AUDIO SOURCE SEPARATION USING SIGNAL PROCESSING AND MACHINE LEARNING TECHNIQUES

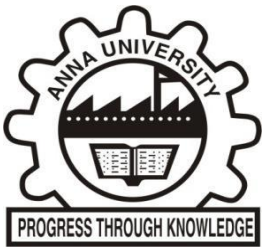
A THESIS

Submitted by

KUMAR M

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY



FACULTY OF INFORMATION AND COMMUNICATION ENGINEERING

ANNA UNIVERSITY

CHENNAI 600 025

JUNE 2020

ANNA UNIVERSITY

CHENNAI 600 025

BONAFIDE CERTIFICATE

The research work embodied in the present Thesis entitled **"AUDIO SOURCE SEPARATION USING SIGNAL PROCESSING AND MACHINE LEARNING TECHNIQUES"** has been carried out in the Department of Electronics and Communication Engineering, PSNA College of Engineering and Technology, Dindigul. The work reported herein is original and does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion or to any other scholar.

I understand the University's policy on plagiarism and declare that the thesis and publications are my own work, except where specifically acknowledged and has not been copied from other sources or been previously submitted for award or assessment.

KUMAR M
RESEARCH SCHOLAR

Dr. VE. JAYANTHI
SUPERVISOR
Professor
Department of Electronics and Communication Engineering
PSNA College of Engineering and Technology
Dindigul - 624 622

ABSTRACT

Blind Source Separation (BSS) is a signal processing technique used to extract the sources from their observed mixture signals without knowing the nature of mixing process and information about the sources. BSS is used in a broad range of applications including wireless communication, biomedical signal analysis, image processing and text document analysis. This work aims to separate the sources from real-time recorded speech mixture signals using suitable algorithms in a typical office noisy environment.

As a first step, FastICA algorithm that is capable to separate the sources from its overdetermined instantaneous mixture signals is analyzed. However the underdetermined convolutive mixture signals closely replicate the real-time recorded mixture signals. The FastICA algorithm involves some limitations in separating the real-time mixture signals. FastICA algorithm is not suitable for the separation of the real-time recorded mixture signals because of the presence of background noise sources and the convolutive mixing nature of acoustic sources. The convolutive mixture signals are converted into the frequency domain using Short Time Fourier Transform (STFT). The observed mixture signals become instantaneous mixture signals in the frequency domain. A complex FastICA algorithm is proposed to extract the sources from its overdetermined convolutive mixture signals overcoming the highlighted limitations suggested above. The Permutation problem occurred while processing the mixture signals in the frequency domain is solved by using a two-pass method with the correlation between the power ratios of the estimated source components. The proposed complex FastICA algorithm separates all the sources equally well when compared with the existing techniques. Also, the proposed permutation alignment technique reduces the misaligned frequency bins from 20% to 16%.

However, the proposed complex FastICA algorithm has the capability to separate the convolutive mixture signals when the number of microphones is at least equal to the number of sources. This condition cannot be satisfied in all practical situations. In order to overcome this limitation, this research contribution proposes two methods for the separation of the underdetermined convolutive mixture signals. The algorithms are mixing matrix estimation method using the Single Source Point (SSP) detection algorithm and Time-Frequency (TF) mask construction method using capsule networks respectively. Single source active points are identified using the normalized TF vectors of the observed mixture signals. The mixing matrix is estimated using Single Source Points (SSPs). As the matrix is non-square for underdetermined mixing conditions, the Moore-Penrose pseudo-inverse method is used to invert the mixing matrix. The Moore-Penrose method finds the inverse of the mixing matrix with the least square solution. In the second algorithm, capsule networks are used to learn the TF masks using SSPs and multi-source active points. The conventional Artificial Neural Networks (ANN) learns a particular feature by adjusting its weights for the given scalar inputs. The important drawback of the conventional ANN is that it is failed to learn the spatial relationship between the features. Whereas, the capsule networks accept the vector as input and produces the vector output based on the presence of the particular feature in the given input. The length of the vector represents the probability of the particular feature in the given input and the direction of the vector represents the spatial relationships between the features.

The proposed BSS methods are evaluated by comparing the original sources and estimated sources using the performance parameters of Source to Interference Ratio (SIR), Source to Distortion Ratio (SDR), and Source to Artifact Ratio (SAR). In principle, positive SIR, SDR and SAR values with

the equal SIR and SDR values imply that the separated sources are free from the interferences, noises and artifacts. The results demonstrate the efficiency of the proposed methods when compared with the state o art algorithms. The difference between the SIR and SDR values are varied from 10dB to 2dB for the existing techniques. The difference between the SIR and SDR values reduced to 0.44 dB and 0.01 dB in the proposed SSP detection algorithm and TF mask construction method respectively. The proposed TF mask construction method outperforms with increase of SIR and SDR values up to 1dB when compared with the SSP detection algorithm.

The research in overall provides contributions in terms of a permutation alignment technique based on the correlation between the power ratios of the estimated source components, the complex FastICA algorithm for the separation of the convolutive mixture signals, the SSP detection algorithm for underdetermined convolutive mixture signals and TF mask construction method using capsule networks. The use of capsule networks for the blind source separation process is a new initiative in the field of signal processing and paves the way for further improvement in this field.

ACKNOWLEDGEMENT

I owe a deep sense of gratitude to the Lord Almighty for blessing me with abundance help through various sources and whose grace was the moving spirit behind my effort in making this thesis a reality.

I am deeply indebted to my supervisor, **Dr.VE.Jayanthi**, Professor, Department of Electronics and Communication Engineering, PSNA College of Engineering and Technology, Dindigul, who has provided me with endless support, guidance and advice throughout my research. Her support and guidance have not only helped me to become a better researcher but has also helped me to become a better person. She has continuously inspired me as a person and my way of thinking.

I am extremely grateful to **Dr.J.William,** Professor and Head, Department of Electronics and Communication Engineering, Agnel Institute of Technology and Design, Goa, who has provided support to initiate this research. I am immensely grateful to the Doctoral Committee members **Dr.P.Palanisamy**, Professor, Department of Electronics and Communication Engineering, National Institute of Technology, Trichy and **Dr. N. Baskar**, Professor, Department of Mechanical Engineering, Saranathan College of Engineering, Trichy for their valuable inputs.

Special thanks must go to my Family Members, who have always been there for me and provided unconditional love and support throughout the highs and lows of my life. I truly love you all. Once again, I would like to thank all the above persons and many others whom I have not mentioned for being there for me throughout this research period in my life.

KUMAR M

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	**iii**
	LIST OF TABLES	**xii**
	LIST OF FIGURES	**xiii**
	LIST OF SYMBOLS AND ABBREVIATIONS	**xvii**
1	**INTRODUCTION**	**1**
	1.1 MOTIVATION	1
	1.2 TYPES OF MIXING PROCESS	1
	1.2.1 Instantaneous Mixing	2
	1.2.2 Convolutive Mixing	3
	1.3 TYPES OF BSS	5
	1.4 APPLICATIONS OF BSS	7
	1.4.1 Blind Identification Methods	7
	1.4.2 Telecommunication Applications	8
	1.4.3 Biomedical Applications	9
	1.4.4 Audio Applications	10
	1.5 CONTRIBUTIONS	11
	1.6 ORGANIZATION OF THE THESIS	12
2	**LITERATURE SURVEY**	**15**
	2.1 HISTORY OF BSS	15
	2.2 INDEPENDENT COMPONENT ANALYSIS (ICA)	16
	2.2.1 ICA model	17

CHAPTER NO.	TITLE	PAGE NO.
	2.2.2 JADE	18
	2.2.3 Infomax ICA	18
	2.2.4 FastICA	19
	2.2.5 Complex Valued ICA	20
	2.2.6 ICA for Convolutive Mixtures	21
	2.3 PERMUTATION ALIGNMENT TECHNIQUES	22
	2.3.1 DOA Based Permutation Alignment	22
	2.3.2 Correlation Based Permutation Alignment	23
	2.4 SPARSE COMPONENT ANALYSIS	24
	2.4.1 Degenerate Unmixing Estimation Technique (DUET) method	25
	2.4.2 TIme Frequency Ratio of Mixtures (TIFROM) Method	27
	2.4.3 Single Source Point (SSP) Detection Methods	29
	2.5 MACHINE LEARNING TECHNIQUES	32
	2.5.1 Supervised Learning	32
	2.5.2 Unsupervised Learning	35
	2.6 EVALUATION OF BSS ALGORITHMS	36
3	**OVERDETERMINED CONVOLUTIVE BLIND SOURCE SEPARATION USING COMPLEX FASTICA ALGORITHM**	**41**
	3.1 BSS OF INSTANTANEOUS SPEECH MIXTURE SIGNALS	41
	3.2 STATISTICAL INDEPENDENCE	43
	3.2.1 Information Theoretic Approach	43

CHAPTER NO.	TITLE	PAGE NO.
	3.2.2 Non-Gaussianity	44
	3.2.3 Whitening	45
	3.2.3.1 Kurtosis	46
	3.2.3.2 Negentropy	47
	3.2.3.3 Maximum likelihood estimation	49
3.3	OPTIMIZATION OF THE CONTRAST FUNCTIONS	51
	3.3.1 Simple Gradient Method	51
	3.3.2 Natural Gradient Method	51
	3.3.3 Newton's Method	52
3.4	FASTICA ALGORITHM	54
	3.4.1 Iteration Using Deflation Approach	54
	3.4.2 Iteration Using Symmetric Approach	55
3.5	DRAWBACKS OF EXISTING METHODS	56
	3.5.1 Noisy Sources	56
	3.5.2 Complex Sources	56
	3.5.3 Nonlinear Mixtures	57
	3.5.4 Permutation Ambiguity	57
	3.5.5 Scaling Ambiguity	58
3.6	PROPOSED METHOD	58
	3.6.1 Complex FastICA Algorithm	59
	3.6.2 Permutation Alignment Algorithm	61
3.7	RESULTS AND DISCUSSION	65
	3.7.1 Experimental Setup	65
	3.7.2 Results of Instantaneous BSS Using FastICA	65

CHAPTER NO.	TITLE	PAGE NO.
	3.7.3 Results of Convolutive BSS Using Complex FastICA Algorithm	72
	3.7.4 Results of Permutation Alignment Technique	74
	3.8 CONCLUSION	76
4	**UNDERDETERMINED CONVOLUTIVE BLIND SOURCE SEPARATION USING SINGLE SOURCE POINT DETECTION ALGORITHM**	**78**
	4.1 INTRODUCTION	78
	4.1.1 Basic Principles of Sparse Component Analysis	79
	4.2 EXISTING METHODS FOR UNDERDETERMINED BSS	80
	4.2.1 Single Source Point (SSP) Detection Methods	81
	4.2.2 Mixing Matrix Estimation Methods	84
	4.2.3 Source Estimation Methods	85
	4.3 DRAWBACKS OF EXISTING TECHNIQUES	86
	4.4 PROPOSED METHOD FOR UNDERDETERMINED BSS	87
	4.4.1 Underdetermined Instantaneous BSS Using SSP Detection Algorithm	87
	4.4.2 Underdetermined Convolutive BSS Using SSP Detection Algorithm	90
	4.5 RESULTS AND DISCUSSIONS	93
	4.5.1 Experimental Setup	93

CHAPTER NO.	TITLE	PAGE NO.
	4.5.2 Experimental Results for Underdetermined Instantaneous Mixtures	94
	4.5.3 Experimental Results for Underdetermined Convolutive Mixtures	100
	4.6 CONCLUSION	103
5	**UNDERDETERMINED CONVOLUTIVE BLIND SOURCE SEPARATION USING CAPSNET**	**105**
	5.1 INTRODUCTION	105
	5.1.1 System Model	106
	5.1.2 Deep Neural Learning	106
	5.1.3 TF Masks	109
	5.2 CAPSULE NETWORKS	111
	5.2.1 Dynamic Routing Between Capsules	115
	5.3 PROPOSED METHOD	117
	5.4 RESULTS AND DISCUSSIONS	121
	5.4.1 Experimental Setup	121
	5.4.2 Results	122
	5.5 CONCLUSION	127
6	**CONCLUSION AND FUTURE WORK**	**128**
	6.1 CONCLUSION	128
	6.2 FUTURE WORK	131
	REFERENCES	**132**
	LIST OF PUBLICATIONS	**144**

LIST OF TABLES

TABLE NO.	TITLE	PAGE NO.
3.1	Performance of the FastICA algorithm	68
3.2	Computational Complexity of the FastICA algorithm	69
4.1	Results of the underdetermined instantaneous BSS algorithm	96
4.2	Performance of the underdetermined convolutive BSS algorithm with microphone spacing of 5cm	100
4.3	Performance of the underdetermined convolutive BSS algorithm with microphone spacing of 1m	101
5.1	Comparison between the capsule and the neuron	114
5.2	Results of the proposed system for synthetic convolutive mixture signals with microphone spacing of 5cm	123
5.3	Results of the proposed system for synthetic convolutive mixture signals with microphone spacing of 1m	125

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE NO.
1.1	General instantaneous mixing processes	2
1.2	General convolutive mixing processes	4
1.3	Types of mixtures based on the number of sources and microphones	6
1.4	Types of BSS problems	7
2.1	Computation of capsule neuron output vector	35
2.2	DNN architecture for source separation	36
3.1	BSS of general instantaneous mixture signals	41
3.2	Permutation problem in the BSS	61
3.3	Proposed methods for BSS of overdetermined/determined convolutive mixing signals	61
3.4	The computation of amplitude correlation between the adjacent frequency bins	63
3.5	Proposed two-pass method for computation of adjacent frequency band correlation	64
3.6	Performance of the FastICA algorithm	67
3.7	Limitations of the FastICA algorithm	69
3.8	Time domain results of the FastICA algorithm	70
3.9	Computational complexity of the FastICA algorithm	71

FIGURE NO.	TITLE	PAGE NO.
3.10	Performance comparison of the complex FastICA algorithm	73
3.11	Performance of the complex FastICA algorithm for various room sizes	74
3.12	Results of the permutation alignment algorithms	75
3.13	Performance of the proposed permutation alignment technique	76
4.1	The flow of operations for underdetermined BSS	80
4.2	BSS of underdetermined instantaneous mixtures	87
4.3	BSS of underdetermined convolutive mixtures	90
4.4	Sparsity of the instantaneous mixture signals in the time domain and TF domain	94
4.5	Sparsity of the instantaneous mixture signals using SSPs only	95
4.6	Performance of the proposed algorithm for the male and female instantaneous mixture signals	97
4.7	Time domain representations of one of the sources and the corresponding extracted source	98
4.8	Comparison of the proposed underdetermined instantaneous BSS algorithm with state of art algorithms	99

FIGURE NO.	TITLE	PAGE NO.
4.9	Sparsity of the convolutive mixture signals in the time domain and TF domain	100
4.10	Performance of the proposed algorithm for the convolutive mixture signals recorded with 130ms reverberation	102
4.11	Performance of the proposed algorithm for the convolutive mixture signals recorded with 250ms reverberation	102
4.12	Comparison of the proposed underdetermined convolutive BSS algorithm with state of art algorithms	103
5.1	Two speaker separation based on DNN	111
5.2	Computation in the neuron and in the capsule	115
5.3	Capsule networks architecture for handwritten digit classification	116
5.4	Proposed method for underdetermined BSS using CapsNet	117
5.5	Illustration of the ratio of mixtures when sources are perfectly sparse	120
5.6	Scatter diagram of the output of the primary capsule layer	122
5.7	Performance of the proposed system for synthetic convolutive mixture signals with microphone spacing of 5cm	124

FIGURE NO.	TITLE	PAGE NO.
5.8	Performance of the proposed system for synthetic convolutive mixture signals with microphone spacing of 1m	126
5.9	Comparison of the proposed system with state of art algorithms	127

LIST OF SYMBOLS AND ABBREVIATIONS

ANN	-	Artificial Neural Network
BLSTM	-	Bi-directional Long Short Term Memory network
BSS	-	Blind Source Separation
CASA	-	Computational Auditory Scene Analysis
CNN	-	Convolutional Neural Network
DBN	-	Deep Belief Network
DDFC	-	Dynamic Data Field Clustering
DNN	-	Deep Neural Network
DOA	-	Direction Of Arrival
DTW	-	Dynamic Time Warping
DUET	-	Degenerate Unmixing Estimation Technique
EEG	-	Electoencephalogram
ECG	-	Electrocardiogram
EM	-	Expectation Maximization
FA	-	Factor Analysis
FDSOS	-	Frequency Domain Second Order Statistics
FIR	-	Finite Impulse Response
GMM	-	Gaussian Mixing Model
HF	-	High Frequency
ICA	-	Independent Component Analysis
IBM	-	Ideal Binary Mask
IRM	-	Ideal Ratio Mask
IVA	-	Independent Vector Analysis
JADE	-	Joint Approximate Diagonalization of Eigen matrices
KL	-	Kullback-Leibler
LSTM	-	Long Short Term Memory

LTI	-	Linear Time Invariant
MEG	-	Magnetoencephalography
ML	-	Maximum Likelihood
MLP	-	Multi Layer Perceptrons
MSP	-	Multi-Source Points
MUSIC	-	Multiple Signal Classification
NOSET	-	Number of Sources Estimation Technique
PCA	-	Principle Component Analysis
PR	-	Power Ratio
RFID	-	Radio Frequency Identification
RNN	-	Recurrent Neural Network
SAR	-	Source to Artifact Ratio
SCA	-	Sparse Component Analysis
SDR	-	Source to Distortion Ratio
SIR	-	Source to Interference Ratio
SMM	-	Spectral Magnitude Mask
SNR	-	Source to Noise Ratio
SSP	-	Single Source detection Points
STFT	-	Short Time Fourier Transform
TIFROM	-	Time Frequency Ratio Of Mixtures
TF	-	Time Frequency
UBSS	-	Underdetermined Blind Source Separation
UHF	-	Ultra High Frequency
W-DO	-	W Disjoint Orthogonal
WT	-	Wavelet Transform

CHAPTER 1

INTRODUCTION

1.1 MOTIVATION

Blind Source Separation (BSS) is a basic signal processing technique used to separate the original sources from the observed mixture signals without any prior knowledge about the sources. The term “blind” attests the fact that neither the mixing process nor the information about the sources known. From the introduction of the BSS concepts by Herault *et al.* (1986), a wide range of applications including biomedical signal processing, image processing, wireless communications, seismic signal analysis and text document analysis motivate the development of many algorithms for the separation of the signals from their simple instantaneous mixtures to complex convolutive nonlinear time variant mixtures. However, there are still developments needed to make these algorithms suitable for real complex mixing environments. The main challenges are: unknown number of sources, unequal number of sensors and sources, noisy environment, moving sources and nonlinear mixing process. There are number of papers available which address these problems using different approaches. (Cardoso 1997; Belouchrani *et al.* 1998; Oja 1998; Bingham *et al.* 2000a; Pham 2004; Yilmaz *et al.* 2004; Sawada *et al.* 2010; Chabreil *et al.* 2014; Negro *et al.* 2016; Kitamura *et al.* 2016)

1.2 TYPES OF MIXING PROCESS

Consider the case where two people are talking simultaneously in a room and the mixed signals are picked up by two microphones placed at two

different positions. This example is known as cocktail party problem and it is used often to describe the BSS problem.

The objective of the BSS problem is to separate the speech signals obtained from the microphone recordings without any prior knowledge about the sources, mixing process or microphone positions. In this particular example, the sources are acoustic signals, but it is not necessary for the sources to be restricted to speech. The sources are image or a signals and mixing process are the instantaneous or convolutive. The types of mixing process are described as follows.

1.2.1 Instantaneous Mixing

Instantaneous mixing process means, observed mixture signals at a given moment of time depend on the values of several source signals at the same time. Figure 1.1 depicts the instantaneous mixing process of 'j' number of unknown sources to obtain 'i' number of mixtures assuming that there is no additive noise. A single microphone output is mathematically written as

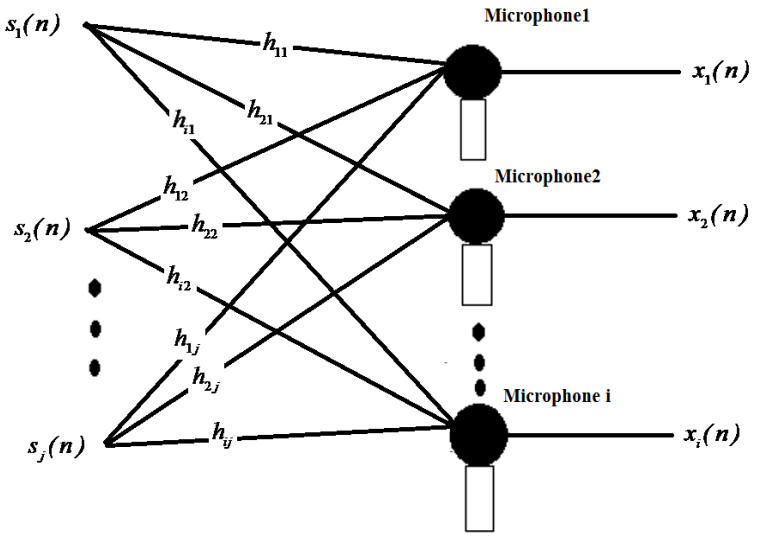


Figure 1.1 General instantaneous mixing processes

$$x_i(n) = h_{i1}s_1(n) + h_{i2}s_2(n) + ... + h_{ij}s_j(n) \quad (1.1)$$

where, $x_i(n)$ is the 'i' th microphone recording at the time index n. $s_1(n)$, $s_2(n)…s_j(n)$ are sources contributed for the mixture signal at the time index n. h indicates the path attenuation. For instance, h_{ij} is the path attenuation from 'j' th source to 'i' th microphone. The equation is simply written in vector-matrix form as

$$\boldsymbol{x(n) = As(n)} \quad (1.2)$$

where, **s(n)** = $[s_1(n), s_2(n), \ldots, s_j(n)]^T$ represents the original sources to be recovered from the observed microphone recordings **x(n)** = $[x_1(n), x_2(n), \ldots, x_i(n)]^T$ and **A** is called as mixing matrix of order i×j which is unknown. The superscript T represents the transpose operator. So the instantaneous BSS is formulated as extracting original sources **s(n)** from the observed mixture signals **x(n)** without knowing any information about **s(n)** and mixing matrix **A**.

$$\boldsymbol{A} = \begin{bmatrix} h_{11} & h_{12} & . & . & . & h_{1j} \\ h_{21} & h_{22} & . & . & . & h_{2j} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ h_{i1} & h_{i2} & . & . & . & h_{ij} \end{bmatrix} \quad (1.3)$$

1.2.2 Convolutive Mixing

Unlike instantaneous mixing process, there are multiple paths exist between any source to microphone in the convolutive mixing process. Therefore sources with different delays may contribute to mixture as the path

length varies between the source and microphone. Figure 1.2 illustrates the general convolutive mixing process assuming that there is no external noise.

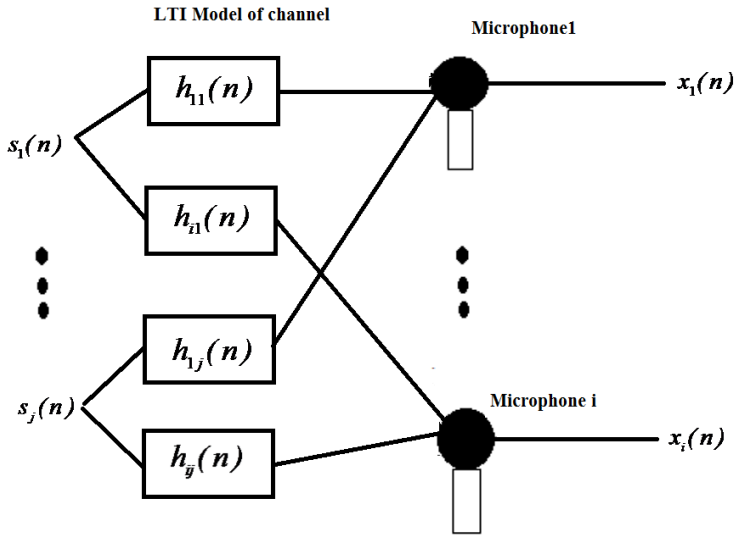


Figure 1.2 General convolutive mixing processes

The source signals s(n) = $[s_1(n), s_2(n), \ldots, s_j(n)]^T$ is unknown and unobservable and x(n) = $[x_1(n), x_2(n), \ldots, x_i(n)]^T$is the observed microphone recordings. The difference between the instantaneous mixing process and convolutive mixing process is that the path attenuation in the former is replaced by Linear Time Invariant (LTI) filter in the latter. The recording of one microphone is mathematically expressed as

$$x_1(n) = \sum_{k=0}^{\infty} h_{11}(k)s_1(n-k) + \sum_{k=0}^{\infty} h_{12}(k)s_2(n-k) + \ldots + \sum_{k=0}^{\infty} h_{1j}(k)s_j(n-k) \tag{1.4}$$

where $h_{ij}(n)$ is the impulse response of 'j' th source to 'i' th microphone. Even though equation (1.4) assumes the infinite length impulse response, the impulse response will perish after certain time duration, typically few seconds in a big auditorium and less than one second in normal room. When length of

LTI filters is finite; the channel is modelled as Finite Impulse Response (FIR) LTI filters. If the length of LTI filters is equal to one, the mixing becomes instantaneous. However, the length of impulse response of a reverberant room is assumed to be L (practically in the order of hundreds of milliseconds), then

$$x_1(n) = \sum_{k=0}^{L-1} h_{11}(k)s_1(n-k) + \sum_{k=0}^{L-1} h_{12}(k)s_2(n-k) + \ldots + \sum_{k=0}^{L-1} h_{1j}(k)s_j(n-k) \tag{1.5}$$

The equation can be simply written in vector-matrix form as

$$\boldsymbol{x(n) = A * s(n)} \tag{1.6}$$

where A is mixing matrix of order i×j and * denotes the convolution operator.

$$\boldsymbol{A} = \begin{bmatrix} h_{11}(n) & h_{12}(n) & . & . & . & h_{1j}(n) \\ h_{21}(n) & h_{22}(n) & . & . & . & h_{2j}(n) \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ h_{i1}(n) & h_{i2}(n) & . & . & . & h_{ij}(n) \end{bmatrix} \tag{1.7}$$

1.3 TYPES OF BSS

BSS may be classified as overdetermined/determined, underdetermined and single channel BSS based on the number of sources and microphones involved in the mixing processes. Figure 1.3 illustrates the types of mixtures based on the number of sources and microphones. Most of the work during the early period of BSS research is for overdetermined/determined instantaneous mixtures (no. of sources ≤ no. of mixtures), which resulted in many great algorithms for the separation of over determined and determined mixtures. The number of unknowns (sources) is less than the number of known (mixtures) in over determined mixture has

mathematically tractable problem. Whereas, the number of unknowns (sources) are greater than the number of known (mixtures) in underdetermined mixture is mathematically ill posed problem. However, it is potential to separate the sources from underdetermined mixture by utilizing some prior information about the sources such as sparseness property of the speech sources. Single channel mixture is a special case of underdetermined mixture and it is a challenging problem in BSS as it is mathematically intractable.

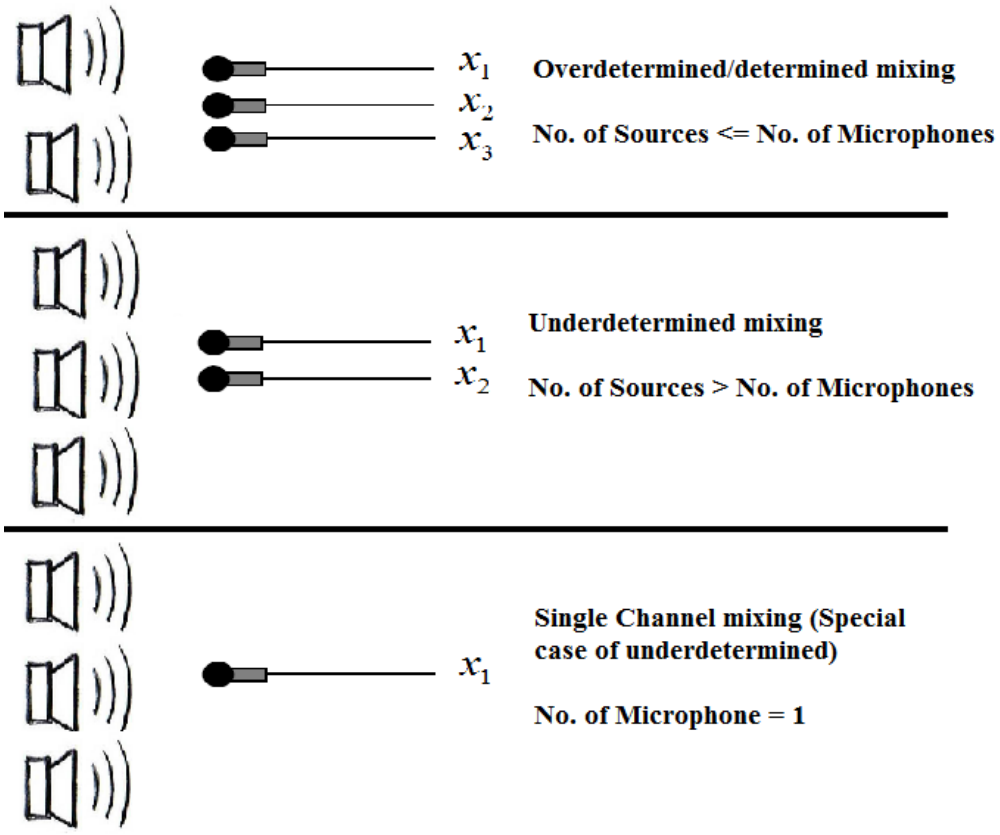


Figure 1.3 Types of mixtures based on the number of sources and microphones

Figure 1.4 consolidates the various types of BSS problems based on the mixing process as well as the number of sources and mixtures

involved in the mixing process. The convolutive mixtures are converted into frequency domain to utilize the strategic of instantaneous BSS solution. Time domain solution to the convolutive mixtures involves complex computations and instability problems.

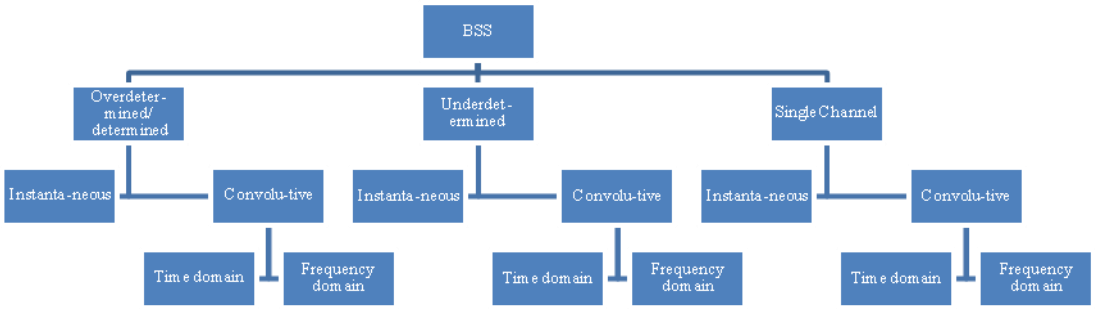


Figure 1.4 Types of BSS problems

1.4 APPLICATIONS OF BSS

In each application domain, generally BSS requires some priors knowledge and hence BSS turns out to be semi-bind source separation. While designing efficient BSS algorithms, it is important to make use of these priors. Finally, the results are appropriate only if such priors are fulfilled by extracted sources.

1.4.1 Blind Identification Methods

Blind identification refers to the recognition of people, objects, and animals in the everyday-life situations. The approach used to perform such identifications is based on some electronic systems like Radio Frequency Identification (RFID). RFID recognizes the object based on the demodulation of tag input to the field of the RFID receiver. However, when two tags are simultaneously positioned in the RF field of the receiver, the demodulated

signal is a mixture of two components and cannot be decoded by the receiver. Therefore this system is unable to recognize the simultaneously present objects/people.

This drawback is avoided based on the BSS techniques by replacing the receiver with two antennas and two decoders which produce two different mixture signals. These mixture signals are processed by the BSS technique to identify the simultaneously presented tags. The separation of high number of tag signals (more than two) are achieved by having the same receiver with two antennas and two decoders with underdetermined BSS technique.

The sources are the content of tag memory and therefore may be considered as statistically independent sources. Moreover, the RFID receiver reads the source from the tags that are close to it (within few cm). So the ideal model of the blind identification system is corresponding to the simplest form of instantaneous BSS. If the number of tags simultaneously presented to the receiver is at most two, instantaneous determined BSS model is used to separate the mixture of signals. Otherwise, instantaneous underdetermined BSS is required to separate the mixture of signals.

1.4.2 Telecommunication Applications

There are sharp rises in the utilization of wireless communications for civilians as well as surveillance applications in the last few decades. This increasing development of radio communication generates the increasing need of spectrum monitoring by state control agencies. The spectrum control problem blindly analyzes all the received signals which are detected in the band of frequencies at the receiver. This requires a pre-processing of the data

in isolating the different components of the observed mixture signals, i.e. Blind Source Separation (BSS).

Moreover, the sources propagate through multi-path propagation channels due to multiple reflections on ionospheric layers (HF), on natural scatterers such as mountains, etc and on buildings (UHF in urban areas). Spectrum control systems are generally to use convolutive BSS technique. Besides, the number of sources is unknown in many practical situations. In such cases, overdetermined/determined convolutive BSS are not suitable for practical conditions and hence underdetermined convolutive BSS technique is more suitable for this mixing model.

