_{i,o}^{(k)}|^2)\Big] \qquad (2.6.33)$$

To reduce the computation, $\lambda_i^{(k)}$ can be removed by multiplying the numerator of Equation (2.6.32) on both sides of the equation and with normalisation, the learning rule is obtained as:

$$
\begin{aligned}
\mathbf{w}_i^{(k)} \leftarrow & E\Big[F'\Big(\sum_{k=1}^{K}|\hat{s}_{i,o}^{(k)}|^2\Big) + |\hat{s}_{i,o}^{(k)}|^2 F''\Big(\sum_{k=1}^{K}|\hat{s}_i^{(k)}|^2\Big)\Big]\mathbf{w}_i^{(k)} \\
& - E\Big[(\hat{s}_{i,o}^{(k)})^* F'\Big(\sum_{k=1}^{K}|\hat{s}_{i,o}^{(k)}|^2\Big)\mathbf{x}^{(k)}\Big]
\end{aligned}
\tag{2.6.34}
$$

The symmetric decorrelation scheme is employed to construct the unmixing matrix $\mathbf{W}^{(k)}$ for all sources:

$$
\mathbf{W}^{(k)} \leftarrow (\mathbf{W}^{(k)}(\mathbf{W}^{(k)})^H)^{-1/2}\mathbf{W}^{(k)}. \tag{2.6.35}
$$

The non-linear score function for the FastIVA algorithm is derived from the source prior selected to model the speech signals. The separation performance of the algorithm depends on the accuracy of the source prior model. A detailed discussion on the choice of the source prior and its affects upon the separation performance of the FastIVA can be found in Chapter 5. The super Gaussian distribution used by the original natural gradient IVA algorithm [20] can be used as a source prior for the FastIVA algorithm. Assuming zero mean and unity variance is considered as unity, the non-linear score functions can be derived as:

$$
F(\sum_{k'=1}^{K}|\hat{s}_i^{(k')}|^2) = \sqrt{\Big(\sum_{k'=1}^{K}|\hat{s}_i^{(k')}|^2\Big)} \tag{2.6.36}
$$

## 2.7 Summary

In this chapter background theory related to the convolutive blind source separation (CBSS) problem was introduced. First, the various BSS mixing models were discussed. Then the major SOS and HOS frequency domain BSS (FD-BSS) techniques for solving BSS were reviewed and their ambiguities were discussed. The Parra-Spence algorithm based on SOS FD-BSS was discussed. Two main HOS FD-BSS algorithms, namely the independent component analysis (ICA) and independent vector analysis (IVA), were explained in detail. Finally, the fast fixed point IVA (FastIVA) was introduced.

The data sets including the speech signals and room impulse responses as well as the performance measures used to evaluate the separation performance of the different BSS algorithms are discussed in the next chapter.

Chapter 3

# DATA SETS, IMPULSE RESPONSE MODELS AND EVALUATION CRITERIA

## 3.1 Introduction

The evaluation of speech blind source separation (BSS) algorithms requires an experimental setup which, commonly, includes speech source signals, acoustic room environments and separation performance criteria. In this chapter the datasets and techniques employed for source separation of convolutive speech mixtures, presented in this thesis, are outlined. Firstly, the dataset, from which the speech signals are obtained, is described. Then, the different room models, deployed, are discussed. Finally, the performance measures, used to evaluate and analyse the separation performance of the algorithms are explained.

## 3.2 TIMIT Acoustic-Phonetic Continuous Speech Corpus

The speech source signals used for all experiments throughout this thesis are obtained from the DARPA TIMIT (Texas Instruments (TI) and Massachusetts Institute of Technology (MIT)) Acoustic-Phonetic Continuous Speech Corpus [66]. The TIMIT corpus is a standard database of phonetically-balanced English speech signals. It is widely used to provide speech data for the development and evaluation of speech recognition systems, acoustic-phonetic studies and the evaluation of speech separation algorithms [67].

TIMIT contains recordings of 630 male and female speakers of eight major American English dialects. Ten phonetically rich sentences spoken by each speaker were recorded, giving a total of 6300 utterances. The utterances were recorded using a Sennheiser close-talking microphone and sampled at 16 kHz rate with 16-bit resolution speech waveforms of various lengths. However, they were downsampled to 8 kHz for all experiments in this thesis [68].

## 3.3 Room Impulse Responses

In real room environments, audio signals captured by acoustic sensors, are convolutive mixtures of the source signals. The convolution is due to the time delays and attenuation of the sound signals resulting from reflections in closed reverberant environments [64]. Such a mixing situation is generally modelled with room impulse responses (RIRs) from the sound sources to the sensors [69]. The degree of mixing depends on the reverberation time of the room and direct to reverberant ratio (DRR). The room reverberation time (RT) is the time period required for the energy of an impulse response to decay below a certain level in decibels (dB). $RT_{60}$ is a commonly used reverberation time which corresponds to a decay of the impulse response to 60 dB from its initial level [70]. Methods of measuring reverberation time and decay curves can be found in [70]. In this thesis, the RIRs used to evaluate various BSS algorithms have been obtained from three different models (All downsampled to 8 kHz in the experiments).

### 3.3.1 Image Source Method

The image source method (ISM) is a simulation method for small rooms based on an approximate image expansion for non-rigid-wall enclosures [71]. The model assumed is a simple rectangular room with a source-to-receiver impulse response calculated using a time domain image expansion method. The generated impulse responses are synthetic and thus are not deemed suitable for robust evaluation of acoustic BSS algorithms in real life environments. The uncertainties of the ISM are analysed in [72,73]. However, they are useful for comparative studies as they provide flexible acoustic environments i.e. different experimental setups can be realised by controlling some parameters. Figure 3.1 illustrates an example of a simulated room environment with dimensions $(7m \times 5m \times 2.75m)$ and $RT_{60}$ of 200 ms and the corresponding four room impulse responses.

### 3.3.2 Real Room Impulse Responses

Two types of real RIRs, termed binaural room impulse responses (BRIRs), that provide a robust evaluation of BSS algorithms in realistic scenarios have also been employed throughout the thesis.

#### 3.3.2.1 Binaural Room Impulse Responses (BRIRs) (Shinn-Cunningham)

These binaural room impulse responses (BRIRs) are real RIRs which were recorded in a classroom, with dimensions $(5m \times 9m \times 3m)$, using a dummy Knowles Electronics Manikin for Acoustic Research (KEMAR) to emulate a human head in a real acoustic environment [74]. The inter-ear distance on the KEMAR is 15 cm. The KEMAR was placed at four different locations (centre, back, ear, and corner) with the ears

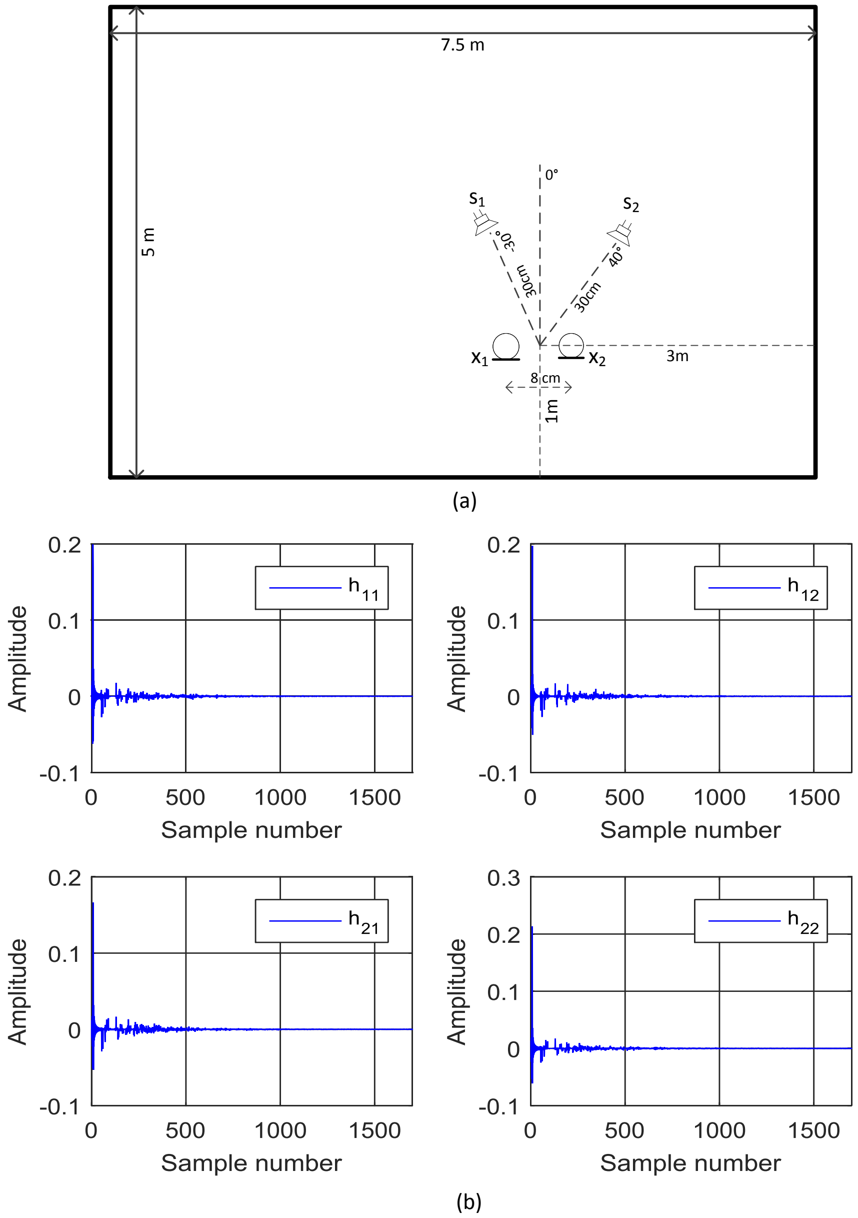


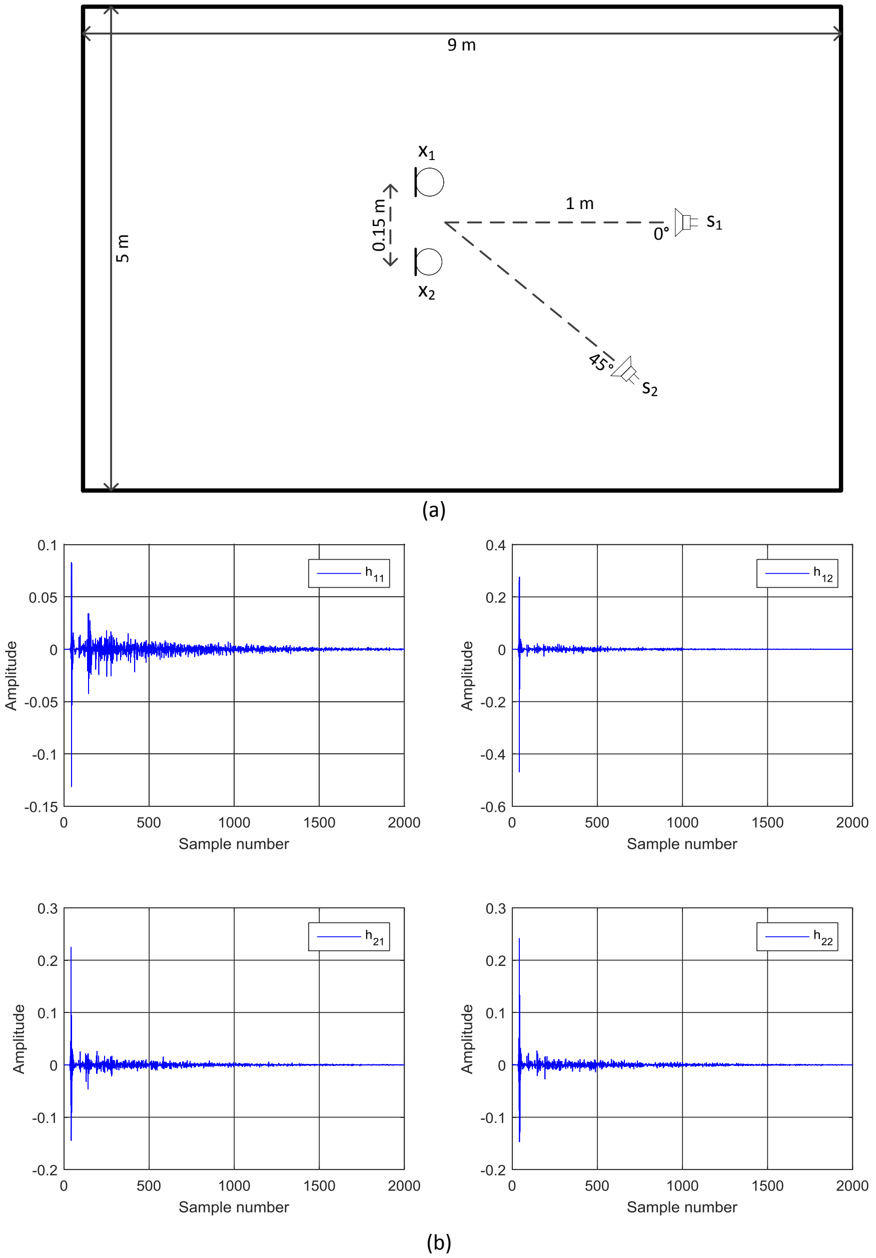
**Figure 3.1.** Simulated ISM room (a) Room environment showing the locations of sources and microphones. The heights of the sources and microphones are 1.5 m. (b) Impulse responses.

at a height of 1.5 m above the floor and the sound sources were placed at the same horizontal plane. For each head location, the BRIRs were measured for seven source azimuths $(0°, 15°, 30°, 45°, 60°, 75°$, and $90°)$ at three source distances (0.15 m, 0.40 m, and 1 m) relative to the centre point between the ears. All measurements were repeated three times, with equipment disassembled and reassembled between the measurements.

In all experiments in this thesis, only the centre location [2.5 m, 4.5 m, 1.5 m] of the KEMAR is considered with a measured reverberation time $RT_{60}$ of 565 ms and a sampling rate of 44.1 kHz. In order to increase reliability, the BRIRs were averaged over the three repeated measurements for each source location. An example of the room environment, with the head placed in the centre of the room, is illustrated in Figure 3.2(a). Source $s_1$ is placed at 0° and source $s_2$ at 45° at a distance of 1 m from the centre of the head. The corresponding four room impulses are shown in Figure 3.2(b).

#### 3.3.2.2 Binaural Room Impulse Responses (BRIRs) (Hummersone)

These BRIRs were measured in real room environments exploiting different enclosure designs. They were recorded using a Cortex Instruments Mk.2 Head and Torso Simulator (HATS) to emulate a human head [75]. The sound sources were placed around the HATS on an arc in the median plane at the same height as the ears with a 1500 mm radius between ±90° and measurements were taken at 5° intervals. The BRIRs were recorded at 48 kHz sampling rate and also resampled to 16 kHz. A summary of the acoustical properties of each of the rooms is provided in Table 3.1. The different rooms are described next.



(a)

(b)

**Figure 3.2.** Example of BRIR (Shinn-Cunningham). (a) Room environment (b) Impulse responses.

**Table 3.1.** Room acoustical properties, including $RT_{60}$, Initial Time Delay Gap (ITDG), Direct-to-Reverberant Ratio (DRR) and clarity index $C_{te}$.

| Room | $RT_{60}$ (ms) | ITDG (ms) | DRR (dB) | $C_{te}$ (50 ms) (dB) |
|---|---|---|---|---|
| A | 320 | 8.72 | 6.09 | 16.5 |
| B | 470 | 9.66 | 5.31 | 11.4 |
| C | 680 | 11.9 | 8.82 | 17.4 |
| D | 890 | 21.6 | 6.12 | 9.43 |

#### Room A

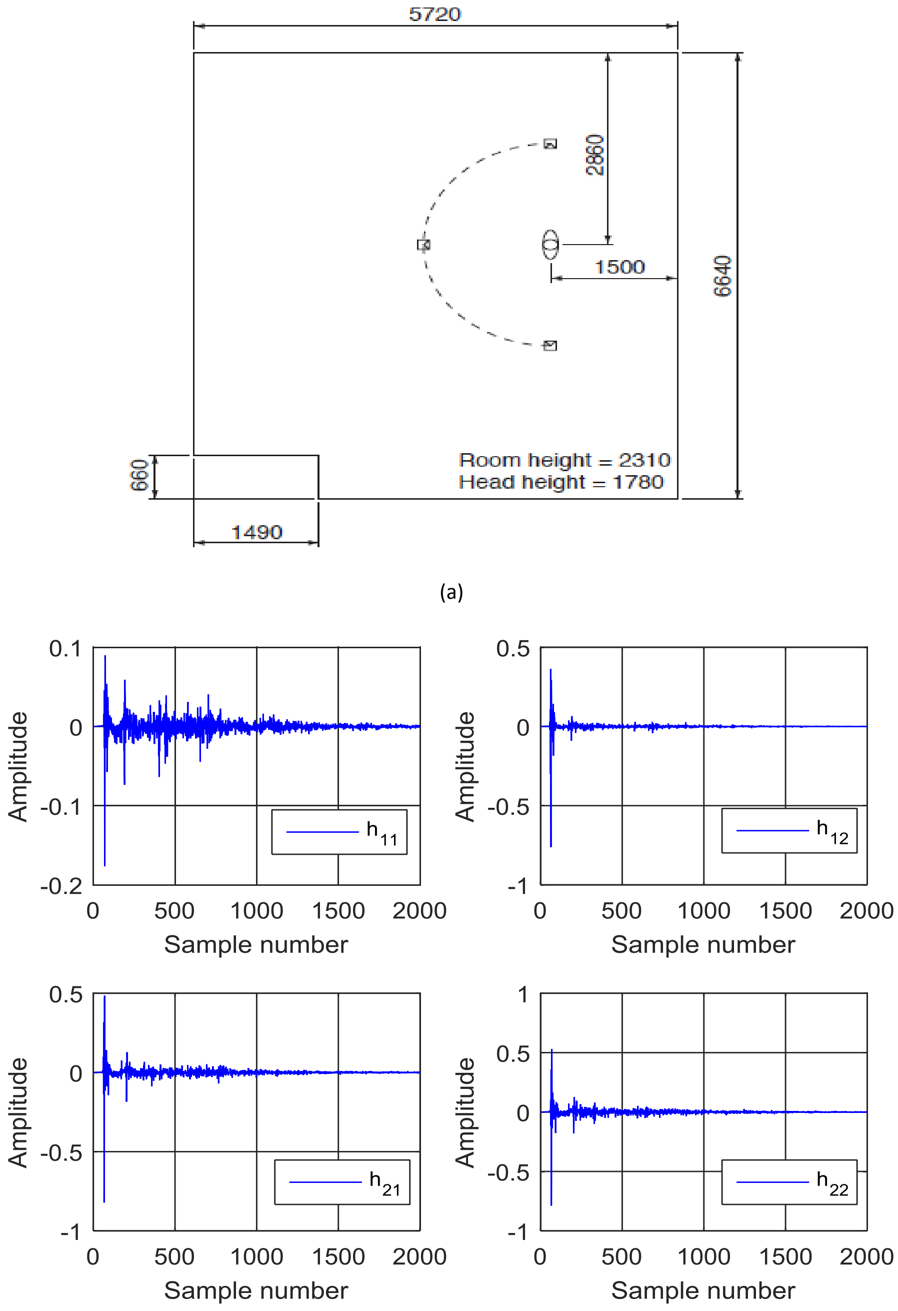
Room A is a typical medium-sized office that seats 8 people and has a small $RT_{60}$ of 320ms. The room layout and dimensions are given in Figure 3.3 along with an example of the room impulse responses where source $s_1$ is placed at 0° and source $s_2$ at 45°.

#### Room B

Room B is a medium-small class room with a relatively long $RT_{60}$ of 470 ms. The room layout and dimensions are given in Figure 3.4 along with an example of the room impulse responses where source $s_1$ is placed at 0° and source $s_2$ at 45°.

#### Room C

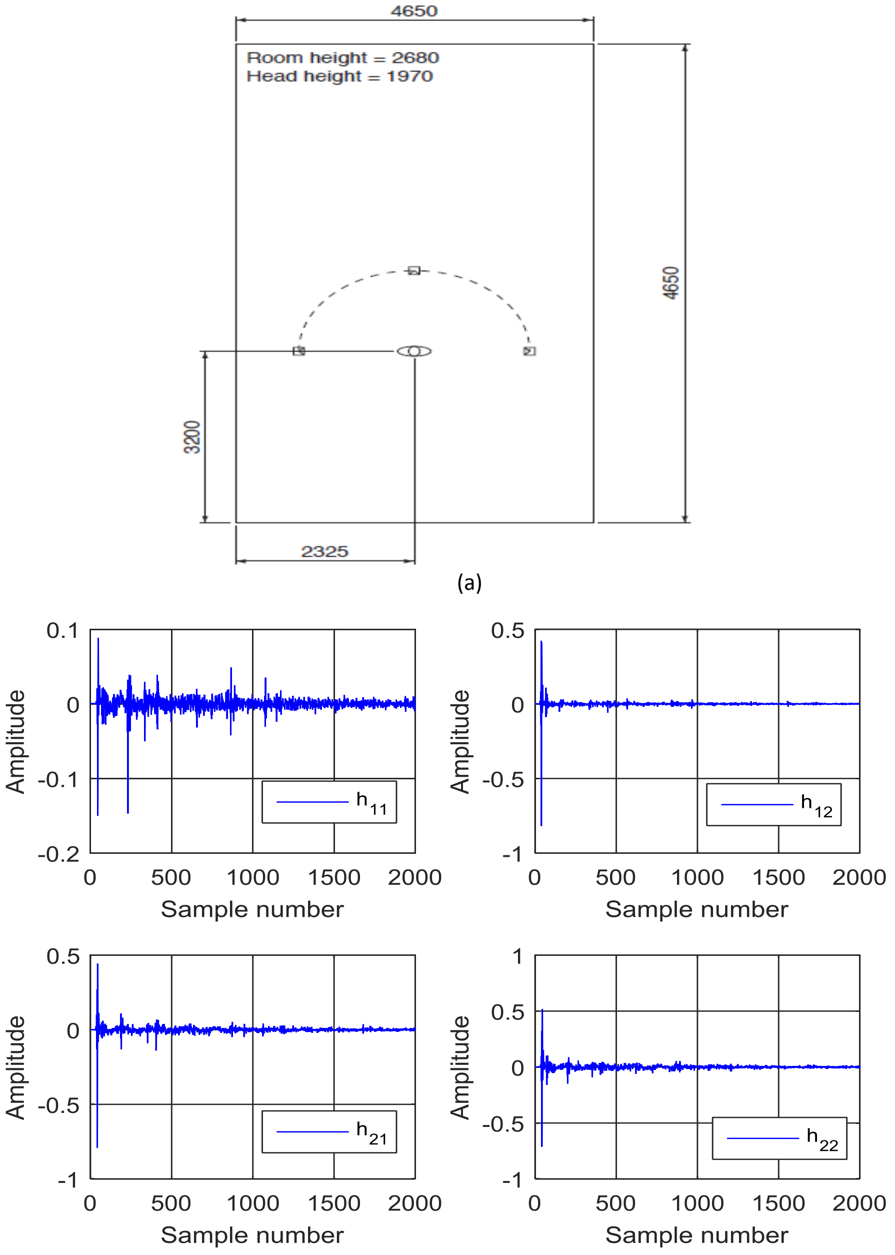
Room C is a large cinema-style lecture theatre that has 428 seating with longer $RT_{60}$ of 680 ms. The room layout and dimensions are given in Figure 3.5 along with an example of the room impulse responses where source $s_1$ is placed at 0° and source $s_2$ at 45°. The shaded area indicates banked seating and the room height is the height of the room at the HATS location.



(a)

(b)

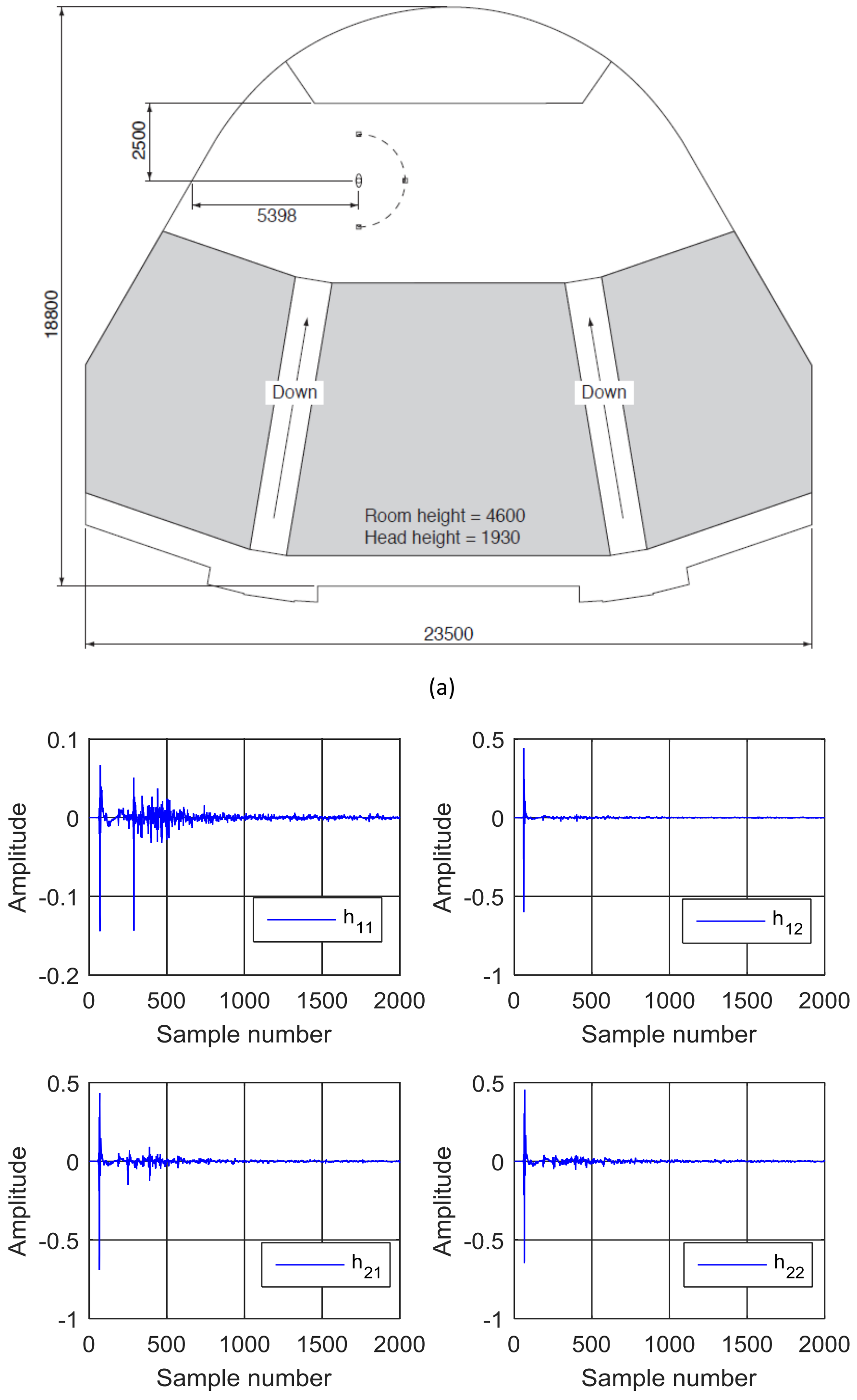
**Figure 3.3.** Room A (a) 2D Plan and HATS location [75]. (b) Impulse responses source $s_1$ at 0° and source $s_2$ at 45°.



(a)

(b)

**Figure 3.4.** Room B (a) 2D Plan and HATS location [75]. (b) Impulse responses source $s_1$ at 0° and source $s_2$ at 45°.



(a)

(b)

**Figure 3.5.** Room C (a) 2D Plan and HATS location [75]. (b) Impulse responses source $s_1$ at 0° and source $s_2$ at 45°.

#### Room D

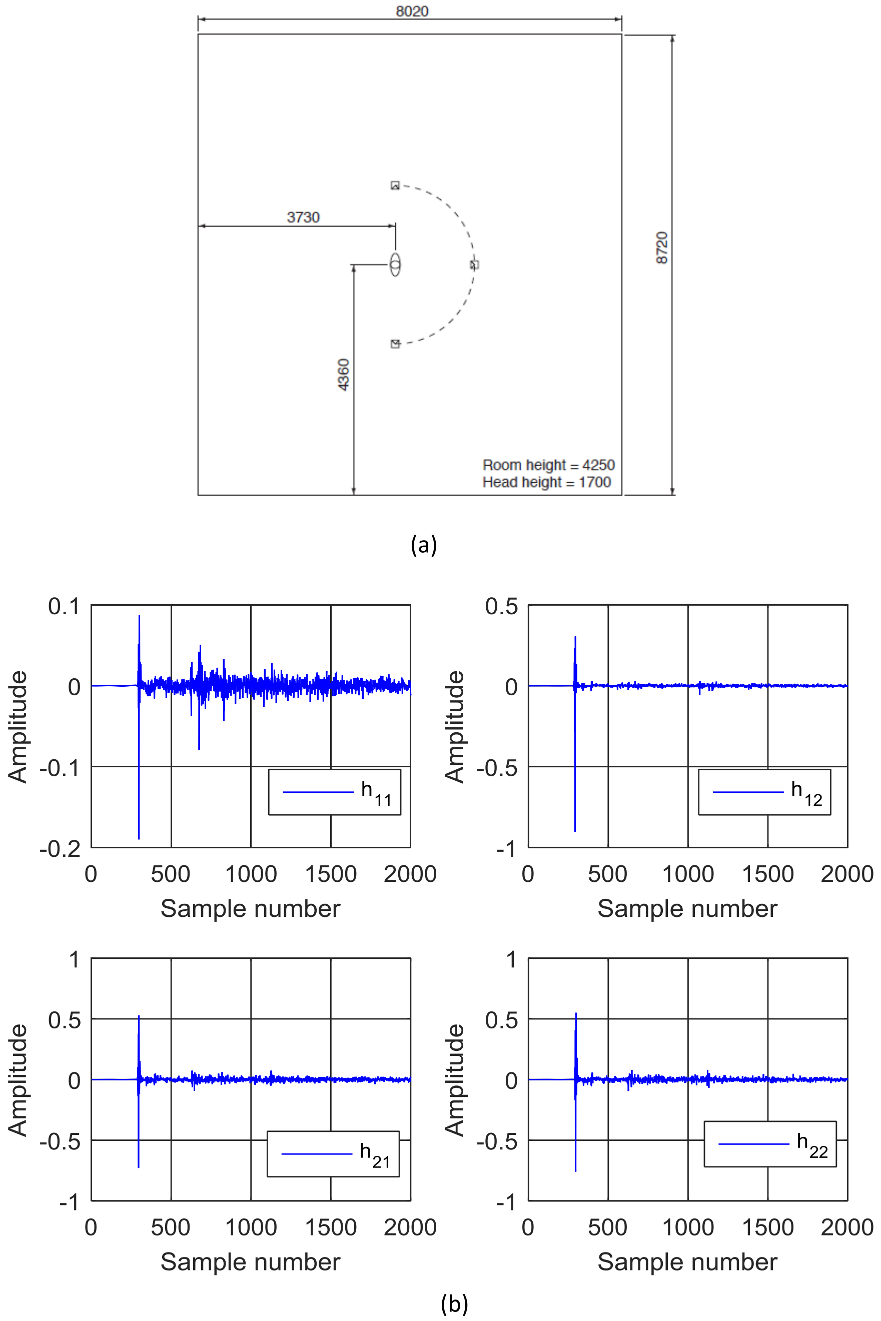
Room D is a typical medium-large sized seminar and presentation space with a very high ceiling and very long $RT_{60}$ of 890 ms. The room layout and dimensions are given in Figure 3.6 along with an example of the room impulse responses where source $s_1$ is placed at $0^\circ$ and source $s_2$ at $45^\circ$.

