STATISTICAL SINGLE CHANNEL SOURCE SEPARATION

Abd Majid Darsono

A thesis submitted to the Newcastle University for the degree of Doctor of Philosophy

School of Electrical and Electronic Engineering

Faculty of Science, Agriculture and Engineering

December 2012

NEWCASTLE UNIVERSITY

SCHOOL OF ELECTRICAL AND ELECTRONIC ENGINEERING

I, Abd Majid Darsono, confirm that this thesis and work presented in it are my own achievement.

I have read and understand the penalties associated with plagiarism.

Signed:

Date:

ABSTRACT

Single channel source separation (SCSS) principally is one of the challenging fields in signal processing and has various significant applications. Unlike conventional SCSS methods which were based on linear instantaneous model, this research sets out to investigate the separation of single channel in two types of mixture which is nonlinear instantaneous mixture and linear convolutive mixture. For the nonlinear SCSS in instantaneous mixture, this research proposes a novel solution based on a two-stage process that consists of a Gaussianization transform which efficiently compensates for the nonlinear distortion follow by a maximum likelihood estimator to perform source separation. For linear SCSS in convolutive mixture, this research proposes new methods based on nonnegative matrix factorization which decomposes a mixture into two-dimensional convolution factor matrices that represent the spectral basis and temporal code. The proposed factorization considers the convolutive mixing in the decomposition by introducing frequency constrained parameters in the model. The method aims to separate the mixture into its constituent spectral-temporal source components while alleviating the effect of convolutive mixing. In addition, family of Itakura-Saito divergence has been developed as a cost function which brings the beneficial property of scale-invariant. Two new statistical techniques are proposed, namely, Expectation-Maximisation (EM) based algorithm framework which maximizes the log-likelihood of a mixed signals, and the maximum a posteriori approach which maximises the joint probability of a mixed signal using multiplicative update rules. To further improve this research work, a novel method that incorporates adaptive sparseness into the solution has been proposed to resolve the ambiguity and hence, improve the algorithm performance. The theoretical foundation of the proposed solutions has been rigorously developed and discussed in details. Results have concretely shown the effectiveness of all the proposed algorithms presented in this thesis in separating the mixed signals in single channel and have outperformed others available methods.

.

ACKNOWLEDGEMENT

Alhamdulillah. Thanks to Allah SWT, whom with His willing giving me the opportunity to complete this thesis.

This thesis is a collection of not only hard work, perseverance and continuous efforts in the past four years, but also encouragement, cooperation and support from many people. I would like to take an opportunity to acknowledge these people.

First and foremost, I would like to express my utmost gratitude to my primary supervisor Dr. Wai Lok Woo and second supervisor Profesor Satnam Dlay for giving me opportunity to pursue a PhD at Newcastle University. My deepest appreciation for my supervisors who nurtures me to become an independent researcher, trained me to critically analyse scientific issues and helped me understand concepts of signal processing. I am fortunate to have them as my supervisors who are never tired in giving me invaluable support, guidance and encouragement.

I would like also to extend my gratefulness to my research colleagues Bin Gao, Imran, and Norhaslinda who really helped me in giving me ideas and constructive suggestions for my research through our discussion.

I am also appreciated very much to my employer, Universiti Teknikal Malaysia Melaka (UTeM) and Ministry of Higher Education of Malaysia for their financial support for my study which made my research possible.

Last but not least, my deepest thankfulness goes to my beloved wife, Izwanni and my daughters, Nurin and Nura for their endless love, understanding, sacrifice, care and support. Thank you very much.

ABBREVIATIONS/ACRONYMS

SS	Source Separation
BSS	Blind Source Separation
SCSS	Single Channel Source Separation
CPP	Cocktail Party Problem
CASA	Computational Auditory Scene Analysis
ICA	Independent Component Analysis
PCA	Principal Component Analysis
ISA	Independent Subspace Analysis
EMD	Empirical mode Decomposition
IMF	Intrinsic Mode Functions
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
FHMM	Factorial Hidden Markov Model
PNL	Post-nonlinear
PDF	Probability Density Function
CDF	Cumulative Density Function
TF	Time-Frequency
FIR	Finite Impulse Response
STFT	Short Time Fourier Transform
IS	Itakura-Saito

LS	Least Square
KL	Kullback-Leibler
GMM	Gaussian Mixture Model
NMF	Nonnegative Matrix Factorization
NMF2D	Two-Dimensional Nonnegative Matrix Factorization
SNMF2D	Sparse Two-Dimensional Nonnegative Matrix Factorization
FCNMF2D	Frequency Constrained Two-Dimensional Nonnegative Matrix Factorization
FC-SNMF2D	Frequency Constrained Sparse Two-Dimensional Nonnegative Matrix Factorization
ML	Maximum Log-likelihood
MAP	Maximum a Posterior
Fro	Frobenius Norm
IBM	Ideal Binary Mask
WDO	Windowed Disjoint Orthogonality
SIR	Signal-to-Interference Ratio
SDR	Signal-to-Distortion Ratio
SAR	Source-to-Artiacts Ratio
PSR	Preserved Signal Ratio

LIST OF CONTENTS

LIST OF PUBLICATIONS

- A.M. Darsono, Bin Gao, W.L. Woo, S.S. Dlay, "Nonlinear single channel source separation", *International Symposium on communications systems, networks and digital signal processing (CSNDSP 2010)*, 2010, pp: 507-511.
- A.M. Darsono, Bin Gao, W.L. Woo, S.S. Dlay, "Sparsity aware machine learning algorithm for single channel convolutive source separation", *Submitted to IEEE Transaction on Systems, Man and Cybernetics, Part B: Cybernetics*.
- A.M. Darsono, Bin Gao, W.L. Woo, S.S. Dlay, "Frequency constrained nonnegative matrix factorization for single channel convolutive mixture", *Submitted to IEEE Transaction on Neural Networks and Learning Systems.*
- A.M. Darsono, Bin Gao, W.L. Woo, S.S. Dlay, " Machine learning algorithm for frequency-constrained nonnegative matrix factorization", *Submitted to 38th International Conference on Acoustics, Speech, and Signal Processing (ICASSP),2013*

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

a	Mixing parameter.
$c_m(t)$	m^{th} intrinsic mode function (IMF) .
$C_{IS}(.)$	Itakura-Saito (IS) divergence cost function.
$C_{KL}(.)$	Kullback-Leibler (KL) divergence cost function.
$C_{LS}(.)$	Least square (LS) distance cost function.
$\mathbf{C}_k$	k^{th} component of the mixture.
e	Noise.
$E[.]$	Expectation.
f	Frequency or frequency bin.
$f(.)$	Nonlinear mixing function.
$F_v(v)$	Cumulative density function (cdf) of v.
$F_z(z)$	Cumulative density function (cdf) of z.
$g(.)$	Inverse of $f(.)$
$\mathbf{H}$	Temporal code matrix.
$\mathbf{H}^{\phi}$	ϕ^{th} slice of temporal code matrix.
j	j^{th} source.
L^{ICA}	Likelihood function of Independent Component Analysis (ICA)
m	Mask.
m^{ICA}	Basis function in ICA
n	Time slots.

n_{iter}	Number of iterations.
N_o	Number of observed signals.
N_s	Number of independent sources.
$p(.)$	Probability density function.
Q_k^{ML}	k^{th} minus hidden-data log likelihood.
$r_M^{EMD}(t)$	Final residue in empirical mode decomposition (EMD).
$\mathbf{s}_j(t)$	Basis coefficient vector of source j^{th} in ICA
S_a	Amplitude similarity.
S_f	Frequency similarity.
t	Time or sample index.
$\underline{\mathbf{u}}$	Mean vector.
$\mathbf{U}$	Frequency constrained of NMF2D
$v(t)$	Single channel of post-nonlinear mixture
$\mathbf{V}_k$	Posterior power of mixture component, $\mathbf{C}_k$.
w^{ISA}	Time-varying weight in independent Subspace Analysis (ISA).
$\mathbf{W}$	Spectral basis matrix.
$\mathbf{W}^{\tau}$	τ^{th} slice of spectral basis matrix.
$x_{f,n}$	Source signals in time-frequency (TF) domain.
$x(t)$	Source signals in time domain.
$\lvert\mathbf{X}_j\rvert^{.2}$	Power spectogram matrix of j^{th} source.
$\lvert\mathbf{X}_j^{im}\rvert^{.2}$	Power spectogram matrix of j^{th} source image.

$y_{f,n}$	Mixture or observation in time-frequency (TF) domain.
$y(t)$	Mixture or observation in time domain.
$\lvert\mathbf{Y}\rvert^2$	Power spectogram matrix of mixture $y(t)$.
$z(t)$	Gaussianized time domain signal
$\mathbf{z}^{ISA}$	Basis vector in independent Subspace Analysis (ISA).

Greek Symbols

$\mathbf{\Lambda}$	Sparsity parameters matrix
λ	Sparsity parameter
λ^{post}	Posterior variance
$\mathbf{\Sigma}$	Covariance matrix.
$\xi(.\lvert.)$	Gamma probability density function
$\xi_m^H(f)$	Cross spectrum of Hilbert spectrum.
$\theta_m^H(f)$	Spectral projection of Hilbert spectrum.
$\boldsymbol{\theta}$	Set of all parameter in estimation using EM.
η	Learning gain or learning rate.
$\vartheta(.)$	Weighted Wiener filters.
$\Phi(.)$	Cumulative density function of Gaussian.
$\varphi(.)$	Gradient ascent adaptation of coefficient density.
μ^{post}	Posterior mean
ψ	Threshold value for ascertaining the convergence.

CHAPTER 1

INTRODUCTION

1.1 Background of Source Separation

Source separation (SS) has received wide attention and has been a topic of investigation for over two decades. SS problem refers to the statistical technique of separating a mixture of underlying source signals. Cocktail party problem (CPP) [1-2] is the classic example of SS problem to address the phenomenon associated with human auditory system that when a number of people are talking simultaneously with the present of background interferences and noise like in cocktail party, humans have the ability to focus their listening attention on a single speaker. During the last decade, many researchers and scientists have attempted to tackle this problem and remarkable developments have been achieved in the area of SS [3-12]. It has become one of the promising and exciting topics with solid theoretical foundations and potential applications in the fields of signal processing, neural computation and advanced statistics. SS has been successfully applied in various fields, such as speech enhancement, biomedical image processing, image processing, remote sensing, communication systems, exploration seismology, geophysics, econometrics, data mining and neural networks. Despite all these applications, so far there are no machines produced that can perform SS in the manner of human listening. It remains an open problem and demands further research investigation.

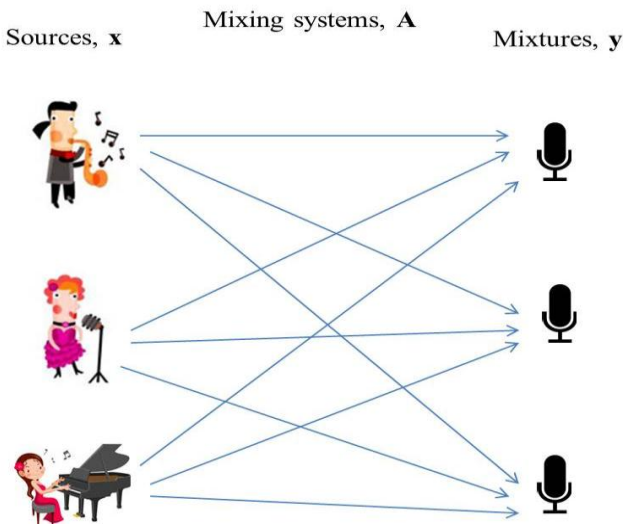


Figure 1.1: A simplified scenario of the source separation problem with three audio sources and three microphones

1.1.1 Source separation problem formulation

A general SS problem can be illustrated as in Figure 1.1 which is mathematically defined according to the conventional linear model as follows: A set of observations $\mathbf{y}(t) = \left[y_1(t)\, y_2(t) \cdots y_{N_o}(t) \right]^{\mathrm{T}}$ which are random processes is generated as a mixture of underlying source signals $\mathbf{x}(t) = \left[x_1(t)\, x_1(t) \cdots x_{N_s}(t) \right]^{\mathrm{T}}$ according to:

$$\begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_{N_o}(t) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N_s} \\ a_{21} & a_{22} & \cdots & a_{2N_s} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N_o1} & a_{N_o2} & \cdots & a_{N_oN_s} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_{N_s}(t) \end{bmatrix} \Leftrightarrow \mathbf{y}(t) = \mathbf{A}\mathbf{x}(t) \tag{1.1}$$

where **A** is the unknown mixing matrix of dimension $N_o \times N_s$ and t is the time or sample index. The technique of SS aims to estimate both the original sources $\mathbf{x}(t) = \left[x_1(t)\, x_1(t) \cdots x_{N_s}(t) \right]^{\mathrm{T}}$ and the mixing matrix **A** using only the observations $\mathbf{y}(t) = \left[y_1(t)\, y_2(t) \cdots y_{N_o}(t) \right]^{\mathrm{T}}$. It is noted that (1.1) represents a simplified model which may not be true representation of the real environment. In order to represent a realistic model, issues such as propagation delay of signals, nonlinear distortions, and noise should be taken into account and evaluated. Hence the need for further research has led to various branches of research in SS.

1.1.2 Classification of source separation

A literature review of current reports shows that there exist three main classifications of SS. These include: Linear and Nonlinear SS; Instantaneous and Convolutive SS; Overdetermined and Underdetermined SS.

1.1.2.1 Linear and Nonlinear SS

Linear algorithms were a popular choice and dominate the SS research field because of its simplicity in analysis and its explicit separability. As defined in (1.1), linear SS assumes that the mixture is represented by a linear combination of source signals [3], [13]. Nevertheless, nonlinearity occurs in practical problems which leads to the issue of model mismatch [14-18] and this consequently led to the emergence of

models based on nonlinear SS. This model represent more accurate model of a realistic environment by taking nonlinear distorted signals into consideration.

A general nonlinear SS model can be expressed as:

$$\mathbf{y}(t) = \mathbf{f}\left(\mathbf{x}(t)\right) \tag{1.2}$$

where $\mathbf{y}(t)$ and $\mathbf{x}(t)$ are defined in (1.1) and $\mathbf{f}(.)$ is the nonlinear mixing process. The nonlinear SS problem is more complicated than the linear SS problem because the principle of linear superposition of the source signals is violated. Under general nonlinear condition statistical independence is not sufficiently strong to recover the sources without any distortions and there always exist infinite solutions in nonlinear SS problems if the nonlinear mixing function $\mathbf{f}(.)$ is not constrained [19]. Therefore, some form of constraints need to be imposed and nonlinear mixing models with constraints are generally preferred over a general model. An example of a constrained nonlinear model is the post-nonlinear model [20-22] which can be expressed as

$$y_i(t) = f_i\left(\sum_{j=1}^{N_s} a_{ij} x_j(t)\right) \tag{1.3}$$

where $f_i(.)$ is a nonlinear distortion function and $\sum_{j=1}^{N_s} a_{ij} x_j(t)$ is a linear mixing process.

1.1.2.2 Instantaneous and Convolutive SS

The convolutive mixture occurs when observed signals consist of combinations of multiple time-delayed versions of the original source signals and/or mixed signals themselves. Without the presents of time delays results in the instantaneous mixture

of observed signals which is expressed in (1.1). In linear convolutive mixture models, each element of the mixing matrix **A** in (1.1) becomes a filter instead of a scalar, which results in the observation signals represented by a linear combination of a set of time-delayed source signals, expressed as

$$\begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_{N_o}(t) \end{bmatrix} = \begin{bmatrix} a_{11,\rho} & a_{12,\rho} & \cdots & a_{1N_s,\rho} \\ a_{21,\rho} & a_{22,\rho} & \cdots & a_{2N_s,\rho} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N_o1,\rho} & a_{N_o2,\rho} & \cdots & a_{N_oN_s,\rho} \end{bmatrix} \begin{bmatrix} x_1(t-\rho) \\ x_2(t-\rho) \\ \vdots \\ x_{N_s}(t-\rho) \end{bmatrix} \Leftrightarrow \mathbf{y}(t) = \sum_{\rho=0}^{L-1} \mathbf{A}(\rho)\mathbf{x}(t-\rho) \quad (1.4)$$

where $\mathbf{A}(\rho)$ is the finite-impulse response (FIR) of some causal filters and L-1 is the length of the filter. From (1.4), it can be seen that the simplest conventional linear instantaneous SS model can actually be considered as a special case of linear convolutive SS when $\mathbf{A}(\rho)$, $0 \le \rho \le L-1$ is reduced to **A**.

1.1.2.3 Overdetermined and Undetermined SS

Overdetermined SS refers to the separation problem when the number of observed signals are more than the number of independent sources $(N_o > N_s)$. On the other hand, when the number of observed signals is smaller than the number of independent sources $(N_o < N_s)$, this becomes underdetermined SS. Overdetermined problem can be easily solved by reducing to determined SS as in (1.1). As for the underdetermined BSS [23-25], the separation quality in terms of interference suppression and signal distortion is still not as good as with determined SS. This is

particularly true if wideband signals such as speech signals are involved. The difficulty is that in contrast to determined SS the solution of underdetermined BSS goes beyond system identification. Even if the mixing system is fully identified, additional effort is required to separate the mixtures. The problem is even more challenging if only one channel is available where we need to imposed constraints such as sparsity and non-negativity on the observed signal in order to perform separation.

1.1.3 Applications of SS

Source separation research has attracted a substantial amount of attention in both the academic field as well as the industry area due to the diverse promising and exciting applications. Tremendous developments have been achieved in the application of SS, particularly in audio processing, wireless communication, medical signal processing, geophysical exploration and image enhancement/ recognition [26-41]. In audio processing, SS problem refers to cocktail party problem where the voice or sound is extracted from the recorded mixture of several microphones. Similar example of SS problem in the field of radio communication where it involve the observations which correspond to the outputs of several antenna elements in response to several transmitters which represents the original signals. In the analysis of medical signals, Electroenchephalogram (EEG), Magnetoencephalogram (MEG) and Electrocardiogram (ECG) [5, 13, 26, 28] data represents the observations and SS is used as a signal processing tool to assist noninvasive medical diagnosis. For example in chemometrices application such as in [13], SS has been applied to determine the

spectra and concentration profiles of chemical components in an unresolved mixture. SS has also been applied to the data analysis in other areas such as finance, seismology and telecommunications. It has even been conjectured that SS will have a role in the analysis of the cosmic microwave background [42], potentially helping to elucidate the very origins of the universe. Further evidence of the SS applications can be found in [29-41].

1.1.4 Single channel source separation

This thesis focuses on the special case of underdetermined source separation (SS) problem when only one observation is available called single channel source separation (SCSS). For many practical applications such as audio scenarios, generally only one channel recording is available in the hardware and in such cases conventional source separation techniques are not appropriate. This is the most exciting case seen from hearing instrument industry point of view such that the specific applications [43] are described as follow:

1. It is often desirable to process a single instrument in a recording. For example, in a single microphone recording of vocals and acoustic guitar, we might want to adjust the volume of the guitar or shift the pitch of the vocals. Thus, if the individual instruments can be distinguished in a mixture, they can be processed individually.
2. Speech recognition in the presence of noise, particularly heavy non-stationary noise, is a challenging problem. Speech recognition performance could improve if

the speech can be distinguished from the noise and perform recognition on the portion of the mixture that corresponds to speech.

3. It has been observed that people with the perceptive hearing loss suffer from insufficient speech intelligibility. It is difficult for them to pick up the target speech, in particular, when there exist some interfering sounds nearby. However, amplification of the signal is not sufficient to increase the intelligibility of the target speech as all signals (both target and interference) are amplified. For this application scenario, SCSS is highly desirable to produce a clean target speech to these hearing impaired people.
4. Musicians often spend large amounts of time trying to listen to a song and learn the part of a specific instrument by ear. This task becomes more difficult when the given piece of music has numerous instruments (which if often the case). If the instrument of interest can be extracted, it could simplify the task of the musician. In practice, this is common problem for guitar players that try to learn their parts from recordings of bands.
5. Automatic music transcription of polyphonic music is a challenging problem. If each of the instruments in the mixture can be modeled, they can be transcript individually.
6. A number of music information retrieval (MIR) tasks involve extracting information from individual sources. For example, guitar and piano parts could be good indicators of the key of the song. However, the percussion part will rarely have any useful information for this task. Although, the sound mixture can be directly used for many of these tasks, extracting the information from the right source could improve the performance.

Other field such as neuroscience (spike sorting) [44, 45] seeks to elucidate concerns the mechanisms used by dedicated parts of brains that perform specific tasks can also be performed by using a single channel. This leads to the SCSS research area where the problem can be treated as one observed signal mixed with several unknown sources as shown in Figure 1.2. Important inspiration can be taken from the human auditory system, which possesses a powerful ability to segregate and separate incoming sounds.

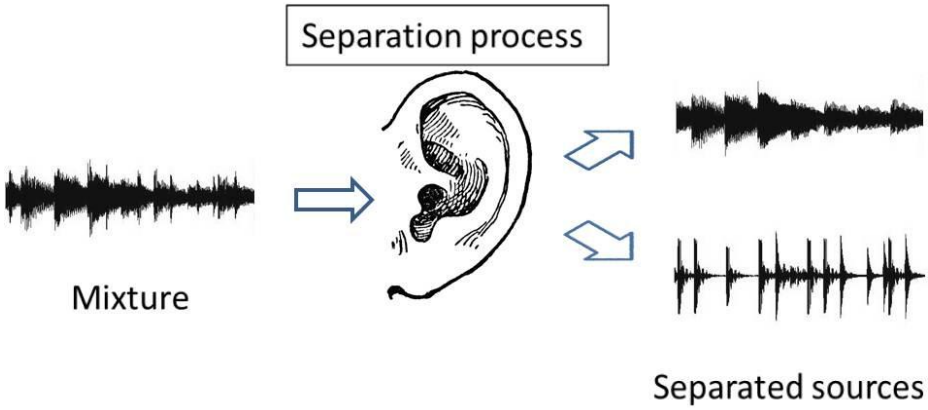


Figure 1.2: Single channel source separation problem where a single mixture of multiple sources is separated into their components.

For linear instantaneous mixing in time domain, the single channel mixture, *y* of the sources, *x* can be model as

$$y(t) = \sum_{j=1}^{J} x_j(t) \tag{1.5}$$

where $j = 1,...,J$ denotes number of sources and the goal is to estimate the sources $x_j(t)$ when only the observation signal $y(t)$ is available. SCSS is an underdetermined problem and for some cases its solution requires additional information about the

sources. For example, it is evident that in the case of linear instantaneous with two sources $\mathbf{x}_1 = \bar{\mathbf{x}}$ and $\mathbf{x}_2 = \mathbf{y} - \bar{\mathbf{x}}$ is a solution for any $\bar{\mathbf{x}}$, and it is necessary to use additional information about the sources to constrain the problem.

In the time-frequency domain (TF), the mixture (1.5) can be expressed as,

$$y_{f,n} = \sum_{j=1}^{J} x_{j,f,n} \tag{1.6}$$

where $y_{f,n}$ and $x_{j,f,n}$ denote TF components which can be obtained by applying short time Fourier transform (STFT). Here, the time slots are given by $n = 1,2,...,N$ while frequencies are given by $f = 1,2,...,F$. Note that in (1.6), each component is a function of *n* and *f* variables and as such, the power spectrogram is defined as the squared magnitude of (1.6):

$$\left|y_{f,n}\right|^2 = \sum_{j=1}^{N_s}\sum_{k=1}^{K} x_{j,f,n} x_{k,f,n}^{*} + 2\sum_{j=1}^{J} \mathrm{Re}\left[x_{j,f,n}\right] \tag{1.7}$$

Assuming the windowed disjoint orthogonality (WDO) of the sources i.e. $x_{j,f,n} x_{k,f,n}^{*} = 0$ for all f and n with $j \neq k$, (1.7) can be written as

$$\left|y_{f,n}\right|^2 = \sum_{j=1}^{J} \left|x_{j,f,n}\right|^2 \tag{1.8}$$

Thus a matrix representation for (1.8) is given as follows:

$$\left|\mathbf{Y}\right|^{.2} = \sum_{j=1}^{J} \left|\mathbf{X}\right|_j^{.2} \tag{1.9}$$

where $\left|\mathbf{Y}\right|^{.2} = \left[\left|y_{f,n}\right|^2\right]_{n=1,2,...,N}^{f=1,2,...,F}$ and $\left|\mathbf{X}\right|_j^{.2} = \left[\left|x_{j,f,n}\right|^2\right]_{n=1,2,...,N}^{f=1,2,...,F}$ are two-dimensional matrices (row and column vectors represent the time and slots and frequencies or frequency bins respectively) which denotes the power TF representation of (1.6). The superscript

"$\cdot$" is element-wise operation. A common choice of calculating the TF is by using STFT but there is also alternative option for computing TF by employing a scale which has a high resolution at lower frequencies and a low resolution at higher frequencies, e.g., constant-Q transform, gammatone filterbank, or a mel scale.

1.2 Objectives of Thesis

The aims of this thesis are to investigate SCSS methods in terms of its fundamental theory, assumptions, applications and limitations as well as further develop new frameworks of single channel source separation (SCSS). In addition, three novel algorithms, one tailored specifically for SCSS of post-nonlinear instantaneous mixture and another two algorithms for linear convolutive mixture have been proposed. Rigorous mathematical derivations and simulations are carried out to validate the effectiveness of the proposed algorithms. The objectives of this thesis are listed as follows:

i.) To present a unified perspective of the widely used existing SCSS methods. The theoretical aspects of SCSS are presented to provide sufficient background knowledge relevant to the thesis.

ii.) To find useful signal analysis algorithms that have desirable properties unique to SCSS problem and how these properties can be advantaged to relax the constraints posed by the problem.

iii.) To develop a new SCSS algorithm for post-nonlinear instantaneous mixture which addresses the following:

- Non-stationary and temporal correlation of the source signals.

- Formulation of an iterative learning process to update model parameters and estimate source signals.
- Estimation of nonlinear distortions by Gaussianization technique.
- Delivery of effectiveness performance by the separation algorithm in various mixture conditions.

iv.) To develop novel methods for SCSS in linear convolutive mixture which addresses the following:

- Non-stationary, spectral coherence and temporal correlation of the audio signals.
- Formulation of an iterative learning process to update model parameters and estimate source signals.
- Delivery of enhanced accuracy and evidence in the form of comparisons to existing counterpart algorithms based on synthesized and real audio signals.

1.3 Thesis Outline

This research is carried out with the focus predominantly on single channel in post-nonlinear instantaneous mixtures and linear convolutive mixture. Three novel generative methods for SCSS will be proposed in this thesis. Real time testing will be conducted and the results should give superior performance over other existing approaches.

In Chapter 2, an overview of recent SCSS methods is given, which is a major class of SCSS methods. The start of this chapter is by introducing SCSS general frameworks. Main current SCSS methods namely model-based SCSS, independent

subspace analysis (ISA), Empirical mode decompositions (EMD), computational auditory scene analysis (CASA) and nonnegative matrix factorization (NMF) have been reviewed in this chapter.

In Chapter 3, a new SCSS method is developed to separate source in post-nonlinear instantaneous mixture. The proposed model is a linear mixture of the independent sources followed by an element-wise post-nonlinear distortion function. In addition, the chapter develops a novel solution that efficiently compensates for the nonlinear distortion and performs source separation. The proposed solution is a two-stage process that consists of a Gaussianization transform and a maximum likelihood estimator for the sources. Simulations have been carried out to verify the theory and evaluate the performance of the proposed algorithm.

In Chapter 4, a new SCSS method is proposed for linear convolutive mixture. Novel matrix factorization algorithms are proposed to decompose an information-bearing matrix (TF representation of mixture) into two-dimensional convolution of factor matrices that represent the spectral basis and temporal code of the sources. In the proposed methods, frequency constraint is imposed onto the model in order to compensate for the distortion caused by the convolutive mixing. In addition, a family of Itakura-Saito (IS) divergence has been derived for estimation of the sources. Two new algorithms have been proposed where the first algorithm method based on expectation maximisation (EM) algorithm framework which maximising the log-likelihood of a mixed signals. As for the second algorithm, it is based on the maximum a posteriori (MAP) approach which maximises the joint probability of the mixing channel, spectral basis and temporal codes conditioned on the mixed signal using multiplicative update rules. Simulation of feature extraction and audio source

separation application have been carried out to investigate the effectiveness of the proposed method

In Chapter 5, a new SCSS method is developed to separate a linear audio convolutive mixture where the proposed method was developed to take into account the reverberation environment in real audio application. The proposed method further improved the frequency constraint two-dimensional nonnegative matrix factorization by imposing sparsity constraint to solve the ambiguity problem in matrix decomposition. An adaptive sparseness approach is derived to compute the sparsity parameter for optimising the matrix factorization. Several comparisons and simulation are carried out to investigate the accomplishment of the proposed method.

This thesis is concluded in Chapter 6. This chapter presents the closing remarks as well as future possibilities for research.