1.4.3 Biomedical Applications

Advances in digital signal processing and data recordings technologies have enabled recording and analysis of vast amounts of multidimensional biomedical data. The ultimate goal is to extract crucial features from the recorded data. The use of BSS in biomedical systems is now prevalent, since it decomposes the data as individual sources. However, biomedical signals pose a challenge as it is often difficult to determine original sources, that could be used to evaluate the accuracy of the BSS technique, or even its meaning. Therefore, it is difficult to model the mixing process as either an instantaneous or convolutive one. Similarly, it is difficult to fix the number of sources.

EEG and MEG are two techniques that measure the scalp electric potentials and the magnetic fields respectively. Indeed, these signals are very often recorded in the presence of noise signals, which are decomposed into internal and external noises. The external noise is the system noise. The internal noise includes all normal physiological activities that may generate

electrical currents but that are not required for the study. A common example of physiological signal which is not required but present, while recording EEG signal is the electric potential linked to movements of the eyes and eye blinks. BSS is employed to remove the unwanted physiological signals from EEG signal.

The electrocardiogram (ECG) records the electrical activity of the heart and it is recorded with surface electrodes, placed on the chest, arms and limbs. The standard 12-lead ECG is widely used by the physician when waveform morphology is required. Although ECG recording techniques are very effective, the noises and artifacts present in the recorded ECG signals. Indeed, in many practical situations, the ECG signal is affected by different types of noise sand artifacts, such as sinusoidal 50/60Hz power-line, electrode movements and broken wire contacts, but also interfering physiological signals such as muscle movements and breathing. Therefore, it is natural to consider BSS technique for ECG analysis. This includes the removal of artifacts and noises from the recorded ECG signal and extracting fetal ECG from the maternal recordings.

1.4.4 Audio Applications

Audio signal processing is one of the earliest fields in the blind source separation problem. Indeed, most available audio signals are mixtures of several sources. Although good separation may currently be obtained for some simple synthetic instantaneous as well as convolutive mixtures. The separation of real-world signals in this field remains difficult to achieve. It should therefore be stressed, that acoustics is among the most difficult application fields of source separation under research.

The mixing process results from the simultaneous transmission of acoustic signals through air from the sources to the microphones. The microphones convert the acoustic signals into electric signals. In normal situations, this mixing process is linear. The effect of acoustic-electric conversion by microphone is merged with the propagation mechanism. The output of each microphone is superposition of the sources contributed. The contribution of the each source to the microphone is determined by the convolution of the original source signal and impulse response of the channel from source to microphone. Hence the acoustic source separation may be considered as convolutive BSS.

1.5 CONTRIBUTIONS

This dissertation provides the theoretical background for the blind source separation problems and their solutions. A common evaluation strategy is used to compare the performance of the BSS algorithms. Acoustic source separation is one of the difficult separation problems which is considered for conducting experiments. Specific contributions of this thesis are:

1. Complex FastICA algorithm is proposed to separate the overdetermined convolutive mixture signals in the frequency domain.

2. The permutation alignment algorithm is proposed to solve the permutation problem that occurred in the frequency domain processing based on the correlation of power ratio of mixture signals. A two-pass method is used to control the misaligned frequency bins are guiding the subsequent frequency bins for misalignment.

3. A single source detection point (SSP) algorithm is proposed for underdetermined convolutive mixture signals. SSPs are identified for each frequency bins. These SSPs are used to estimate mixing filter coefficients based on K means clustering algorithms. The Pseudo-inverse of mixing filter coefficients are used to extract the original sources from underdetermined convolutive mixture signals.

4. A machine learning technique using capsule network is proposed for underdetermined convolutive BSS. Time Frequency (TF) mask is constructed based on the learning of capsule networks and then it is used to extract the sources. The results are compared with SSP based BSS algorithm.

1.6 ORGANIZATION OF THE THESIS

This thesis mainly addresses three problems in BSS; the first one is BSS of the overdetermined convolutive mixtures, the second one is a solution to the permutation problem occurred in the frequency domain processing of convolutive mixtures and the third one is BSS of underdetermined convolutive mixtures using sparse component analysis and capsule networks. In this thesis, complex FastICA algorithm is proposed to separate the sources from overdetermined convolutive mixtures. The algorithm uses maximum likelihood function as a measure of statistical independence between the extracted sources and it is an extension of popular FastICA algorithm to handle the complex values. The BSS of overdetermined convolutive mixtures are importance because of convolutive mixtures is more closely replicates the real environment of acoustic signals than instantaneous mixtures. The permutation problem is the major drawback of the frequency domain processing of the convolutive mixtures. In this thesis, a two pass algorithm

proposed based on the correlation of power ratio of adjacent frequency bins to the permutation problem.

The separation of sources from their mixtures, when the number of microphones is smaller than the number of sources is practically importance. Two methods are proposed in this thesis to extract the sources from the underdetermined convolutive mixtures. They are (i) single source detection point algorithm based on the sparse component analysis and (ii) construction TF mask based on unsupervised machine learning technique using capsule networks. Finally, the performance of the both methods is compared in terms of the separation accuracy.

The thesis is organized as follows. Chapter 1 introduces various types of BSS and its applications. Chapter 2 reviews the various approaches used in the literature for the BSS. The review of the FastICA algorithm and its limitations in solving the real recording of mixture signals are presented in chapter 3. Further, complex FastICA algorithm is proposed along with the solution to the permutation problem is occurred in the course of frequency domain processing of convolutive mixtures. In chapter 4, single source detection point algorithm is proposed for the estimation of mixing matrix for underdetermined convolutive mixtures. As the mixing matrix of underdetermined case is non square, the sources are extracted by applying pseudo inverse of mixing matrix to the mixtures. In chapter 5, a new initiative in the field of BSS presented by constructing the TF mask for underdetermined convolutive mixtures using capsule networks. Finally, the results are compared in terms of separation accuracy. Chapter 6 concludes the thesis with future research directions.

Chapter 1: Introduction

Chapter 2: Literature Survey

Chapter 3: Overdetermined convolutive blind source separation using complex FastICA algorithm

Chapter 4: Underdetermined convolutive blind source separation using single source point detection algorithm

Chapter 5: Underdetermined convolutive blind source separation using CapsNet

Chapter 6: Conclusion and Future work.

CHAPTER 2

LITERATURE SURVEY

2.1 HISTORY OF BSS

The pioneering work on blind source separation is an adaptive algorithm proposed by Jutten *et al.* (1991). The adaptive rule tests the independence of the extracted sources using non-linear function is called as contrast function. This approach is further developed by Comon *et al.* (1991); Cichocki *et al.* (1992); Karhunen *et al.* (1993) and others. Comon (1994) introduced the concept of Independent Component Analysis (ICA) and proposed a non-linear function to the minimization of mutual information between the observed mixture signals.

On the other hand, Linsker (1992); Becker *et al.* (1992) and others proposed unsupervised learning rules for blind source separation using information-theory approach. The algorithm maximizes the mutual information between the inputs and outputs of the neural networks. Bell *et al.* (1995) and Roth *et al.* (1996) separately obtained stochastic gradient learning rules for this maximization of mutual information. Further, Bell *et al.* (1995) is the first, using the information-theory approach for the separation of the sources. The adaptive methods are more credible from the neural processing viewpoint than the cumulant based contrast function. A similar algorithm is proposed by Cardoso *et al.* (1996) to the information-theory approach but it is non-neural algorithm.

Lee *et al.* (2000) provided a unifying structure to the source separation problem by giving the relationship of the different algorithms to

each other. Relative gradient (Cardoso *et al.* 1996) or natural gradient (Amari *et al.* 1997) is used to optimize the non-linear function so as to maximize the independence between the sources.

2.2 INDEPENDENT COMPONENT ANALYSIS (ICA)

The objective of ICA is to get independent sources from only sensor observations that are unknown linear mixtures of the independent source signals. Principal Component Analysis (PCA) decorrelates the second-order statistics of the observed signals. ICA decorrelates not only second-order statistics but also higher-order statistics. Due to this the independent sources are extracted. So, independence is a harder condition than uncorrelatedness.

In accordance with the central limit theorem, when independent random variables are added, their sum tends towards a Gaussian distribution even if the original variables themselves are non-Gaussian. Most of the algorithms are in ICA, directly or indirectly minimize the mutual information between the estimated sources. This is same as the maximization of the negentropy. Negentropy is a measure of non-Gaussianity of the estimated sources. Exact calculation of negentropy is computationally demanding and hence most of the ICA algorithm is viewed as an approximation of negentropy such as used in Hyvarinen (1998) and Hyvarinen *et al.* (2001).

Synchronous source separation cannot be solved by using ICA methods since synchronous sources are statistically dependent. Independent phase analysis Almeida *et al.* (2011a) and phase-locked matrix factorization Almeida *et al.* (2011b) is two-step algorithms for synchronous source separation. ICA performs well only in noiseless case and with little added white Gaussian noise. ICA performs well for overdetermined instantaneous

mixtures but not for underdetermined, since the model becomes mathematically ill-posed in the latter case.

2.2.1 ICA model

The ICA model can be mathematically defined as

$$\boldsymbol{x} = \boldsymbol{As} + \boldsymbol{n} \tag{2.1}$$

where A is a full rank M×N matrix whose elements are unknown and n is additive background noise. M is the number of observations and N is the number of sources contributed to mixtures. ICA takes one of the three forms determined (M=N), underdetermined (M<N) and overdetermined (M>N). The aim of ICA is to estimate s by

$$\boldsymbol{y} = \boldsymbol{Wx} \tag{2.2}$$

such that y is the estimate of s. W is N×M unmixing matrix. In the ICA model, there are two uncertainties (i) the energies of the sources cannot be determined and (ii) the order of the sources cannot be determined. A common approach to ICA consists of pre-processing the observed signals, measuring non-Gaussianity of the estimated sources and optimizing the objective function maximizes non-Gaussianity. Most of the methods in the literature either maximizes the negentropy or minimizes the mutual information between the estimated sources. There are three popular methods for ICA namely Joint Approximate Diagonalization of Eigen matrices (JADE) (Cardoso *et al.* 1993), Infomax (Bell *et al.* 1995) and FastICA (Hyvarinen *et al.* 1997 and Hyvarinen 1999).

2.2.2 JADE

JADE is based on the diagonalization of the fourth-order cumulant of the pre-processed input data. The problem with JADE is the determination of mixing matrices when dealing with high-dimensional data. JADE is based on the estimate of kurtosis using cumulants. Ziegaus *et al.* (2004) determined the mixing matrix by the neural implementation of JADE. Natural gradient learning for M=N is the real steepest-descent method in the differential geometry of non-singular matrices. Natural gradient learning is extended to the overdetermined and underdetermined cases in Amari (1999). The contrast function of JADE is given by

$$E = -\ln\left|\det W\right| - \sum_{i=1}^{N} \log p_i\left(y_i(t)\right) \tag{2.3}$$

The natural gradient ICA algorithm iteratively finds the minimum of the above contrast function and it is given by

$$W(t+1) = W(t) + \eta\left[I - g\left(y(t)\right)y^T(t)\right]W(t) \tag{2.4}$$

where η > 0 and $g_i(y_i) = -d\log p(y_i)/dy_i$.

2.2.3 Infomax ICA

The infomax approach (Bell *et al.* 1995) maximizes the mutual information between the sources and the observations. For determined ICA, the infomax approach is equivalent to the maximization of entropies of estimated sources and minimizing the mutual information between the estimated sources.

$$y_i = f_i\left(w_i^T x\right) \tag{2.5}$$

where $f_i(.)$ is a squashing function. The natural gradient solution is obtained as

$$W^T(t+1) = W^T(t) + \eta\left[I - 2g(y)y^T\right]W^T(t) \quad (2.6)$$

where η is the learning rate and g(y) = tanh(y) for source signals. Infomax can be viewed as equivalent to maximum likelihood (ML) (Cardoso 1997) assuming some given a priori marginal distributions y_i.

Infomax is better appropriate for the estimation of super-Gaussian sources. Infomax is called as either gradient infomax or natural gradient infomax (Amari 1998) based on the optimization technique used. The algorithm proposed by Park *et al.* (2006) is a trade-off between natural gradient and gradient Infomax.

2.2.4 FastICA

FastICA is a popular fixed-point ICA algorithm (Hyvarinen *et al.* 1997 and Hyvarinen 1998). It is a measure of non-Gaussianity of the estimated sources using kurtosis or the negentropy and the contrast function is optimized using Newton's method. FastICA accomplishes consistent and at least quadratic convergence. It is considered as a fixed-point algorithm for maximum likelihood estimation of the ICA. FastICA is computationally simple and requires less memory space. FastICA does not depend on any user-defined factors and quickly converge to the most precise solution. The FastICA algorithm finds all non-Gaussian independent sources irrespective of their probability distributions whereas infomax is suitable for super-Gaussian source extraction.

The original FastICA algorithm is based on kurtosis. Since kurtosis is based on fourth-order cumulant of the estimated sources, it is sensitive to

outliers. Therefore, negentropy is used as the contrast function which is a more robust choice than kurtosis. A statistical analysis of the deflation based FastICA algorithm is given in Ollila (2010). FastICA can be implemented either using the sequential mode (deflation approach) or symmetric mode. Symmetric mode FastICA extracts all the independent sources simultaneously. Whereas deflation approaches extract the independent sources one by one. FastICA algorithm extract all the non-Gaussian independent sources for any non-linear function g(.) Hyvarinen (1998) suggested three different non-linear functions

$$g_4(s) = \log\cosh(s) \tag{2.7}$$

$$g_5(s) = -\exp(-s^2/2) \tag{2.8}$$

$$g_6(s) = s^4/4 \tag{2.9}$$

Giannakopoulos *et al.* (1999) conducted a widespread experimental comparison on different types of ICA algorithms including JADE, infomax, FastICA, and natural gradient ICA.

2.2.5 Complex Valued ICA

Complex valued ICA is needed for separating the independent sources from convolutive mixtures in the frequency domain. In the complex ICA model, all the sources are assumed to have zero mean and unit variance with uncorrelated real and imaginary parts. The extension of the JADE to the complex sources is straightforward due to the use of fourth order cumulant in the algorithm. Similarly the extension of negentropy to the complex valued sources based on third and fourth order cumulants is discussed in Comon (1994).

ICA algorithms based on the complex weighted neural network are developed to separate complex valued statistically independent sources (Friori 2000; Friori 2003). The extension of FastICA algorithm to the complex-valued sources is called as c-FastICA (Bingham *et al.* 2000b). c-FastICA is also having the cubic global convergence property like FastICA (Ristaniemi *et al.* 2002). However, c-FastICA is only applicable for second-order circular sources. The following contrast function is used in the c-FastICA for measuring the non-Gaussianity.

$$J(w) = E\left[g\left(\left| w^H z \right|^2 \right) \right] \tag{2.10}$$

where g(.) is any nonquadratic even function like $g(y) = y^2$.

The complex ICA algorithms for the separation of non-circular sources are introduced in Douglas (2007), Li *et al.* (2008), Novey *et al.* (2008a) and Novey *et al.* (2008b). Complex ICA is implemented by maximizing the complex kurtosis contrast function using fixed-point update, gradient update, or Newton update. Similar algorithms are proposed in Novey *et al.* (2008a) using negentropy contrast functions with gradient-descent and a quasi-Newton optimization. It gives superior performance with circular and noncircular sources.

2.2.6 ICA for Convolutive Mixtures

ICA algorithm is directly applied to convolutive mixtures in time domain (Amari *et al.* 1997; Kawamoto *et al.* 1998). This approach gives good separation, if the algorithm converges. But the convergence of ICA is poor when compared with instantaneous ICA and sometimes it may lead to instability problem in the separation of convolutive mixtures. Also ICA is

computationally expensive when applying for convolutive mixtures having long FIR filters.

The separation can be performed in the frequency domain with the benefits of instantaneous BSS. The complex instantaneous BSS is applied to each frequency bin in Smaragdis (1998), Murata *et al.* (2001) and Schobben *et al.* (2002). Any complex valued ICA is used to separate the sources frequency bin wise. But frequency domain BSS contains the permutation problem. Independent vector analysis resolves frequency domain BSS successfully without the permutation problem between the frequencies by using dependencies of frequency bins (Kim 2010).

2.3 PERMUTATION ALIGNMENT TECHNIQUES

Permutation problem would occur in any frequency-domain BSS if the observed signals are processed frequency bin. So there is a necessity of permutation alignment technique to realign the frequency bins in the estimated sources. Generally, there are two kinds of methods in the literature to solve the permutation ambiguity for complex BSS in the frequency-domain. One group of permutation alignment technique is based on the Direction Of Arrival (DOA) estimation techniques; the other group is based on the similarity between the STFT coefficients of adjacent frequency bins. A robust and precise method is proposed by Sawada *et al.* (2004).

2.3.1 DOA Based Permutation Alignment

Ikram *et al.* (2005) proposed a permutation alignment technique based on the directivity pattern of microphone arrays. A simple BSS technique called as Frequency Domain Second-Order Statistics (FDSOS) implemented in Ikram *et al.* (2005). The FDSOS algorithm aligns the frequency bins based on the null location of the two channel directivity

pattern. This algorithm is applicable for the microphone spacing is less than a half wavelength of operating frequency. Toyama *et al.* (2009) modified to the DOA method based on phase linearity of the frequency domain ICA demixing matrix. However, the method has a difficulty around some points where two linear curves of the phase response meet. Permutation ambiguity is solved using two methods in Mallis *et al.* (2017). The first one is using the likelihood Ratio Jump solution and the second one is Multiple Signal Classification (MUSIC) as a pre-processing step before applying the first method. The algorithm converges for more frequency bins compared to the first method for same number of iterations. The algorithm offers robust solution for more than two sensors and sources. However the algorithm is implemented for determined BSS problem.

2.3.2 Correlation Based Permutation Alignment

Sawada *et al.* (2010) proposed a method for underdetermined BSS consists of two stages. In the first step, the frequency domain mixture samples are clustered into each source by an Expectation Maximization (EM) algorithm. The second stage solves the permutation ambiguity occurred in the first step using the likelihood of samples belonging to the clustered source. Reju *et al.* (2010) proposed solution for the permutation problem based on the correlation between envelops of STFT coefficients of the estimated sources. The correlation matrix is calculated as

$$C_{\bar{s}_1\bar{s}_2} = \begin{bmatrix} R_{s_1s_1} & R_{s_1s_2} \\ R_{s_2s_1} & R_{s_2s_2} \end{bmatrix} \tag{2.11}$$

where $R_{s_is_j}$ is the correlation matrix of estimated sources. Sarmiento *et al.* (2015) proposed a permutation alignment technique using a contrast function which measures global similarity of the speech spectrum.

The contrast function depends on two tuning parameters α and β and helps to prevent the propagation errors during the alignment of frequency bins. The maximum of the contrast function is achieved when the estimated sources are properly aligned in all the frequency bins. But the algorithm requires proper selection of tuning parameters α and β empirically. Saito *et al.* (2015) proposed signal separation algorithm in noisy and reverberant environments. To tackle the permutation problem, the correlation between the power ratios of inter frequency bins is used. Lv *et al.* (2017) proposed a permutation algorithm based on Dynamic Time Warping (DTW). The algorithm uses the fact that there is a high similarity between the adjacent frequency bins of independent components. But this algorithm is computationally complex compared to the algorithms that use envelope correlation and higher-order statistics.

2.4 SPARSE COMPONENT ANALYSIS

Sparsity is a natural property of many real signals like speech, activities of the brain and heart beat. The goal of sparse coding is to learn an over complete basis set that characterizes each signal point as a sparse combination of the basis vectors. In a sparse representation, a small number of points contain a large portion of the energy. Sparse recovery aims to reconstruct sparse signals from a set of underdetermined linear mixtures. Many signals can be efficiently characterized by sparse coding like audio, images, and video. Sparse representations of signals are essential in fields such as blind source separation, signal analysis, signal compression, and sampling. Since the ICA technique can be applied only for determined/overdetermined observations, Sparse Component Analysis (SCA) is used for the separation of underdetermined observations of microphone recordings. Also, SCA based methods do not require the statistical independence of the sources. Many algorithms are been proposed to deal with

the Underdetermined Blind Source Separation (UBSS) problem of which SCA is an effective method (Bofill *et al.* 2001; Sadhu *et al.* 2013). Based on SCA, UBSS can obtain encouraging results by utilizing the sparsity of signals (Bofill *et al.* 2001; Georgiev *et al.* 2005). The idea of BSS based on Time-Frequency (TF) points first reported by Belouchrani *et al.* (1998). Short-time Fourier Transform (STFT) and Wavelet Transform (WT) are usually employed to describe the signal in the TF domain (Sadhu *et al.* 2013; Abrard *et al.* 2005; Esmaeilbeig *et al.* 2016; Hu *et al.* 2016). Many Single Source Point (SSP) detection methods are proposed to improve the precision of mixing matrix estimation.

It is assumed that the number of microphones is two and the number of sources are three without affecting the generalization of the underdetermined mixing process where M<N. For convenience, the equation of the underdetermined instantaneous microphone recording is given here.

$$x_1(t) = h_{11}s_1(t) + h_{12}s_2(t) + h_{13}s_3(t) \tag{2.12}$$

$$x_2(t) = h_{21}s_1(t) + h_{22}s_2(t) + h_{23}s_3(t) \tag{2.13}$$

2.4.1 Degenerate Unmixing Estimation Technique (DUET) method

Yilmaz *et al.* (2004) proposed the solution to the underdetermined convolutive BSS problem by having only two mixture signals. There are two assumptions held (i) the sources in the observations are satisfying perfectly W-Disjoint Orthogonal (W-DO) property in the TF domain (ii) the sources in the observations are satisfying approximately W-DO property in TF domain. W-DO is the property of the mixture signal in which at any TF point there exists only one source is nonzero and all the remaining sources are zero.

$$X_1(t,f) = H_{11}(f)S_1(t,f) + H_{12}(f)S_2(t,f) + H_{13}(f)S_3(t,f) \tag{2.14}$$

$$X_2(t,f) = H_{21}(f)S_1(t,f) + H_{22}(f)S_2(t,f) + H_{23}(f)S_3(t,f) \tag{2.15}$$

Assuming the observations satisfies W-DO property, the equation are written at any TF point as,

$$X_1(t,f) = H_{1j}(f)S_j(t,f) \tag{2.16}$$

$$X_2(t,f) = H_{2j}(f)S_j(t,f) \tag{2.17}$$

The ratio between the two observations may be calculated as

$$R_{21}(t,f) = \frac{X_2(t,f)}{X_1(t,f)} = \frac{H_{2j}(f)}{H_{1j}(f)} \tag{2.18}$$

The set of ratios for entire TF points will have three unique numbers as there are three sources in the contribution of the mixture signals. Based on the ratio of mixture, a mask is constructed to extract the source signals.

$$M_j(t,f) = \begin{cases} 1; S_j(t,f) \neq 0 \\ 0; otherwise \end{cases} \tag{2.19}$$

Hence, it is enough that the two mixture signals is sufficient enough to separate the number of sources if it satisfies the W-DO property. But it is a stringent condition to satisfy by the real mixture signals. So the condition is relaxed and it is known as approximate W-DO. For each nonzero TF point, the ratio of mixtures is calculated. A union of pairs, magnitude and phase is constructed for the ratio of mixtures. As the mixture is not strictly satisfying the W-DO property, there will not be three unique numbers for the three sources. Based on the percentage of approximate W-DO, there will be a

large number of pairs of magnitude and phase will be around those unique pairs.

So, a histogram is constructed to locate the peaks for the pairs of magnitude and phase. There should be N peaks for N number of sources with peak location approximately equal to the true mixing parameters. The mask is constructed using the peak location of the histogram and then it is applied to the mixture signals for the reconstruction of the sources.

DUET has the limitations that it is applicable for microphones separated at most $c/2f_m$ where c is the velocity of the signal and f_m is the operating signal frequency. For acoustic signal separation, if c = 340m/s and f_m = 4000Hz, then the maximum distance between the two microphones is 4.25cm. Rickard (2007) gives the extension of DUET for the separation of mixtures regardless of the microphone spacing.

2.4.2 TIme Frequency Ratio of Mixtures (TIFROM) Method

Abrard (2005) used TIme-Frequency Ratio Of Mixtures (TIFROM) for the separation of underdetermined instantaneous mixtures.

$$x_1(t) = a_{11}s_1(t) + a_{12}s_2(t) \tag{2.20}$$

$$x_2(t) = a_{21}s_1(t) + a_{22}s_2(t) \tag{2.21}$$

The source separation is achieved by successive source cancellation method as mentioned in the equation (2.22).

$$cx_2(t) = \beta s_1(t) + a_{12}s_2(t) \tag{2.22}$$

However, this method has the restrictions in the number of sources which cannot be satisfied in reality. The factors like noise, a large number of sources, and a high degree of sources overlapping in the TF domain lead to the violation of sparsity conditions required for DUET and TIFROM methods. Li *et al.* (2006) proposed the TIFROM method where the TF overlapping condition is relaxed. It requires only some adjacent non-overlapping TF regions for each source and TF coefficients are overlapped in other regions. Note that at each of the non-overlapping TF regions, the ratio of the TF coefficients is unique and constant. To convert the observed mixture signals into the TF domain, wavelet packets are applied to each mixture signals. The mixing matrix is estimated by using the ratio of TF coefficients at non-overlapping TF regions. Zhen *et al.* (2017) further relaxed the constraint of the algorithm by assuming that there are some TF points where only one source is dominant than all other sources.

$$X(u,v) = S_i(u,v)a_i \tag{2.23}$$

where a_i is the orientation of the mixing matrix at the particular TF point. If a_i is other than the TF point and if the same source is dominant then X (u, v) is

$$X(u,v) = aX(\psi,\omega) \tag{2.24}$$

Sparse coefficients are calculated using l1 homotopy of TF points. If the TF vector of the mixture contains only one non-zero element, the particular TF point is added to the set Ω. The mixing matrix is identified by applying the clustering technique to the set Ω and it is applied only for underdetermined instantaneous BSS. Zhang *et al.* (2017) proved that if only one source contributes the power to the mixture signals, then

$$\sum_{p=1}^{M-1}\sum_{q=p+1}^{M}\left|\frac{|X_q(t_a,f_a)|}{|X_p(t_a,f_a)|}-1\right|=0 \tag{2.25}$$

But in reality, this condition is seldom valid. Hence the equation is compared with a small threshold value to find the dominant source. The mixing matrix is estimated by applying the clustering procedure to the ratio of mixtures. The technique is also be used to identify the Direction Of Arrival (DOA) estimation. But DOA fails at the separation of lower frequency sources.

2.4.3 Single Source Point (SSP) Detection Methods

The Single Source Points (SSPs) refer to those TF points with only one source contribution for mixture signals. Reju *et al.* (2009) proposed a new method for detecting single source points which used absolute directions of real parts and imaginary parts of TF points. If the absolute directions of TF points are the same, these points are taken as SSPs. If only one source is active at some TF points,

$$R[X(t_2,f_2)]=a_2R[S_2(t_2,f_2)] \tag{2.26}$$

$$I[X(t_2,f_2)]=a_2I[S_2(t_2,f_2)] \tag{2.27}$$

At a multi-source point (MSP),

$$R[X(t_3,f_3)]=a_2R[S_2(t_3,f_3)]+a_3R[S_3(t_3,f_3)] \tag{2.28}$$

$$I[X(t_3,f_3)]=a_2I[S_2(t_3,f_3)]+a_3I[S_3(t_3,f_3)] \tag{2.29}$$

The direction of $R[X(t_3,f_3)]$ will be equal to $I[X(t_3,f_3)]$ if and only if

$$\frac{R[S_1(t_3,f_3)]}{I[S_1(t_3,f_3)]}=\frac{R[S_2(t_3,f_3)]}{I[S_2(t_3,f_3)]} \tag{2.30}$$

At the SSP, the direction of modulus of mixture vectors in the TF domain will be the same as those of column vectors of the mixing matrix (Xiao *et al.* 2005). Using SSPs, the mixing matrix is estimated by the clustering procedure. Similar methods are proposed in Li *et al.* 2011, Thiagarajan *et al.* (2013) and Xu *et al.* (2014). Aissa-el-bey *et al.* (2007) has a single source point selection method in the TF domain using STFT. Further k-means clustering is used to estimate the mixing matrix. Loesch *et al.* (2008) proposed the algorithm for detecting the number of sources using SSP. The algorithm is known as Number of Sources Estimation Technique (NOSET). SSP is detected based on the eigenvalue ratio of the STFT of mixtures. Then the number of sources is detected using histogram of the eigenvalue ratio. Sun *et al.* (2016) improved the absolute direction methods in Esmaeilbeig *et al.* (2016) by eliminating low energy points in the TF domain. Lu *et al.* (2019) takes the points for each source where

$$|S_i(t,f)|>|S_j(t,f)| \ \forall \text{ j}\neq\text{i} \tag{2.31}$$

At any TF point (u,v) if only one source is active, then

$$X(u,v)=a_iS_i(u,v) \tag{2.32}$$

and x(u,v) will be collinear with a_i. At any other TF point (ψ,ω), if the same source is only active, then

$$X(\psi,\omega)=a_iS_i(\psi,\omega) \tag{2.33}$$

Hence,

$$X(u,v) = rX(\psi,\omega) \tag{2.34}$$

where r is real coefficient. By normalizing both sides,

$$\tilde{X}(u,v) = \tilde{X}(\psi,\omega) \tag{2.35}$$

Lu *et al.* (2019) proposed an SSP identification method based on the normalized TF coefficients as given below.

$$1 - \left| \left\langle \tilde{X}(u,v), \tilde{X}(\psi,\omega) \right\rangle \right| < \varepsilon \tag{2.36}$$

However, these SSP detection methods (Kim *et al.* 2009, Reju *et al.* 2009, Thiagarajan *et al.* 2013; Xu *et al.* 2014; Lu *et al.* 2019) and TF point detection methods (Peng *et al.* 2010; Yang *et al.* 2013; Peng *et al.* 2015) are suitable only for underdetermined instantaneous mixtures. Cho *et al.* (2011) proposed SSP detection algorithm based on the ratio of mixtures.

$$S_{s,k} = \left\{ [t,k],[t+1,k] \left\| \frac{X_2(t,k)X_2^*(t+1,k)}{X_1(t,k)X_1^*(t+1,k)} \right\| \prec \varepsilon \right. \tag{2.37}$$

After SSP detection, the mixing matrix is estimated by grouping the values of $S_{s,k}$. Even after the accurate estimation of the mixing matrix, it is not simple to extract the sources from the underdetermined mixture because the matrix is non-square. Cho *et al.* (2011) used the Moore-Penrose method to find the inverse of the non-square mixing matrix. Moore-Penrose pseudo inverse matrix provides the least-squares solution to the matrix inversion that lacks a unique solution. Guo *et al.* (2017) used local directional density

detection and Dynamic Data Field Clustering (DDFC) after finding SSPs. However, the above-mentioned SSPs identification methods are mainly based on the property of single SSP, which will lead to low estimation accuracy in noisy cases.

2.5 MACHINE LEARNING TECHNIQUES

Signal processing and machine learning techniques are two domains that are used to pact with different challenges in BSS. There are two learning strategies in BSS, supervised and unsupervised learning. Supervised learning separates the sources using labelled training data for different sources. So there is a need for training data along with mixed signals to train the unmixing system. Hence it is truly not 'blind' separation. On the other hand, unsupervised source separation does not require the training data. Hence this technique is truly blind source separation since only observed signals are enough for the source separation. Such learning is important for music signal separation where voice is separated from musical instruments (Ozerov *et al.* 2007; Raj *et al.* 2007; Li *et al.* 2007; Durrieu *et al.* 2011).

2.5.1 Supervised Learning

Deep Neural Network (DNN) is introduced as a special technique for blind source separation of non-linear mixed signals by solving a supervised regression problem (Wang *et al.* 2012; Wang *et al.* 2013; Grais *et al.* 2014). During the training stage, the sources are separated so as to minimize regression errors. Huang *et al.* (2014) proposed a deep recurrent neural network to learn temporal structures of the sources. Deep learning based on the artificial neural network consists of multiple hidden layers that learn the high-level abstraction behind the source and characterizes the complex nonlinear relationship between the sources and observations. Such a

Deep Neural Network (DNN) is successfully used for a number of regression and classification systems including speech recognition (Hinton *et al.* 2012; Yu *et al.* 2012) language processing, music information retrieval.

The deep architecture in DNN is effective because the computational units in all the layers follow the same functions, i.e., the affine transformation and nonlinear activation. Basically, the mapping between the spectra of source signals and observations is really complex. The performance of source separation based on linear models like ICA is likely to be bounded. A nonlinear and deep model is overcome this limitation of the linear model. After training a deep model, the lower level of learning is used for different purposes. In addition, DNN is having many fully-connected layers where the parameter space is large. In some cases, a DNN trained with random initialization give poor responses than the linear model. Consequently, a serious issue in DNN training is to find a good initialization and a fast convergence of deep model structure, so it is used for the separation of unknown sources from the observations. A Deep Belief Network (DBN) is a probabilistic model that gives a reliable initialization for DNN. DBN is an unsupervised learning method to reconstruct the inputs. The layers of DBN extract the required features from the observations. This DBN is further united with a supervised regression problem in the application of source separation. Deep clustering is a technique which blindly separates the unknown sources in underdetermined mixtures. The training procedure of these works (Hershey *et al.* 2016; Isik *et al.* 2016) assumes that ideal binary masks for each source are available. These binary masks are used to train a multi-layer Bi-directional Long Short Term Memory network (BLSTM) (Schuster *et al.* 1997).

Hinton *et al.* (2018) introduced a completely new type of neural network called as capsule network. In addition, Sabour *et al.* (2017) proposed an algorithm called “dynamic routing between capsules” to train the capsule networks. Convolutional neural networks are one of the deep learning techniques used for image classification for a long time. However, Hinton *et al.* (2018) identified an important and fundamental drawback of convolutional neural networks. For example, consider a non-technical example of classifying an image as to whether face or not. Convolutional neural networks are trained for identifying an image as a face based on the presence of components like eyes, nose, mouth, etc. But the convolutional networks failed to learn the orientation and relative spatial relationship between the components.