### 3.4 Performance Measures

The performance of a BSS algorithm can be evaluated using different measures that measure the quality of the separation process. The separation performance is evaluated in terms of the separation magnitude and the convergence performance. These performance measures can be classified as objective and subjective measures.

The objective evaluation measures compute numerically the quality of the estimation method. To calculate these measures, the original system parameters, the individual source signals and the mixing process are required, which are not available in real-life BSS procedure. In this case, subjective measures are used instead.

In this section three typical objective performance measures, used in audio BSS, are presented; the signal to interference ratio (SIR), signal to distortion ratio (SDR) and performance index (PI) as well as a subjective measure known as perceptual evaluation of speech quality (PESQ).



(a)

(b)

**Figure 3.6.** Room D (a) 2D Plan and HATS location [75]. (b) Impulse responses source $s_1$ at 0° and source $s_2$ at 45°.

### 3.4.1 Signal to Interference Ratio (SIR) and Signal to Distortion Ratio (SDR)

The source to interference ratio (SIR) and source to distortion ratio (SDR) are two measures, provided by the SiSec toolbox [76], to evaluate the separation performance of BSS algorithms. The performance measures are computed for each estimated source $\hat{s}_i$ by comparing it to a true source $s_i$ of $N$ sources from $M$ mixtures. The estimated source $\hat{s}_i$ may be compared with all the sources $(\hat{s}_{i'})_{1\leq i'\leq N}$ and the true source may be selected as the one that gives the best results. The estimated source $\hat{s}_i$ is decomposed based on the following model:

$$\hat{s}_i = s_{target} + e_{interf} + e_{noise} + e_{artif} \tag{3.4.1}$$

where $s_{target}$ is the part of the estimated source $\hat{s}_i$ which represents a distorted version of the original source $s_i$, $e_{interf}$ is the interference introduced by the other sources, $e_{noise}$ is the noise error term and $e_{artif}$ is the artifacts error term which represents unknown errors, such as distortion caused by the separation algorithm. These four terms should represent the part of $\hat{s}_i$ perceived as coming from the source of interest $s_i$, from other undesirable sources $(s_{i'})_{i'\neq i}$, from sensor noises $(n_j)_{1\leq j\leq M}$ and from other artifacts. The decomposition terms of $\hat{s}_i$ in Equation (3.4.1) are determined as:

$$s_{target} = \mathbf{P}_{s_i}\hat{s}_i \tag{3.4.2}$$

$$e_{interf} = \mathbf{P}_{\mathbf{s}}\hat{s}_i - \mathbf{P}_{s_i}\hat{s}_i \tag{3.4.3}$$

$$e_{noise} = \mathbf{P}_{\mathbf{s},\mathbf{n}}\hat{s}_i - \mathbf{P}_{\mathbf{s}}\hat{s}_i \tag{3.4.4}$$

$$e_{artif} = \hat{s}_i - \mathbf{P}_{\mathbf{s},\mathbf{n}}\hat{s}_i \tag{3.4.5}$$

where $\mathbf{P}_{\mathbf{x}}$ denotes a matrix of orthogonal projection onto a subspace spanned by the vectors $\mathbf{x}$ denoted by $\Pi\{\mathbf{x}\}$. The above three orthogonal projectors are defined as:

$$\mathbf{P}_{s_i} = \Pi\{s_i\} \tag{3.4.6}$$

$$\mathbf{P}_{\mathbf{s}} = \Pi\{(s_{i'})_{1\leq i'\leq N}\} \tag{3.4.7}$$

$$\mathbf{P}_{\mathbf{s},\mathbf{n}} = \Pi\{(s_{i'})_{1\leq i'\leq N}, (n_j)_{1\leq j\leq M}\} \tag{3.4.8}$$

The SDR and SIR are defined as numerical performance criteria by computing energy ratios of the estimated sources expressed in decibels (dB). The SIR takes into consideration only the interference introduced by the other sources $e_{interf}$ on the estimated source. It is defined as:

$$\text{SIR} = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf}||^2} \tag{3.4.9}$$

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

The SDR takes into consideration all three decomposition terms of estimated source $\hat{s}_i$; $e_{interf}$, $e_{noise}$ and $e_{artif}$. It is defined as:

$$\text{SDR} = 10\log_{10}\frac{||s_{target}||^2}{||e_{interf}+e_{noise}+e_{artif}||^2} \tag{3.4.10}$$

The values of both the SIR and SDR are directly proportional to the quality of source separation. The higher the value, the better the estimation. The SIR and SDR values at the observations are, normally, considered to be 0 dB based on the assumption that all the sources have identical variance at the microphones.

### 3.4.2 Performance Index

The performance index (PI) is a widely used performance measure in BSS evaluation. PI measures the quality of either the estimated separating matrix $\mathbf{W}$ or the estimated mixing matrix $\mathbf{H}$. It is calculated at each frequency bin and is based on the overall system matrix $\mathbf{G} = \mathbf{WH}$ which is insensitive to permutation and scaling ambiguities. The PI is defined as function of $\mathbf{G}$ as follows [77]:

$$\begin{aligned} PIG(\mathbf{G}) = &\left[\frac{1}{N}\sum_{i=1}^{n}\left(\sum_{j=1}^{m}\frac{|G_{ij}|}{max_j|G_{ij}|}-1\right)\right]+ \\ &\left[\frac{1}{M}\sum_{j=1}^{m}\left(\sum_{i=1}^{n}\frac{|G_{ij}|}{max_i|G_{ij}|}-1\right)\right] \end{aligned} \tag{3.4.11}$$

where $G_{ij}$ is the element at $i$th row and $j$th column of $\mathbf{G}$.

The lower bound value for PI is zero and the upper bound value depends on the normalisation factor. The lower the value of PI, the better the separation performance with PI=0 gives best separation performance.

#### 3.4.2.1 Permutation Measurement (PM)

The PI can measure the separation performance at each frequency bin, but it is insensitive to permutation. Thus, it cannot evaluate the permutation performance. The permutation measurement (PM) [78], which is sensitive to permutation, is used to evaluate the permutation performance. For a two-input two-output model, the PM is given as [79]:

$$PM = |G_{11}G_{22}| - |G_{12}G_{21}| \tag{3.4.12}$$

For a permutation free FDCBSS $PM > 0$.

### 3.4.3 Perceptual Evaluation of Speech Quality

Perceptual evaluation of speech quality (PESQ) is a subjective method of measuring speech quality used to evaluate the separation performance of speech BSS algorithm. PESQ was predominantly developed to model subjective tests commonly used in telecommunications to assess the voice quality by human beings [80]. Basically, PESQ predicts subjective Mean Opinion Scores (MOS) by comparing the estimated (output) speech signals with the original versions (input) of these speech signals.

To perform the measure, a group of listeners rate the quality of the speech signals by selecting one of five levels, ranging from 1(bad) to 5 (excellent), as shown in Table 3.2. Then, the arithmetic average of the assigned numbers is taken to represent the MOS. After the PESQ analysis, a score is given ranging from 0.5 to 4.5, as demonstrated by the diagram in Figure 3.7. A higher score means a better speech quality. Thus, 0.5 denotes a very poor separation performance and 4.5 an excellent separation performance [81].

**Table 3.2.** Speech quality scale.

| Quality of the speech | Score |
|---|---|
| Bad | 1 |
| Poor | 2 |
| Fair | 3 |
| Good | 4 |
| Excellent | 5 |

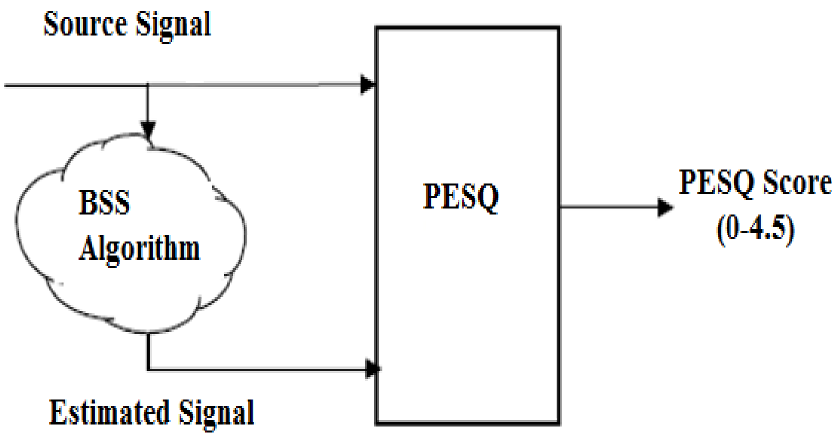


**Figure 3.7.** PESQ score measurement model

## 3.5 Summary

Different techniques and settings associated with speech source separation systems were outlined in this chapter. The groundwork to implement and evaluate a BSS system was laid including datasets, room impulse models and separation performance criteria. The various parameters involved in the mixing process of speech sources, when captured by microphones in enclosed settings, were examined and illustrated with examples. In addition the separation performance measures used to evaluate speech BSS were presented.

Following in this thesis is the first contribution chapter where the independent vector analysis (IVA) algorithm is applied and analysed with various source priors in real room environments, discussed in this chapter.

Chapter 4

# INDEPENDENT VECTOR ANALYSIS WITH VARIOUS SOURCE PRIORS, FOR APPLICATION IN REAL ROOM ENVIRONMENTS

## 4.1 Introduction

An application of blind source separation (BSS) is the separation of audio sources that have been mixed and then captured by multiple microphones in a real room environment. Speech and audio signal mixtures in a real reverberant environment are generally convolutive mixtures due to time delays resulting from sound propagation over space and the multi-path generated by reflections of sound off different objects in enclosed settings. While time-domain BSS algorithms have been developed to perform the task, they can be computationally expensive due to the multichannel convolution operations involved. Transforming the task into the frequency domain can simplify the convolutive BSS process, as convolution in time becomes multiplication in frequency [1]. With the application of short-time Fourier transforms (STFTs), convolutive mixtures in the time domain can be approximated as multiple instantaneous mixtures in the frequency domain. So, separation is performed in each frequency bin with a simple instantaneous separation matrix. The noise free FD-BSS model is described as:

$$\mathbf{x}^{(k)} = \mathbf{H}^{(k)}\mathbf{s}^{(k)} \tag{4.1.1}$$

$$\hat{\mathbf{s}}^{(k)} = \mathbf{W}^{(k)}\mathbf{x}^{(k)} \tag{4.1.2}$$

where $\mathbf{x}^{(k)} = [x_1^{(k)}, x_2^{(k)}, \cdots, x_M^{(k)}]^T$, $\mathbf{s}^{(k)} = [s_1^{(k)}, s_2^{(k)}, \cdots, s_N^{(k)}]^T$ and $\hat{\mathbf{s}}^{(k)} = [\hat{s}_1^{(k)}, \hat{s}_2^{(k)}, \cdots, \hat{s}_N^{(k)}]^T$ are the observed signal vector, the source signal vector and the estimated signal vector in the frequency domain at the $k$th frequency bin respectively and $(.)^T$ denotes the transpose operator. $\mathbf{H}^{(k)}$ is the $M \times N$ mixing matrix and $\mathbf{W}^{(k)}$is the $N \times M$ transfer function of the unmixing filter matrix at the $k$th frequency bin, $k = 1, 2, ..., K$, and $K$ is the number of frequency bins.

Independent component analysis (ICA) is one of the most popular early methods to solve the BSS problem of instantaneous mixtures [22]. ICA is a statistical method for extracting mutually independent sources from their mixtures, based on the assumption that source signals are statistically independent and thus learns the unmixing matrix by maximizing the independence among the estimated signals. Instantaneous mixing is the simplest mixing scenario, for which early BSS algorithms, including the standard ICA, were designed. Real-world acoustic environments lead to convolutive mixing of the sources when measured at the acoustic sensors, and the degree of mixing is significant when the reverberation time $RT_{60}$ of the room is high ($> 300ms$). Such algorithms will have limited practical applicability in speech separation problems unless additional effort is made on the system implementation.

In the standard ICA approach, the source signal prior is defined independently at each frequency bin. When the standard ICA methods are applied to FD-BSS, however, the well-known permutation problem arises [23–26]. That is, the grouping of separated frequency components which originate from the same source. This problem is due to the permutation indeterminacy of ICA and has resulted in extensive research [82, 83] that proposed techniques to address the permutation ambiguity. These are primarily based upon either higher level dependencies or additional geometric information about the system setup. An approach is smoothing the frequency-domain filter [82], limiting the filter length in the time domain [39] and for coloured signals the inter-frequency correlation between the signal envelopes was utilised [84,85]. Also, direction of arrival estimation was exploited [23, 60] and video tracking of sources was suggested [61]. Although these methods perform well under certain conditions, they may not provide good performance in general conditions.

Independent vector analysis (IVA) has been proposed for FD-BSS, as a new ICA formulation, to mitigate the permutation problem [20,86]. The method has been proven successful with its application to convolutive mixtures of speech signals. IVA is an extension of ICA from univariate components to multivariate components. It utilises the statistical independence among multivariate signals as well as the statistical inter-frequency dependency between the frequency bins of each multivariate signal. Hence, new source priors that measure the independence among multivariate signals have been proposed and the speech was modelled with multivariate probability density functions (PDF). A super Gaussian source prior was proposed in the original IVA method [20]. In [87], a chain type overlapped source prior was introduced. Another implementation of a multivariate super Gaussian source prior in the time domain was proposed in [88]. Also, a multivariate generalized Gaussian source prior was adopted [89].

This chapter presents extensive evaluations that compare the separation performance of the various FD-BSS techniques, namely the ICA and IVA algorithms [90]. The FastICA implementation [58] of the ICA algorithm is adopted. The IVA algorithm is evaluated with the original multivariate super Gaussian source prior proposed in [20]. In addition, the multivariate Student's t distribution is adopted as a source prior to improve the performance of the IVA method. The multivariate Student's t distribution has heavier tails as compared with the multivariate Laplacian distribution, which can be advantageous in modelling frequency domain non-stationary speech signals [92–94], as the distribution for a frequency domain speech signal is commonly a heavy tail distribution.

The algorithms are evaluated using simulated room impulse responses based on the image source method (ISM) [71] and, more importantly, in real room environments using binaural room impulse responses (BRIRs) [74]. Real recorded speech signals, from the TIMIT acoustic-phonetic continuous speech corpus [66], are used as the source signals. The separation performance is measured objectively by the signal to distortion ratio (SDR) [76] and subjectively by the perceptual evaluation of speech quality (PESQ) [81]. The detailed evaluations confirm significant improvement in separation performance of the IVA algorithm exploiting the multivariate Student's t source prior.

## 4.2 Independent Component Analysis (ICA)

In ICA, source signals are assumed to be statistically independent. If the sources are statistically independent, the joint probability density function $p(s_1, \cdots, s_N)$ equals the product of the marginal distributions of the sources:

$$p(s_1, \cdots, s_N) = \prod_{i=1}^{N} p(s_i) \tag{4.2.1}$$

where $p(s_i)$ is the marginal distribution of the i$th$ source.

Fundamentally, ICA relies upon a statistical criterion expressed in terms of a contrast function which requires to be either minimised or to be maximised as well as an optimisation technique to carry out the minimisation or maximisation of the contrast function [6]. One of the most popular procedures for contrast based ICA and instantaneous BSS is the FastICA algorithm [5, 58].

### 4.2.1 Fast Fixed-Point ICA Algorithm

In the FastICA algorithm, higher order statistics (HOS) are implicitly embedded into the algorithm by arbitrary non-linearities. In the one-unit version of FastICA, the contrast function is expressed as follows:

$$J_G(\mathbf{w}) = EG(|\mathbf{w}^H \mathbf{z}|^2) \tag{4.2.2}$$

where $\mathbf{z}$ is the whitened vector, $\mathbf{w}$ is an n-dimensional column vector of the separating matrix $\mathbf{W}$, $G$ is a smooth even function and $(\cdot)^H$ denotes Hermitian transpose.

It is desirable that the estimator given by the contrast function is robust against outliers. The slow growth of $G$ with the increase of its argument provides a more robust estimator. The function $G$ can take three different forms:

$$G_1(y) = \sqrt{a_1 + y} \tag{4.2.3}$$

$$G_2(y) = log(a_2 + y) \tag{4.2.4}$$

$$G_3(y) = \frac{1}{2}y^2 \tag{4.2.5}$$

where $a_1$ and $a_2$ are some small positive arbitrary constants deployed to avoid division by or the logarithm of zero. $G_1$ and $G_2$ grow slower than $G_3$:

$$G_1^{'}(y) = \frac{1}{2\sqrt{a_1 + y}} \tag{4.2.6}$$

$$G_2^{'}(y) = \frac{1}{a_2 + y} \tag{4.2.7}$$

$$G_3^{'}(y) = y \tag{4.2.8}$$

The FastICA finds a direction vector, i.e. a unit vector $\mathbf{w}$ such that $\mathbf{w}^H\mathbf{z}$ maximises non-Gaussianity. The optima of $EG(|\mathbf{w}^H\mathbf{z}|^2)$ under the constraint $E|\mathbf{w}^H\mathbf{z}|^2 = \|\mathbf{w}\|_2^2 = 1$ are obtained at points where:

$$\nabla EG(|\mathbf{w}^H\mathbf{z}|^2) - \beta\nabla|\mathbf{w}^H\mathbf{z}|^2 = 0 \tag{4.2.9}$$

where $\beta \in \mathbb{R}$, $\|.\|$ is the Euclidian norm and $\nabla$ is the gradient which is computed with respect to real and imaginary parts separately. The Newton method [44] is used to solve the equation and the fixed point algorithm for one unit can be written as [58]:

$$\mathbf{w}^{+} = E\{\mathbf{z}(\mathbf{w}^{H}\mathbf{z})^{*}G^{'}(|\mathbf{w}^{H}\mathbf{z}|^{2})\} - E\{G^{'}(|\mathbf{w}^{H}\mathbf{z}|^{2})\} + |\mathbf{w}^{H}\mathbf{z}|^{2}G^{''}(|\mathbf{w}^{H}\mathbf{z}|^{2})\mathbf{w} \tag{4.2.10}$$

where $G^{'}()$ is the derivative of $G()$, $G^{''}()$ is the derivative of $G^{'}()$ and $(.)^{*}$ is the complex conjugate.

$$\mathbf{w}_{new} = \frac{\mathbf{w}^{+}}{\|\mathbf{w}^{+}\|} \tag{4.2.11}$$

In order to prevent units from converging to the same maxima, the outputs are decorrelated after every iteration. This can be accomplished based on Gram-Schmidt-like decorrelation [5]; that is after estimating $\mathbf{w}_1, \cdots, \mathbf{w}_p$, during the estimation of unit $\mathbf{w}_{(p+1)}$ after every iteration step, subtract from $\mathbf{w}_{(p+1)}$ the projections of the previously estimated $p$ vectors and normalise:

$$\mathbf{w}_{(p+1)} = \mathbf{w}_{(p+1)} - \sum_{j=1}^{P} \mathbf{w}_j \mathbf{w}_j^{H} \mathbf{w}_{(p+1)} \tag{4.2.12}$$

$$\mathbf{w}_{(p+1)} = \frac{\mathbf{w}_{(p+1)}}{\|\mathbf{w}_{(p+1)}\|} \tag{4.2.13}$$

The Independent Vector Analysis (IVA) algorithm will be discussed in the next Section.

## 4.3 Independent Vector Analysis (IVA)

In the IVA algorithm, the sources are considered to be multidimensional random vectors, not just single variables as in ICA. Since the elements of a random vector are related to each other, elements within a source vector are dependent as well as correlated [20]. The algorithm allows independence between multivariate source signals represented as random vectors, and retains dependency between the source signals within each source vector. It also considers that multivariate signals have a multidimensional mixing linear model [20].

Compared with ICA methods, the interfrequency dependencies depend on a modified model for the source signal prior based on higher order dependencies across frequencies. The IVA method defines each source prior as a multivariate super-Gaussian distribution which is a simple extension of the independent Laplacian distribution. Thus, it can preserve the higher order interfrequency dependencies and structures of frequency components. It therefore mitigates the permutation problem and improves the separation performance of sources [20]. In addition to the dependency model which captures interfrequency dependencies, the IVA approach proposes a new cost function that measures the independence among multivariate signals with multivariate probability density functions (PDFs). It is an extension of mutual information between multivariate random variables. The learning algorithm for the parameters of the separating filters is derived by minimising the cost function [20].

### 4.3.1 Cost Function

The IVA algorithm deploys a multivariate cost function to separate multivariate sources from multivariate observations. A multivariate source prior attempts to remove dependency between the sources and retains the dependency between different frequency bins of each source. In IVA, the independence is measured by the Kullback-Leibler (KL) divergence between the exact joint probability density function of the estimated source vectors $p(\hat{\mathbf{s}}_1, \cdots, \hat{\mathbf{s}}_N)$ and the product of marginal probability density functions of the individual source vectors $\prod_{1=1}^{N} q(\hat{\mathbf{s}}_i)$ [20]:

$$C = KL\Big(p(\hat{\mathbf{s}}_1, \cdots, \hat{\mathbf{s}}_N) \| \prod_{1=1}^{N} q(\hat{\mathbf{s}}_i)\Big) \tag{4.3.1}$$

$$= const. - \sum_{k=1}^{K} \log |det(\mathbf{W}^{(k)})| - \sum_{i=1}^{N} E \log q(\hat{\mathbf{s}}) \tag{4.3.2}$$

where $det(.)$ is the matrix determinant operator, $|.|$ denotes the absolute value and $E[.]$ represents the statistical expectation operator. The random variables are multivariate.

The source prior $q(\hat{\mathbf{s}})$ in the cost function is a vector across all frequency bins. Each source is multivariate and the cost function would be minimised when the dependency between the source vectors is removed but the dependency between the components of each vector can be retained. The nonlinear score function for IVA algorithm will be discussed next.

### 4.3.2 Natural Gradient IVA

The gradients for the coefficients $(\Delta \mathrm{w}_{ij}^{(k)})$ are obtained by minimising the KL cost function using a gradient descent method [42], as follows [20]:

$$\Delta \mathrm{w}_{ij}^{(k)} = \sum_{l=1}^{N} (I_{il} - E\Big[\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}, \cdots, \hat{\mathbf{s}}_i^{(k)})\hat{\mathbf{s}}_l^{*(k)}\Big]\mathrm{w}_{ij}^{(k)} \tag{4.3.3}$$

where $I_{il}$ is the identity matrix ($I_{il} = 1$ when $i = l$) and ($I_{il} = 0$ when $i \neq l$). The nonlinear score function vector $\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}, \cdots, \hat{\mathbf{s}}_i^{(k)}) = [\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}), \cdots, \varphi^{(k)}(\hat{\mathbf{s}}_i^{(N)})]^T$ is defined as:

$$\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}, \cdots, \hat{\mathbf{s}}_i^{(k)}) = -\frac{\partial \log q(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)})}{\partial \hat{\mathbf{s}}_i^{(k)}} \tag{4.3.4}$$

The coefficients of the separating matrices $\mathrm{w}_{ij}^{(k)}$ can be updated by the batch update rule as [20]:

$$\mathrm{w}_{ij}^{(k)new} = \mathrm{w}_{ij}^{(k)old} + \eta \Delta \mathrm{w}_{ij}^{(k)} \tag{4.3.5}$$

where $\eta$ is the learning rate.

$\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}, \cdots, \hat{\mathbf{s}}_i^{(k)})$ is a multivariate score function which preserves the dependency structure across the frequency bins. The score function is obtained from the multivariate source prior to model the speech sources in the frequency domain. The following subsection introduces a multivariate source prior which uses a super Gaussian distribution proposed in the IVA method [20] to model the source vectors.

### 4.3.3 Multivariate super Gaussian Source Prior

The IVA algorithm proposed in [20] defines the source prior as a dependent multivariate super-Gaussian distribution in the form:

$$p(\mathbf{s}_i) = \alpha \exp\Bigg( -\sqrt{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i)} \Bigg) \tag{4.3.6}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are respectively the mean vector and covariance matrix of the *ith* source signal. Assuming zero mean vector $\boldsymbol{\mu}_i$ and identity covariance matrix $\Sigma_i$, equation (4.3.6) can be rewritten as:

$$p(\mathbf{s}_i) = \alpha \exp\left( - \sqrt{\sum_{k=1}^{K} \left| \frac{\hat{s}_i{}^{(k)}}{\sigma_i{}^{(k)}} \right|^2} \right) \tag{4.3.7}$$

where $\sigma_i{}^{(k)}$ is the standard deviation of the $i$th source at the $k$th frequency bin that determines the scale of each element of a source vector. Assuming unity standard deviation $\sigma_i^{(k)}$, the multivariate nonlinear score function used by the algorithm to extract the $i$th source at the $k$th frequency is obtained as [20]:

$$\varphi^{(k)}(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)}) = -\frac{\partial \log q(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)})}{\partial \hat{s}_i^{(k)}} = \frac{\partial \sqrt{\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2}}{\partial \hat{s}_i^{(k)}} \tag{4.3.8}$$

$$\varphi^{(k)}(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)}) = \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2}} \tag{4.3.9}$$

The function was the proposed form of a multivariate score function for separating source signals in the original IVA algorithm [20]. The following subsection introduces a multivariate source prior which uses the Student's t distribution to model the source vectors.

### 4.3.4 Multivariate Student's t Source Prior

It has been found that the t copula [95] is suitable for modeling the dependence structure for frequency domain speech signals [96]. Thus, a multivariate source prior based on the Student's distribution is proposed to model the speech sources in the IVA algorithm. The heavy tails of the Student's t distribution make it more fit to model the high amplitude data within the spectrum of non-stationary speech signals. The univariate Students t distribution takes the form:

$$p(s_i^{(k)}) = \frac{\Gamma(\frac{\nu+K}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1 + \frac{|s_i^{(k)}|^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{4.3.10}$$

where $\Gamma(.)$ is the Gamma function and $\nu$ is the degree of freedom. The plots in Figure 4.1 display the heavier tails of the univariate Student's t distribution, for all values of the parameter $\nu$, as compared with the super Gaussian distribution. The height of tails of the Student's t distribution is inversely proportional to the value degree of freedom parameter $\nu$. The lower the value of $\nu$ the heavier the tails. As $\nu$ increases the PDF approaches the Gaussian distribution PDF [96]. A plot of a bivariate Student's t distribution, with degrees of freedom ($\nu$) set to four, is given in Figure 4.2. The degrees of freedom parameter $\nu$ adjusts the variance and leptokurtic nature of the PDF [94].

The multivariate Student's t distribution, adopted as a source prior for the IVA algorithm, takes the form [97]:

$$p(\mathbf{s}_i) \propto \left(1 + \frac{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1}(\mathbf{s}_i - \boldsymbol{\mu}_i)}{\nu}\right)^{-\frac{\nu+K}{2}} \tag{4.3.11}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean and the covariance matrix, respectively and $\nu$ is the degrees of freedom parameter. The heavier tails of the distribution makes it suitable for certain types of speech signals [93].

The multivariate Student's t distribution can be shown to model the higher-order dependencies between frequency bins in the IVA approach. The marginal probability density function (PDF) is a univariate Student's t distribution. The product of the marginal probability density functions is not the same as the joint density function in Equation (4.3.11) when the covariance matrix is diagonal ($p(\hat{\mathbf{s}}_1, \cdots, \hat{\mathbf{s}}_N) \neq \prod_{1=1}^{N} p(\hat{\mathbf{s}}_i)$). Therefore, the variables of the multivariate Student's t

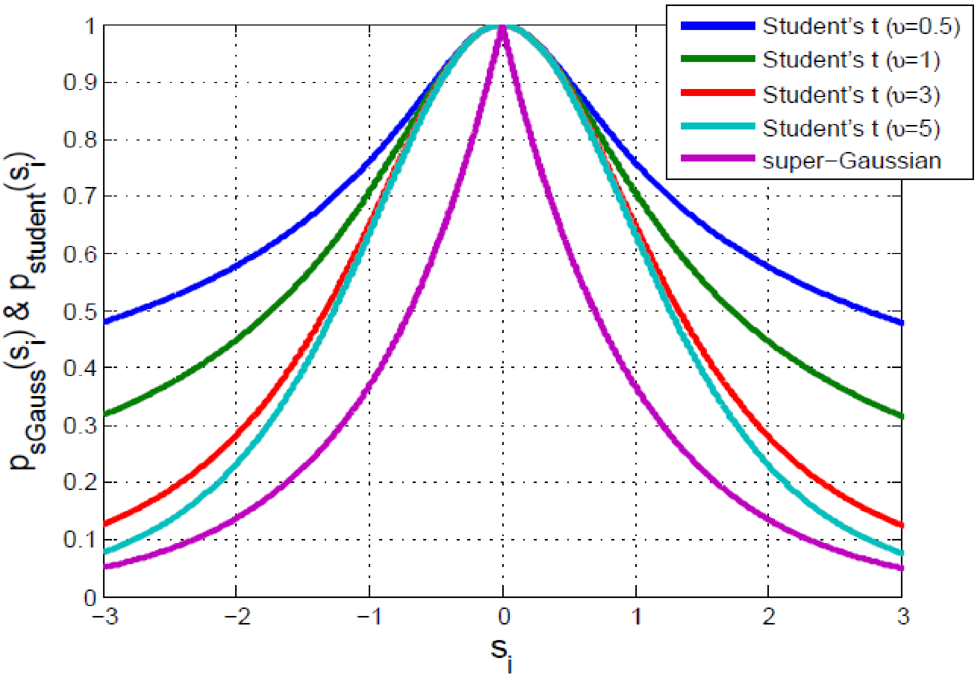


**Figure 4.1.** Univariate Student's t distribution as a function of the degrees of freedom parameter($\nu$) and univariate super-Gaussian distribution.

distribution are dependent and it can be used as a source prior for the IVA algorithm to retain the dependence across the frequency bins.