1.4 Thesis Contributions

The SCSS problem has been continually discussed and many approaches have been proposed by researchers. However, these approaches assumed that the mixture is linear instantaneous mixture and therefore are limited by restrictive assumptions which are against realistic situations. The contribution of this thesis is to generate novel solutions for SCSS in two types of mixtures which are post-nonlinear instantaneous mixture and linear convolutive mixture. Hence, the proposed methods overcome the limitations associated with the conventional approaches. This thesis presents three novel methods with a significant improvement in performance in terms of both accuracy and versatility. The following outlines the contribution of this thesis:

i.) A unified view for the existing SCSS methods based on linear mixing model.

ii.) A new framework for SCSS of post-nonlinear instantaneous mixture using two-stage approach has been derived based on underdetermined independent component analysis (ICA) approach. The proposed model is a linear mixture of the independent sources followed by an element-wise post-nonlinear distortion function using Gaussianization transform. The proposed Gaussianization transform offers a simple yet an effective solution in linearising the nonlinearity. In separation stage, the proposed algorithm used basis adapted by ICA learning rules and best characteristic features are incorporated to find a sparse solutions.

iii.) A novel frequency constrained two-dimensional nonnegative matrix factorization (FCNMF2D) for SCSS in convolutive mixture is proposed with the following features.

- Most of the SCSS approaches have been proposed assume that the original sources is instantaneously mixed, which is not realistic in a real application. In audio application for example, the sound or speech signals received by microphone/receiver are exposed to the reverberations in a room which will degrade quality and characteristics of sound. Therefore, the assumption that the mixture is instantaneous adopted by the existing SCSS approaches is violated. Our research in this thesis is to remedy these drawbacks and the formulation of a SCSS model that accounts for convolutive mixing is proposed.
- The proposed model allows over-complete representation by allowing many spectral and temporal shifts which are not inherent in the nonnegative matrix factorization (NMF).

- A family of Itakura-Saito (IS) divergence has been developed to extract the spectral basis and temporal code. The proposed factorization is scale invariant whereby the lower energy components in the TF representation can be treated with equal importance as the higher energy components. Within the context of SCSS, this property enables the spectral-temporal features of the sources that are characterised by a large dynamic range to be estimated with higher accuracy. This is to be compared with the conventional matrix factorization based on Least Square (LS) distance and Kullback-Leibler (KL) divergence where both methods favor the high energy components but neglect the low energy components.
- Two new algorithms introduced using Quasi Expectation-Maximisation (Quasi-EM) and multiplicative update (MU) method. These algorithms will respectively be termed as Quasi-EM FCNMF2D and MU FCNMF2D. The Quasi-EM FCNMF2D algorithm provides reliability in the solution where it avoids zeros in the factors which avoid the solution to be trapped at local minima. On the other hand, the MU FCNMF2D algorithm does not a priori exclude the zero coefficients in the factors. Since zero coefficients are invariant under MU FCNMF2D, if the MU FCNMF2D algorithm attains a fixed point solution with zero entries, then it cannot be determined since the limit point is a stationary point.

iv.) A novel framework to solve SCSS for convolutive mixture based on FCNMF2D with sparsity awareness is proposed. The proposed method extends the MU FCNMF2D algorithm by imposing the sparseness constraints in the model. Imposing sparseness is necessary to give unique and realistic representations of

the non-stationary signals such as an audio signal. Unlike the conventional sparsity constraint solutions in two-dimensional sparse NMF (SNMF2D) where the sparsity parameter is set manually, the proposed model imposes sparseness on temporal code **H** element-wise so that each individual code has its own distribution. Therefore, the sparsity parameter can be individually optimised for each code which will overcome the problem of under-sparse or over-sparse factorization. In addition, each sparsity parameter in the proposed model is learnt and adapted as part of the matrix factorization.

CHAPTER 2

LITERATURE OF SINGLE CHANNEL SOURCE SEPARATION

For the last decade, many approaches have been developed in solving SCSS problem and given the nature of its problem which is inherently underdetermined; its solution relies on making appropriate assumptions concerning the sources and can be categorised either as supervised or unsupervised. These SCSS methods whether they are supervised or unsupervised, have been proved to produce clear defined separability and provide practical applications for data processing especially for linear instantaneous case. In this chapter, existing learning approaches for SCSS are reviewed as well as the relationships among these approaches will be discussed.

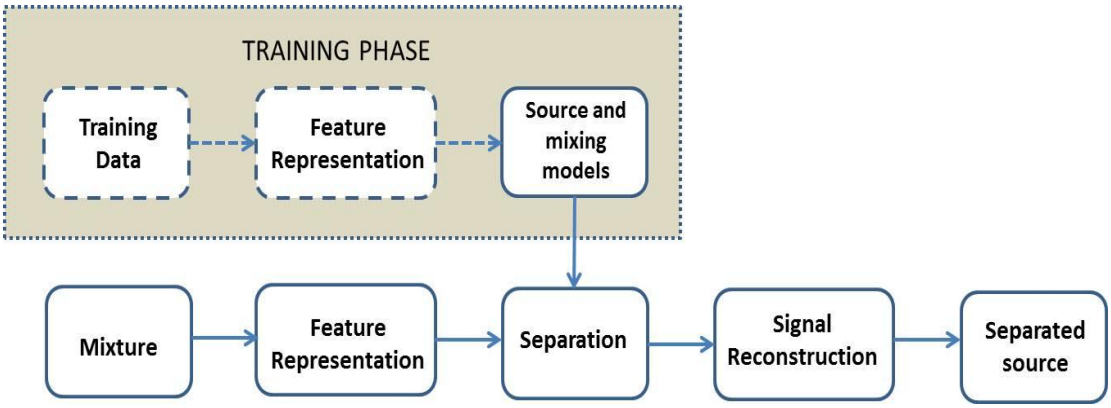


Figure 2.1: Schematic diagram of a general SCSS system

A schematic diagram of a general SCSS system is illustrated in Figure 2.1. This shows general tasks in development of SCSS algorithm. Note that for unsupervised SCSS methods, the training phase block is excluded from the workflow. The input to the source separation system is the mixed signal, $y(t)$. For supervised SCSS approaches, training data are needed for some or all of the source signals. Firstly, the mixture is transformed into an appropriate signal representation using e.g. short-time Fourier transform (STFT) or wavelet transform, in which the signal separation is performed. Signal representations can be chosen in order to highlight main characteristics in the signal that helps discriminate between sources. The transformations such as STFT and wavelet transforms are useful to produce sparse data in the transformed domain, and this can lead to effective separation algorithms. Then, the source models are either constructed directly based on knowledge of the signal sources, or by learning from training data. These models are used to capture properties of the sources and mixing process to effectively allow the sources to be separated, and have a convenient parametric form to allow efficient separation. In the separation stage, the models and data are combined to yield estimates of the sources, either directly or through a signal reconstruction step. As for unsupervised SCSS approaches, since the separation are done without relying on training information, the training phase block are not needed in the separation process.

After the signals are separated in the representation domain, separated signals must be reconstructed in the original signal domain. This can be attained using a filtering technique where the separated source is used to construct a filter that is applied to the signal mixture. This filter can be a time varying Wiener filter, binary or soft masking in a transform domain, etc. It is important to choose an appropriate

signal representation such that adequate signal reconstruction is achievable. Five main approaches in SCSS will be discussed in this chapter namely model-based SCSS, independent subspace analysis (ISA), empirical mode decomposition (EMD), computational auditory scene analysis (CASA) and nonnegative matrix factorization (NMF).

2.1 Model-based SCSS

Model-based SCSS techniques are similar to model-based single channel speech enhancement (SE) techniques [46]–[50]. In this case, SCSS can be considered as an SE problem in which both the target and interference with similar probabilistic characteristics which are non-stationary sources must be estimated. This is a supervised method and generally, the following procedures are commonly applied in model-based SCSS techniques. First, the training phase will generate the patterns of the sources. From the obtained patterns, combinations pattern that model the observation signal are chosen. Finally, the selected patterns are either directly used to estimate the sources [51]–[53] or used to build filters which when imposed on the observation signal result in an estimate of the sources [54]–[56].

Further work on model-based SCSS method has been proposed in [57] where it exploits the hidden Markov models (HMM) or other algorithms such as e.g. nonnegative matrix factorization, sparse code, etc. to generate codebook of audio signals. The HMM based methods are widely used and the heart of these frequency model based SCSS methods is the approximation of the posterior $p\left(\underline{\mathbf{x}}_{j,n},...,\underline{\mathbf{x}}_{J,n}\left|\underline{\mathbf{y}}_{n}\right.\right)$ by Gaussian distribution [54, 58]. The posterior distribution can be expressed as:

$$p\left(\underline{\mathbf{x}}_{j,n},...,\underline{\mathbf{x}}_{J,n}\left|\underline{\mathbf{y}}_n\right.\right) \propto p\left(\underline{\mathbf{y}}_n\left|\underline{\mathbf{x}}_{j,n},...,\underline{\mathbf{x}}_{J,n}\right.\right)\prod_{j=1}^{I} p\left(\underline{\mathbf{x}}_{j,n}\right) \tag{2.1}$$

where $\underline{\mathbf{x}}_{j,n}=\begin{bmatrix}\left|X_j\left(1,n\right)\right|^2\\ \vdots\\ \left|X_j\left(F,n\right)\right|^2\end{bmatrix}, \underline{\mathbf{x}}_{J,n}=\begin{bmatrix}\left|X_J\left(1,n\right)\right|^2\\ \vdots\\ \left|X_J\left(F,n\right)\right|^2\end{bmatrix}$ and $\underline{\mathbf{y}}_n=\begin{bmatrix}\left|Y\left(1,n\right)\right|^2\\ \vdots\\ \left|Y\left(F,n\right)\right|^2\end{bmatrix}$ are the power spectrum vectors. The priori information for the sources in probability density functions (pdf) is assumed as Gaussian mixture models (GMM) is defined as:

$$p\left(\underline{\mathbf{x}}_{j,n}\right)=\sum_{k_j}^{Q_j}\omega_{j,k_j}N_{gauss}\left(\underline{\mathbf{x}}_{j,n};\underline{\mathbf{u}}_{j,k_j},\boldsymbol{\Sigma}_{j,k_j}\right) \tag{2.2}$$

$$\text{where } N_{gauss}\left(\underline{\mathbf{x}}_{j,n};\underline{\mathbf{u}}_{j,k_j},\boldsymbol{\Sigma}_{j,k_j}\right)=\frac{\exp\left(-\frac{1}{2}\left(\underline{\mathbf{x}}_{j,n}-\underline{\mathbf{u}}_{j,k_j}\right)^{\mathbf{T}}\boldsymbol{\Sigma}_{j,k_j}^{-1}\left(\underline{\mathbf{x}}_{j,n}-\underline{\mathbf{u}}_{j,k_j}\right)\right)}{\left(2\pi\right)^{F/2}\det\left(\boldsymbol{\Sigma}_{j,k_j}\right)^{1/2}} \tag{2.3}$$

where $\underline{\mathbf{u}}_{j,k_j}$ is the mean vector, $\boldsymbol{\Sigma}_{j,k_j}$ is the diagonal covariance matrix with $\boldsymbol{\Sigma}_{j,k_j}=\text{diag}\left(\sigma_{j,k_j}^2\left(f\right)\right)$, $\omega_{j,k_j}\geq 0$ is the weight (satisfying $\sum_{k_j}\omega_{j,k_j}=1$), Q_j denotes number of components in the model of sources x_j , 'det' denotes determine and '**T**' is matrix transpose. In frequency model based SCSS method, $\underline{\mathbf{u}}_{j,k_j}$ and $\boldsymbol{\Sigma}_{j,k_j}$ of each source are trained before separation process.

In [55], Wiener filter is used to derive an estimation of the sources, given the mixture in the GMM setting. The estimation can be expressed as weighted Wiener filters where the weights are adaptive. Considered case of two sources, weighting probabilities is expressed as

$$\vartheta_{k_1,k_2}(n) \propto \omega_{1,k_1}\omega_{2,k_2} \times \prod_f N_{gauss}\left(Y(f,n);\left\{\sigma^2_{1,k_1}(f)+\sigma^2_{2,k_2}(f)\right\}\right) \tag{2.4}$$

where $Y(f,n) \sim N\left(0, \text{diag}\left(\sigma^2_{1,k_1}(f)+\sigma^2_{2,k_2}(f)\right)\right)$. Then, the posterior mean estimator is used to estimate the sources such that

$$\hat{X}_1(f,n) = \sum_{k_1=1}^{Q_1}\sum_{k_2=1}^{Q_2} \vartheta_{k_1,k_2}(n) \frac{\sigma^2_{1,k_1}(f)}{\sigma^2_{1,k_1}(f)+\sigma^2_{2,k_2}(f)} Y(f,n)$$

$$\hat{X}_2(f,n) = \sum_{k_1=1}^{Q_1}\sum_{k_2=1}^{Q_2} \vartheta_{k_1,k_2}(n) \frac{\sigma^2_{2,k_2}(f)}{\sigma^2_{1,k_1}(f)+\sigma^2_{2,k_2}(f)} Y(f,n) \tag{2.5}$$

The parameter $\left\{\sigma^2_{j,k_j}\right\}$ is estimated in a training phase in which excerpts of each source are provided. By using Expectation-Maximisation (EM) algorithm parameter $\left\{\sigma^2_{j,k_j}\right\}$ is estimated such that $\sigma^2_{j,k_j}(f) = \dfrac{\sum_n \vartheta_{k_j}(n)\left|X_j(f,n)\right|^2}{\sum_n \vartheta_{k_j}(n)}$.

In the case of GMM, the prior weights of the Gaussian densities are kept constant. Hidden Markov Models (HMM) with mixture of Gaussian conditional densities, of order *L*, can be seen as a generalisation of GMM, in which the prior weights at time slot *n* depend on the active HMM state, which corresponds to the component index $q_j(\tau)$ at previous times $\tau = n-1,....,n-L$. The HMM density for the source x_j is given by,

$$p\left(X_j(f,n)\right) = \sum_{k_j}^{Q_j} \omega_{j,k_j,q_j(n-1),...,q_j(n-L)} N_{gauss}\left(X_j(f,n);\left\{\sigma^2_{j,k_j}\right\}\right) \tag{2.6}$$

In [54], the factorial hidden Markov model (FHMM) was employed to separate the mixture. FHMM consists of two or more underlying Markov chains (the hidden states) which evolve independently. In this approach, a HMM/GMM is learned for each source on isolated training data, and to separate sources the most likely joint state sequence is separated. Element-wise max observation model is applied for efficient inference of a models with a large number of states. To estimate the separated sources, author in [54] proposes a re-filtering technique based on a binary mask.

For FHMM, good separation requires detailed source models that might employ thousands of full spectral states e.g. in [54], GMM with 8000 states were required to accurately represent one person's speech for a source separation task. The large state space required because it attempts to capture every possible instance of the signal. However, these model based SCSS techniques are computationally intensive not only for training the prior parameters but also for presenting many difficult challenges during both the learning and separation stages.

2.2 Independent Subspace Analysis

Another method in SCSS is based on independent subspace analysis (ISA) techniques [59-61] which motivate from independent component analysis (ICA) but relaxes the constraint that requires at least as many mixture observation signals as sources. ISA was originally proposed by [59] for images processing application and then was extending for SCSS by author in [61] for audio source separation application. The proposed approach in [61] extends an ISA method by extracting the statistically independent subspaces from the projection of a one-dimensional signal

onto a manifold. In addition, the approach also introduces the use of dynamics components to represent non-stationary signals where sources are tracked by similarity of dynamic components over small time steps.

In subspace analysis methods, firstly, the instantaneous mixture is transformed to the time-frequency domain using the short-time Fourier transform (STFT). Then, the time-frequency space of the mixed signal is decomposed as the sum of independent source subspaces. Given the poser spectrogram of mixture TF representation $|\mathbf{Y}|^{.2}$, each time frame of the mixture power spectrogram is expressed as a weighted sum of P independent basis vectors, $\mathbf{z}_{\rho}^{ISA}$:

$$\underline{\mathbf{y}}_n = \sum_{\rho=1}^{P} w_{\rho,n}^{ISA} \mathbf{z}_{\rho}^{ISA} \tag{2.7}$$

where $\rho = [1,...,P]$ denotes the index of basis vectors, P is the number of basis vectors, $\underline{\mathbf{y}}_n$ is the STFT transformed observed signal vector, and n is the time slot index. Each basis vector is weighted by a time-varying scalar $w_{\rho,n}^{ISA}$. The appropriate number of basis vectors ρ is found by singular value decomposition and applying a threshold on the decreasing sorted eigenvalues. Thus each source is spanned by a subset of such basis vectors that define a subspace which is a matrix with basis vectors in columns $\mathbf{Z}_i^{ISA} = \left[\mathbf{z}_{1,i}^{ISA},...,\mathbf{z}_{P^i,i}^{ISA}\right]$. In ISA methods, the weight coefficients are obtained by projection of the input $\underline{\mathbf{y}}_n$ onto each basis component in the subspace. Assume orthonormal components, namely:

$$\mathbf{w}_i^{ISA^{\mathbf{T}}} = \mathbf{Z}_i^{ISA^{\mathbf{T}}} \underline{\mathbf{y}}_n \tag{2.8}$$

This is the projection of $\underline{\mathbf{y}}_n$ on to the subspace spanned by the basis vectors $\mathbf{Z}_i^{ISA}$. By successively projecting onto each of the *I* sets of basis vectors, thus the $\underline{\mathbf{y}}_n$ is decomposed to sums of independent subspaces as:

$$\underline{\mathbf{y}}_n = \sum_{i=1}^{I} \mathbf{Z}_i^{ISA} \mathbf{w}_i^{ISA^{\mathbf{T}}} \tag{2.9}$$

To extend the method to all time frames of power spectrogram can be estimated as $|\mathbf{X}|_i^2 = \mathbf{Z}_i^{ISA} \mathbf{W}_i^{ISA^{\mathbf{T}}}$ where $\mathbf{W}_i^{ISA} = \left[\mathbf{w}_{i,n}^{ISA}, ..., \mathbf{w}_{i,N}^{ISA}\right]$. Finally, use inverse STFT to reconstruct $|\mathbf{X}|_i^2$ back to time domain source. As for approach in [3], the independent feature vectors are assigned to sources based on a similarity measure. The similarity is represented in an ixigram, which measures the mutual similarity of components in an audio segment as independent cross-entropy matrix. The pair-wise similarity measure is approximated by the symmetric Kullback-Leibler distance, resulting in a symmetric distance matrix. Grouping is performed by a clustering procedure using the dissimilarities in the distance matrix. The source signals can be reconstructed using the weights and the source dependent basis vectors.

Nevertheless, these ISA techniques employ the STFT to construct the TF plane which leads to remarkable amount of cross-spectral terms due to the harmonic assumption and the window overlapping between successive time frames. This drawback implies that it is difficult to represent the mixture as the sum of individual sources subspaces. The separation efficiency [57] is greatly affected by the cross-spectral energy introduced by STFT. In addition, this approach is only appropriate when the underlying sources have disjoint spectral support which guarantees that the ICA bases will be linearly independent. If the sources have overlapping support in

frequency, as is generally the case for mixtures of speech signals, then the separation algorithm must utilize strong prior information to obtain high quality separations. Another limitation of subspace analysis based SCSS techniques is that this process works well on extracting drums from a mixture because they tend to account for most of the variance in musical signals. However, because of the way in which model represents the data, it is limited to stationary pitch sound such as drums.

2.3 Empirical Mode Decomposition

The empirical mode decomposition (EMD) has recently gained reputation as a method for analysing nonlinear and non-stationary time series data. By combining other data analysis tools, EMD can be employed to separate the audio sources from a single mixture. Molla and Hirose [62] proposed a subspace decomposition based SCSS method using EMD and Hilbert spectrum (HS). The EMD decompose the mixture into sum of band-limited functions termed as intrinsic mode functions (IMFs), namely:

$$y(t) = \sum_{m=1}^{M} c_m(t) + r_M^{EMD}(t) \tag{2.10}$$

where $c_m(t)$ is the m^{th} IMF, M is the total number of IMFs, and $r_M^{EMD}(t)$ is the final residue. Constructing the Hilbert spectrum for both mixed and IMFs, this gives $\mathbf{Y}^H = \left[y_{f,n}^H \right]_{t_s=1,2,\ldots,N}^{f=1,2,\ldots,F}$ and $\mathbf{C}_m^H = \left[c_{m,f,n}^H \right]_{n=1,2,\ldots,N}^{f=1,2,\ldots,F}$. By computing the spectral projection vectors between the mixture and individual IMF components, this is defined as:

$$\theta_m^H(f) = \frac{\left|\xi_m^H(f)\right|^2}{\varphi_y^H(f)\varphi_{m,c}^H(f)} \tag{2.11}$$

where $\xi_m^H(f)$ is the cross spectrum of *y(t)* and $c_m(t)$, $\varphi_y^H(f) = \sum_{n=1}^{N} Y_{f,n}^H$ and $\varphi_{m,c}^H(f) = \sum_{n=1}^{N} c_{m,f,n}^H$. Thus we can arrange the spectral projection vectors as individual column of a matrix $\mathbf{D}^H = \left[\boldsymbol{\theta}_1^H, ..., \boldsymbol{\theta}_M^H\right]$ and then derive spectral independent bases from $\mathbf{D}^H$ by applying principal component analysis (PCA) and ICA. Once these sets of spectral independent bases are obtained, the KL divergence *k*-means clustering is used to group the bases into (number of sources) subsets. Finally, synthesis time domain estimated sources $\tilde{x}_j(t)$.

The performance of the above EMD based SCSS method rely very heavily on the derived basis vectors which are only stationary over time. Therefore, good separation results can be obtained only if basis vectors are statistically independent over time. For some source (e.g. male and female speeches), the features can be very similar and hence, it becomes difficult to obtain the independent basis vectors by PCA or ICA.

2.4 Computational Auditory Scene Analysis

Another method for SCSS that has been widely studied is based on a computational model of the human auditory scene and its processing in the brain. The goal in computational auditory scene analysis (CASA) [63-69] is to incorporate as much information as the human auditory system is using and replicate the process by

exploiting signal processing approaches (e.g. notes in music recordings) and grouping them into auditory streams using psycho-acoustical cues. In CASA methods, after an appropriate transform such as the short-time Fourier transform (STFT) or cochleagram TF representation, low level perceptual cues are used to segment a time-frequency representation of a mixture into regions consistent with being generated by a single source. The grouping cues [70] can be summarised as follow:

- Common onset and offset – sound component with the same onset time and to a lesser extent, the same offset time, tend to be grouped into same unit by auditory system. Since unrelated sounds seldom start or stop at exactly the same time, the auditory system assumes that components showing a "common fate" are likely to have origin in a common single source. To find the starting of musical events (e.g. notes, chords and etc.), the spectral flux can be used as the onset detection function defined as,

$$SF(n) = \sum_{k=0}^{N_H/2} H_w\left(\left|X\left(n,k\right)\right|^2 - \left|X\left(n-1,k\right)\right|^2\right) \tag{2.12}$$

where $X\left(n,k\right)$ represents the k^{th} frequency bin of the n^{th} frame of power magnitude in dB of the STFT, $H_w = \frac{x+|x|}{2}$ is the half-wave rectifier function and N_H is the Hamming window size. A peak at time at $t = \frac{nH_w}{f_s}$ is selected as an onset if it satisfies the following conditions:

$$\begin{aligned} &SF(n) \geq SF(k) \quad \forall k : n-w \leq k \leq n+w \\ &SF(n) > \frac{\sum_{k=n-mw}^{n+w} SF(k)}{mw+w+1} \times thres \end{aligned} \tag{2.13}$$

where *w* is the size of the window used to find a local maximum, *m* is the multiplier so that the mean is computed over the larger range before the peak and *thres* is threshold relative to the local mean that the peak must reach in order to adequately prominent to be selected as an onset.

- Amplitude and frequency similarity - Amplitude and frequency features of the sound components are the most basic similarities explored by the auditory system. Accordingly, the edge weight connecting two peaks p_l^k and p_m^{k+n} will depend on their amplitude and frequency proximities. Amplitude and frequency similarities, S_a and S_f respectively, are defined as follows,

$$S_a\left(p_l^k, p_m^{k+n}\right) = \exp\left[-\left(\frac{a_l^k - a_m^{k+n}}{\sigma_a}\right)\right]$$
$$S_f\left(p_l^k, p_m^{k+n}\right) = \exp\left[-\left(\frac{f_l^k - f_m^{k+n}}{\sigma_f}\right)\right] \quad (2.14)$$

where the Euclidean distances are modelled as two Gaussian functions. The amplitudes are measured in dB while the frequencies are measured in Barks.

- Harmonicity – A wide variety of sounds produced by humans are harmonic. When a body vibrates with a periodic movement, its vibrations create an acoustic pattern whose frequency components are multiples of a common fundamental i.e. harmonics of a fundamental frequency. Interestingly, the auditory system tends to group a set of harmonically related acoustic components into single event. In [68], harmonicity principle is modelled using smoothed harmonic map, $h(i, j)$ which provides harmonic similarity between spectral lines *i* and *j* where $h(i, j) = 1$ if $\frac{i}{j}$ or $\frac{j}{i}$. Since the spectrograms are noisy, the smoothing process is performed. The smoothed harmonic similarity

between spectral lines $i+\delta i$ and $j+\delta j$ due to harmonic similarity is given by $sh(i,j)=h(i,j)\times N\left(i,pi;\delta i\right)\times N\left(j,pj;\delta j\right)$ where $N\left(\mu,\sigma;s\right)$ is the Gaussian function with parameters μ, σ and p is a constant. The map is normalised so that $\sum_j sh(i,j)=1$. Then, the similarity between lines i and j at time t is defined as

$$S_t(i,j)=\left|X(t,i)\right|sh(i,j) \tag{2.15}$$

where $\left|X(t,i)\right|$ denotes the magnitude spectrum at time t and spectral line i.

- Common modulation – If a sound source exhibits amplitude or frequency modulation, it is expected that all of its components show similar modulation manifestations.
- Spatial proximity – one of the best generalisations that can be made about independent sound sources is that they normally occupy distinct positions in space. As a consequence, sound source location could provide the strongest cue in the construction of an ecological representation of the surrounding sound environment.

For more detailed review of CASA grouping cues, please refer to [70]. Time-frequency cells with consistent cues are then grouped together using a masking technique which then be used for separation. More progressive machine learning approaches have also been proposed for segmentation such as in [71] where segments sources using spectral clustering of perceptual grouping cues. However, these perceptual cues are only applicable to harmonic signals such as voiced speech and thus are limited in terms of the types of signals they can separate. In addition, most of CASA methods cannot efficiently segregate instruments playing in the same pitch

range into different streams. In current approaches, because of the difficulty in grouping process, in order to ease the task, one of the essential signals is assumed fully voice. Furthermore, CASA approaches suffer the problems where separability in speech separation is rather limited for unvoiced speech and the formant structure is not explicitly used as a feature [72].

2.5 Nonnegative Matrix Factorization

One of the popular techniques in SCSS that recently attract much attention is nonnegative matrix factorization (NMF) [73-80]. NMF is a factorization technique for decomposing data with nonnegative constraint on the component. In many fields, real components contains nonnegative element such as the microarray data, amplitude spectra, pixel intensities and the natural images [81, 82]. For example, for signal cryptosystem, in order to get an accurate result in decryption, the plaintext sources need to be nonnegative before encryption [83]. Therefore, in the analysis of mixtures of such data, non-negativity of the individual components is a reasonable constraint. NMF gives a more part based decomposition [84] and the decomposition is unique under certain conditions [85], making it unnecessary to impose the constraints in the form of orthogonality and independence which have led to a significant interest in NMF. Because of its simplicity and applicability, NMF has been successfully applied in the fields of signal/image processing such as for application of audio source separation [86] [87], automatic music transcription [88] and pattern recognition [89]. In image processing or pattern recognition, the images/objects is expressing as a linear combination of basis objects which used to extract features. While in source

separation problem, the magnitude or power spectrogram of a sources signal is taken as a data matrix,

2.5.1 Conventional NMF

The key point in this NMF technique is to model the power of mixture as a product of two nonnegative matrices in time- frequency (TF) domain using e.g. Short-time Fourier transform (STFT). In conventional NMF, for analysis on the single input recordings, (1.8) yielding a TF representation which can be decomposed as:

$$\left|\mathbf{Y}\right|^2 \approx \mathbf{WH} \tag{2.16}$$

where $\left|\mathbf{Y}\right|^2 \in \Re_+^{F\times N}$ is the power TF representation of mixture $y(t)$ which is the product of two nonnegative matrices, $\mathbf{W} \in \Re_+^{F\times K}$ and $\mathbf{H} \in \Re_+^{K\times N}$ and F and N represent the frequency bins and time slots in matrix $\mathbf{Y}$, respectively. K is usually chosen such that $FK + KN << FN$, hence reducing the data dimension. If K is chosen to be $K=N$, no benefit is achieved in terms of representation. Thus the idea is to determine $K<N$ so the data matrix $\mathbf{W}$ can be compressed and reduced to its integral components such as $\mathbf{W}$ is a matrix containing only a set of spectral basis vectors, and $\mathbf{H}$ is an encoding matrix which describes the amplitude of each basis vector at each time point. NMF is used to estimate the spectral bases and temporal code of the sources signal. In the combination of the basis, NMF does not allow negative elements in either the signal basis or the signal coefficients/dictionary. Thus, it represents the signals only by additions of weighted signal basis, such that no cancellations can occur. This agrees with the intuitive idea of building the whole as the sum of its parts.