The main component of convolutional neural networks is a convolutional layer. The convolutional layer detects the important features in the inputs. The deep layers learn simple features like edges and gradients whereas higher-level layers are combined simple features into complex features. Finally, dense layers at the final stage combine all high-level features and decide the image classification. Higher-level features are calculated based on the weighted sum of lower-level features. Lower-level features mean here that the output of the layer nearer to the input. The orientation or relative spatial information is missed while calculating the higher-level features are using lower-level features. This is an important limitation in the learning processes of the convolutional neural networks.

Hinton *et al.* (2018) pointed out that in order to correctly classify the inputs, it is important to learn the hierarchical relationship between the features. When these relationships are built into the input of the layer then it becomes easier for the neural networks to learn the features correctly. Capsule

networks learn the features with their hierarchical relations. A capsule is a group of neurons whose activity vector represents the instantiation parameters of a specific object or feature. The length of the activity vector represents the probability of the object present and the direction of the vector represents the orientation of the feature. Conventional convolutional neural networks accept the scalar as the inputs and produce the scalar outputs after multiplying weight vectors and applying non-linear activation function. Whereas capsule networks process the vector inputs and produces the vector outputs after weighting and squashing function.

The algorithm used to learn the instantiation parameters into the capsule networks is “dynamic routing between capsules” proposed by Sabour *et al.* (2017). The lower level capsules send its output to higher level capsules when the feature exist in the input.

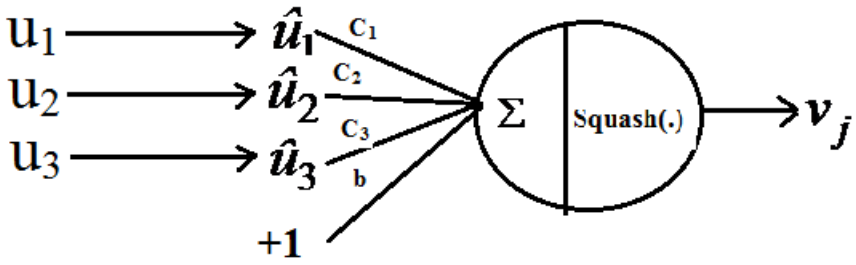


Figure 2.1 Computation of capsule neuron output vector

2.5.2 Unsupervised Learning

In a real environment, sounds and noises are recorded at the microphone in the presence of room reverberation. So, the underdetermined convolutive mixtures more closely replicate the real environment than underdetermined instantaneous mixtures. The task of underdetermined

convolutive BSS is an emerging technique in machine learning for signal processing. In recent years, deep learning is mostly used as they increase the accuracy of the mask estimation (Isik *et al.* 2016; Yu *et al.* 2017; Kolbaek *et al.* 2017). Feature extraction and separation processes are incorporated into the convolutional neural networks (Fu *et al.* 2018).

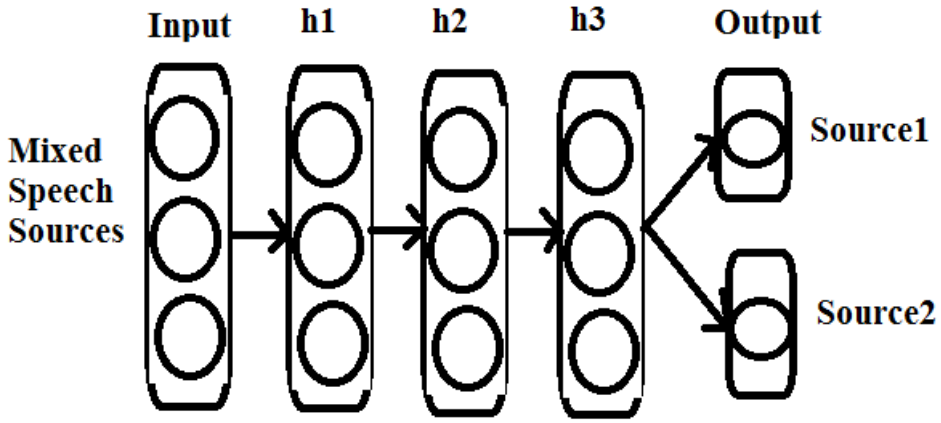


Figure 2.2 DNN Architecture for source separation

2.6 EVALUATION OF BSS ALGORITHMS

BSS Eval is a Matlab toolbox to evaluate the performance of blind source separation algorithms with a framework where the original sources are available. The measures are based on the breakdown of each estimated source signals into a number of components corresponding to original source, interference from unwanted sources, and algorithmic artifacts. The evaluation is valid for any type of sources like audio, biomedical signals, any type of mixed signals like instantaneous, convolutive, and any algorithm like ICA, SCA, time-frequency masking. Vincent *et al.* (2006) proposed a method for the performance measurements of blind source separation algorithms.

In order to evaluate several successful methods such as ICA, SCA, and Computational Auditory Scene Analysis (CASA) and compare the performance measures of several algorithms, there is a necessity of a common framework for evaluation. The BSS eval toolbox assumes that the true source signals and noise signals are known. Performance measures of the algorithm computed for each estimated source $\hat{s}_j$ by comparing it to the original source s_j. The computation of the performance involves two steps. In the first step, each of the estimated sources $\hat{s}_j$ is decomposed into

$$\hat{s}_j = s_{target} + e_{interf} + e_{noise} + e_{artif} \tag{2.38}$$

where $s_{target} = f(s_j)$ is a version of s_j modified by an allowed distortion $f \in F$ and e_{interf}, e_{noise}, and e_{artif} are interferences, noises and artifacts respectively. In the second step, energy ratios are computed to evaluate the relative amount of each of these four components.

The mixing matrix A is time-invariant instantaneous matrix in the instantaneous mixing process and hence the sources are extracted by applying the time-invariant instantaneous unmixing matrix W to the mixture signals. In this process, $\hat{s}_j$ is decomposed as

$$\hat{s}_j = (\boldsymbol{WA})_{jj} s_j + \sum_{j' \neq j} (\boldsymbol{WA})_{jj'} s_{j'} + \sum_{i=1}^{m} W_{ji} n_i \tag{2.39}$$

As $(WA)_{jj}$ is a time-invariant gain, the three terms s_{target}, e_{interf} and e_{noise} are identified respectively from the equation (2.39). However, the mixing matrix A and unmixing matrix W are generally unknown and hence orthogonal projection method is used to define the terms s_{target}, e_{interf}, $e_{noise}$and e_{artif}. Let us denote $\prod(y_1, y_2, ..., y_k)$ the orthogonal projector onto the

subspace spanned by the vectors y_1, y_2, … ,y_k. The projector is T×T matrix, where T is the length of the vectors. The three orthogonal projectors are:

$$P_{s_j} := \prod\{s_j\} \tag{2.40}$$

$$P_s := \prod\left\{\left(s_{j'}\right)_{1\le j'\le n}\right\} \tag{2.41}$$

$$P_{s,n} := \prod\left\{\left(s_{j'}\right)_{1\le j'\le n},(n_i)_{1\le i\le m}\right\} \tag{2.42}$$

Using these projectors, the estimated source is decomposed into four terms.

$$s_{target} := P_{s_j}\hat{s}_j \tag{2.43}$$

$$e_{interf} := P_s\hat{s}_j - P_{s_j}\hat{s}_j \tag{2.44}$$

$$e_{noise} := P_{s,n}\hat{s}_j - P_s\hat{s}_j \tag{2.45}$$

$$e_{artif} := \hat{s}_j - P_{s,n}\hat{s}_j \tag{2.46}$$

By using the equations (2.40) to (2.46), the numerical performance of the blind source separation algorithm is computed by energy ratios expressed in decibels (dB). The following terms are used to compute the performance of the algorithm.

Source to Distortion Ratio (SDR)

$$SDR = 10\log_{10}\frac{\left\|S_{target}\right\|^2}{\left\|e_{interf} + e_{noise} + e_{artif}\right\|^2} \tag{2.47}$$

Source to Interference Ratio (SIR)

$$SIR = 10\log_{10}\frac{\left\|S_{target}\right\|^2}{\left\|e_{interf}\right\|^2} \tag{2.48}$$

Source to Artifact Ratio (SAR)

$$SAR = 10\log_{10}\frac{\left\|S_{target} + e_{interf} + e_{noise}\right\|^2}{\left\|e_{artif}\right\|^2} \tag{2.49}$$

Source to Noise Ratio (SNR)

$$SNR = 10\log_{10}\frac{\left\|S_{target} + e_{interf}\right\|^2}{\left\|e_{noise}\right\|^2} \tag{2.50}$$

For example, consider an instantaneous noisy 2×2 mixture signals. If the estimated source $\hat{s}_1 = \varepsilon s_1 + s_2 + e_{noise}$ with $\|\varepsilon s_1\| << \|s_2\|$ and $\|e_{noise}\| \approx \|\varepsilon s_1\|$. From this, the estimated source is dominated by interfering signals and contains negligible amount of noise as well as target source. The performance measures of this example will be SIR = -∞ and SNR = + ∞ since, SNR has e_{intef} in the numerator. Similarly, SAR is independent of the SIR and SNR. If SDR = + ∞ which implies that the estimated source is completely free from interferences, noises, and artifacts. It is important to note here that the numerical precision is lower for high-performance algorithms than the lower one. When the signals are analog in nature, the precision of the performance depends on the number of bits per sample.

The performance measures are depending a lot on the number of delays and time frames chosen for decomposition. The value of SDR mainly depends on the allowed distortions. For example, F and F' are two families of allowed distortions with F ⊂F', the SDR of the given estimated source will be higher allowing the distortions in F' than in F. To compare the

performance measures with subjective auditory performance measurements, SIR, SNR, and SAR seem to be better related to the auditory notion when the time-varying filter decomposition is used. The decomposition with few delays of time-invariant filters is not able to extract the perceived interferences. But the interferences split into e_{interf} and e_{artif}. On the converse, the decomposition with long delays of time-invariant filters extracts all the interferences and artifacts into e_{interf}. Hence an optimum delayed time-invariant filter is used for decomposition. Vincent *et al.* (2006) suggested the time-invariant filters with delays such as L = 256, 512 seem preferable.

CHAPTER 3

OVERDETERMINED CONVOLUTIVE BLIND SOURCE SEPARATION USING COMPLEX FASTICA ALGORITHM

3.1 BSS OF INSTANTANEOUS SPEECH MIXTURE SIGNALS

The focus of this chapter is on the BSS of overdetermined/determined convolutive mixture signals that supported the complex FastICA algorithm. But the BSS algorithm development during the initial period of BSS research is under the overdetermined/determined instantaneous mixture signals. So it's convenient to know the BSS of instantaneous mixture signals before moving to BSS of convolutive mixture signals. Figure 3.1 shows the BSS of the general instantaneous mixture signals.

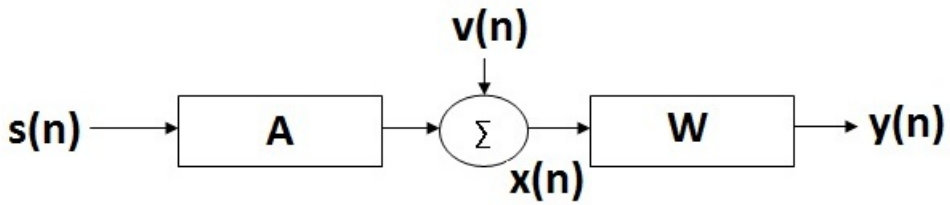


Figure 3.1 BSS of general instantaneous mixture signals

x(n) = $[x_1(n), x_2(n), \ldots, x_M(n)]^T$ denotes the observed speech mixture signals from the microphone recordings. M is the number of microphones used for recording the mixture signals placed at the various places of the room and n is the discrete-time index of the signal. **s(n)** = $[s_1(n), s_2(n), \ldots, s_N(n)]^T$ are the source signals making up of the mixture signals, where N is the number of sources and it is an unknown value

in many practical situations. The instantaneous mixing system is described by a matrix **A** of order M × N and it is called as mixing matrix in the BSS process. **v(n)** denotes the uncorrelated additive noise sources and it's also contributes to the mixture signals **x(n)** apart from **s(n)**.

$$\boldsymbol{x(n) = As(n) + v(n)} \tag{3.1}$$

Noiseless instantaneous mixing model is the simplest BSS method and it is described as

$$\boldsymbol{x(n) = As(n)} \tag{3.2}$$

Inverse mixing matrix **W** of order N × M separates all the sources from the observed mixture signals **x(n)**. As BSS is the process of extracting the sources by only having the observed mixture signals without knowing the mixing matrix **A**, it is necessary to have some prior information about the sources and/or mixing system. The success and efficiency of the BSS algorithm are with the reliability of these prior informations about the sources and mixing process. So selection of the proper model for applications is necessary for efficient source separation. However, energy of the source signals cannot be estimated, instead the scaled version of the sources only can be estimated.

Similarly, the order of the separated sources is random. These two indeterminate states are called as scaling and permutation ambiguity in the BSS. For speech signals, the prior information about the sources are namely, (i) speech sources originating from various sources are statistically independent (ii) speech signals are non-stationary for long duration and pseudo-stationary for a short duration. These characteristics are reviewed shortly in the following section.

The algorithms for the separation of overdetermined mixture signals consist of two-step functions: (i) selection of a suitable contrast function for the measure of statistical independence between the estimated sources (ii) development of an algorithm for minimization and maximization of the contrast function so that it satisfies the statistical independence of the estimated sources.

3.2 STATISTICAL INDEPENDENCE

One of the most commonly used assumptions in the BSS is the original source signals in the mixture are independent. In general, higher-order statistics are used for estimating independent components from the observed mixture signals (Mendel 1991; Cardoso 1999), where the separation is based on the minimization of the fourth-order cumulant. In these types of algorithms, there should not be more than one Gaussian source has zero higher-order moments. The common higher-order statistics are discussed as follows.

3.2.1 Information Theoretic Approach

In the information-theoretic approach, the joint probability density of the independent sources is the same as the product of their individual marginal distributions. In other words, the independent sources do not carry any mutual information between them. This can be achieved by maximizing the entropy of the individual estimated sources. If the sum of the entropy of the individual estimated source is the same as the joint entropy, then the sources are independent. The probability densities of the sources are approximated by some non-linear functions. The maximization of the entropy is the same as maximizing the Kullback-Leibler (KL) divergence between the densities of the estimated sources.

3.2.2 Non-Gaussianity

According to the central limit theorem, when more signals are mixed together, the resultant mixture signal is more Gaussian than the individual Gaussianity of the sources. As the sources of the speech are statistically independent, measuring non-Gaussianity of the estimated sources is the same measuring of the statistical independence. Here, it is important to note that the independent components are not correlated, but not all the uncorrelated components are independent. Principle Component Analysis (PCA) removes the second-order correlation between the estimated sources which gives the uncorrelated components as the output. ICA removes all the higher-order correlations between the estimated sources so that the result will be independent sources. It can be seen that the higher-order statistical methods are connected to one another in some ways.

For example consider **y(n)** = **Wx(n)** is the estimated source of **s(n)** from the observed mixture signals **x(n)**. Then the mutual information between the estimated sources is

$$I(\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N) = \sum_{i=1}^{N} H(\boldsymbol{y}_i) - H(\boldsymbol{x}) - log|det\boldsymbol{W}| \tag{3.3}$$

In equation (3.3), the mutual information is based on the sum of the entropy of the estimated sources and entropy of the observed signals **x(n)**. As the last term is constant, minimizing mutual information is the same as minimizing the entropy of the estimated sources. The estimated source from the observed signals is described as

$$y_i(n) = \sum_{j=1}^{M} w_{ij} x_j(n) \quad i = 1,2,3,...,N \tag{3.4}$$

where w_{ij} are the elements of the unmixing matrix **W**. w_{ij} are optimized such that the linear transformation produces one of the estimated source $y_i(n)$. The correctness of the elements w_{ij} is verified by measuring the non-Gaussianity of the estimated source $y_i(n)$. If the estimated $y_i(n)$ is truly statistically independent, then its non-Gaussianity is maximum. So searching the weights w_{ij} such that the estimated source has maximum non-Gaussianity would separate the independent component.

3.2.3 Whitening

Before applying the mixture signals to the BSS algorithm, it is important to pre-process the signals. Typical pre-processing methods are Principle Component Analysis (PCA), Factor Analysis (FA) and whitening. A zero-mean random vector is called white when its elements are uncorrelated and its variance is unity. In fact, there are many ways to make any random vector as the white vector. For the observed mixture signals **x(n)**, it is necessary to find the linear transformation matrix **V** so that the mixture signals are transformed into white.

$$\boldsymbol{z(n)}_{M\times 1} = \boldsymbol{V}_{M\times M}\,\boldsymbol{x(n)}_{M\times 1} \tag{3.5}$$

The transformation matrix **V** is calculated by Eigen vectors and Eigen values of the covariance matrix of the mixture signals **x(n)**. Eigen Value Decomposition (EVD) is used to obtain the Eigen vectors and Eigen values.

$$\boldsymbol{C}_{M\times M} = cov\left[\boldsymbol{x(n)}\right] \tag{3.6}$$

$$\boldsymbol{C}_{M\times M} = \boldsymbol{E}_{M\times M}\,\boldsymbol{D}_{M\times M}\,\boldsymbol{E}^{T}_{M\times M} \tag{3.7}$$

where **E** is Eigen vectors as its column and **D** is a diagonal matrix having Eigen values as its diagonal elements.

$$\boldsymbol{V}_{M\times M} = \boldsymbol{D}_{M\times M}^{-1/2}\boldsymbol{E}_{M\times M}^{T} \tag{3.8}$$

3.2.3.1Kurtosis

Kurtosis is a classical measure of non-Gaussianity and it is based on the fourth-order cumulant of the random variable. Kurtosis for zero-mean random variable is defined as

$$kurt[\,\boldsymbol{y(n)}\,] = E\left\{\boldsymbol{y^4(n)}\right\} - 3\left[E\left\{\boldsymbol{y^2(n)}\right\}\right]^2 \tag{3.9}$$

where **y(n)** is the estimated source signals. If the variance of the estimated sources is assumed to be unit variance, then the kurtosis of the sources is just a normalized version of the fourth-order moment. Kurtosis is zero for any Gaussian random variable and nonzero for non-Gaussian random variables. It is positive for super-Gaussian random variables having long tails and more spiky peak and negative for sub-Gaussian random variables. So the absolute value of kurtosis is taken as a measure of non-Gaussianity of the estimated sources. The source signals are supposed to be estimated by using the whitened version of the observed mixture signals.

$$\boldsymbol{y(n)}_{N\times 1} = \boldsymbol{W}_{N\times M}\boldsymbol{z(n)}_{M\times 1} \tag{3.10}$$

$$\boldsymbol{y(n)}_{N\times 1} = \boldsymbol{W}_{N\times M}\boldsymbol{V}_{M\times M}\boldsymbol{x(n)}_{M\times 1} \tag{3.11}$$

$$\boldsymbol{B}_{N\times N}\boldsymbol{s(n)}_{N\times 1} = \boldsymbol{W}_{N\times M}\boldsymbol{V}_{M\times M}\boldsymbol{A}_{M\times N}\boldsymbol{s(n)}_{N\times 1} \tag{3.12}$$

$$y_i(n) = \sum_{j=1}^{N} b_{ij}^T s_j(n) = \boldsymbol{b}_i^T \boldsymbol{s(n)} \tag{3.13}$$

The estimated source signal $y_i(n)$ is equal to one of the source signals **s(n)** ,if only one element of the matrix **B** is unity in each row and all others are zero. This shows that the source signals can be estimated up to permutation. The kurtosis of the estimated source signal is given as

$$kurt[\, y_i(n)] = kurt[\, \boldsymbol{w}_i^T \boldsymbol{z(n)}\,] \tag{3.14}$$

The kurtosis value reaches the maximum if only one source contributes to the estimated source. As the variances of the sources are unknown, it is assumed that the estimated sources have unit variance.

$$E\left[y_i^2(n) \right] = \|w_i\|^2 = 1 \tag{3.15}$$

So now the BSS problem is formulated as a constrained numerical optimization problem and it is given as "Maximize kurt[$\mathbf{w}_i^T\mathbf{z(n)}$] subject to the constraint $\|\mathbf{w}_i\|^2 = 1$".

3.2.3.2Negentropy

Negentropy is used as a measure of non-Gaussianity in the information theory approach. Negentropy gives the difference in entropy between the given data distribution and a standard Gaussian distribution having the same mean and variance. Negentropy $J[y_i(n)]$ is defined as

$$J\left[y_i(n) \right] = H\left[y_G(n) \right] - H\left[y_i(n) \right] \tag{3.16}$$

where $H[y_G(n)]$ is the entropy of the standard Gaussian random variable and $H[y_i(n)]$ is the entropy of the estimated random variable. Gaussian random

variables have the maximum entropy compared to any other non-Gaussian distributed random variables. The linear transformation of a random variable changes the kurtosis value whereas negentropy is not affected. But, negentropy is computationally complex, as the integration of probability density function involved in the negentropy function and a kernel estimator is used to estimate the probability density function. The effectiveness of the negentropy approach depends on the correct choice of kernel estimation parameters. As the estimation of the probability density function is difficult and some approximation of negentropy function is used in practice. A reliable and flexible approximation for negentropy is proposed by Pearlmutter *et al.* (1997).

$$J\left[y_i(n)\right] \approx \rho\left[E\left\{f_i\left(y_i\right)\right\} - E\left\{f_i\left(y_G\right)\right\}\right] \tag{3.17}$$

where ρ is constant and $f_i(.)$ is any non-quadratic function. If the non-quadratic function is $y_i^4(n)$, then it leads to kurtosis again. The following non-quadratic function is proved as robust estimators in Hyvarinen *et al.* (2001).

$$f_i\left[y_i(n)\right] = log\, cosh\left[y_i(n)\right] \tag{3.18}$$

$$f_i\left[y_i(n)\right] = -exp\left[-y_i^2(n)/2\right] \tag{3.19}$$

So now the BSS problem is formulated as a constrained numerical optimization problem and it is given as "Maximize $J[\mathbf{w}_i^T\mathbf{z(n)}]$ subject to the constraint $||\mathbf{w}_i||^2 = 1$".

3.2.3.3Maximum likelihood estimation

Maximum likelihood is a popular method in the estimation of the independent sources. Considering the simplest ICA model **x(n)** = **As(n),** the probability density function of the observed mixture signals is described as

$$p_x\left[\boldsymbol{x(n)}\right]=\frac{1}{\left|det\,\boldsymbol{A}\right|}p_s\left[\boldsymbol{s(n)}\right]=\left|det\,\boldsymbol{B}\right|p_s\left[\boldsymbol{Bx(n)}\right] \tag{3.20}$$

where **B** = $\mathbf{A}^{-1}$. As the speech sources are statistically independent, the cumulative density of the mixture signals is the product of the marginal densities of the sources.

$$p_x\left[\boldsymbol{x(n)}\right]=\prod_{i=1}^{N}p_s\left[\boldsymbol{b}_i^T\boldsymbol{x(n)}\right]\left|det\,\boldsymbol{B}\right| \tag{3.21}$$

If there are T samples observed in the microphone recordings, then the likelihood of the matrix **B** is given as

$$L\left[\boldsymbol{B}\right]=p_x\left[\boldsymbol{x(n)}\,|\,\boldsymbol{b}_i\right]=\prod_{n=1}^{T}\prod_{i=1}^{N}p_s\left[\boldsymbol{b}_i^T\boldsymbol{x(n)}\right]\left|det\,\boldsymbol{B}\right| \tag{3.22}$$

Log-likelihood function is often considered rather than likelihood function for computational convenience. So the log-likelihood function of the matrix **B** is written as

$$log\,L[\,\boldsymbol{B}\,]=T\,log\left|det\,\boldsymbol{B}\right|+\sum_{n=1}^{T}\sum_{i=1}^{N}log\left[\,p_i\left\{\boldsymbol{b}_i^T\boldsymbol{x(n)}\right\}\right] \tag{3.23}$$

$$\frac{1}{T}log\,L\left[\boldsymbol{B}\right]=log\left|det\,\boldsymbol{B}\right|+E\left\{\sum_{i=1}^{N}log\left[\,p_i\left(\boldsymbol{b}_i^T\boldsymbol{x(n)}\right)\right]\right\} \tag{3.24}$$

Here log-likelihood function depends on the separation matrix **B** and marginal densities of the estimated sources. The estimation of the densities of the sources is a non-parametric problem. The non-Gaussianity is used to solve this non-parametric problem. Approximate densities of either super-Gaussian or sub-Gaussian sources are used for all unknown sources. Hyvarinen *et al.* (2001) states that if $p_i[s_i(n)]$ is assumed density of the estimated source, then the ML estimator is locally consistent provided,

$$E\left\{s_i(n)g_i\left[s_i(n)\right]-g_i'[s_i(n)]\right\}>0 \tag{3.25}$$

where $g_i\left[s_i(n)\right]=\frac{\partial}{\partial s_i(n)}p_i\left[s_i(n)\right]$. Consider the following log densities for super-Gaussian and sub-Gaussian probability density functions respectively.

$$log\ p_i\left[s_i(n)\right]=a-log\,cosh\left[p_i(s_i)\right] \tag{3.26}$$

$$log\ p_i\left[s_i(n)\right]=b-\left\{s_i^2(n)/2-log\,cosh\left[p_i(s_i)\right]\right\} \tag{3.27}$$

The choice between the super-Gaussian and sub-Gaussian function is based on the computation of the equation (3.25). If the value of the equation (3.25) is greater than zero, then the super-Gaussian probability density function is used for likelihood calculation. Otherwise, the sub-Gaussian probability density function is used for likelihood computation. So now the BSS problem is formulated as a constrained numerical optimization problem and it is given as "Maximize log L[**B**] subject to the constraint $\|\mathbf{w}_i\|^2 = 1$".

3.3 OPTIMIZATION OF THE CONTRAST FUNCTIONS

The BSS problem is formulated as a constrained numerical optimization problem regardless of the contrast function. The contrast functions are kurtosis, negentropy and maximum likelihood estimation to measure the statistical independence between the sources. The algorithm used for the minimization or maximization of the contrast function can be the simple gradient method, natural gradient method or Newton's method.

3.3.1 Simple Gradient Method

The simple gradient method is an optimization algorithm used to minimize the contrast function by iteratively moving in the direction of steepest descent. The parameters to be updated in all our contrast functions are the elements of unmixing matrix **W**. The size of the steps used in each iteration for updating these parameters is called as learning rate of the algorithm. With a high learning rate, the algorithm reaches the convergence faster at the risk of overshooting the lowest point. This may lead to the instability of the algorithm. With a very low learning rate, the problem of overshooting is avoided at the cost of a slow convergence rate. So it is difficult to choose the proper learning rate for the algorithm. Also, a common learning rate will not be suitable for the separation of different mixture signals.

3.3.2 Natural Gradient Method

A simple gradient algorithm has the possibility of convergence around the local estimates. But in reality, there can be an optimized result around some other regions. So in the sense, fixed learning rate cannot achieve the optimum result when the solution has many local minimal and one global minimum. The natural gradient defines the learning rate of the parameters

based on the distance in the distribution space but not based on the distance between the parameter spaces. The simple gradient algorithm assumes that the parameter space as a flat surface and computes the distance between the two points as Euclidean distance. The algorithm assumes that the straight line between the two points is the shortest distance. But in practice, the parameter surface is not flat, more often it is a curved surface. In the natural gradient algorithm, the mathematics of the curved space is used and it's known as Riemannian geometry. In general, natural gradient ICA algorithms are used for the online processing of the mixture signals. The basic equation for the natural gradient method is

$$\boldsymbol{W} \leftarrow \boldsymbol{W} + \eta\left[\boldsymbol{I} - g(\boldsymbol{y})\boldsymbol{y}^{T}\right]\boldsymbol{W} \tag{3.28}$$

where η is the learning rate of the algorithm and W is the unmixing matrix. $g(\mathbf{y}) = [g(y_1), g(y_2), \ldots, g(y_N)]^T$ is a nonlinear contrast function. However, it is difficult to obtain the bound on the natural gradient step for the stability and convergence of the algorithm.

3.3.3 Newton's Method

In optimization, Newton's method is an iterative method that uses twice differentiable contrast function to find the optimized points of the contrast function. The solution can be a minimum or maximum value of the contrast function.

The iteration method in Newton's is given as

$$\boldsymbol{w}_i \leftarrow \boldsymbol{w}_i + \left[\frac{\partial^2 J(\boldsymbol{w}_i)}{\partial \boldsymbol{w}_i^{\,2}}\right]\frac{\partial J(\boldsymbol{w}_i)}{\partial \boldsymbol{w}_i} \tag{3.29}$$

where J($\mathbf{w}_i$) is the contrast function to be optimized. Let consider the above-said contrast functions to measure the statistical independence and their optimization using Newton's method one by one. First consider the constrained optimization problem using kurtosis function i.e., "Maximize kurt[$\mathbf{w}_i^T\mathbf{z(n)}$] subject to the constraint $\|\mathbf{w}_i\|^2 = 1$".

$$\frac{\partial kurt\left[\boldsymbol{w}_i^T\boldsymbol{z(n)}\right]}{\partial \boldsymbol{w}_i} = E\left\{\boldsymbol{z(n)}\left[\boldsymbol{w}_i^T\boldsymbol{z(n)}\right]^3\right\} \tag{3.30}$$

The constraint of the optimization problem is satisfied by bounding $\mathbf{w}_i$ to unit norm after every iteration.

$$\boldsymbol{w}_i \leftarrow \boldsymbol{w}_i / \|\boldsymbol{w}_i\| \tag{3.31}$$

Next consider the constrained optimization problem using negentropy function. Deriving Newton's method of optimization,

$$\frac{\partial J\left[\boldsymbol{w}_i^T\boldsymbol{z(n)}\right]}{\partial \boldsymbol{w}_i} \alpha\left[E\left\{\boldsymbol{z(n)}\right\} f_i'\left[\boldsymbol{w}_i^T\boldsymbol{z(n)}\right]\right] + \beta\boldsymbol{w}_i \tag{3.32}$$

$$\frac{\partial^2 J\left[\boldsymbol{w}_i^T\boldsymbol{z(n)}\right]}{\partial \boldsymbol{w}_i^2} \alpha\left[E\left\{\boldsymbol{z(n)}\right\} f_i''\left[\boldsymbol{w}_i^T\boldsymbol{z(n)}\right]\right] + \beta \tag{3.33}$$

where $f_i(.)$ is the non-quadratic function as given in equation (3.18) and (3.19).

Next, consider the constrained optimization problem using maximum likelihood function. The equation (3.33) is utilized for maximum likelihood function optimization also by replacing the non-quadratic function

$f_i(.)$ by either equation (3.26) or (3.27) based on the results of the equation (3.25).

3.4 FASTICA ALGORITHM

One of the most popular solutions for linear instantaneous BSS is FastICA. The algorithm is simple and converges faster than other ICA algorithms. FastICA is an iterative fixed-point algorithm derived from one of the above-said contrast functions. The estimation of the separation matrix is done iteratively.

3.4.1 Iteration Using Deflation Approach

The FastICA algorithm using the deflation approach estimates one row of the separation matrix **W** as a vector $\mathbf{w}^T$ by optimizing one of the contrast functions. The pre-processed mixture signals are used as the input to the FastICA algorithm. Assume that **z(n)** is a whitened version of the observed mixture signals and $\mathbf{w}^T$ is one of the rows of the separation matrix **W**. Estimation of '**w**' continues iteratively with the following steps.

1. Center the observed microphone signals to make its mean zero.

$$x_i(n) = x_i(n) - E\left[x_i(n)\right] \text{ for } i = 1,2,3,...,M$$

2. Linearly transform of the observed microphone signals and it is whitened.

$$\boldsymbol{z(n) = Vx(n)}$$

3. Choose an unmixing matrix randomly.

4. Compute the source signals by using unmixing matrix and observed microphone recordings.

$$y(n) = Wz(n)$$

5. Update the unmixing matrix based on the contrast function calculation.

6. Normalize the row of the unmixing matrix to avoid repeated extraction of the independent sources.

$$w_i \leftarrow w_i + \left[\frac{\partial^2 J(w_i)}{\partial w_i^2}\right]\frac{\partial J(w_i)}{\partial w_i}$$

7. If the unmixing matrix is not converged, then go to step 4.

3.4.2 Iteration Using Symmetric Approach

The independent components can be extracted one by one as discussed in the deflationary approach or can extract all the sources simultaneously. In the deflationary approach, it is important to assure that each vector '**w**' of the separation matrix is orthogonal to each other. This is done after each iteration before normalizing the vector. In the symmetric approach, the algorithm computes all the vectors in one iteration and the matrix is orthogonalized before proceeding to the next iteration. Step 6 of the algorithm is replaced by the following equation for symmetric extraction of the sources.

$$W \leftarrow \left(WCW^T\right)^{-1/2} W \qquad (3.34)$$

The asymptotic convergence of the FastICA algorithm is at least quadratic and generally cubic if the noiseless ICA model given in equation (3.2) holds.

3.5 DRAWBACKS OF EXISTING METHODS

3.5.1 Noisy Sources

The estimation of the noiseless instantaneous overdetermined/determined ICA model converges faster and gives the solution to the separation problem. When the noiseless model is extended to the noisy ICA model, the overdetermined/determined mixing process does not hold anymore as noises act as additional sources. If noises are additional sources contributed to mixing signals, then the number of sources is more than the number of microphones used for recording the mixture signals. Hence the mixing process becomes underdetermined instantaneous mixtures inherently. The underdetermined instantaneous BSS cannot be handled by simple ICA techniques as the number of knowns (mixture signals) is less than the number of unknowns (sources). Practically it is not possible to increase the number of microphones more and more apart from a certain limit and furthermore it is not possible to predict the number of sources in advance. Moreover, it is not possible in practice to divide the observed data into signals and noise sources.