Assuming zero mean $\boldsymbol{\mu}_i$ and identity covariance matrix $\boldsymbol{\Sigma}_i$ due to the orthogonality of Fourier bases, Equation (4.3.11) can be rewritten as:

$$p(\mathbf{s}_i) \propto \left(1 + \frac{\sum_{k=1}^{K} |s_i^{(k)}|^2}{\nu}\right)^{-\frac{\nu+K}{2}} \tag{4.3.12}$$

The nonlinear multivariate score function for the multivariate Student's t distribution to extract the $i$th source at the $k$th frequency can be derived using the NG-IVA as:

$$\varphi^{(k)}(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)}) = -\frac{\partial \log q(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)})}{\partial \hat{s}_i^{(k)}} \tag{4.3.13}$$

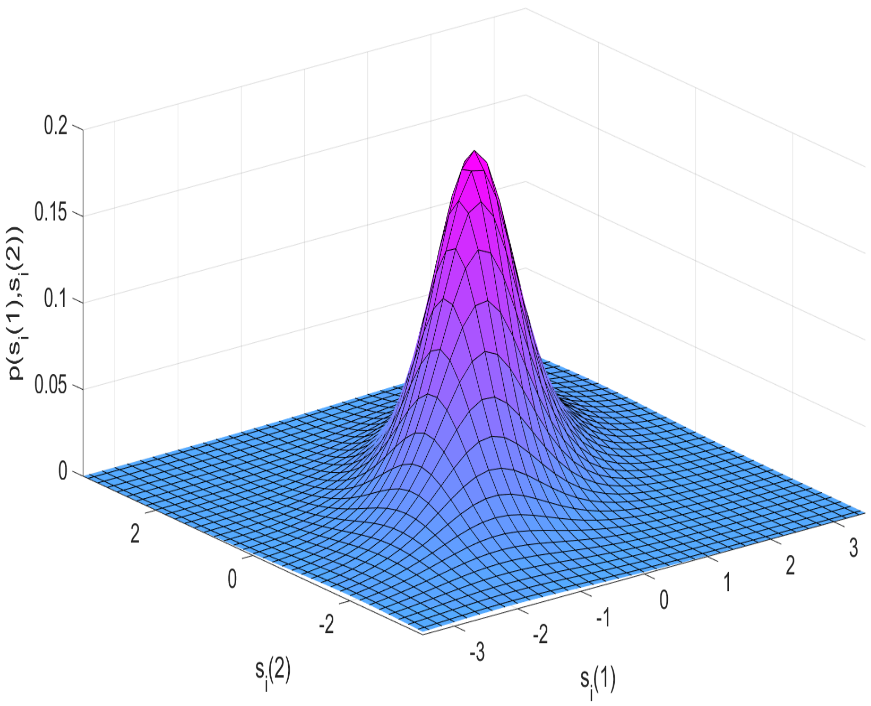


**Figure 4.2.** Bivariate Student's t distribution with degrees of freedom ($\nu = 4$).

$$\varphi^{(k)}(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)}) = -\frac{\partial \log \left(1 + \frac{\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2}{\nu}\right)^{-\frac{\nu+K}{2}}}{\partial \hat{s}_i^{(k)}} \tag{4.3.14}$$

$$\varphi^{(k)}(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)}) = \frac{\nu + K}{\nu} \frac{\hat{s}_i^{(k)}}{1 + (\frac{1}{\nu}) \sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2} \tag{4.3.15}$$

The constant $\frac{\nu+K}{\nu}$ can be absorbed by the learning rate $\eta$ in the update equation. A normalised score function is:

$$\varphi^{(k)}(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)}) = \frac{\hat{s}_i^{(k)}}{1 + (\frac{1}{\nu}) \sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2} \tag{4.3.16}$$

The separation performance of the different algorithms including the selection of the degree of freedom for the Student's t source prior will be discussed in the results section.

## 4.4 Experimental Results

In this section, the separation performance of the different separation algorithms is evaluated. The FastICA algorithm and the IVA algorithm with the original super Gaussian source prior and the new multivariate Students t source prior are evaluated using simulated room impulse responses and real room impulse responses. The simulated room impulse responses (RIRs) are based on the image source method (ISM) [71], which are artificial and do not represent a real life room environment. They are however normally used for performance comparison purposes. The real room impulses response are called binaural room impulse responses (BRIRs), which were recorded in a real classroom environment with very high $RT_{60}$ of 565ms [74]. The following smoothing function is used for the FastICA because it grows at a slow rate that gives a more robust estimator [58]:

$$G(y) = \sqrt{a+y}, \qquad G'(y) = \frac{1}{2\sqrt{a+y}} \tag{4.4.1}$$

where $a = 0.1$.

Selecting the degrees of freedom $\nu$, for the Student's t distribution, is a challenging task, as the source prior is used to model the speech mixtures instead of the separate original speech signals. Increasing $\nu$ results in lighter tails of the distribution. As $\nu \to \infty$, the Student's t distribution approaches the Gaussian distribution. Therefore, the value of $\nu$ should not be set too high. The separation performance of the IVA algorithm using the Student's t source prior was tested with different values of $\nu$. The outcome of this empirical procedure was that the value four for $\nu$ produces the best separation performance.

Two different speech signals of length of approximately four seconds were chosen randomly from the TIMIT dataset [66] and convolved into two mixtures using both room impulse responses. These mixtures were then separated using the different algorithms. The separation performance of the algorithms was measured using the objective measure of signal to distortion ratio (SDR) [76] in decibels (dB) as well as the subjective measure of perceptual evaluation of speech quality (PESQ) [81]. The evaluation results of the IVA method with new Student's t source prior are compared with the original IVA method [20] with the super Gaussian source prior as well as the FastICA method [58]. To improve the reliability of results, the SDR values at each source position were averaged over five speech mixtures. Results of the FastICA with post-processing are not included as the IVA algorithm does not require such processing.

### 4.4.1 Evaluation with the Image Source Method (ISM)

In these experiments, the room impulse responses are generated using the image source method (ISM) [71]. The $RT_{60}$ was set to 200ms. The different experiment parameters are given in Table 4.1. The sources were moved to six different positions around the room and five different sets of speech signals were used for the evaluation at each position.

The speech mixtures were created using the ISM RIRs and then separated using the FastICA [58] and IVA [20] algorithms with both source priors. The results of the IVA with the Student's t source prior are compared with the IVA method and the FastICA method. The separation performance of the three algorithms for both sources, expressed in SDR (dB), at the six source positions, is shown in Figure 4.3. The

**Table 4.1.** Experiment Parameters for ISM

| | |
|---|---|
| Sampling rate | 8kHz |
| STFT frame length | 1024 |
| Degrees of freedom | 4 |
| Reverberation time | 200 ms |
| Positions of Microphones | [3.42 2.60 1.50] m and [3.48 2.60 1.50] m |
| Room dimensions | 7m x 5m x 3m |
| Source signal duration | 4 s (TIMIT) |

separation performance (SDR) is averaged over the five speech signal mixture pairs at the six source positions. The SDR values are generally high because the room impulse responses are synthetic with a quite low $RT_{60}$.

The negative SDR values for the ICA algorithm demonstrate the inability of the algorithm to mitigate for the permutation problem without any pre-processing or post-processing. The results for the original IVA algorithm confirm the algorithm addressed the permutation ambiguity. The IVA algorithm with Student's t source prior showed a decent improvement over the original IVA algorithm at all six source positions. Table 4.2 shows the average SDR in dB of the two sources at the six positions. The average recorded separation performance improvement using the new Student's t source prior is approximately 0.78 dB compared with the original IVA method.

### 4.4.2 Evaluation with Binaural Room Impulse Responses (BRIRs)

In these experiments, the ICA algorithm and the IVA algorithm with both source priors are evaluated using the BRIRs, which were recorded using a dummy head to simulate the effect of a human

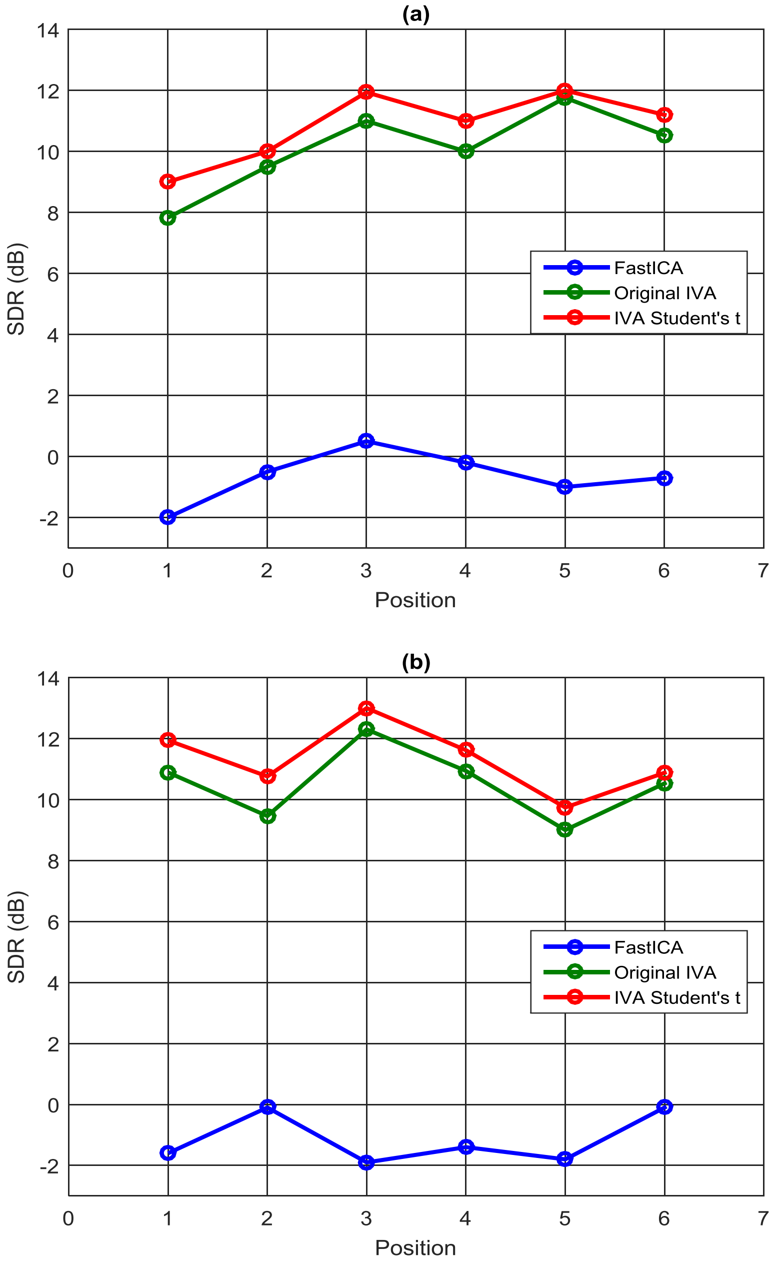


**Figure 4.3.** The graph shows the separation performance SDR (dB) at six different source positions using ISM. SDR was averaged over five mixtures. a) Source1 b) Source2. The performance of the ICA algorithm is poor due to the permutation probelm. The Student's t source prior consistently enhances the separation performance of the IVA algorithm.

**Table 4.2.** SDR (dB) values for the IVA algorithm with both source priors using ISM responses. The Student's t source prior for the IVA shows improvement for all mixture.

| Position | Original IVA | IVA Student's t | Improvement |
|---|---|---|---|
| Position-1 | 9.36 | 10.47 | 1.11 |
| Position-2 | 9.48 | 10.38 | 0.90 |
| Position-3 | 11.65 | 12.47 | 0.82 |
| Position-4 | 10.47 | 11.31 | 0.84 |
| Position-5 | 10.38 | 10.87 | 0.49 |
| Position-6 | 10.53 | 11.04 | 0.51 |

head in a real acoustic environment [74]. The BRIRs are real room recordings with very high $RT_{60}$ of 565ms. They provide realistic evaluation of the separation performance of BSS algorithms in highly reverberant environments. BRIRs were measured for 21 different relative source locations, consisting of all combinations of seven source azimuths $(0^\circ, 15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ, 90^\circ)$ and three source distances (0.15m, 0.40m, and 1m) from the centre point between the ears of the head. All measurements were repeated on three separate occasions, with equipment taken down and reassembled in between. In order to increase the reliability of the experiments, the room impulse responses were averaged over the three measurements.

The room layout and experimental setup are illustrated in Figure 4.4. The microphones were placed at the centre of the room with inter-microphone distance of 15cm. The sources were placed at 1 m from the centre of the microphones as this is approximately the critical distance. The first source $s_1$ was placed at a fixed position perpendicular to both microphones at angle $(0^\circ)$ and the second source $s_2$ was placed at all six different angles $(15^\circ$ to $90^\circ)$ relative to source $s_1$ in the room. Changing source positions represents speakers moving in the room which provide

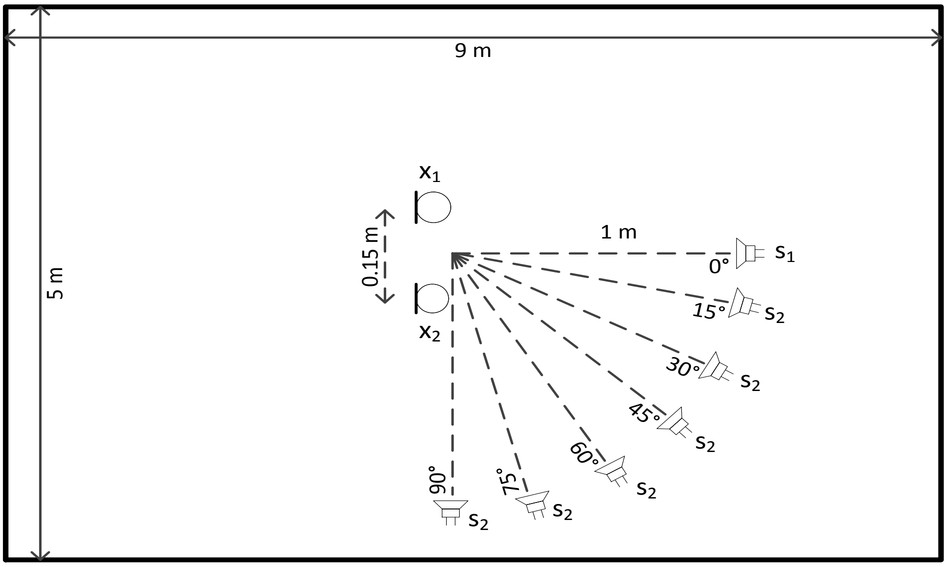


**Figure 4.4.** 2D plan of room and experimental setup for the BRIRs showing locations of sources and microphones.

thorough evaluation of separation performance of the algorithms. The summary of different parameters used in the experiments is provided in Table 4.3.

The speech mixtures were created by using the BRIRs with high $RT_{60}$ of 565ms and then separated using the three algorithms; the FastICA algorithm [58] and the IVA algorithm [20] with the Student's t source prior and the original super Gaussian source prior. The separation performance of the three algorithms for both sources, expressed in SDR (dB), at different angles, is shown in Figure 4.5. The graphs show the average SDR of five mixture signals at each angle for each source.

The results show the poor separation performance of the ICA algorithm where SDR is always negative as no pre-processing or post-processing applied which is generally required for this method to address the permutation problem. The original IVA algorithm with the multivariate super Gaussian source prior demonstrated its capability to

**Table 4.3.** Experiment parameters for BRIRs.

| | |
|---|---|
| Sampling rate | 8kHz |
| STFT frame length | 1024 |
| Degrees of freedom | 4 |
| Reverberation time | 565 ms |
| Room dimensions | 9 m x 5 m x 3.5 m |
| Source signal duration | 4 s (TIMIT) |
| Source distance | 1 m |

mitigate the permutation ambiguity. The IVA algorithm with Student's t source prior showed a considerable improvement on the original IVA algorithm at all six source positions. Table 4.4 shows the average SDR in dB of the two sources at the six angles. The average recorded separation performance improvement using the new Student's t source prior is approximately 1.31 dB compared with the original IVA method. It is worth noting the Student's t source prior provided better separation improvement with the real BRIRs than with the simulated ISM. This confirms the suitability of the Student's t distribution to model speech signals in real life scenarios.

**Table 4.4.** SDR (dB) values for both source priors for the IVA method using real BRIRs. The Student's t source prior shows improvement at all separation angles.

| Angle | Original IVA | IVA with Student's t | Improvement (dB) |
|---|---|---|---|
| 15° | 2.71 | 3.01 | 0.30 |
| 30° | 3.55 | 4.22 | 0.67 |
| 45° | 2.54 | 4.30 | 1.76 |
| 60° | 4.00 | 5.53 | 1.53 |
| 75° | 3.91 | 4.91 | 1.00 |
| 90° | 3.60 | 5.20 | 2.60 |

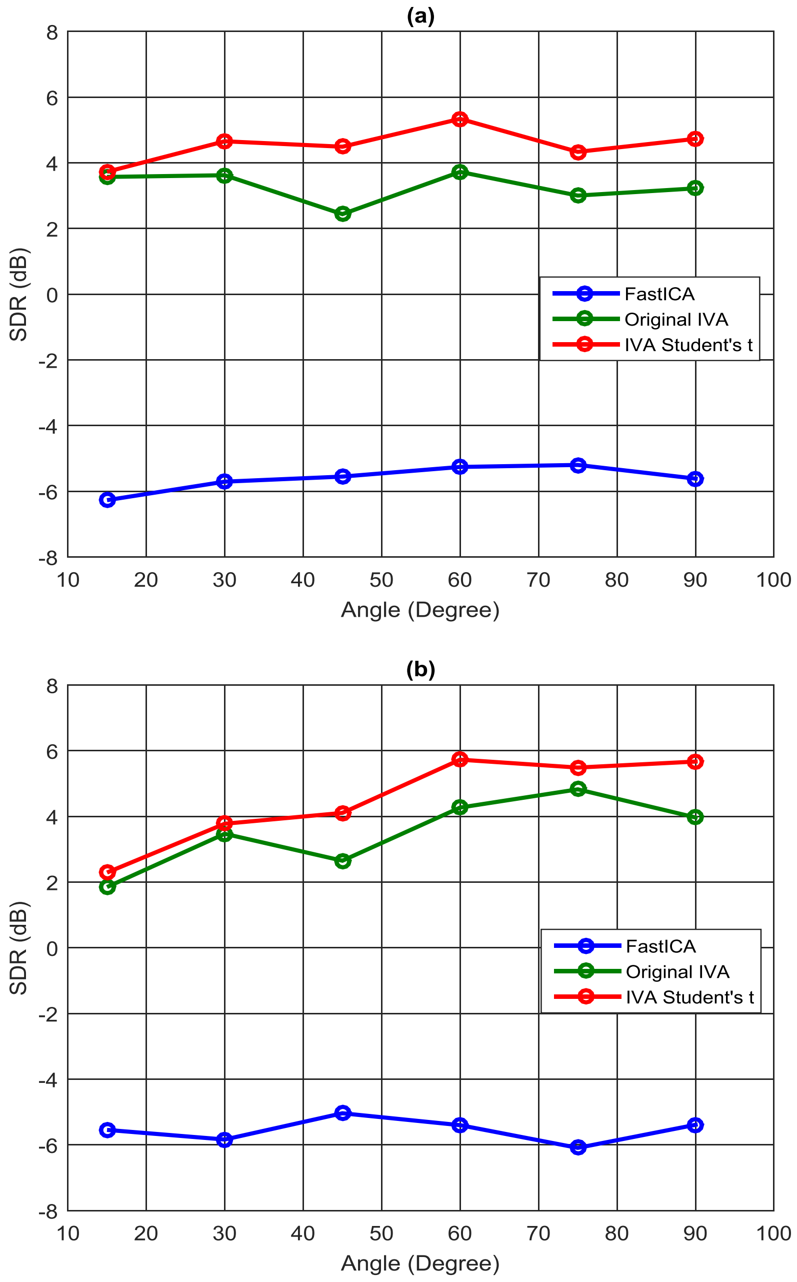


**Figure 4.5.** The graph shows the separation performance SDR (dB) at six different separation angles using real BRIRs. a) Source1 b) Source2. Results were averaged over five mixtures. The Student's t source prior enhances the separation performance of the IVA algorithm at all separation angles.

The subjective measure of perceptual evaluation of speech quality (PESQ) [81] is used to measure the separation performance of the algorithm using the BRIRs. The PESQ is a commonly used measure to examine the quality of the separated signal as it compares the original signals. A score is given between 0-4.5, 0 for very poor separation and 4.5 for excellent separation.

The signals were separated from mixtures using the IVA method with both source priors in the same settings as in Table 4.3. The PESQ scores for separated signals were measured as shown in Table 4.5. All the PESQ scores for each mixture are the average of PESQ scores of five speech mixtures at the six different source location azimuths varying from (15 ° to 90 °). The PESQ scores for the Student's t source prior is compared with the PESQ score of the estimated signals separated by the original IVA method in the same settings. This subjective study confirms the improved separation performance for the IVA method with the Student's source prior. The average separation performance improvement PESQ score is approximately 0.75 (35%).

**Table 4.5.** PESQ scores for the IVA algorithm with the two source priors. The Student's t source prior enhances the separation performance of the IVA algorithm at all source locations

| Angle | super Gaussian source prior | Student's t source prior |
|---|---|---|
| 15° | 1.72 | 2.13 |
| 30° | 2.14 | 2.65 |
| 45° | 1.65 | 2.71 |
| 60° | 2.42 | 3.15 |
| 75° | 2.35 | 2.93 |
| 90° | 2.19 | 3.32 |

## 4.5 Summary

In this chapter, the separation performance of the ICA and IVA techniques for FD-BSS was evaluated. The heart of the IVA method is the multivariate source prior used to model the speech signals because the non-linear score function used to retain the inter-frequency dependency is obtained from the PDF of the source prior. A new multivariate Student's t source prior was introduced for the IVA algorithm. The multivariate Student's t distribution was proven to better model the spectrum of speech signals. The tails of the Student's t distribution can be tuned to closely match the generally heavy tail distribution of the frequency domain speech signals due to the high amplitude data points. Real recorded speech signals and real room environments were used to evaluate the performance of the various algorithms. The experimental results in the highly reverberant real room environments, confirm that the proposed Student's t source prior consistently improves the separation performance of the IVA algorithm. Also, the results demonstrated the limitation of the ICA algorithm in CBSS speech separation applications without additional algorithmic scheme in order to correct the serious permutation ambiguity. Due to the poor performance of the ICA method for convolutive mixtures, it is not considered for the remainder of the thesis.

In the next chapter, a new energy driven multivariate mixed source prior with clique based dependency structure for the IVA algorithm is introduced. The proposed source prior is a mixture of the original multivariate super Gaussian distribution and the multivariate Student's t distribution. The Student's t distribution is used to model the high amplitude and the original super Gaussian distribution to model the lower amplitude components of the speech signal. The mixing ratio is adjusted according to the energy of the observed mixtures.

Chapter 5

# ENERGY DRIVEN MIXED SOURCE PRIOR FOR THE INDEPENDENT VECTOR ANALYSIS ALGORITHM

## 5.1 Introduction

The independent vector analysis algorithm (IVA) is a frequency domain technique that solves, algorithmically, the permutation problem in blind source separation (BSS) by preserving the dependency within each source vector. The separation performance of the IVA algorithm relies on the multivariate source prior adopted to model the sources. The model is used to derive the nonlinear score function that retains the dependency between different frequency bins [21]. Various statistical models to represent the statistical dependence within the IVA method have been proposed [88, 89, 91]. Statistical models that can enhance the dependency structure within each source vector would improve the separation performance of the IVA method.

In this chapter, a new enhanced multivariate source prior for the IVA algorithm is introduced. The proposed source prior is a mixture of two distributions, instead of a single distribution; namely the original multivariate super Gaussian distribution as in [20] and the multivariate Student's t distribution. Human speech is highly random in nature and can have variable amplitude components [77]. The Student's t distribution is a super Gaussian distribution with heavier tails which is proven efficient in modelling certain types of speech signals [94]. In the proposed source prior, the Student's t distribution models the high amplitude components of the speech signal [94] and the original super Gaussian distribution is used to model the lower amplitude components. In order for the mixed source prior to adapt to different types of speech signals, it is empowered with an energy driven scheme that adjusts the weight of each distribution according to the energy of the observed mixtures.

Moreover, the process exploits the frequency bins dependency structure to enhance the separation performance. The adjacent frequency bins generally have much stronger dependency as compared to distant frequency bins [87, 98]. Therefore, the fully connected frequency bin structure is decomposed into smaller cliques whilst retaining adequate overlap between adjacent cliques. The new energy driven mixed source prior with clique based dependency structure is evaluated in different real room environments. The results confirm that this approach consistently improves the separation performance of the IVA algorithm.

## 5.2 Source Prior for the IVA method

The IVA algorithm models independence between sources using a cost function which retains the inherent frequency dependency within each source vector, whilst removing the dependency among the sources [21]. The cost function is minimised when the vector sources are independent while preserving dependency within the components of each source vector. The learning algorithm is derived by minimising the cost function using the gradient descent algorithm. The nonlinear multivariate score function $\varphi^{(k)}$, which maintains dependency between frequency bins for source $\hat{s}_i$, is written in the general case as [20]:

$$\varphi^{(k)}({\hat{s}_i}^{(1)} \cdots {\hat{s}_i}^{(k)} \cdots \hat{s}_i^{(K)}) = -\frac{\partial \text{log} q({\hat{s}_i}^{(1)} \cdots {\hat{s}_i}^{(k)} \cdots \hat{s}_i^{(K)})}{\partial {\hat{s}_i}^{(k)}} \tag{5.2.1}$$

The particular nonlinear score function is based on the source prior selected to represent the frequency domain information of the sources. The performance of the IVA algorithm greatly depends on the multivariate model used as a source prior. Therefore, the selection of a suitable multivariate source prior plays a major role in the IVA algorithm.

### 5.2.1 The super Gaussian Source Prior

In the original IVA method [20], the source prior representing the interfrequency dependencies is a dependent multivariate super-Gaussian distribution in the form [20]:

$$p(\mathbf{s}_i) = \alpha \exp\left( - \sqrt{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i)} \right) \tag{5.2.2}$$

where $(.)^H$ denotes Hermitian transpose, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are respectively the mean vector and covariance matrix of the $i$th source signal at the $k$th frequency bin. Equation (5.2.2) shows there is a variance dependency between the frequency bins. In other words, the variance of one frequency component is directly proportional to the variance for other frequency components. By setting a zero mean vector and the covariance matrix to identity matrix because the frequency bins are uncorrelated due to the orthogonality of Fourier bases, the source prior of Equation (5.2.2) can be written as:

$$p(s_i) = \alpha \exp\left( -\sqrt{\sum_{k=1}^{K} \left| \frac{\hat{s}_i{}^{(k)}}{\sigma_i{}^{(k)}} \right|^2} \right) \tag{5.2.3}$$

where $\sigma_i{}^{(k)}$ is the standard deviation of the $i$th source at the $k$th frequency bin. By setting $\sigma_i{}^{(k)}$ to unity, the original nonlinear multivariate score function can be derived as [20]:

$$\varphi^{(k)}(\hat{s}_i{}^{(1)} \ldots \hat{s}_i{}^{(K)}) = \frac{\partial \sqrt{\sum_{k=1}^{K} \left| \hat{s}_i{}^{(k)} \right|^2}}{\partial \hat{s}_i{}^{(k)}} = \frac{\hat{s}_i{}^{(k)}}{\sqrt{\sum_{k=1}^{K} |\hat{s}_i{}^{(k)}|^2}} \tag{5.2.4}$$

Equation (5.2.4) represents the multivariate score function used as interdependency model for the original IVA method with the super Gaussian multivariate source prior. However, as discussed before, this score function is not unique. It depends on the types of sources selected to model the source signals. A source prior based on the Student's t distribution is presented in the next section.

### 5.2.2 The Student's t Source Prior

The multivariate Student's t distribution is well suited to model certain types of speech signals [93]. The multivariate Student's t distribution, when adopted as a source prior for the IVA algorithm, takes the form:

$$p(\mathbf{s}_i) \propto \left(1 + \frac{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i)}{\nu}\right)^{-\frac{\nu+K}{2}} \tag{5.2.5}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i^{-1}$ are the mean and the inverse covariance matrix, respectively and $\nu$ is the degrees of freedom parameter, which can tune the variance and the leptokurtic nature of the Student's t distribution [94]. The tails of the distribution becomes heavier when the degrees of freedom parameter $\nu$ decreases which makes it suitable for certain types of speech signals [93].

A score function for the original IVA method can be derived from the general IVA score function (5.2.1) 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 nonlinear multivariate score function is obtained as:

$$\varphi^{(k)}(\hat{s}_i{}^{(1)} \cdots \hat{s}_i^{(K)}) = \frac{\hat{s}_i^{(k)}}{1 + (\frac{1}{\nu}) \sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2} \tag{5.2.6}$$

The separation performance of the IVA method can potentially be improved by using a source prior that combines different distributions instead of a conventional single distribution source prior. Hence, a new mixed source prior that can adapt to different speech sources is proposed in detail in the next section.