Commonly use cost functions introduced in [84] for NMF are the Least Square (LS) distance and generalised Kullback-Leibler (KL) divergence which are expressed as:

$$\begin{aligned} &\text{Least Square}: && C_{LS}\left(\left|\mathbf{Y}\right|^{.2}\left\|\hat{\mathbf{Y}}\right|^{.2}\right)=\frac{1}{2}\sum_{f,n}\left(\left|\mathbf{Y}_{f,n}\right|^{2}-\left|\hat{\mathbf{Y}}_{f,n}\right|^{2}\right)^{2} \\ &\text{Kullback-Leibler}: && C_{KL}\left(\left|\mathbf{Y}\right|^{.2}\left\|\hat{\mathbf{Y}}\right|^{.2}\right)=\sum_{f,n}\left(\left|\mathbf{Y}_{f,n}\right|^{2}\log\frac{\left|\mathbf{Y}_{f,n}\right|^{2}}{\left|\hat{\mathbf{Y}}_{f,n}\right|^{2}}-\left|\mathbf{Y}_{f,n}\right|^{2}+\left|\hat{\mathbf{Y}}_{f,n}\right|^{2}\right) \end{aligned} \tag{2.17}$$

where $\left|\hat{\mathbf{Y}}\right|^{.2}=\mathbf{WH}$. In (2.17), C_{LS} is equivalent to the maximum likelihood estimation of **W** and **H** in additive independent and identically distributed (i.i.d.) Gaussian noise and C_{KL} is equivalent to assuming a Poisson noise model for the data. In [84], the authors minimise the chosen cost function by initializing the entries of **W** and **H** with random positive values, and then by using multiplicative update rules, **W** and **H** are updated iteratively. Each update decreases the value of the cost function until the algorithm converges. The update rule of **W** and **H** for LS divergence is given by:

$$\mathbf{W}\leftarrow\mathbf{W}\bullet\frac{\left|\mathbf{Y}\right|^{.2}\mathbf{H}^{\mathrm{T}}}{\mathbf{WHH}^{\mathrm{T}}}\quad\text{and}\quad\mathbf{H}\leftarrow\mathbf{H}\bullet\frac{\mathbf{W}^{\mathrm{T}}\left|\mathbf{Y}\right|^{.2}}{\mathbf{W}^{\mathrm{T}}\mathbf{WH}} \tag{2.18}$$

where '$\bullet$' denote the element-wise multiplication and '$\frac{\mathbf{A}_{NMF}}{\mathbf{B}_{NMF}}$' denotes the element-wise division of matrices $\mathbf{A}_{NMF}$ and $\mathbf{B}_{NMF}$. The multiplicative update rule for KL divergence is given by:

$$\mathbf{W}\leftarrow\mathbf{W}\bullet\frac{\left(\left|\mathbf{Y}\right|^{.2}./\,\mathbf{WH}\right)\mathbf{H}^{\mathrm{T}}}{\mathbf{1H}^{\mathrm{T}}}\quad\text{and}\quad\mathbf{H}\leftarrow\mathbf{H}\bullet\frac{\left(\left|\mathbf{Y}\right|^{.2}./\,\mathbf{WH}\right)\mathbf{W}^{\mathrm{T}}}{\mathbf{1W}^{\mathrm{T}}} \tag{2.19}$$

where '**1**' is an all-one *F* by *N* matrix and './' denote the element-wise division.

The extensions of NMF have also been proposed by using other families of parameterised cost functions, such as the Beta divergence [90] and Cziszar's divergences [91] for the separation of audio signals. After factorization, the recovered j^{th} source in TF representation can be estimated as $\left|\tilde{\mathbf{X}}\right|_j^{.2} \approx \mathbf{W}_j\mathbf{H}_j$ where $\mathbf{W}_j$ represent the spectral basis of j^{th} source in TF representation and $\mathbf{H}_j$ represents the code for each spectral basis element. Regardless of the cost function being used, in order to achieve audio source separation, some methods are required to group the basis functions by source or instrument. As discussed in [92] practically if the sources overlap in the TF domain, it is difficult to obtain the correct clustering.

2.5.2 Convolutive NMF

Normally, the temporal relationship between multiple observations over nearby intervals of time is discovered using a convolutive generative model. Motivated by this, author in [93] extends the conventional model in (2.16) by introducing the convolutive NMF model (also known as nonnegative matrix deconvolution (NMFD)) where each object has a sequence of successive spectral and corresponding activation pattern across time such that

$$\left|\mathbf{Y}\right|^{.2} \approx \sum_{\tau=0}^{\tau_{\max}} \mathbf{W}^{\tau} \overrightarrow{\mathbf{H}}^{\tau} \tag{2.20}$$

where $\tau_{\max}$ is the length of each spectrum sequence and $\mathbf{W}^{\tau}$ represent the τ^{th} slice of spectral basis $\mathbf{W}$ The arrow sign in $\overrightarrow{\mathbf{H}}^{\tau}$ denotes the right shift operator which moves each element in the matrix by τ column to the right, with zero filling on the right i.e.

$$\mathbf{B}=\begin{bmatrix}1 & 2 & 3\\ 4 & 5 & 6\end{bmatrix},\quad \overrightarrow{\mathbf{B}}^{1}=\begin{bmatrix}0 & 1 & 2\\ 0 & 4 & 5\end{bmatrix},\quad \overrightarrow{\mathbf{B}}^{2}=\begin{bmatrix}0 & 0 & 1\\ 0 & 0 & 4\end{bmatrix}$$

Now $\left|\hat{\mathbf{Y}}\right|^{.2}$ of cost function in (2.17) for the convolutive generative function is expressed as $\left|\hat{\mathbf{Y}}\right|^{.2}=\sum_{\tau=0}^{\tau_{\max}}\mathbf{W}^{\tau}\overrightarrow{\mathbf{H}}^{\tau}$. The new cost function can be viewed as a set of $\tau_{\max}$ conventional NMF operations that are summed to produce the final result. Consequently, as opposed to updating $\mathbf{W}$ and $\mathbf{H}$ as in conventional NMF, $\tau_{\max}+1$ matrices require an update $\mathbf{W}_0,...,\mathbf{W}_{\tau_{\max}}$ and $\mathbf{H}$. Using the multiplicative update rules, the resultants update for $\mathbf{W}$ and $\mathbf{H}$ of LS distance is expressed as

$$\mathbf{W}^{\tau}\leftarrow\mathbf{W}^{\tau}\bullet\frac{\left|\mathbf{Y}\right|^{.2}\overrightarrow{\mathbf{H}}^{\tau\mathrm{T}}}{\left|\hat{\mathbf{Y}}\right|^{.2}\overrightarrow{\mathbf{H}}^{\tau\mathrm{T}}}\quad\text{and}\quad\mathbf{H}\leftarrow\mathbf{H}\bullet\frac{\mathbf{W}^{\tau\mathrm{T}}\overleftarrow{\left|\mathbf{Y}\right|^{.2}}^{\tau}}{\mathbf{W}^{\tau\mathrm{T}}\overleftarrow{\left|\hat{\mathbf{Y}}\right|^{.2}}^{\tau}}\tag{2.21}$$

As for the multiplicative update rule of KL divergence is given by:

$$\mathbf{W}^{\tau}\leftarrow\mathbf{W}^{\tau}\bullet\frac{\left(\left|\mathbf{Y}\right|^{.2}./\left|\hat{\mathbf{Y}}\right|^{.2}\right)\overrightarrow{\mathbf{H}}^{\tau\mathrm{T}}}{\mathbf{1}\overrightarrow{\mathbf{H}}^{\tau\mathrm{T}}}\quad\text{and}\quad\mathbf{H}\leftarrow\mathbf{H}\bullet\frac{\left(\overleftarrow{\left|\mathbf{Y}\right|^{.2}}^{\tau}./\overleftarrow{\left|\hat{\mathbf{Y}}\right|^{.2}}^{\tau}\right)\mathbf{W}^{\tau\mathrm{T}}}{\mathbf{1}\mathbf{W}^{\tau\mathrm{T}}}\tag{2.22}$$

For each τ , $\mathbf{H}$ and each $\mathbf{W}^{\tau}$ are update at every iteration. That way, the factors can be optimise in parallel and account for their interplay.

2.5.3 Two-dimensional nonnegative matrix factorization

The recently developed two-dimensional NMF factorization (NMF2D) model [94] extends the NMF model to be two-dimensional convolution of **W** and **H**. This model is the extension of NMFD model of (2.20). The factorization is based on a model that represents temporal structure and pitch change. In audio source separation for example, the model represents each instruments by a single time-frequency profile convolved in both time and frequency by a time-pitch weight matrix. This model radically reduces the number components need to model various instruments and effectively solves the SCSS problem. NMF2D model can be formulated as

$$|\mathbf{Y}|^{\cdot 2} \approx \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{W}^{\tau} \overset{\downarrow\phi}{\mathbf{H}}^{\phi} \overset{\rightarrow\tau}{} \tag{2.23}$$

The matrix $\mathbf{W}^{\tau}$ represents the τ^{th} slice spectral basis and $\mathbf{H}^{\phi}$ represents the ϕ^{th} slice of temporal code for each spectral basis element. The superscript upper arrow sign in $\overset{\downarrow\phi}{\mathbf{W}}^{\tau}$ denotes downward shift operator which moves each element in the matrix by ϕ row down. By the same token, the arrow sign in $\overset{\rightarrow\tau}{\mathbf{H}}^{\phi}$ denotes the right shift operator which moves each element in the matrix by τ column to the right. This can be interpreted as follows, i.e:

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix}, \quad \overset{\downarrow 1}{\mathbf{B}} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix}, \quad \overset{\rightarrow 2}{\mathbf{B}} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

The 3D-representation for **W** and **H** are shown in Figure 2.2. **W** has been sliced in frontal (i.e. τ^{th}-slice) and vertical (i.e. j^{th}-slice) directions. It is true that **W** can also have horizontal slice representation but this has not been shown since we don't need it

for NMF2D model. As for **H**, it has been sliced in frontal (i.e. ϕ^{th}-slice) and horizontal directions (i.e. j^{th}-slice) .

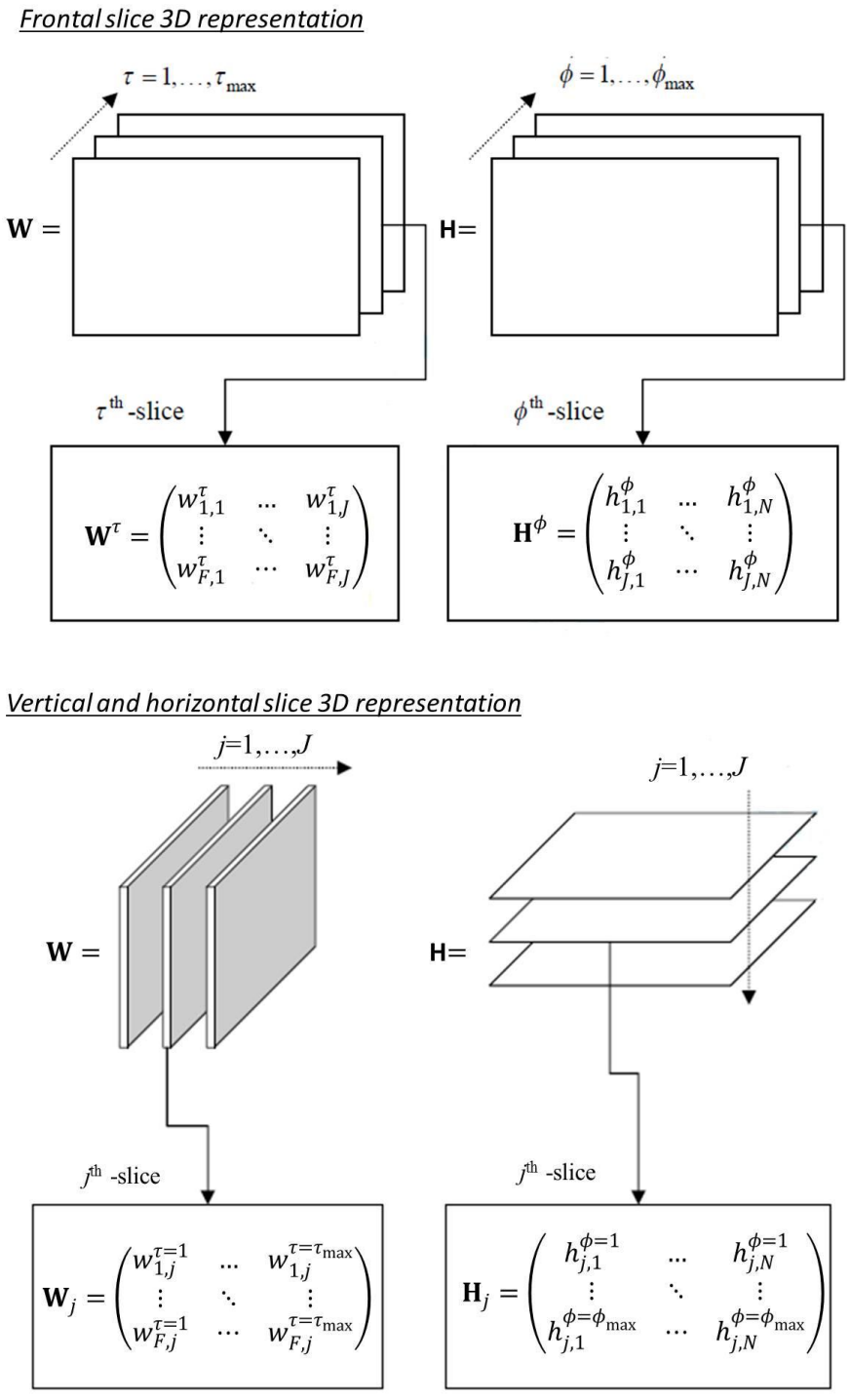


Figure 2.2: 3D-representation for **W** and **H**

We replace $\left|\hat{\mathbf{Y}}\right|^{.2} = \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{W}^{\tau} \overset{\downarrow\phi}{\mathbf{H}}^{\phi} \overset{\rightarrow\tau}{}$ in cost function of (2.17) for NMF2D model, the multiplicative update rules for LS distance of $\mathbf{W}$ and $\mathbf{H}$ can be expressed as

$$\mathbf{W}^{\tau} \leftarrow \mathbf{W}^{\tau} \bullet \frac{\overset{\uparrow\phi}{\left|\mathbf{Y}\right|}^{.2} \overset{\rightarrow\tau}{\mathbf{H}}^{\phi\,\mathrm{T}}}{\overset{\uparrow\phi}{\left|\hat{\mathbf{Y}}\right|}^{.2} \overset{\rightarrow\tau}{\mathbf{H}}^{\phi\,\mathrm{T}}} \quad \text{and} \quad \mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi} \bullet \frac{\overset{\downarrow\phi}{\mathbf{W}}^{\tau\,\mathrm{T}} \overset{\leftarrow\tau}{\left|\mathbf{Y}\right|}^{.2}}{\overset{\downarrow\phi}{\mathbf{W}}^{\tau\,\mathrm{T}} \overset{\leftarrow\tau}{\left|\hat{\mathbf{Y}}\right|}^{.2}} \tag{2.24}$$

For KL divergence, the multiplicative update rule is given by:

$$\mathbf{W}^{\tau} \leftarrow \mathbf{W}^{\tau} \bullet \frac{\left(\overset{\uparrow\phi}{\left|\mathbf{Y}\right|}^{.2} ./ \overset{\uparrow\phi}{\left|\hat{\mathbf{Y}}\right|}^{.2}\right) \overset{\rightarrow\tau}{\mathbf{H}}^{\phi\,\mathrm{T}}}{\mathbf{1} \overset{\rightarrow\tau}{\mathbf{H}}^{\phi\,\mathrm{T}}} \quad \text{and} \quad \mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi} \bullet \frac{\left(\overset{\leftarrow\tau}{\left|\mathbf{Y}\right|}^{.2} ./ \overset{\leftarrow\tau}{\left|\hat{\mathbf{Y}}\right|}^{.2}\right) \overset{\downarrow\phi}{\mathbf{W}}^{\tau\,\mathrm{T}}}{\mathbf{1} \overset{\downarrow\phi}{\mathbf{W}}^{\tau\,\mathrm{T}}} \tag{2.25}$$

It is important to note that the NMF2D model has certain ambiguities between the factors $\mathbf{W}^{\tau}$ and $\mathbf{H}^{\phi}$. For example in audio source separation, there exists an adverse shift of the time-pitch signature if a time-frequency signature of an instrument is shifted in time or frequency (if disregard edge effect). In order to improve the shift ambiguity, it can be useful to shift $\mathbf{W}^{\tau}$ and $\mathbf{H}^{\phi}$ during the recursive computation e.g. such that the geometric mean value of the row coefficients in $\mathbf{W}^{\tau}$ and the column coefficients of $\mathbf{H}^{\phi}$ are centered.

2.6 Summary

In this chapter, a general framework and the main approaches in SCSS have been reviewed. Since it is a necessary to use additional information about the sources to constrain the problem in SCSS, all approaches discussed here can be considered as forms of constrained optimisation such as mutual statistical independence, and non-

negativity. In source separation, it have been stated that if the number of observations decreases then the similarity of the underlying sources increases. Hence more constrained must be apply in the separation methods. This is a great challenge for the extreme case in separating single channel mixtures of multiple sources. Method like model-based SCSS which is the supervised SCSS methods is more reliable and accurate since they rely on access to source-specific training data to learn constraints in the form of source model. In addition, these methods can be used for all types of mixture if the prior knowledge or training data of the source models are provided. However, in most real applications, only observation signal is available in such case the supervised SCSS methods cannot separate it efficiently because of the lack of the prior knowledge of source models it is necessary to use additional information about the sources to constrain the problem. Hence, SCSS approaches of like ISA, EMD, CASA and NMF have been proposed to resolve this problem. Since no training data are needed, these methods have the advantage of being less computationally intensive. So far, the SCSS methods discussed in this chapter basically are based on conventional linear instantaneous mixture model. These methods dominate the current literature due to their simplicity and are computationally inexpensive. However, this assumption is often violated and may not be an accurate representation of the actual observed mixture. This problem have leads the research direction in focusing on the issue of how to develop the SCSS methods for all types of mixtures environment with high accuracy separation performance. As for unsupervised SCSS method, the issue on how to learn source model and automatic detect the number of sources when only a mixture is available need to be resolve. In this thesis, three novel SCSS methods will be developed where the first one is for the case of post-nonlinear instantaneous

mixture using a supervised SCSS method and the other two methods is an unsupervised method for solving the case of linear convolutive mixture. The design of each method will be described in the next three chapters.

CHAPTER 3

NONLINEAR SINGLE CHANNEL SOURCE SEPARATION IN INSTANTANEOUS MIXTURE

The SCSS methods explain in Chapter 2 by far were based on linear instantaneous model. Linear model is known for its simplicity and ease of implementation. These methods show very good performance in separating the signal for linear mixture. However, in practical applications such as in speech recognition, music transcription or telecommunications, the transmitted signals are often received by nonlinear receiver such as carbon-button microphones [95-97] or antennas [98-100]. Hence, in the real environments, the mixed signals are more likely to be nonlinear or subject to some kind of nonlinear distortions due to sensor sensitivity. Therefore, the assumption that the mixture is linear adopted by the existing SCSS approaches is violated and may not characterise the actual observed signals accurately. The need of an accurate representation of the distorted signals has resulted in the emergence of SCSS for nonlinear mixture model. So far, no method in SCSS has been proposed to solve the nonlinear problem.

In this chapter, a new model of nonlinear single channel source separation is proposed. The proposed model is a linear mixture of the independent sources followed by an element-wise post-nonlinear distortion function. In addition, the research develops a novel solution that efficiently compensates for the nonlinear distortion and performs source separation. The proposed solution is a two-stage process that consists

of a Gaussianization transform and a maximum likelihood estimator for the sources. The chapter also discusses the theory behind the proposed solution. Simulations have been carried out to verify the theory and evaluate the performance of the proposed algorithm. Results obtained have shown the effectiveness of the algorithm even in presence of the strong nonlinearity.

The organisation of the chapter is as follow. Section 3.1 describes the underdetermined and the post-nonlinear single channel model. Section 3.2 explains the proposed solution for two stage process. Discussion of the Gaussianization transform performances in compensates the nonlinearity and experimental results are analysed in Section 3.3. Finally, Section 3.4 summarises the work in this chapter.

3.1 Background

3.1.1 Nonlinear single channel instantaneous mixture model

For linear mixture of SCSS [101, 102], suppose that the observed time domain signal $y(t)$ is mixed with *J* independent sources

$$y(t) = a_1 x_1(t) + a_2 x_2(t) + ... + a_3 x_J(t) \tag{3.1}$$

where $x_j(t)$ is the t^{th} sampled value of j^{th} source signal, a_j is the mixing gain of sources. The goal is to recover $x_j(t)$ given only single input $y(t)$.

In this chapter, we propose the nonlinear single channel mixture that is modelled by the post-nonlinear (PNL) mixing model as described in the following:

$$v(t) = f\left(a_1 x_1(t) + a_2 x_2(t) + \ldots + a_J x_J(t)\right) \tag{3.2}$$

where $f(.)$ is an invertible nonlinear function. The basis functions and coefficients learned by independent component analysis (ICA) constitute an efficient representation of the given time-ordered sequences of a sound source by estimating the maximum likelihood densities. For each source signal, a K-sample vector $\mathbf{x}_j(t) = \left[x_j(t)\ x_j(t+1)\ \cdots\ x_j(t+J-1) \right]^{\mathrm{T}}$ can be expressed as a linear combination of basis functions such that

$$\mathbf{x}_j(t) = \sum_{k=1}^{K} m_{jk}^{ICA} s_{jk}(t) = \mathbf{M}_j^{ICA} \mathbf{s}_j(t) \tag{3.3}$$

where K is the number of basis functions, m_{jk}^{ICA} is the k^{th} basis function of j^{th} source, and $s_{jk}(t)$ is the coefficient. The transform between $\mathbf{x}_j(t)$ with coefficient vector, $\mathbf{s}_j(t)$ is assumed to be reversible with

$$\mathbf{s}_j(t) = \mathbf{W}_j^{ICA} \mathbf{x}_j(t) \tag{3.4}$$

where $\mathbf{W}_j^{ICA}$ is the inverse of the basis matrix, $\mathbf{W}_j^{ICA} = \left(\mathbf{M}_j^{ICA}\right)^{-1}$. Then, the ICA algorithm can be exploited for capturing the source coefficient density. Generalised exponential density model [103] is expressed as

$$p\left(s|q_a,u,\sigma\right)=\frac{q^a\sigma^{-\frac{1}{q_a}}}{2\Gamma\left(\frac{1}{q_a}\right)}\exp\left(-\left|\frac{s-u}{\sigma}\right|^{q_a}\right) \tag{3.5}$$

where $\Gamma(.)$ denotes the gamma function. The coefficients are determined by parameters mean *u*, exponent q_a and variance $\sigma=\sqrt{E\left[\left(s-u\right)^2\right]}$ where *E*[.] denotes the expectation. By using maximum log likelihood estimator, gradient ascent adaptation of coefficient density is obtained such that

$$\varphi(s)=\frac{\partial \log p\left(s|q_a,u,\sigma\right)}{\partial s} \tag{3.6}$$

In our proposed method, for simplicity, mean and variance are zero and unit, respectively.

The PNL model represents the important subclass of the general nonlinear model which is simpler and widely applicable. PNL structure is popular due to its ability to model some systems reasonably well such as those that involve the use of nonlinear sensors [104]. The PNL model that will be used is shown in Figure 3.1 where two independent sources are mixed together and then are distorted by a nonlinear function.

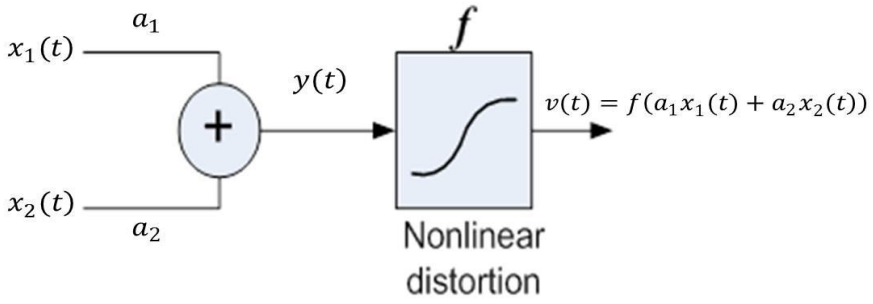


Figure 3.1: Post-nonlinear mixing model for SCSS

3.2 Proposed Separation Method

In this research work, the sources are estimated in a two-stage approach as shown in Figure 3.2. In the first stage, the nonlinearity of the observation $\mathbf{v}(t)$ is equalised using the nonlinear transform, $g(v) = \hat{f}^{-1}(v)$ with the linearised signal is given by $z(t) = g\left(v(t)\right)$. This is followed by the second stage where the linear separation algorithm based on maximum likelihood (ML) [101] will be used to obtain an estimated sources of $\hat{x}_1(t)$ and $\hat{x}_2(t)$.

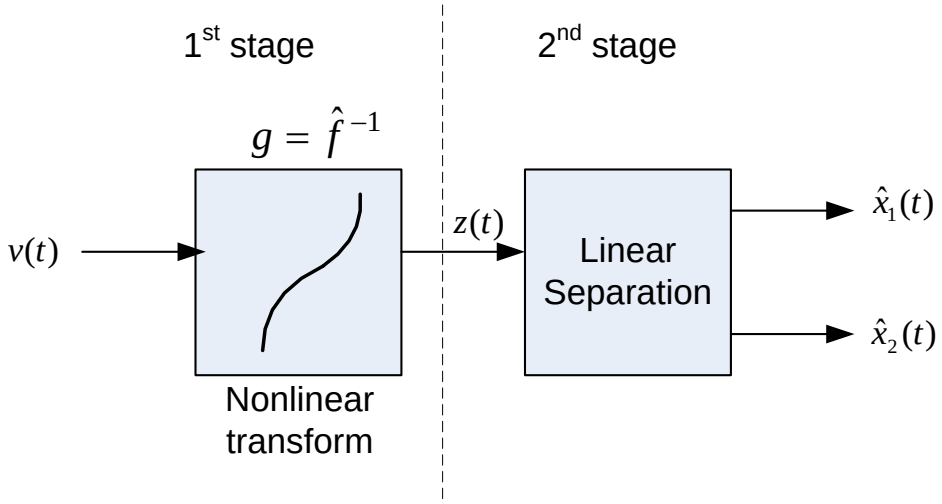


Figure 3.2: Proposed two-stage nonlinear SCSS

3.2.1 Nonlinearity compensation

In order for the signals in nonlinearly distorted mixture to be accurately separated, the nonlinearity of the mixture must be compensated. To this end, we propose a linearisation technique known as the Gaussianization transform. The main impetus of using the proposed technique comes from the following principle: Firstly, we observe that the sources from the observation are statistically independent and non-Gaussian.

According to the central limit theorem, when the sources are mixed the resulting mixture tends toward a Gaussian distribution. However, the post-nonlinearity *f*(.) distorts the amplitude distribution of the linear mixture and subsequently transform it to a non-Gaussian distribution. Thus, the non-Gaussian behaviour of the observation *v*(*t*) to some extents is directly attributed to the nonlinearity *f*(.) in the mixture. As such, the nonlinearity can be compensated by finding a suitable transformation such that the output returns to a Gaussian distribution. Therefore, it is clear that by using this principle, the estimation problem can be readily split into two tasks where the first task is to compensate for the nonlinear distortion and the second task is to seek separation from the compensated signals where separation can be assumed to be linear [105-107]. In this chapter, we constrained the number of sources to be two for illustration purpose only. In reality, the number of sources will be considerably more than two and in such case, our proposed method will work even more efficiently.

Although the nonlinearity *f*(.) is unknown, it is possible to determine the inverse function $g_j(.)$ by finding a suitable transformation which convert the component $v_j(t)$ to the Gaussian random variable. The goal is to find $g_j(.)$ such that

$$g_j\left(x_j\right) \sim N\left(0, \sigma_j^2\right) \qquad \text{for} \quad j = 1, ..., J \tag{3.7}$$

where $\sigma_j^2 = 1$ due to the usual scaling indeterminacies.

Consider $F_v(v)$ which denote the cumulative density function (cdf) as

$$F_v(v) = \int_{-\infty}^{v} p_v v(t) dt \tag{3.8}$$

with $p_v(v)$ denote probability density function (pdf) of v, $p_v(v) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right)$.

Assuming cdf of z, $F_z(z)$ is continuos and strictly increasing so that the inverse mapping exist, then $F_z(z)$ can be express as,

$$\begin{aligned} F_z(z) &= F_v(v) \\ F_z\left(g(v)\right) &= F_v(v) \\ g(v) &= F_z^{-1} F_v(v) \end{aligned} \tag{3.9}$$

Since the desired distribution of $z(t)$ is Gaussian, and the cdf of Gaussian is expressed as, $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} v^{-\frac{t^2}{2}} dt$ the expression of F_z^{-1} can be replaced by the inverse of Gaussian cdf, $\Phi(z)^{-1}$. Hence, the nonlinear mapping $g(.)$ can be expressed as

$$g = \Phi^{-1} \circ F_v \tag{3.10}$$

If the mixed signal, $y(t)$ are closely Gaussian distribution, then the Gaussianized signal should estimate the signal perfectly with $z(t) \approx y(t)$.