3.5.2 Complex Sources

Separation of complex-valued signals is a frequently occurring problem in many signal processing applications. For example, the separation of speech sources in the real environment is not instantaneous but convolutive in nature. Separation of convolutive mixed source signals requires the

computation of complex values in the frequency domain. The simple FastICA algorithm cannot separate the complex sources.

3.5.3 Nonlinear Mixtures

The mixture signals are passed through certain non-linearity before it is being used for applications. For example, speech sources are recorded using microphones. The microphones provide not only the linear mixing of the speech signals but also introduce certain non-linearity to the mixture signals. Therefore, it is necessary to separate the sources from the observed non-linear mixtures. A simple ICA algorithm cannot separate the non-linear mixture signals.

3.5.4 Permutation Ambiguity

The order of the extracted sources from the ICA algorithms cannot be determined. Consider the simple noiseless ICA model in equation (3.2).

$$s_i(n) = \sum_{j=1}^{M} w_{ij} x_j(n) \tag{3.35}$$

Both **A** and **S** are unknown in the equation, the order of the terms in the equation (3.35) is changed freely and the sources are numbered as the first source and so on. This indeterminacy is called as the permutation problem and it is severed when the signals are processed in the frequency domain. Because of this indeterminacy, blind identification of the signal is not possible. Mathematically, the permutation problem is described by using the mixing matrix **A** and the separation matrix **W** as

$$\boldsymbol{AW} = \boldsymbol{P} \tag{3.36}$$

where **P** is a permutation matrix.

3.5.5 Scaling Ambiguity

The energy of the extracted sources cannot be determined. Since both **A** and **S** are unknown, the effect of multiplying a constant with any source estimate is cancelled by dividing the corresponding column of the mixing matrix by the same constant. This indeterminacy is called scaling ambiguity in ICA. So it is assumed that each extracted sources have unit variance.

$$E\left[s_i^2\right]=1 \tag{3.37}$$

3.6 PROPOSED METHOD

The speech mixture signals observed from the microphone recordings are convolutive mixture signals in nature. Hence simple FastICA algorithm cannot be used for convolutive BSS as it is suitable for the separation of instantaneous mixture signals only. Nevertheless, convolutive mixture signals in the time domain is converted into the frequency domain by using Fourier transform and hence the mixture signals become instantaneous. But the FastICA algorithm is not directly apply to the frequency domain separation as it involves the complex values in the frequency domain. We have extended the FastICA algorithm for the separation of sources from convolutive mixture signals. The problem is complicated since the frequency domain separation involves complex numbers and the frequency domain separation problem creates the permutation problem whereas the permutation problem is absent in the source separation of instantaneous mixture signals.

The proposed FastICA algorithm is different from the conventional FastICA algorithm in the sense of extending it to the complex numbers and hence named as complex FastICA algorithm. The algorithm also provides the

permutation alignment by the proposed two-pass method. Also, the speech signals are not completely stationary, but they are locally stationary. Hence Short Time Fourier Transform (STFT) is used to convert the time domain mixture signals into the frequency domain. Bingham *et al.* (2000) proposed the method for extending the independent component analysis technique to the complex-valued signals.

3.6.1 Complex FastICA Algorithm

The noiseless complex ICA model equivalent to the equation (3.2) is given as

$$\boldsymbol{X(t, f) = H(f)S(t, f)} \tag{3.38}$$

where **X(t,f)** is the observed mixture signals transformed into Time-Frequency (TF) vectors using STFT. **H(f)** is the frequency response of the path between the source and microphone. It is assumed that the impulse responses of the paths between the source and microphone are time-invariant. **S(t,f)** is TF vectors of the source signals to be separated from the observed mixture signals. The observed complex TF points are represented as X = U+iV where U and V are real-valued random variables. The mean of the observed signals is E{X} = E{U} + i E{V}. The observed mixture signals are uncorrelated if $E\{X_1X_2^*\} = E\{X_1\}E\{X_2^*\}$ where X_2^* is the conjugate of the observed signal. The covariance of the zero-mean complex signals **X(t,f)** = $[X_1(t,f), X_2(t,f),\ldots,X_M(t,f)]$ is

$$E\left\{XX^{H}\right\}=\begin{bmatrix} C_{11} & C_{12} & . & . & . & C_{1M} \\ C_{21} & C_{22} & . & . & . & C_{2M} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ C_{M1} & C_{M2} & . & . & . & C_{MM} \end{bmatrix} \tag{3.39}$$

Where $C_{ij} = E\{X_i X_j^{*}\}$ and X^H denotes the Hermitian of X. The contrast function is generally described as follows.

$$J(\boldsymbol{w}) = E\left\{G\left(\left|\boldsymbol{w}^{H}\boldsymbol{X(t,\,f)}\right|^{2}\right)\right\} \tag{3.40}$$

where G is the non-linear function as discussed in the kurtosis, negentropy and maximum likelihood functions and '**w**' is the complex weight vector and $E\{|\mathbf{w}^H\mathbf{X(t,f)}|^2\} = 1$. For example, if the non-linear function $G(x) = x^2$ is chosen, then J(**w**) measures the kurtosis of the extracted sources. So the problem is now described as a constrained optimization problem.

To Maximize $\sum_{j=1}^{N} J(w_j)$ with respect to w_j under the constraint $E\left\{\left(\boldsymbol{w_k}^{H}\boldsymbol{X(t,\,f)}\right)\left(\boldsymbol{w_j}^{H}\boldsymbol{X(t,\,f)}\right)^{*}\right\} = \delta_{jk}$. The complex FastICA algorithm is summarized as follows:

1. Convert the observed microphone signals into frequency domain using STFT.

2. Center the observed signals to make its mean as zero.

$$X_i(t,f) = X_i(t,f) - \left\{E\left[Re\left(X_i(t,f)\right) + E\left[Im\left(X_i(t,f)\right)\right]\right]\right\}$$

for $i = 1,2,3,...,M$

3. Linearly transform the observed microphone signals so that it is whitened.

$$Z(t,f) = VX(t,f)$$

4. Choose an unmixing matrix randomly.

5. Compute the source signals by using unmixing the matrix and observed microphone recordings.

$$Y(t,f) = W^H Z(t,f)$$

6. Update the unmixing matrix based on the contrast function calculated.

7. Estimate all the independent components simultaneously using symmetric decorrelation approach.

$$W \leftarrow W\left(W^H W\right)^{-1/2}$$

8. If the unmixing matrix is not converging, then go to step 4.

3.6.2 Permutation Alignment Algorithm

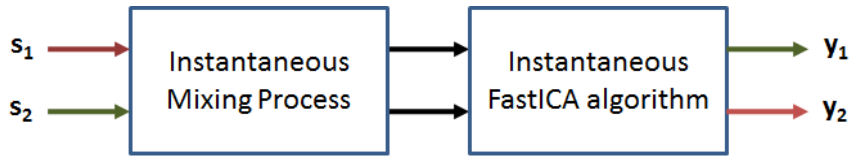


Figure 3.2 Permutation problem in the BSS

The order of the extracted sources from the FastICA algorithms cannot be determined. Figure 3.2 shows the change in the order of the output from the FastICA algorithm. This indeterminacy is called the permutation problem and it is severed when the signals are processed in the frequency

domain. Before converting the separated signals back into the time domain, the correct order of the frequency components is required. Otherwise, the separated signals are not contain the frequency components from the same source and hence the objective of BSS vanishes. Hence, a good permutation alignment algorithm is required for effective BSS.

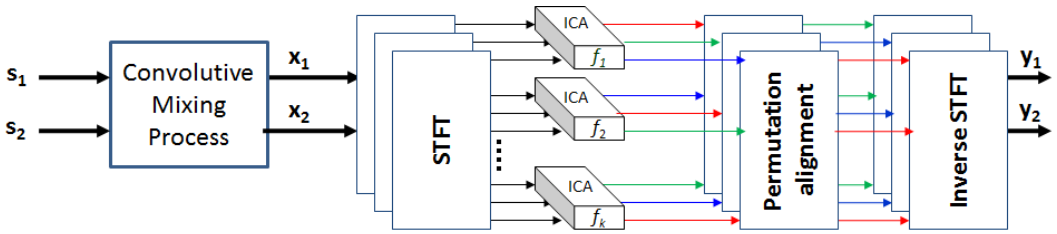


Figure 3.3 Proposed methods for BSS of overdetermined/determined convolutive mixing signals

Figure 3.3 shows the proposed method for the BSS of convolutive mixture signals. Instead of taking the Fourier transform, STFT is computed for the observed mixture signals using the windowing technique. The length of the window must satisfy the stationary property of the source signals. A long window size gives better frequency resolution, but it affects the basic requirements of the stationary property of the source signals. A short window size results in poor frequency resolution. Optimum window size should be chosen to satisfy the stationary property of the source signals and frequency resolution. It is proved that 16ms to 32ms long window is optimum for analysis of speech sources (Yilmaz *et al.* 2004). For each frequency bins, the complex FastICA algorithm is applied to find the independent components from the observed signals. Due to the permutation problem in the FastICA algorithm, the order of components will be random for each frequency bins. The figure shows that there are three independent components existing in each frequency bin. The order of the independent components are random at $f_1, f_2,\ldots f_k$.

In order to align the order of independent components at each frequency bins, the amplitude correlation between the adjacent frequency bins are used in Reju *et al.* (2010), Sawada *et al.* (2010), Sarmiento *et al.* (2015), Lv *et al.* (2017). Figure 3.4 illustrates the amplitude correlation between the adjacent frequency bins.

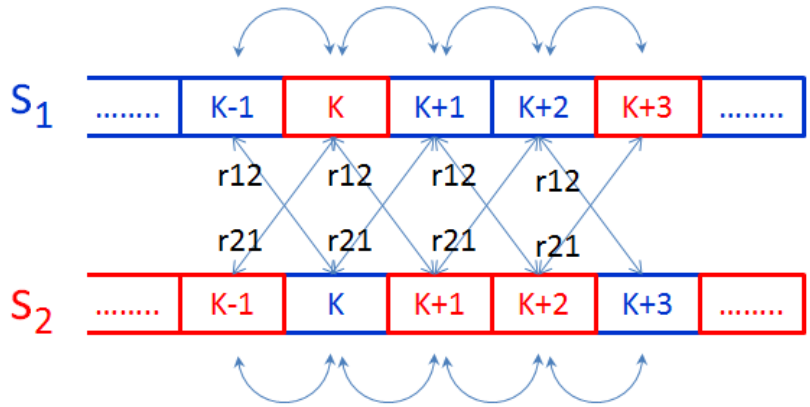


Figure 3.4 The computation of amplitude correlation between the adjacent frequency bins

The adjacent frequency bins, for example $(K-1)^{th}$ and K^{th} frequency bins are considered for calculating the amplitude correlation between them. If both the frequency bins are from the same source, then its correlation is high. Otherwise, the correlation is lower. If the correlation between the adjacent frequency bins is lower than some fixed threshold value then the frequency bins are interchanged.

$$\rho_{12} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \tag{3.41}$$

If r_{11} and $r_{22} < r_{12}$ and r_{21}, then the frequency bins are interchanged. But the problem with this method is that if the correlation values of r_{11}, r_{22} are moderate value, then it is difficult to make the decision whether to swap the

frequency bins or not. Also if one frequency bin is misaligned due to this problem, then subsequent frequency bins would also result in the wrong alignment of the frequency bins. Hence to avoid such subsequent misalignment of the frequency bins, a two-pass method is proposed in this work.

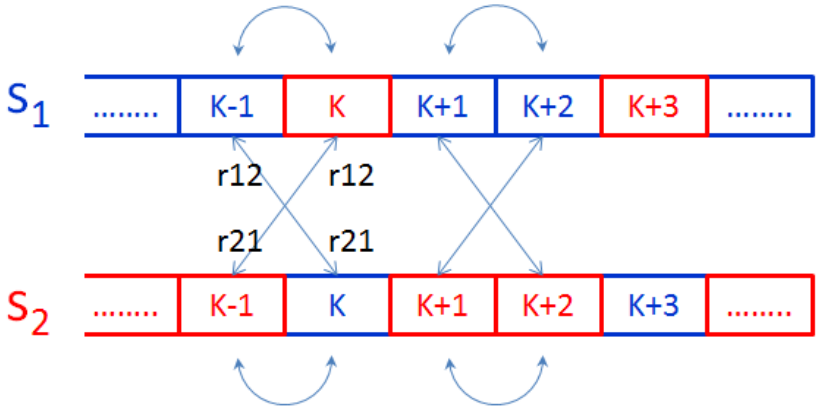


Figure 3.5 Proposed two-pass method for computation of adjacent frequency band correlation

In the proposed method, the correlation between the power ratios is used with two-pass calculations to reduce the number of misaligned frequency bins.

$$PR_i(t,f) = \frac{\|Y_i(t,f)\|^2}{\sum_{k=1}^{N}\|Y_k(t,f)\|^2} \tag{3.42}$$

In the first pass, if the correlation values of r_{12} and r_{21} are high (typically >0.9), then such frequency bins are interchanged with high confidence. If the correlation values of r_{12} and r_{21} are moderate, then such frequency bins are left unchanged and skipped for subsequent correlation calculation. By doing so, subsequent misalignment is reduced. In the second

pass, the less confident frequency bins are correlated with the centroid of the confident frequency bands. Based on these correlation values, the frequency bins are interchanged or left as it is.

3.7 RESULTS AND DISCUSSION

3.7.1 Experimental Setup

Noisy speech data sets from Vincent *et al.* (2009a) is used to evaluate the efficiency of various BSS algorithms. Data sets containing synthetic mixture signals as well as live recorded mixture signals are used for evaluation. Two channel mixtures of two speech sources and real background noises sampled at 16 KHz are considered for testing our algorithm. The mixture as well as source signals are of 10 seconds duration. The direction of arrival of each speech sources is different in each mixture and the Signal to Noise Ratio (SNR) is drawn randomly between -17dB to +12dB. The signals are recorded in the reverberant environment. The recording room dimensions of the room1, room2, room3 and room4 are 1.5×2×2.5 m (chamber), 10×8×3 m (conference room), 3×3×2.5 m (medium size conference room) and 3.55×4.45×2.5 m (chamber) respectively. The microphones are placed at 50cm and 1m from the walls randomly. The sources are placed randomly in the room with an average distance of 2m between the sources.

3.7.2 Results of Instantaneous BSS Using FastICA

Initially two sources without noise, one male speech source and another female speech source are taken for testing the FastICA algorithm. The two sources are mixed instantaneously by using a random mixing matrix,

$$A = \begin{bmatrix} 5 & 2 \\ 3 & 4 \end{bmatrix} \tag{3.43}$$

The real-time mixture signals are recorded in the typical office environment using the same two sources used in the synthetic mixture signals with background noises. The estimated sources using FastICA algorithm and original sources are used to evaluate the performance of the algorithm. Source to Distortion Ratio (SDR), Source to Interference Ratio (SIR) and Source to Artifact Ratio (SAR) values in decibels (dB) are measured for the evaluation of the algorithm. Table 3.1 gives the performance of the algorithm on both synthetic mixture signals and real-time mixture signals.

The FastICA algorithm performs well for the separation of the synthetically mixed signals. The performances of any BSS algorithm are measured by comparing the values of SDR, SIR and SAR and their equations are presented here for convenience.

Source to Distortion Ratio (SDR)

$$SDR = 10\log_{10}\frac{\left\|S_{target}\right\|^2}{\left\|e_{interf} + e_{noise} + e_{artif}\right\|^2} \tag{3.44}$$

Source to Interference Ratio (SIR)

$$SIR = 10\log_{10}\frac{\left\|S_{target}\right\|^2}{\left\|e_{interf}\right\|^2} \tag{3.45}$$

Source to Artifact Ratio (SAR)

$$SAR = 10\log_{10}\frac{\left\|S_{target} + e_{interf} + e_{noise}\right\|^2}{\left\|e_{artif}\right\|^2} \tag{3.46}$$

If SIR is equal with SDR, then the equations (3.44) and (3.45) implies that the noises and artifacts are absent in the estimated sources. The absence of the artifact is confirmed with the values of SAR. Hence the results

of the FastICA algorithm using kurtosis, negentropy and maximum likelihood functions succeed in the separation of the synthetic instantaneous mixture signals. Figure 3.6 illustrates the results of the FastICA algorithm using the three contrast functions. The time-domain results of the FastICA algorithm are given in Figure 3.8.

However, when the real-time mixture signals with background noises are applied as input to the FastICA algorithm, the performance of the algorithm degrades considerably. The results in the Table 3.1 shows SIR>>SDR, which imply the presence of the noises and artifacts in the estimated sources. The negative values of SDR and SAR also infer the degradation in the performance of the algorithm.

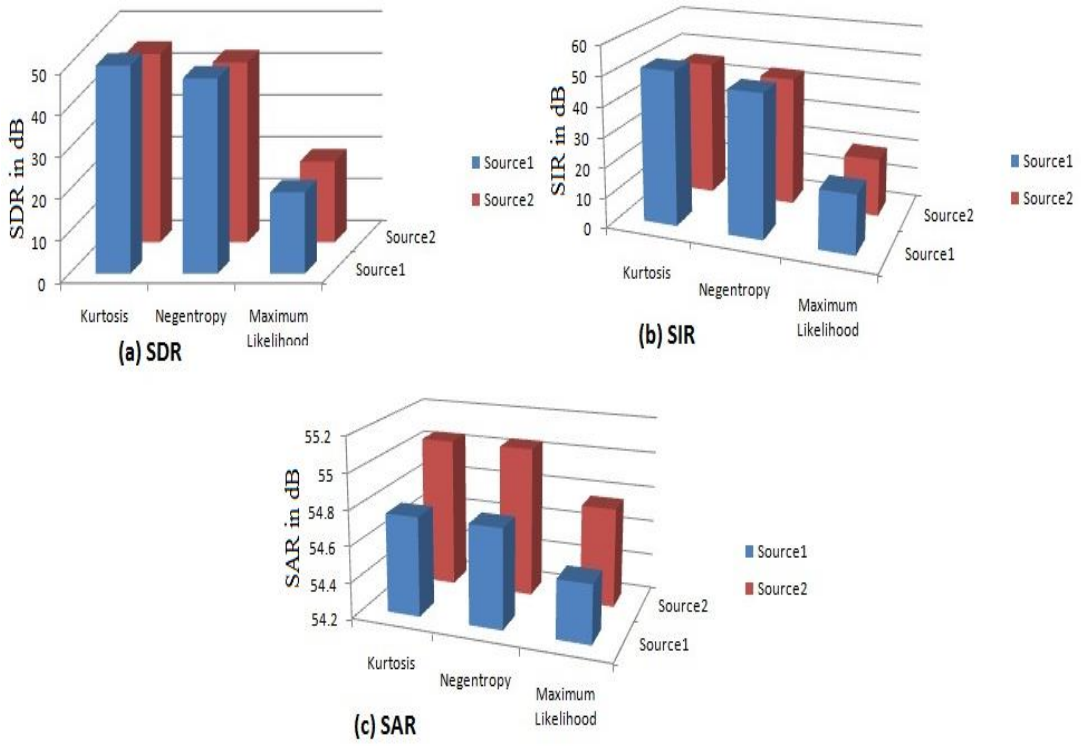


Figure 3.6 Performance of the FastICA algorithm

Table 3.1 Performance of the FastICA algorithm

Parameter	Types of Mixture signals	Source	Kurtosis	Negentropy	Maximum likelihood function
SDR (dB)	Synthetic mixture signals	Source 1	49.7015	45.5588	19.4378
		Source 2	45.0607	43.0009	19.2976
SIR (dB)		Source 1	51.3274	47.4021	19.4392
		Source 2	45.5089	43.2810	19.2988
SAR (dB)		Source 1	54.7559	54.7559	54.5249
		Source 2	55.0454	55.0454	54.7559
SDR (dB)	Real time mixture signals	Source 1	-13.9451	13.9581	-18.3843
		Source 2	-18.6029	-18.6478	-13.9400
SIR (dB)		Source 1	10.8432	10.6818	4.0665
		Source 2	3.3453	3.2067	10.8848
SAR (dB)		Source 1	-13.5870	-13.5870	-16.9230
		Source 2	-16.9230	-16.9230	-13.5870

Figure 3.7 shows the limitations of the FastICA algorithm in the separation of real-time mixture signals.

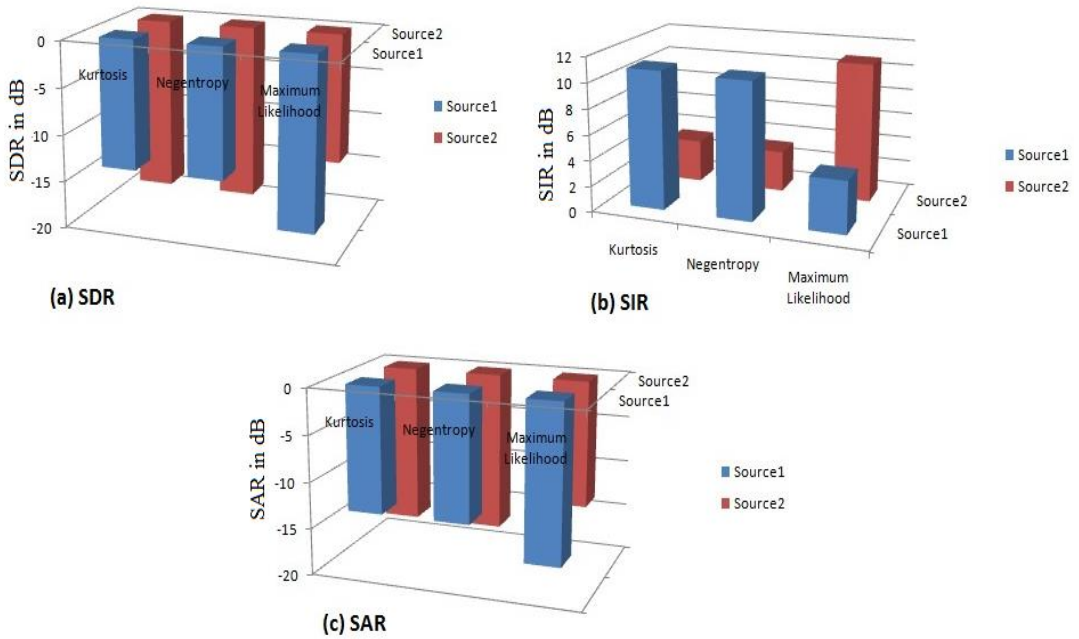


Figure 3.7 Limitations of the FastICA algorithm

Table 3.2 Computational Complexity of the FastICA algorithm

Computational Complexity	Types of mixture signals	Kurtosis	Negentropy	Maximum likelihood function
No. of iterations to converge	Synthetic instantaneous mixture signals	3	3	2
Computation time in sec per GHz CPU		0.6	0.106	0.143
No. of iterations to converge	Real-time mixture signals	4	5	2
Computation time in sec per GHz CPU		0.61331	0.122	0.1003

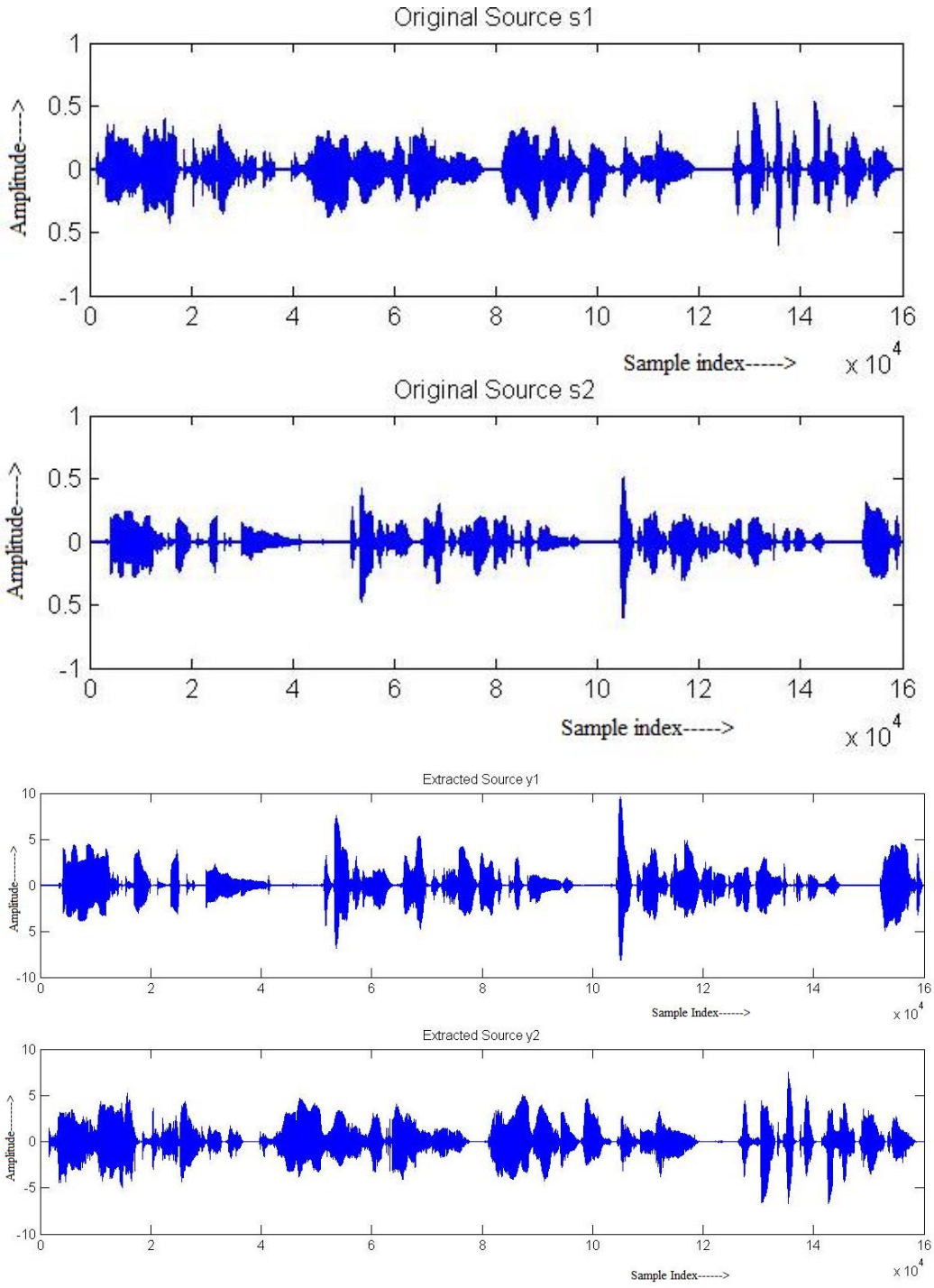
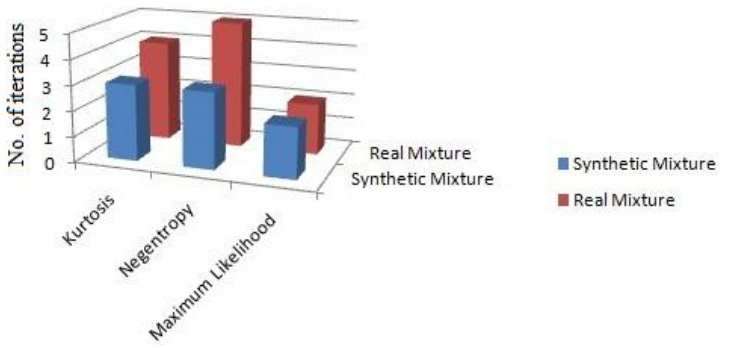
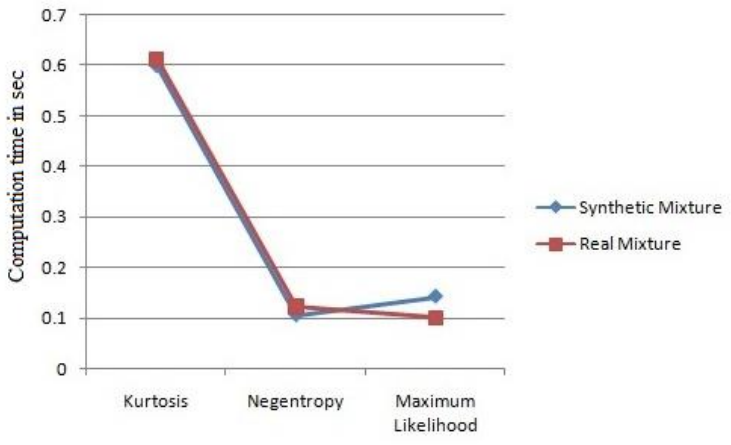


Figure 3.8 Time domain results of FastICA algorithm

In general, the FastICA algorithm converges at least quadratic. The computational requirement of the FastICA algorithm is tested using number of iterations and computation time in second per GHz CPU. Table 3.2 and Figure 3.9 illustrates the computational complexity of the FastICA algorithm. The computation time is calculated in seconds per 10-second long audio excerpt per GHz CPU. The computation time of the FastICA algorithm is as low as 0.1 sec for processing 10-sec long speech mixtures. For 32 second long mixture signals FastICA takes 11.82 sec computation time. The infomax and JADE algorithm requires 25.79 sec and 48.26 sec respectively for the processing of the same mixture signals (Sahonero-Alvarez et.al. 2017).



(a) Number of iterations

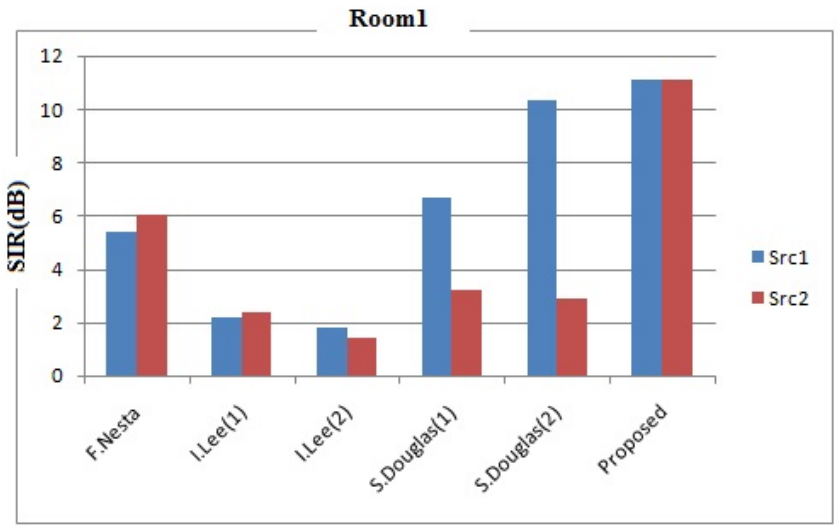


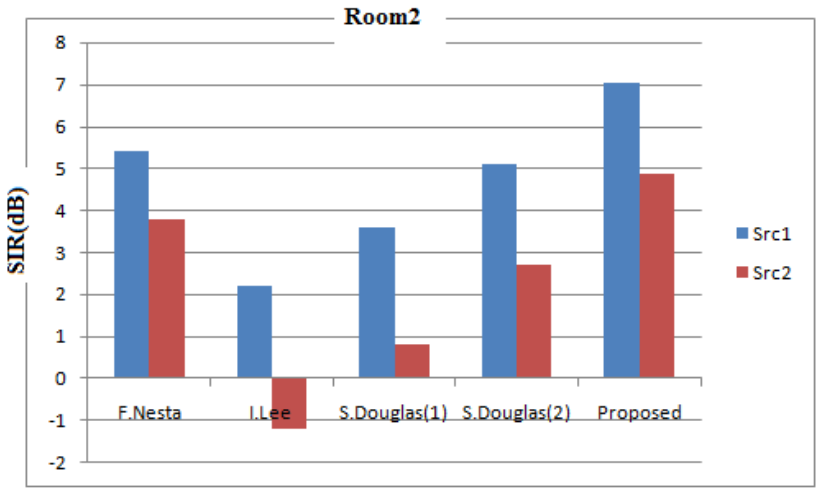
(b) Computation time

Figure 3.9 Computational complexity of FastICA algorithm

3.7.3 Results of Convolutive BSS Using Complex FastICA Algorithm

Simple FastICA algorithm is not able to separate the real-time mixture signals because of two reasons (i) the presence of background noises while recording the signals (ii) the real-time mixture signals are not instantaneous in nature, but convolutive mixture signals. Even in the absence of noises also, simple FastICA algorithm is not able to separate the source signals as its basic assumption is that the mixing process is instantaneous.