### 5.2.3 Mixed Source Prior for the IVA Method

Speech signals are statistically nonstationary and their statistical properties can vary from a signal to another. Therefore, a single distribution may not be suitable to model all speech sources. To improve the separation performance of the IVA algorithm, a mixed source prior is proposed. The new multivariate source prior uses a mixture of the original super Gaussian and Student's t distributions. Owing to its heavy tails, the Student's t distribution is deployed to model the high amplitude information in the speech sources. The super Gaussian distribution is used to model the remaining information [99]. The new mixed multivariate source prior for the IVA algorithm takes the following form:

$$p(\mathbf{s}_i) = \lambda_d.f_{St} + (1-\lambda_d).f_G \tag{5.2.7}$$

where $f_{St}$ and $f_G$ are respectively the multivariate Student's t distribution and the multivariate super Gaussian distributions, $\lambda_d \ \epsilon \ [0,1]$ is a weighting parameter that determines the ratio of each distribution in the mixed source prior at frequency bin $k$. Replacing the multivariate Student's t by Equation (5.2.5) and the original multivariate super Gaussian by Equation (5.2.2), the new source prior is written as:

$$\begin{aligned} p(\mathbf{s}_i) =& \lambda_d \left(1 + \frac{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i)}{\nu}\right)^{-\frac{\nu+K}{2}} + \\ & (1-\lambda_d) \exp\left( - \sqrt{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i)}\right) \end{aligned} \tag{5.2.8}$$

The nonlinear score function for the IVA algorithm with the mixed source prior can be derived from Equation (5.2.8). Using Equations (5.2.4) and (5.2.6), the overall non linear score function for source $\hat{\mathbf{s}}_i$ can be written as:

$$\begin{aligned}\varphi^{(k)}(\hat{s}_i{}^{(1)}\cdots\hat{s}_i{}^{(K)}) \propto &\lambda_d\left(\frac{\hat{s}_i^{(k)}}{1+\frac{1}{\nu}\sum_{k=1}^{K}|\hat{s}_i^{(k)}|^2}\right)\\ &+(1-\lambda_d)\left(\frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{k=1}^{K}|\hat{s}_i^{(k)}|^2}}\right)\end{aligned} \tag{5.2.9}$$

The nonlinear score function for the IVA method using the mixed source prior is a multivariate function that can retain the inter-frequency dependency as all the frequency bins are accounted for during the learning process. The weight of each distribution in the source prior can be determined for particular speech signals by adjusting the value of $\lambda_d \,\epsilon\, [0, 1]$; $\lambda_d = 1$ yields a pure Student's t distribution and $\lambda_d = 0$ yields a pure super Gaussian distribution as a source prior. The separation performance of the IVA algorithm with this new mixed multivariate source prior is evaluated and discussed in Section 5.4. The new mixed source prior is also adopted for the fast version of the IVA algorithm and it is discussed in detail in the next section.

## 5.3 The Mixed Source Prior for the FastIVA algorithm

The proposed mixed source prior is also adopted as a source prior for the FastIVA method which is a fast converging version of the IVA method. The Newton's method, which can converge quadratically, is used as a learning gradient. The objective function used by the FastIVA algorithm is given as [43]:

$$J_{FastIVA} = \sum_{i=1}^{N}\Big[E[F(\sum_{k=1}^{K}|\hat{s}_i^{(k)}|^2)] - \sum_{k=1}^{K}\lambda_i^{(k)}(\mathbf{w}_i^{(k)H}\mathbf{w}_i^{(k)} - 1)\Big] \tag{5.3.1}$$

where $\mathbf{w}_i^H$ is the $i$th row of the unmixing matrix $\mathbf{W}$, and $\lambda_i$ is the $i$th Lagrange multiplier. $F(\cdot)$ represents the nonlinear function which is the summation of the desired signals in all frequency bins. This nonlinear score function may take several different forms as explained in [43]. Using the appropriate normalisation, the learning rule for the FastIVA method can be derived as:

$$\begin{aligned}\mathbf{w}_i^{(k)} \leftarrow & E\Big[F^{'}(\sum_{k'=1}^{K}|\hat{s}_{i,o}^{(k')}|^2) + |\hat{s}_{i,o}^{(k)}|^2 F^{''}(\sum_{k'=1}^{K}|\hat{s}_i^{(k')}|^2))\Big]\mathbf{w}_i^{(k)} \\ & - E\Big[(\hat{s}_{i,o}^{(k)})^* F^{'}(\sum_{k'=1}^{K}|\hat{s}_{i,o}^{(k')}|^2)\mathbf{x}^{(k)}\Big]\end{aligned} \tag{5.3.2}$$

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

$$\mathbf{W}^{(k)} \leftarrow (\mathbf{W}^{(k)}(\mathbf{W}^{(k)})^H)^{-1/2}\mathbf{W}^{(k)}. \tag{5.3.3}$$

The nonlinear score function $F(\cdot)$ can take different forms based on the source prior it is derived from. The selection of the source prior is crucial to the separation performance of the algorithm.

### 5.3.1 The super Gaussian Source Prior

A particular super Gaussian distribution is used as a source prior for the FastIVA algorithm as in [43]. Assuming unity variance and zero mean, this super Gaussian source prior is written as:

$$F\Bigg(\sum_{k'=1}^{K}|\hat{s}_i{}^{(k)}|^2\Bigg) = \sqrt{\sum_{k'=1}^{K}|\hat{s}_i{}^{(k')}|^2} \tag{5.3.4}$$

With the appropriate normalisation, the nonlinear score function for the FastIVA method using the original super Gaussian distribution as a source prior can be derived as follows:

$$F''\Big(\sum_{k'=1}^{K}|\hat{s}_i{}^{(k)}|^2\Big) = \Bigg(\frac{1}{\sqrt{\sum_{k'=1}^{K}|\hat{s}_i{}^{(k')}|^2}}\Bigg)^3 \tag{5.3.5}$$

Similar to the normal IVA method, the separation performance of the FastIVA method can be further improved by carefully selecting an appropriate source prior.

### 5.3.2 The Student's t Source Prior

The source prior using the Student's t distribution takes the form:

$$p(\mathbf{s}_i) \propto \left(1 + \frac{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1}(\mathbf{s}_i - \boldsymbol{\mu}_i)}{\nu}\right)^{-\frac{\nu+K}{2}} \tag{5.3.6}$$

The nonlinear score function for the FastIVA method can be derived from the source prior in (5.3.6). When the covariance matrix is set to an identity matrix due to Fourier bases, the mean is assumed to be zero and with appropriate normalisation, the nonlinear multivariate score function for the Student's t source prior based FastIVA algorithm can be written as:

$$F''(\sum_{k'=1}^{K}|\hat{s}_i^{(k')}|^2) = \frac{1-\sum_{k'=1}^{K}|\hat{s}_i^{(k')}|^2}{\big(1+\sum_{k'=1}^{K}|\hat{s}_i^{(k')}|^2\big)^2} \tag{5.3.7}$$

This nonlinear multivariate score function will preserve the inter-frequency dependency as all the frequency bins are accounted for during the learning process. The separation performance of the FastIVA method can also be improved by using the new mixed source prior that can be adjusted based on different speech sources. The approach is explained in the next section.

### 5.3.3 Mixed source prior for the FastIVA Method

A mixture of the original multivariate super Gaussian and multivariate Student's t source priors is also adopted as a source prior for the FastIVA method [100]. The latter accounts for the high amplitude samples and the former for the lower samples as explained in Section 5.2.3. The new mixed multivariate source prior for the FastIVA method can be expressed, in general form, as:

$$p(\mathbf{s}_i) = \lambda_d . f_{St} + (1 - \lambda_d) . f_G \tag{5.3.8}$$

When $f_{St}$ is replaced by the multivariate Student's t distribution in Equation (5.3.6) and $f_G$ is replaced with the original super Gaussian distribution in Equation (5.3.4), the general equation for the new source prior takes the form:

$$\begin{aligned} p(\mathbf{s}_i) =& \lambda_d \left( 1 + \frac{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i)}{\nu} \right)^{-\frac{\nu+K}{2}} \\ &+ (1 - \lambda_d) \left( \sqrt{\sum_{k'=1}^{K} |\hat{s}_i{}^{(k')}|^2} \right) \end{aligned} \tag{5.3.9}$$

The nonlinear multivariate score function for the FastIVA method can be derived using the mixed multivariate source prior shown in equation (5.3.9). The overall score function for the FastIVA method based on the new mixed source prior for source $\hat{s}_i$ can be obtained as:

$$
\begin{aligned}
F''\Big(\sum_{k'=1}^{K}|\hat{s}_i^{(k')}|^2\Big) =&(\lambda_d)\Bigg(\frac{1-\sum_{k'=1}^{K}|\hat{s}_i^{(k')}|^2}{\big(1+\sum_{k'=1}^{K}|\hat{s}_i^{(k')}|^2\big)^2}\Bigg)\\
&+(1-\lambda_d)\Bigg(\frac{1}{\sqrt{\sum_{k'=1}^{K}|\hat{s}_i{}^{(k')}|^2}}\Bigg)^3
\end{aligned}
\qquad (5.3.10)
$$

Equation (5.3.10) represents the nonlinear multivariate score function for the FastIVA algorithm with $\lambda_d$ as a weighting parameter, which can be used to control the ratio of both distributions in the mixed source prior to cater for different types of speech signals. The separation performance of the FastIVA algorithm with this new mixed multivariate source prior is evaluated and discussed in the next section.

## 5.4 Experimental Results

The new mixed source prior for the IVA and FastIVA methods is evaluated using two different room impulse responses. Firstly, it is evaluated with simulated room impulse responses (RIRs) based on the image source method (ISM) [71]. These RIRs are synthetic and do not represent a real life room environment. They are however normally used to compare the performance of different algorithms. For more robust evaluation of the algorithms, the proposed mixed source prior is further evaluated with real binaural room impulse responses (BRIRs), which were recorded in a real classroom environment with very high $RT_{60}$ of 565ms by Shinn, et al. [74]. The value of the degrees of freedom for the Student's t distribution was set to the value of four, which is empirically found to to yield 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 original super Gaussian distribution and the Student's t distribution in the mixed source prior.

Two different speech signals of length of approximately four seconds were chosen randomly from the TIMIT dataset [66] and convolved into two mixtures. These mixtures were then separated by using the IVA and FastIVA algorithms with the new mixed source prior. The separation performance of the algorithms was measured using the objective measure of signal to distortion ratio (SDR) in decibels (dB). The evaluation results using both room impulse responses are compared with the original IVA method [20] and original FastIVA method [43] with the super Gaussian source prior.

### 5.4.1 Evaluation with the Image Source Method (ISM)

In these experiments, the room impulse responses are generated using the image source method (ISM) [71]. The $RT_{60}$ was set to 200ms. The different experiment parameters are given in Table 5.1. For increased reliability, five different sets of speech signals were used and sources were placed at six source positions in the same room. The separation performance (SDR) is averaged over the six positions for the five speech signal mixture pairs. For each set of speech signals the SDR values are averaged for both estimated source signals.

#### 5.4.1.1 The IVA Algorithm

The separation performance results of the original IVA method [20] and the IVA method with the new source prior are shown in Table 5.2. Since the room impulse responses are simulated at relatively low $RT_{60}$ of 200ms, the obtained SDR values are generally high.

**Table 5.1.** Experiment Parameters for ISM

| | |
|---|---|
| Sampling rate | 8kHz |
| STFT frame length | 1024 |
| Weighting parameter | 0.5 |
| Degrees of freedom | 4 |
| Reverberation time | 200 ms |
| Positions of Microphones | [3.48 2.50 1.50] m and [3.44 2.50 1.50] m |
| Room dimensions | 7m x 5m x 3m |
| Source signal duration | 4 s (TIMIT) |

The results confirm that the new mixed source prior based IVA method produces better separation performance than the original IVA method, for all the five sets of speech signals. The average separation performance improvement using the new mixed source prior is approximately 0.92 dB compared with the original IVA method.

**Table 5.2.** SDR (dB) values for both source priors for the original IVA method for five speech mixtures with ISM [71]. SDR was averaged over six source positions. The mixed source prior shows improvement over the super Gaussian source prior for all mixtures.

| | s-Gaussian [20] | Mixed Source Prior | Improvement (dB) |
|---|---|---|---|
| Set-1 | 9.24 | 10.38 | 1.14 |
| 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 |

#### 5.4.1.2 The FastIVA Algorithm

The separation performance results of the original FastIVA method with the super Gaussian source prior [43] and the FastIVA method with the new source prior for all five speech mixtures are shown in Table 5.3. Again, the obtained SDR values are relatively high due to the relatively low $RT_{60}$ of 200ms.

**Table 5.3.** Improvement in separation performance of the FastIVA algorithm with new source prior in terms of SDR (dB) for five speech mixtures using the ISM [71]. SDR was averaged over six source positions. The proposed source prior shows improvement over the super Gaussian source prior for all mixtures

| | s-Gaussian [43] | 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 |

The results demonstrate that, for all the mixtures, the proposed mixed source prior improves the separation performance of the FastIVA method. It improves the average separation performance of the FastIVA method by approximately 1.15 dB.

### 5.4.2 Evaluation with Binaural Room Impulse Responses (BRIRs)

In these experiments, the separation performance of the IVA and FastIVA methods with new mixed source prior is evaluated with BRIRs, which were obtained from [74]. As these BRIRs are real room recordings with high $RT_{60}$ of 565ms, they provide realistic evaluation of the separation performance of BSS algorithms in highly reverberant environments. In order to evaluate the separation performance of the proposed mixed source prior, five different source location azimuths $(15^\circ, 30^\circ, 45^\circ, 60^\circ, 75^\circ)$ relative to the first source, in the room, were considered. The room layout and experimental setup are illustrated in Figure 5.1. Changing source positions represents speakers moving in the room which provide thorough evaluation of separation performance of the algorithm. All the experiments at all the source location

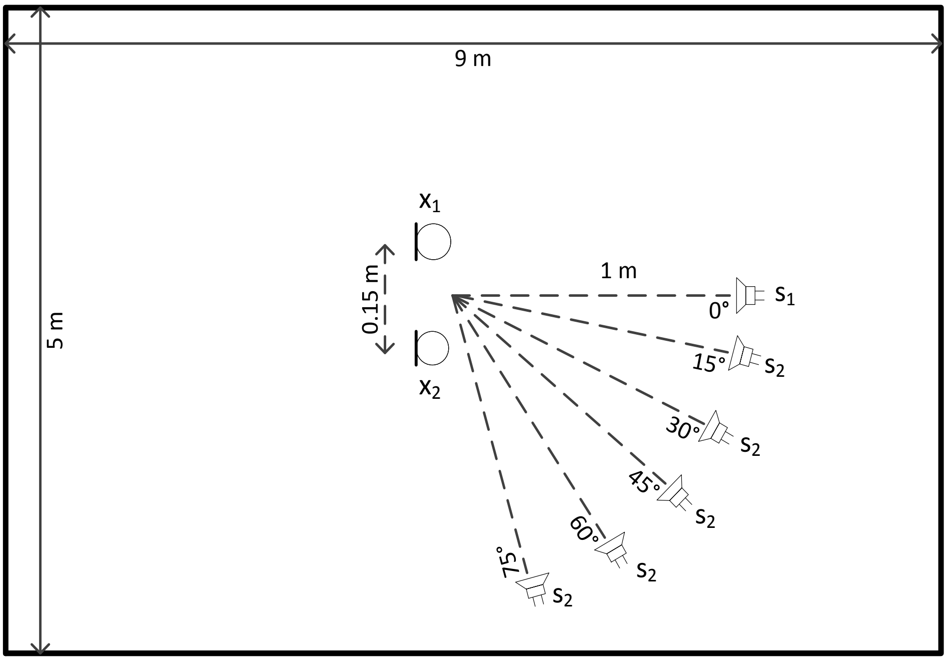


**Figure 5.1.** 2D plan of room and experimental setup for evaluating the mixed source prior using BRIRs, showing locations of sources and microphones.

azimuths were repeated three times to improve the reliability of the results. The summary of different parameters used in the experiments is provided in Table 5.4.

**Table 5.4.** Experiment parameters for BRIRs.

| | |
|---|---|
| Sampling rate | 8kHz |
| STFT frame length | 1024 |
| Weighting parameter | 0.5 |
| Degrees of freedom | 4 |
| Reverberation time | 565 ms |
| Room dimensions | 9 m x 5 m x 3.5 m |
| Source signal duration | 4 s (TIMIT) |
| Source distance | 1 m |

The speech mixtures were created using the BRIRs with high $RT_{60}$ of 565ms and then separated using the IVA and FastIVA algorithms with the new mixed multivariate source prior and the results were compared with the original IVA method [20] and original FastIVA method [43], respectively. In order to improve the reliability of results, the SDR values at each angle were averaged over eighteen speech mixtures.

#### 5.4.2.1 The IVA Algorithm

The separation performance of the IVA algorithm, expressed in SDR, using both source priors is shown in Figure 5.2. The data show variable separation performance at different angles with the best performance at angle 45°. The bar plots confirm that the IVA method with new mixed source prior has better separation performance compared to the original IVA method at all five source positions. The average recorded separation performance improvement using the new mixed source prior is approximately 0.85 dB compared with the original IVA method.

#### 5.4.2.2 The FastIVA Algorithm

Figure 5.3 shows the SDR separation performance, in (dB), of both FastIVA algorithms at five different positions. The data reveal that the FastIVA algorithm with the new mixed source prior enhances the separation performance at all azimuth angles. On average, the new mixed source prior improves the separation performance of the FastIVA method by approximately 0.9 dB using the real BRIRS.

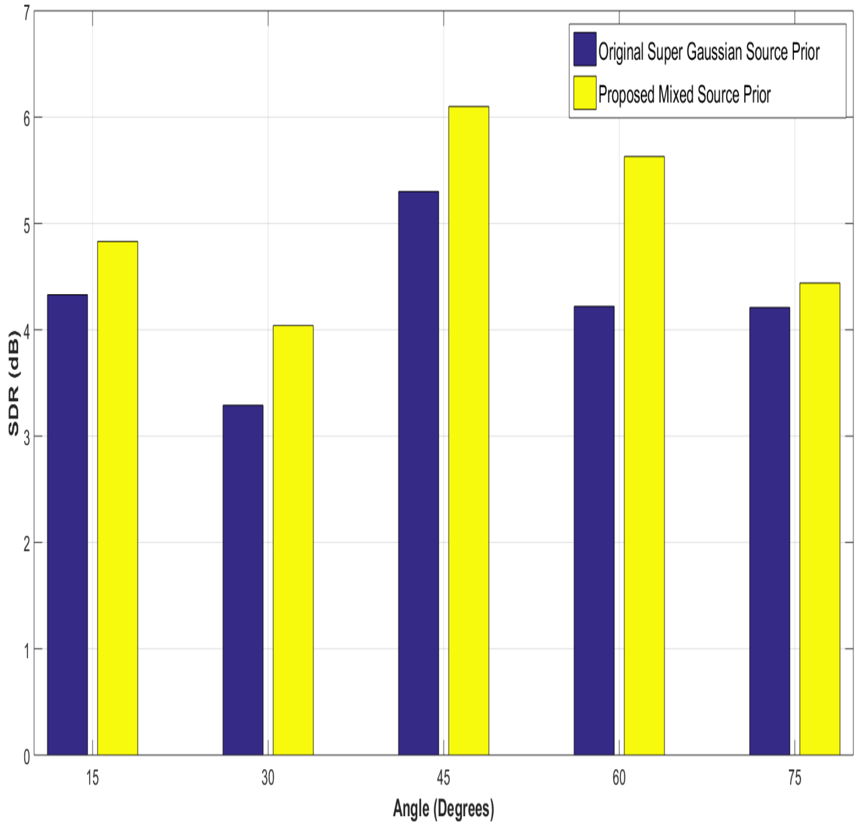


**Figure 5.2.** The graph shows the SDR (dB) values at five different separation angles. Real BRIRs from [74] were used. Results were averaged over eighteen mixtures. The mixed source prior enhances the separation performance of the IVA algorithm at all separation angles.

The performance of the proposed mixed 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 mixed source prior that can adapt to different speech mixtures is proposed in the next section.

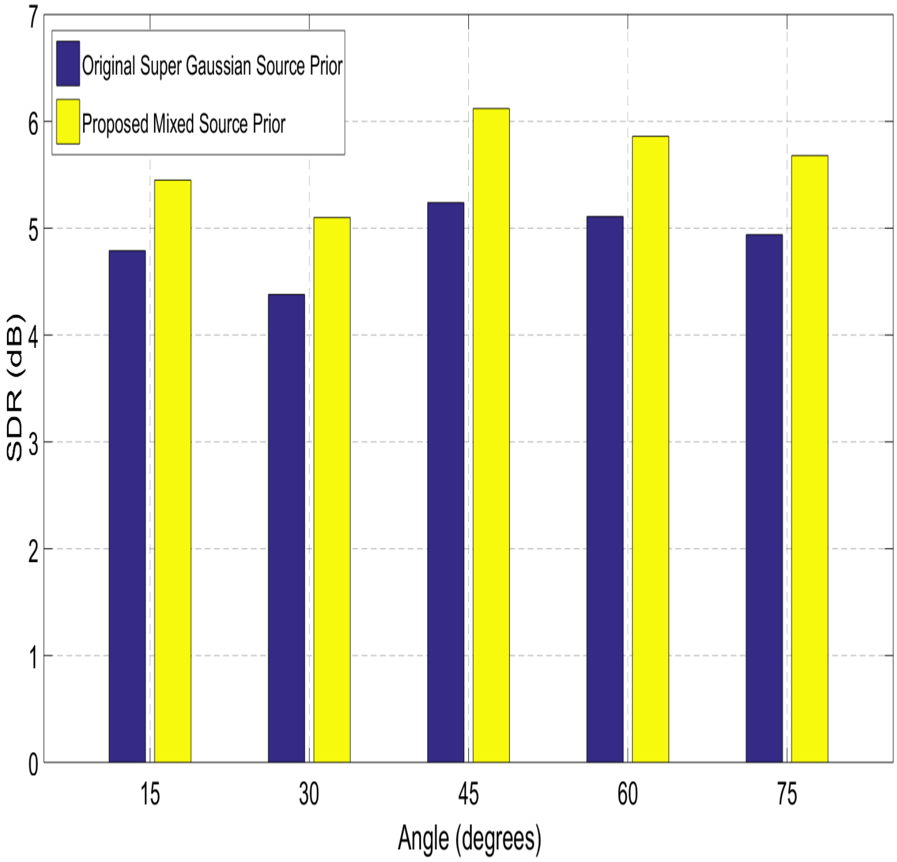


**Figure 5.3.** The bar graph provides SDR (dB) for the FastIVA method [43] and the proposed mixed source prior FastIVA for five different angles. All the SDR values are averaged over eighteen random mixtures. Real BRIRs from [74] were used. The new mixed source prior enhance the separation performance at all separation angles.

## 5.5 Energy Driven Mixed Source Prior for the Original IVA Method

In the mixed source prior for the original IVA method, equal weights were given to both the Student's t and the original super Gaussian distributions for all speech sources. As speech signals have different statistical properties, selecting a ratio for both distributions based on the variations in the speech sources, can potentially improve the separation performance of the technique. In this section, the source prior is modified so that the weights of both distributions in the mixed source prior are adjusted automatically according to the energy of the observed speech mixture signals. This technique is found to be successful only with access to mixture signals not the original sources [101]. To enhance the separation performance of the algorithm, the structure of the dependency model is exploited. The dependency among neighbouring frequency bins is generally stronger and much weaker between distant frequency bins [98]. Hence, a clique based approach is adopted where the fully connected frequency spectrum is decomposed into smaller cliques retaining considerable overlap between adjacent cliques.

### 5.5.1 Clique Based IVA Method

In the original IVA method, the inter-frequency dependency is preserved by using the multivariate source prior. It adopts a spherically symmetric dependency model that assigns the same kind of dependency to neighbouring frequency components and to frequency components that are located far apart. The model can be depicted as a total clique, as shown in Figure 5.4 (a), where all of the line connections represent the same weight of dependency. Such a source prior does not model speech accurately because, in real speech signals, the dependency of neighbouring frequency components is much stronger than that of distant frequency components [98].

Therefore, in order to enhance the frequency dependency within the IVA method, the single and fully connected statistical model is divided to several overlapping cliques of fixed size. This dependency model

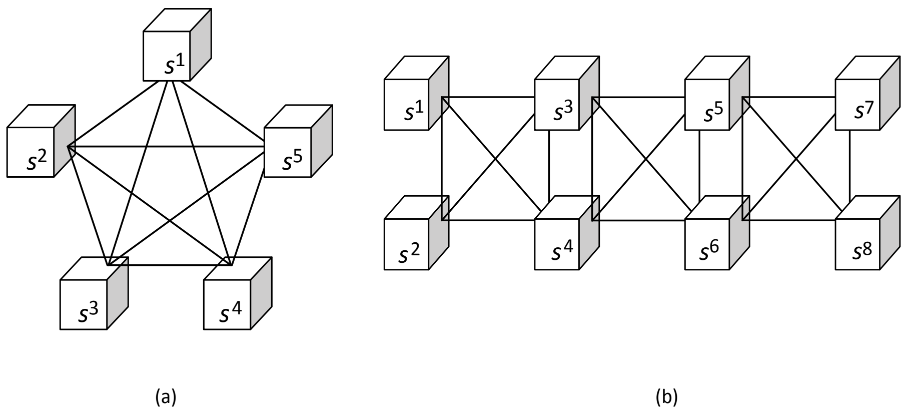


**Figure 5.4.** The IVA dependency models [102]. The line connections of each clique represent a fixed spherical dependency weight. (a) A global clique to represent spherical dependency. (b) A chain of cliques to represent dependency propagation through the overlaps of the chains.

is locally spherical and the dependency among the frequency components is propagated through overlaps of cliques so that the dependency between the components weakens as the distance separating them increases [102]. The model is depicted in Figure 5.4 (b).

The multivariate probability density function of clique based dependency model can be written in the form [98]:

$$p(\mathbf{s}_i) \propto \exp\left( -\sum_{c=1}^{C} \sqrt{\sum_{k=f_c}^{l_c} \left| \frac{\hat{\mathbf{s}}_i^{(k)}}{\sigma_i{}^{(k)}} \right|^2} \right) \tag{5.5.1}$$

where $f_c$ and $l_c$ are the first and last indices of the $c$th clique, respectively. $C$ is the number of cliques. This new dependency structure consists of several cliques of fixed and identical size and the centre frequency increases with clique propagation. For instance, in order to deploy the clique based dependency structure for the case of 1024 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 $[f_1, l_1] = [1, 256], [f_2, l_2] = [17, 272], \ldots, [f_c, l_c] = [769, 1024]$. The model improves the dependency structure for the IVA method because the strength of the dependency between the frequency bins is obtained as a function of the distance between them with some overlap. Consequently, it would improve the separation performance of the IVA method with the new energy driven mixed source prior. Finding the energy of the measured speech signals and tuning the mixed source prior accordingly is discussed in the next section.

### 5.5.2 Energy Calculation of Measured Speech Mixtures

As discussed in Section 5.2.3, the mixed source prior for the original IVA method is given as:

$$\begin{aligned} p(\mathbf{s}_i) =& \lambda_d \Bigg(1 + \frac{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i)}{\nu}\Bigg)^{-\frac{\nu+K}{2}} + \\ & (1 - \lambda_d) \exp\Bigg( - \sqrt{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i)}\Bigg) \end{aligned} \tag{5.5.2}$$

The corresponding nonlinear score function for the above mixed source prior can be obtained as:

$$\begin{aligned} \varphi^{(k)}(\hat{s}_i{}^{(1)} \cdots \hat{s}_i{}^{(K)}) \propto & \lambda_d \Bigg( \frac{\hat{s}_i^{(k)}}{1 + \frac{1}{\nu} \sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2} \Bigg) \\ & + (1 - \lambda_d) \Bigg( \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2}} \Bigg) \end{aligned} \tag{5.5.3}$$

The parameter $\lambda_d$, in the score function (5.5.3), is a weighting parameter, that defines the ratio of the Student's t and the original super

Gaussian distributions in the mixed source prior. The observed speech mixtures can have different energies due to the different statistical properties of the speech source signals. Therefore, the mixed source prior can adapt to model different speech mixtures, if the value of $\lambda_d$ is tuned in line with their local energy. The weighting parameter becomes frequency dependent i.e. $\lambda_d^{(k)}$ and is estimated according to the energy of the observed speech mixture. The frequency bins are divided into smaller non-overlapping blocks because different frequency ranges can have different energy. To model a particular block, an appropriate value for $\lambda_d^{(k)}$ can be selected. $\lambda_d^{(k)}$ is calculated as the normalised energy of the speech mixtures in the frequency domain blocks. The normalised energy of a particular block can be calculated as follows:

$$E_b = \frac{1}{E_t}\left(\sum_{k=f_b}^{l_b} ||\mathbf{x}_p^{(k)}||^2\right) \tag{5.5.4}$$

where $E_b$ is the energy of the particular block, $E_t$ is the total energy of the source mixture and $||(\cdot)||$ denotes Euclidean norm. $f_b$ is the first index of the block and $l_b$ is its last index. $\mathbf{x}_p^{(k)}$ denotes the vector of all frequency components $k$, calculated by dividing the entire speech observation into subblocks indexed by $p$.

The energy of a particular block is, generally, a measure of the amplitude information of the mixture signals in that block. The higher the energy the higher the amplitude. As a result, the value of the weighting parameter $\lambda_d^{(k)}$ is selected to reflect amplitude level within each block. As a high energy indicates high amplitude information, $\lambda_d^{(k)}$ is tuned so that the percentage of the Student's t distribution in the mixed source prior is higher than that of the original super Gaussian distribution.

The Student's t distribution can improve the modelling of the high amplitude information due to its heavy tail nature. Likewise, when the energy of a particular block is relatively low, it indicates lack of high amplitude information. Therefore, in order to appropriately model the speech sources, the mixed source includes higher percentage of the original super Gaussian distribution than the student's t distribution. The weighting parameter $\lambda_d^{(k)}$ is tuned to assign more weight to the original super Gaussian distribution in the mixed source prior. The energy driven mixed source prior should provide more accurate model to the underlying non-stationary speech signals by accustoming to the nature of the measured speech mixture. This, consequently, results in separation performance improvement of the IVA method. The experimental setup and the evaluation results of the performance of the new energy driven mixed source prior based IVA algorithm are discussed in the next section.

### 5.5.3 Experimental Results

The proposed energy driven mixed Student's t-original super Gaussian source prior for the original IVA method is evaluated using three different types of room impulse responses; a synthetic room impulse response and two real room impulse responses are employed. It is evaluated with simulated room impulse responses (RIRs) based on the image source method (ISM) [71]. These RIRs do not represent a real life room environment, but they are useful in comparing the performance of different algorithms. For real life environment, the proposed mixed source is further evaluated with two real room impulse responses; BRIRs [74] and Hummerstone [75].