3.2.2 Source Estimation: Maximum Likelihood

After recovery of the nonlinearity, the linear SCSS approach based on maximum likelihood (ML) [101] was used to solve the single channel source separation problem. The main idea approach for SCSS is based on assuming that the audio source signal can be represented by a set of ICA basis functions which can be learned by training. These bases are then used to discriminate between sources using an assumption of source independence and in the probability model. The basis functions imply inherent

types of non-Gaussian signals. Thus the separation algorithm use hybrid of maximum likelihood (ML) and maximum a posteriori (MAP) estimators to recover the independent components

Considered two sources signal $x_1(t)$ and $x_2(t)$, according to ICA, the sources are said to be statistically independent if only the probability density function (pdf),

$$p_{x_1(t),x_2(t)}\left(y(t)\middle|\hat{x}_1(t),\hat{x}_2(t)\right)=p_{x_1(t)}\left(x_1(t)\right)p_{x_2(t)}\left(x_2(t)\right) \tag{3.11}$$

where number of samples, $t=1,\ldots,T$. The sources vectors are passed through the fixed basis filter $\mathbf{W}_j^{ICA}$ to generate set of basis coefficients,

$$\begin{aligned} p\left(x_j(1),x_j(2),\ldots,x_j(T)\middle|\mathbf{W}_j^{ICA}\right)&=\prod_{t=1}^{T}p\left(\mathbf{x}_j(t)\middle|\mathbf{W}_j^{ICA}\right)\\ &=\prod_{t=1}^{T}\prod_{k=1}^{K}p\left(\mathrm{s}_{jk}(t)\right)\left|\det\left(\mathbf{W}_j^{ICA}\right)\right| \end{aligned} \tag{3.12}$$

The likelihood function, L^{ICA} can be expressed as,

$$\begin{aligned} L^{ICA}&=\log\prod_{t=1}^{T}p\left(\mathrm{x}_1(t)\middle|\mathbf{W}_1^{ICA}\right)p\left(\mathrm{x}_2(t)\middle|\mathbf{W}_2^{ICA}\right)\\ &=\log\prod_{t=1}^{T}\left\{\prod_{k=1}^{K}p\left(s_{1k}(t)\right)\left|\det\left(\mathbf{W}_1^{ICA}\right)\right|\prod_{k=1}^{K}p\left(s_{2k}(t)\right)\left|\det\left(\mathbf{W}_2^{ICA}\right)\right|\right\}\\ &\propto\sum_{t=1}^{T}\sum_{k=1}^{K}\left(\log p\left(s_{1k}(t)\right)+\log p\left(s_{2k}(t)\right)\right) \end{aligned} \tag{3.13}$$

Gradient ascent method was exploited to find an optimised value of $x_1(t)$.

$$\begin{aligned}\frac{\partial L^{ICA}}{\partial x_1(t)} &= \sum_{t=1}^{T}\sum_{k=1}^{K}\left[\frac{\partial \log p\left(s_{1k}(t)\right)}{\partial x_1(t)} + \frac{\partial \log p\left(s_{2k}(t)\right)}{\partial x_1(t)}\right] \\ &= \sum_{t=1}^{T}\sum_{k=1}^{K}\left[\frac{\partial \log p\left(s_{1k}(t)\right)}{\partial p\left(s_{1k}(t)\right)}\frac{\partial p\left(s_{1k}(t)\right)}{\partial x_1(t)} + \frac{\partial \log p\left(s_{2k}(t)\right)}{\partial\left(s_{2k}(t)\right)}\frac{\partial\left(s_{2k}(t)\right)}{\partial x_2(t)}\frac{\partial x_2(t)}{\partial x_1(t)}\right]\end{aligned} \tag{3.14}$$

For case with the two source, the observe mixture is stated $y(t) = a_1 x_1(t) + a_2 x_2(t)$, then At every time *t* every source signal can be expressed by the counterpart, $x_1(t) = \frac{y(t) - a_2 x_2(t)}{a_1}$ and $x_2(t) = \frac{y(t) - a_1 x_1(t)}{a_2}$. Thus, it will result in the differential of $\frac{\partial x_2^t}{\partial x_1^t} = -\frac{a_1}{a_2}$. Equation (3.14) then become,

$$\frac{\partial L^{ICA}}{\partial x_1(t)} = \sum_{n_n=1}^{K}\sum_{k=1}^{K}\left[\varphi\left(s_{1k}\left(t_n\right)\right) w_{1kn_n}^{ICA} - \frac{a_1}{a_2}\varphi\left(s_{2k}\left(t_n\right)\right) w_{2kn_n}^{ICA}\right] \tag{3.15}$$

where $t_n = t - n_n + 1,\ \forall t \in [1,T]$ and $\forall n_n \in [1, N_n]$. Scalar $w_{jkn_n}^{ICA}$ is the n_n th of $\mathbf{w}_{jk}^{ICA}$ which is the component of adjustment in change from k^{th} filter output to source *j*. For $\varphi(s) = \frac{\partial \log p\left(s_{jk}(t)\right)}{\partial s_{jk}(t)}$, it can be obtained from (3.6). Similar formulation are applied to the second source,

$$\frac{\partial L^{ICA}}{\partial x_2(t)} = \sum_{n_n=1}^{K}\sum_{k=1}^{K}\left[-\frac{a_2}{a_1}\left(\varphi\left(s_{1k}(t_n)\right) w_{1kn_n}^{ICA} + \varphi\left(s_{2k}(t_n)\right) w_{2kn_n}^{ICA}\right)\right] \tag{3.16}$$

Then, the update process of the sources can be written as

$$x_j^{(\text{new})}(t) = x_j^{(\text{old})}(t) + \eta \frac{\partial L^{ICA}}{\partial x_j(t)} \tag{3.17}$$

where η is a learning gain.

Next step is to estimate the scaling factors, a_j by finding the maximum the posteriori (MAP) values. From (3.13), L^{ICA} is differentiated with respect to a_1 as follow:

$$\begin{aligned}\frac{\partial L^{ICA}}{\partial a_1} &= \sum_{t=1}^{T}\sum_{k=1}^{K}\left[\frac{\partial \log p\left(s_{1k}(t)\right)}{\partial a_1}+\frac{\partial \log p\left(s_{2k}(t)\right)}{\partial a_1}\right] \\ &= \sum_{t=1}^{T}\sum_{k=1}^{K}\left[\varphi\left(s_{1k}(t)\right)\frac{\partial s_{1k}(t)}{\partial a_1}+\varphi\left(s_{2k}(t)\right)\frac{\partial s_{2k}(t)}{\partial a_2}.\frac{\partial a_2}{\partial a_1}\right] \\ &= \sum_{t=1}^{T}\sum_{k=1}^{K}\left[\varphi\left(s_{1k}(t)\right)\frac{\partial s_{1k}(t)}{\partial a_1}-\varphi\left(s_{2k}(t)\right)\frac{\partial s_{2k}(t)}{\partial a_2}\right] \\ &= \sum_{t=1}^{T}\sum_{k=1}^{K}\left[-\varphi\left(s_{1k}(t)\right)\frac{s_{1k}(t)}{a_1}+\varphi\left(s_{2k}(t)\right)\frac{s_{2k}(t)}{a_2}\right]\end{aligned} \tag{3.18}$$

where the partial derivative of $s_{jk}(t)$ with respect to a_j given by $\frac{\partial s_{jk}(t)}{\partial a_j} = a_j s_{jk}(t).\frac{\partial}{\partial a_j}\left(\frac{1}{a_j}\right) = -\frac{s_{jk}(t)}{a_j}$. The update scaling factor of a_1 can be written as,

$$a_1^{(\text{new})} = h_a\left(a_1^{(\text{old})} + \eta_{a.}\frac{\partial L^{ICA}}{\partial a_1}\right) \tag{3.19}$$

we force the sum of the factors to be constant, such that $a_1 + a_2 = 1$. The value of a_2 is completely dependent on the value of a_1. Thus we can substitute. And for the of the a_2,

$$a_2^{(\text{new})} = 1 - a_1^{(\text{new})} \tag{3.20}$$

where η_a is a learning gain and h_a is a limiting function. Table 3.1 shows the summary of the proposed algorithm.

Table 3.1: Algorithm of two-stage nonlinear SCSS

First stage

Observation: $y(t) = a_1 x_1(t) + a_2 x_2(t)$

Set the initial value of a_1 and a_2 where $\sum_{j=1}^{2} a_j = 1$ and $a_j \neq 0$

For t=1:T,

1. Introduce the post-nonlinearity distortion, *f(.).*, $v(t) = f\left(y(t)\right)$

2 .Find cdf, F_v and inverse cdf of Gaussian distribution, Φ^{-1}

3. Estimate the nonlinearity mapping, $g = \Phi^{-1} \circ F_v$ to find *z(t)*

Second stage

Input*: z(t)*

For t=1:T , k=1:K and j=1,2

1. Compute $\mathbf{s}_j(t) = \mathbf{W}_j^{ICA}\mathbf{x}_j(t)$ and $\varphi(s) = \dfrac{\partial \log p\left(s_{jk}(t)\right)}{\partial s_{jk}(t)}$

2. Update the sources signal, $x_j^{(\text{new})}(t) = x_j^{(\text{old})}(t) + \eta \dfrac{\partial L^{ICA}}{\partial x_j(t)}$

3. Update the scaling factors, $a_1^{(\text{new})} = h_a\left(a_1^{(\text{old})} + \eta_{a.} \dfrac{\partial L^{ICA}}{\partial a_1}\right)$ and $a_2^{(\text{new})} = 1 - a_1^{(\text{new})}$

4. Repeat steps 1 to 3 until convergence

3.3 Results and Analysis

3.3.1 Experiment setup

The proposed separation was tested on recorded audio signals. All recordings and processings were conducted using a PC with Intel Core 2 Duo CPU 5250 @ 1.5GHz and 2GB RAM. For mixture generation, three type of mixtures were used i.e. mixture of piano andflute; mixture of piano andmale speech; mixture of male and female speech. All mixtures are sampled at 16 kHz sampling rate. The objective was to separate the sources in the nonlinear mixture. The values of signal gain a_1 and a_2 both were fixed at 0.5. The number of iterations needed for the algorithm to converge was 200. Independent component analysis (ICA) algorithm (e.g. FastICA) [106] was used to obtain the basis filter $\mathbf{W}_j^{ICA} = \left(\mathbf{M}_j^{ICA}\right)^{-1}$ and source coefficient density was modelled using generalised Gaussian parameter. The basis functions obtained in this simulation was based on [102] which used best characteristic features that being extracted from cross-correlation matrix. Cross correlation was used to identify the characteristically most similar features inherent in the audio signals.

3.3.2 Quality evaluation

In order to get a good representation of errors which may occur in SCSS, several measures are proposed in [109], each investigating a certain property of the error, e.g. interference energy or distortion. The estimated sources is divided as follows

$$\hat{x}(t) = x_{\text{target}}(t) + e_{\text{interf}}(t) + e_{\text{noise}}(t) + e_{\text{artif}}(t) \tag{3.21}$$

where $x_{\text{target}}(t)$ is the target source, and $e_{\text{interf}}(t), e_{\text{noise}}(t)$ and $e_{\text{artif}}(t)$ are the interference i.e unwanted sources, noise and artifact errors, respectively. The decomposition of the estimated source signal was based on orthogonal projections. It was done up to a constant scaling factor. A procedure to calculate the pure source specific energy contained in the separated source signal is described in [109,110]. Performance criteria in decibels are introduced as follow:

1. Source-to-distortion ratio (SDR) – the SDR measures the ratio of the target energy to all unwanted distortions contained in the signal

$$\text{SDR} = 10\log_{10}\frac{\sum_{t=1}^{T}\left\|x_{\text{target}}(t)\right\|^2}{\sum_{t=1}^{T}\left\|e_{\text{interf}}(t)+e_{\text{noise}}(t)+e_{\text{artif}}(t)\right\|^2} \tag{3.22}$$

2. Source-to-interference ratio (SIR) – the SIR measures the ratio between the target source component to all other source components in the mixture. In other words, the residual energy of one source given all others is computed as

$$\text{SIR} = 10\log_{10}\frac{\sum_{t=1}^{T}\left\|x_{\text{target}}(t)\right\|^2}{\sum_{t=1}^{T}\left\|e_{\text{interf}}(t)\right\|^2} \tag{3.23}$$

3. Source-to-artifact ratio (SAR) – the SAR estimates the amount of distortions, defined as

$$\text{SAR} = 10\log_{10}\frac{\sum_{t=1}^{T}\left\|x_{\text{target}}(t)+e_{\text{interf}}(t)+e_{\text{noise}}(t)\right\|^2}{\sum_{t=1}^{T}\left\|e_{\text{artif}}(t)\right\|^2} \tag{3.24}$$

SDR is a global measure that as it accounts for both SIR and SAR. The goal is to maximise SIR (as this measure the actual separation) while trying to keep SAR as high as possible (in order to prevent the introduction of artifacts).

3.3.3 Evaluation of proposed algorithm

In this experiment, the proposed method is evaluated by comparing the performance of linear algorithm (without linearising the nonlinearity stage) and proposed PNL algorithm in nonlinear instantaneous mixture. We will also show the importance of nonlinearity compensation in reducing the distortion in the separated sources.

3.3.3.1 Gaussianization transform

The motivation behind the Gaussianization transform is that the linearly mixed signals before nonlinear transformation are approximately Gaussian distributed due to the Central Limit Theorem. Here, the performance of Gaussianization to compensate the nonlinearity is evaluated. In Figure 3, the histogram of the signal is plotted along with a Gaussian model where its parameters are calculated from the data. The larger the deviation between the histogram and the Gaussian model signifies larger deviation from Gaussianity. Firstly, two non-Gaussian distribution sources (e.g. piano sound and flute sound) as shown in Figure 3.3(a) and 3.3(b) respectively are linearly mixed. Theoretically it tends to be more Gaussian distributed as shown in Figure 3.3(c). The post-nonlinear distortion is applied to the mixed signal using the following nonlinearity of $f(y) = 0.3y + \tanh(3y)$ which is the bounded nonlinear function. This

distortion causes the Gaussian distribution of the mixed signal to deviate from Gaussianity as in Figure 3.3(d). Without using any knowledge of the nonlinearity, the Gaussianization transform inverts the nonlinearly distorted signal and restores the distribution to the Gaussian pdf. The histogram of Gaussianized signal in Figure 3.3(e) proves the performance of the transformation. The line shown in the Figure 3.3 was a histogram of a Gaussian model fit using the mean and variance of a signal. In Figure 3.4, relationship between the signals of $y(t)$, $v(t)$ and $z(t)$ are shown using a scatter plot. A linear relationship between the Gaussianized signal and mixed signal, $y(t)$ can be seen clearly in Figure 3.4(c). From both Figure 3.3 and 3.4, it indicates that the nonlinear distortion has been successfully been linearised using Gaussianization transform.

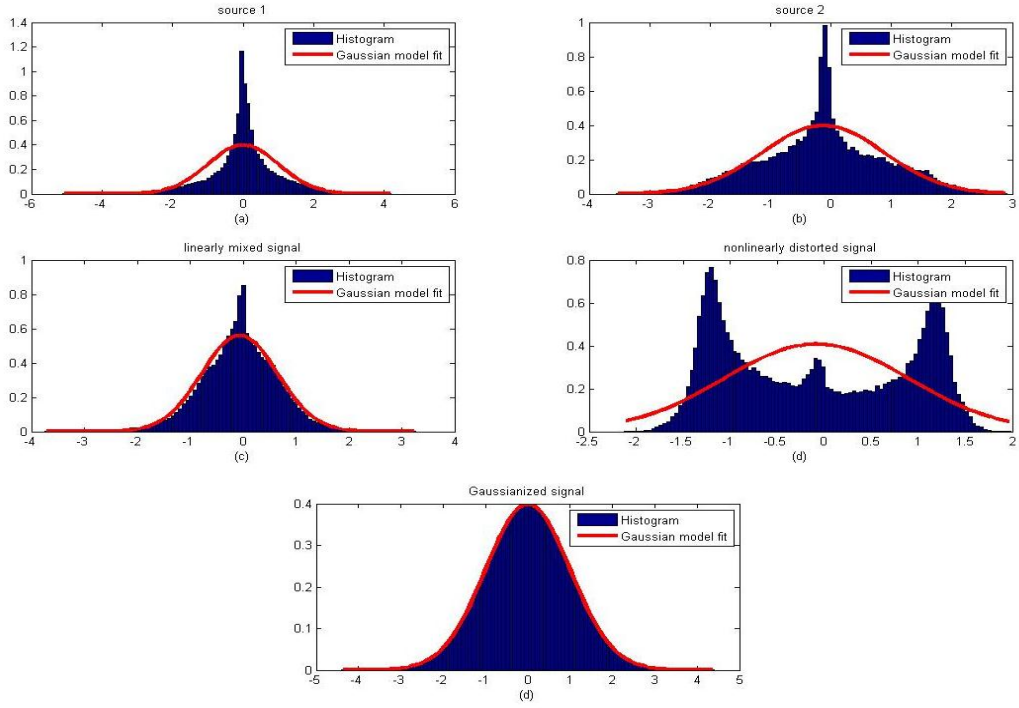


Figure 3.3: Histogram of (a) piano sound (b) flute sound (c) linearly mixed signal (d) nonlinearly distorted signal and (e) Gaussianized signal

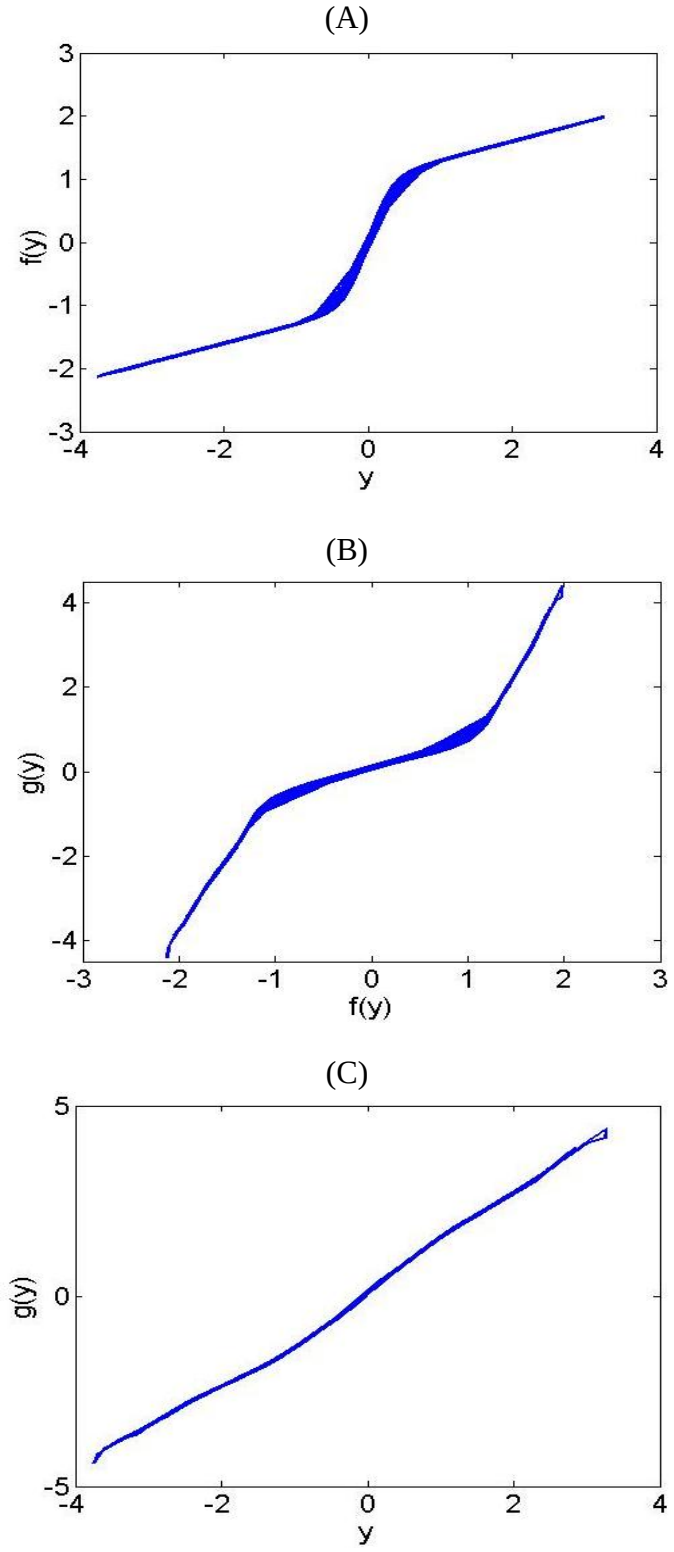


Figure 3.4: Scatter plot of (A) nonlinear functions *f*(.), (B) Inverse function *g*(.) in relation to *f*(.), (C) Gaussianized mixture.

3.3.3.2 Source separation result

Figure 3.5 shows the separation result using proposed algorithm. It can be seen that the proposed algorithm shows the capability to separate the single mixture and recover the piano and flute sound very well in nonlinear mixture. Comparing with the separation result of the linear algorithm in Figure 3.6, without the Gaussianization, the results are affected by the nonlinearity distortion which has resulted in poorer separation. Table 3.2 shows the comparison performance in SDR, SIR and SAR. Note that this is an average result of source 1 and source 2 of the mixture. In the overall, the proposed algorithm shows a very good separation results in a nonlinear environment with the good SDR, SIR and SAR values compared with the linear algorithm for all types of mixture. By using our proposed algorithm, average SDR improvement of 2.2dB, 0.9dB and 2.5dB have been achieved for piano-flute, piano-male and male-female mixture, respectively comparing with linear algorithm. In the proposed algorithm, the Gaussianization transform used to compensate the nonlinearity was proved to be effective and useful in producing improved performance in nonlinear mixture.

In separating the single mixture, the SCSS using ML algorithm achieved a good performance because the proposed algorithm using basis adapted by ICA learning rules [108] i.e. FastICA. This prior information from the basis and their corresponding pdfs are the key to obtaining a faithful MAP based inference algorithm. As in this experiment, the basis obtained using the best characteristic features which is already proved to be better in separation if the less number of bases used [102]. One of the most useful properties is that resulting in decompositions which are often intuitive and

easy to interpret because they are sparse. For single channel using ML approach, the coefficients of the basis functions have the higher degree of sparseness.

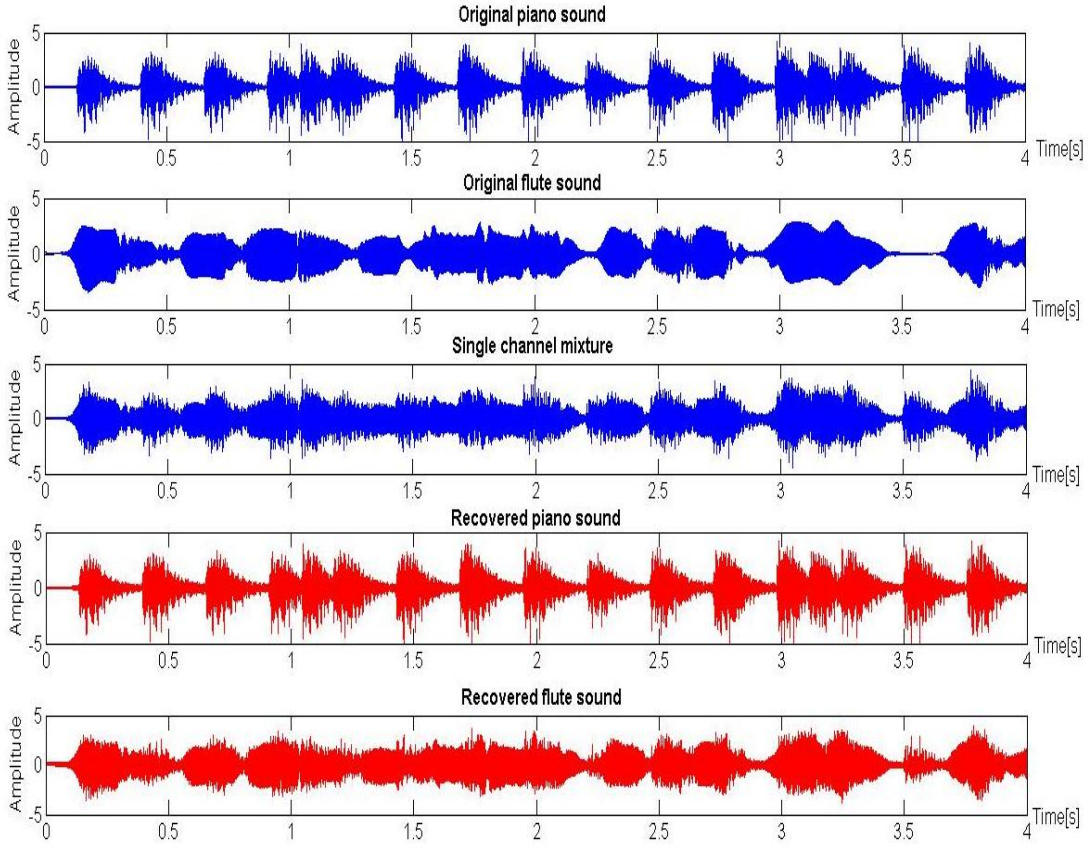


Figure 3.5: Separation results of nonlinear mixture using PNL algorithm

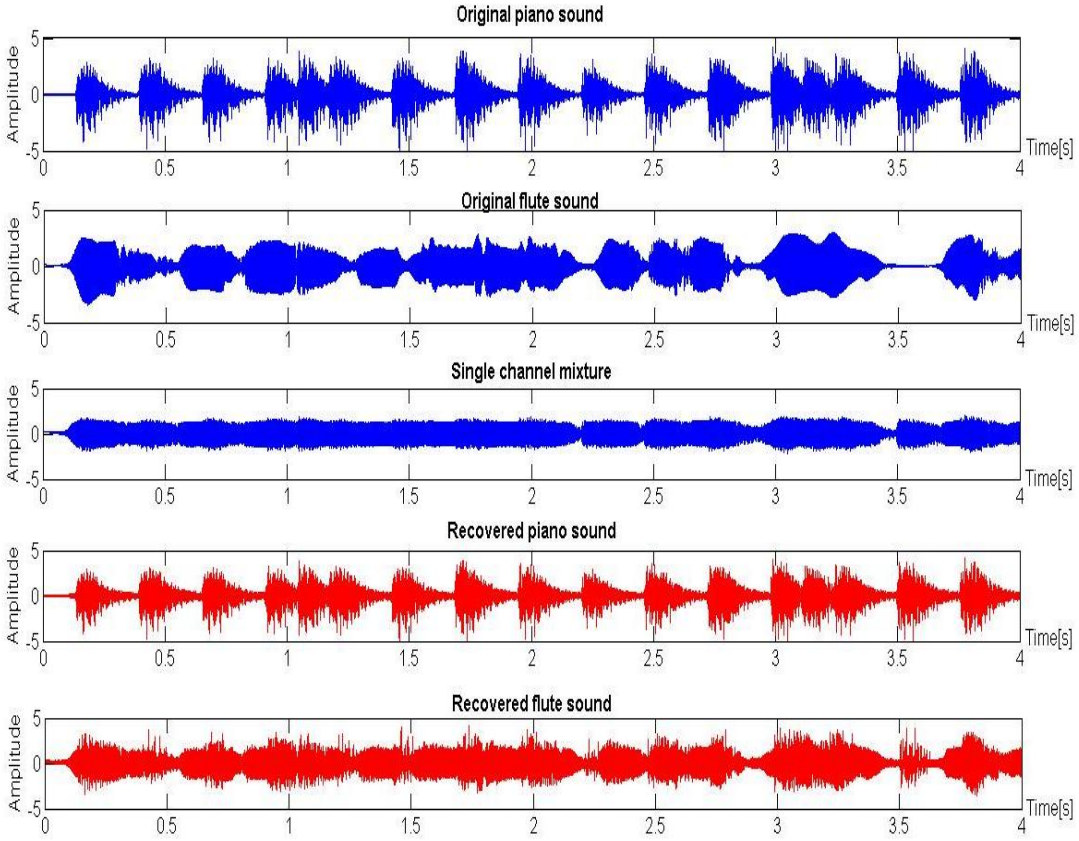


Figure 3.6: Separation results of nonlinear mixture using linear algorithm.

TABLE 3.2: Performance comparison of proposed method with linear algorithm in nonlinear mixture

Mixture	Algorithm	SDR	SIR	SAR
Piano-Flute	Linear	11.0	10.5	19.5
	Proposed method	13.2	14.5	23.2
Piano-Male	Linear	9.5	9.9	16.5
	Proposed method	10.4	13.1	17.4
Male-Female	Linear	9.6	10.2	19.0
	Proposed method	12.1	12.4	21.6

3.3.3.3 Experiment using nonlinear handset model

As mentioned in [95-97], low quality microphones such as those made from carbon materials are subject to nonlinear distortions especially when the input audio is of large amplitude. Applications that involve the use of carbon microphone include mobile handsets and speaker telephones. Consequently, if these microphones are used then the nonlinearity introduced by these microphones cannot be ignored and must be taken into account in the mixing model. In this section, we intend to model the nonlinearity conducted in the study of [95-97] corresponding to carbon-button handset mapper. The nonlinearity is define as a piecewise function with variable boundary points, given by

$$f(y) = \begin{cases} f_+ & for \quad y > y_+ \\ y(j) + y^3(j) & for \quad y_- \le y \le y_+ \\ f_- & for \quad y > y_- \end{cases} \tag{3.25}$$

where the output saturation levels, f_+ and f_- correspond to the input levels y_+ and y_- respectively. From Table 3.3, the results obtained show the same pattern as in Table 3.2 with the performance of PNL algorithm is always shows better performance compared with the linear algorithm. For all type of mixtures, an average of SDR improvement recorded at 1.8dB for the proposed method compare with the linear algorithm. Again, it proves that for nonlinearly distorted signal, by compensating the nonlinearity using Gaussianization, the separation in single channel can yield a good performance.