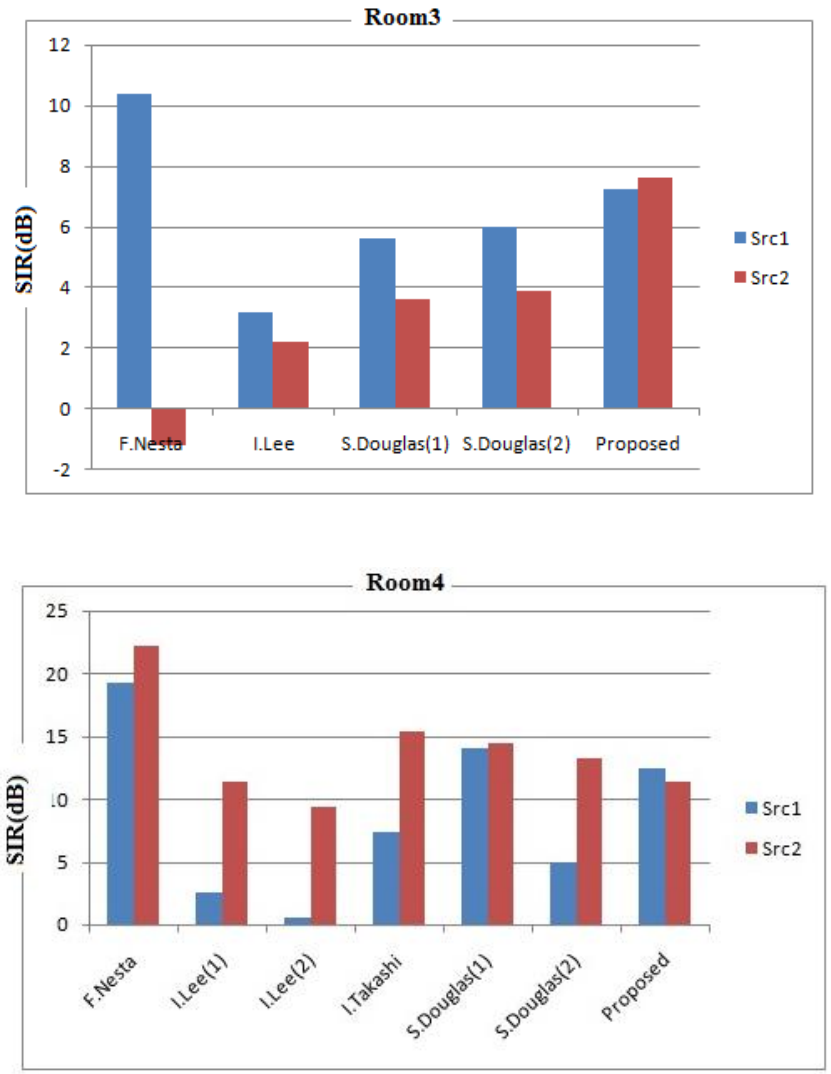


Figure 3.10 Performance comparison of the complex FastICA algorithm

In order to test the proposed complex FastICA algorithm for the separation of the convolutive mixture signals, the proposed algorithm is compared with many state of art BSS algorithms presented in Vincent *et al.* (2009a). Figure 3.10 compares the performance of the proposed method with various BSS algorithms. Since the reference mono signals are not provided in the results, the results of the algorithm are evaluated using the sources that contribute for first mixture signals. Hence only SIR results are provided.

In all the four different size rooms, the proposed method improves the SIR considerably when compared with the other BSS algorithms. Moreover, the SIR of the source 1 and source 2 are equally well in all the conditions. This means that the algorithm separates both the sources equally well. Figure 3.11 shows the results of the proposed algorithm for various room sizes.

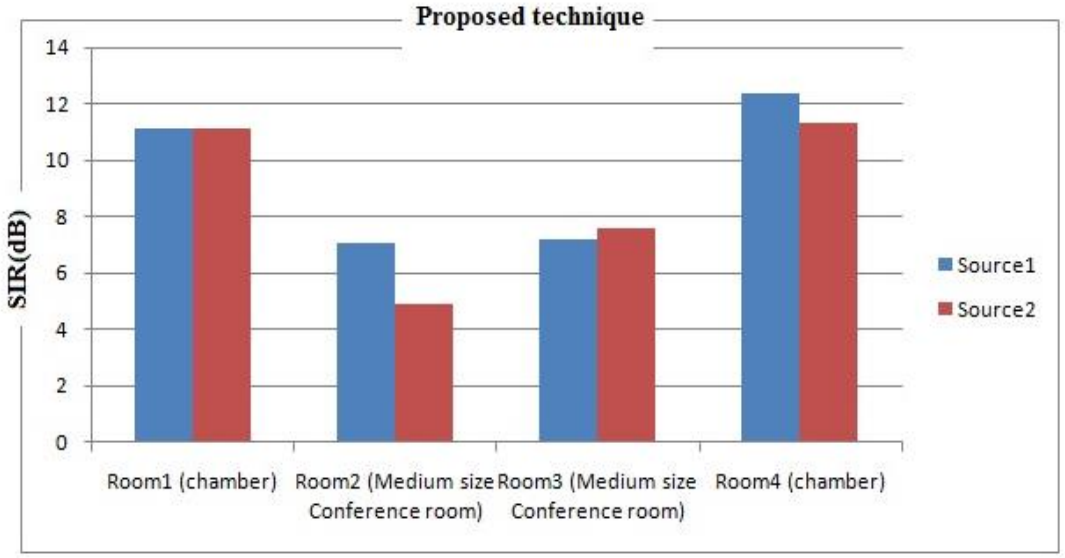


Figure 3.11 Performance of the complex FastICA algorithm for various room sizes

3.7.4 Results of Permutation Alignment Technique

In this section, the result of the proposed permutation alignment technique is discussed. Figure 3.12 shows the frequency bin index versus the separated sources. The value '1' indicates that the separated source frequency bin is matched with that of the original source frequency bins. i.e., the frequency bin is correctly aligned. The value '0' indicates that the corresponding frequency bins are misaligned when compared with original source frequency bin. Figure 3.13 compares the proposed method with

adjacent frequency band correlation method. The proposed method reduces the misaligned frequency bins from 20% to 16%.

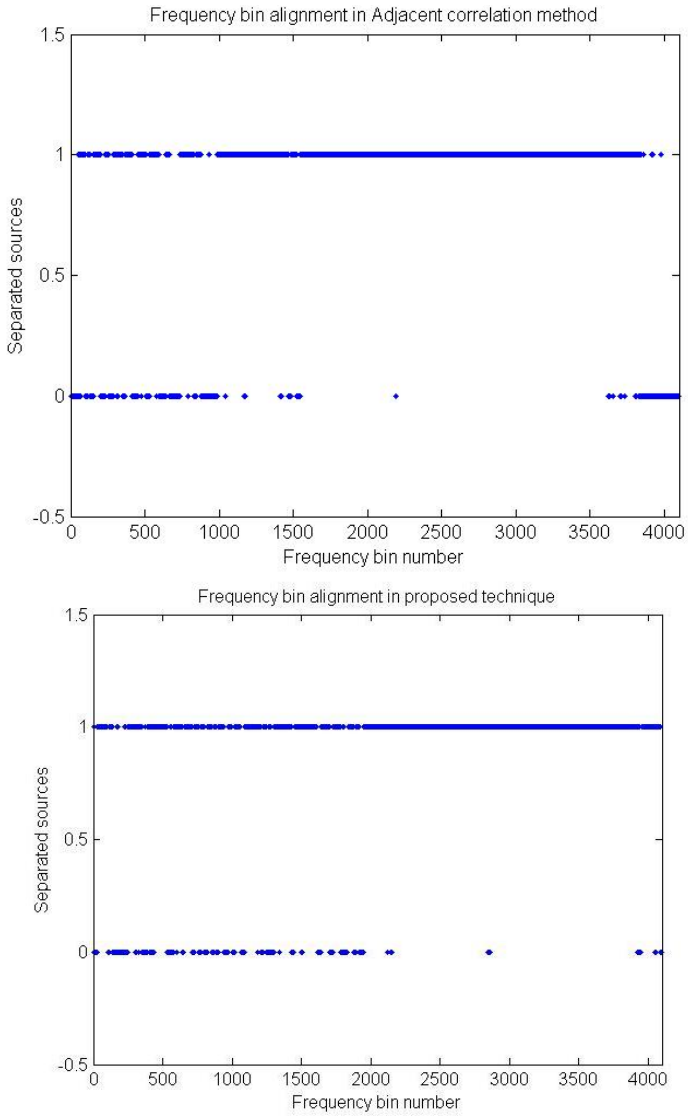


Figure 3.12 Results of the Permutation alignment algorithms

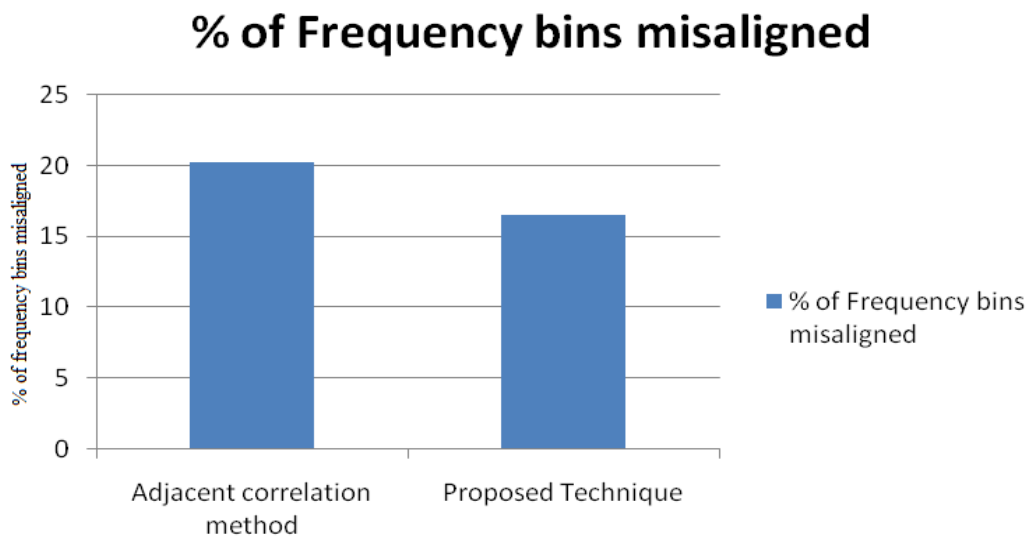


Figure 3.13 Performance of the proposed permutation alignment technique

3.8 CONCLUSION

The FastICA algorithm is successful to separate the sources from its overdetermined/determined instantaneous mixture signals. But the speech mixture signals are not instantaneous in nature, and more often it is modelled by convolutive mixture signals. Hence it is necessary to extend the FastICA algorithm for the overdetermined convolutive mixture signals separation. The convolutive mixing process becomes instantaneous when converting the convolutive mixture signals into the frequency domain. So the processing of these signals involves complex values. Therefore, a simple FastICA algorithm cannot be directly applied to it. A complex FastICA algorithm is proposed for the BSS of overdetermined convolutive mixture signals. Another problem occurred while reconstructing the sources from the frequency domain to the time domain is the permutation problem. Two-pass permutation alignment technique based on the correlation between the power ratios of the adjacent frequency bin is proposed. The results demonstrate the efficiency of the

proposed algorithm in the separation of the convolutive mixture signals when compared with the state of art algorithms.

However, the proposed complex FastICA algorithm has the capability to separate the convolutive mixture signals when the number of microphones is at least equal to the number of sources. This condition cannot be satisfied in all practical situations. In order to overcome this limitation, two methods are proposed for the separation of the underdetermined convolutive mixture signals and it is analyzed in the next chapter.

CHAPTER 4

UNDERDETERMINED CONVOLUTIVE BLIND SOURCE SEPARATION USING SINGLE SOURCE POINT DETECTION ALGORITHM

Complex FastICA algorithm performance is good, when the number of microphones is at least equal to the number of sources. But practically this condition is not satisfied in all situations. In order to overcome this limitation, two methods are proposed for the separation of the underdetermined convolutive mixture signals. The algorithms are mixing matrix estimation method using the Single Source Point (SSP) detection algorithm and Time-Frequency (TF) mask construction method using capsule networks. Mixing matrix estimation method using SSP detection algorithm is discussed and analyzed in this chapter.

4.1 INTRODUCTION

This chapter addresses the problem of estimation of the mixing matrix and hence the sources from their underdetermined mixtures. Complex FastICA algorithms and other ICA based algorithms are successful in separating the sources from overdetermined/determined mixture signals when the model is noiseless. In a typical environment, it is impossible to record the signals without noise. The presence of noise in the recorded mixture signals violates the basic requirements of ICA. The recorded mixture signals will not be overdetermined mixtures as noises act as additional sources. Hence ICA based algorithms have the limitations in separating the real-time mixture signals. When number of sources is more than the number of mixtures, underdetermined BSS technique is required.

ICA separates the sources using only the observed mixture signals by utilizing the prior information of statistical independence between the sources. Likewise, Sparse Component Analysis (SCA) uses the sparsity of the sources in the transform domain as prior information for BSS. The advantage of the SCA method is that, it is possible to estimate the mixing matrix in addition to the separation of the source signals. In the framework of SCA, the sources are assumed to be sparse which means only few points are nonzero and contain most of the information of the signal. The BSS techniques assume that the sparsity of the sources is fully or partially disjoint.

4.1.1 Basic Principles of Sparse Component Analysis

The underdetermined instantaneous mixing process is mathematically described as

$$\boldsymbol{x(n) = As(n)} \tag{4.1}$$

where $\mathbf{x(n)} = [x_1(n), x_2(n), \ldots, x_M(n)]^T$ are the M mixed signals, **A** is the mixing matrix of order M ×N and $\mathbf{s(n)} = [s_1(n), s_2(n), \ldots, s_N(n)]^T$ are the sources. For the underdetermined mixtures, M is assumed as less than N.

When mixing process is underdetermined, sparse component analysis is used. The sparsity of the speech signals are more in the frequency domain than in the time domain (Bofill *et al.* 2001). Hence the transformation of the observed signals into frequency domain is necessary to utilize the sparse domain better. Short Time Fourier Transform (STFT) is used to transform the speech signals into TF domain as the speech signals are pseudo-stationary.

The idea of using TF points for BSS is reported by Belouchrani *et al.* (1998) and this algorithm is for the separation of sources

from overdetermined mixtures. The algorithm is extended to underdetermined mixtures in Nguyen *et al.* (2001).The algorithm proposed in Abrard *et al.* (2005) is based on the complex ratio of mixture signals in TF domain and it is called as TIme Frequency Ratio Of Mixtures (TIFROM). Single Source Point (SSP) is detected based on the complex ratio of the mixtures. If only one source is active, then such ratio of mixtures will be constant wherever the particular is only active.

Hence identifying the constant complex ratio of mixtures is equivalent to detecting the SSPs. These constant values are used to estimate the cancelling coefficients and these cancelling coefficients are used to estimate the sources further. The advantage of sparsity based methods is that (i) the underdetermined mixing matrix is estimated (ii) the sources are extracted from the observed mixture signals by using the estimated mixing matrix.

4.2 EXISTING METHODS FOR UNDERDETERMINED BSS

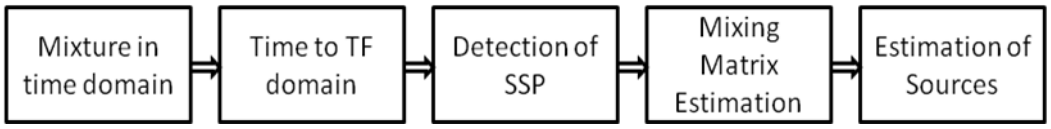


Figure 4.1 The flow of operations for underdetermined BSS

Figure 4.1 illustrates the flow of operations in the process of underdetermined BSS. The underdetermined instantaneous mixing model in equation (4.1) is transformed into TF points by using STFT and it is given as

$$\boldsymbol{X(t, f) = AS(t, f)} \tag{4.2}$$

where **X(t,f)** and **S(t,f)** are STFT coefficients respectively in the frequency bin f and at time t. **A** is the mixing matrix and it is assumed as time-invariant.

4.2.1 Single Source Point (SSP) Detection Methods

For our convenience without affecting the generality of the underdetermined mixing condition, it is assumed that there are only three sources and two microphones record the mixture of the signals i.e., N = 3, M = 2. At any point on the TF plane, the mixture signals are given as

$$\begin{bmatrix} X_1(t_1,f_1) \\ X_2(t_1,f_1) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} S_1(t_1,f_1) \\ S_2(t_1,f_1) \\ S_3(t_1,f_1) \end{bmatrix} \tag{4.3}$$

The equation is rewritten by scaled version of the source signals as

$$\begin{bmatrix} X_1(t_1,f_1) \\ X_2(t_1,f_1) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ \frac{a_{21}}{a_{11}} & \frac{a_{22}}{a_{12}} & \frac{a_{23}}{a_{13}} \end{bmatrix} \begin{bmatrix} a_{11}S_1(t_1,f_1) \\ a_{12}S_2(t_1,f_1) \\ a_{13}S_3(t_1,f_1) \end{bmatrix} \tag{4.4}$$

As the sources are assumed as sparse in the TF domain, i.e., only a few points are having most of the information, there are possibilities of TF points with only one source component is active. For example, if the source s_1 is only active component at (t_1,f_1) point and remaining sources are zero i.e., $S_1(t_1,f_1) \neq 0$ and $S_2(t_1,f_1) = S_2(t_1,f_1) = 0$.

$$\begin{bmatrix} X_1(t_1,f_1) \\ X_2(t_1,f_1) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ \frac{a_{21}}{a_{11}} & \frac{a_{22}}{a_{12}} & \frac{a_{23}}{a_{13}} \end{bmatrix} \begin{bmatrix} a_{11}S_1(t_1,f_1) \\ 0 \\ 0 \end{bmatrix} \tag{4.5}$$

The ratio between the two mixtures signals are utilized to find the one of the elements of the mixing matrix.

$$\frac{X_2(t_1,f_1)}{X_1(t_1,f_1)}=\frac{a_{21}}{a_{11}} \tag{4.6}$$

The value of the ratio between the two mixtures signals is unique for each source if only one source is active and all other sources are zero. By grouping the ratio of the mixture signals into three clusters by some clustering procedure, it is possible to estimate the mixing matrix value. But this is possible only if the observed mixture signals are perfectly sparse either in the time domain or frequency domain. The real-time mixture signals are not perfectly sparse either in the time domain or frequency domain.

Reju *et al.* (2009) proposed a method for detecting single source active points based on the real and imaginary parts of the ratio between the observed mixture signals. If only one source s_1 is active at (t_1,f_1) point in the TF plane, then the observed mixture signal is

$$\begin{bmatrix} X_1(t_1,f_1) \\ X_2(t_1,f_1) \end{bmatrix}=a_{11}S_1(t_1,f_1)\begin{bmatrix} 1 \\ \frac{a_{21}}{a_{11}} \end{bmatrix} \tag{4.7}$$

The real and imaginary parts of the equation (4.7) is given as

$$Real\begin{bmatrix} X_1(t_1,f_1) \\ X_2(t_1,f_1) \end{bmatrix}=Real\{a_{11}S_1(t_1,f_1)\}\begin{bmatrix} 1 \\ \frac{a_{21}}{a_{11}} \end{bmatrix} \tag{4.8}$$

$$Imag\begin{bmatrix} X_1(t_1,f_1) \\ X_2(t_1,f_1) \end{bmatrix} = Imag\{a_{11}S_1(t_1,f_1)\}\begin{bmatrix} 1 \\ \frac{a_{21}}{a_{11}} \end{bmatrix} \quad (4.9)$$

The absolute directions of equation (4.8) and (4.9) will be same and equal to the direction of the mixing matrix elements $[1\ a_{21}/a_{11}]^T$. Similarly consider another point (t_2, f_2) in the TF plane where only source s_2 is active,

$$Real\begin{bmatrix} X_1(t_2,f_2) \\ X_2(t_2,f_2) \end{bmatrix} = Real\{a_{12}S_2(t_2,f_2)\}\begin{bmatrix} 1 \\ \frac{a_{22}}{a_{12}} \end{bmatrix} \quad (4.10)$$

$$Imag\begin{bmatrix} X_1(t_2,f_2) \\ X_2(t_2,f_2) \end{bmatrix} = Imag\{a_{12}S_2(t_2,f_2)\}\begin{bmatrix} 1 \\ \frac{a_{22}}{a_{12}} \end{bmatrix} \quad (4.11)$$

The absolute directions of equation (4.10) and (4.11) are same and equal to the direction of the mixing matrix elements $[1\ a_{22}/a_{12}]^T$. Now consider another TF point (t3, f3), where both the sources are active, then the real and imaginary part of the observed mixture signals are,

$$Real\begin{bmatrix} X_1(t_3,f_3) \\ X_2(t_3,f_3) \end{bmatrix} = Real\{a_{11}S_1(t_3,f_3)\}\begin{bmatrix} 1 \\ \frac{a_{21}}{a_{11}} \end{bmatrix} + Real\{a_{12}S_2(t_3,f_3)\}\begin{bmatrix} 1 \\ \frac{a_{22}}{a_{12}} \end{bmatrix} \quad (4.12)$$

$$Imag\begin{bmatrix} X_1(t_3,f_3) \\ X_2(t_3,f_3) \end{bmatrix} = Imag\{a_{11}S_1(t_3,f_3)\}\begin{bmatrix} 1 \\ \frac{a_{21}}{a_{11}} \end{bmatrix} + Imag\{a_{12}S_2(t_3,f_3)\}\begin{bmatrix} 1 \\ \frac{a_{22}}{a_{12}} \end{bmatrix} \quad (4.13)$$

Reju *et al.* (2009) proved empirically that there is negligible chance to have equal absolute directions of the real and imaginary parts of the

Multi-Source Points (MSP). Hence Single Source Point (SSP) is detected based on the difference between the ratios are given in equation (4.14).

$$\left|1-\frac{Real\{\boldsymbol{X(t,f)}\}}{Imag\{\boldsymbol{X(t,f)}\}}\right|=0 \tag{4.14}$$

However, in practice the probability of only one source is active with all other source amplitude equal to zero is very less. Hence there is a small mismatch between the absolute direction of the real and imaginary parts of the observed mixture signals. So the condition in the equation (4.14) is relaxed as

$$\left|1-\frac{Real\{\boldsymbol{X(t,f)}\}}{Imag\{\boldsymbol{X(t,f)}\}}\right|<0.01 \tag{4.15}$$

Lu *et al.* (2019) proposed a similar method for identifying the single source points.

4.2.2 Mixing Matrix Estimation Methods

After obtaining SSPs, the mixing matrix is estimated by clustering techniques. It is worth noting here that the number of sources is assumed to be known in advance before applying the clustering techniques. Many clustering techniques are proposed in the literature for estimating the mixing matrix. Hierarchical clustering procedure is used in Reju *et al.* (2009). K means and Fuzzy C means clustering procedures are also used to estimate the mixing matrix. Lu *et al.* (2019) proposed Density Peak Clustering (DPC) procedure to find the mixing matrix. DPC is improved to identify the cluster centers and number of clusters automatically. Hence by using DPC, it is used to identify the number of sources automatically in this method.

4.2.3 Source Estimation Methods

Even though the exact mixing matrix is identified, it is not simple to estimate the sources in underdetermined cases since the mixing matrix is not square. The common approach to sparse BSS is that l1 norm decomposition method (Bofill *et al.* 2001; Hu *et al.* 2016) as

$$\min_{s(t)} \|s(t)\|_1 \text{ such that } x(t) = As(t) \tag{4.16}$$

In fact, minimizing $\|s(t)\|_1$ means that finding the shortest path to **x(t)** by having mixing matrix **A** from all possible solutions. This l1 norm decomposition method is easily be extended to multi-dimensional cases.

$$\min_{S(t,f)} \|S(t,f)\|_1 \text{ such that } X(t,f) = AS(t,f) \tag{4.17}$$

The sources are extracted based on the Minimum Mean Square Error (MMSE) methods also as described in Cho *et al.* (2011).

$$S(t,f) = A^{\dagger} X(t,f) + Vz \tag{4.18}$$

where $A^{\dagger}$ is pseudo-inverse of the mixing matrix **A, V** is an N×(N-M) matrix whose columns are bases of **A. z** is an arbitrary vector of size (N-M)×1. The problem of estimation of **S(t, f)** now becomes the estimation of the arbitrary vector **z**. The vector **z** is derived by minimizing MMSE as

$$\hat{z}_{ms} \approx \frac{1}{z_p} \sum_{m=1}^{nC_m} p\left(S = A^{\dagger} X + Vz_m^{*}\right) z_m^{*} \tag{4.19}$$

The sources are estimated with minimum mean square error using the vectors $\hat{z}_{ms}$ as

$$\hat{\boldsymbol{S}}_{ms} = \boldsymbol{A}^{\dagger}\boldsymbol{X} + \boldsymbol{V}\hat{\boldsymbol{z}}_{ms} \tag{4.20}$$

4.3 DRAWBACKS OF EXISTING TECHNIQUES

Previous methods in the underdetermined BSS are suitable for instantaneous mixture signals. These methods estimate the mixing matrix based on the detection of SSPs. After detecting the SSPs, they are grouped using clustering procedure and the cluster centers are used to estimate the mixing matrix. The main disadvantage of these methods is low estimation accuracy of the mixing matrix for insufficient sparse sources. Practically, the sparse of the sources is reduced under noisy conditions. Hence the SSPs identification and mixing matrix estimation accuracy reduced when the sources are contaminated by the noise.

The sparsity of the sources is affected when the mixture signals are convolutive. The instantaneous mixing process has only one path between the sources to microphone. So the source amplitude reaches the microphone only once with path attenuation. The recorded mixture signals naturally have the sparsity of the sources. But while recording the speech sources in the typical room environment, the recorded signals are not be instantaneous in nature. Because of the room reverberations, the recorded signals are modelled by the convolutive mixture signals. The convolutive mixing process has multiple paths between the sources to microphone. So the same source is received in the microphone multiple times at different delays. This affects the sparsity of the signals in convolutive mixtures. Hence the separation of underdetermined convolutive BSS is difficult when compared with underdetermined instantaneous BSS. The existing techniques using SSP identification mainly focus on the BSS of the instantaneous mixture signals.

4.4 PROPOSED METHOD FOR UNDERDETERMINED BSS

Two algorithms are proposed for underdetermined BSS based on SSP detection method, one for instantaneous and another for convolutive mixture signals. Even though convolutive mixing process more closely replicates the real situation of the acoustic recording of the signals, instantaneous mixing process is also be useful for some applications. Because the model of the mixing process is chosen mainly by the application for which the BSS is used. Hence for unifying approach both instantaneous and convolutive BSS algorithms are presented.

4.4.1 Underdetermined Instantaneous BSS Using SSP Detection Algorithm

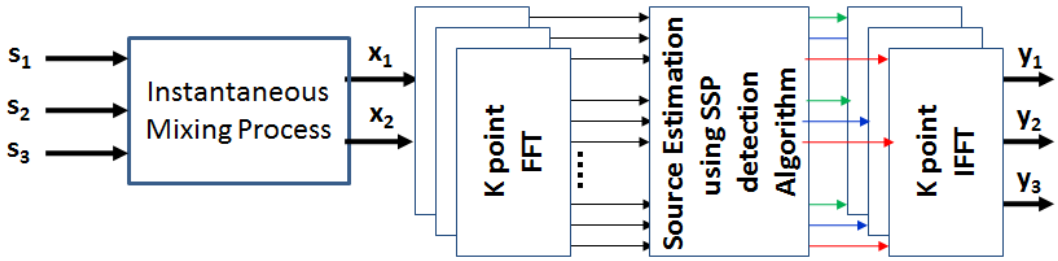


Figure 4.2 BSS of underdetermined instantaneous mixtures

The instantaneous mixture signals are applied as the input to this algorithm. In the first step, the signals are converted into TF domain using short-time Fourier transform, as the sparsity is well exploited in the TF domain rather than in the time domain. The main part of this algorithm is the detection of single source points. Let assume that the number of sources is three and the number of observed mixture signals is two. The observed mixture signals in TF domain are described as

$$X_1(t,f) = a_{11}S_1(t,f) + a_{12}S_2(t,f) + a_{13}S_3(t,f) \quad (4.21)$$

$$X_2(t,f) = a_{21}S_1(t,f) + a_{22}S_2(t,f) + a_{23}S_3(t,f) \tag{4.22}$$

At any TF point, say (t_1,f_1) if only one source is active, then

$$\begin{bmatrix} X_1(t_1,f_1) \\ X_2(t_1,f_1) \end{bmatrix} = a_{11}S_1(t_1,f_1)\begin{bmatrix} 1 \\ \frac{a_{21}}{a_{11}} \end{bmatrix} \tag{4.23}$$

Equation (4.23) shows that the $\mathbf{X(t_1,f_1)}$ will be collinear with $[1\ a_{21}/a_{11}]^T$. Now assume that at another TF point (t_2, f_2) the same source is active.

$$\begin{bmatrix} X_1(t_2,f_2) \\ X_2(t_2,f_2) \end{bmatrix} = a_{11}S_1(t_2,f_2)\begin{bmatrix} 1 \\ \frac{a_{21}}{a_{11}} \end{bmatrix} \tag{4.24}$$

From equation (4.24), it is clear that the $\mathbf{X(t_2,f_2)}$ is also collinear with $[1\ a_{21}/a_{11}]^T$. From equations (4.23) and (4.24), all the SSPs where only the same source is active, the mixture TF points will be collinear.

$$\boldsymbol{X(t_1, f_1)} = r\boldsymbol{X(t_2, f_2)} \tag{4.25}$$

where r is constant. By normalizing the vectors on both sides,

$$\tilde{\boldsymbol{X}}\boldsymbol{(t_1, f_1)} = \tilde{\boldsymbol{X}}\boldsymbol{(t_2, f_2)} \tag{4.26}$$

where $\tilde{\boldsymbol{X}}$ represents the normalized vector. Hence, all the vectors are normalized and the SSPs are identified by using the following equation.

$$\left|1 - \left\langle \tilde{\boldsymbol{X}}\boldsymbol{(t_1, f_1)}\tilde{\boldsymbol{X}}\boldsymbol{(t_2, f_2)} \right\rangle\right| < \varepsilon \tag{4.27}$$

where $\langle . \rangle$ is the scalar product between the vectors and ε is a very small threshold value close to zero. The threshold ε is needed as the probability of only one source is active with all other source amplitude equal to zero is very less.

After detecting enough number of SSPs using the equation (4.27), the points are grouped by using k-means clustering procedure. The mixing matrix is estimated using cluster centers. As underdetermined mixing signals are processed, the mixing matrix is not square. Hence it is not possible to find the exact solution to the inverse of the mixing matrix. A simple way is that Moore-Penrose pseudo-inverse are used to find the inverse of the non-square mixing matrix. The pseudo-inverse provides the least square solution to the many possible solutions of the inverse of the mixing matrix. Moore-Penrose pseudo-inverse provides the Euclidean norm solution to the inverse of the mixing matrix. The estimated inverse of the mixing matrix is directly applied to the observed mixture signals in TF domain. Finally, the separated signals are converted back to time domain. The algorithm steps are given as follows.

1. Convert the observed mixture signals into TF domain. $\boldsymbol{X(t, f) = AS(t, f)}$

2. Normalize all the TF vectors $\boldsymbol{X(t, f)}$ to $\tilde{\boldsymbol{X}}\boldsymbol{(t, f)}$.

3. Search across all the TF vectors for the single source points (SSPs) using the normalized vectors obtained in step 2. $\left|1-\left\langle \tilde{\boldsymbol{X}}(t_1, f_1)\tilde{\boldsymbol{X}}(t_2, f_2)\right\rangle\right| < \varepsilon$

4. Apply k-means clustering procedure to the obtained SSPs after eliminating the outliers if any.

5. Estimate the mixing matrix $\tilde{\boldsymbol{A}}$ using the cluster centers.

6. Find the Moore-Penrose pseudo-inverse of the estimated mixing matrix. $\boldsymbol{B} = \tilde{\boldsymbol{A}}^{\dagger}$

7. Apply the separation matrix directly to the observed mixture signals.

8. Convert the separated signals back into time domain using inverse STFT.

4.4.2 Underdetermined Convolutive BSS Using SSP Detection Algorithm

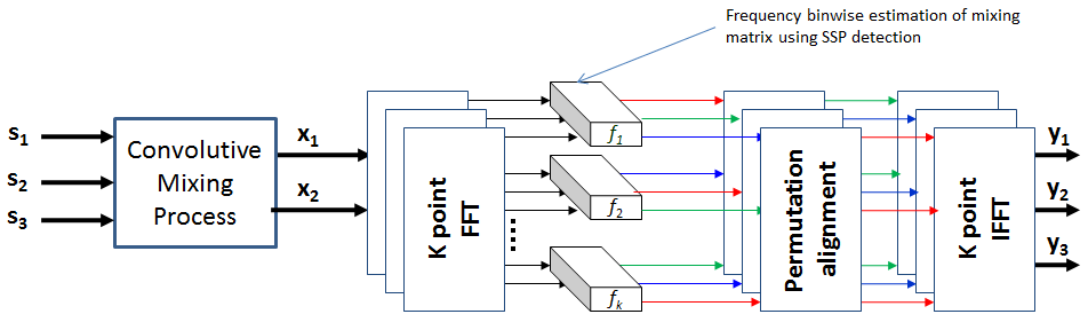


Figure 4.3 BSS of underdetermined convolutive mixtures

This algorithm gives the BSS of the underdetermined convolutive mixture signals. Without affecting the generality, let assume that there are three sources contribute for the two mixture signals. The convolutive mixture signals are mathematically described as

$$x_1(n) = h_{11}(n) * s_1(n) + h_{12}(n) * s_2(n) + h_{13}(n) * s_3(n) \tag{4.28}$$

$$x_2(n) = h_{21}(n) * s_1(n) + h_{22}(n) * s_2(n) + h_{23}(n) * s_3(n) \tag{4.29}$$

where $h_{ij}(n)$ is the time-invariant impulse response of the channel from the source j to the microphone i. $x_1(n)$ and $x_2(n)$ are the mixture signals of the sources $s_1(n)$, $s_2(n)$ and $s_3(n)$ in the time-domain. * denotes the convolution

operator. In the first step, the observed signals are converted into TF domain using STFT.

$$X_1(t,f) = H_{11}(f)S_1(t,f) + H_{12}(f)S_2(t,f) + H_{13}(f)S_3(t,f) \quad (4.30)$$

$$X_2(t,f) = H_{21}(f)S_1(t,f) + H_{22}(f)S_2(t,f) + H_{23}(f)S_3(t,f) \quad (4.31)$$

At any TF point, say (t_1, f_1) if only one source i.e., s_1 is active, then

$$X_1(t_1,f_1) = H_{11}(f_1)S_1(t_1,f_1) \quad (4.32)$$

$$X_2(t_1,f_1) = H_{21}(f_1)S_1(t_1,f_1) \quad (4.33)$$

The above equation is simply written in vector form as

$$\boldsymbol{X(t_1, f_1)} = S_1(t_1,f_1)\boldsymbol{H_1(f_1)} \quad (4.34)$$

where $\mathbf{H_1(f_1)}$ = $[1\ H_{21}(f_1)/H_{11}(f_1)]^T$. Now consider some other TF point in the same frequency bin say (t_2, f_1) where the same source s_1 is only active.

$$\boldsymbol{X(t_2, f_1)} = S_1(t_2,f_1)\boldsymbol{H_1(f_1)} \quad (4.35)$$

Equations (4.34) and (4.35) are collinear with $\mathbf{H_1(f_1)}$ = $[1\ H_{21}(f_1)/H_{11}(f_1)]^T$. A remarkable note here is that the mixing matrix now depends on the frequency whereas it is constant in the instantaneous case. But the mixing matrix is considered as constant within the same frequency bin provided good frequency resolution in the TF domain. Hence in any frequency bin if there are two points with the same active source, then

$$\boldsymbol{X(t_1, f_1)} = r\boldsymbol{X(t_2, f_1)} \quad (4.36)$$

where r is constant. After normalizing the vectors in the TF plane,

$$\tilde{\boldsymbol{X}}\boldsymbol{(t_1, f_1)} = \tilde{\boldsymbol{X}}\boldsymbol{(t_2, f_1)} \quad (4.37)$$

Hence, the single source points are identified at each frequency bins by using the following equation,

$$\left|1-\left\langle \tilde{X}(t_1, f_1)\tilde{X}(t_2, f_1)\right\rangle\right| < \varepsilon \qquad (4.38)$$

where ε is a very small threshold value to compensate the noises in the mixture signals. To separate the sources from the underdetermined convolutive mixture signals, each frequency bins should have atleast a pair of N number of single source points where N is the number of sources.