The value of the degrees of freedom for the Student's t distribution was set to the value of four, which is empirically found to yield best separation performance. Two different speech signals of length of approximately four seconds were chosen randomly from the TIMIT dataset [66] and convolved into two mixtures. The weighting parameter $\lambda_d^{(k)}$ in the mixed source prior was tuned according to the energy of the measured speech mixtures as explained in Section 5.5.2. These mixtures were then separated using the IVA algorithm with new energy driven mixed source prior. The separation performance of the algorithms was measured using the objective measure of signal to distortion ratio (SDR) in decibels (dB) as well as the subjective measure of perceptual evaluation of speech quality (PESQ) [81]. The separation performance was compared with the fixed mixed source prior, the original super Gaussian source prior and Student's t source prior for the original IVA method [20].

#### 5.5.3.1 Evaluation with Image Source Method (ISM)

In these experiments, the room impulse responses are generated using the image source method (ISM) [71]. The different experiment parameters are given in Table 5.5. The $RT_{60}$ was set to 250ms to provide higher reverberation time than used for the fixed mixed source prior. For increased reliability, five different sets of speech signals were used and sources were placed at six source positions in the same room. The separation performance (SDR) is averaged over the six positions for the five speech signal mixture pairs. For each set of speech signals the SDR values are averaged for both estimated source signals. The separation performance of the new energy mixed source prior is compared with the performance of the original super Gaussian source prior.

**Table 5.5.** Experiment Parameters for ISM

| | |
|---|---|
| Sampling rate | 8kHz |
| STFT frame length | 1024 |
| Degrees of freedom | 4 |
| Reverberation time | 250 ms |
| Positions Microphones | [3.48 2.50 1.50] m and [3.44 2.50 1.50] m |
| Room dimensions | 7m x 5m x 3m |
| Source signal duration | 4 s (TIMIT) |

The separation performance results of the IVA method with the new energy driven mixed source prior and the original IVA method [20] are shown in Table 5.6. The relatively high SDR values are due to the relatively low $RT_{60}$ of 250ms. The results confirm that the new energy driven mixed source prior IVA method produces better separation performance than the original IVA method, for all the five sets of speech signals. The average separation performance improvement using the proposed source prior is approximately 1 dB.

**Table 5.6.** SDR (dB) values for both source priors for the original IVA method with image room impulse response [71]. The energy driven mixed source prior shows improvement for all mixtures.

| | s-Gaussian [20] | Proposed Source Prior | Improvement (dB) |
|---|---|---|---|
| Set-1 | 8.58 | 9.53 | 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 |

#### 5.5.3.2 Evaluation with Real Room Impulse Responses from Hummerstone

In these experiments, the separation performance of the IVA method with the proposed energy driven mixed source prior is evaluated with real room impulse response, which were obtained from Hummerstone [75]. These are room impulse responses recorded in real life environment in four different rooms. The types of the four rooms and their respective reverberation time $RT_{60}$ are shown in Table 5.7. The four rooms have different sizes, geometry and reverberation time $RT_{60}$. Therefore, they offer extensive evaluation of algorithm over a range of reverberation times and room settings. In addition, in each room, the source location azimuths relative to the second source can be varied from $-90\,^{\circ}$ to $90\,^{\circ}$, allowing for evaluation at different positions for moving sources.

**Table 5.7.** Room types and the respective $RT_{60}$. Hummerstone

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

For this set of experiments, the source location azimuths in step of $15\,^{\circ}$ was considered from $15\,^{\circ}$ to $90\,^{\circ}$ in all the rooms. The separation performance was measured objectively with SDR in dB. The mixtures were then separated by using the new energy driven mixed source prior and its separation performance is compared with the original IVA method [20]. The separation performance of both methods for all four rooms is shown in Figure 5.5. The results at all the separation angles are the average of twelve mixtures for increased reliability of the results as in [101].

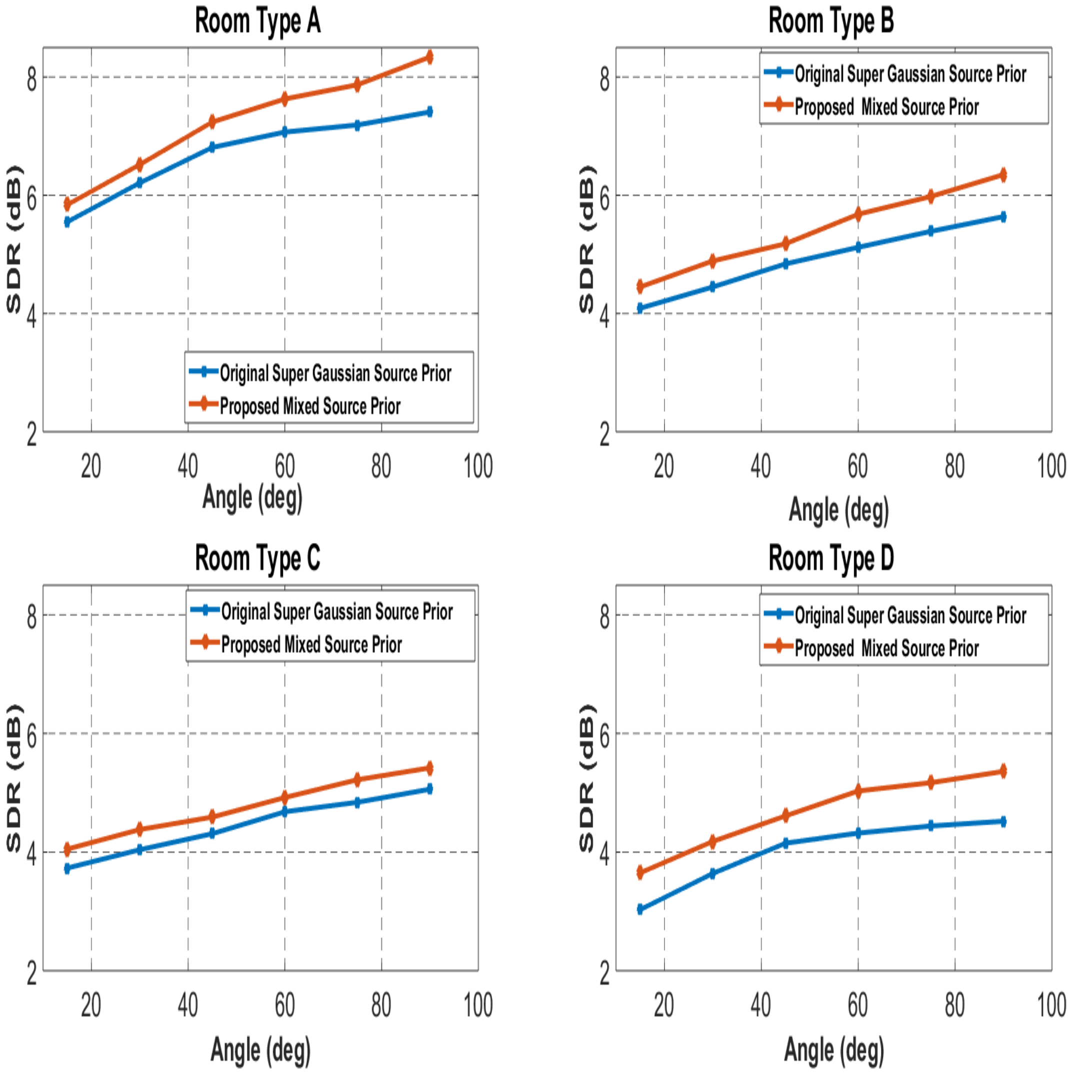


**Figure 5.5.** The separation performance (SDR) for different rooms. SDR values (dB) were averaged over twelve mixtures at each separation angle. Energy driven mixed source prior enhance the separation performance of the IVA algorithm in all types of reverberant conditions.

The results in Figure 5.5 reveal that the energy driven mixed source prior consistently improves the separation performance of the IVA algorithm in all rooms at all the separation angles. In room A, the SDR values for both the algorithms is the highest as the $RT_{60}$ is the lowest at 320ms. The SDR values drop for both algorithms as the reverberation time $RT_{60}$ rises (Rooms B, C and D). However, in all cases, the proposed technique performs better than the original IVA method at all the separation angles. This demonstrates the robustness of the technique in different settings including extremely difficult and highly reverberant environments. The proposed source prior tunes the weight of both distributions according to the energy of the measured mixtures and it provides better separation compared with the original IVA method [20] at all the separation angles. The average separation performance improvement using the proposed source prior is approximately 0.6 dB.

#### 5.5.3.3 Evaluation with Binaural Room Impulse Responses (BRIRs)

In these experiments, the separation performance of the IVA method with new energy driven mixed source prior is evaluated with BRIRs, which were obtained from [74]. As these BRIRs are real room recordings with high $RT_{60}$ of 565ms, they provide realistic evaluation of the separation performance of BSS algorithms in highly reverberant environments. The length of the speech signals was chosen to be approximately five seconds. Six different source location azimuths varying from $15\,^\circ$ to $90\,^\circ$ with a step of $15\,^\circ$ relative to the first source, in the room, were considered. The room layout and experimental setup are illustrated in Figure 5.6. Changing source positions represents speakers moving in the room which provide thorough evaluation of separation

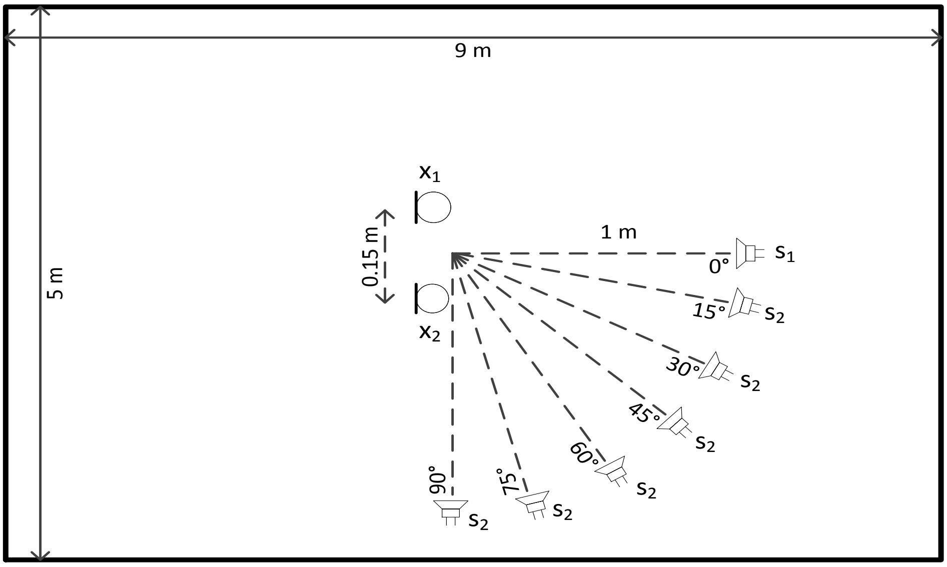


**Figure 5.6.** 2D plan of room and experimental setup for for evaluating the energy driven mixed source prior using BRIRs, showing locations of sources and microphones.

performance of the algorithm. All the experiments at all the source location azimuths were repeated three times to improve the reliability of the results. The summary of different parameters used in the experiments is provided in Table 5.8.

**Table 5.8.** Different parameters used in experiments (BRIRs).

| | |
|---|---|
| Sampling rate | 8kHz |
| STFT frame length | 1024 |
| Degrees of freedom | 4 |
| Reverberation time | 565 ms (BRIRs) |
| Room dimensions | 9 m x 5 m x 3.5 m |
| Source signal duration | 5 s (TIMIT) |
| Source distance | 1 m |

The speech mixtures were created by using the BRIRs with high $RT_{60}$ of 565ms and then separated using the IVA algorithms with the proposed mixed multivariate source prior and the results were compared with the original IVA method [20], in terms of SDR and PESQ. In order to improve the reliability of evaluation, the results at each angle were averaged over eighteen speech mixtures.

The separation performance of the IVA algorithm at different angles was measured objectively using SDR in dB. The results using both source priors is shown in Figure 5.7. The data show variable separation performance at different angles and confirm that the IVA method with new energy driven mixed source prior consistently achieves improved separation performance compared to the original IVA method at all six source positions. The average recorded separation performance improvement using the new mixed source prior is approximately 1 dB compared with the original IVA method.

Additionally, in this experiment, a subjective measure of perceptual evaluation of speech quality (PESQ) is used to measure the separation performance of the algorithm using the BRIRs. The PESQ is a commonly used measure to examine the quality of the separated signal as it compares the estimated (original) signals. A score is given between 0-4.5, 0 for very poor separation and 4.5 for excellent separation.

The signals were separated from mixtures using the energy driven mixed source prior based IVA method in the same settings as in Table 5.8. The PESQ scores for separated signals were measured as shown in Table 5.9. All the PESQ scores for each mixture are the average of PESQ scores for six different source location azimuths varying from ($15^\circ$ to $90^\circ$). The PESQ scores for the proposed energy driven source

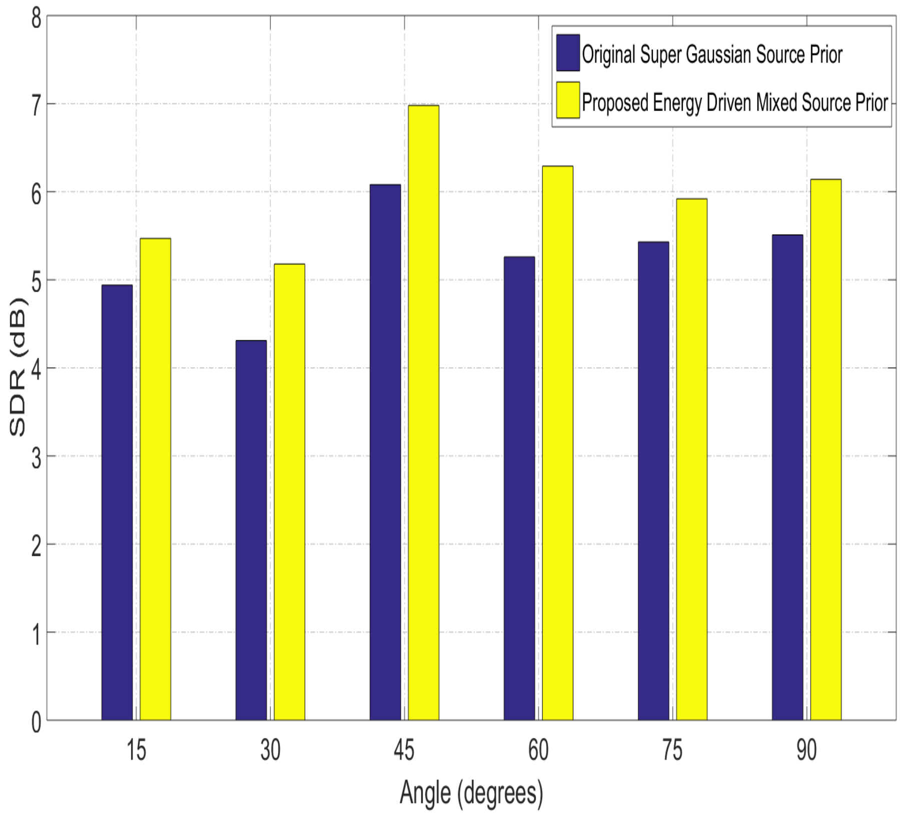


**Figure 5.7.** Separation performance in terms of SDR (dB) values for the energy based mixed source prior and the original IVA algorithm. The energy based mixed source prior enhances the separation performance of the IVA algorithm at all source locations in the room.

prior is compared with the PESQ score of the estimated signals separated by the original IVA method in the same settings. This subjective study also confirms the improved separation performance for the IVA method with the energy driven mixed source prior. The average separation performance improvement PESQ score is approximately 0.25.

### 5.5.4 Comparative Evaluation of Different Source Priors

The final set of experiments will establish the advantage of automatically adapting the weight of distributions in the mixed source prior

**Table 5.9.** PESQ values for the IVA algorithm with the two source priors. PESQ scores are averaged over six different locations in the room. The proposed scheme enhances the separation performance of the IVA algorithm at all source locations

| | Original Source Prior [21] | 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 |

based on the mixture energy. The proposed source prior is compared with the fixed mixed source prior and the Student's t source prior. For all three methods, the mixtures were created by using the room impulse response generated by the BRIRs [74] with reverberation time $RT_{60}$ of 565ms. The experimental settings are similar to the parameters given in Table 5.8. 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 were separated using the IVA method with the three source priors at six different source location azimuths ($15\,^\circ$ to $90\,^\circ$). The separation performance in terms of SDR is shown in Figure 5.8. For increased reliability of evaluation, the SDR value at each angle were averaged over twelve speech mixtures. It is evident, from plots in Figure 5.8, the superiority of the proposed energy driven mixed source prior based IVA method over the other two methods. The adjustment of mixing ratio according to the statistical properties of the measured mixtures, makes the technique well suited to model different types of speech sources which leads to separation performance enhancement of the IVA algorithm. The data also illustrate that the fixed mixed source prior outperforms the Student's t source prior.

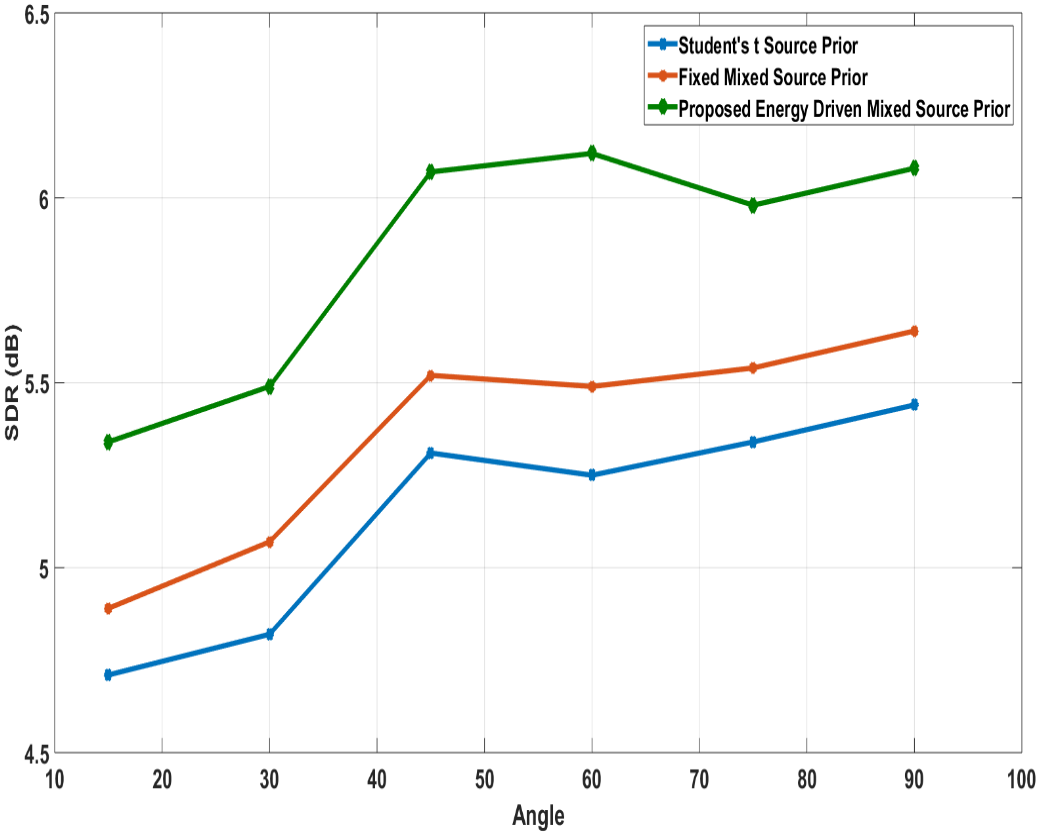


**Figure 5.8.** The separation performance of the IVA algorithm with three different source priors. Mixture were generated by BRIRs. The energy based mixed source prior is superior for all source locations.

## 5.6 Summary

In this chapter, a new enhanced multivariate source prior for the IVA algorithm was introduced as a mixture of two distributions, instead of single distribution. The source prior is constructed by mixing the original multivariate super Gaussian distribution as in [20] and the multivariate Student's t distribution with a certain ratio. Human speech is highly random in nature and can have variable amplitude components [77]. In the proposed source prior, the Student's t distribution has been used to model the high amplitude components of the speech signal [94] and the original super Gaussian distribution to model the lower amplitude components. In order for the mixed source prior to adapt to different types of speech signals, it was empowered with an energy driven scheme that adjusts the weight of each distribution according to the energy of the observed mixtures as well as a clique based dependency model.

This mixed source prior was adopted for the IVA and the FastIVA algorithms and compared with different single distribution source priors. The detailed experimental studies using simulated and real room environment with different reverberation times confirmed consistent separation performance improvement of the energy driven mixed source prior based IVA.

In the following chapter, online IVA algorithm for speech separation is introduced. The algorithm is enhanced by an adaptive learning scheme to improve the performance in terms of convergence time and steady state separation and accuracy. Two source priors are used to evaluate the proposed scheme. A switched source prior technique, that combines the advantages of both source priors, is proposed.

# Chapter 6

# ONLINE IVA WITH ADAPTIVE LEARNING FOR SPEECH SEPARATION USING VARIOUS SOURCE PRIORS IN REAL ROOM ENVIRONMENTS

## 6.1 Introduction

Independent vector analysis (IVA) is a method to tackle BSS in the frequency domain. The technique has proven efficient in separating independent speech signals from convolutive mixtures [20]. It solves, algorithmically, the problematic permutation problem inherent in independent component analysis (ICA) [58]. IVA extends ICA from a univariate source signal model to a multivariate one. The multivariate source prior models statistical inter dependency across the frequency bins of each source.

The original IVA method proposed in [20] runs in an offline batch manner where the entire set of input samples is gathered before calculating the parameters. This approach is not applicable to practical online systems. A block-based approach can be applied to implement a real time BSS system [103]. However, this approach encompasses heavy computational load. A fully online version of the IVA algorithm was proposed in [104] which exploits the multivariate super Gaussian distribution as a sources prior. The algorithm is suitable for practical embedded systems, where the coefficients of the separation filter are updated at every time frame. The auxiliary version of the algorithm was implemented in [105]. Online IVA with Student's t source prior for convolutive speech mixtures was proposed in [106]. Previously, several implementations of online ICA based techniques were proposed in [107–112]. However, they entail additional post processing techniques to address the permutation problem which may render them unsuitable for embedded systems applications due to the added computational complexity.

Often online IVA methods use a fixed learning rate to update the unmixing matrix. If the learning rate is set to a high value, the solution converges faster with large fluctuations. For small learning rate value, the convergence is slower with smoother solution. In this chapter, the contribution is to introduce a new adaptive learning scheme to improve the performance of the online IVA in terms of convergence time and steady state separation performance and accuracy [113, 114]. The scheme exploits gear-shifting to combine the advantages of the high and small values of the learning rate. The learning rate is controlled by a Frobenius norm as a measure of the proximity to the target solution, which is extracted from the learning gradient adopted. Two source priors are used to model the speech signals; the super-Gaussian distribution proposed in the original IVA [20] based on a spherically symmetric Laplace (SSL) distribution and a generalized Gaussian distribution proposed in [89] which exploits fourth order inter-frequency correlation and was previously only tested on the batch IVA. Finally, a switched source prior that combines the better performance attributes of both source priors, is introduced.

Firstly, the proposed scheme is evaluated using simulated room impulse responses based on the image source method (ISM) [71] which was used by the original online IVA algorithm [104]. Then, it is evaluated in real room environments using two different room impulse responses; binaural room impulse responses (BRIRs) [74] and Hummerstone real impulse responses [75]. Real recorded speech signals, from the TIMIT acoustic-phonetic continuous speech corpus [66], are used as the source signals. The separation performance is measured objectively by the signal to distortion ratio (SDR) [76]. The learning algorithm for the online IVA is discussed in the next section.

## 6.2 Online IVA Learning Algorithm

The noise free frequency domain convolutive blind source separation (FD-CBSS) batch IVA mixing and separation models are described as [20]:

$$x_j^{(k)}[n] = \sum_{i=1}^{N} h_{ji}^{(k)} s_i^{(k)}[n] \tag{6.2.1}$$

$$\hat{s}_i^{(k)}[n] = \sum_{j=1}^{M} \mathrm{w}_{ij}^{(k)} x_j^{(k)}[n] \tag{6.2.2}$$

where $x_j^{(k)}[n]$, $s_i^{(k)}[n]$ and $\hat{s}_i^{(k)}[n]$ are respectively the $j$th observation value of $M$ observations, the $i$th source signal and $i$th estimated source of $N$ sources at the $k$th frequency bin. $h_{ji}^{(k)}[n]$ and $\mathrm{w}_{ij}^{(k)}[n]$ are the mixing and unmixing filter coefficients at the *k-th* frequency bin respectively. $k = 1, 2, \cdots, K$, and $K$ is the number of frequency bins. $n$ is the short-time Fourier transform (STFT) time block index.

The coefficients of the unmixing matrices $\mathrm{w}_{ij}^{(k)}$ can be updated by the batch update rule as [20]:

$$\mathrm{w}_{ij}^{(k)new} = \mathrm{w}_{ij}^{(k)old} + \eta \Delta \mathrm{w}_{ij}^{(k)} \tag{6.2.3}$$

where $\Delta \mathrm{w}_{ij}^{(k)}$ is the gradient for the coefficients and $\eta$ is the learning rate.

In the online IVA algorithm, the coefficients of the unmixing model are updated at every time frame. Thus, the time frame index $n$ is introduced to the unmixing model [104] as follows:

$$\hat{s}_i^{(k)}[n] = \sum_{j=1}^{M} \mathrm{w}_{ij}^{(k)}[n] x_j^{(k)}[n] \tag{6.2.4}$$

where $\mathrm{w}_{ij}^{(k)}[n]$ is the unmixing coefficient at time block $n$, frequency bin $k$ between source $i$ and microphone $j$, $\hat{s}_i$ is the $i$th estimated source, and $x_j^{(k)}[n]$ is the observation value of $M$ observations at time frame $n$ and frequency bin $k$. Therefore, the coefficients of the online unmixing filter are updated as follows:

$$\mathrm{w}_{ij}^{(k)}[n+1] = \mathrm{w}_{ij}^{(k)}[n] + \eta \Delta \mathrm{w}_{ij}^{(k)}[n] \tag{6.2.5}$$

where $\Delta \mathrm{w}_{ij}^{(k)}[n]$ is the gradient at the current time frame $n$. This is the critical variation between the batch IVA and online IVA algorithms and will be discussed in subsection (6.2.1).

The IVA algorithm deploys a cost function for multivariate random variables that attempts to remove dependency between the sources and retain the dependency between frequency bins of each source. To measure independence, the cost function of the IVA ($C$) uses the Kullback-Leibler ($KL(\cdot)$) divergence between the joint probability density function (PDF) of the estimated source vectors $p(\hat{\mathbf{s}}_1, \cdots, \hat{\mathbf{s}}_N)$ and the product of their marginal individual PDFs $\prod_{1=1}^{N} q(\hat{\mathbf{s}}_i)$ [20]:

$$C = KL\Big(p(\hat{\mathbf{s}}_1, \cdots, \hat{\mathbf{s}}_N) \| \prod_{1=1}^{N} q(\hat{\mathbf{s}}_i)\Big) \tag{6.2.6}$$

$$= const. - \sum_{k=1}^{K} \log |det(\mathbf{W}^{(k)})| - \sum_{i=1}^{N} E \log q(\hat{\mathbf{s}}) \tag{6.2.7}$$

where $det(\cdot)$ is the matrix determinant operator, $|\cdot|$ denotes the absolute value and $E[\cdot]$ represents the statistical expectation operator.

The source prior $q(\hat{\mathbf{s}})$ in the cost function is a vector across all frequency bins. Each source is multivariate and the cost function would be minimised when the dependency between the source vectors is removed but the inherent frequency dependency between the components of each vector can be retained. The online natural gradient IVA will be derived next.

### 6.2.1 Online Natural Gradient IVA

The learning algorithm for the parameters of the unmixing filters is derived by minimising the $KL(\cdot)$ divergence. A gradient descent method [42] is employed by differentiating the cost function $C(\cdot)$ with respect to the separating filter coefficients (matrices) ($\mathrm{w}_{ij}^{(k)}$). The gradients for the coefficients ($\Delta\mathrm{w}_{ij}^{(k)}$) can be obtained [20]:

$$\Delta\mathrm{w}_{ij}^{(k)} = -\frac{\partial C}{\partial \mathrm{w}_{ij}^{(k)}} \tag{6.2.8}$$

$$= \mathrm{w}_{ij}^{-H(k)} - E\Big[\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}, \cdots, \hat{\mathbf{s}}_i^{(k)})\hat{\mathbf{x}}_j^{*(k)}\Big] \tag{6.2.9}$$

where $(\mathbf{W}^{(k)^{-1}})^H \equiv \mathrm{w}_{ij}^{-H(k)}$, $(\cdot)^*$ is the complex conjugate and $(\cdot)^H$ denotes Hermitian transpose.

The natural gradient [42] can be obtained by multiplying the gradient matrices $\Delta\mathbf{W}^{(k)} \equiv \{\Delta\mathrm{w}_{ij}^{(k)}\}$ with the scaling matrices $\mathbf{W}^{(k)H}\mathbf{W}^{(k)}$:

$$\Delta\mathrm{w}_{ij}^{(k)} = \sum_{l=1}^{N}(I_{il} - E\Big[\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}, \cdots, \hat{\mathbf{s}}_i^{(k)})\hat{\mathbf{s}}_l^{*(k)}\Big]\mathrm{w}_{ij}^{(k)} \tag{6.2.10}$$

where $I_{il}$ is the $il$-th element of the identity matrix $\mathbf{I}$ ($I_{il} = 1$ when $i = l$ and $I_{il} = 0$ when $i \neq l$). $\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}, \cdots, \hat{\mathbf{s}}_i^{(k)}) = [\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}), \cdots, \varphi^{(k)}(\hat{\mathbf{s}}_i^{(N)})]^T$ is the nonlinear score function vector.