TABLE 3.3: Performance comparison using polynomial carbon-button nonlinearity

Mixture	Algorithm	SDR	SIR	SAR
Piano-Flute	Linear	10.6	11.2	18.0
	Proposed method	12.0	13.3	22.4
Piano-Male	Linear	8.5	8.9	15.5
	Proposed method	9.0	9.5	16.1
Male-Female	Linear	8.9	9.2	16.9
	Proposed method	10.5	13.7	18.0

3.4 Summary

In this chapter, a new statistical model for the separation of SCSS in post-nonlinear instantaneous has been proposed. The post-nonlinear mixture model is popular not only due to its simplicity in analysis, but also widely applicable. In the proposed techniques, it combines the Gaussianization transform and the time-domain maximum likelihood separation algorithm. In the proposed method, Gaussianization transform inverts the nonlinearly distorted signal and restores the distribution to the Gaussian pdf so that the mixture can be efficiently separated by the linear separation algorithm. The ML approach has been developed to estimate the model parameters and source signal is estimated by MAP approach. The best characteristic features have been used to generate basis function efficiently. From the experiments, the proposed method shows significant performance with high SDR value in nonlinear mixture compare with the one of linear algorithm. Besides that, Gaussianization transform has performed very well in recovering the loss of signal information due to the nonlinearity.

CHAPTER 4

LINEAR SINGLE CHANNEL SOURCE SEPARATION IN CONVOLUTIVE MIXTURE USING QUASI-EM AND MULTIPLICATIVE UPDATE FREQUENCY CONSTRAINED NONNEGATIVE MATRIX FACTORIZATION

Conventional SCSS approaches inherently assume that the original sources have been mixed instantaneously, which in some real applications is not realistic. In addition, they do not exploit the redundancy of channel in an optimal way. For example in audio application, the sound or speech signals received by microphone/receiver are exposed to the reverberations in a room which will degrade the quality and characteristics of sound. Therefore, the assumption that the mixture is instantaneous adopted by the existing SCSS approaches is violated. The aim of this chapter is to remedy these drawbacks and to formulate a SCSS model that accounts for convolutive mixing is proposed. Our work is based on NMF technique which yields a decomposition that may benefit many tasks not only for BSS but to other applications such as in pattern recognition [111], cryptography [112], data mining [113] and binary test data [114]. Several methods have been proposed for SCSS in convolutive mixture problems such as in [115] and [116]. In [115], the convolutive mixture is divided into finite number of subbands using parallel bank of finite impulse

response (FIR) filter before applying empirical mode decomposition (EMD) to produce intrinsic mode function (IMF) in each subband. For this method, the good separation results only can be obtained if basis vectors are statistically independent in each subband. Especially, if the features of the sources are similar, it is difficult to obtain the independent basis vector so that the separation performance will be degraded. In [116], the autocorrelation is used to estimates the delay coefficient in each reflected path and then an artificial stereo mixture is generated from the single channel mixture before applying azimuth discrimination and synthesis (ADRess) algorithm for separation. The difficulty associated with measuring the delay coefficient increased if the time delayed coefficient increased. This is somehow the limitation of the method because the performances of the algorithm rely too heavily on delay estimation coefficient. In NMF-based technique, the authors in [117] have proposed an algorithm for convolutive mixture BSS but the model is focusing on multichannel source separation. In addition, this method requires a posterior binding step so as to group the basis functions according to the sources. Furthermore, the numbers of component per source must be selected *a priori* which may otherwise results in significant degradation in the separation performance. Depending on the type of source signal, the higher the variability of the signal the more components are needed.

In this chapter, a novel development of two-dimensional NMF (NMF2D) using Itakura-Saito (IS) divergence for single channel convolutive mixture is proposed. Two algorithms are proposed using Quasi Expectation-Maximisation (Quasi-EM) and multiplicative update (MU) method. The proposed solutions are an unsupervised method which separates sources from single convolutive channel without using the

training data from the original sources. The first algorithm method based on EM algorithm framework which maximises the log-likelihood of a mixed signals. As for the second algorithm, it is based on the maximum a posteriori (MAP) approach which maximises the joint probability of the mixing channel, spectral basis and temporal codes conditioned on the mixed signal using MU rules. NMF2D model extends the NMF model to be a two-dimensional convolution of **W** and **H**. The factorization is based on a model that represents temporal structure and pitch change which occur when an instrument plays different notes. In audio source separation, the model represents each instrument compactly by a single time-frequency convolved in both time and frequency by a time-pitch weight matrix. This model dramatically decreases the number of components needed to model various instruments and effectively solves the SCSS problem.

In the proposed method, frequency constrained is imposed onto the model in order to compensate for the distortion caused by the convolutive mixing and this model is term as frequency constrained NMF2D (FCNMF2D). The Itakura-Saito (IS) divergence in the proposed algorithms will allows a more precise representation of the factorization where low-energy components cost as much as the high-energy components. This is to be compared with the Kullback-Leibler (KL) and Least Square (LS) divergence whose estimation of low-energy components is often discounted in support of the high-energy components.

The chapter is organised as follows: In Section 4.1, single channel convolutive mixture model in the time-frequency (TF) domain is introduced. In Section 4.2 derivation of two new algorithms of frequency constrained two dimensional sparse NMF is detailed. The results of experimental tests and analysis in feature extraction

and source separation are presented in Section 4.3. Finally, Section 4.4 concludes the chapter.

4.1 Background

4.1.1 Single channel convolutive mixture model

For instantaneous audio mixing in the time domain, the single channel mixture, *y* of the sources, *x* can be model as

$$y(t) = \sum_{j=1}^{J} a_j x_j(t) + e(t) \tag{4.1}$$

where a_j is the mixing filters with t=1,2,...,T and j=1,2,..,J denotes the time index and the number of sources respectively and $e(t)$ is an additive noise. From (4.1), the convolutive mixing model for single channel is described in the following,

$$y(t) = \sum_{j=1}^{J} \sum_{\rho=0}^{L-1} a_j(\rho) x_j(t-\rho) + e(t) \tag{4.2}$$

where $a_j(\rho)$ is the finite-impulse response (FIR) of some causal filters. The time domain convolutive mixture in (4.2) are then projected to time-frequency domain using short-time Fourier transform (STFT) function such that

$$y_{f,n} = \sum_{j=1}^{J} a_{j,f,n} x_{j,f,n} + e_{f,n} \tag{4.3}$$

where $y_{f,n}$ and $x_{j,f,n}$ are the time-frequency components of the corresponding time signals, $a_{j,f,n}$ denotes the value of complex-valued discrete Fourier transform of filter $a_j(\rho)$, f=1,2,...,F and n=1,2,...,N denotes the frequency bin and time frame index respectively. For all ρ, filter length L needs to be shorter than the length of the window used in STFT in order to avoid the similarity effect from the convolution. Assuming the mixing channel is time-invariant, $a_{j,f,n} = a_{j,f}$ and (4.3) can be written as

$$y_{f,n} = \sum_{j=1}^{J} a_{j,f} x_{j,f,n} + e_{f,n} \tag{4.4}$$

Then the power spectrogram is defined as the squared magnitude of (4.4) which can be expressed as

$$\left|y_{f,n}\right|^2 = \sum_{j=1}^{J}\sum_{k=1}^{K} a_{j,f} a_{k,f}^{*} x_{j,f,n} x_{k,f,n}^{*} + 2\sum_{j=1}^{J} \mathrm{Re}\left[a_{j,f} x_{j,f,n} e_{f,n}^{*}\right] + \left|e_{f,n}\right|^2 \tag{4.5}$$

Assuming the windowed disjoint orthogonality (WDO) of the sources and the noise i.e. $x_{j,f,n} x_{k,f,n}^{*} = 0$ and $x_{j,f,n} e_{f,n}^{*} = 0$ for all f and n with $j \neq k$, (4.5) can be written as

$$\left|y_{f,n}\right|^2 = \sum_{j=1}^{J} \left|a_{j,f}\right|^2 \left|x_{j,f,n}\right|^2 + \left|e_{f,n}\right|^2 \tag{4.6}$$

In the matrix form, model (4.6) can rewritten such that

$$\begin{aligned} \left|\mathbf{Y}\right|^{.2} &= \sum_{j=1}^{J} \left|\mathbf{A}_j\right|^{.2} \left|\mathbf{X}_j\right|^{.2} + \left|\mathbf{E}\right|^{.2} \\ &= \sum_{j=1}^{J} \left|\mathbf{X}_j^{im}\right|^{.2} + \left|\mathbf{E}\right|^{.2} \end{aligned} \tag{4.7}$$

where $\left|\mathbf{Y}\right|^{.2} = \left\{\left|y_{f,n}\right|^2\right\}_{fn}$, $\left|\mathbf{X}_j\right|^{.2} = \left\{\left|x_{j,f,n}\right|^2\right\}_{jfn}$, $\left|\mathbf{E}\right|^{.2} = \left\{\left|e_{f,n}\right|^2\right\}_{fn}$ and $\left|\mathbf{A}_j\right|^{.2} = \mathrm{diag}\left(\left|\mathbf{a}_j\right|^{.2}\right)$ with $\left|\mathbf{a}_j\right|^{.2} = \left\{\left|a_{j,f}\right|^2\right\}_f$. The superscript "**.**" is element-wise operation and "diag" is an operator that converts a vector to a diagonal matrix. In single channel blind source separation for convolutive mixture, given the power spectrogram of observed mixture, $\left|\mathbf{Y}\right|^{.2}$, we are interested in estimating the sources image $\mathbf{X}_j^{im}$ and the mixing matrix $\mathbf{A}_j$. The source image is defined as $\mathbf{X}_j^{im} = \mathbf{A}_j\mathbf{X}_j$. Estimating the original sources directly from a single channel convolutive mixture is an ill-posed problem since it requires the inversion of multiple mixing channels from the observed signal *y*(*t*) alone. Subsequently, any estimator obtained in this way will not be statistical consistent and is prone to high levels of discontinuities and artifacts in the estimated sources. In many audio and music processing, filtering the original sources is desirable for enhancing the perceptible quality of signal and in creating an immersive experience [118-120]. In many cases, it suffices to separate the mixture into its constituent parts characterised by the source images (i.e. decomposing the mixture signal into the independent source images) than estimating the original source signals. In addition, the signal-to-distortion ratio (SDR) performance of the estimated source image is better than using the estimated source signal since the algorithm preserves the integrity of signal in the source image more than that of the estimated source signals.

4.1.2 Itakura-Saito divergence properties

Itakura-Saito (IS) divergence was obtained by Itakura and Saito [121] from the maximum likelihood (ML) estimation of short-time speech spectra under autoregressive modeling. The expression of the IS divergence is given by

$$d_{IS}\left(a|b\right)=\frac{a}{b}-\log\frac{a}{b}-1 \tag{4.8}$$

IS divergence is a limit case of the β-divergence [90] which is defined as:

$$D_{\beta}\left(a|b\right)=\begin{cases}\dfrac{1}{\beta\left(\beta-1\right)}\left(a^{\beta}+\left(\beta-1\right)b^{\beta}-\beta ab^{\beta-1}\right)\ ,\ \beta\in\Re\backslash\left\{0,1\right\}\\ a\left(\log a-\log b\right)+\left(b-a\right)\ ,\ \beta=1\\ \dfrac{a}{b}-\log\dfrac{a}{b}-1\ ,\ \beta=0\end{cases} \tag{4.9}$$

In the context of NMF as well, was separately constructed so as to interpolate between the KL divergence (β = 1) and the Euclidean distance (β = 2). The following property holds for any value of β,

$$D_{\beta}\left(\gamma a|\gamma b\right)=\gamma^{\beta}D_{\beta}\left(a|b\right) \tag{4.10}$$

This implies that the IS divergence is scale-invariant i.e. $D_{IS}\left(\gamma a|\gamma b\right)=D_{IS}\left(a|b\right)$ and is the only one of the β-divergence family to possess this property. Scale invariance means that same relative weight is given to small and large coefficients of $\left|\mathbf{Y}\right|^{.2}$ in the sense that a bad fit of the factorization for a low-power coefficient $\left|\mathbf{Y}\right|^{.2}$ will cost as much as a bad fit for a higher-power coefficient $\left|\mathbf{Y}\right|^{.2}$. In contrast, factorizations obtained with $\beta > 0$ (such as with the Euclidean distance or the KL divergence) will rely more heavily on the largest coefficients, and less precision is to be expected in the estimation of the low-power components. The scale invariance of the IS divergence is

relevant to decomposition of audio spectra, which typically exhibit exponential power decrease along the frequency and also usually comprise low-power transient components such as note attacks, together with higher-power components such as tonal parts of sustained.

4.2 Proposed Separation Method

In this section, two new algorithms will be developed, namely the *Quasi-EM FCNMF2D* and the *MU FCNMF2D*. The former algorithm optimises the parameters of the signal model using the Expectation-Maximisation approach whereas the latter is directly based on the multiplicative update rule using gradient descent. To facilitate the derivation of these algorithms, let first consider the signal model in terms of the power TF representation.

4.2.1 Sources Model

Since we are dealing with single channel recording in convolutive mixing, a natural time-frequency domain will be to use the log-frequency spectrogram generated using the constant-Q transform [122]. The log-frequency spectrogram is more suitable compared with classic spectrogram where signals are decomposed to components of linearly spaced frequencies. It is desirable especially for audio application since it provides resolution that is geometrically related to the frequency. By using the constant-Q transform, the twelve-tone equal tempered scale divides each octave into twelve half notes where the frequency ratio between each successive half note is

equal. The frequency of the note which is d_Θ half note above can be written as $f_{d_\Theta} = f_{\text{fund}}.2^{d_\Theta/24}$ where f_{fund} is the fundamental frequency of the note. By imposing the logarithmic scale, this gives $\log f_{d_\Theta} = \log f_{\text{fund}} + \frac{d_\Theta}{24}\log 2$ which shows that in log-frequency spectrogram, the musical octave notes are linearly space or constant.

A new model based on (4.7) is proposed as our mixture model where the case of noise free environment is considered with all elements is non-negative. For the proposed model, the (log-frequency) power spectrogram of each j^{th} source image $\left|\mathbf{X}_j^{im}\right|^{.2}$ is defined as a product of nonnegative matrices of source signal model with frequency constrained as follows:

$$\left|\mathbf{X}_j^{im}\right|^{.2} \approx \sum_{j=1}^{J}\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}} \mathbf{U}_j \overset{\downarrow\phi}{\mathbf{W}_j^{\tau}} \overset{\rightarrow\tau}{\mathbf{H}_j^{\phi}} \tag{4.11}$$

where $\mathbf{U}_j = \text{diag}\left(\left|\mathbf{a}_j\right|^{.2}\right)$ is the frequency constrained of the NMF2D model introduced to resolve the distortion due to the convolutive mixing matrix. $\mathbf{W}_j^{\tau}$ is the j^{th} column of $\mathbf{W}^{\tau}$ which represents the spectral basis of the j^{th} source and $\mathbf{H}_j^{\phi}$ is the j^{th} row of $\mathbf{H}^{\phi}$ which represents the temporal code for each spectral basis element. $\mathbf{W}^{\tau} = \left\{ w_{f,j}^{\tau} \,|\, f = 1,...,F \text{ and } j=1,..,J \right\}$ represents the τ^{th} slice of basis $\mathbf{W}$ and $\mathbf{H}^{\phi} = \left\{ h_{j,n}^{\phi} \,|\, j=1,..,J \text{ and } n=1,...,N \right\}$ represent ϕ^{th} slice of temporal code $\mathbf{H}$. In (4.11), the superscript upper arrow sign in $\overset{\downarrow\phi}{\mathbf{W}_j^{\tau}}$ denotes downward shift operator which moves each element in the matrix by ϕ row down. Concurrently, the arrow sign in $\overset{\rightarrow\tau}{\mathbf{H}_j^{\phi}}$ denotes

the right shift operator which moves each element in the matrix by τ column to the right. The terms $\tau_{\max}$ and $\phi_{\max}$ are the maximum number of τ shifts and ϕ shifts respectively. In the following, two novel algorithms are proposed to estimate the parameter of $\mathbf{U}_j$, $\overset{\downarrow\phi}{\mathbf{W}}_j^{\tau}$ and $\overset{\rightarrow\tau}{\mathbf{H}}_j^{\phi}$ from the mixture.

The task of source separation is to estimate the source image spectrograms as well as the mixing system, $\mathbf{U}$. However, if the solution for the convolutive mixture by using just the NMF2D method [94] without any constraint on the channel, this will introduce error distortion in the obtained solution. To prove this, let us consider the NMF2D as follow and rewrite the expression in terms of the mixing channel:

$$\begin{aligned}\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}}\overset{\downarrow\phi}{\mathbf{W}}_j^{\tau}\overset{\rightarrow\tau}{\mathbf{H}}_j^{\phi} &= \sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}}\overset{\downarrow\phi\rightarrow\phi}{\mathbf{U}}_j\left(\overset{\downarrow\phi\rightarrow\phi}{\mathbf{U}}_j\right)^{-1}\overset{\downarrow\phi}{\mathbf{W}}_j^{\tau}\overset{\rightarrow\tau}{\mathbf{H}}_j^{\phi} \\ &= \sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}}\overset{\downarrow\phi\rightarrow\phi}{\mathbf{U}}_j\,\widehat{\overset{\downarrow\phi}{\mathbf{W}}}_j^{\tau}\overset{\rightarrow\tau}{\mathbf{H}}_j^{\phi}\end{aligned} \tag{4.12}$$

where $\widehat{\overset{\downarrow\phi}{\mathbf{W}}}_j^{\tau} = \left(\overset{\downarrow\phi\rightarrow\phi}{\mathbf{U}}_j\right)^{-1}\overset{\downarrow\phi}{\mathbf{W}}_j^{\tau}$ and $\overset{\downarrow\phi\rightarrow\phi}{\mathbf{U}}_j$ refers to sliding all the elements in $\mathbf{U}_j$ one step along the diagonal direction while the initial resulting empty diagonal component is inserted with 1. Note that $\overset{\downarrow\phi}{\mathbf{W}}_j^{\tau}$ and $\overset{\downarrow\phi\rightarrow\phi}{\mathbf{U}}_j\,\widehat{\overset{\downarrow\phi}{\mathbf{W}}}_j^{\tau}$ give the same set of spectral bases. Let us now compare (4.11) with (4.12) and it is immediately apparent that $\overset{\downarrow\phi\rightarrow\phi}{\mathbf{U}}_j = \mathbf{U}_j$ if and only if $\mathbf{U}_j = \mathbf{I}$ in which case this correspond to an instantaneous or anechoic mixture. However, in convolutive mixture $\mathbf{U}_j \neq \mathbf{I}$ and hence $\overset{\downarrow\phi\rightarrow\phi}{\mathbf{U}}_j \neq \mathbf{U}_j$ in which case the solutions in (4.11) and (4.12) become different. This discrepancy will lead to

distortion in the decomposition of convolutive mixture. From above, it can be argued that on the one hand, $\overset{\downarrow\phi}{\mathbf{W}}_j^{\tau}$ in the NMF2D allows an efficient representation of the source spectral contents which are the desirable attributes in SCSS. On the other hand, because the mixing channels are no longer frequency-flat (i.e. $\mathbf{U}_j \neq \mathbf{I}$), $\overset{\downarrow\phi}{\mathbf{W}}_j^{\tau}$ will inadvertently introduce distortion due to the frequency shifting of the spectral basis (i.e. the ϕ-shift). To remedy this, it is imperative that the frequency distribution of the spectral basis be constrained through the diagonal matrix $\mathbf{U}_j$ as in (4.11) to avoid the distortion generated when $\overset{\downarrow\phi}{\mathbf{W}}_j^{\tau}$ shifts downwards.

From the model in (4.11) two new algorithms are proposed, namely, one based on Quasi-EM framework and the other using multiplicative update rule which will be explained in the next sub-section.

4.2.2 Formulation of Quasi-EM FCNMF2D

Following generative model is considered which is defined as:

$$\mathbf{y}_n = \sum_{k=1}^{K} \mathbf{c}_{k,n} \quad , \quad \forall n = 1,\ldots,N \quad \mathbf{c}_{k,n} = \left[c_{k,1,n},\ldots,c_{k,F,n}\right]^{\mathbf{T}} \tag{4.13}$$

where $\mathbf{y}_n \in \mathbb{C}^{F\times 1}$, $\mathbf{c}_{k,n} \in \mathbb{C}^{F\times 1}$ and $N_c\left(u,\Sigma\right)$ denotes the proper complex Gaussian distribution and the components $\mathbf{c}_{1,n,}\ldots_{,}\mathbf{c}_{K,N}$ are both mutually and individually independent. The Expectation-Maximisation (EM) framework is developed for the ML estimation of $\boldsymbol{\theta} = \left\{\mathbf{U},\mathbf{W}^{\tau},\mathbf{H}^{\phi}\right\}$. Due to the additive structure of the generative model

(4.13), the parameters describing each component $\mathbf{C}_k = \left[\mathbf{c}_{k,1}, \ldots, \mathbf{c}_{k,N}\right]$ can be updated separately. We now consider a partition of the parameter space $\boldsymbol{\theta} = \bigcup_{k=1}^{K} \boldsymbol{\theta}_k$ as $\boldsymbol{\theta}_k = \left\{\mathbf{u}_k, \mathbf{w}_k^{\tau}, \mathbf{h}_k^{\phi}\right\}$ where $\mathbf{w}_k^{\tau}$ is the k^{th} column of $\mathbf{W}^{\tau}$ and $\mathbf{h}_k^{\phi}$ is the k^{th} row of $\mathbf{H}^{\phi}$. The EM algorithm works by formulating the conditional expectation of the negative log likelihood of $\mathbf{C}_k$ as

$$Q_k^{ML}\left(\boldsymbol{\theta}_k \middle| \boldsymbol{\theta}'\right) = -\int_{\mathbf{C}_k} p\left(\mathbf{C}_k \middle| \mathbf{Y}, \boldsymbol{\theta}'\right) \log p\left(\mathbf{C}_k \middle| \boldsymbol{\theta}_k\right) d\mathbf{C}_k \tag{4.14}$$

where $\boldsymbol{\theta}'$ always contains the most recent parameter values of $\left\{\mathbf{U}, \mathbf{W}^{\tau}, \mathbf{H}^{\phi}\right\}$.

4.2.2.1 Expressions of the E- and M-step

One iteration of the EM algorithm includes computing the E-step and maximising the M-step $Q_k^{ML}\left(\boldsymbol{\theta}_k \middle| \boldsymbol{\theta}'\right)$ for $k = 1, \ldots, K$. The minus hidden-data log likelihood is defined as:

$$\begin{aligned} -\log p\left(\mathbf{C}_k \middle| \boldsymbol{\theta}_k\right) &= -\sum_{n=1}^{N}\sum_{f=1}^{F} \log N_c\left(c_{k,f,n} \middle| 0, \sum_{\tau,\phi} u_{k,f} w_{f-\phi,k}^{\tau} h_{k,n-\tau}^{\phi}\right) \\ &\overset{c}{=} \sum_{n=1}^{N}\sum_{f=1}^{F} \log\left(\sum_{\tau,\phi} u_{k,f} w_{f-\phi,k}^{\tau} h_{k,n-\tau}^{\phi}\right) + \frac{\left|c_{k,f,n}\right|^2}{\sum_{\tau,\phi} u_{k,f} w_{f-\phi,k}^{\tau} h_{k,n-\tau}^{\phi}} \end{aligned} \tag{4.15}$$

where '$\overset{c}{=}$' in the second line denotes equality up to constant terms. Then, by virtue of (4.13), the hidden-data posterior also has a Gaussian form as

$p(\mathbf{C}_k \mid \mathbf{Y}, \boldsymbol{\theta}) = \prod_{n=1}^{N} \prod_{f=1}^{F} N_c\left(c_{k,f,n} \middle| \mu_{k,f,n}^{post} \lambda_{k,f,n}^{post}\right)$ where $\mu_{k,f,n}^{post}$ and $\lambda_{k,f,n}^{post}$ are the posterior mean and variance of $c_{k,f,n}$ given as:

$$\mu_{k,f,n}^{post} = \frac{\sum_{\tau,\phi} u_{k,f} w_{f-\phi,k}^{\tau} h_{k,n-\tau}^{\phi}}{\sum_{\tau,\phi,l} u_{l,f} w_{f-\phi,l}^{\tau} h_{l,n-\tau}^{\phi}} y_{f,n}$$

$$\lambda_{k,f,n}^{post} = \frac{\sum_{\tau,\phi} u_{k,f} w_{f-\phi,k}^{\tau} h_{k,n-\tau}^{\phi}}{\sum_{\tau,\phi,l} u_{l,f} w_{f-\phi,l}^{\tau} h_{l,n-\tau}^{\phi}} \sum_{\tau,\phi,l \neq k} u_{l,f} w_{f-\phi,l}^{\tau} h_{l,n-\tau}^{\phi} \tag{4.16}$$

Thus, the E-step merely includes computing the posterior power $\mathbf{V}_k$ of component $\mathbf{C}_k$, defined as $[\mathbf{V}_k]_{f,n} = v_{k,f,n} = \left| \mu_{k,f,n}^{post} \right|^2 + \lambda_{k,f,n}^{post}$. The M-step can be treated as one-component NMF problem:

$$\begin{aligned} Q_k^{ML}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}') &\doteq \sum_{n=1}^{N} \sum_{f=1}^{F} \log\left(\sum_{\tau,\phi} u_{k,f} w_{f-\phi,k}^{\tau} h_{k,n-\tau}^{\phi} \right) + \frac{\left| \mu_{k,f,n}^{post'} \right|^2 + \lambda_{k,f,n}^{post'}}{\sum_{\tau,\phi} u_{k,f} w_{f-\phi,k}^{\tau} h_{k,n-\tau}^{\phi}} \\ &\doteq \sum_{n=1}^{N} \sum_{f=1}^{F} d_{IS}\left(|\mu_{k,f,n}^{post'}|^2 + \lambda_{k,f,n}^{post'} \middle| \sum_{\tau,\phi} u_{k,f} w_{f-\phi,k}^{\tau} h_{k,n-\tau}^{\phi} \right) \end{aligned} \tag{4.17}$$

where $d_{IS}(\cdot \mid \cdot)$ is the IS divergence and is formally defined as $d_{IS}(a \mid b) = (a/b) - \log(a/b) - 1$. The IS divergence has the property of scale invariant i.e. $d_{IS}(\gamma a \mid \gamma b) = d_{IS}(a \mid b)$ for any γ. This implies that any low energy components (a, b) will bear the same relative importance as the high energy ones $(\gamma a, \gamma b)$. This is particularly important to situations where $|\mathbf{Y}|^{\cdot 2}$ is characterised by large dynamic range such as the audio short-term spectra.