The mixing matrix is estimated for each frequency bins separately and their inverse is calculated by using the Moore-Penrose pseudo-inverse method. The pseudo-inverse of the mixing matrix is applied to the observed mixture signals to estimate the sources at each frequency bin. The permutation problem is inherent as the sources are estimated at each frequency bins separately. Hence before converting the signals back into the time domain using inverse STFT, the permutation alignment algorithm is needed to align the components order at each frequency bin. The permutation alignment algorithm proposed in chapter 3 is used here for this purpose. The permutation alignment is done based on the two-pass method using the correlation between the power ratios of adjacent frequency bins. The step by step procedure for the underdetermined convolutive BSS is given as follows.

1. Convert the observed mixture signals into TF domain. $X(t, f) = AS(t, f)$
2. Normalize the TF vectors at each frequency bin $X(t, f)$ to $\tilde{X}(t, f)$.
3. Find the single source points (SSPs) at each frequency bin using the normalized vectors obtained in step 2. $\left|1-\left\langle \tilde{X}(t_1, f_1)\tilde{X}(t_2, f_1)\right\rangle\right| < \varepsilon$

4. Apply k-means clustering procedure to the obtained SSPs at each frequency bins after eliminating the outliers if any.

5. Estimate the mixing matrices $\tilde{\boldsymbol{A}}$ using the cluster centers. The number of matrices is equal to the number of frequency bins used.

6. Find the Moore-Penrose pseudo-inverse of the estimated mixing matrices for each estimated mixing matrices. $\boldsymbol{B} = \tilde{\boldsymbol{A}}^{\dagger}$

7. Apply the separation matrix directly to the observed mixture signals in TF domain.

8. Apply the permutation alignment algorithm as described in chapter 3 to align the components order at each frequency bin.

9. Convert the separated signals back into time domain using inverse STFT.

4.5 RESULTS AND DISCUSSIONS

4.5.1 Experimental Setup

The speech sources and mixture signals have downloaded from the signal separation evaluation campaign website (Source: https://sisec.inria.fr/). The room dimensions are 4.45×3.55×2m. The room reverberation is set to either 130ms or 250ms. The algorithms are tested by using either three male speech mixture signals or three female speech mixture signals or three music mixtures. The distance between the two microphones is set to either 5cm or 1m. The distance between the microphone center and the sources are varied between 80cm and 1.2m. The direction of arrival of the source to the microphone is varying between -60 degrees to +60 degrees. For each mixing condition, the mixture signals are generated from different sets of source signals placed at different positions in the room.

4.5.2 Experimental Results for Underdetermined Instantaneous Mixtures

The blind separation of the underdetermined instantaneous mixture signals are considered in this section. Before applying the mixture signals to the proposed BSS algorithm, let verify the sparsity of the observed mixture signals by using scatter diagram as shown in Figure 4.4. The figure shows the sparsity of the mixture signals in both the time domain and in the TF domain. Obviously, it can be seen from the figure that the normally mixed signals do not provide the sparsity property in the time domain. The improvement in the sparsity can be seen when converting the mixture signals into TF domain using STFT. But this sparsity is not sufficient for source separation. The figure 4.5 shows the orientation of the sources clearly, while keeping the single source points and excluding the other points in the TF domain.

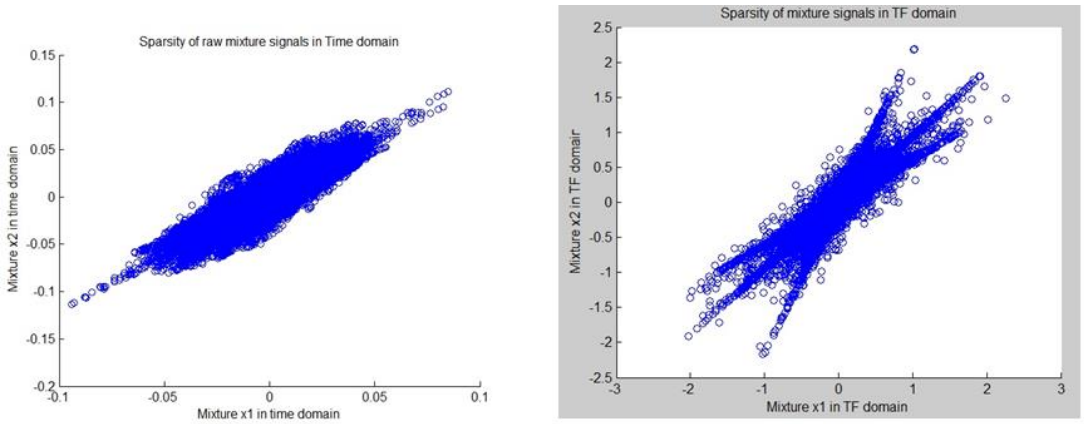


Figure 4.4 Sparsity of the instantaneous mixture signals in the time domain and TF domain

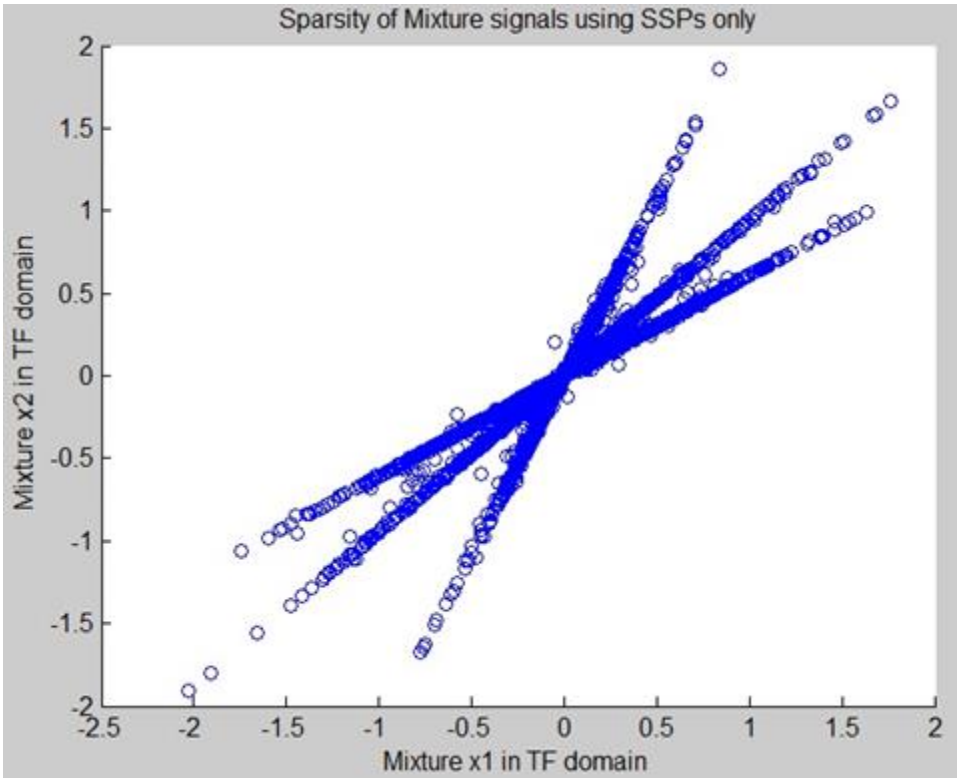


Figure 4.5 Sparsity of the instantaneous mixture signals using SSPs only

Table 4.1 gives the results of the underdetermined instantaneous BSS of mixture signals created from three female speech sources or three male speech sources. The algorithm is evaluated using the parameters SIR, SDR and, SAR all in decibels (dB). As discussed in chapter 3, If SIR is equal with SDR, then it implies that the noises and artifacts are absent in the estimated sources. The absence of the artifact is confirmed with the values of SAR. Thus, it is clear from the Figure 4.6 that the sources have been successfully separated. The success of the algorithm cannot be determined solely by the value of SIR, SDR and SAR. Besides, the difference between the SIR and SDR is critical to measure the success of the algorithm. The difference between the SIR and the SDR is proportional to the interferences and artifacts in the extracted sources. Therefore the equivalent value of the SIR and the SDR indicates the absence of interferences and artifacts in the

extracted sources. On this basis, the Figure 4.6 demonstrates the success of the proposed algorithm.

Table 4.1 Results of the underdetermined instantaneous BSS algorithm

Parameter	**Types of Mixture signals**	**Source1**	**Source2**	**Source3**
SIR (dB)	Female3	15.24	3.04	10.10
SDR (dB)	Instantaneous	15.21177	3.045776	10.10529
SAR (dB)	Mixture signals	37.05	40.78	41.58
SIR (dB)	Male3	14.80	3.09	10.18
SDR (dB)	Instantaneous	14.80	3.09	10.18
SAR (dB)	Mixture signals	49.09	59.86	59.16

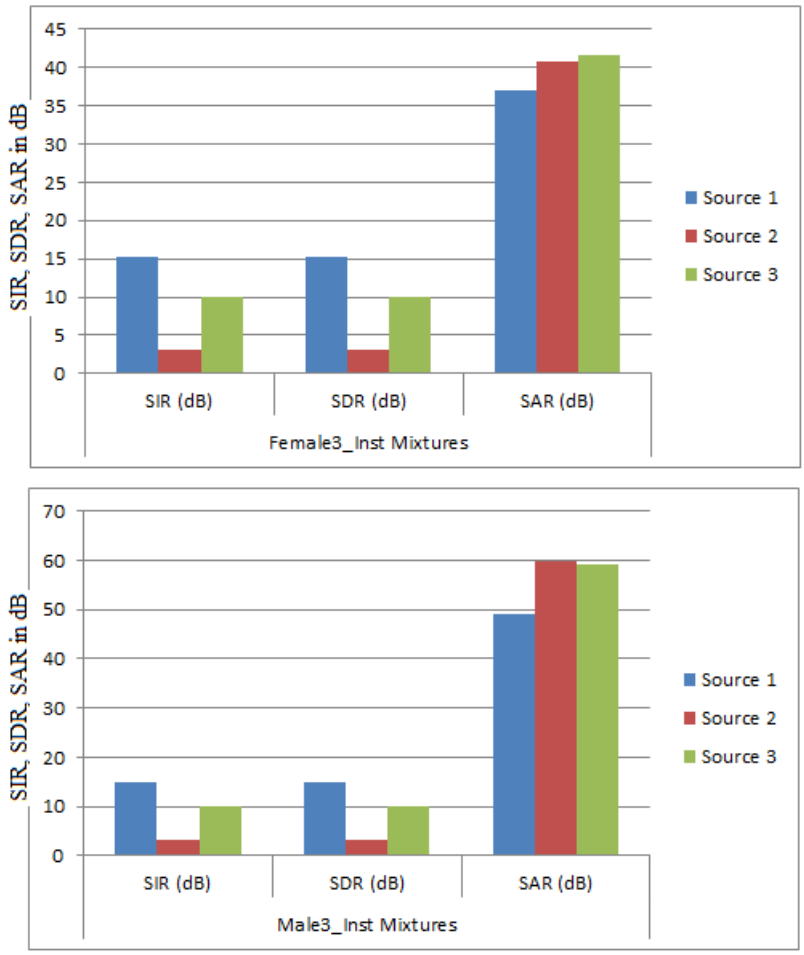


Figure 4.6 Performance of the proposed algorithm for the male and female instantaneous mixture signals

Figure 4.7 shows one of the sources and its extracted version from the underdetermined instantaneous BSS algorithm in the time domain.

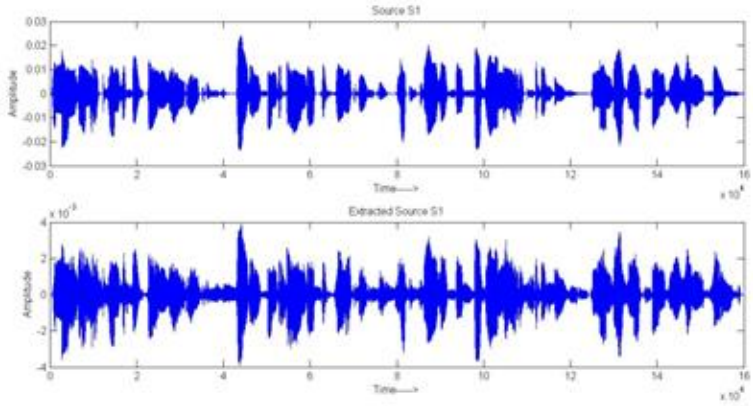


Figure 4.7 Time domain representations of one of the sources and the corresponding extracted source

Figure 4.8 compares the performance of the proposed algorithm with its state of art algorithms. The following algorithms are used for comparison.

A. Nesbit – Extension of sparse, adaptive signal decomposition to blind source separation (Nesbit *et al.* 2009)

A. Ozerov – Non-negative matrix factorization method (Ozerov *et al.* 2009)

E. Vincent – Underdetermined audio source separation via local Gaussaian modeling (Vincent *et al.* 2009b)

M. Cobos – Stereo audio source separation using based on time-frequency masking (Cobos *et al.* 2008)

S.Arberet – Blind spectral-GMM estimation for underdetermined instantaneous audio source separation (Arberet *et al.* 2009)

Z. El Chami – A new model based underdetermined source separation (El Chami *et al.* 2008)

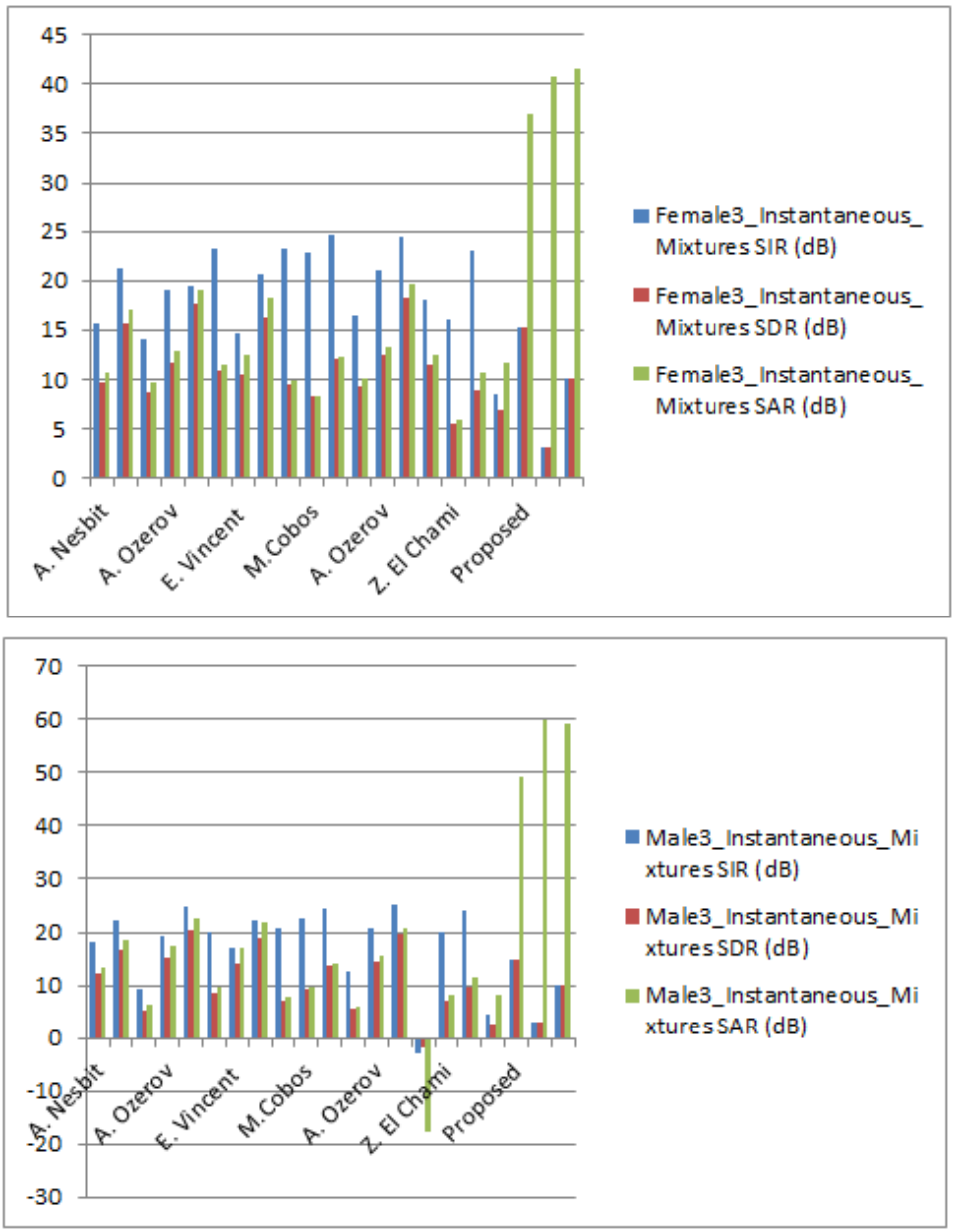


Figure 4.8 Comparison of the proposed underdetermined instantaneous BSS algorithm with state of art algorithms

4.5.3 Experimental Results for Underdetermined Convolutive Mixtures

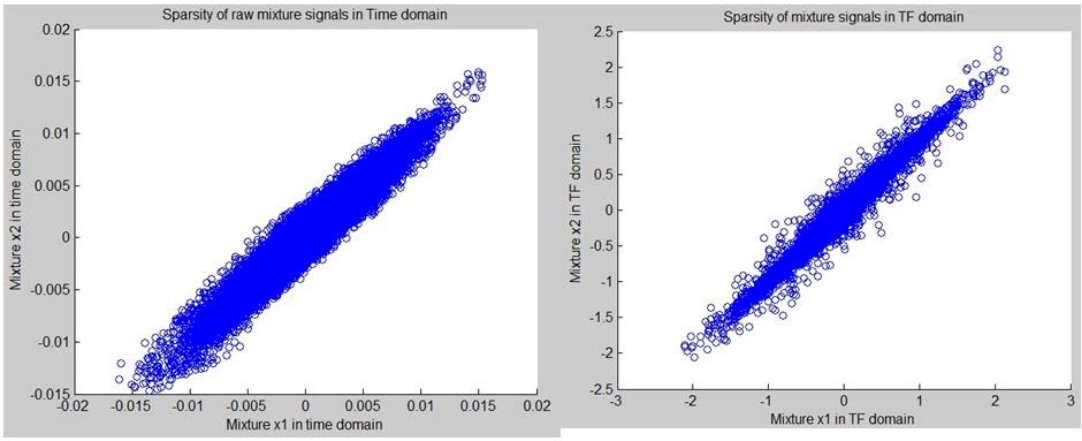


Figure 4.9 Sparsity of the convolutive mixture signals in the time domain and TF domain

Table 4.2 Performance of the underdetermined convolutive BSS algorithm with microphone spacing of 5cm

Types of Mixture signals	Source	130ms Reverberation			250ms Reverberation		
		SIR (dB)	SDR (dB)	SAR (dB)	SIR (dB)	SDR (dB)	SAR (dB)
Female3 convolutive mixture signals	Source 1	5.98	4.24	20.19	6.27	6.34	17.32
	Source 2	4.37	4.50	23.11	3.31	3.50	21.24
	Source 3	4.11	4.37	20.07	4.62	4.76	22.72
Male3 convolutive mixture signals	Source 1	4.64	4.89	20.20	4.07	4.44	18.65
	Source 2	3.08	3.27	21.27	3.36	3.65	19.29
	Source 3	4.17	4.27	24.11	4.11	4.36	20.48

Table 4.3 Performance of the underdetermined convolutive BSS algorithm with microphone spacing of 1m

Types of Mixture signals	Source	130ms Reverberation			250ms Reverberation		
		SIR (dB)	SDR (dB)	SAR (dB)	SIR (dB)	SDR (dB)	SAR (dB)
Female3 convolutive mixture signals	Source 1	2.53	2.66	22.65	3.32	3.52	20.97
	Source 2	4.00	4.09	25.17	4.87	3.01	22.86
	Source 3	5.61	5.69	26.40	5.97	5.27	19.91
Male3 convolutive mixture signals	Source 1	2.58	2.63	26.84	3.39	3.59	20.73
	Source 2	4.51	4.62	23.78	4.15	4.25	24.53
	Source 3	5.06	5.13	26.38	5.74	5.86	23.42

Figure 4.9 shows the sparsity of the observed convolutive mixed signals in both time domain and TF domain. The figure illustrates that convolutive mixture signals do not have the sparsity property like the instantaneous mixture signals in the time domain or the TF domain. The performance of the algorithm is tested using various convolutive mixture signals made up of the male and female speech sources, recorded in the 130ms reverberation environment and 250ms reverberation environment. Table 4.2 and 4.3 gives the results of the underdetermined convolutive BSS algorithm with microphone spacing of 5cm and 1m respectively. Figure 4.10 and Figure 4.11 show the results with room reverberation of 130ms and 250ms respectively. All the results have approximately equal SIR and SDR with positive SAR. These results demonstrate the effectiveness of the proposed algorithm in separating the sources from its underdetermined convolutive mixture signals.

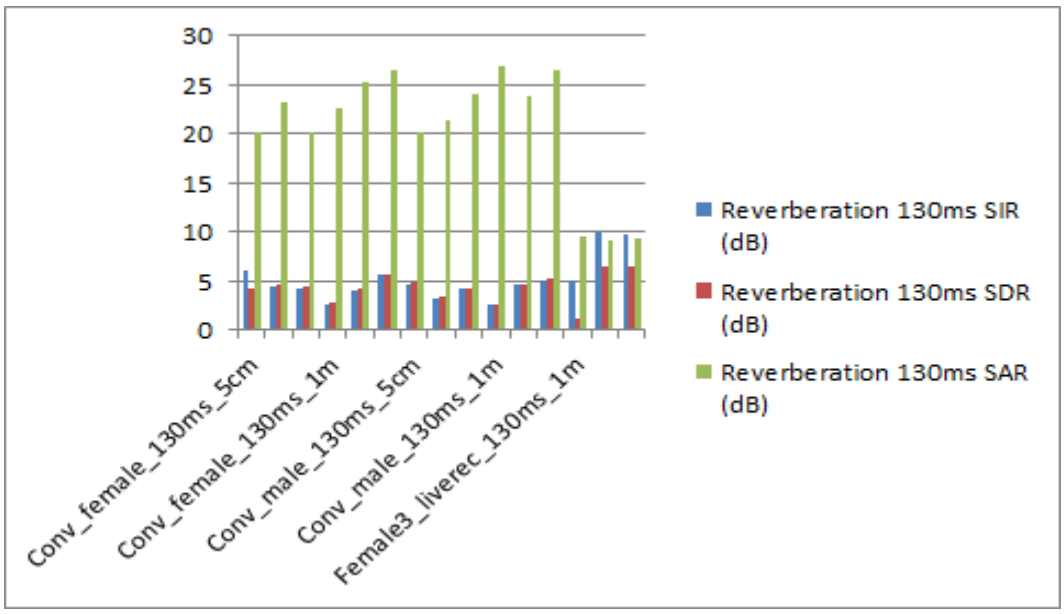


Figure 4.10 Performance of the proposed algorithm for the convolutive mixture signals recorded with 130ms reverberation

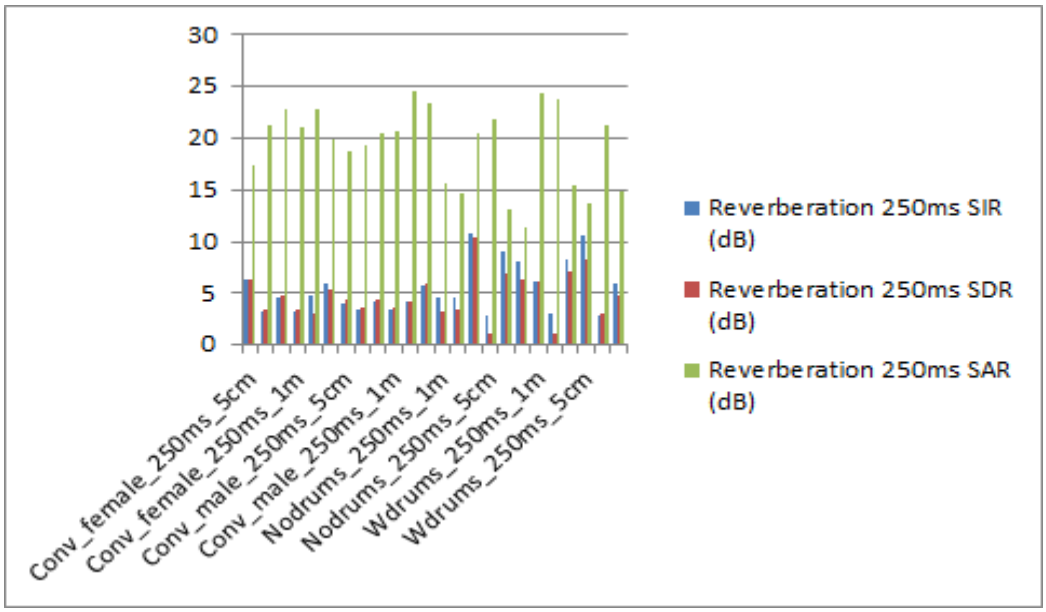


Figure 4.11 Performance of the proposed algorithm for the convolutive mixture signals recorded with 250ms reverberation

Figure 4.12 shows the comparisons of the proposed algorithm with its counterparts. Although the SIR value of M.Cobos and the binary mask is high, the difference between the SDR and SIR is high which denotes the interferences are present in the separated signals. The equal SIR and SDR values of the proposed method demonstrate that the absence interferences in the separated signals. Hence the proposed method performs better than the other techniques in the underdetermined convolutive BSS.

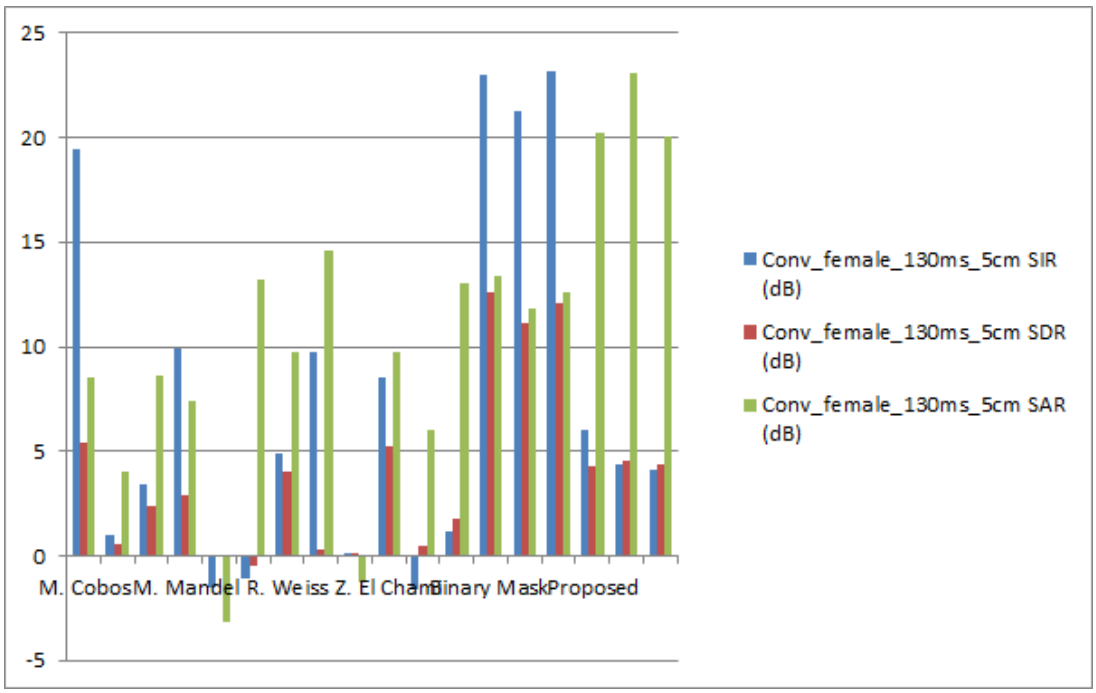


Figure 4.12 Comparison of the proposed underdetermined convolutive BSS algorithm with state of art algorithms

4.6 CONCLUSION

In this chapter, the blind source separation of the underdetermined instantaneous mixture signals and the underdetermined convolutive mixture signals are proposed using single-source point detection algorithms. The sparsity property does not exist naturally in the observed mixture signals.

When the mixture signals are converted into the TF domain, the sparsity increases somewhat still, it is not enough to separate the source signals. The condition is even worse for the underdetermined convolutive mixture signals. Hence the single-source point detection algorithm increases the sparsity well enough to separate the sources from the underdetermined mixture signals. The single-source point detection algorithm is applied for each frequency bin when the observed mixture signal is convolutive. The algorithm is tested on the various mixture signals. The various mixture signals are generated by varying the microphone spacing either 5cm or 1m and room reverberations set to either 130ms or 250ms. The results demonstrate the effective separation of the sources from the mixture signals without interferences and artifacts when compared with state of art algorithms. Whether the separated sources are free from the interferences, noises and artifacts of the proposed BSS methods are evaluated by comparing the original sources and estimated sources using the performance parameters such as Source to Interference Ratio (SIR), Source to Distortion Ratio (SDR), and Source to Artifact Ratio (SAR).

CHAPTER 5

UNDERDETERMINED CONVOLUTIVE BLIND SOURCE SEPARATION USING CAPSNET

Capsule networks are used to learn the TF masks using SSPs and multi-source active points. The conventional Artificial Neural Networks (ANN) learns a particular feature by adjusting its weights for the given scalar inputs. The important drawback of the conventional ANN is failed to learn the spatial relationship between the features. Whereas, the capsule networks accept the vector as input and produces the vector output based on the presence of the particular feature in the given input.

5.1 INTRODUCTION

This chapter tackles with the problem of BSS from their underdetermined convolutive mixtures using capsule networks. BSS is actually an ill-posed problem and hence it cannot be solved without prior information. The problem is solved based on the mixing model (either instantaneous or convolutive), statistical independence, sparseness property of the sources, the number of sources and microphones (either overdetermined or underdetermined). Recent studies are revealed Deep Neural Networks (DNNs) able to model complex functions and perform well on various applications. The DNNs are utilized for the source separations that act on the magnitude of the Short-Time Fourier Transform (STFT). The DNNs are used to predict the sources spectrogram or to predict a TF mask. The source signals are estimated by multiplying the mixture signals and TF mask.

5.1.1 System Model

The underdetermined convolutive mixing model is described as

$$x_i(n) = \sum_{j=1}^{N} \sum_{k=0}^{L-1} h_{ij}(k) s_j(n-k) \quad i = 1,2,3,...,M \tag{5.1}$$

where M is the number of mixture signals observed from the microphones and N is the number of source signals. L is the length of room reverberations which is typically in the order of seconds. The observed mixture signals are converted into the TF domain using short-time Fourier transform and it is expressed as

$$X_i(t,f) = \sum_{j=1}^{N} H_{ij}(f) S_j(t,f) \qquad i = 1,2,3,...,M \tag{5.2}$$

where t is time frame and f is frequency bin. The disjoint orthogonal property of the source signals are utilized for the generation of the TF mask. But in general, the sources are not perfectly disjoint in the TF domain. For the estimation of the TF masks, it is assumed that the sources have atleast as many single source points as the number of sources and each frequency bin has atleast one single source point belonging to each of the sources. This is a valid assumption and can easily be satisfied by the sources compared to disjoint orthogonal conditions.

5.1.2 Deep Neural Learning

More recently deep learning techniques are used in the source separation. The source separation algorithm using machine learning techniques is inspired by the concept of the TF masking. The TF masking uses a two-dimensional weight to the mixture signals in the TF domain in order to extract the sources. An ideal binary mask represents whether a target

source dominates a TF point in the mixture signals. An ideal binary mask is equivalent to a binary classification of the target sources. In the case of an ideal binary mask, the mask is used as the target function during the training process. Since the formation of speech separation as a classification problem, the supervised speech separation is advanced by using large training data and computing resources. Supervised separation improves its performance by using a deep learning process. Supervised separation performance depends on the following parameters: (i) learning machines used (ii) Features extracted from the available sources (iii) training targets. Speech separation includes the separation of the target source from the interfering speech sources, noises, room reverberations or combinations of these.