For the online algorithm, the expectation in Equation (6.2.10) is omitted and the resultant online scored correlation $\Re_{il}^{(k)}$ at the current time frame $n$ is defined as [104]:

$$\Re_{il}^{(k)}[n] = \varphi^{(k)}\Big(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)}\Big)\hat{s}_l^{(k)*}[n] \tag{6.2.11}$$

Thus, the online natural gradient at the current time frame $n$ is given as:

$$\Delta\mathrm{w}_{ij}^{(k)}[n] = \sum_{l=1}^{N}\Big(I_{il} - \Re_{il}^{(k)}[n]\Big)\mathrm{w}_{ij}^{(k)}[n] \tag{6.2.12}$$

### 6.2.2 Nonholonomic Constraint

A nonholonomic constraint is adopted in [104] to address the problem of the gradient fluctuation due to energy changes in source signals. The gradient is obtained by replacing the identity matrix $\mathbf{I}$ with a diagonal matrix $\mathbf{\Lambda}^{(k)}$ based on the scored correlation ($\Lambda_{ii}^{(k)}[n] = \Re_{ii}^{(k)}$ and $\Lambda_{il}^{(k)}[n] = 0$ when $i \neq l$). Equation (6.2.12) becomes:

$$\Delta \mathrm{w}_{ij}^{(k)}[n] = \sum_{l=1}^{N} \left( \Lambda_{il}^{(k)}[n] - \Re_{il}^{(k)}[n] \right) \mathrm{w}_{ij}^{(k)}[n] \tag{6.2.13}$$

As the diagonal elements of $(\Lambda_{il}^{(k)}[n] - \Re_{il}^{(k)}[n])$ are always zero, the algorithm is more robust to large changes in the local magnitude of the source signals and thus converges faster. In addition, the constraint cuts $N$ multiplications at each frequency bin. The online update rule with a normalised learning rate is given by:

$$\mathrm{w}_{ij}^{(k)}[n+1] = \mathrm{w}_{ij}^{(k)}[n] + \eta \sqrt{(\xi^{(k)}[n])^{-1}} \Delta \mathrm{w}_{ij}^{(k)}[n] \tag{6.2.14}$$

where $(\sqrt{(\xi^{(k)}[n])^{-1}})$ is a normalisation factor and $\xi^{(k)}[n]$ is defined as:

$$\xi^{(k)}[n] = \beta \xi^{(k)}[n-1] + (1-\beta) \sum_{j=1}^{N} |x_j^{(k)}[n]|^2 / N \tag{6.2.15}$$

where $\beta \ \epsilon \ [0,1]$ is a smoothing factor which improves the robustness of the algorithm as it counts for the mean energy across the mixtures. Normally, a small positive arbitrary constant is added to $\xi^{(k)}[n]$ to avoid division by zero. In the following section, a new adaptive learning scheme, to enhance the separation performance for the online IVA algorithm, is introduced. Instead of a fixed learning rate, the proposed scheme controls the learning rate of the algorithm as a function of the proximity to the target solution.

### 6.2.3 Adaptive Online IVA Learning Algorithm

The gradient $\Delta \mathrm{w}_{ij}^{(k)}[n]$ converges to zero as $\Lambda_{il}^{(k)}[n]$ approaches $\Re_{il}^{(k)}[n]$ i.e. $(\Lambda_{il}^{(k)}[n] - \Re_{il}^{(k)}[n])$ approaches zero. We therefore assign:

$$g^{(k)}[n] = \left\| \mathbf{\Lambda}^{(k)}[n] - \boldsymbol{\Re}^{(k)}[n] \right\|_F \tag{6.2.16}$$

where $\| \cdot \|_F$ denotes the Frobenius norm.

The descending behaviour of $g^{(k)}[n]$ is utilised as a gear-shifting type operator for the learning algorithm. In the initial stages the learning rate is set to a high value to move faster towards the solution. Then it decreases as the system converges to reduce the fluctuations and improve stability. A new normalised learning rate at time frame $n$ is defined as:

$$\eta^{(k)}[n] = \eta_0 \frac{g^{(k)}[n]}{g^{(k)}[1]} \tag{6.2.17}$$

where $\eta_0$ is the initial learning rate. $\eta^{(k)}[n]$ will start with the initial value $\eta_0$ for the first frame and then it decreases as $n$ increases. In a non-stationary environment $g^{(k)}[1]$ could be reinitialised. Then $\eta^{(k)}[n]$ is smoothed:

$$\eta^{(k)}[n] = \lambda \eta^{(k)}[n-1] + (1-\lambda)\eta^{(k)}[n] \tag{6.2.18}$$

where $\eta \ \epsilon \ [0, 1]$ is a smoothing factor. The online update equation is adjusted accordingly as:

$$\mathrm{w}_{ij}^{(k)}[n+1] = \mathrm{w}_{ij}^{(k)}[n] + \eta^{(k)}[n]\sqrt{(\xi^{(k)}[n])^{-1}}\Delta\mathrm{w}_{ij}^{(k)}[n] \qquad (6.2.19)$$

The multivariate source priors used to model the speech source signals in the IVA algorithm are discussed in the following section.

## 6.3 Multivariate Source Priors

The nonlinear multivariate score function vector for the IVA, $\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}, \cdots, \hat{\mathbf{s}}_i^{(k)}) = [\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}), \cdots, \varphi^{(k)}(\hat{\mathbf{s}}_i^{(N)})]^T$, that maintains the dependencies between the frequency bins with each source vector, is given in the general form:

$$\varphi^{(k)}(\hat{\mathbf{s}}_i^{(1)}, \cdots, \hat{\mathbf{s}}_i^{(k)}) = -\frac{\partial \log q(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)})}{\partial \hat{\mathbf{s}}_i^{(k)}} \qquad (6.3.1)$$

The performance of the IVA algorithm greatly depends on the multivariate inter-dependency model used as a source prior. The particular nonlinear score function depends on the source prior selected to model the speech signals. The selection of a suitable score function is integral to the performance of the IVA method. In this work, two multivariate source priors are used for the online IVA algorithm, which will be discussed next.

### 6.3.1 Super Gaussian Source Prior

The IVA algorithm proposed in [20] defines the source prior as a dependent multivariate super-Gaussian distribution in the form:

$$p(\mathbf{s}_i) = \alpha \exp\Big( - \sqrt{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1} (\mathbf{s}_i - \boldsymbol{\mu}_i)}\Big) \qquad (6.3.2)$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are respectively the mean vector and covariance matrix of the $i$th source signal. This source prior indicates variance dependency between the frequency bins. Since the frequency bins are uncorrelated due to the orthogonality of Fourier bases, the mean vector $\boldsymbol{\mu}_i$ can be set to zero and the covariance matrix $\boldsymbol{\Sigma}_i$ to an identity diagonal matrix. Therefore, Equation (6.3.2) can be rewritten as:

$$p(\mathbf{s}_i) = \alpha \exp\left( - \sqrt{\sum_{k=1}^{K} \left| \frac{\hat{s}_i{}^{(k)}}{\sigma_i{}^{(k)}} \right|^2} \right) \tag{6.3.3}$$

where $\sigma_i{}^{(k)}$ is the standard deviation of the $i$th source at the $k$th frequency bin that determines the scale of each element of a source vector. Assuming unity standard deviation $\sigma_i^{(k)}$:

$$p(\mathbf{s}_i) = \alpha \exp\left( - \sqrt{\sum_{k=1}^{K} | \ \hat{s}_i{}^{(k)}|^2} \right) \tag{6.3.4}$$

The corresponding multivariate nonlinear score function vector to extract the $i$th source at the $k$th frequency is obtained as [20]:

$$\varphi^{(k)}(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)}) = - \frac{\partial \log\left( - \exp\left( \sqrt{\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2} \right) \right)}{\partial \hat{s}_i^{(k)}} \tag{6.3.5}$$

$$\varphi^{(k)}(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)}) = \frac{\partial \sqrt{\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2}}{\partial \hat{s}_i^{(k)}} = \frac{\hat{s}_i^{(k)}}{\sqrt{\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2}} \tag{6.3.6}$$

This is a multivariate function which takes into account the inter-frequency dependency between the source vectors in the learning process [20]. However, the form of the multivariate score function may vary based on different types of dependency. The following subsection introduces an alternative multivariate source prior which uses a generalized Gaussian distribution to model the source vectors.

### 6.3.2 Generalized Gaussian Source Prior

A new multivariate generalized Gaussian distribution was found to be suitable as source prior for IVA algorithm [89]. The family of multivariate generalized Gaussian distributions has the form:

$$p(\mathbf{s}_i) = \alpha \exp\Bigg( - \Big(\frac{1}{\alpha}\sqrt{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1}(\mathbf{s}_i - \boldsymbol{\mu}_i)}\Big)^{\beta}\Bigg) \tag{6.3.7}$$

The distribution parameters are set as $\beta = \frac{2}{3}$ and $\alpha = 1$, to yield a particular source prior in the form:

$$p(\mathbf{s}_i) = \alpha \exp\Bigg( - \sqrt[3]{(\mathbf{s}_i - \boldsymbol{\mu}_i)^H \boldsymbol{\Sigma}_i^{-1}(\mathbf{s}_i - \boldsymbol{\mu}_i)}\Bigg) \tag{6.3.8}$$

The source prior can preserve the dependency across the frequency bins within each source vector [20], similar to the original super Gaussian distribution used to derive the IVA algorithm. The distribution has heavier tails as compared with the original super Gaussian distribution used in [20] as illustrated by univariate distributions shown in Figure 6.1. It can better model certain types of statistically nonstationary speech signals [92] and is more robust to outliers present in such signals [89], which can achieve better separation performance. A multivariate generalized Gaussian distribution is plotted in Figure 6.2.

With the assumption that the mean vector $\boldsymbol{\mu}_i$ of the sources is zero and the covariance matrix $\boldsymbol{\Sigma}_i$ is an identity diagonal matrix (due to the

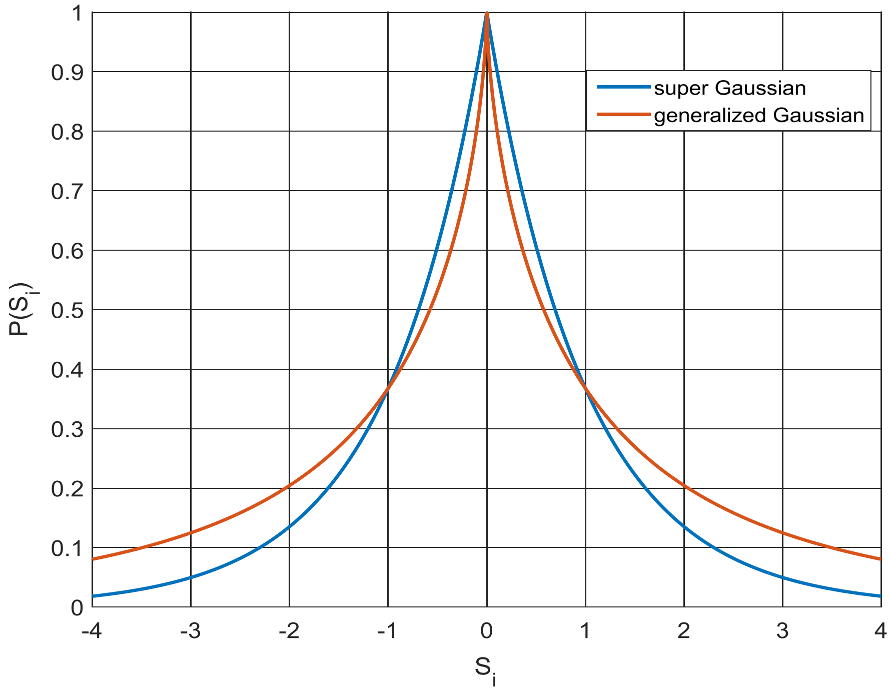


**Figure 6.1.** Univariate generalized-Gaussian distribution and univariate super-Gaussian distribution.

orthogonality of Fourier bases), Equation (6.3.8) can be rewritten as:

$$p(\mathbf{s}_i) = \alpha \exp\left( - \sqrt[3]{\sum_{k=1}^{K} \left| \frac{\hat{s}_i{}^{(k)}}{\sigma_i{}^{(k)}} \right|^2} \right) \tag{6.3.9}$$

Assuming unity standard deviation $\sigma_i^{(k)}$:

$$p(\mathbf{s}_i) = \alpha \exp\left( - \sqrt[3]{\sum_{k=1}^{K} | \ \hat{s}_i{}^{(k)} |^2} \right) \tag{6.3.10}$$

The proposed source prior is applied to derive the corresponding multivariate nonlinear score function vector to extract the $i$th source at the $k$th frequency is obtained as:

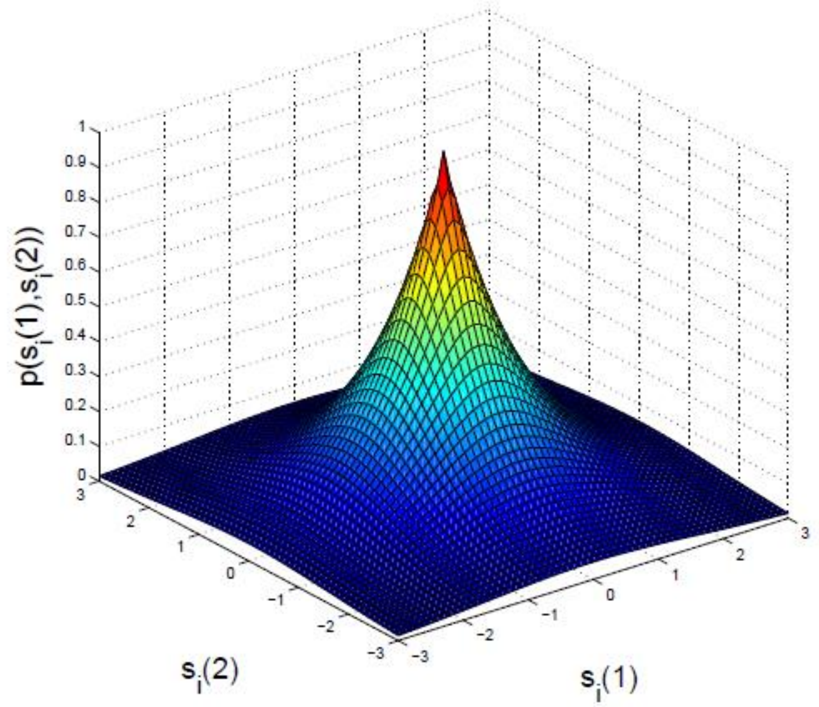


**Figure 6.2.** Bivariate generalized Gaussian distribution.

$$\varphi^{(k)}(\hat{s}_i^{(1)},\cdots,\hat{s}_i^{(k)}) = -\frac{\partial \log q(\hat{s}_i^{(1)},\cdots,\hat{s}_i^{(k)})}{\partial \hat{s}_i^{(k)}} \tag{6.3.11}$$

$$\varphi^{(k)}(\hat{s}_i^{(1)},\cdots,\hat{s}_i^{(k)}) = -\frac{\partial \log\left(-\exp\left(\sqrt[3]{\sum_{k=1}^{K}|\hat{s}_i^{(k)}|^2}\right)\right)}{\partial \hat{s}_i^{(k)}} \tag{6.3.12}$$

$$\varphi^{(k)}(\hat{s}_i^{(1)},\cdots,\hat{s}_i^{(k)}) = \frac{\partial \sqrt[3]{\sum_{k=1}^{K}|\hat{s}_i^{(k)}|^2}}{\partial \hat{s}_i^{(k)}} \tag{6.3.13}$$

$$\varphi^{(k)}(\hat{s}_i^{(1)},\cdots,\hat{s}_i^{(k)}) = \frac{2}{3}\frac{\hat{s}_i^{(k)}}{\sqrt[3]{(\sum_{k=1}^{K}|\hat{s}_i^{(k)}|^2)^2}} \tag{6.3.14}$$

The constant $(\frac{2}{3})$ can be absorbed by the learning rate $\eta$ in the update equation. The normalised nonlinear score function becomes:

$$\varphi^{(k)}(\hat{s}_i^{(1)}, \cdots, \hat{s}_i^{(k)}) = \frac{\hat{s}_i^{(k)}}{\sqrt[3]{(\sum_{k=1}^{K} |\hat{s}_i^{(k)}|^2)^2}} \tag{6.3.15}$$

The above derived nonlinear score function is multivariate which takes into account independence between the source vectors and dependency between the frequency bins within each source vector in the learning process. In addition, it introduces fourth order relationships between different frequency components within each source vector capturing more information describing the dependency structure. Therefore, it can better model the inter-frequency dependency which can improve the separation performance of the algorithm [65]. The experimental setup and the evaluation results of the performance of the new adaptive learning scheme for the IVA algorithm with both source priors are presented in the next section.

## 6.4 Experimental Results

Experiments were conducted to evaluate the separation performance of the online IVA algorithm with the proposed adaptive learning scheme using both source priors. A two-input (speaker) two-output (microphone) (TITO) system under spatially stationary conditions was adopted. The different algorithms are evaluated using simulated and real room impulse responses. The simulated room impulse responses (RIRs) are based on the image source method (ISM) [71]. As they are simulated responses, they do not characterise a real life room environment, but they are useful for initial comparative studies. For vigorous evaluation of the algorithms, they are further evaluated with real binaural room impulse responses (BRIRs) which were recorded in a real classroom environment with very high $RT_{60}$ of 565ms [74]. Real recorded speech signals, from the TIMIT acoustic-phonetic continuous speech corpus [66], were used as the source signals.

Speech signals obtained from the TIMIT database [66] were concatenated to form longer speech signals of length up to 300 seconds to allow for convergence. Two different randomly selected speech signals were convolved with both corresponding room impulse responses and then mixed to generate the mixture signals at the microphones. These mixtures were then separated using the different algorithms. The objective measure of signal to distortion ratio (SDR) [76] in decibels (dB) is used to evaluate the separation performance of the algorithms. The separation performance was evaluated in terms of the convergence time and the steady state SDR. The steady state is considered to be the average SDR of the last 50 seconds and the convergence time is defined as the time it takes the algorithm to reach 80% of the final steady state SDR. The accuracy of the steady state SDR is measured by the standard deviations from the average steady state SDR. It is assumed the speakers would remain stationary for the duration of the speech signals.

For all experiments, a 2048-point FFT with Hanning window and sampling rate of 8kHz were used to convert the time signals to the frequency domain to ensure it is sufficient to cover the time domain room impulse responses. The magnitude of the Frobenius norm $g^{(k)}[n]$ in Equation (6.2.16) is calculated as the average magnitude of the first 256 frequency bins at each time frame $n$. The smoothing factor $\beta$ in Equation (6.2.15) was chosen to be 0.5 as in [104] and the smoothing factor $\lambda$ in Equation (6.2.18) was empirically set to 0.99. For all algorithms the leaning rates $\eta$ and $\eta_0$ were set to the largest values that make the system converge fast whilst maintaining stability for all source positions. Larger values of $\eta$ may make the algorithm converge faster but produces high fluctuations in the steady state, which may lead to instability. The general experimental conditions are given in Table 6.1.

The evaluation results of the new scheme using both room impulse responses are compared with the original online IVA method [104] with both the super Gaussian source prior and the generalized source prior. The Student's t distribution, when adopted as a source prior for the online IVA, produced inconsistent results as compared with the batch IVA, due to the sensitivity of the distribution. To make the evaluation more reliable, each experiment was conducted ten times. The system performance was considered to be the average performance produced by the ten mixtures. For each set of speech signals the SDR values are averaged for both estimated source signals. Initial values of the separation filter matrix were chosen as identity matrix at each frequency bin.

**Table 6.1.** Experimental conditions for all room impulse responses

| | |
|---|---|
| Sampling rate | 8kHz |
| Length of the FFT | 2048 |
| Window type | Hanning |
| Source signal duration | 300 s (TIMIT) |
| Smoothing factor $\beta$ | 0.5 |
| Smoothing factor $\lambda$ | 0.99 |

### 6.4.1 Evaluation with the Image Source Method (ISM)

In these experiments, the room impulse responses are generated using the image source method (ISM) [71]. The $RT_{60}$ was set to 200ms. Fig-

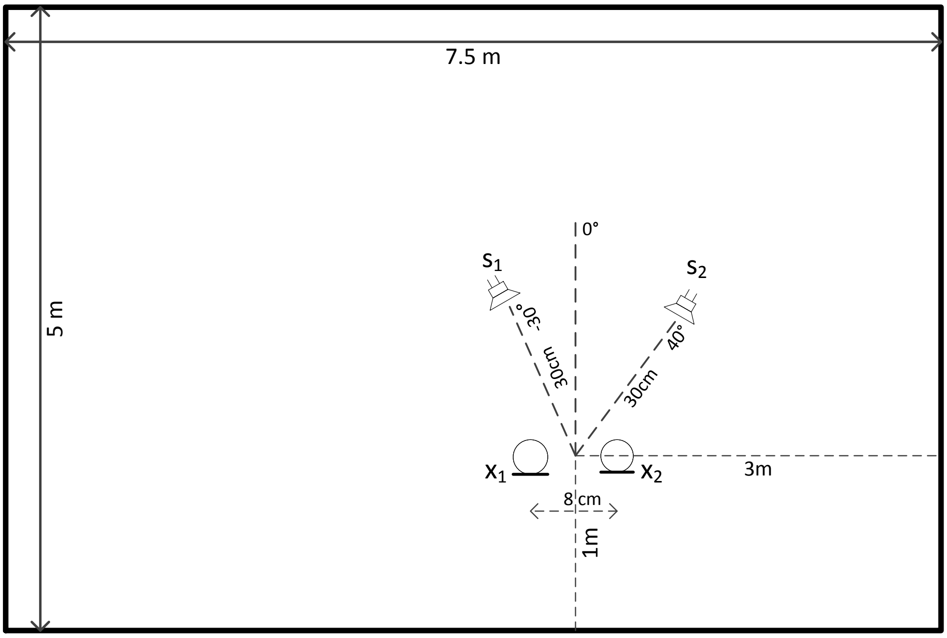


**Figure 6.3.** 2D plan of the simulated room environment showing the locations of sources and microphones. The heights of the sources and microphones are 1.5m.

ure 6.3 shows the environment of the simulated room and the locations of sources and microphones. The centre of the microphones was 1m and 3m away from each wall with inter-microphone distance of 8 cm. The locations of the sources were 30 cm at $-30°$ and $40°$ from the centre of the microphones. All the heights of the sources and microphones are 1.5 m. The leaning rates $\eta$ and $\eta_0$ were set to the largest values that maintain the system stability for all source positions. $\eta$ was set to 1.0 and $\eta_0$ to 3.0. The different experiment parameters for this method are given in Table 6.2. Only one source location is considered for this method. The purpose is to test the proposed scheme and the new generalized Gaussian source prior and compare the results with the original online IVA algorithm with the super Gaussian source prior. The detailed evaluation will be using real room environments in the following subsection.

**Table 6.2.** Experiment Parameters for ISM

| | |
|---|---|
| Room dimensions | 7m x 5m x 2.75m |
| Reflection coefficients | 0.79 |
| Reverberation time | 200 ms |
| Positions of Microphones | [2.96 1 1.5] m, [3.04 1 1.5] m |
| Inter-microphone distance | 8 cm |
| Positions of the Sources | $-30°, 40°$ |
| Source distance | 30 cm |
| Sound propagation speed | 343 m/s |
| $\eta$ for original method | 1.0 |
| $\eta_0$ for proposed scheme | 3.0 |

The source signals were separated from the generated mixtures using the online IVA algorithms with the proposed scheme using the super Gaussian and the generalized Gaussian source priors. The results were compared with the original IVA algorithm using both source priors. Figure 6.4 shows the behaviour of the proposed adaptive learning operator $\eta_{(k)}[n]$ as function of the frame number for one mixture. It starts at $\eta_0$ then it decreases exponentially with frame number as a function of the proximity to the solution. This demonstrates the validity of the scheme.

Figure 6.5 shows the SDR convergence plots for the proposed scheme as compared with the original online IVA algorithm using both source priors over a period of 100 seconds. The plots demonstrate a considerable separation performance improvement in terms of the convergence speed and the steady state separation performance. The proposed scheme provides faster convergence speed and higher SDR value with less fluctuations in both cases. Figure 6.6 compares the performance of the proposed scheme using both source priors. The plots show the generalized Gaussian source prior provides faster convergence whereas the

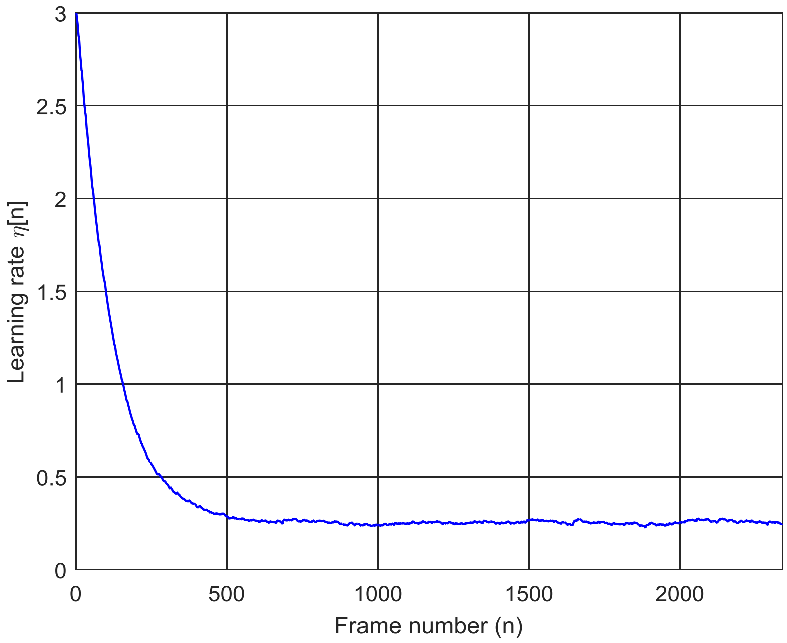


**Figure 6.4.** The behaviour of proposed adaptive learning operator as function of the frame number for one mixture.

super Gaussian source prior provides better steady state performance. The numerical results and analysis will follow.

The convergence time in seconds, steady state SDR in (dB) and steady state SDR standard deviation in (dB) for the online IVA algorithm with the proposed scheme are compared with the same measures of the original IVA algorithm using the super Gaussian and generalized Gaussian source priors. The results are shown in Table 6.3 and Table 6.4 respectively. The results exhibit consistent performance improvement of the proposed scheme in all three measures; convergence time, steady state SDR value and accuracy. The algorithm converges faster to a higher SDR value with smaller fluctuations using both source priors.

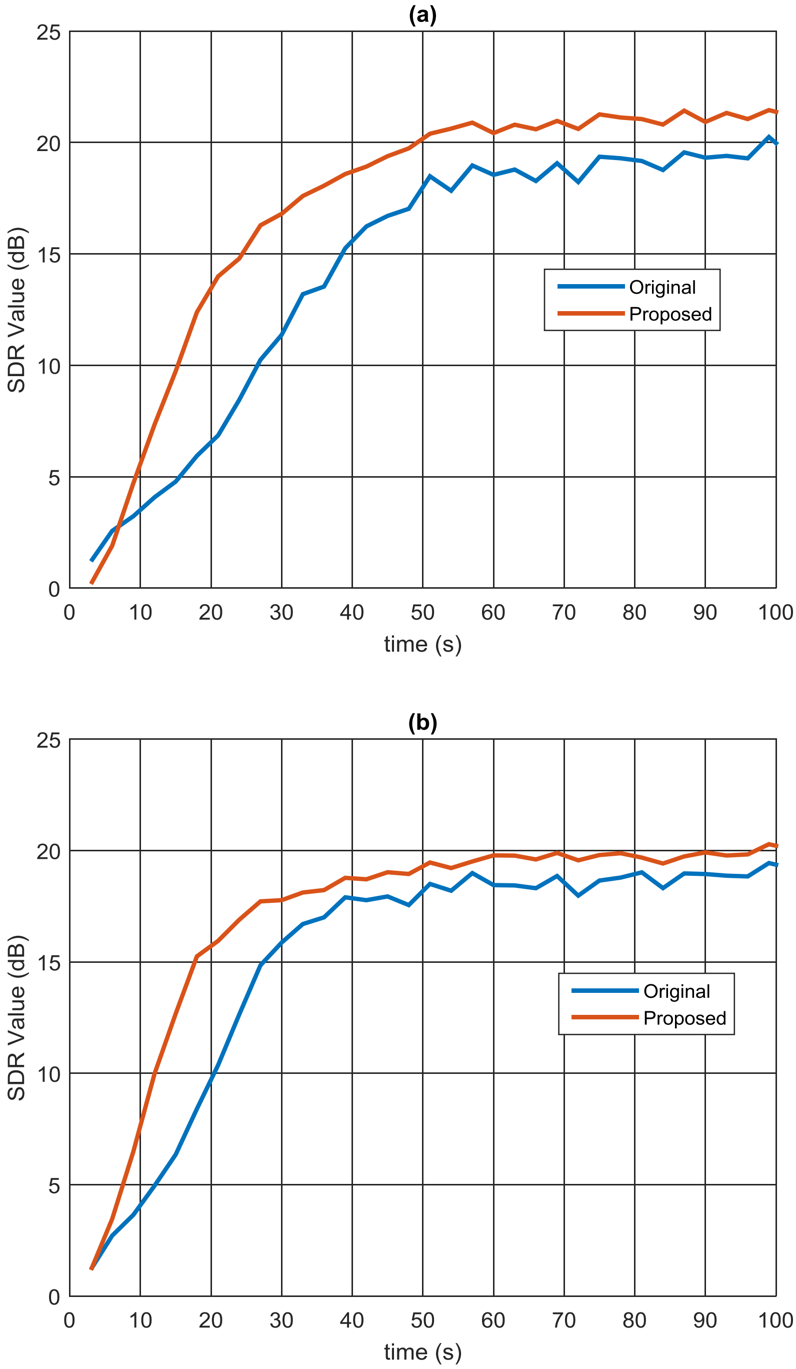


**Figure 6.5.** SDR convergence plots for the proposed scheme using each source prior over a period of 100 seconds (a) super Gaussian source prior (b) generalized Gaussian source prior.