4.2.2.2 Estimation of the spectral basis and temporal code

The spectral basis and temporal code can be obtained from (4.17). The derivative of a given element of $g_{k,f,n} = \sum_{\tau,\phi} u_{k,f} w^{\tau}_{f-\phi,k} h^{\phi}_{k,n-\tau}$ with respect to $u_{k,f}$, $w^{\tau}_{f,k}$ and $h^{\phi}_{k,n}$ is given by:

$$\begin{aligned}
\frac{\partial g_{k,f,n}}{\partial u_{k',f'}} &= \frac{\partial \sum_{\tau,\phi} u_{k,f} w^{\tau}_{f-\phi,k} h^{\phi}_{k,n-\tau}}{\partial u_{k',f'}} = \sum_{\tau,\phi} w^{\tau}_{f'-\phi,k'} h^{\phi}_{k',n-\tau} \\
\frac{\partial g_{k,f,n}}{\partial w^{\tau'}_{f',k'}} &= \frac{\partial \sum_{\tau,\phi} u_{k,f} w^{\tau}_{f-\phi,k} h^{\phi}_{k,n-\tau}}{\partial w^{\tau'}_{f',k'}} = u_{k',f} h^{f-f'}_{k',n-\tau'} \quad \text{where } f' = f - \phi \\
\frac{\partial g_{k,f,n}}{\partial h^{\phi'}_{k',n'}} &= \frac{\partial \sum_{\tau,\phi} u_{k,f} w^{\tau}_{f-\phi,k} h^{\phi}_{k,n-\tau}}{\partial h^{\phi'}_{k',n'}} = u_{k',f} w^{n-n'}_{f-\phi',k'} \quad \text{where } n' = n - \tau
\end{aligned} \tag{4.18}$$

The derivatives of (4.17) corresponding to $u_{k,f}$, $w^{\tau}_{f,k}$ and $h^{\phi}_{k,n}$ is then obtained as:

$$\begin{aligned}
\frac{\partial Q^{ML}_{k}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}')}{\partial u_{k',f'}} &= \frac{\partial}{\partial u_{k',f'}} \sum_{f,n} \log\left(g_{k,f,n}\right) + \frac{v'_{k,f,n}}{g_{k,f,n}} \\
&= \sum_{n} \frac{1}{g_{k,f,n}} \sum_{\tau,\phi} w^{\tau}_{f'-\phi,k'} h^{\phi}_{k',n-\tau} + \frac{v'_{k,f,n}}{g^2_{k,f,n}} \sum_{\tau,\phi} w^{\tau}_{f'-\phi,k'} h^{\phi}_{k',n-\tau} \\
&= \sum_{n} \left(\frac{g_{k,f,n} - v'_{k,f,n}}{g^2_{k,f,n}} \right) \sum_{\tau,\phi} w^{\tau}_{f'-\phi,k'} h^{\phi}_{k',n-\tau}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial Q^{ML}_{k}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}')}{\partial w^{\tau'}_{f',k'}} &= \frac{\partial}{\partial w^{\tau'}_{f',k'}} \sum_{f,n} \log\left(g_{k,f,n}\right) + \frac{v'_{k,f,n}}{g_{k,f,n}} \\
&= \sum_{f,n} \frac{1}{g_{k,f,n}} u_{k',f} h^{f-f'}_{k',n-\tau'} + \frac{v'_{k,f,n}}{g^2_{k,f,n}} u_{k',f} h^{f-f'}_{k',n-\tau'} \\
&= \sum_{\phi,n} \left(\frac{g_{k,f'+\phi,n} - v'_{k,f'+\phi,n}}{g^2_{k,f'+\phi,n}} \right) u_{k',f'+\phi} h^{\phi}_{k',n-\tau'}
\end{aligned}$$

$$\begin{aligned}\frac{\partial Q_k^{ML}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}')}{\partial h_{k',n'}^{\phi'}} &= \frac{\partial}{\partial h_{k',n'}^{\phi'}} \sum_{f,n} \log\left(g_{k,f,n}\right) + \frac{v'_{k,f,n}}{g_{k,f,n}} \\ &= \sum_{f,n} \frac{1}{g_{k,f,n}} u_{k',f} w_{f-\phi',k'}^{n-n'} + \frac{v'_{k,f,n}}{g_{k,f,n}^2} u_{k',f} w_{f-\phi',k'}^{n-n'} \\ &= \sum_{\tau,f} \left(\frac{g_{k,f,n'+\tau} - v'_{k,f,n'+\tau}}{g_{k,f,n'+\tau}^2} \right) u_{k',f} w_{f-\phi',k'}^{\tau}\end{aligned} \tag{4.19}$$

Unlike the conventional EM algorithm, it is not possible to directly set $\partial Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}')/w_{f',k'}^{\tau'} = 0$, $\partial Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}')/h_{k',n'}^{\phi'} = 0$ and $\partial Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}')/u_{k',f'} = 0$ because of the nonlinear coupling between $u_{k,f}$, $w_{f,k}^{\tau}$ and $h_{k,n}^{\phi}$ via $v'_{k,f,n}$. Thus, closed form expressions for estimating $u_{k,f}$, $w_{f,k}^{\tau}$ and $h_{k,n}^{\phi}$ cannot be accomplished. To overcome this problem, the following update rules are used and unify it as part of the M-step. For each of individual component, standard gradient descent method is applied with

$$\boldsymbol{\theta}_k \leftarrow \boldsymbol{\theta}_k \cdot \left(\frac{\left[\nabla Q_k^{ML}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}')\right]_-}{\left[\nabla Q_k^{ML}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}')\right]_+} \right) \tag{4.20}$$

where $\nabla Q_k^{ML}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}') = \left[\nabla Q_k^{ML}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}')\right]_+ - \left[\nabla Q_k^{ML}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}')\right]_-$. For each $u_{k,f}$, $w_{f,k}^{\tau}$ and $h_{k,n}^{\phi}$ variables, we have:

$$\begin{aligned}\left[\nabla Q_k^{ML}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}')\right]_-^{\mathbf{U}} &= \sum_n \left(g_{k,f'+\phi,n}\right)^{-2} v'_{k,f'+\phi,n} \sum_{\tau,\phi} w_{f'-\phi,k'}^{\tau} h_{k',n-\tau}^{\phi} \\ \left[\nabla Q_k^{ML}(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}')\right]_+^{\mathbf{U}} &= \sum_n \left(g_{k,f'+\phi,n}\right)^{-1} \sum_{\tau,\phi} w_{f'-\phi,k'}^{\tau} h_{k',n-\tau}^{\phi}\end{aligned} \tag{4.21}$$

and

$$\begin{aligned}\left[\nabla Q_k^{ML}\left(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}'\right)\right]_-^{\mathbf{W}} &= \sum_{\phi,n} u_{k',f'+\phi}\left(g_{k,f'+\phi,n}\right)^{-2} v'_{k,f'+\phi,n} h^{\phi}_{k',n-\tau'} \\ \left[\nabla Q_k^{ML}\left(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}'\right)\right]_+^{\mathbf{W}} &= \sum_{\phi,n} u_{k',f'+\phi}\left(g_{k,f'+\phi,n}\right)^{-1} h^{\phi}_{k',n-\tau'}\end{aligned} \tag{4.22}$$

and

$$\begin{aligned}\left[\nabla Q_k^{ML}\left(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}'\right)\right]_-^{\mathbf{H}} &= \sum_{\tau,f} u_{k',f} w^{\tau}_{f-\phi',k'}\left(g_{k,f,n'+\tau}\right)^{-2} v'_{k,f,n'+\tau} \\ \left[\nabla Q_k^{ML}\left(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}'\right)\right]_+^{\mathbf{H}} &= \sum_{\tau,f} u_{k',f} w^{\tau}_{f-\phi',k'}\left(g_{k,f,n'+\tau}\right)^{-1}\end{aligned} \tag{4.23}$$

Inserting (4.21) into (4.20) leads to

$$u_{k',f'} \leftarrow u_{k',f'}\left(\frac{\sum_n g^{-2}_{k,f,n} v'_{k,f,n} \sum_{\tau,\phi} w^{\tau}_{f'-\phi,k'} h^{\phi}_{k',n-\tau}}{\sum_n g^{-1}_{k,f,n} \sum_{\tau,\phi} w^{\tau}_{f'-\phi,k'} h^{\phi}_{k',n-\tau}}\right) \tag{4.24}$$

Similarly, the updates for $d^{\tau}_{f,k}$ give

$$w^{\tau'}_{f',k'} \leftarrow w^{\tau'}_{f',k'}\left(\frac{\sum_{\phi,n} u_{k',f'+\phi} g^{-2}_{k,f'+\phi,n} v'_{k,f'+\phi,n}\ h^{\phi}_{k',n-\tau'}}{\sum_{\phi,n} u_{k',f'+\phi} g^{-1}_{k,f'+\phi,n} h^{\phi}_{k',n-\tau'}}\right) \tag{4.25}$$

and as for $h^{\phi}_{k,n}$, the update is given by

$$h^{\phi'}_{k',n'} \leftarrow h^{\phi'}_{k',n'}\left(\frac{\sum_{\tau,f} u_{k',f} w^{\tau}_{f-\phi',k'} g^{-2}_{k,f,n'+\tau} v'_{k,f,n'+\tau}}{\sum_{f,\tau} u_{k',f} w^{\tau}_{f-\phi',k'} g^{-1}_{k,f,n'+\tau}}\right) \tag{4.26}$$

It can be verified that the above update rules have an advantage of ensuring the non-negativity constraints of $u_{k,f}$, $w^{\tau}_{f,k}$ and $h^{\phi}_{k,n}$ are always maintained during every

iteration. Defining $[\mathbf{P}_k]_{f,n} = p_{k,f,n} = \sum_{\tau,\phi} w_{f-\phi,k}^{\tau} h_{k,n-\tau}^{\phi}$, the update rules for (4.24), (4.25) and (4.26) can be written in matrix notation as

$$\mathbf{u}_k \leftarrow \mathbf{u}_k \left(\frac{\left(\left(\mathbf{G}_k \right)^{.-2} \cdot \mathbf{V}_k \cdot \mathbf{P}_k \right) \mathbf{1}_{N\times 1}}{\left(\left(\mathbf{G}_k \right)^{.-1} \cdot \mathbf{P}_k \right) \mathbf{1}_{N\times 1}} \right) \tag{4.27}$$

where vector $\mathbf{1}_{N\times 1}$ is a N-vector of ones. For $\mathbf{w}_k^{\tau}$ update, it is written as

$$\mathbf{w}_k^{\tau} \leftarrow \mathbf{w}_k^{\tau} \left(\frac{\sum_{\phi} \mathrm{diag}(\overset{\uparrow\phi}{\mathbf{u}}_k) \left(\left(\overset{\uparrow\phi}{\mathbf{G}}_k \right)^{.-2} \cdot \overset{\uparrow\phi}{\mathbf{V}}_k \right) \overset{\rightarrow\tau^{\mathrm{T}}}{\mathbf{h}_k^{\phi}}}{\sum_{\phi} \mathrm{diag}(\overset{\uparrow\phi}{\mathbf{u}}_k) \left(\overset{\uparrow\phi}{\mathbf{G}}_k \right)^{.-1} \overset{\rightarrow\tau^{\mathrm{T}}}{\mathbf{h}_k^{\phi}}} \right) \tag{4.28}$$

And similarly for $\mathbf{h}_k^{\phi}$

$$\mathbf{h}_k^{\phi} \leftarrow \mathbf{h}_k^{\phi} \left(\frac{\sum_{\tau} \left(\mathrm{diag}(\mathbf{u}_k) \overset{\downarrow\phi}{\mathbf{w}}_k^{\tau} \right)^{\mathrm{T}} \left(\left(\overset{\leftarrow\tau}{\mathbf{G}}_k \right)^{.-2} \cdot \overset{\leftarrow\tau}{\mathbf{V}}_k \right)}{\sum_{\tau} \left(\mathrm{diag}(\mathbf{u}_k) \overset{\downarrow\phi}{\mathbf{w}}_k^{\tau} \right)^{\mathrm{T}} \left(\overset{\leftarrow\tau}{\mathbf{G}}_k \right)^{.-1}} \right) \tag{4.29}$$

Table 4.1 presents the main steps of the proposed method for Quasi-EM FCNMF2D where $\psi = 10^{-6}$ is the threshold for ascertaining the convergence.

Table 4.1: Quasi-EM FCNMF2D algorithm

Input: $|\mathbf{Y}|^{.2}$

Output: $\mathbf{U}$, $\mathbf{W}^{\tau}$ and $\mathbf{H}^{\phi}$.

Initialise $\mathbf{U}$, $\mathbf{W}^{\tau}$ and $\mathbf{H}^{\phi}$ with nonnegative random values.

Compute initialise cost value *cost*(1) using (4.15)

for *iter=1:no. of iterations*

for *k=1:K*

E-step

-Compute $v_{k,f,n} = \left| \mu_{k,f,n}^{post} \right|^2 + \lambda_{k,f,n}^{post}$ using (4.16)

M-step

- Update $\mathbf{u}_k$ using (4.27) for all τ and ϕ

- Update $\mathbf{w}_k^{\tau}$ using (4.28) for all τ and ϕ and normalise $\mathbf{d}_k^{\tau}$

- Update $\mathbf{h}_k^{\phi}$ using (4.29) for all τ and ϕ and normalise $\mathbf{h}_k^{\phi}$

end

Compute cost value using (4.15)

end

stopping criterion: $\frac{cost(iter-1) - cost(iter)}{cost(iter)} < \psi$

4.2.3 Formulation of Multiplicative Update FCNMF2D

We consider the following generative model defined as:

$$|\mathbf{Y}|^{.2} = \left(\sum_{j=1}^{J} \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{U}_j \overset{\downarrow\phi}{\mathbf{W}_j^{\tau}} \overset{\rightarrow\tau}{\mathbf{H}_j^{\phi}} \right) \bullet \mathbf{E} \qquad (4.30)$$

where $\mathbf{E}$ is a scalar of multiplicative independent and identically-distributed (i.i.d.) Gamma noise with unit mean i.e. $p(\mathbf{E}_{f,n}) = \xi(\mathbf{E}_{f,n} \mid \alpha, \beta)$ where $\xi(\mathbf{E}_{f,n} \mid \alpha, \beta)$ denotes the Gamma probability density function (pdf) defined as:

$\xi(\mathbf{E}_{f,n} \mid \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \left(\mathbf{E}_{f,n}\right)^{\alpha-1} \exp\left(-\beta \mathbf{E}_{f,n}\right), \mathbf{E}_{f,n} \geq 0$. Next, we define $\mathbf{U} = \left[\mathbf{U}_1 \mathbf{U}_2 \cdots \mathbf{U}_I\right]$ $\mathbf{W} = \left[\mathbf{W}^1 \mathbf{W}^2 \cdots \mathbf{W}^{\tau_{\max}}\right]$ and $\mathbf{H} = \left[\mathbf{H}^1 \mathbf{H}^2 \cdots \mathbf{H}^{\phi_{\max}}\right]$. Under the independent and identically distributed (i.i.d.) noise assumption, the term $-\log p\left(\left|\mathbf{Y}\right|^{\cdot 2} \mid \mathbf{U}, \mathbf{W}, \mathbf{H}\right)$ becomes

$$-\log p\left(\left|\mathbf{Y}\right|^{\cdot 2} \mid \mathbf{U}, \mathbf{W}, \mathbf{H}\right) = \frac{-\sum_{n=1}^{N}\sum_{f=1}^{F} \log \xi\left(\frac{|\mathbf{Y}|_{f,n}^{\cdot 2}}{\sum_{j=1}^{J}\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}} \mathbf{U}_{j,f} \overset{\downarrow\phi}{\mathbf{W}}{}^{\tau}_{f,j} \overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi}_{j,n}} \middle| \alpha, \beta \right)}{\sum_{j=1}^{I}\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}} \mathbf{U}_{j,f} \overset{\downarrow\phi}{\mathbf{W}}{}^{\tau}_{f,j} \overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi}_{j,n}}$$

$$\overset{c}{=} d_{IS}\left(|\mathbf{Y}|_{f,n}^{2} \middle| \sum_{j=1}^{J}\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}} \mathbf{U}_{j,f} \overset{\downarrow\phi}{\mathbf{W}}{}^{\tau}_{f,j} \overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi}_{j,n} \right) \qquad (4.31)$$

where '$\overset{c}{=}$' in the second line denotes equality up to constant terms. Thus, the cost function is $C_{IS} = -\log p\left(\left|\mathbf{Y}\right|^{\cdot 2} \mid \mathbf{U}, \mathbf{W}, \mathbf{H}\right)$. The derivatives of (4.31) corresponding to $\mathbf{U}$, $\mathbf{W}^{\tau}$ and $\mathbf{H}^{\phi}$ are described as follows:

For $u_{j,f}$, the derivatives is given by

$$\begin{aligned}\frac{\partial C_{IS}}{\partial u_{j,f}} &= -\sum_{n}\left|y_{f,n}\right|^{2} z_{f,n}^{-2}\sum_{f,n} w_{f-\phi,j}^{n-n'} h_{j,n-\tau}^{f-f'} + \sum_{n} z_{f,n}^{-1}\sum_{f,n} w_{f-\phi,j}^{n-n'} h_{j,n-\tau}^{f-f'} \\ &= -\sum_{n}\left|y_{f,n}\right|^{2} z_{f,n}^{-2}\sum_{\phi,\tau} w_{f-\phi,j}^{\tau} h_{j,n-\tau}^{\phi} + \sum_{n} z_{f,n}^{-1}\sum_{\phi,\tau} w_{f-\phi,j}^{\tau} h_{j,n-\tau}^{\phi}\end{aligned} \tag{4.32}$$

where $z_{f,n} = \sum_{j}\sum_{\tau}\sum_{\phi} u_{j,f} w_{f-\phi,j}^{\tau} h_{j,n-\tau}^{\phi}$. As for $w_{f,j}^{\tau}$ and $h_{j,n}^{\phi}$, the derivative of the component are given by

$$\begin{aligned}\frac{\partial C_{IS}}{\partial w_{f',j'}^{\tau'}} &= -\sum_{f,n}\left|y_{f,n}\right|^{2} z_{f,n}^{-2}\sum_{f,n} u_{j',f} h_{j',n-\tau'}^{f-f'} + \sum_{f,n} z_{f,n}^{-1}\sum_{f,n} u_{j',f} h_{j',n-\tau'}^{f-f'} \\ &= -\sum_{\phi,n}\left|y_{f'+\phi,n}\right|^{2} z_{f'+\phi,n}^{-2} u_{j',f'+\phi} h_{j',n-\tau'}^{\phi} + \sum_{\phi,n} \tilde{v}_{f'+\phi,n}^{-1} u_{j',f'+\phi} h_{j',n-\tau'}^{\phi},\end{aligned} \tag{4.33}$$

Similarly,

$$\begin{aligned}\frac{\partial \tilde{C}_{IS}}{\partial h_{j',n'}^{\phi'}} &= -\sum_{f,n}\left|y_{f,n}\right|^{2} z_{f,n}^{-2}\sum_{f,n} u_{j',f} \tilde{w}_{f-\phi',j'}^{n-n'} + \sum_{f,n} z_{f,n}^{-1}\sum_{f,n} u_{j',f} w_{f-\phi',j'}^{n-n'} \\ &= -\sum_{f,\tau}\left|y_{f,n'+\tau}\right|^{2} z_{f,n'+\tau}^{-2} u_{j',f} w_{f-\phi',j'}^{\tau} + \sum_{f,\tau} z_{f,n'+\tau}^{-1} u_{j',f} w_{f-\phi',j'}^{\tau}\end{aligned} \tag{4.34}$$

For each of individual component, standard gradient descent method is applied with

$$u_{j,f} \leftarrow u_{j,f} - \eta_u \frac{\partial C_{IS}}{\partial u_{j,f}},\quad w_{f',j'}^{\tau'} \leftarrow w_{f',j'}^{\tau'} - \eta_w \frac{\partial C_{IS}}{\partial w_{f',j'}^{\tau'}} \quad \text{and} \quad h_{j',n'}^{\phi'} \leftarrow h_{j',n'}^{\phi'} - \eta_h \frac{\partial C_{IS}}{\partial h_{j',n'}^{\phi'}} \tag{4.35}$$

where η_u, η_w, and η_h are the positive learning rate. Based on [84], the positive learning rate can be set to the followings:

$$\eta_u = \frac{u_{j,f}}{\sum_{n} z_{f,n}^{-1}\sum_{\phi,\tau} w_{f-\phi,j}^{\tau} h_{j,n-\tau}^{\phi}},\quad \eta_w = \frac{w_{f',j'}^{\tau'}}{\sum_{\phi,n} z_{f'+\phi,n}^{-1} u_{j',f'+\phi} h_{j',n-\tau'}^{\phi}}$$
$$\text{and}\quad \eta_h = \frac{h_{j',n'}^{\phi'}}{\sum_{f,\tau} \tilde{v}_{f,n'+\tau}^{-1} u_{j',f} \tilde{w}_{f-\phi',j'}^{\tau}} \tag{4.36}$$

Using (4.35) and (4.36), the multiplicative update (MU) rules are obtained where for $u_{j,f}$, the update is given by

$$u_{j,f} \leftarrow u_{j,f} \left(\frac{\sum_n \left| y_{f,n} \right|^2 z_{f,n}^{-2} \sum_{\phi,\tau} w_{f-\phi,j}^{\tau} h_{j,n-\tau}^{\phi}}{\sum_n z_{f,n}^{-1} \sum_{\phi,\tau} w_{f-\phi,j}^{\tau} h_{j,n-\tau}^{\phi}} \right) \tag{4.37}$$

Similarly, the MU rules for $w_{f,j}^{\tau}$ gives

$$w_{f',j'}^{\tau'} \leftarrow w_{f',j'}^{\tau'} \left(\frac{\sum_{\phi,n} \left| y_{f'+\phi,n} \right|^2 z_{f'+\phi,n}^{-2} u_{j',f'+\phi} h_{j',n-\tau'}^{\phi}}{\sum_{\phi,n} z_{f'+\phi,n}^{-1} u_{j',f'+\phi} h_{j',n-\tau'}^{\phi}} \right) \tag{4.38}$$

and as for $h_{j,n}^{\phi}$, the update is given by

$$h_{j',n'}^{\phi'} \leftarrow h_{j',n'}^{\phi'} \left(\frac{\sum_{f,\tau} \left| y_{f,n'+\tau} \right|^2 z_{f,n'+\tau}^{-2} u_{j',f} w_{f-\phi',j'}^{\tau}}{\sum_{f,\tau} z_{f,n'+\tau}^{-1} u_{j',f} w_{f-\phi',j'}^{\tau}} \right) \tag{4.39}$$

In the matrix notation, the update rules for (4.37), (4.38) and (4.39) can be written as

$$\mathbf{u}_j \leftarrow \mathbf{u}_j . \frac{\left(\mathbf{Z}^{\cdot -2} . \left| \mathbf{Y} \right|^{.2} . \mathbf{P}_j \right) \mathbf{1}_{N \times 1}}{\left(\mathbf{Z}^{\cdot -1} . \mathbf{P}_j \right) \mathbf{1}_{N \times 1}} \tag{4.40}$$

where $\mathbf{P}_j = \sum_{\tau,\phi} \mathbf{W}_{f,j}^{\tau \downarrow \phi} \mathbf{H}_{j,n}^{\phi \rightarrow \tau}$ and vector $\mathbf{1}_{N \times 1}$ is a *N*-vector of ones. For **W** update, it is written as

$$\mathbf{W}_j^{\tau} \leftarrow \mathbf{W}_j^{\tau} . \left(\frac{\sum_{\phi} \mathrm{diag}(\overset{\uparrow\phi}{\mathbf{u}}_j) \left(\left(\overset{\uparrow\phi}{\mathbf{Z}} \right)^{.-2} . \left| \overset{\uparrow\phi}{\mathbf{Y}} \right|^{.2} \right) \mathbf{H}_j^{\phi \rightarrow\tau\,\mathrm{T}}}{\sum_{\phi} \mathrm{diag}(\overset{\uparrow\phi}{\mathbf{u}}_j) \left(\overset{\uparrow\phi}{\mathbf{Z}} \right)^{.-1} \mathbf{H}_j^{\phi \rightarrow\tau\,\mathrm{T}}} \right) \tag{4.41}$$

And similarly for **H,**

$$\mathbf{H}_j^{\phi} \leftarrow \mathbf{H}_j^{\phi} . \frac{\sum_{\tau} \left(\mathrm{diag}(\mathbf{u}_j) \overset{\downarrow\phi}{\mathbf{W}}_j^{\tau} \right)^{\mathrm{T}} \left(\left(\overset{\leftarrow\tau}{\mathbf{Z}} \right)^{.-2} . \left| \overset{\leftarrow\tau}{\mathbf{Y}} \right|^{.2} \right)}{\sum_{\tau} \left(\mathrm{diag}(\mathbf{u}_j) \overset{\downarrow\phi}{\mathbf{W}}_j^{\tau} \right)^{\mathrm{T}} \left(\overset{\leftarrow\tau}{\mathbf{Z}} \right)^{.-1}} \tag{4.42}$$

Table 4.2 presents the main steps of the proposed method for MU FCNMF2D .

Table 4.2: MU FCNMF2D algorithm

Input: $\left|\mathbf{Y}\right|^{.2}$

Output: $\mathbf{U}$, $\mathbf{W}^{\tau}$ and $\mathbf{H}^{\phi}$

Initialise $\mathbf{U}$, $\mathbf{W}^{\tau}$ and $\mathbf{H}^{\phi}$ with nonnegative random values.

Compute initialise cost value *cost*(1) using (4.31)

for *iter=1:no. of iterations*

Compute $p_{f,n} = \sum_{j,\tau,\phi} w_{f-\phi,j}^{\tau} h_{j,n-\tau}^{\phi}$ and $z_{f,n} = \sum_{j,\tau,\phi} u_{j,f} w_{f-\phi,j}^{\tau} h_{j,n-\tau}^{\phi}$

-Update $\mathbf{u}_j$ using (4.40).

-Update $\mathbf{W}_j^{\tau}$ using (4.41) for all τ and ϕ and normalise $\mathbf{W}_j^{\tau}$

-Update $\mathbf{H}_j^{\phi}$ using (4.42) for all τ and ϕ and normalise $\mathbf{H}_j^{\phi}$

Compute cost value using (4.31).

end

stopping criterion: $\frac{cost(iter-1) - cost(iter)}{cost(iter)} < \psi$

For Quasi-EM FCNMF2D algorithm, it avoids zeros in the factors where $\mathbf{W}^{\tau}$ and $\mathbf{H}^{\phi}$ cannot take entries equal to zero. If $w_{f,k}^{\tau}$ or $h_{k,n}^{\phi}$ is zero, it will give infinite value to the cost function. On the other hand, this is not feature shared by the MU FCNMF2D algorithm, which does not priori exclude zero coefficients in $\mathbf{W}^{\tau}$ and $\mathbf{H}^{\phi}$ (except for $\mathbf{Z}=0$ which lead to a division by zero). Since zero coefficients are invariant under MU FCNMF2D. If the MU FCNMF2D algorithm attains a fixed point solution with zero entries, then it cannot be determined since the limit point is a stationary point [87, 123]. Consequently, the resulting factorizations rendered by these algorithms are not equivalent. This is the drawback of MU update, once a parameter is exactly zero, it remains zero. For this reason, the Quasi-EM FCNMF2D algorithm can be considered more reliable for updating $\mathbf{W}^{\tau}$ and $\mathbf{H}^{\phi}$.

4.3 Results and Analysis

The proposed algorithms were tested on two applications which were feature extraction and source separation of audio signals. For feature extraction, the objective was to extract the basis and code from the convolutive mixed data using a toy data. The toy data mixture was generated using round and cross patterns which overlap in TF domain. It was artificially generated using MATLAB. As for source separation, the proposed algorithm were tested for real application where the objective was to separate an audio mixture in convolutive mixture. All experiments and analysis were performed using a PC with Intel core i3 M380 @ 2.53GHz and 3GB RAM.

4.3.1 Feature extraction of toy data

In this sub-section, the proposed algorithms will be tested to evaluate their ability in extracting the basis and code from a simulated mixed data. The simulated data were generated to have high degree of pattern overlap which occupied area of low frequency, middle frequency and high frequency in the log-frequency spectrogram. Figure 4.1 shows the true factors of basis in vertical panels and code in horizontal panels of the simulated mixed patterns. The basis **W** consists of three circles and three crosses features. These features were convolved with the code **H** given at the top panels of the figure to yield the mixture of both patterns. Then, hamming window was applied to the mixture to produce a convolutive mixture of the data matrix **Y**. For this experiment the convolutive factors were selected such that $\tau = \{0,\ldots,16\}$ and $\phi = \{0,\ldots,16\}$. In this experiment, we will compare the performance between Quasi-EM FCNMF2D and MU FCNMF2D and investigate the effects of frequency constrained, **U** for both algorithms. To observe effect of **U**, we will evaluate the proposed algorithm where **U** is constant by simply setting $\mathbf{U}_j = \mathbf{I}$ while the adaptation of **W** and $\mathbf{H}$ still follows the proposed methodology.