Over the last decade, deep learning algorithms improved its performance of supervised speech separation by a considerable margin. The most popular neural network is a MultiLayer Perceptrons (MLPs) that has feed forward connections from the input layer to the output layer with fully connected networks. An MLP is trained by using the backpropagation algorithm where the neuron's weights are aligned so that the prediction error is reduced through gradient descent. The error is measured by the cost function between the target output and actual output. The common cost function is Mean Square Error (MSE).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}_i\right)^2 \tag{5.3}$$

where y_i and $\hat{y}_i$ are the target output and actual output respectively. The performance of the classification increases if the number of layers increases in the MLP. However, it is difficult to train a DNN with multiple hidden layers because of the vanishing gradient problem. Vanishing gradient problems refer to the error vanishes at the lower layer when the prediction error is propagated

from the higher layer (nearer to the output layer) to the lower layer (nearer to the input layer) using backpropagation algorithm. This is the reason for using MLPs with a single hidden layer.

A class of feedforward neural networks is known as Convolutional Neural Networks (CNNs) which is suitable for pattern recognition problems. A typical CNN architecture consists of convolutional layers and subsampling layers. A convolutional layer learns a local feature from its input regardless of the previous layer. Subsampling layers followed by the convolution layers perform local averaging or maximization of the convolution layer output. CNN is well trained by the backpropagation algorithm, even though CNN is a classification of DNN.

Recurrent Neural Networks (RNNs) uses both feedforward and feedback connections between the hidden layers. RNNs consider the whole input samples simultaneously and learn the features accordingly. A speech signal has a strong correlation between the successive frames in the time domain. Therefore, RNNs used to learn the features from the time-domain structure of the speech signals. RNNs is treated as infinite depth DNNs. Hence RNNs trained with backpropagation algorithm may result in vanishing gradient problem. To overcome this problem, Long Short Term Memory (LSTM) is used to smooth the progress of learning over time. A memory cell in LSTM has three gates (i) input gate (ii) forget gate (iii) output gate. An input gate controls the input data to be maintained from the current input and the forget gate maintains amount of data to be retained for the further learning process. These gates are helpful to improve the RNNs training process. From all these overviews, the DNNs is considered as any neural networks with atleast two hidden layers whereas the neural networks with only one hidden layer are called MLPs.

5.1.3 TF Masks

In the speech separation process through neural networks learning, there are two types of targets used (i) masking based targets (ii) mapping based targets. Masking based training uses the time-frequency relations between clean speech signals, interference signals and noises. Mapping based training uses the spectral relationship between clean speech signals, interference signals, and noises. The first training target used in speech separation is called the ideal binary mask. The Ideal Binary Mask (IBM) is defined as

$$IBM(t,f) = \begin{cases} 1, if SNR(t,f) > \gamma \\ 0, otherwise \end{cases} \tag{5.4}$$

where t and f represents time and frequency respectively. The IBM value assigns binary 1 if the SNR of the particular TF point is greater than the threshold γ. There will be as many numbers of IBMs as the target sources.

Ideal Ratio Mask (IRM) uses soft version of the IBM instead of hard decision on each TF points of the corresponding target sources.

$$IRM(t,f) = \left(\frac{S(t,f)^2}{S(t,f)^2 + N(t,f)^2} \right)^{\beta} \tag{5.5}$$

where $S(t, f)^2$ represents energy of the target source and $N(t, f)^2$ represents the energy of other sources including the noises. β is used to scale the mask. However, the signals S(t, f) and N(t, f) should be uncorrelated and hence it is not suitable for convolutive mixture signals.

Spectral Magnitude Mask (SMM) is defined, based on the magnitude of STFT of the target source to other sources.

$$SMM(t,f) = \frac{|S(t,f)|}{|Y(t,f)|} \tag{5.6}$$

where |S(t,f)| and |Y(t,f)| represents the magnitude of STFT of the target source and all other sources respectively. Apart from masking and mapping-based approaches, the recent deep learning techniques map the temporal structures rather than TF representation. Zhao *et al.* (2017) proposed a two-stage DNN where the first stage finds the ratio masking and the second stage performs spectral mapping. Huang *et al.* (2014) introduced DNN for speaker separation. The authors used both DNN and RNN for two speaker separation and a masking layer is added to the network based on the fact that the summation of the spectra of the sources will not be equal to the spectra of the mixture signals.

$$\hat{s}_1(t) = \frac{|\hat{s}_1(t)|}{|\hat{s}_1(t)| + |\hat{s}_2(t)|} Y(t) \tag{5.7}$$

$$\hat{s}_2(t) = \frac{|\hat{s}_2(t)|}{|\hat{s}_1(t)| + |\hat{s}_2(t)|} Y(t) \tag{5.8}$$

where |Y(t)| is the mixture spectrum.

Speaker independent separation is done by unsupervised clustering where TF points are clustered into distinct features dominated by individual speakers. Huang *et al.* (2014) combined DNN based feature learning and spectral clustering for individual speaker separation.

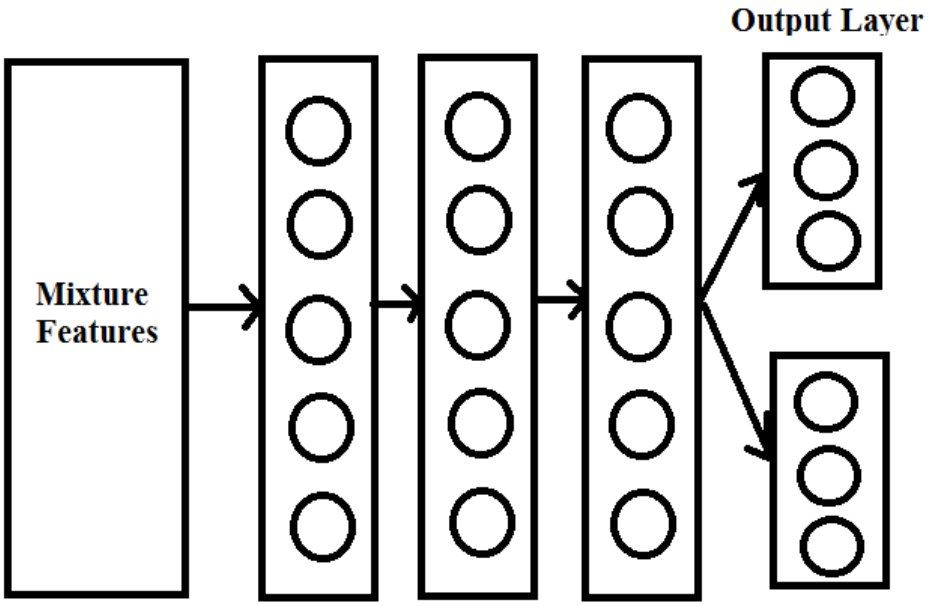


Figure 5.1 Two speaker separation based on DNN

Chapter 2 addresses the problem based on the statistical independence between the sources, for overdetermined mixing signals. Chapter 3 estimates the mixing matrices by increasing the sparse property of the observed mixture signals using single-source points and hence the source signals are estimated using the mixing matrix. This chapter addresses the problem of underdetermined blind source separation by constructing the mask based on the unsupervised learning of the capsule networks.

5.2 CAPSULE NETWORKS

Hinton *et al.* (2018) introduced a completely new type of neural network called a capsule network. In addition to this, Sabour *et al.* (2017) proposed an algorithm called “dynamic routing between capsules” to train the capsule networks. Convolutional neural networks are one of the deep learning techniques used for image classification for a long time. However, Hinton *et al.* (2018) identified an important and fundamental drawback of convolutional neural networks. For CNN, the presence of the objects is

important, but it does not consider the spatial relationship between the objects. Also, it does not take the translational and rotational variance of the objects while classifying the objects. Hence CNN failed to learn the orientation and relative spatial relation between the objects which leads to the misclassification of the image.

Lower layers (nearer to the input layer) learn the simple features such as edges and gradients whereas higher layers (nearer to the output layer) combine all these features into more complex features. The final classification of the image is done based on the complex features. High-level features combine lower level features as a weighted sum of them. There are no orientation features such as rotational and translational features in this setup. CNNs use max-pooling layers to detect higher-order features. Hinton *et al.* (2018) describes max pooling operations as a big mistake in CNNs. Computer graphics construct the visual image based on the internal hierarchical representation of geometric data. The internal representation of the image is stored in the computer memory as an array of geometrical objects and relative position and orientation of these objects. Then some software is used to convert these stored data into visual images. This is called rendering in computer graphics.

Human brains recognize the images by doing the opposite of the rendering and it is called inverse graphics. In order to classify correctly, the images and recognition problem, it is important to learn the hierarchical relationships between the object parts. The capsule networks learn these hierarchical relationships as a pose matrix. Another advantage of the capsule network is that, it requires only a fraction of the data that CNN uses for learning. The algorithm used to train the capsule network is called “dynamic routing between capsules”.

Artificial neurons give a scalar output. Convolutional layers in CNN use kernel's weight across the entire input volume and then the output is produced as a matrix. The matrix replicates the presence of a particular feature across the input volume. The outputs of all kernel's matrices are applied to the max-pooling layer where the largest number in each region is considered for further processing. Hence max-pooling loses valuable information at this point and does not encode the relative spatial relationships between the features. In the capsule networks, the capsules preserve all the information about the feature in the form of vector whereas neurons output scalar value. Capsules encode the probability of the detection of the particular feature for the given input volume as the length of the vector and state of the feature is encoded as the direction of the vector. If the detected feature moves somewhere in the image correspondingly the orientation of the vector changes without changing the length of the vector. This is called activity equivariance. This property is missed in the conventional artificial neural networks and CNN, but incorporated in the capsule networks.

The difference between the capsule and the neuron is that the capsule preserves the features in the vector form whereas the neuron detects the features in the scalar value. Table 5.1 and Figure 5.2 summarize the difference between the capsule and the neuron.

Table 5.1 Comparison between the capsule and the neuron

Input from low level capsule / neuron		Vector (u_i)	Scalar (x_i)
Operation	Affine transform	$\hat{\boldsymbol{u}}_{j\|i} = \boldsymbol{W}_{ij}\boldsymbol{u}_i$	-
	Weighting and sum	$\boldsymbol{s}_j = \sum_i c_{ij}\hat{\boldsymbol{u}}_{j\|i}$	$a_j = \sum_i w_i x_i + b$
	Nonlinear activation	$\boldsymbol{v}_j = \frac{\\|\boldsymbol{s}_j\\|^2}{1+\\|\boldsymbol{s}_j\\|^2}\frac{\boldsymbol{s}_j}{\\|\boldsymbol{s}_j\\|}$	$h_j = f(a_j)$
Output		Vector ($\boldsymbol{v}_j$)	Scalar (h_j)

The input vector of the capsules from the previous layer capsules denotes the probabilities and state of the objects corresponding to lower-level capsules. These input vectors are multiplied by corresponding weight matrices **W** that encode the spatial relationship between the lower level features. After multiplication by these matrices, the position and the probability of the higher- level features can be identified. In the conventional neural networks, the weights are learned by the backpropagation algorithm. But in the capsule networks, the weight matrices are learned using the "dynamic routing between capsules" algorithm. The nonlinear active function ensures the length of the vector is not more than one, but its direction remains the same.

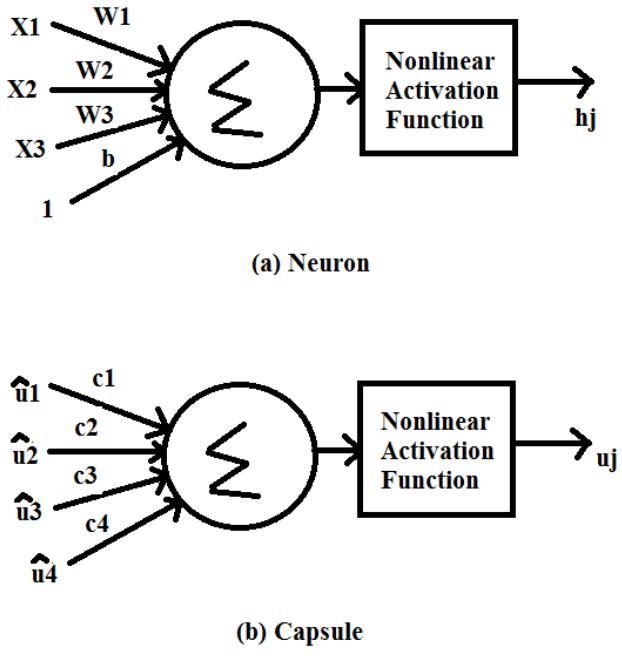


Figure 5.2 Computations in the neuron and in the capsule

5.2.1 Dynamic Routing Between Capsules

A capsule in the lower-level layers needs to send its output to a capsule in the higher-level layers and it is decided by a dynamic routing algorithm. The decision is made by the scalar weights c_{ij}, where c_{ij} represents the scalar weight from the output vector of lower level capsule 'i' to higher-level capsule 'j'. For each lower-level capsule 'i', the sum of all weights to higher-level capsules will be equal to one. These weights are iteratively determined by a dynamic routing algorithm. If the number of iterations is more, there are chances for over learning. Hence the number of iterations should be less. Normally three routing iterations are used in practice.

Originally Sabour *et al.* (2017) used the capsule networks for the classification of handwritten images as shown in Figure 5.3.

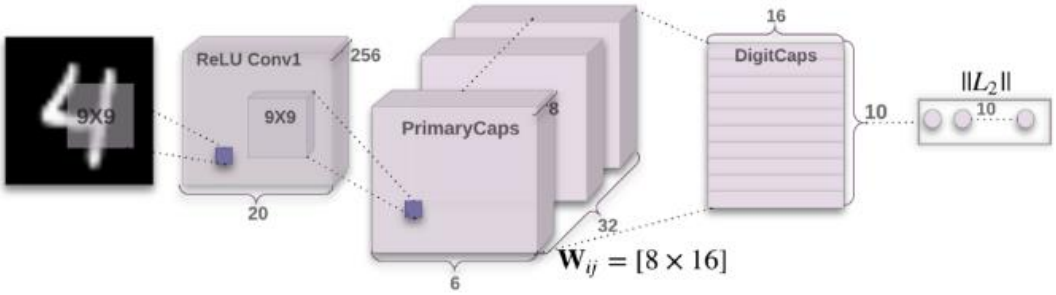


Figure 5.3 Capsule networks architecture for handwritten digit classification Source: Sabour *et al.* (2017)

In the convolution layer, 256 kernels with size 9×9×1 are used, followed by Rectified Linear activation Unit (ReLU). So totally there are 20992 tuneable parameters are used to get the output of 20×20×256 tensors. The output of the convolution layer is fed as the input to the primary capsule layer. There are 32 capsules in this layer. Each capsule uses 9×9×256 kernels to the 20×20×256 input volumes and produces 6×6×8 output tensor. Hence 32 capsules produce the output volume of size 6×6×8×32. There are totally 5308672 parameters in this layer. The output of the primary capsule layer is applied as input to the digit caps. There are ten digit capsules. Each capsule receives the 6×6×8×32 tensor as input. Each of the vector input is mapped into 16-dimensional vectors using 8×16 weight matrices. Finally, the loss function calculates 10 one-hot encoded vectors. The loss is zero if the correct digit cap predicts the correct label with a probability greater than 0.9. Similarly, the loss is zero if the mismatching digit cap predicts incorrect label with probability less than 0.1. Otherwise the loss function will be non-zero either for prediction of the correct label with probability less than 0.9 or prediction of the incorrect label with a probability more than 0.1. Decoder is used to construct the correct handwritten image from the 16-dimensional

output vector of the digit caps. The decoder creates an image of size 28×28 pixels from the 16-dimensional vector inputs.

5.3 PROPOSED METHOD

In this work, the capsule network is used for the separation of the speech signals from its underdetermined convolutive mixture signals. This is the first time that the capsule network is used to separate speech signals. The TF masking based training is followed in this process. The objective of this work is to separate the power dominant source from the interfering speech sources and background noises. The observed mixture signals converted into Time-Frequency domain using STFT. The hamming window of the length 30ms is used for taking STFT. Even though a longer window increases the frequency resolution, the length of the window cannot be increased, since the speech signals are non-stationary. At the same time, a short window may increase the residual noises occurred due to STFT. The TF points of the mixture signals are given as input to the neural networks. The equation is rewritten here for convenience.

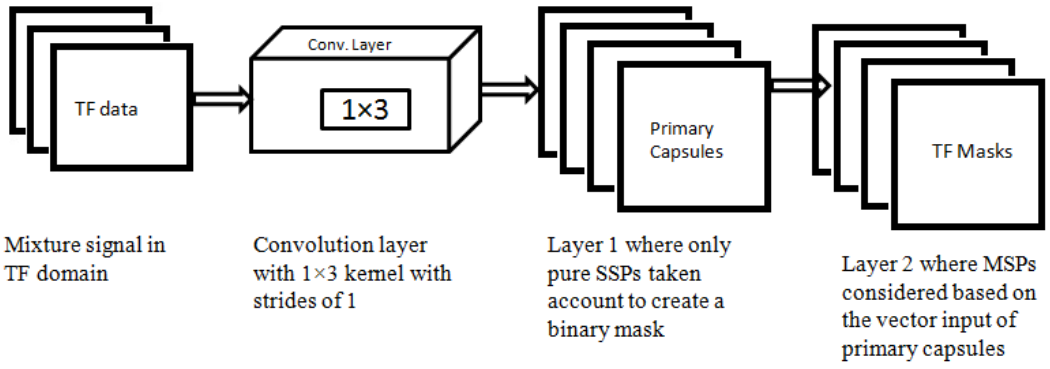


Figure 5.4 Proposed method for underdetermined BSS using CapsNet

$$\boldsymbol{X_i(t, f) = \sum_{j=1}^{N} H_{ij}(f) S_j(t, f)} \qquad i = 1,2,3,...,M \tag{5.9}$$

Without losing the generality of the undetermined mixture signals, it is assumed that there are two mixture signals of three sources. After converting the mixture signals into TF domain, it is represented as

$$\boldsymbol{X_i(t, f) = \sum_{j=1}^{3} H_{ij}(f) S_j(t, f)} \qquad i = 1,2 \tag{5.10}$$

While converting the observed mixture signals into the TF domain, the sparseness property is better utilized by the neural network for source separation. The equation (5.10) can be written in matrix form as

$$\begin{bmatrix} X_1(t,f) \\ X_2(t,f) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ \dfrac{H_{21}(f)}{H_{11}(f)} & \dfrac{H_{22}(f)}{H_{12}(f)} & \dfrac{H_{23}(f)}{H_{13}(f)} \end{bmatrix} \begin{bmatrix} H_{11}(f) S_1(t,f) \\ H_{12}(f) S_2(t,f) \\ H_{13}(f) S_3(t,f) \end{bmatrix} \tag{5.11}$$

Ideal Ratio Mask (IRM) is used to classify the TF points as the single source active points and multi-source active points. One of the observed mixture signals is taken as reference signal and the ratio between the reference signal and all other mixture signals is obtained. At a particular TF point, if only one source i.e., s_1 is active then the observed mixture signals will be

$$X_1(t,f) = H_{11}(f) S_1(t,f) \tag{5.12}$$

$$X_2(t,f) = H_{21}(f) S_1(t,f) \tag{5.13}$$

The ratio between the two mixture signals is given as

$$\frac{X_2(t,f)}{X_1(t,f)} = \frac{H_{21}(f)}{H_{11}(f)} \tag{5.14}$$

Similarly, if source 2 and source 3 are only active, then the ratio is given in equation (5.15) and (5.16).

$$\frac{X_2(t,f)}{X_1(t,f)} = \frac{H_{22}(f)}{H_{12}(f)} \tag{5.15}$$

$$\frac{X_2(t,f)}{X_1(t,f)} = \frac{H_{23}(f)}{H_{13}(f)} \tag{5.16}$$

At a particular frequency bin, there may exist three different groups for three different sources, if the source signals are perfectly sparse. However, the source signals are not perfectly sparse in practice. Here, a set is constructed based on magnitude and phase of the ratio mixtures.

$$\Omega(t,f) = \left\{ \left| \frac{X_2(t,f)}{X_1(t,f)} \right|, angle\left(\frac{X_2(t,f)}{X_1(t,f)} \right) \right\} \tag{5.17}$$

The convolution layer with kernel size 1×3 with stride of one is used to find the single source active points. Figure 5.5 shows the scatter diagram of ratio of mixture signals for perfectly sparse signals.

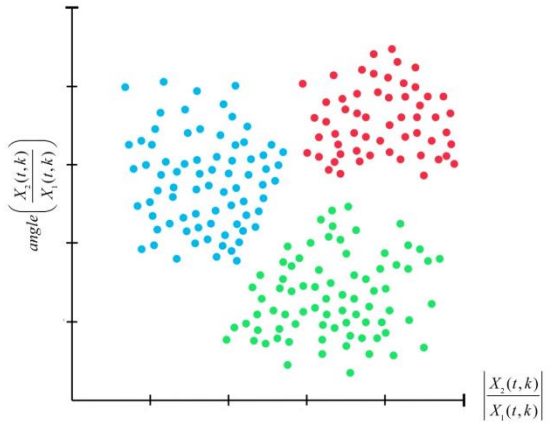


Figure 5.5 Illustration of the ratio of mixtures when sources are perfectly sparse

The size of the set $\Omega(t,f)$ is T×F×N, where T is the number of frames, F is the number of frequency bins in the STFT and N is the number of sources to be separated. The vector of $\Omega(t,f)$ is the input to the primary capsule layer. The lower level capsule (capsule in the primary capsule layer) is used to identify the single source active points based on the input vectors. The lower level capsules calculate the probability of being a particular TF point belongs to any one source. Each of the points in $\Omega(t,f)$ is connected to three capsules with weights c_{ij}. For each lower-level capsule 'i', the sum of all weights to higher-level capsules will be equal to one. These weights are iteratively determined by a dynamic routing algorithm. If the TF point belongs to SSP i.e., only one source is active, then the length of the capsule is one and the remaining capsule lengths are zero. The length of the capsule denotes the probability of the TF point which belongs to only one source.

However in practice, not all the TF points belong to SSP. i.e., more than one source may be active at the same TF point and such TF point is mentioned here as Multi-Source Points (MSP). If the particular TF point belongs to MSP, then the sum of the length of all capsules in the TF point will

be one. Depending upon the length of the capsules, TF masks are constructed for each source. As it is assumed that there are two mixture signals of three sources, three TF masks are constructed by the second layer based on the vector input from the primary capsule layer.

The sources are reconstructed by multiplying the TF mask with any one of the observed mixture signals. As the signals are processed in TF points, the permutation problem occurs while reconstructing the original source signals. Therefore, before multiplying the TF masks with the mixture signals, frequency bins are aligned by using the permutation alignment technique as proposed in chapter 3.

5.4 RESULTS AND DISCUSSIONS

5.4.1 Experimental Setup

The data sets used in our experiments are downloaded from SiSEC 2015 website. The data sets contain synthetic convolutive mixtures as well as live recording mixture signals. The data sets contain two types of mixture signals with reverberation time of 130ms and 250ms. The distance between the two microphones that record the speech signals is 5cm and 1m. Both male and female speech mixture signals are used in our experiments. The recording room dimensions for both synthetic convolutive mixture signals and live recordings are 4.45×3.55×2.5m.

In order to produce synthetic convolutive mixture signals, the room's impulse response is measured and it is synthetically convoluted with individual speech signals. The angle of direction speech signals that are played for recording ranges between -60 and +60 degrees.

5.4.2 Results

Figure 5.6 illustrates the scatter diagram of the output of the primary capsule layer. It clearly shows that there are three sources in the mixture signals, based on the magnitude and phase of the ratio of mixture signals. Further, the remaining TF points are considered as MSP. Based on the position of the TF point, the length of the capsules in the second layer will be computed. The length of the capsule represents how much percentage of the corresponding source contributes to the MSP. The TF mask is finalized, based on the length of the capsule in the second layer. Table 5.2 and Figure 5.7 show the results of the proposed system for the BSS of convolutive mixture signals with the microphone spacing of 5cm.

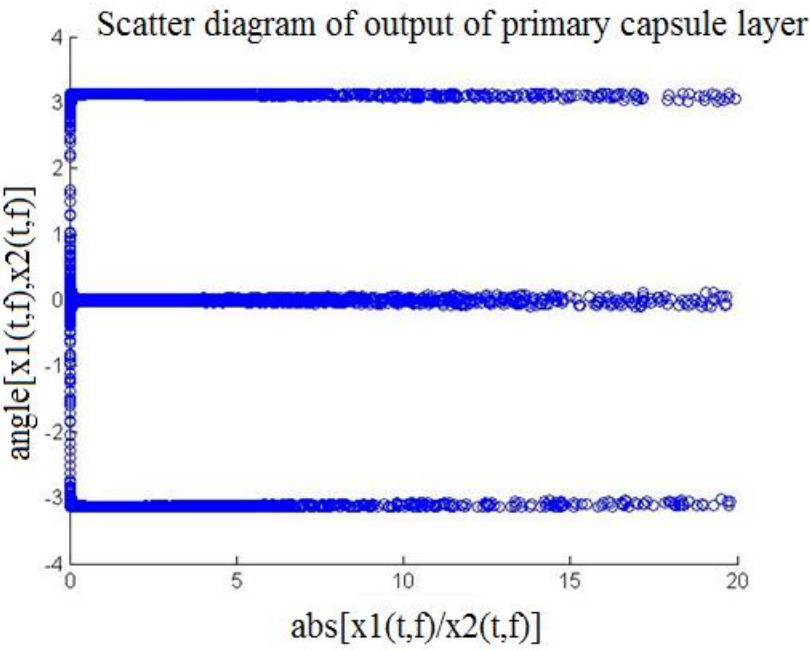


Figure 5.6 Scatter diagram of the output of the primary capsule layer

Table 5.2 Results of the proposed system for synthetic convolutive mixture signals with microphone spacing of 5cm

Types of mixture signals	Source	130ms reverberation			250ms reverberation		
		SIR (dB)	SDR (dB)	SAR (dB)	SIR (dB)	SDR (dB)	SAR (dB)
3 Female speech mixture signal	s1	5.89	5.90	35.26	5.06	5.07	33.71
	s2	4.93	4.93	38.41	3.32	3.32	36.97
	s3	4.80	4.81	36.60	3.25	3.26	34.36
3 Male speech mixture signal	s1	4.27	4.28	33.94	5.89	5.90	33.11
	s2	3.20	3.20	40.527	3.56	3.56	38.92
	s3	4.17	4.17	35.99	4.32	4.33	33.99

Table 5.3 and Figure 5.8 show the results of the proposed system for the BSS of convolutive mixture signals with the microphone spacing of 1m. All these results demonstrate the success of the proposed algorithm in the separation of underdetermined convolutive mixture signals. The SIR and SDR values in all these experiments are approximately equal and confirm that there are no interferences and artifacts in the separated signals. When the microphone spacing is increased, the SIR and SDR of the extracted signals are reduced. But it does not affect the separation process, even though the scaling of the separated signals is reduced.

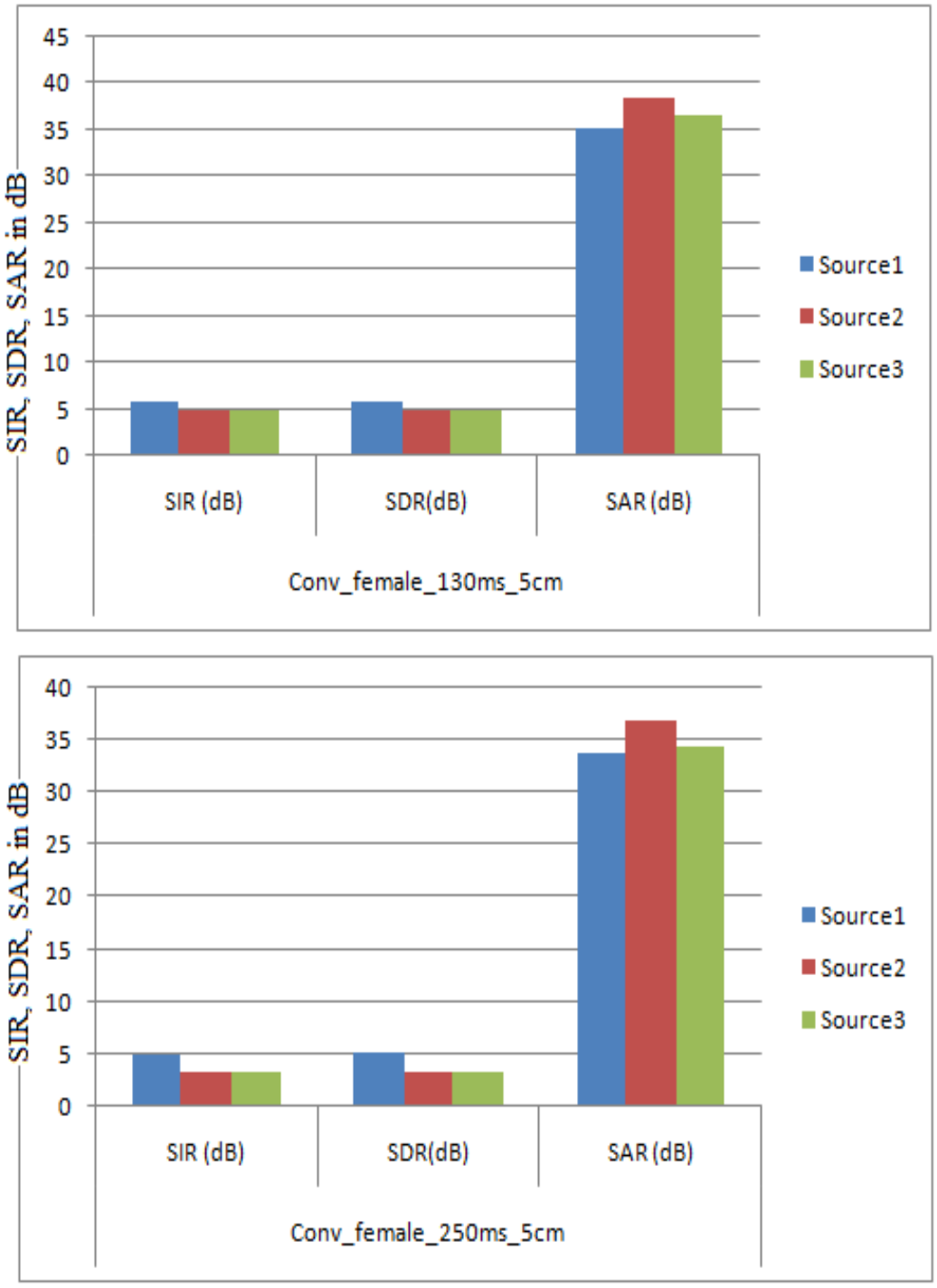


Figure 5.7 Performance of the proposed system for synthetic convolutive mixture signals with microphone spacing of 5cm

Table 5.3 Results of the proposed system for synthetic convolutive mixture signals with microphone spacing of 1m

Types of mixture signals	Source	130ms reverberation			250ms reverberation		
		SIR (dB)	SDR (dB)	SAR (dB)	SIR (dB)	SDR (dB)	SAR (dB)
3 Female speech mixture signal	s1	0.50	0.51	32.09	2.97	2.99	30.49
	s2	6.06	6.16	25.43	6.77	6.92	23.25
	s3	7.46	7.34	25.41	6.36	6.20	23.62
3 Male speech mixture signal	s1	2.41	2.51	23.71	3.73	3.85	23.99
	s2	4.30	4.32	31.69	4.24	4.27	30.82
	s3	6.91	6.75	23.77	6.02	6.18	23.04

Figure 5.9 shows the comparison of the proposed results with the state of art algorithms. The SIR is much greater than the SDR of these algorithms, which implies that there are noises present in the extracted signals, whereas the proposed system separates the sources without interferences and artifacts. When compared with our proposed method in chapter 4 using single source point detection algorithm, the proposed method using capsule network increases the SIR and SDR of the separated signals considerably.