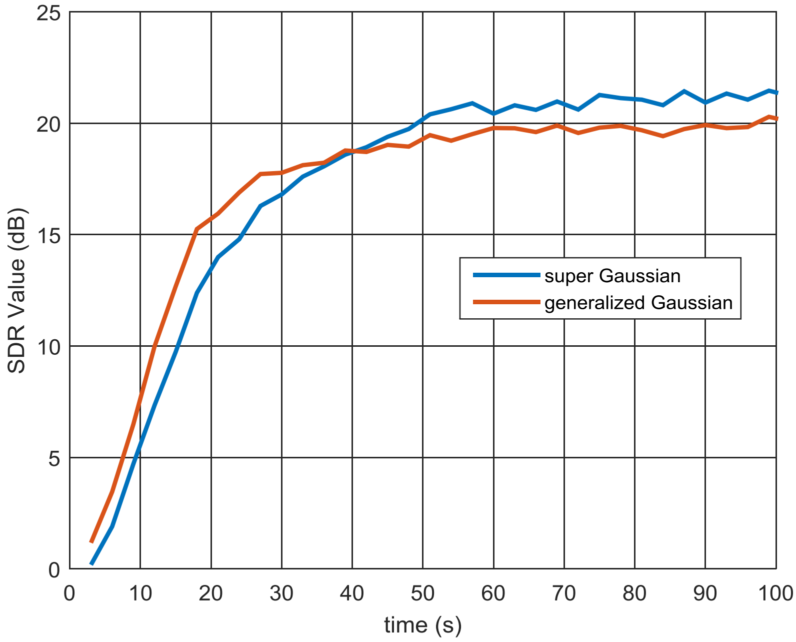


**Figure 6.6.** Comparison of SDR convergence plots using the proposed scheme with both source priors over a period of 100 seconds.

**Table 6.3.** Performance measures of the on line IVA with the proposed scheme and the original IVA method using the super Gaussian source prior.

| Algorithm | Original | Proposed | Improvement |
|---|---|---|---|
| Convergence time (s) | 40 | 26 | 15 |
| Steady State SDR (dB) | 19.97 | 21.11 | 1.14 |
| Standard deviation (dB) | 0.343 | 0.232 | 0.111 |

**Table 6.4.** Performance measures of the on line IVA with the proposed scheme and the original IVA method using the generalized Gaussian source prior.

| Algorithm | Original | Proposed | Improvement |
|---|---|---|---|
| Convergence time (s) | 27 | 18 | 9 |
| Steady State SDR (dB) | 19.34 | 20.38 | 1.04 |
| Standard deviation (dB) | 0.245 | 0.228 | 0.017 |

The online IVA with the proposed scheme using the super Gaussian source prior converges faster than the original IVA by approximately 15 seconds (35%). The average steady state SDR improvement is approximately 1.14 dB with standard deviation reduction of 0.111 dB (32%). With the generalized Gaussian source prior, it converges faster by 9 second (33%), taking into account the original IVA algorithm converges faster using the generalized source prior as illustrated in the plots and tables. The proposed scheme with the generalized Gaussian source prior converges faster than with the super Gaussian source prior by approximately 8 seconds (30%). The average steady state SDR improvement is approximately 1.04 dB with standard deviation reduction of 0.017 dB (7%). The scheme converges fastest using the generalized Gaussian source prior and accomplishes better steady separation performance using the super Gaussian source prior. Next, the algorithm is evaluated in real room environments.

### 6.4.2 Evaluation with Binaural Room Impulse Responses (BRIRs)

In these experiments, the room impulse responses are obtained from the BRIRs, which were recorded using a dummy head to simulate the effect of a human head in a real acoustic environment [74]. The BRIRs are real room recordings with very high $RT_{60}$ of 565ms. They provide realistic evaluation of the separation performance of BSS algorithms in highly reverberant environments. BRIRs were measured for 21 different relative source locations, consisting of all combinations of seven source azimuths $(0°, 15°, 30°, 45°, 60°, 75°, 90°)$ and three source distances (0.15m, 0.40m, and 1m) from the centre point between the ears of the head. All measurements were repeated on three separate occasions, with equipment taken down and reassembled in between. In order to increase the reliability of the experiments, the room impulse responses were averaged over the three measurements.

The room layout and experimental setup with the locations of sources and microphones are illustrated in Figure 6.7. The microphones were placed at the centre of the room with inter-microphone distance of 15cm. The sources were placed at 0.4 m from the centre of the microphones. The first source $s_1$ was placed at a fixed position perpendicular to both microphones at angle (0°) and the second source $s_2$ was placed at five different angles (15° to 75°), with 15° increment, relative to source $s_1$ in the room. Changing source positions represents speakers moving in the room which provide thorough evaluation of separation performance of the algorithms. The leaning rates $\eta$ and $\eta_0$ were set to the largest values that maintain the system stability for all source positions. $\eta$ was set to 0.5 and $\eta_0$ to 2 respectively. These values are lower than the values assigned for the ISM method due to the high reverberant real room environment. The summary of different parameters used in the experiments for this method is provided in Table 6.5.

**Table 6.5.** Experiment parameters for BRIRs.

| | |
|---|---|
| Room dimensions | 9 m x 5 m x 3.5 m |
| Reverberation time | 565 ms |
| Positions of Microphones | Centre of the room |
| Inter-Microphone distance | 0.15m |
| Source 1 position | 0° |
| Source 2 positions | 15°, 30°, 45°, 60°, 75° |
| Source distance | 0.4 m |
| $\eta$ for original method | 0.5 |
| $\eta_0$ for proposed scheme | 2.0 |

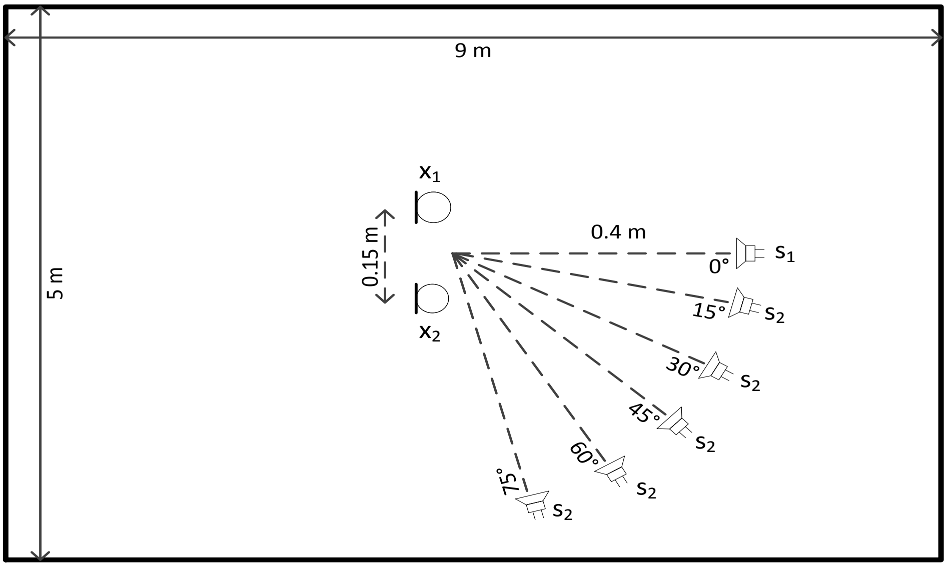


**Figure 6.7.** 2D plan of room environment and experimental setup for the BRIRs showing locations of sources and microphones.

The sources were separated from the generated mixtures using the various algorithms with the super Gaussian and the generalized Gaussian source priors. Figure 6.8 shows the behaviour of the proposed adaptive learning operator $\eta_{(k)}[n]$ as function of the frame number for one mixture. It decreases exponentially with frame number such that it starts at $\eta_0$ then it drops as a function of the proximity to the solution. This further confirms the validity of the scheme in real life scenarios. Figure 6.9 shows the SDR convergence plots for the proposed scheme as compared with the original online IVA algorithm using both source priors with source $s_2$ at angle 45° over a period of 100 seconds. The plots demonstrate a significant improvement in the convergence time as well as in the steady state separation performance in terms of SDR value and accuracy (smoothness). Figure 6.10 compares the performance of the proposed scheme using both source priors. The plots show the gen-

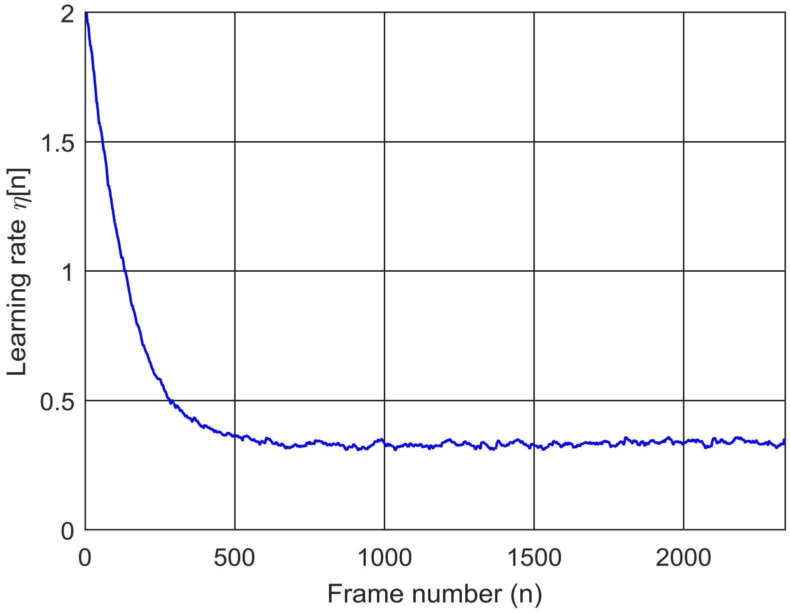


**Figure 6.8.** The behaviour of proposed adaptive learning operator as function of the frame number for one mixture.

eralized Gaussian source prior provides faster convergence whereas the super Gaussian source prior provides better steady state performance. The data for all source positions will presented in table format along with numerical analysis.

The convergence times in seconds, steady state SDR in (dB) and steady state SDR standard deviation in (dB) for different algorithms at different source angles are respectively shown in Table 6.6, Table 6.7 and Table 6.8.

The results exhibit consistent performance improvement of the proposed scheme in all three measures; convergence time, steady state SDR and accuracy. Table 6.6 shows consistent and considerable improved performance of the proposed scheme in terms of the convergence speed. It reduces the convergence time by approximately an average of 20.4 seconds (46%) using the super-Gaussian source prior (minimum

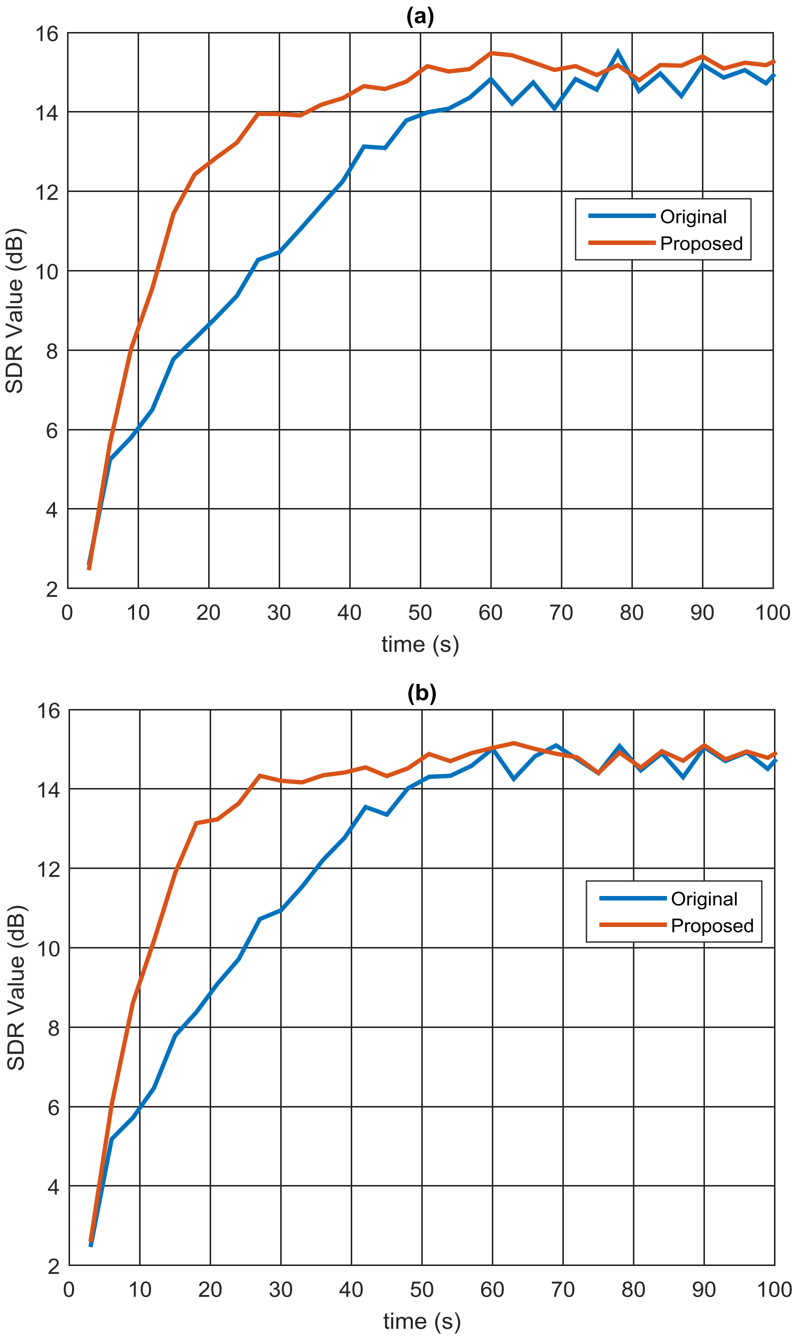


**Figure 6.9.** SDR convergence plots for the proposed scheme using each source prior with source $s_2$ at angle 45° over a period of 100 seconds (a) super Gaussian source prior (b) generalized Gaussian source prior.

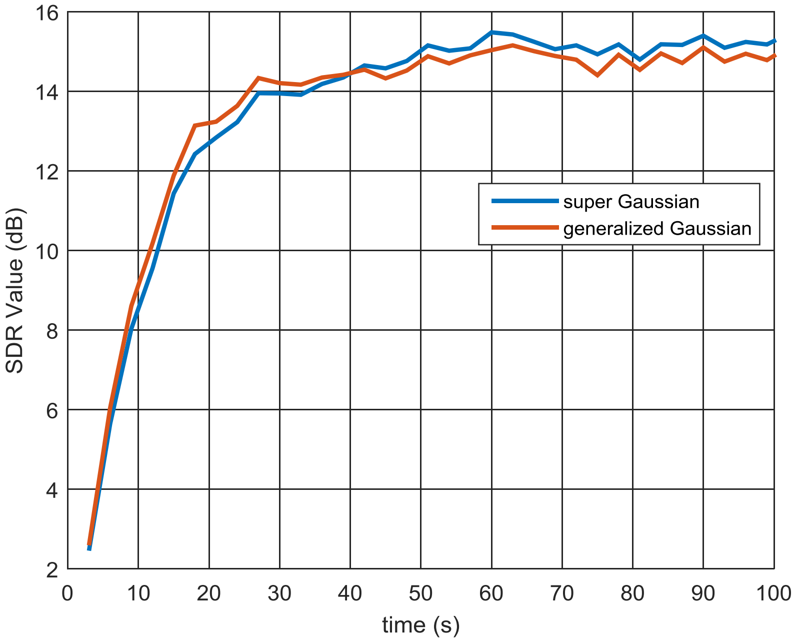


**Figure 6.10.** Comparison of SDR convergence plots using the proposed scheme for both source priors with source $s_2$ at angle 45° over a period of 100 seconds.

**Table 6.6.** Convergence time in seconds for various algorithms at different source $s_2$ positions

| **Algorithm** | **15°** | **30°** | **45°** | **60°** | **75°** | **Average** |
|---|---|---|---|---|---|---|
| s-Gaussian | 75 | 42 | 38 | 35 | 31 | 44.2 |
| s-Gaussian/Adaptive | 50 | 22 | 17 | 16 | 14 | 23.8 |
| Improvement | 25 | 20 | 21 | 19 | 17 | 20.4 |
| g-Gaussian | 75 | 40 | 35 | 30 | 25 | 41 |
| g-Gaussian/Adaptive | 40 | 17 | 15 | 14 | 14 | 20 |
| Improvement | 35 | 23 | 20 | 16 | 11 | 21 |

33% and maximum 55%) as compared with the original IVA algorithm and by an average of 21 seconds (51%) using the generalized Gaussian source prior (minimum 44% and maximum 57%) as compared with the original online IVA using the same source prior. The online IVA algorithm with the proposed scheme using the generalized Gaussian

**Table 6.7.** Average steady state SDR in (dB) for various algorithms at different source $s_2$ positions

| **Algorithm** | **15°** | **30°** | **45°** | **60°** | **75°** | **Average** |
|---|---|---|---|---|---|---|
| s-Gaussian | 9.25 | 13.20 | 14.94 | 15.82 | 16.33 | 13.91 |
| s-Gaussian/Adaptive | 9.39 | 13.44 | 15.25 | 16.10 | 16.68 | 14.17 |
| Improvement | 0.14 | 0.24 | 0.31 | 0.28 | 0.35 | 0.26 |
| g-Gaussian | 9.18 | 13.18 | 14.85 | 15.71 | 16.22 | 13.83 |
| g-Gaussian/Adaptive | 9.24 | 13.25 | 14.95 | 15.78 | 16.32 | 13.91 |
| Improvement | 0.06 | 0.07 | 0.10 | 0.07 | 0.10 | 0.08 |

**Table 6.8.** Average steady state SDR standard deviation in (dB) for various algorithms at different source $s_2$ positions

| **Algorithm** | **15°** | **30°** | **45°** | **60°** | **75°** | **Average** |
|---|---|---|---|---|---|---|
| s-Gaussian | 0.357 | 0.310 | 0.289 | 0.277 | 0.295 | 0.306 |
| s-Gaussian/Adaptive | 0.214 | 0.186 | 0.160 | 0.138 | 0.166 | 0.173 |
| Improvement | 0.143 | 0.124 | 0.129 | 0.139 | 0.129 | 0.133 |
| g-Gaussian | 0.312 | 0.322 | 0.294 | 0.292 | 0.305 | 0.305 |
| g-Gaussian/Adaptive | 0.203 | 0.211 | 0.190 | 0.158 | 0.153 | 0.183 |
| Improvement | 0.109 | 0.111 | 0.104 | 0.134 | 0.152 | 0.122 |

source prior converges faster than the original algorithm using the super Gaussian source prior by approximately 24.2 seconds (55%). The proposed scheme with the generalized Gaussian source prior converges faster than with the super-Gaussian source prior. The former is faster, on average, by 3.8 seconds (16%).

In terms of the steady state performance, the online IVA algorithm with the proposed scheme converges to a higher SDR value with smaller fluctuations using both source priors as illustrated by Figure 6.9 and Tables 6.7 and 6.8. This demonstrates the success of the adaptive learning scheme in reducing the learning rate as the algorithm convergence to the target solution. The average steady state SDR improvements are approximately 0.26 dB and 0.08 dB using the super-Gaussian source prior and the generalized Gaussian source prior respectively as compared with the original algorithm with the respective source priors. The online IVA algorithm with the proposed scheme using the super Gaussian source prior achieves better steady separation performance as compared with the original algorithm using the generalized Gaussian source prior by approximately 0.34 dB. The proposed scheme with the super Gaussian source prior outperforms that with the generalized Gaussian source prior in the steady state separation performance by approximately 0.26 dB.

The smoothness of the steady state SDR using the proposed scheme in Figure 6.9 and the data in Table 6.8 confirm the improved accuracy expressed in standard deviation. The average steady state accuracy provided by the scheme is approximately 0.173 dB and 0.183 dB using the super Gaussian source prior and the generalized Gaussian source prior, respectively, as compared with the original online IVA, where the average accuracy was approximately 0.306 dB and 0.305 dB. This yields approximate percentage improvement of (43%) and (40%). The accuracy of the super Gaussian source is slightly higher than that of the generalized Gaussian source prior.

The results show that the generalized Gaussian source prior outperforms the super Gaussian source prior in terms of the convergence speed and vice versa when it comes to the steady state performance. Based on these outcomes, a switching technique between the two source priors that acquires the best aspect of each distribution is proposed in the following section.

## 6.5 Switched Source Prior

The results obtained from evaluating the online IVA algorithm with the adaptive learning scheme using the super Gaussian source prior and generalized Gaussian source prior demonstrated a superiority of the first in terms of the steady state SDR and of the latter in terms of the convergence time. As a result, a new source prior switching technique is introduced. The technique initially starts the learning algorithm with the generalized Gaussian source prior to achieve faster convergence time and then switches to the super Gaussian source prior as the algorithm approaches the steady state to yield higher separation performance. This switching process is controlled by the adaptive learning scheme. While the adaptive learning rate $\eta^{(k)}[n]$ is still greater than a threshold value (a ratio of the initial learning rate $\eta_0$), the generalized Gaussian source prior is used. Once it reaches that threshold it switches to the super Gaussian source prior. The switched source prior takes the following form:

$$P(\mathbf{s}_i) = \begin{cases} f_{gG} & \text{if } \eta^{(k)}[n] \geq \eta_{TH} \\ f_{sG} & \text{otherwise} \end{cases} \tag{6.5.1}$$

where $f_{gG}$ is the multivariate generalized Gaussian distribution and $f_{sG}$ is the multivariate super Gaussian distributions. $\eta_{TH}$ is the threshold learning rate at the switching point:

$$\eta_{TH} = \alpha\eta_0 \tag{6.5.2}$$

where $\alpha$ is the ratio and set to around 0.25 (25%).

## 6.6 Experimental Results

The separation performance of the online IVA algorithm with the proposed adaptive learning scheme switched source prior is evaluated using different room impulse responses and settings. Firstly, the source prior switching technique is applied to the same date sets in Section 6.4 using the ISM method and the BRIRs. Then, the adaptive learning scheme and the switched source prior are further evaluated with real room impulse responses obtained from Hummerstone [75], which provide different real evaluation environments. The general experimental conditions are given in Table 6.1. The separation performance was measured objectively with SDR in dB.

### 6.6.1 Image Source Method (ISM)

The source prior switching technique is evaluated using the room impulse response based on the image source method ISM [71]. The algorithm was applied to the same experimental setup in subsection 6.4.1. The different experiment parameters for this method are given in Table 6.2. $\alpha$ in Equation (6.5.2) was set to 0.2. Figure 6.11 shows the SDR convergence plot using the proposed technique and the plots of the individual source priors from Figure 6.6. It is evident the plot takes the better performance features of each source prior i.e faster convergence time and higher steady state separation performance.

The new separation performance measures, namely the convergence time in seconds, steady state SDR in (dB) and steady state SDR standard deviation in (dB), for the online IVA algorithm with the proposed switched source prior compared with the measures of the individual source priors are shown in Table 6.9. The results demonstrate the su-

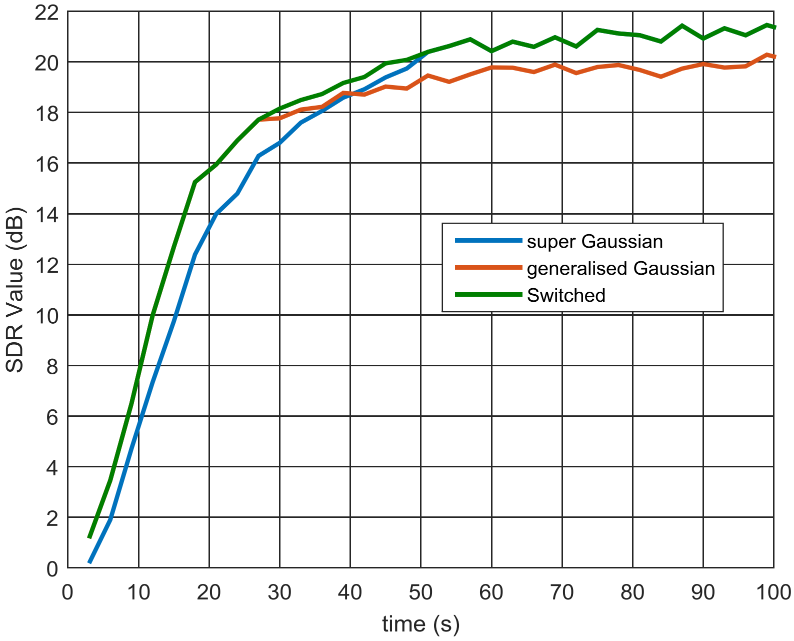


**Figure 6.11.** SDR convergence plot for the proposed scheme with switched source prior using ISM over a period of 100 seconds.

periority of the proposed technique in all three measures. Although the convergence time is slightly slower (by $2s$) than the generalized Gaussian source prior, it corresponds to higher steady state SDR and faster than the super Gaussian source prior.

**Table 6.9.** Performance measures of the on line IVA with the proposed scheme using all source priors.

| Source Prior | s-Gaussian | g-Gaussian | Switched |
|---|---|---|---|
| Convergence time (s) | 26 | 18 | 20 |
| Steady State SDR (dB) | 21.11 | 20.38 | 21.12 |
| Standard deviation (dB) | 0.232 | 0.228 | 0.230 |

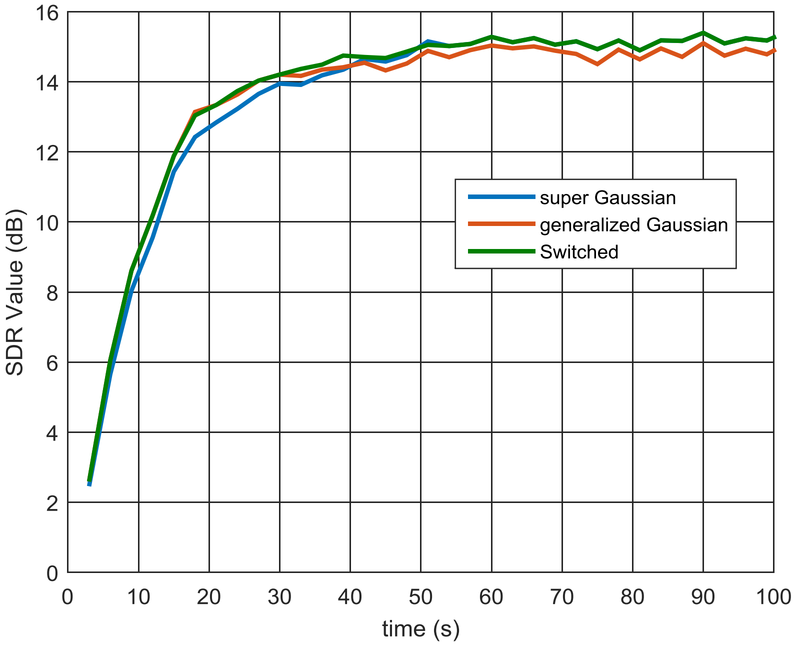


**Figure 6.12.** SDR convergence plot for the proposed scheme with switched source prior using BRIRs, source $s_2$ at angle 45° over a period of 100 seconds.

### 6.6.2 Binaural Room Impulse Responses (BRIRs)

The source prior switching technique is evaluated using the binaural real room impulse response (BRIR) [74]. The algorithm was applied to the same experimental setup in subsection 6.4.2. The summary of different parameters used in the evaluation for this method is provided in Table 6.5. $\alpha$ in Equation (6.5.2) was set to 0.25. Figure 6.12 shows the SDR convergence plot using the proposed technique and the plots of the individual source priors provided from Figure 6.10. It is evident the plot takes the better properties of each source prior i.e faster convergence time and higher steady state separation performance.

The average new separation performance measures, namely the convergence time in seconds, steady state SDR in (dB) and steady state

SDR standard deviation in (dB), for the online IVA algorithm with the proposed switched source prior compared with the measures of the individual source priors are shown in Table 6.10. The results demonstrate the superiority of the proposed technique in all three measures.

**Table 6.10.** Performance measures of the online IVA with the proposed scheme using all source priors.

| Source Prior | s-Gaussian | g-Gaussian | Switched |
|---|---|---|---|
| Convergence time (s) | 24 | 20 | 20.2 |
| Steady State SDR (dB) | 14.17 | 13.91 | 14.16 |
| Standard deviation (dB) | 0.173 | 0.183 | 0.175 |

### 6.6.3 Binaural Real Room Impulse Responses from (Hummerstone)

The proposed adaptive learning scheme with the switched source prior technique for the online IVA method is evaluated with other real room impulse responses, which were obtained from Hummerstone [75]. These are room impulse responses recorded in real life environment in four different rooms. The four rooms have different sizes, geometry and reverberation time $RT_{60}$. In each room, the source location azimuths relative to the second source can be varied from $-90^\circ$ to $90^\circ$, allowing for evaluation at different source positions around the room. Therefore, they offer broad evaluation environments for speech BSS algorithms over a range of experimental parameters and room settings. The types of the four rooms and their respective reverberation time $RT_{60}$ are shown in Table 6.11. The plans of the rooms can be found in Chapter 3.

**Table 6.11.** Room types and the respective $RT_{60}$. Hummerstone

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

Two different randomly selected speech signals, from the TIMIT database [66], were convolved with the corresponding room impulse response and then mixed to generate the mixture signals at the microphones. The mixtures were separated using the original online IVA algorithm, online IVA algorithm with adaptive learning the the super Gaussian source prior and with the generalized Gaussian source prior as well as the online IVA algorithm with the switched source prior. The separation performance of the different online IVA algorithm implementations is evaluated, compared and analysed. The performance measures adopted are the convergence time and the steady state SDR value and accuracy as defined the previous experiments. The experimental parameters, used for this method, are similar to the parameters used in the BRIR method provided in Tables 6.1 and 6.5 with the exception of the reverberation time and room layout.

In this set of experiments, the proposed adaptive learning scheme and the switched source prior demonstrated consistent performance improvement across the rooms and source locations. Figure 6.13 shows an example of the SDR convergence plots for the proposed scheme as compared with the original online IVA algorithm using both source priors over a period of 150 seconds. This example is selected from Room C with source $s_2$ at angle 45°. The plots exhibit large fluctuations due the higher reverberation time $RT_{60}$ of $680ms$ and different room design (cinemastyle lecture theatre). The plots show noticeable improvements in the convergence time and the steady state separation performance in terms of SDR value and accuracy (smoothness). This demonstrates the robustness of the scheme in different room settings including extremely difficult and highly reverberant environments.