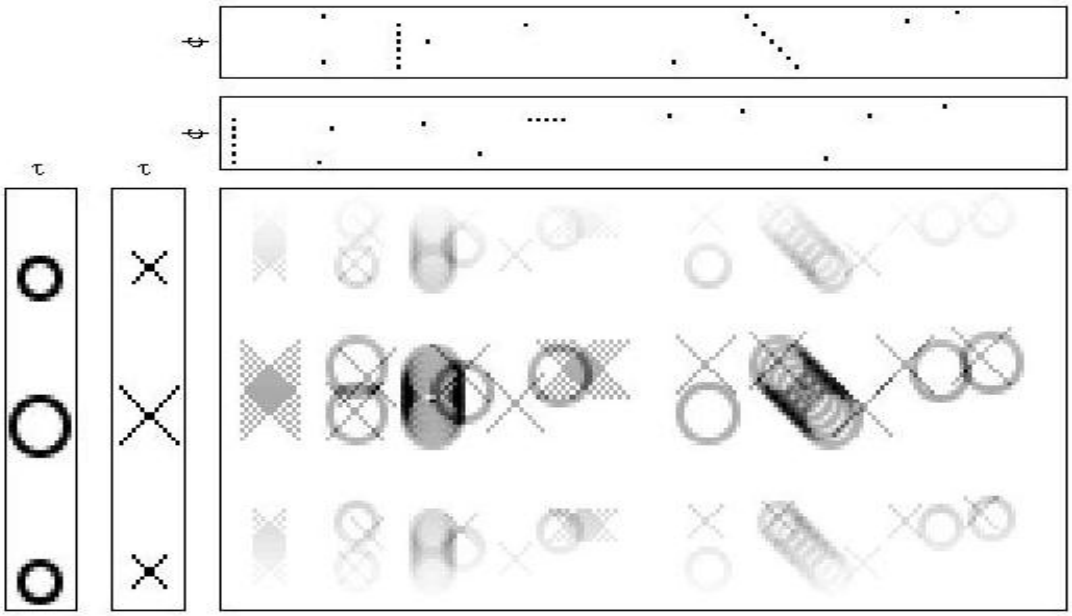


Figure 4.1: True factor of basis and code of the simulated convolutive mixed data

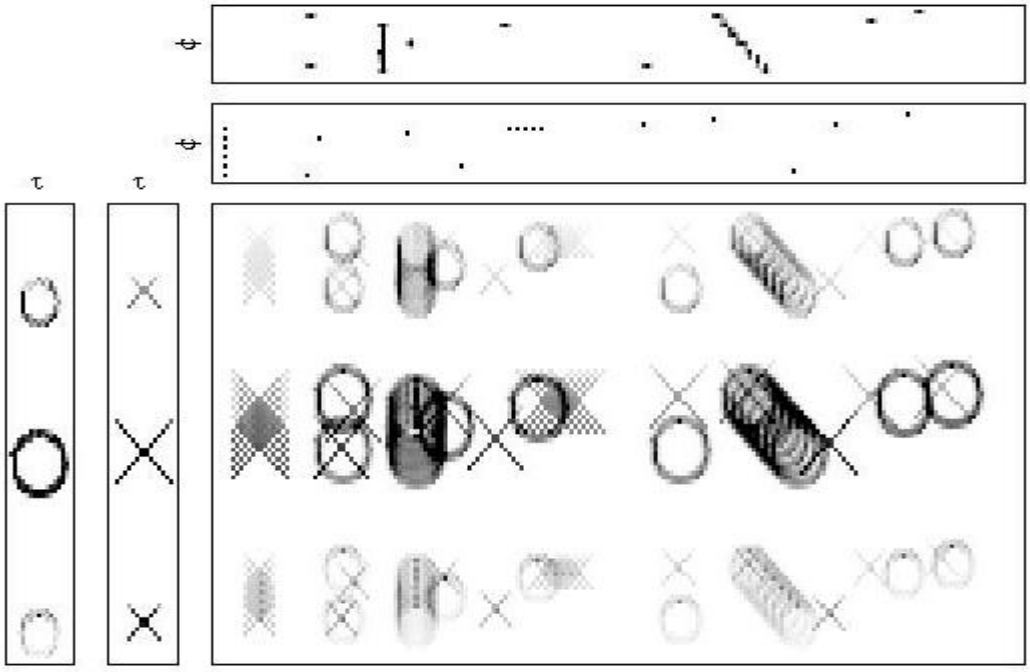


Figure 4.2: The estimated results using Quasi-EM FCNMF2D

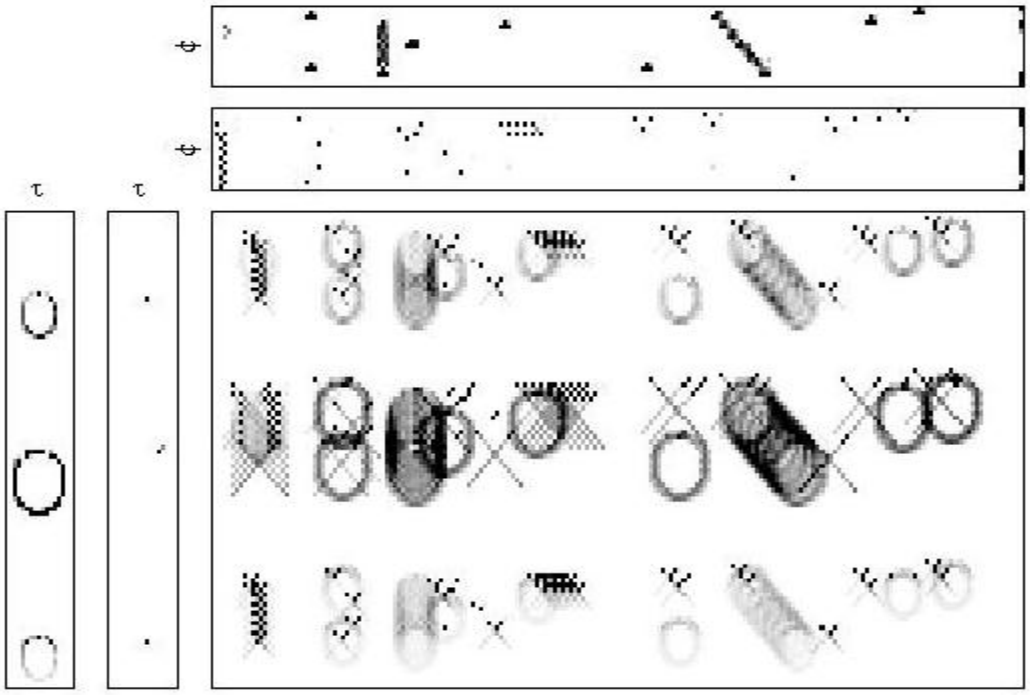


Figure 4.3: The estimated results using MU FCNMF2D

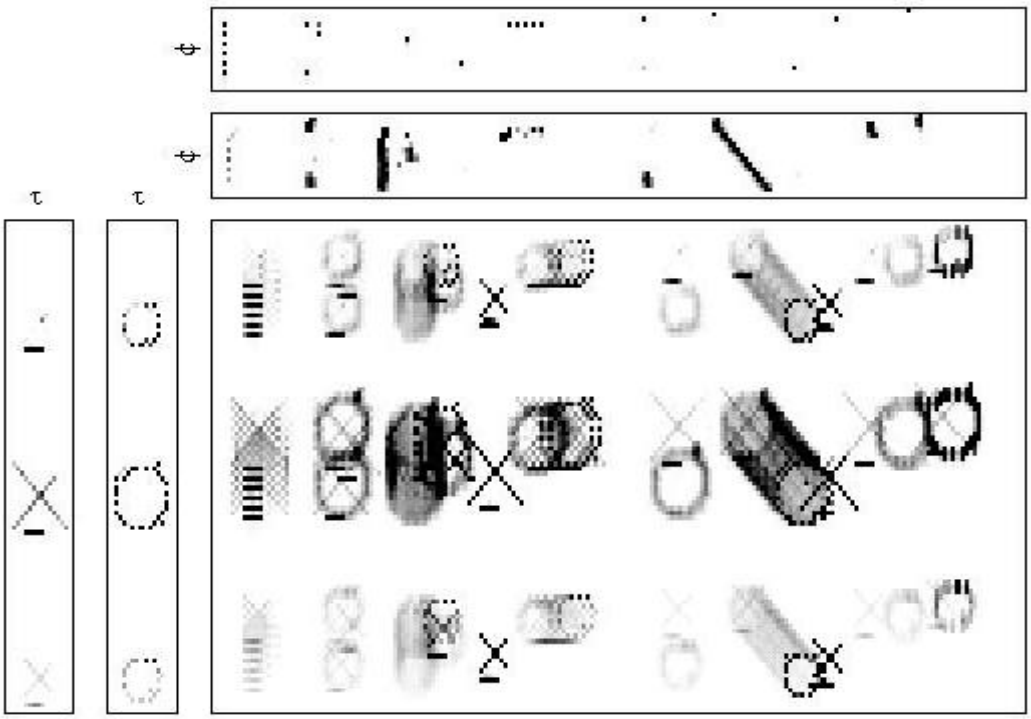


Figure 4.4: The estimated results using Quasi-EM NMF2D (**U**=**I**)

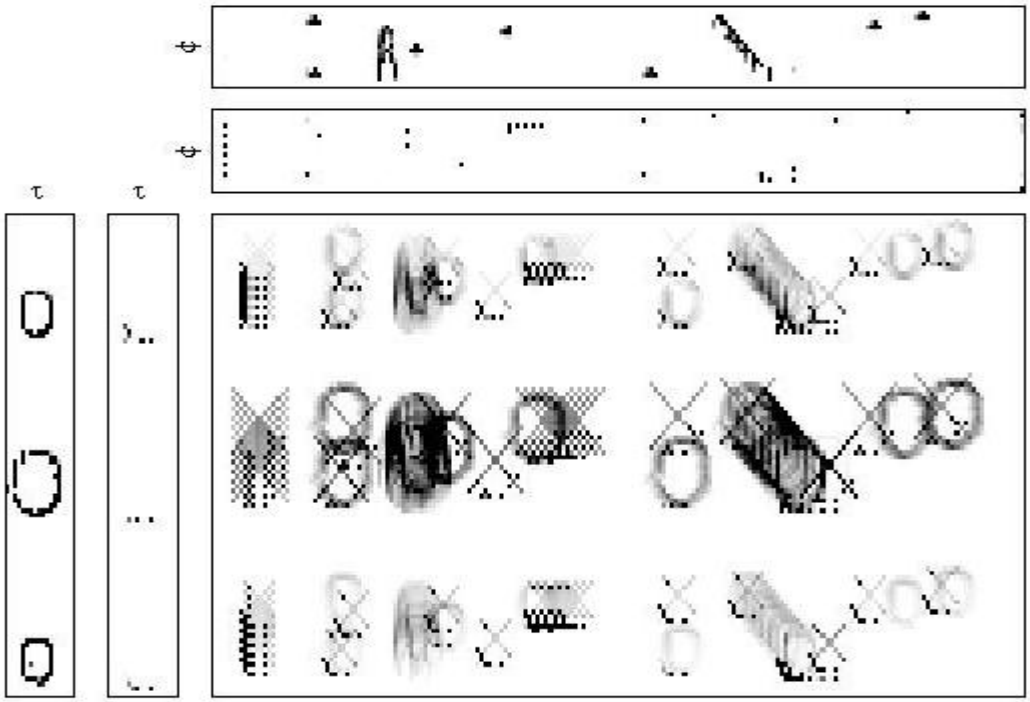


Figure 4.5: The estimated results using MU NMF2D (**U**=**I**)

Figure 4.2 and 4.3 show the matrix factorization results for both Quasi-EM FCNMF2D and MU FCNMF2D, respectively. It can be seen from the figures that the factorization using Quasi-EM FCNMF2D show better extraction and reconstruction performance compare with MU FCNMF2D. Every code was assign almost correctly to each basis feature and the estimation of convolutive mixing parameter has shown a good outcomes. As for MU FCNMF2D, the poorer performance indicates that the MU FCNMF2D could be trapped in local minima which will result incorrect extraction of codes and thereby causing the crosses and circles features missing from the figure. Figure 4.4 and 4.5 show the matrix factorization when **U**=**I** for both Quasi-EM and MU of proposed algorithms. It clearly shows that the feature extraction for both case cannot fully recovered the basis features and the codes accordingly. This is due to the assumption that **U** has the uniform value which is not true for convolutive mixture.

4.3.2 Blind source separation

In this sub-section, the proposed method is tested on audio signals. To generate mixed signal, polyphonic music containing piano and trumpet was analysed. The mixture was approximately 5s long and sampled at 16kHz. In this experiment, STFT using 2048-point Hamming window with 50% overlap was used and this gave 175 frequency bins in the log-frequency spectrogram within the range of 50Hz to 8kHz with 24 bins per octave. This corresponded to twice the resolution of the equal source signal scale.

We generated synthetic convolutive mixture of the sources using the Room Simulation (Roomsim) toolbox [124]. In this experiment, Roomsim simulates an omnidirectional microphone in a room of dimension 4.45m x 3.00m x 3.00m. The receiver, source 1 and source 2 are located 1.2m from the floor. The distance between the sources and the microphone is 2m with source 1 and source 2 located 1m between each other. The Roomsim toolbox integrates frequency dependent absorption at the reflective surfaces and in the airspace of the room to model the reverberation at a receiver. After considering the appropriate setting of surface absorption for each surface that model the actual room, the impulse response for source 1 and source 2 in our system is created as shown in Figure 4.6. The reverberation time (RT60) for the 1000Hz band was calculated to be 0.35s.

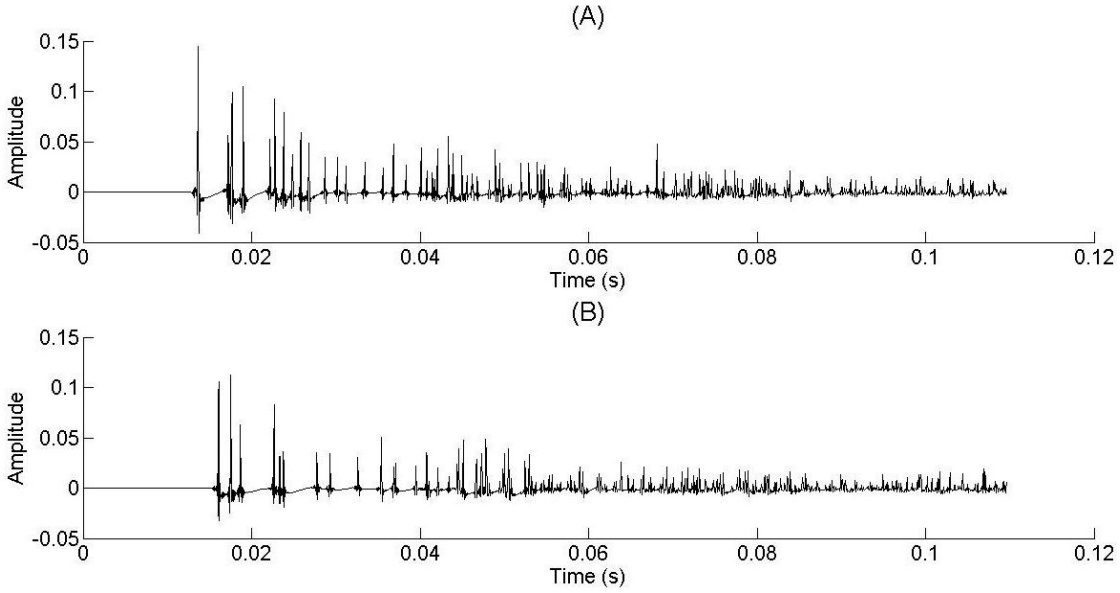


Figure 4.6: Impulse response of: (A) channel 1 and (B) channel 2

For audio separation, after conducting the Monte-Carlo experiments over 100 independent realizations of the mixture, the parameters of the convolutive factors of τ and ϕ shifts are set to be $\tau_{\max} = 8$ and $\phi_{\max} = 32$. This is the best attainable parameter setting to represent the temporal code and spectral basis in the factorization for most of music signals since audio signal have higher variability and require higher number of τ-shift and ϕ-shift to capture the temporal dependency of the frequency pattern in audio signal. Figure 4.7 shows the original TF domains of the source images of piano and trumpet as well as its convolutive mixture. The piano and trumpet play a different short melodic with a different distinct note. We can see that both piano and trumpet overlap in time while the piano notes are scattered and interspersed between frequencies with the trumpet notes. To evaluate the proposed algorithm, the performance will be measured using the signal-to-distortion ratio (SDR), source-to-artifacts ratio (SAR) and source-to-interference ratio (SIR) which measures an overall

sound quality of the source separation. The MATLAB implementation of these measurements can be found in [109, 110].

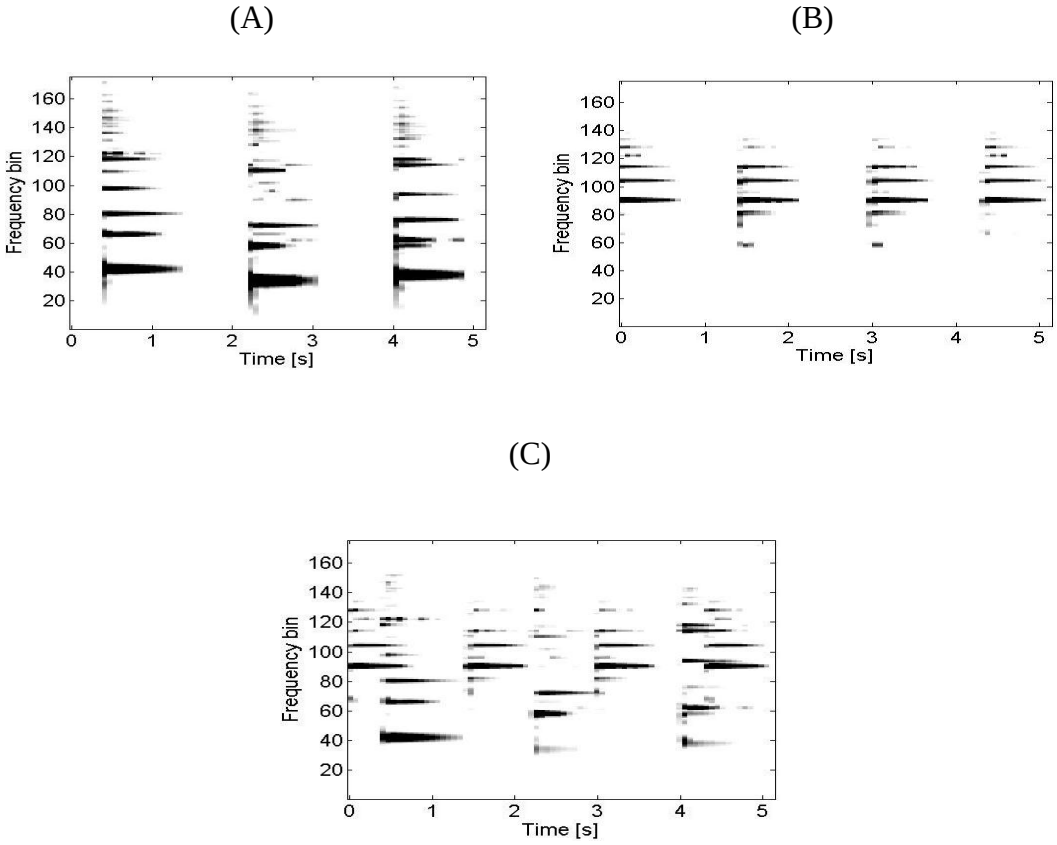


Figure 4.7: Log-frequency spectrogram of (A) piano, (B) trumpet and (C) convolutive mixed signal.

4.3.2.1 Sources estimation

In this experiment, from convolutive mixture $|\mathbf{Y}|^{\cdot 2}$, we seek the two estimated sources images which are $\left|\hat{\mathbf{X}}_1^{img}\right|^{\cdot 2} = \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} \mathbf{U}_1 \overset{\downarrow\phi}{\mathbf{W}_1^{\tau}} \overset{\rightarrow\tau}{\mathbf{H}_1^{\phi}}$ and $\left|\hat{\mathbf{X}}_2^{img}\right|^{\cdot 2} = \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} \mathbf{U}_2 \overset{\downarrow\phi}{\mathbf{W}_2^{\tau}} \overset{\rightarrow\tau}{\mathbf{H}_2^{\phi}}$.

Then, by using binary masking technique [125] and defining the masking matrix as $\mathbf{M}_j = \left\{ m_{j,f,n} \,|\, f = 1,...,F \text{ and } j = 1,...,J \right\}$ where

$$m_{j,f,n} = \begin{cases} 1, & if \quad \left|\hat{x}_{j,f,n}^{im}\right|^2 > \left|\hat{x}_{k,f,n}^{im}\right|^2 \\ 0, & Otherwise \end{cases} \tag{4.43}$$

the time domain estimated signal $\hat{\mathbf{x}}_j$ is obtained by resynthesizing $\mathbf{M}_i$ with the mixture $|\mathbf{Y}|^2$ i.e. $\hat{\mathbf{x}}_j = \text{resynthesize}\left(\mathbf{M}_j \bullet |\mathbf{Y}|^2\right)$. Here, 'resynthesize' signifies the inverse mapping of log-frequency axis to the original frequency axis and then followed by inverse STFT back to the time domain.

4.3.2.2 Comparison between Quasi-EM FCNMF2D and MU FCNMF2D

In this sub-section, we compare the performance of source separation between both proposed algorithms of Quasi-EM FCNMF2D and MU FCNMF2D. Figure 4.8 shows the separation result in log-frequency spectrogram for both proposed algorithms. Compared with original image sources in Figure 4.7, it is visually clear that separation of MU FCNMF2D in Figure 4.8(A) and 4.8(B) led to poor result since the factorization still contains the mixed signal (indicated by the box marked area). This is because in MU FCNMF2D algorithm, the stationary point cannot be determined properly where it not promised to be local minimum. As for the Quasi-EM FCNMF2D, it has yielded the better performance with the source images almost fully recovered. This is due to the property of Quasi-EM algorithm which guaranteed the convergence to the local minimum.

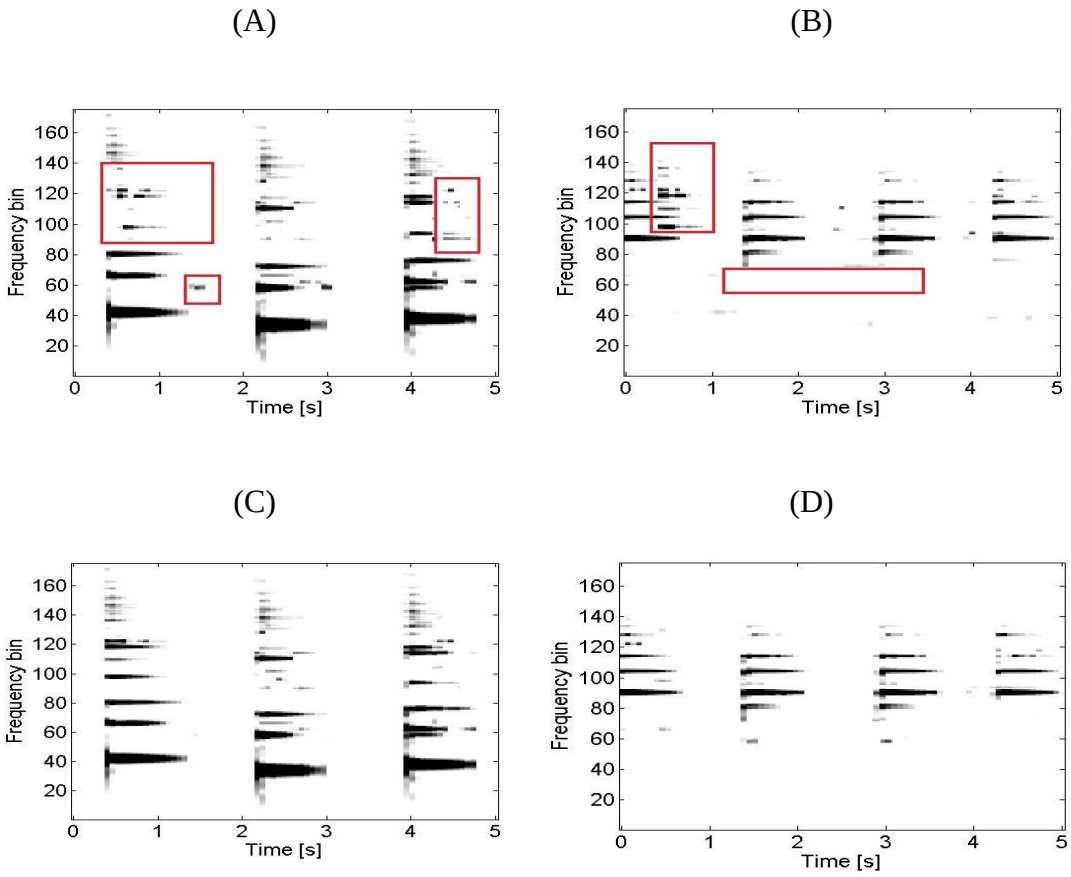


Figure 4.8: Separated sound in log-frequency spectrogram (A)-(B) piano and trumpet sound using MU FCNMF2D (C)-(D) piano and trumpet sound using Quasi-EM FCNMF2D.

From Table 4.3, in general both proposed algorithms deliver decent results with the performance of SDR, SIR and SAR that can be considered good. Over 10 dB of SDR measurement have been recorded for both proposed methods. However, performance of Quasi-EM FCNMF2D algorithm is superior compare to MU FCNMF2D with the average SDR improvement of 3.7dB per source. In percentage, this translates to an average improvement of 31%. Again, this indicates that MU

FCNMF2D was trapped in local minimum which affect the performance of the separation.

Table 4.3: Separation results for FCNMF2D methods

Algorithms	Separated piano			Separated trumpet		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
MU FCNMF2D	13.4	20.3	14.5	10.2	13.1	14.2
Quasi-EM FCNMF2D	16.2	22.7	17.4	14.8	22.3	15.6

4.3.2.3 Effect of frequency mixing, U

In this sub-section, the impact of frequency mixing, **U** in both proposed algorithms is demonstrated. To observe this, we will evaluate the proposed algorithm where **U** is constant by simply setting $\mathbf{U}_j = \mathbf{I}$ while the adaptation of $\mathbf{W}$ and $\mathbf{H}$ still follows the proposed methodologies. Figure 9 shows the separation result of the same convolutive mixture from previous experiment. Comparing with original sources in Figure 4.7, by discounting the frequency variation in the channels in both algorithms, the plots are clear to show that errors have accumulated in both separated sounds (highlighted by the marked box) where some components have been attributed incorrectly. This is due to the assumption that **U** has the uniform value for all frequency which is not true for convolutive mixture. Consequently, this has led to misrepresentation of spectral basis and temporal in the separation. Comparing the separated signal in the plot of Figure 4.9 with the plots of FCNMF2D in Figure

4.8(A)-(D), it is undoubtedly show the vital of constraining **U** in order to reduce the errors as well as to obtain the accurate result in separation.

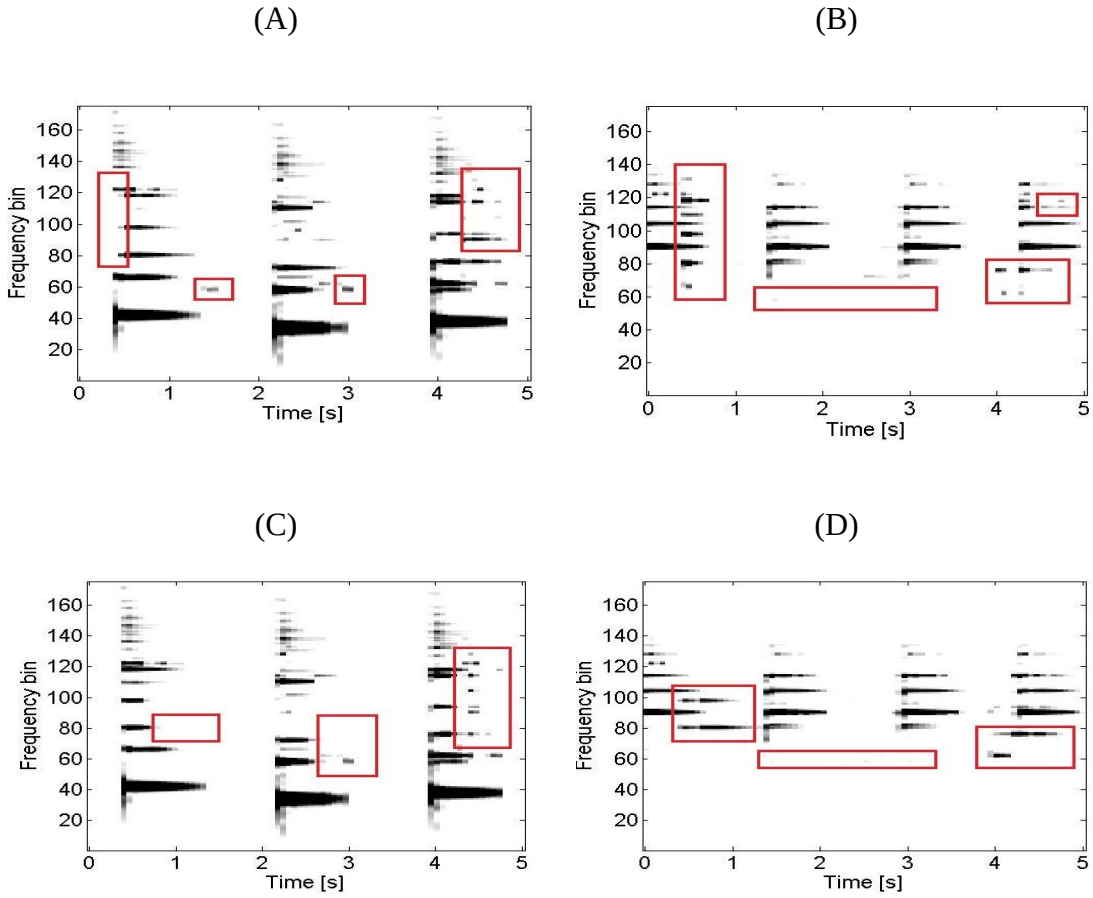


Figure 4.9: Separated sound in log-frequency spectrogram for the case of without updating **U** (A)-(B) piano and trumpet sound using MU NMF2D (C)-(D) piano and trumpet sound using Quasi-EM NMF2D.

Table 4.4 shows the separation performance of proposed method if we set **U**=**I**. Comparing Table 4.3 with Table 4.4, through updating the frequency mixing parameter, the SDR performance of the algorithm has improved by 2.9 dB for separated piano and 5 dB for separated trumpet for Quasi-EM FCNMF2D which

translates to an average of 17% per source. For the MU FCNMF2D, an improvement of 1.2 dB for separated piano and 0.7dB for separated trumpet which indicate 10% improvement per source. This improvement indicates that for convolutive mixture, it is crucial for frequency distribution of the spectral basis be constrained through **U** to avoid the distortion in the decomposition.