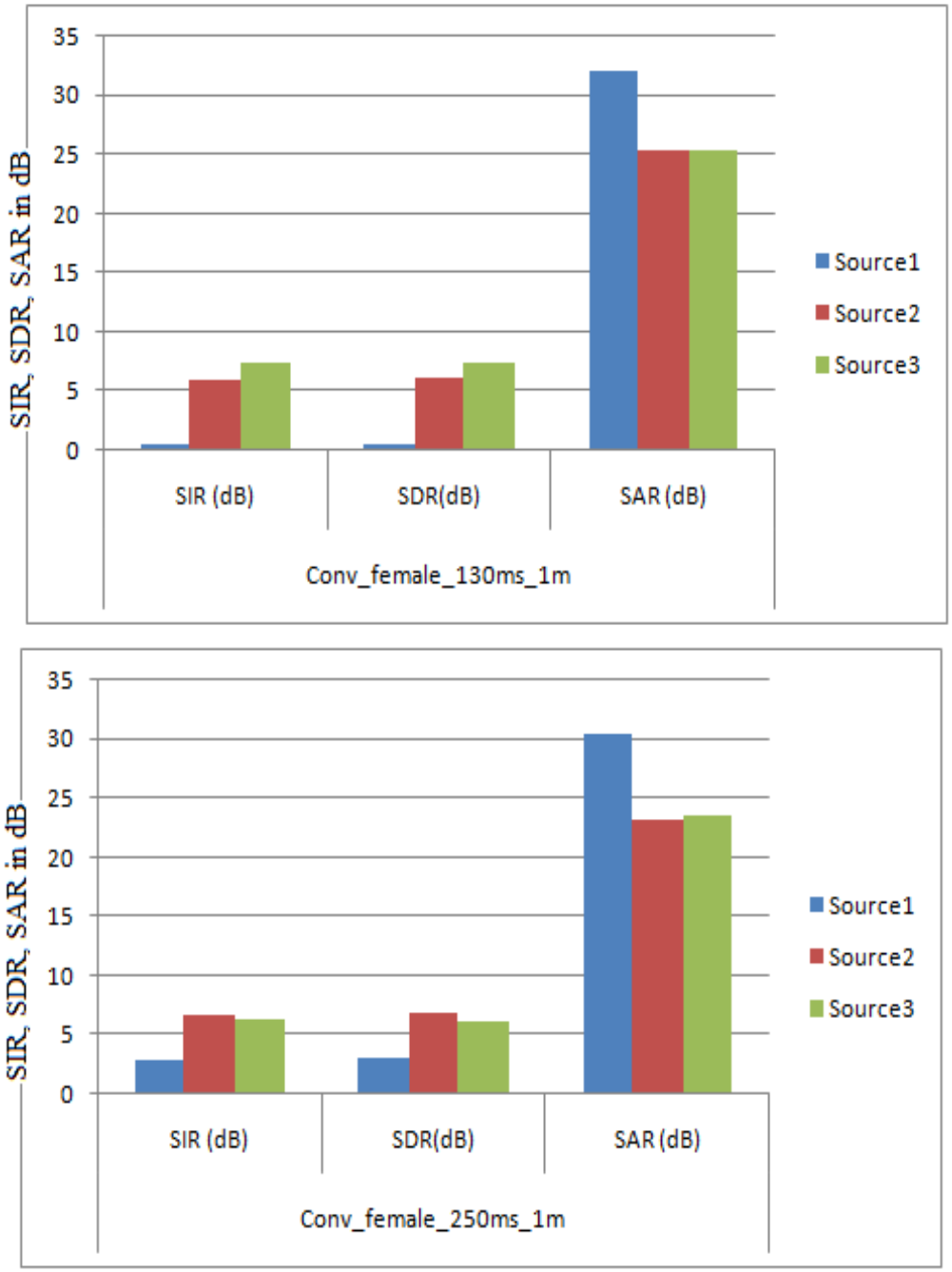


Figure 5.8 Performance of the proposed system for synthetic convolutive mixture signals with microphone spacing of 1m

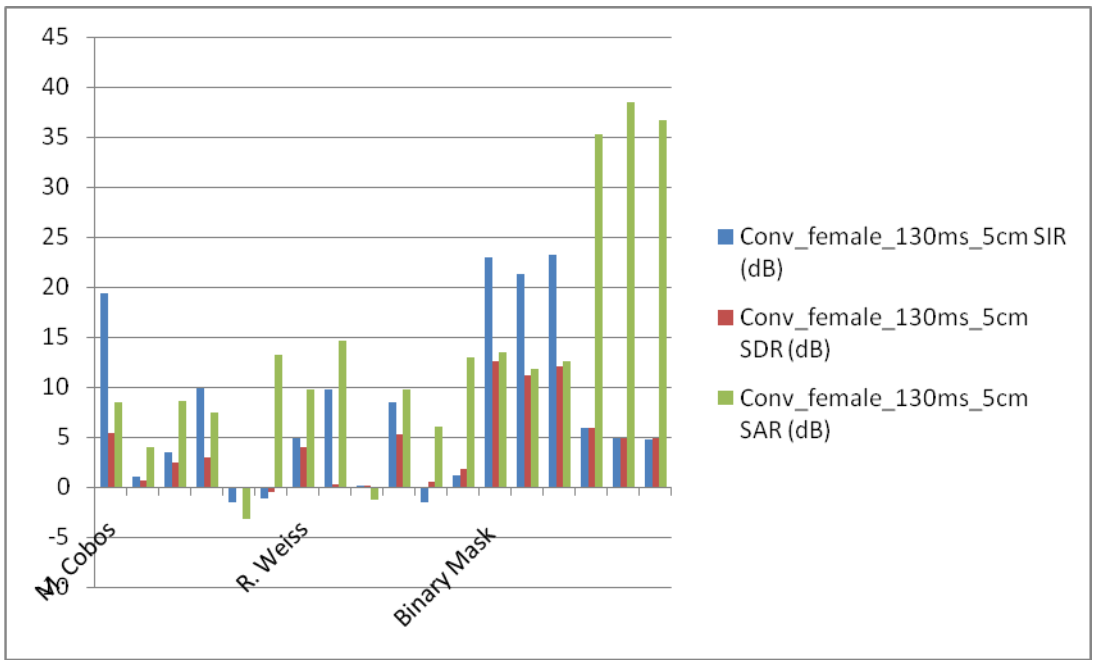


Figure 5.9 Comparison of the proposed system with state of art algorithms

5.5 CONCLUSION

In this chapter, the blind source separation of the underdetermined convolutive mixture signals capsule network is proposed. The capsule networks differ from other artificial neural networks in the sense that the vectors are used as input and the probability of the feature is calculated as the length of the capsules and the spatial relationship is calculated as direction of the vector. The proposed system is tested on the various mixture signals. The various mixture signals are generated by varying the microphone spacing either 5cm or 1m and room reverberations set to either 130ms or 250ms. The results demonstrate the effective separation of the sources from the mixture signals without interferences and artifacts, when compared with state of art algorithms. Furthermore, the SIR and the SDR are increased, when compared with the results of the proposed SSP detection algorithm in chapter 4.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

The research has concentrated on extracting the speech sources from its observed mixture signals in area of blind source separation. Blind source separation is used for a variety of applications at different situations. Therefore, a blind source separation approach common to all applications is practically complex and not yet available. The major challenge in speech application and nature of mixing sources are considered. Depending on the number of microphones, mixing sources used and how the sources are mixed, the BSS solution is divided into overdetermined, determined BSS, underdetermined BSS or instantaneous BSS, convolutive BSS respectively. It is important to note here that one cannot exclude anyone of the above BSS methods based on its performance because the methods are driven by applications. For example, the conventional FastICA algorithm is more suitable for the BSS with overdetermined instantaneous mixtures rather than other algorithms. So this thesis provides solutions to the separation of overdetermined instantaneous mixture, overdetermined convolutive mixture, underdetermined instantaneous mixture and underdetermined convolutive mixture signals.

The conventional FastICA algorithm is proved to be more suitable for the overdetermined instantaneous mixture signals separation. The computational complexity of the algorithm is another challenge. Limitations of this algorithm include: (i) the background noises are also recorded with the mixture signals which affects the overdetermined condition. As noises act as

additional sources the overdetermined mixture becomes underdetermined and hence conventional FastICA cannot be used in practice. (ii) The observed speech mixture signals are not instantaneous in nature but convolutive. Because of these reasons, conventional FastICA algorithm cannot be used for the separation of practical speech mixture signals.

Initially, the solution is proposed for convolutive mixture signals in time domain has convergence issues and stability problems. The convolutive mixture signals can be converted into instantaneous mixture signals by transforming the time domain to the frequency domain. As the speech sources are stationary only for short duration. Short Time Fourier Transform (STFT) is used to convert the mixture signals into the frequency domain. A complex FastICA algorithm is proposed to separate the overdetermined convolutive mixture signals as a complex numbers involve in the frequency domain. The complex FastICA algorithm is applied frequency bin to estimate the sources in the frequency domain. The order of the estimated source components at each frequency bin is random and in general called permutation problem. Hence before converting the estimated sources back into the time domain, the permutation alignment technique is needed to order the estimated source components correctly. A two-pass permutation alignment algorithm is proposed based on the correlation between the power ratios of estimated source components. The estimated sources are compared with the original sources to compute the performance of the algorithm. The results are demonstrated the effectiveness of the proposed algorithm compared to state of art algorithms in terms of Source to Interference Ratio (SIR). The proposed complex FastICA algorithm increases the SIR value from 2 dB to 10dB when compared with the existing algorithms and it separates both the sources equally well. The proposed permutation alignment technique reduces the

misaligned frequency bins from 20% to 16% compared to existing permutation technique.

The background noises present in the mixture signals violate the overdetermined condition and hence underdetermined convolutive mixture signals closely replicates the practical speech mixture signals. Therefore, an algorithm is proposed to estimate the mixing matrix and sources based on the Single Source Point (SSP) detection. The SSP detection algorithm is applied frequency binwise to estimate the mixing matrix due to the varying nature of the frequencies. The estimated mixing matrix is non-square for underdetermined mixtures. Hence the Moore-Penrose pseudo-inverse method is used to invert the mixing matrix and thereby to estimate the sources. Once again the permutation problem occurs while reconstructing the sources is solved by using two-pass permutation alignment algorithm based on the correlation between the power ratios of estimated source components. The results show that the separated sources are free from the interferences and artifacts. The success of the algorithm is proved by comparing it with its equivalent algorithms. The proposed algorithm has equal SIR and SDR values. The difference between the SIR and SDR values for the existing methods are from 10dB to 2dB. The difference is reduced to 0.44 dB in the proposed technique.

A Time-Frequency (TF) mask estimation method based on capsule networks is proposed for the underdetermined convolutive mixture signals. The conventional artificial neural networks in general accept a scalar value as the input and produce a scalar output based on its learning from the inputs. The artificial neural networks failed to learn the spatial relationship between the features. To overcome this drawback a capsule network is used to learn the features as well as their spatial relationships. Capsule networks accept the

vector as the input and produce the vector output. The length of the capsule represents the probability of the existence of the particular feature and the direction of the vector represents the position of the feature. The capsule networks construct the TF mask based on SSP and multi-source points. The results illustrate the effectiveness of the method compared to SSP detection algorithm in terms of increased SIR and Source to Distortion Ratio (SDR). The difference between the SIR and SDR values for the existing methods are from 10dB to 2dB. The difference is reduced to 0.01 dB in the proposed technique. Moreover, the proposed TF mask construction method increases the SIR values up to 1dB when compared with our previous SSP detection algorithm.

6.2 FUTURE WORK

The research can be extended in the following directions:

Single-channel blind source separation is a method of separating sources from only one observed mixture signal. At least two mixture signals are required for separate the undetermined convolutive mixture using the SSP detection algorithm and TF mask construction using the capsule network. So, at least two microphones are needed to record the mixture signals. Only a few works have been reported in the literature for single-channel BSS. Hence the single-channel BSS algorithm may be implemented using supervised learning of the neural networks.

Single source signals can be separated by using semi-supervised learning method when only samples of the one source are available.

Research can be extended to reduce the computational complexity of the proposed SSP detection and TF mask construction algorithm.

REFERENCES

1. Abrard, F & Deville, Y 2005, 'A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources', Signal Processing, vol. 85, pp. 1389 – 1403.

2. Aissa-el-bey, A, Linh-Trung, N, Abed-Meraim, K, Belouchrani, A & Grenier, Y 2007, 'Underdetermined blind source separation of nondisjoint sources in the Time-Frequency domain', IEEE transactions on signal processing, vol. 55, no. 3, pp. 897 – 907.

3. Almeida, M, Schleimer, J. –H, Bioucas-Dias, J & Vigario, R 2011a, 'Source separation and clustering of phase locked subspaces', IEEE transactions on neural networks, vol. 22, no. 9, pp. 1419 – 1434.

4. Almeida, M, Vigario, R & Bioucas-Dias, J 2011b, 'Phase locked matrix factorization', In proceedings of the 19th European Signal Processing Conference (EUSIPCO), Barcelona, Spain, pp. 1728 -1732.

5. Amari, S, Douglas, SC, Cichocki, A & Yang, HH 1997, 'Multichannel blind deconvolution and equalization using the natural gradient', In proceedings of the IEEE workshop on signal processing advances in wireless communications, pp. 101 – 104.

6. Amari, S 1998, 'Natural gradient works efficiently in learning', Neural computation, vol. 10, no. 2, pp. 251 – 276.

7. Amari, S 1999, 'Natural gradient learning for over- and under-complete bases in ICA', Neural computation, vol. 11, pp. 1875 – 1883.

8. Arberet, S, Ozerov, A, Gribonval, R & Bimbot, F 2009, 'Blind spectral-GMM estimation for underdetermined instantaneous audio source separation', In proceedings of ICA.

9. Becker, S & Hinton, GE 1992, 'Self organizing neural network that discovers surfaces in random-dot stereograms', Nature, vol. 355, pp. 161 – 163.

10. Bell, AJ & Sejnowski, TJ 1995, 'An information maximization approach to blind separation and blind deconvolution', Neural computation, vol. 7, no. 6, pp. 1129 – 1159.

11. Belouchrani, A & Amin, MG 1998, 'Blind source separation based on time-frequency signal representations', IEEE transactions on signal processing, vol. 46, no. 11, pp. 2888 – 2897.

12. Bingham, E & Hyvarinen, A 2000a, 'A fast and fixed-point algorithm for independent component analysis of complex valued signals', International journal on Neural Systems, vol. 10, no. 1, pp. 1-8.

13. Bingham, E & Hyvarinen, A 2000b, 'ICA of complex valued signals: A fast and robust deflationary algorithm', In proceedings of the International joint conference on Neural Networks (IJCNN), Como, Italy, vol. 3, pp. 357 – 362.

14. Bofill, P & Zibulevsky, M 2001, 'Underdetermined blind source separation using sparse representations', Signal Processing, vol. 81, pp. 2353 – 2362.

15. Cardoso, JF 1997, 'Infomax and maximum likelihood for blind source separation', IEEE Signal Processing Letters, vol. 4, no. 4, pp. 112-114.

16. Cardoso, JF 1999, 'Higher-order contrasts for independent component analysis', Neural computation, vol. 11, no. 1, pp. 157 – 192.

17. Cardoso, JF, Laheld, BH 1996, 'Equivariant adaptive source separation', IEEE transactions on signal processing, vol. 44, no. 12, pp. 3017 – 3030.

18. Cardoso, JF & Souloumiac, A 1993, 'Blind beamforming for non-Gaussian signals', IEEE proceedings – F, vol. 140, no.6, pp. 362 – 370.

19. Chabreil, G, Kleinsteuber, M, Moreau, E, Shen, H, Tichavsky, P & Yeredor, A 2014, 'Joint Matrices Decompositions and Blind Source Separation: A survey of methods, identification, and applications', IEEE Signal Processing Magazine, vol. 31, no. 3, pp. 34-43.

20. Cho, J, Choi, J & Yoo, CD 2011, 'Underdetermined convolutive blind source separation using a novel mixing matrix estimation and MMSE based source estimation', IEEE international workshop on machine learning for signal processing, Beijing, China.

21. Cichocki, A & Unbehauen, R 1992, 'Neural networks for solving systems of linear equations and related problem', IEEE transactions on circuits and systems I: fundamental theory and applications, vol. 39, no. 2, pp. 124 – 138.

22. Cobos, M & Lopez, JJ 2008, 'Stereo audio source separation based on time-frequency masking and multilevel thresholding', Digital signal processing, vol. 18, pp. 960 – 976.

23. Comon, P, Jutten, C & Herault, J 1991, 'Blind separation of sources, part II: Problems statement', Signal Processing, vol. 24, no. 1, pp. 11-20.

24. Comon, P 1994, 'Independent component analysis – A new concept?', Signal processing, vol. 36, no. 3, pp. 287 – 314.

25. Douglas, SC 2007, 'Fixed-point algorithms for the blind separation of arbitrary complex valued non-Gaussian signal mixtures', EURASIP journal on Advances in signal processing, Article ID 36525, pp. 1-15.

26. Durrieu, JL, David, B & Richard, G 2011, 'A musically motivated mid-level representation for pitch estimation and musical audio source separation', IEEE journal of selected topics in signal processing, vol. 5, no. 6, pp. 1180 – 1191.

27. El Chami, Z, Pham, ADT, Serviere, C & Guerin, A 2008, 'A new model based underdetermined source separation', In proceedings of IWAENC.

28. Esmaeilbeig, M, Sheikhzadeh, H & Razzazi, F 2016, 'A novel and fast algorithm for solving permutation in convolutive BSS based on real and imaginary decomposition', Circuits Systems and Signal processing, vol. 35, pp. 4532 – 4549.

29. Friori, S 2000, 'Blind separation of circularly distributed sources by neural extended APEX algorithm', Neurocomputing, vol. 34, pp. 239 – 252.

30. Friori, S 2003, ‘Extended Hebbian learning for blind separation of complex-valued sources’, IEEE transactions on circuits and systems II, vol. 50, no. 4, pp. 195 – 202.

31. Fu, SW, Wang, TW, Tsao, Y , Lu, X & Kawai, H 2018, ‘End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks’, IEEE/ACM transactions on Audio, Speech, and Language Processing, vol. 26, no. 9, pp. 1570 – 1584.

32. Georgieav, P, Theis, F & Cichocki, A 2005, ‘Sparse component analysis and blind source separation of underdetermined mixtures’, IEEE transactions on neural networks, vol. 16, pp. 992 – 996.

33. Giannakopoulos, X, Karhunen, J & Oja, E 1999, ‘An experimental comparison of neural algorithms for independent component analysis and blind separation’, International journal of neural systems, vol. 9, no. 2, pp. 99 – 114.

34. Grais, E. M, Sen, M. U & Erdogan, H 2014, ‘Deep neural networks for single channel source separation’, In: proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 3762 – 3766.

35. Guo, Q, Ruan, G & Nan, P 2017, ‘Underdetermined mixing matrix estimation algorithm based on single source points’, Circuits, systems and signal processing, vol. 36, pp. 4453 – 4467.

36. Herault, J & Jutten, C 1986, ‘Space or time adaptive signal processing by neural network models’, in Proceedings of the American Institute of Physics Conference Proceedings, New York, vol. 151, pp. 206.

37. Hershey, JR, Chen, Z, Le Roux, J & Watanabe, S 2016, ‘Deep clustering discriminative embeddings for segmentation and separation’, in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 31 – 35.

38. Hinton, G, Deng, L, Yu, D, Dahl, G, Mohamed, A, Jaitly, N, Senior, A, Vanhoucke, V, Nguyen, P, Sainath, T & Kingsbury, B 2012, ‘Deep neural networks for acoustic modeling in speech recognition – four research group share their views’, IEEE signal processing magazine, vol. 29, no. 6, pp. 82 – 97.

39. Hinton, GE, Sabour, S & Frosst, N 2018, 'Matrix capsules with EM routing', in Proceedings of ICLR 2018.

40. Hu, C, Yang, Q, Huang, M & Yan, W 2016, 'Sparse component analysis based under-determined blind source separation for bearing fault feature extraction in wind turbine gearbox', IET Renewable power generation, vol. 11, pp. 330-337.

41. Huang, PS, Kim, M, Hasegawa-Johnson, M & Smaragdis, P 2014, 'Deep learning for monaural speech separation', In: proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1562 – 1566.

42. Hyvarinen, A 1998, 'New approximations of differential entropy for independent component analysis and projection pursuit', In M. I. Jordan, M. J. Kearns & S. A. Solla (Eds), Advances in neural information processing, Cambridge, MA: MIT Press, vol. 10, pp. 273- 279.

43. Hyvarinen, A 1999, 'Fast and robust fixed point algorithm for independent component analysis', IEEE transactions on Neural Networks, vol. 10, no. 3, pp. 626 – 634.

44. Hyvarinen, A, Hopyer, PO & Ink, M 2001, 'Topographic independent component analysis', Neural computation, vol. 13, pp. 1527-1558.

45. Hyvarinen, A, Karhunen, J & Oja, E 2001, 'Independent component analysis', New York, Wiley.

46. Hyvarinen, A & Oja, E 1997, 'A fast fixed point algorithm for independent component analysis', Neural computation, vol. 9, no. 7, pp. 1483 – 1492.

47. Ikram, MZ & Morgan, DR 2005, 'Permutation inconsistency in blind speech separation: Investigations and solutions', IEEE transactions on Speech and Audio processing, vol. 13, no. 1, pp. 1 – 13.

48. Isik, Y, Le Roux, J, Chen, Z, Watanabe, S & Hershey, JR 2016, 'Single channel multi-speaker separation using deep clustering', in Annual conference of the International Speech Communication Association (INTERSPEECH).

49. Jutten, C & Herault, J 1991, 'Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture', Signal Processing, vol. 24, no. 1, pp. 1-10.

50. Karhunen, J & Joutsensalo, J 1993, 'Learning of robust principal component subspace', In proceedings of international conference on neural networks (IJCNN -93), Nagoya, Japan.

51. Kawamoto, M, Matsuoka, K & Ohnishi, N 1998, 'A method of blind separation for convolved nonstationary signal', Neurocomputing, vol. 22, pp. 157 – 171.

52. Kim, SG & Yoo, CD 2009, 'Underdetermined blind source separation based on subspace representation', IEEE transaction on Signal processing, vol. 57, pp. 2604 – 2614.

53. Kim, T 2010, 'Real-time independent vector analysis for convolutive blind source separation', IEEE transactions on circuits and systems I, vol. 57, no. 7, pp. 1431 – 1438.

54. Kitamura, D, Ono, N, Sawada, H, Kameoka, H &Saruwatari, H 2016, 'Determined Blind Source Separation Unifying Independent Vector Analysis and Nonnegative Matrix Factorization', IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 9, pp. 1626-1641.

55. Kolbaek, M, Yu, D, Tan, ZH & Jensen, J 2017, 'Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks' IEEE/ACM transaction on Audio, Speech, and Language Processing, vol. 25, no. 10, pp. 1901 – 1913.

56. Lee, TW, Girolami, M, Bell, AJ & Sejnowski, TJ 2000, 'A unifying information-theoretic framework for independent component analysis', Computers and mathematics with applications, vol. 39, pp. 1 – 21.

57. Li, H & Adali, T 2008, 'A class of complex ICA algorithms based on kurtosis cost function', IEEE transactions on neural networks, vol. 19, no. 3, pp. 408 – 420.

58. Li, Y, Amari, SI, Cichocki, A, Ho, DWC & Xie, S 2006, 'Underdetermined blind source separation based on sparse representation', IEEE transactions on signal processing, vol. 54, no. 2, pp. 423 – 437.

59. Li, H, Shen, YH, Wang, JG & Ren, XS 2011, 'Estimation of the complex-valued mixing matrix by single-source-points detection with less sensors than sources', Transactions on emerging telecommunication technologies, vol. 23, pp. 137 – 147.

60. Li, Y & Wang, D 2007, 'Separation of singing voice from music accompaniment for monaural recordings', IEEE transactions on Audio, Speech, and Language Processing, vol. 15, no. 4, pp. 1475 – 1487.

61. Linsker, R 1992, 'Local synaptic learning rules suffice to maximize mutual information in a linear network', Neural computation, vol. 4, no. 5, pp. 691 – 702.

62. Loesch, B & Yang, B 2008, 'Source number estimation and clustering for underdetermined blind source separation', In International Workshop on Acoustic Echo and Noise Control (IWAENC).

63. Lu, J, Cheng, W & He, D & Zi, Y 2019, 'A novel underdetermined blind source separation method with noise and unknown source number', Journal of sound and vibration, vol. 457, pp. 67 – 91.

64. Lv, Z, Zhang, B Wu, X, Zhang, C & Zhou, B 2017, 'A permutation algorithm based on dynamic time warping in speech frequency-domain blind source separation', Speech communication, vol. 92, pp. 132 – 141.

65. Mallis, D, Sgouros, T & Mitianoudis, N 2017, 'Convolutive audio source separation using robust ICA and intelligent evolving permutation ambiguity solution', Evolving systems, vol. 9, no. 4, pp. 315 – 329.

66. Mendel, JM 1991, 'Tutorial on higher-order statistics (spectra) in signal processing and system theory: Theoretical results and some applications', Proceedings of the IEEE, vol. 79, pp. 278 – 305.

67. Murata, N, Ikeda, S & Ziehe, A 2001, 'An approach to blind source separation based on temporal structure of speech signals', Neurocomputing, vol. 41, pp. 1 – 24.

68. Negro, F, Muceli, S, Castronovo, A M, Holobar, A & Farina, D 2016, 'Multi-channel intramuscular and surface EMG decomposition by convolutive blind source separation', Journal of Neural Engineering, vol. 13, no. 2, pp. 1-17.

69. Nesbit, A, Vincent, E & Plumbley, MD 2009, 'Extension of sparse, adaptive signal decompositions to semi-blind audio source separation', In proceedings of ICA.

70. Nguyen, LT, Belouchrani, A, Abed-Merain, K & Boashash, B 2001, 'Separating more sources than sensors using time-frequency distributions', In proceedings of the international symposium on signal processing and its applications, pp. 583-586.

71. Novey, M & Adali, T 2008a, 'Complex ICA by negentropy maximization', IEEE transactions on neural networks, vol. 19, no. 4, pp. 596 – 609.

72. Novey, M & Adali, T 2008b, 'On extending the complex FastICA algorithm to noncircular sources', IEEE transactions on signal processing, vol. 56, no. 5, pp. 2148 – 2154.

73. Oja, E 1998, 'From neurallearning to independent components', Neurocomputing, vol. 22, no. 1-3, pp. 187-199.

74. Ollila, E 2010, 'The deflation-based FastICA estimator: statistical analysis revisited', IEEE transactions on signal processing, vol. 58, no. 3, pp. 1527 – 1541.

75. Ozerov, A & Fevotte, C 2009, 'Multichannel nonnegative matrix factorization in convolutive mixtures with application to blind source separation', In proceedings of ICASSP (2009).

76. Ozerov, A, Philippe, P, Bimbot, F & Gribonval, R 2007, 'Adaptation of Bayesian models for single channel source separation and its application to voice/music separation in popular songs', IEEE transactions on Audio, Speech, and Language processing, vol. 15, no. 5, pp. 1564 – 1578.

77. Park, HM, Oh, SH & Lee, SY 2006, 'A modified infomax algorithm for blind signal separation', Neurocomputing, vol. 70, pp. 229 – 240.

78. Pearlmutter, Barak A & Lucas C. Parra, 1997, 'Maximum likelihood blind source separation: A context-sensitive generalization of ICA.', Advances in neural information processing systems, pp. 613-619.

79. Peng, D & Xiang, Y 2010, 'Underdetermined blind separation of non-sparse sources using spatial time-frequency distributions', Digital Signal Processing, vol. 20, pp. 581 – 596.

80. Peng, TL, Chen, Y & Liu, ZL 2015, 'A time-frequency domain blind separation method for underdetermined instantaneous mixture', Circuits, systems and signal processing, vol. 34, pp. 3883 – 3895.

81. Pham, DT 2004, 'Fast algorithms for mutual information based independent component analysis', IEEE Transactions on Signal Processing, vol. 52, no. 10, pp. 2690-2700.

82. Raj, B, Smaragdis, P, Shashanka, M & Singh, R 2007, 'Separating a foreground singer from background music', In: Proceedings of International Symposium on Frontiers of Research on Speech and Music.

83. Reju, VG, Koh, SN & Soon, IY 2009, 'An algorithm for mixing matrix estimation in instantaneous blind source separation', Signal Processing, vol. 89, pp. 1762 – 1773.

84. Reju, VG, Koh, SN & Soon, IY 2010, 'Underdetermined convolutive blind source separation via Time-Frequency masking', IEEE transactions on Audio, Speech and Language processing, vol. 18, no. 1, pp. 101 – 116.

85. Rickard, S 2007, 'The DUET blind source separation algorithm', in: Makino, S, Sawada, H, Lee, TW (eds) Blind Speech Separation, Signals and Communication Technology, Springer, Dordrecht.

86. Ristaniemi, T & Joutsensalo, J 2002, 'Advanced ICA based receivers for block fading DS-CDMA channels', Signal Processing, vol. 82, no. 3, pp. 417 – 431.

87. Roth, Z & Baram, Y 1996, 'Multidimensional density shaping by sigmoids', IEEE transactions on neural networks, vol. 7, no. 5, pp. 1291 – 1298.

88. Sabour, S, Frosst, N & Hinton, G.E 2017, 'Dynamic routing between capsule', arXiv: 1710.09829v2.

89. Sadhu, A, Hazra, B & Narasimhan, S 2013, 'Decentralized modal identification of structures using parallel factor decomposition and sparse blind source separation', Mechanical systems and signal processing, vol. 41, pp. 396 – 419.

90. Sahonero-Alvarez, G & Calderon, H 2017, 'A comparison of SOBI, FastICA, JADE and Infomax algorithms', in proceedings of the 8th international multi-conference on complexity, informatics and cybernetics.

91. Saito, S, Oishi, K & Furukawa, T 2015, 'Convolutive blind source separation using an iterative least squares algorithm for non-orthogonal approximate joint diagonalization', IEEE/ACM transactions on Audio, Speech and Language processing, vol. 23, no. 12, pp. 2434 – 2448.

92. Sarmiento, A, Duran-Diaz, I, Cichocki, A & Cruces, S 2015, 'A contrast function based on generalized divergence for solving the permutation problem in convolved speech mixtures', IEEE/ACM transactions on Audio, Speech and Language processing, vol. 23, no. 11, pp. 1713 – 1726.

93. Sawada, H, Araki, S & Makino S 2010, 'Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment', IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 3, pp. 516-527.

94. Sawada, H, Mukai, R, Araki, S & Makino, S 2004, '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, no. 5, pp. 530 – 538.

95. Schobben, L & Sommen, W 2002, 'A frequency domain blind signal separation method based on decorrelation', IEEE transactions on signal processing, vol. 50, no. 8, pp. 1855 – 1865.

96. Schuster, M & Paliwal, KK 1997, 'Bidirectional recurrent neural network', IEEE transactions on signal processing, vol. 45, no. 11, pp. 2673 – 2681.

97. SiSec 2016, Underdetermined-speech and music mixtures. Available from: https://sisec.inria.fr/sisec-2016/2016-underdetermined-speech-and-music-mixtures/. [17 March 2020].

98. Smaragdis, P 1998, 'Blind separation of convolved mixtures in the frequency domain', Neurocomputing, vol. 22, pp. 21 – 34.

99. Sun, J, Li, Y, Wen, J & Yan, S 2016, 'Novel mixing matrix estimation approach in underdetermined blind source separation', Neurocomputing, vol. 173, no. 3, pp. 623 – 632.

100. Thiagarajan, JJ, Karthikeyan, NR & Spanias, A 2013, 'Mixing matrix estimation using discriminative clustering for blind source separation', Digital signal processing, vol. 23, pp. 9 – 18.

101. Toyama, K & Plumbley, MD 2009, 'Using phase linearity in frequency-domain ICA to tackle permutation problem', In proceedings of ICASSP 2009.

102. Vincent, E, Araki, S & Bofill, P 2009a, 'The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation', in proceedings of International Conference on Independent Component Analysis and Signal Separation.

103. Vincent, E, Arberet, S & Gribonval, R 2009b, 'Underdetermined instantaneous audio source separation via local Gaussian modeling', In proceedings of ICA.

104. Vincent, E, Gribonval & Fevotte, C 2006, 'Performance measurement in blind audio source separation' IEEE transactions on Audio, Speech and Language processing, vol. 14, no. 4, pp. 1462 – 1469.

105. Wang, Y & Wang, D 2012, 'Cocktail party processing via structured prediction', In: Advances in Neural Information Processing Systems (NIPS), vol. 25, MIT press, pp. 224 – 232.

106. Wang, Y & Wang, D 2013, 'Towards scaling up classification based speech separation', IEEE transactions on Audio, Speech and Language Processing, vol. 21, no. 7, pp. 1381 – 1390.

107. Xiao, M, Xie, S & Fu, Y 2005, 'A novel approach for underdetermined blind sources separation in frequency domain', ISSN 2005 Advances in neural networks, pp. 484 – 489.

108. Xu, JD, Yu, XC, Hu, D & Zhang, LB 2014, 'A fast mixing matrix estimation method in the wavelet domain', Signal processing, vol. 95, pp. 58- 66.

109. Yang, L, Lv, J & Xiang, Y 2013, 'Underdetermined blind source separation by parallel factor analysis in time-frequency domain', Cognitive Computing, vol. 5, pp. 207 – 214.

110. Yilmaz, O & Rickard, S 2004, 'Blind separation of speech mixtures via time–frequency masking', IEEE Transaction on Signal Processing, vol. 52, no. 7, pp. 1830–1847.

111. Yu, D, Hinton, G, Morgan, N, Chien, JT & Sagayama, S 2012, 'Introduction to the special section on deep learning for speech and language processing', IEEE transactions on Audio, Speech, and Language processing, vol. 20, no. 1, pp. 4 – 6.

112. Yu, D, Kolbaek, M, Tan, ZH & Jensen, J 2017, 'Permutation invariant training of deep models for speaker independent multi-talker speech separation', in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 241 – 245.

113. Zhang, C, Wang, Y & Jing, F 2017, 'Underdetermined blind source separation of synchronous orthogonal frequency hopping signals based on single source point detection', Sensors, vol. 17, no. 9, pp. 2074.

114. Zhao, Y, Wang, ZQ & Wang, DL 2017, 'A two-stage algorithm for noisy and reverberant speech enhancement', in Proceedings of ICASSP, pp. 5580 - 5584.

115. Zhen, L, Peng, D, Yi, Z, Xiang, Y & Chen, P 2017, 'Underdetermined blind source separation using sparse coding', IEEE transactions on neural networks and learning systems, vol. 28, no. 12, pp. 3102 – 3108.

116. Ziegaus, Ch & Lang, EW 2004, 'A neural implementation of the JADE algorithm (nJADE) using higher order neurons', Neurocomputing, vol. 56, pp. 79 – 100.

LIST OF PUBLICATIONS

International Journals

1. **Kumar, M** & Jayanthi, VE 2020, 'Underdetermined blind source separation using CapsNet', Soft Computing, vol. 24, no. 12, pp. 9011 – 9019. https://doi.org/10.1007/s00500-019-04430-4. **(Impact Factor: 2.784).**

2. **Kumar, M** & Jayanthi, VE 2020, 'Blind source separation using kurtosis, negentropy and maximum likelihood functions', International Journal of Speech Technology, vol. 23, no. 1, pp. 13 – 21. https://doi.org/10.1007/s10772-019-09664-z.