Figure 6.14 shows the SDR convergence plot of the proposed switched source prior and the plots of the individual source priors. The plots confirm the generalized Gaussian source prior provides faster convergence whereas the super Gaussian source prior provides better steady state performance. They also confirm the success of the switched source prior as it achieves the best combination of performance measures. The various separation performance measures, the convergence time in seconds, steady state SDR in (dB) and steady state SDR standard deviation in (dB), for the various online IVA algorithms are compared in Table 6.12. The results demonstrate the superiority of the proposed technique in all three measures.

**Table 6.12.** Performance measures of the on line IVA with the proposed scheme using the switched source prior.

| **Algorithm** | Conv. time (s) | SS SDR (dB) | SS Std. Dev. (dB) |
|---|---|---|---|
| s-Gaussian | 49 | 11.52 | 0.483 |
| s-Gaussian/Adaptive | 33 | 11.97 | 0.343 |
| Improvement | 16 | 0.45 | 0.14 |
| g-Gaussian | 34 | 11.11 | 0.603 |
| g-Gaussian/Adaptive | 21 | 11.63 | 0.243 |
| Improvement | 13 | 0.52 | 0.36 |
| Switched | 21 | 12.07 | 0.283 |

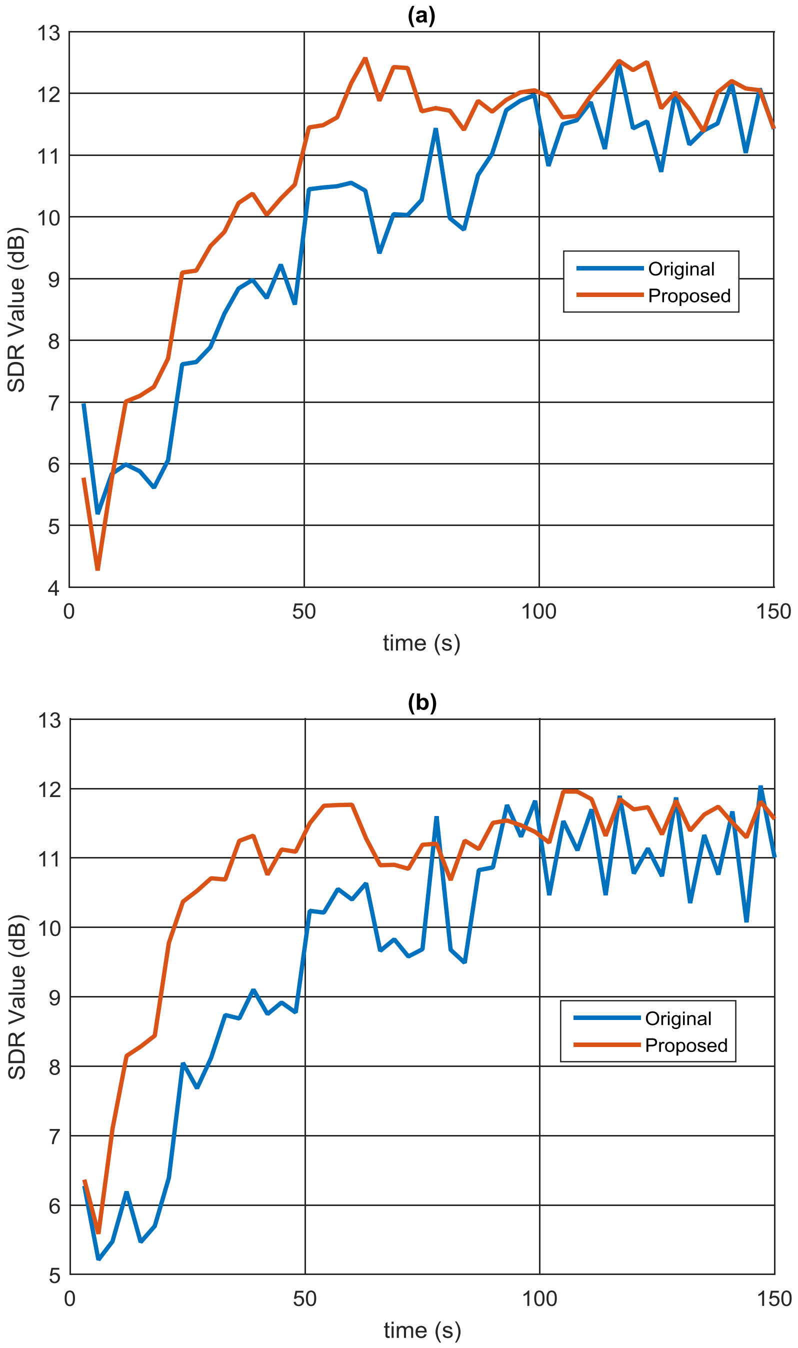


**Figure 6.13.** SDR convergence plots for the proposed scheme using each source prior in room C with source $s_2$ at angle 45° over a period of 150 seconds (a) super Gaussian source prior (b) generalized Gaussian source prior.

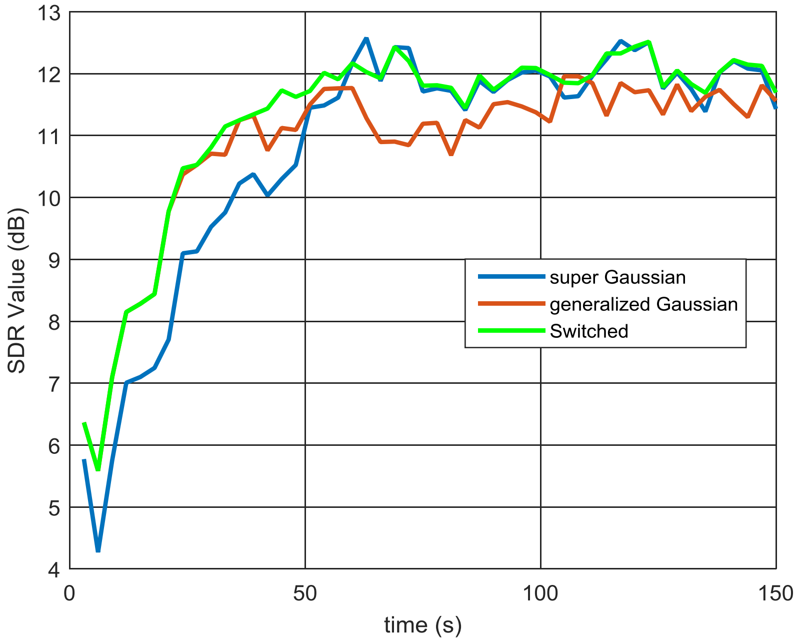


**Figure 6.14.** SDR convergence plot for the proposed switched source prior as compared with the individual source priors, in room C with source $s_2$ at angle 45°, over a period of 150 seconds.

The proposed switched source prior reduces the convergence time of the online IVA algorithm by approximately 28 seconds (57%) and by 13 seconds (38%), improves the steady state SDR by approximately 0.55 dB and 0.96 dB and reduces the error by approximately 0.2 dB (41%) and 0.32 dB (41%), as compared with original algorithm using the super Gaussian and generalized source priors, respectively.

## 6.7 Summary

An online BSS algorithm can be implemented in real time as the system receives the data which make it suitable for embedded systems. Online algorithms based on the ICA technique require a post processing stage to mitigate the permutation problem, which contributes to the computational load. The online IVA algorithm proposed in [104] deploys a multivariate sources prior to, algorithmically, solving the permutation problem by preserving dependency across the frequency bins within each source.

The online IVA method [104] uses a fixed learning rate to update the separation filter coefficients that imposes a trade-off between the convergence time and fluctuations in the steady state. In this chapter, a new robust adaptive learning based scheme was proposed to improve the separation performance of the online IVA algorithm. The scheme introduces a robust way to control the learning rate as the algorithm approaches the target solution. The new algorithm was implemented using the original super-Gaussian distribution [20] and a generalized Gaussian source [89] which was introduced to the online IVA for the first time.

The scheme was tested and compared with the original IVA algorithm in real room environments and real recordings. The experimental results have shown the new scheme yields faster convergence time together with higher and smoother steady state separation performance measured by SDR with both source priors, albeit with a small additional computational cost calculating the learning rate at every time frame. The results also demonstrated that the generalized Gaussian source prior outperforms the super Gaussian source in convergence speed and vice versa in steady state performance. Hence, the switched source prior was introduced to acquire the best aspect of each distribution. In the next chapter, conclusions are drawn from the thesis summarising the contributions and discussing suggestions for future work.

Chapter 7

# SUMMARY, CONCLUSION AND FUTURE WORK

## 7.1 Summary and Conclusions

The work presented in this study represents some promising strides towards the solution of the cocktail party problem (CPP) within the blind source separation (BSS) framework. The aim was mainly to add some novel contributions to enhance the performance of the independent vector analysis (IVA), including its different versions, in separating speech sources from their observed mixtures in real reverberant environments.

The main challenge to blind audio source separation (BASS) is the convolutive mixing of the sources in real room environments. This necessitates conducting the process in the frequency domain (FD) to avoid the computational complexity of the convolution operation in the time domain. In Chapter 2 background theory and fundamentals related to the subject of convolutive blind source separation (CBSS) were introduced. It also highlighted previous related work within the topic and their limitations. Independent component analysis (ICA), a prominent FD-BSS technique, was discussed which led to the permutation problem in FD-BSS. Then, the independent vector analysis

(IVA) algorithm, based on an improved model of the ICA method to address the permutation problem inherent to ICA, was reviewed in its natural gradient form (NG-IVA) and fast fixed point form (FastIVA). The heart of the IVA method is the multivariate source prior used to model the speech signals because the non-linear score function used to retain the inter-frequency dependency is obtained from the probability distribution function (PDF) of the source prior.

In Chapter 3, techniques and settings related to the implementation and evaluation of speech and blind audio source separation (BASS) systems were outlined. Different experimental setups were described, including information on datasets for speech sources, room environments and models as well as the performance parameters used in the evaluation criteria. The separation performance of the different algorithms was mainly measured objectively by signal to distortion ratio (SDR) in dB [76] or subjectively by perceptual evaluation of speech quality (PESQ) (on a scale of 0-4.5) [81] in simulated [71] and binaural real room impulse responses (BRIRs) [74, 75].

The contributions of this thesis satisfy the research objectives outlined in the introduction chapter. The objectives were addressed by introducing new methods to enhance the performance of the IVA algorithm in its various forms. The contributions can be summarised as follows:

1. A new multivariate Student's t distribution the source prior for the batch IVA algorithm.

2. A novel energy driven mixed distribution model as a source prior for the batch IVA algorithm.

**3.** A particular multivariate generalized Gaussian distribution as the source prior for the online IVA algorithm.

**4.** A novel adaptive learning scheme to improve the performance of the online IVA algorithm.

**5.** A novel switched source prior technique for the adaptive learning online IVA algorithm.

In Chapter 4, a new multivariate Student's t distribution is proposed as the source prior for the batch IVA algorithm. A Student's t PDF can better model certain frequency domain non stationary speech signals due to its tail dependency property. The tails of the distribution can be tuned to closely match the generally heavy tail distribution of the frequency domain speech signals due to the high amplitude data points. The chapter, initially, provided an experimental comparison between the batch versions of ICA and IVA. The results demonstrated the poor performance of the standard ICA due to the permutation problem and the IVA directly addresses the problem. Then, the separation performance of the IVA algorithm with the new source prior is compared with the original super Gaussian source prior in simulated and real room environments with a variety of settings. The experimental results confirmed that the proposed Student's t source prior consistently improves the separation performance of the IVA algorithm.

Using simulated room impulse responses [71], the average recorded SDR improvement using the new Student's t source prior was approximately 0.75 dB compared with the original IVA method. In real highly reverberant environment [74], the average recorded SDR improvement was approximately 1.31 dB compared with the original IVA method.

This confirms the suitability of the Student's t distribution to model speech signals in real life scenarios. The subjective study confirmed the improved separation performance for the IVA method with the Student's source prior. The average separation performance improvement PESQ score was approximately 0.75.

In Chapter 5, a novel multivariate source prior for the IVA algorithm was introduced. The proposed source prior is a mixture of two distributions, instead of a single distribution; namely the original multivariate super Gaussian distribution and the multivariate Student's t distribution. Human speech is highly non stationary with variable amplitude components. In the proposed mixed source prior, the Student's t distribution models the high amplitude components and the original super Gaussian distribution is used to model the lower amplitude components of the speech signal. Firstly, equal weights were assigned to both the original super Gaussian distribution and the Student's t distribution in the mixed source prior. Then, it was further enhanced with an energy driven scheme that adjusts the weight of each distribution according to the normalised energy of the observed mixtures at the frequency domain blocks of a clique based dependency model. As a results, the mixed source prior was able to adapt to different statistical properties of speech signals.

The fixed mixed source prior was adopted for the IVA and the FastIVA algorithms and compared with the original single super Gaussian source prior. The detailed experimental studies using simulated [71] and real room environment [74] with different reverberation times confirmed consistent separation performance improvement of the fixed mixed source prior based IVA. Table 7.1 shows the approximate average recorded SDR improvements of both algorithms in simulated and real room environments.

**Table 7.1.** Average SDR (dB) improvements of the IVA and the FastIVA algorithms with the fixed mixed source prior in simulated and real room environments.

| Algorithm | IVA | FastIVA |
|---|---|---|
| Simulated RIRs [71] | 0.92 | 1.15 |
| Real BRIRs [74] | 0.85 | 0.90 |

The energy driven mixed source prior with clique based dependency structure was evaluated in different simulated and real room environments and compared with the original single super Gaussian source prior. The results confirmed that this approach consistently further improved the separation performance of the IVA algorithm. Table 7.2 shows the approximate average recorded SDR improvements of the IVA algorithm in simulated [71] and two types of real room environments [74, 75]. The subjective study also confirmed the improved separation performance for the IVA method with the energy driven mixed source prior. The average separation performance improvement PESQ score was approximately 0.25. The energy driven source prior outperformed the fixed mixed source prior and both single super Gaussian and Student's t source priors.

**Table 7.2.** Average SDR (dB) improvements of the IVA algorithm with the energy driven mixed source prior in simulated and real room environments.

| RIR | SDR improvement (dB) |
|---|---|
| Simulated RIRs [71] | 1.0 |
| Real BRIRs (1) [74] | 1.0 |
| Real BRIRs (2) [75] | 0.6 |

In Chapter 6, a novel adaptive learning scheme was developed for online IVA algorithm to update the separation filter coefficients. The scheme automatically controls the learning rate as a function of proximity to the target solution to achieve a faster convergence and more accurate steady state solution. It starts with the highest possible learning rate and reduces it as the algorithm converges. In addition, a particular multivariate generalized Gaussian distribution was adopted as the source prior for the online IVA algorithm. The nonlinear score function derived from this source prior has an informative and strong dependency structure and thereby improves the separation performance.

The scheme was tested using the original super Gaussian and the generalized Gaussian source priors and compared with the original IVA algorithm in real room environments [74] with various experimental settings. The experimental results demonstrated the robustness of the new scheme in yielding faster convergence time as well as higher and smoother steady state separation performance with both source priors. The results also revealed the generalized Gaussian source prior outperforms the super Gaussian source in convergence speed and vice versa in steady state separation performance. This led to the introduction of a novel switched source prior technique that combines both advantages. The experimental results confirmed the success of the technique to achieve its objective. The approximate average recorded improvements of the adaptive learning online IVA algorithm with various source priors compared with the original online IVA are shown in Table 7.3.

The methods and techniques presented in the contribution chapters of this thesis were implemented in real reverberant room environments. They were also analysed and evaluated with real recoded speech signals

**Table 7.3.** Performance measures improvements of the online IVA with the proposed scheme using all source priors.

| Source Prior | s-Gaussian | g-Gaussian | Switched |
|---|---|---|---|
| Convergence time (s) | 20.4 (46%) | 24.2 (55%) | 24 (54%) |
| Steady State SDR (dB) | 0.26 | 0.08 | 0.25 |
| Standard deviation (dB) | 0.133 (43%) | 0.122 (40%) | 0.131 (43%) |

and a robust evaluation criteria. Therefore, the findings of the thesis can be considered a reliable resource for researchers to build on the suggested ideas and develop future pioneering solutions to the cocktail part problem.

## 7.2 Future Work

There are different potential areas of improvement to the work presented in this thesis. The techniques can be further enhanced and several topics could be further researched. Following are some suggestions for future work:

The performance of the IVA method depends mainly on the dependency model for speech signals. Therefore, alternative multivariate source prior distributions that would improve the separation of speech sources should be further investigated. Moreover, a stronger interfrequency dependency structure may mitigate the permutation problem and potentially improve the separation performance of the algorithm. In the future, other dependency structures needs to be exploited. For example, the cliques based approach could be improved by looking at different ways of grouping the frequency bins into bands that are more suitable for speech signals. Also using different source priors within frequency bins or bands [115] could be investigated.

In this work the degrees of freedom parameter, for the multivariate Students t source prior, was empirically selected. The estimation of the parameter, from only the mixtures, is a challenging task which makes it a potential area of study. As this parameter can be estimated for pure speech signals in different ways, such as the tail index estimation method [116], a potential solution for this problem is to provisionally separate the mixtures, then estimate the degrees of freedom for each source.

When the Student's t source prior was adopted for the online IVA method, the results were inconsistent due to the sensitivity of the distribution. Although good separation performance using the source prior by applying an empirical scaling factor was reported in [106], the applicability of the Student's t source prior for the online IVA needs further investigation.

The mixed super Gaussian Student's t source prior proved efficient in improving the separation performance of the IVA algorithm. It would be interesting to look at mixing different combinations of various distributions including the generalized Gaussian source prior adopted in this thesis. In addition using the mixture models with the online IVA.

The online IVA techniques proposed in this work were simulated in Matlab environment. It would be useful, to implement and test the performance of the enhanced online IVA methods as applications on real time embedded systems.

In this study, to convert the time signals to the frequency domain, a 1024-point FFT was considered for the batch IVA algorithm and 2048-point FFT for the online IVA algorithm. This was to ensure it is sufficient to cover the time domain room impulse responses and the IVA algorithm can maintain good SDR performance values at the outputs. To reduce the computational complexity of the algorithms, reducing the number of frequency bins while maintaining acceptable separation performance could be a subject of future work.

The performance of the IVA algorithm declines in high reverberant room environments [25]. Thus, using a dereverberation method, such as beamforming [117], linear prediction or other methods [118–122], as a preprocessing stage could help to reduce the effect of the high reverberation. However, these methods were mainly developed for one source applications, while the CPP has at least two sources. Combining dereverberation methods with the IVA algorithm is a potential area of research [123].

# References

[1] 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.

[2] S. Makino, T. Lee and H. Sawada, *Blind Speech Separation.* Springer, 2007.

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

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

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

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

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

[8] 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.

[9] 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.

[10] B. M. Sayers and E. C. Cherry, "Mechanism of binaural fusion in the hearing of speech," *J. Acoust. Soc. Am.*, vol. 29, no. 9, pp. 973-987, 1957.

[11] A. S. Bregman, *Auditory Scene Analysis: the perceptual organization of sound.* Cambridge: MIT Press, 1990.

[12] 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.

[13] D. T. Pham, "Blind separation of instantaneous mixture of sources via an independent component analysis," *Signal Processing, IEEE Transactions on*, vol. 44, pp. 2768-2779, 1996.

[14] N. Das, A. Routray, P. K. Dash and D. India, "ICA methods for blind source separation of instantaneous mixtures: A case study," *Neural Information Process.Letters and Reviews*, vol. 11, pp. 225-246, 2007.

[15] 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.

[16] 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.

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

[18] 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.

[19] A. Simpson, G. Roma, and M. D. Plumbley, "Deep karaoke: Extracting vocals from musical mixtures using a convolutional deep neural network," *ArXiv preprint arXiv:1504.04658*, 2015.

[20] 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.

[21] 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.

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

[23] 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.

[24] M. Z. Ikram and D. R. Morgan, "Permutation inconsistency in blind speech separation: Investigation and solutions," *IEEE Trans. Speech Audio Processing,* vol. 13, no. 1, pp. 1-13, 2005.

[25] S. Araki et al., "The fundamental limitation of frequency-domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio Process.,* vol. 11, no. 2, pp. 109-116, 2003.

[26] H. Sawada et al., "Spectral smoothing for frequency-domain blind source separation," In *Proc. Int. Workshop Acoustic Echo and Noise Control (IWAENC),* pp. 311-314, Kyoto, Japan, 2003.

[27] M. Girolami, "Noise reduction and speech enhancement via temporal anti-Hebbian learning," In *Proc. ICASSP, Seattle*, WA, USA, May 12-15, 1998.

[28] H. Sahlin and H. Broman, "Signal separation applied to real world signals," In *Proceedings of 1997 Int. Workshop on Acoustic Echo and Noise Control (IWAENC97)*, London, UK, September, 1997.

[29] T.-W. Lee, A. Bell, and R. Orglmeister, "Blind source separation of real world signals," In *Proc. ICNN*, Houston, TX, June, 1997.

[30] K.-C. Yen and Y. Zhao, "Robust automatic speech recognition using a multichannel signal separation front end," In *Proc. 4th Int. Conf. on Spoken Language Processing (ICSLP'96)*, Philadelphia, PA, October 1996.

[31] 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.

[32] 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.

[33] H. Pan, D. Xia, S. Douglas, and K Smith, “A scalable VLSI architecture for multichannel blind deconvolution and source separation,” In *Proc. IEEE Workshop on Signal Processing Systems*, Boston, MA, October, 1998.

[34] 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.

[35] D. Schobben and P. Sommen, “Transparent communication,” In *Proc. IEEE Benelux Signal Processing Chapter Symposium*, pp. 171-174, Leuven, Belgium, March, 1998.

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

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

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

[43] 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.

[44] A. van den Bos, “Complex gradient and Hessian,” IEE Proc. Vision Image and Signal Process, vol. 141, pp. 380-382. 1994.

[45] J. Hrault, C. Jutten and B. Ans, “Dtection de grandeurs primitives dans un message composite par une architecture de calcul neuromimtique en apprentissage non supervis,” In *Proc. GRETSI*, Nice, France, pp. 1017-1022, 1985.

[46] P. Comon, C. Jutten and J. Herault, “Blind separation of sources, Part II: Problems statement,” *Signal Process*, vol. 24, pp. 11-20, 1991.

[47] S. Roberts and R. Everson, *Independent Component Analysis: Principles and Practice*, Cambridge University Press, 2001.

[48] S. C. Douglas, “Blind separation of acoustic signals,” In *Microphone Arrays Techniques and applications*, Springer, 2001, pp. 355-380.

[49] S. C. Douglas and M. Gupta, “Convolutive Blind Source Separation for Audio Signals,” In *Blind speech separation*, pp. 3-45. Springer, Dordrecht, 2007.

[50] S. Haykin, *Adaptive Filter Theory*, 4th ed., Prentice Hall, 2001.

[51] J. Harris, *Online Source Separation in Reverberant Environments Exploiting Known Speaker Locations*, (PhD thesis, Loughborough University), UK, 2015.

[52] U. A. Lindgren and H. Broman, “Source separation using a criterion based on second-order statistics,” *Signal Processing, IEEE Transactions on*, vol. 46, pp. 1837-1850, 1998.

[53] K. Torkkola, “Blind separation for audio signals-are we there yet?” in *Proc. Workshop on Independent Component Analysis and Blind Signal Separation*, pp. 11-15, 1999.

[54] H. Wu and J. C. Principe, ”Simultaneous diagonalization in the frequency domain (SDIF) for source separation,” in *Proc. ICA*, pp. 245-250, 1999.

[55] A. Souloumiac, “Blind source detection and separation using second order non-stationarity,” in *Acoustics, Speech, and Signal Processing, ICASSP-95., International Conference on*, pp. 1912-1915, 1995.

[56] C. Simon, P. Loubaton and C. Jutten, “Separation of a class of convolutive mixtures: a contrast function approach,” *Signal Process*, vol. 81, pp. 883-887, 2001.

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

[58] 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.

[59] F. Asano, S. Ikeda, M. Ogawa, H. Asoh and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, pp. 204-215, 2003.

[60] M. Z. Ikram and D. R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 881-884, 2002.

[61] 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.

[62] L. Intae, T. Kim, and T. Lee. "Independent vector analysis for convolutive blind speech separation." In *Blind speech separation*, pp. 169-192. Springer, Dordrecht, 2007.

[63] T. Eltoft, T. Kim and T. Lee, "On the multivariate Laplace distribution," *Signal Processing Letters, IEEE*, vol. 13, pp. 300-303, 2006.

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

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

[66] J. S. Garofolo et al., “TIMIT acoustic-phonetic continuous speech corpus,” in *Linguistic Data Consortium*, Philadelphia, 1993.

[67] V. Zue, S. Seneff, and J. Glass, “Speech database development at MIT: TIMIT and beyond,” *Speech Communication*, Vol. 9, No. 4, pp. 351-356, 1990.

[68] M. Weintraub and L. Neumeyer, “Constructing telephone acoustic models from a high-quality speech corpus,” in *Proc. ICASSP94*, vol. 1, pp. 85-88, Adelaide, Australia, 1994.

[69] H. Kuttruff, *Room Acoustics*, CRC Press, 2009.

[70] E. ISO, “3382-2: 2008,” Acoustics. Measurements of room acoustics parameters, Part 2.

[71] 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.

[72] 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.

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

[74] 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.

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

[76] 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.

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

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

[79] S. M. Naqvi, Y. Zhang, T. Tsalaile, S. Sanei, and J. A. Chambers, "A multimodal approach for frequency domain independent component analysis with geometrically-based initialization," in *Signal Processing Conference, 16th European*, pp. 1-5, Lausanne, Switzerland, 2008.

[80] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "PESQ - the new ITU standard for end-to-end speech quality assessment," *109th Audio Engineering Society Convention*, pre-print no. 5260, September 2000.

[81] 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.

[82] P. Smaragdis, “Blind separation of convolved mixtures in the frequency domain,” *Neurocomputing*, vol. 22, no. 1, pp. 21-34, 1998.

[83] D. Yellin and E. Weinstein, “Criteria for multichannel signal separation,” *Signal Processing, IEEE Transactions on*, vol. 42, no. 8, pp. 2158-2168, 1994.

[84] J. Anemueller and B. Kollmeier, “Amplitude modulation decorrelation for convolutive blind source separation,” *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation,* pp. 215-220, 2000.

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

[86] A. Hiroe, “Solution of permutation problem in frequency domain ICA, using multivariate probability density functions,” *Lecture Notes in Computer Science,* vol. 3889, pp. 601-608, 2006.

[87] 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.

[88] 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.

[89] 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.

[90] S. Erateb, W. Rafique, M. Naqvi and J.A. Chambers, "Evaluation of Source Separation Algorithms, including the IVA algorithm with various source priors, using Binaural Room Impulse Responses," *International Conference on Mathematics in Signal Processing, 10th IMA*, Birmingham, UK, 2014.

[91] 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.

[92] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, 1978.

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

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

[95] S. Demarta and A. J. McNeil, “The t copula and related copulas,” *International Statistical Review*, vol. 73, pp. 111-129, 2005.

[96] 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.

[97] 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

[98] 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.

[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] 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 Janerio, 2016.

[101] 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,” *Proc. of Eusipco*, 2016.

[102] G.-J. Jang, I. Lee, and T.-W. Lee, "Independent Vector Analysis using NonSpherical Joint Densities for the Separation of Speech Signals," Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, vol. 2, pp. 629-632, 2007.

[103] R. Mukai, H. Sawada, S. Araki and S. Makino, "Blind Source Separation for moving speech Signals using blockwise ICA and residual crosstalk subtraction," *IEICE Trans. Fundam.*, vol. E87-A, no. 8, pp. 530-538, 2004.

[104] T. Kim, "Real-time independent vector analysis for convolutive blind source separation," *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 57, no. 7, pp. 1431-1438, 2010.

[105] T. Taniguchi, N. Ono, A. Kawamura, and S. Sagayama, "An auxiliary-function approach to online independent vector analysis for real-time blind source separation," in *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2014 4th Joint Workshop on*, IEEE, pp. 107-111, 2014.

[106] 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.

[107] J. Anemüller and T. Gramss, "On-line blind separation of moving sound sources," ICA, 1998.

[108] L. Parra and C. Spence, “On-line convolutive blind source separation of non-stationary signals,” *Journal of VLSI signal processing systems for signal, image and video technology*, vol. 26, no. 1-2, pp. 39-46, 2000.

[109] R. Mukai, H. Sawada, S. Araki, and S. Makino, “Real-time blind source separation for moving speakers using blockwise ICA and residual crosstalk subtraction,” in *Proc. ICA*, pp. 975-980, 2003.

[110] W. Baumann, B.-U. Kohler, D. Kolossa, and R. Orglmeister, “Real time separation of convolutive mixtures,” ICA ’01, pp. 65-69, 2001.

[111] S. Ding, J. Huang, D. Wei, and A. Cichocki, “A near real-time approach for convolutive blind source separation,” *Circuits and Systems I: Regular Papers, IEEE Transactions on*, vol. 53, no. 1, pp. 114-128, 2006.

[112] L. Oliva-Moreno, J. Moreno-Cadenas, L. Flores-Nava, and F. Gomez-Castaneda, “DSP implementation of extended infomax ICA algorithm for blind source separation,” in *Electrical and Electronics Engineering, 2006 3rd International Conference on*, IEEE, pp. 1-4, 2006.

[113] S. Erateb, W. Rafique, J. Harris, M. Naqvi and J.A. Chambers, “Enhanced Online Independent Vector Analysis Algorithm for Speech Separation using Adaptive Learning,” *International Conference on Mathematics in Signal Processing, 11th IMA*, Birmingham, UK, 2016.

[114] S. Erateb, M. Naqvi and J.A. Chambers, "Online IVA with Adaptive Learning for Speech Separation using Various Source Priors," *Sensor Signal Processing for Defence Conference (SSPD)*, pp. 74-78, London, UK, 2017.

[115] Y. Liang, J. Harris, G. Chen, S. Naqvi, C. Jutten, and J. Chambers, "Auxiliary Function Based Independent Vector Analysis Using a Source Prior Exploiting Fourth Order Relationships," *Proc. EUSIPCO*, 2013.

[116] 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.

[117] L. Wang, H. Ding, and F. Yin, "Combining superdirective beamforming and frequency-domain blind source separation for highly reverberant signals," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 113, 2010.

[118] 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.

[119] 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.

[120] 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.

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

[122] 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.

[123] T. Yoshioka, T. Nakatani, M. Miyoshi, and H. G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 69-84, 2011.