Table 4.4: Separation results of proposed method with **U**=**I**

Algorithms	**Separated piano**			**Separated trumpet**		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
MU NMF2D	12.2	21.5	12.9	9.3	12.0	12.8
Quasi-EM NMF2D	13.3	22.1	14.4	9.8	12.1	14.3

4.3.2.4 Separabilty analysis

In the binary masking technique [125], one generates the TF mask corresponding to each source and applies the created mask to the mixture to obtain the estimated source TF representation. In this sub-section, we measure a separability based on the performance of TF masks generated using knowledge of the source and interference TF of mixture. In particular, when the sources do not overlap in the TF domain, an optimum mask $m_{j,f,n}^{opt}$ exists which allows one to extract the j^{th} original source from the mixture as:

$$x_{j,f,n} = m_{j,f,n}^{opt} y_{f,n} \tag{4.44}$$

In order to measure the separability for a given mask, we use two performance criteria: (i) Preserved signal ratio (PSR) which determines how well the mask preserves the source of interest, and (ii) Signal-to-interference ratio (SIR) which indicates how well the mask suppresses the interfering. Both of criteria are introduced as follow:

Given any TF mask $m_{j,f,n}$ such that $0 \le m_{j,f,n} \le 1$ for all *f* and *n*, PSR is defined as:

$$PSR_{m_j}^{\hat{x}_j} = \frac{\left\| m_{j,f,n} \hat{x}_{j,f,n} \right\|_{Fro}^2}{\left\| \hat{x}_{j,f,n} \right\|_{Fro}^2} \tag{4.45}$$

where $\|.\|_{Fro}$ is the Frobenius norm. Note that $PSR_{m_j}^{\hat{x}_j} \le 1$ with $PSR_{m_j}^{\hat{x}_j} = 1$ only if $\text{supp}\, m^{opt} \subseteq \text{supp}\, m$ where 'supp' denotes support. Now we define the interfering sources as $r_j(t) = \sum_{k=1,k\neq j}^{K} x_j(t)$. Then, SIR of the mask, $m_{j,f,n}$ is define as

$$SIR_{m_j}^{\hat{x}_j} = \frac{\left\| m_{j,f,n} \hat{x}_{j,f,n} \right\|_{Fro}^2}{\left\| m_{j,f,n} \hat{r}_{j,f,n} \right\|_{Fro}^2} \tag{4.46}$$

where $\hat{r}_{j,f,n}$ is the TF representations of $\hat{r}_j(t)$. Combining the PSR and SIR into one measure the approximate separability, we define the normalised difference between the signal energy maintained in masking as a measure of separability as:

$$S_{m_j}^{y \to x_j, r_j} = \frac{\left\| m_{j,f,n} x_{j,f,n} \right\|_{Fro}^2}{\left\| x_{j,f,n} \right\|_{Fro}^2} - \frac{\left\| m_{j,f,n} r_{j,f,n} \right\|_{Fro}^2}{\left\| x_{j,f,n} \right\|_{Fro}^2} \tag{4.47}$$

We also define the separability of the mixture with respect to all the *J* sources as:

$$S_{m_1,\ldots,m_J}^{y\to x_1,\ldots,x_J} = \frac{1}{J}\sum_{j=1}^{J} S_{m_j}^{y\to x_j,r_j} \tag{4.48}$$

It can be shown that (4.47) can be expressed as $S_{m_j}^{y\to x_j,r_j} = PSR_{m_j}^{x_j} - PSR_{m_j}^{x_j} \big/ SIR_{m_j}^{x_j}$. Eqn. (4.47) also equivalent to measuring the success of extracting the j^{th} source $x_{j,f,n}$ from the mixture $y_{f,n}$ given the TF mask $m_{j,f,n}$. Similarly, (4.46) measures the success of extracting all the N sources simultaneously from the mixture. When $S_{m_j}^{y\to x_j,r_j} = 1$ (i.e. $PSR_{m_j}^{x_j} = 1$ and $SIR_{m_j}^{x_j} = \infty$), this indicates that the mixture $y(t)$ is separable with respect to the j^{th} source $x_j(t)$. In other words, $x_{j,f,n}$ does not overlap with $r_{j,f,n}$ and the TF mask $m_{j,f,n}$ has perfectly separated the j^{th} source $x_{j,f,n}$ from the mixture $y_{f,n}$. This corresponds to $m_{j,f,n} = m_{j,f,n}^{opt}$ in (4.44). Hence, this is the maximum attainable $S_{m_j}^{y\to x_j,r_j}$ value. For other cases of $PSR_{m_j}^{x_j}$ and $SIR_{m_j}^{x_j}$, we have $S_{m_j}^{y\to x_j,r_j} < 1$. Using this concept, we can extend the analysis for the case of separating J sources. A mixture $y(t)$ is fully separable to all the J sources if and only if $S_{m_1,\ldots,m_J}^{y\to x_1,\ldots,x_J} = 1$ in (4.47). For the case $S_{m_1,\ldots,m_J}^{y\to x_1,\ldots,x_J} < 1$, this implies that some of the sources overlap with each other in the TF domain and therefore, they cannot be fully separated. Thus, $S_{m_1,\ldots,m_J}^{y\to x_1,\ldots,x_J}$ provides the quantitative performance measure for evaluating how separable is the mixture in the TF domain.

To obtain an objective evaluation, we have also included the separation results using ideal binary mask (IBM) [125]. Note that since the IBM is derived directly from the source signals, its separation performance represents the ideal case. Table 4.5 shows the averaged separability performance of piano and trumpet mixture using IBM

and proposed algorithms. Following the listening performing test proposed in [66], we conclude that $S_{m_j}^{y\to x_j,r_j} > 0.85$ leads to acceptable separation performance. Therefore, proposed algorithms in Table 4.5 satisfy this condition. While this is true, without updating **U** for convolutive mixture, MU NMF2D gives only a mediocre level of separability with averaged $S_{m_j}^{y\to x_j,r_j} \approx 0.89$. Compare with constraining the frequency mixing, the separability performance can be seen increase with $S_{m_j}^{y\to x_j,r_j} \approx 0.92$ for MU FCNMF2D which indicates 3% improvement from MU NMF2D. As for Quasi-EM FCNMF2D, the averaged $S_{m_j}^{y\to x_j,r_j} \approx 0.98$ has been achieved which indicate the improvement of 7% from Quasi-EM NMF2D which has $S_{m_j}^{y\to x_j,r_j} \approx 0.91$. This shows the effectiveness of **U** in algorithm for convolutive mixture to obtain the accurate result. We can observe as well that Quasi-EM FCNMF2D is perform superior performance compared to MU FCNMF2D algorithm for all measurement of PSR, SIR and $S_{m_1,m_2}^{y\to x_1,x_2}$. Again, this indicates that Quasi-EM is computationally efficient with only small amount of sources overlap with each other. In addition, the convergence of the Quasi-EM to stationary point can be achieved. Comparing Quasi-EM FCNMF2D with the ideal case of IBM, we can say that the performance of Quasi-EM algorithm is almost imitating the performance of IBM.

Table 4.5: Separability performance

Method	**PSR**	**SIR**	$S_{m_1,m_2}^{y\to x_1,x_2}$
IBM	0.996	220.7	0.992
Quasi-EM FCNMF2D	0.991	151.8	0.976
MU FCNMF2D	0.959	87.4	0.924
Quasi-EM NMF2D	0.932	85.7	0.913
MU NMF2D	0.924	66.5	0.887

4.4 Summary

In this chapter, novel solutions have been presented to separate a mixture in convolutive single channel recording. Two inference techniques have been proposed: a variant of EM algorithm which maximises the joint log-likelihood called Quasi-EM FCNMF2D, and MU rules for the maximisation of individual log-likelihood called MU FCNMF2D. It has also been shown that significant performance improvement has been achieved by updating the frequency mixing parameter for convolutive mixture. Decomposition on feature extraction and blind audio source separation has proven to be exceptional especially for Quasi-EM FCNMF2D algorithm. There are at least three major advantages of proposed method: Firstly, the proposed algorithms contemplate the convolutive mixing model which implies more accurate representation of the actual environment. Secondly, the methods are computationally efficient where it avoids strong constrains of separating sources without prior knowledge of the original

sources. Finally, the IS divergence holds the desirable property of scale invariant that enables low energy components in the log spectrogram bear the same relative importance as the high energy ones.

CHAPTER 5

LINEAR SINGLE CHANNEL SOURCE SEPARATION IN CONVOLUTIVE MIXTURE USING FREQUENCY CONSTRAINED SPARSE NONNEGATIVE MATRIX FACTORIZATION

Previous SCSS method proposed in Chapter 4 which is based on frequency constrained two dimensional nonnegative matrix factorization (FCNMF2D) model has certain ambiguities between the spectral basis, **W** and temporal code, **H**. For example, if the data do not extent the positive octant adequately, a rotation of **W** and opposite rotation of **H** can yield same result. In addition, the structure in **W** can to some extend being place into the signature of the same factor in **H** and vice versa [126, 127]. Hence, it is necessary to impose sparseness to give unique and realistic representations of the non-stationary audio signals. To extend previous FCNMF2D model, a novel two-dimensional frequency constrained sparse nonnegative matrix factorization (FC-SNMF2D) is proposed in this chapter. The method aims to separate the mixture into its constituent spectral-temporal source components while alleviating the effect of convolutive mixing. In addition, we incorporate adaptive sparseness into the solution which bypasses the need of manual selection of the sparseness parameter in conventional matrix factorization methods. Sparseness on **H** is imposed element-wise

so that each individual code has its own distribution. Consequently, the sparsity parameter can be individually optimised for each code. This overcomes the problem of under-sparse and over-sparse factorization. Experimental tests on audio signals have been carried out to verify the proposed FC-SNMF2D model and to evaluate its performance in single channel source separation. We have investigated synthetic convolutive mixture, live-recorded mixture and professional music recording. Results have concretely shown the effectiveness of the proposed framework in separating the signals in reverberant environment.

The chapter is organised as follows: Section 5.1 introduces the background of SNMF2D model. The derivation of proposed separation technique of frequency constrained two dimensional sparse NMF is explained in Section 5.2. In Section 5.3, the results of both experimental and live-recording signals as well as the analysis are presented. Section 5.4 concludes the chapter.

5.1 Background

5.1.1 Two-dimensional sparse nonnegative matrix factorization

Sparse representation is a representation of data where most coefficients are zero. It is proving to be a particularly interesting and powerful tool especially for analysis and processing of audio signals. If each signal to be separated has a sparse representation, then there is a good chance that there will be little overlap between the small sets of coefficients used to represent the different source signals. Therefore by selecting the coefficients used by each source signal, we can restore each of the

original signals with most of the interference from the unwanted signals removed. The use of sparse representation is strongly related to the principle of parsimony, i.e. among all possible accounts the simplest is considered the best. To avoid over fitting Parsimony can be considered a reasonable guiding principle if no formal prior information is given. Hence, NMF2D model [94] can be extended to SNMF2D model [126, 127] where two basic cost functions have been imposed with sparse penalty such that:

Least Square :

$$C_{LS}\left(\left|\mathbf{Y}\right|^{.2}\left|\tilde{\mathbf{P}}\right.\right)=\frac{1}{2}\sum_{f,n}\left(\left|\mathbf{Y}_{f,n}\right|^{2}-\left|\tilde{\mathbf{P}}_{f,n}\right|^{2}\right)^{2}+\lambda f\left(\mathbf{H}\right) \tag{5.1}$$

Kullback-Leibler :

$$C_{KL}\left(\left|\mathbf{Y}\right|^{.2}\left|\tilde{\mathbf{P}}\right.\right)=\sum_{f,n}\left(\left|\mathbf{Y}_{f,n}\right|^{2}\log\frac{\left|\mathbf{Y}_{f,n}\right|^{2}}{\left|\tilde{\mathbf{P}}_{f,n}\right|^{2}}-\left|\mathbf{Y}_{f,n}\right|^{2}+\left|\tilde{\mathbf{P}}_{f,n}\right|^{2}\right)+\lambda f\left(\mathbf{H}\right) \tag{5.2}$$

for $f=1,...,F$, $n=1,...,N$ where $\tilde{\mathbf{P}}=\sum_{\tau,\phi}\overset{\downarrow\phi}{\tilde{\mathbf{W}}}^{\tau}\overset{\rightarrow\tau}{\mathbf{H}}^{\phi}$ and $\tilde{\mathbf{W}}_{f,j}^{\tau}=\mathbf{W}_{f,j}^{\tau}\Big/\sqrt{\sum_{\tau,f}\left(\mathbf{W}_{f,j}^{\tau}\right)^{2}}$ and

$f\left(\mathbf{H}\right)$ can be any function with positive derivative such as $L_{\varsigma}=norm\left(\varsigma>0\right)$ given by $f\left(\mathbf{H}\right)=\left\|\mathbf{H}\right\|_{\varsigma}=\left(\sum_{\phi,j,n}\mathbf{H}_{j.n}^{\phi}\right)^{\frac{1}{\varsigma}}$. This will resolve the ambiguity between the factors by imposing sparseness on $\mathbf{H}^{\phi}$ and forcing the structure onto $\mathbf{W}^{\tau}$. Nevertheless, the disadvantage of SNMF2D is cause by its deficiencies of a generalised criterion for controlling the sparsity of **H**. In SNMF2D, the sparsity parameter is set manually. When SNMF2D imposes uniform sparsity on all temporal codes, this is equivalent to enforcing each temporal code to be identical to a fixed distribution according to the selected sparsity parameter. In addition, by assigning the fixed distribution onto each

individual code, this is equivalent to constraining all codes to be stationary. However, audio signals are non-stationary in the TF domain and have different temporal structure and sparsity. Hence, they cannot be realistically enforced by a fixed distribution. These characteristics are even more pronounced between different types of audio signals. In addition, since SNMF2D introduces many temporal shifts, this will result in more temporal codes to deviate from fixed distribution. Therefore, within the context of SCSS, when SNMF2D imposes uniform sparsity on all the temporal codes, this will inevitably result in under-sparse or over-sparse factorization which will subsequently lead to ambiguity in separating audio mixtures. Thus, the above suggests that the present form sparseness constraint is still technically lacking and is not readily suited for SCSS especially mixtures involving more types of audio signals.

5.2 Proposed Separation Method

In this section, a new algorithm based on frequency constrained sparse two-dimensional nonnegative matrix factorization (FC-SNMF2D) model with adaptive sparseness and channel estimation is proposed for separating single channel convolutive mixture. Unlike [126, 127], in this chapter, Itakura-Saito (IS) divergence will be developed as a cost function as it holds the desirable property of scale invariant where the same relative weight is considered equally for both small and large coefficients of the mixture. This is rather important especially since the decomposition of the mixture involves music or/and speech spectra which contains more low power components mix together with the higher power component such as tonal parts of sustained note.

5.2.1 Frequency constrained SNMF2D

We proposed a new model based on (4.7) as our mixture model and the noise free environment are considered. The (log-frequency) power spectrogram of each j^{th} source image $\left|\mathbf{X}_j^{im}\right|^{.2}$ is product of two nonnegative matrices of source signal model with frequency constrained which is defined as $\left|\mathbf{X}_j^{im}\right|^{.2} \approx \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{U}_j \overset{\downarrow\phi}{\mathbf{W}_j^{\tau}} \overset{\rightarrow\tau}{\mathbf{H}_j^{\phi}}$.

In the proposed model, the factorization of $\mathbf{W}_j^{\tau}$, $\mathbf{H}_j^{\phi}$ and $\mathbf{U}_j$ suffer from the ambiguities between each component during the decomposition process. To overcome this problem, the proposed method incorporated adaptive sparsity constraints on the temporal code **H**, to reassure the ambiguity by forcing the structure onto **W** and Experimental results show that our proposed adaptive sparseness constraint will obtain better separation performance and the details of approach will be explained in the next sub-section.

5.2.2 Cost function with adaptive sparseness

In this sub-section, the cost function of IS divergence for the proposed FC-SNMF2D is formulated. In addition, the impact of adaptive sparseness in the proposed factorization will be analysed. First of all, we may define $\mathbf{U} = \left[\mathbf{U}_1\ \mathbf{U}_2\ ...\ \mathbf{U}_J\right]$, $\mathbf{W} = \left[\mathbf{W}^0\ \mathbf{W}^1\ ...\ \mathbf{W}^{\tau_{\max}}\right]$, $\mathbf{H} = \left[\mathbf{H}^0\ \mathbf{H}^1\ ...\ \mathbf{H}^{\phi_{\max}}\right]$ and $\mathbf{\Lambda} = \left[\mathbf{\Lambda}^0\ \mathbf{\Lambda}^1\ ...\ \mathbf{\Lambda}^{\phi_{\max}}\right]$. $\mathbf{\Lambda}$ is defined as the sparseness parameters which will be imposed onto **H**. To facilitate the

decomposition of $\left|\mathbf{X}_j^{im}\right|^{.2} \approx \sum_{\tau=0}^{\tau_{\max}} \sum_{\phi=0}^{\phi_{\max}} \mathbf{U}_j \overset{\downarrow\phi}{\mathbf{W}}_j^{\tau} \overset{\rightarrow\tau}{\mathbf{H}}_j^{\phi}$, the following generative model is considered:

$$\left|\mathbf{Y}\right|^{.2} = \left(\sum_j \sum_\tau \sum_\phi \mathbf{U}_j \left(\overset{\downarrow\phi}{\mathbf{W}}_j^{\tau} \overset{\rightarrow\tau}{\mathbf{H}}_j^{\phi} \right) \right) \bullet \mathbf{E} \tag{5.3}$$

where "$\bullet$" denotes element-wise product and $\mathbf{E}$ is a matrix multiplicative independent and identically-distributed (i.i.d.) Gamma noise with mean unity. Next, the prior distribution $p(\mathbf{U},\mathbf{W},\mathbf{H})$ is choosing over the factors $\{\mathbf{U},\mathbf{W},\mathbf{H}\}$. By using Bayes' theorem, the posterior can be expressed as follow:

$$p\left(\mathbf{U},\mathbf{W},\mathbf{H}\left|\left|\mathbf{Y}\right|^{.2},\boldsymbol{\Lambda}\right.\right) = \frac{p\left(\left|\mathbf{Y}\right|^{.2}\left|\mathbf{U},\mathbf{W},\mathbf{H}\right.\right) p(\mathbf{U})\, p(\mathbf{W})\, p\left(\mathbf{H}\left|\boldsymbol{\Lambda}\right.\right)}{p\left(\left|\mathbf{Y}\right|^{.2}\right)} \tag{5.4}$$

Since denominator is a constant while $\mathbf{U}$, $\mathbf{W}$ and $\mathbf{H}$ are presumed jointly independent, then the log-posterior can be written as:

$$\log p\left(\mathbf{U},\mathbf{W},\mathbf{H}\left|\left|\mathbf{Y}\right|^{.2},\boldsymbol{\Lambda}\right.\right) = \log p\left(\left|\mathbf{Y}\right|^{.2}\left|\mathbf{U},\mathbf{W},\mathbf{H}\right.\right) + \log p\left(\mathbf{U},\mathbf{W},\mathbf{H}\left|\boldsymbol{\Lambda}\right.\right) + const \tag{5.5}$$

where '*const'* denotes constant. It can be shown that the IS divergence is equivalent to negative log-likelihood estimation in multiplicative independent and identically-distributed (i.i.d.) Gamma noise with unity mean. Defining the gamma probability density function (pdf) as $\xi(\mathbf{E}_{f,n} \mid \alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \left(\mathbf{E}_{f,n}\right)^{\alpha-1} \exp\left(-\beta \mathbf{E}_{f,n}\right), \mathbf{E}_{f,n} \geq 0$, the negative log-likelihood for the first term of the right hand side of (5.5) can be expressed as

$$
\begin{aligned}
&-\log p\left(\left|\mathbf{Y}\right|^{\cdot 2}\middle|\mathbf{U},\mathbf{W},\mathbf{H}\right) \\
&= -\sum_{f,n}\log\xi\left(\left.\frac{\left|\mathbf{Y}\right|^{\cdot 2}_{f,n}}{\sum_{j,\tau,\phi}\mathbf{U}_{j,f}\left(\mathbf{W}^{\tau}_{f,j}\overset{\downarrow\phi}{\mathbf{H}}^{\phi}_{j,n}\overset{\rightarrow\tau}{}\right)}\right|\alpha,\beta\right)\Bigg/\sum_{j,\tau,\phi}\mathbf{U}_{j,f}\left(\mathbf{W}^{\tau}_{f,j}\overset{\downarrow\phi}{\mathbf{H}}^{\phi}_{j,n}\overset{\rightarrow\tau}{}\right) \\
&\overset{c}{=} \beta\sum_{f,n}\left(\frac{\left|\mathbf{Y}\right|^{\cdot 2}_{f,n}}{\sum_{j,\tau,\phi}\mathbf{U}_{j,f}\left(\mathbf{W}^{\tau}_{f,j}\overset{\downarrow\phi}{\mathbf{H}}^{\phi}_{j,n}\overset{\rightarrow\tau}{}\right)}-\frac{\alpha}{\beta}\log\frac{\left|\mathbf{Y}\right|^{\cdot 2}_{f,n}}{\sum_{j,\tau,\phi}\mathbf{U}_{j,f}\left(\mathbf{W}^{\tau}_{f,j}\overset{\downarrow\phi}{\mathbf{H}}^{\phi}_{j,n}\overset{\rightarrow\tau}{}\right)}-1\right)
\end{aligned}
\tag{5.6}
$$

where "$\overset{c}{=}$" denotes e quality up to a positive scale and constant. The ratio α/β is the mean of the Gamma distribution in which when it is equal to unit length, the minus log-likelihood in (5.6) is equal to $D_{IS}\left(\left|\mathbf{Y}\right|^{\cdot 2}\middle\|\left|\hat{\mathbf{Y}}\right|^{\cdot 2}\right)=D_{IS}\left(\left|\mathbf{Y}\right|^{\cdot 2}\middle|\sum_{j}\sum_{\tau}\sum_{\phi}\mathbf{U}_{j,f}\left(\mathbf{W}^{\tau}_{f,j}\overset{\downarrow\phi}{\mathbf{H}}^{\phi}_{j,n}\overset{\rightarrow\tau}{}\right)\right)$ up to a positive scale and constant and $D_{IS}\left(a\middle|b\right)=\frac{a}{b}-\log\frac{a}{b}-1$ is the IS divergence. The term $\log p\left(\mathbf{U},\mathbf{W},\mathbf{H}\middle|\mathbf{\Lambda}\right)$ comprises of the prior distribution of $\mathbf{U}$, $\mathbf{W}$ and $\mathbf{H}$, respectively. The prior over **U** and **W** are flat where each column is assumed to be factor-wise normalised to unit length. The prior over **H** is assumed to be exponentially distributed with decay parameters of $\lambda^{\phi}_{j,n}$ for each element in **H** and it can be expressed as:

$$
\begin{aligned}
p\left(\mathbf{H}\middle|\mathbf{\Lambda}\right) &= \prod_{j}p\left(\mathbf{H}_{j}\middle|\mathbf{\Lambda}_{j}\right) \\
&= \prod_{j}\prod_{n}\prod_{\phi}p\left(h^{\phi}_{j,n}\middle|\lambda^{\phi}_{j,n}\right) \\
&= \prod_{j}\prod_{n}\prod_{\phi}\lambda^{\phi}_{j,n}\exp\left(-\lambda^{\phi}_{j,n}h^{\phi}_{j,n}\right)
\end{aligned}
\tag{5.7}
$$

The minus log-likelihood for prior on **H** is derived such as:

$$\begin{aligned} -\log p\left(\mathbf{H}\middle|\mathbf{\Lambda}\right) &= -\log\left(\prod_{j}\prod_{n}\prod_{\phi}\lambda_{j,n}^{\phi}\exp\left(-\lambda_{j,n}^{\phi}h_{j,n}^{\phi}\right)\right) \\ &= \sum_{j}\sum_{n}\sum_{\phi}\left(\lambda_{j,n}^{\phi}h_{j,n}^{\phi}-\log\lambda_{j,n}^{\phi}\right) \end{aligned} \tag{5.8}$$

By inserting (5.8) into IS divergence derived in (5.6), the proposed cost function to be optimised can be formulated as

$$\begin{aligned} C_{IS} &= D_{IS}\left(\left|\mathbf{Y}\right|^{\cdot 2}\middle|\sum_{j,\tau,\phi}\mathbf{U}_{j,f}\left(\mathbf{W}_{f,j}^{\tau\downarrow\phi}\mathbf{H}_{j,n}^{\phi\rightarrow\tau}\right)\right)+\sum_{j,n,\phi}\left(\lambda_{j,n}^{\phi}h_{j,n}^{\phi}-\log\lambda_{j,n}^{\phi}\right) \\ &= \sum_{f,n}\left(\frac{\left|\mathbf{Y}\right|_{f,n}^{\cdot 2}}{\sum_{j,\tau,\phi}\mathbf{U}_{j,f}\left(\mathbf{W}_{f,j}^{\tau\downarrow\phi}\mathbf{H}_{j,n}^{\phi\rightarrow\tau}\right)}-\log\frac{\left|\mathbf{Y}\right|_{f,n}^{\cdot 2}}{\sum_{j,\tau,\phi}\mathbf{U}_{j,f}\left(\mathbf{W}_{f,j}^{\tau\downarrow\phi}\mathbf{H}_{j,n}^{\phi\rightarrow\tau}\right)}-1\right)+\sum_{j,n,\phi}\lambda_{j,n}^{\phi}h_{j,n}^{\phi}-\sum_{j,n,\phi}\log\lambda_{j,n}^{\phi} \end{aligned} \tag{5.9}$$

where $\lambda_{j,n}^{\phi}$ is the sparsity weight factor and the term $f(\mathbf{H})=\sum_{j}\sum_{n}\sum_{\phi}\lambda_{j,n}^{\phi}h_{j,n}^{\phi}$ forms the L_1-norm regularization which is used to resolve the ambiguity by forcing all structure in $\mathbf{H}$ onto **W**. Therefore, the sparseness of the solution in (5.9) is highly dependent on the regularization parameter $\lambda_{j,n}^{\phi}$.

For conventional sparseness constraint [126-129], the sparsity parameter $\lambda_{j,n}^{\phi}$ is a fixed constant such that $\lambda_{j,n}^{\phi}=\lambda$ for all *j*, *n* and ϕ. In addition, the value of λ is set manually and the temporal code is imposed with uniform sparsity. This will lead to fixed distribution on each temporal code which made them stationary. In practice, many audio signals are non-stationary and to impose stationary assumption is just unrealistic. Furthermore, more temporal codes will deviate from the fixed distribution

since uniform sparsity tends to introduce many temporal shifts. In such circumstances, the obtained factorization will invariably suffer from either over-sparse or under-sparse which consequently lead to ambiguity in the separation process. The proposed algorithm overcome above problem by introducing adaptive sparseness where in each individual element in **H**, the values of $\lambda_{j,n}^{\phi}$ is varied across time. With this technique, it represents the realistic solutions and resolves the ambiguity more efficiently.

5.2.3 Estimation of convolutive mixing, spectral basis and temporal code

Rewriting the proposed cost function (5.9) in terms of the parameters of the FC-SNMF2D as

$$\tilde{C}_{IS} = \sum_{f}\sum_{n}\left(\frac{\left|y_{f,n}\right|^2}{\tilde{z}_{f,n}} - \log\frac{\left|y_{f,n}\right|^2}{\tilde{z}_{f,n}} - 1\right) + \sum_{j,n,\phi} h_{j,n}^{\phi}\lambda_{j,n}^{\phi} - \sum_{j,n,\phi}\log\lambda_{j,n}^{\phi} \tag{5.10}$$

where $\tilde{z}_{f,n} = \sum_{j}\sum_{\tau}\sum_{\phi}\tilde{u}_{j,f}\tilde{w}_{f-\phi,j}^{\tau}h_{j,n-\tau}^{\phi}$ with factor-wise normalised $\tilde{w}_{f,j}^{\tau} = w_{f,j}^{\tau} / \sqrt{\sum_{\tau}\sum_{f}(w_{f,j}^{\tau})^2}$ and $\tilde{u}_{j,f} = u_{j,f} / \sqrt{\sum_{f}(u_{j,f})^2}$. The derivatives of individual component for FC-SNMF2D corresponding to $u_{j,f}$, $w_{f,j}$ and $h_{j,n}$ are described as follows:

For $u_{j,f}$, the derivatives is given by

$$\begin{aligned}\frac{\partial\tilde{C}_{IS}}{\partial u_{j,f}} &= -\sum_{n}\left|y_{f,n}\right|^2\tilde{z}_{f,n}^{-2}\sum_{f,n}\tilde{w}_{f-\phi,j}^{n-n'}h_{j,n-\tau}^{f-f'} + \sum_{n}\tilde{z}_{f,n}^{-1}\sum_{f,n}\tilde{w}_{f-\phi,j}^{n-n'}h_{j,n-\tau}^{f-f'} \\ &= -\sum_{n}\left|y_{f,n}\right|^2\tilde{z}_{f,n}^{-2}\sum_{\phi,\tau}\tilde{w}_{f-\phi,j}^{\tau}h_{j,n-\tau}^{\phi} + \sum_{n}\tilde{z}_{f,n}^{-1}\sum_{\phi,\tau}\tilde{w}_{f-\phi,j}^{\tau}h_{j,n-\tau}^{\phi}\end{aligned} \tag{5.11}$$

As for $w^{\tau}_{f,j}$ and $h^{\phi}_{j,n}$, the derivative of the component are given by

$$\begin{aligned}\frac{\partial \tilde{C}_{IS}}{\partial w^{\tau'}_{f',j'}} &= -\sum_{f,n}\left|y_{f,n}\right|^2 \tilde{z}_{f,n}{}^{-2}\sum_{f,n}\tilde{u}_{j',f}h^{f-f'}_{j',n-\tau'} + \sum_{f,n}\tilde{z}_{f,n}{}^{-1}\sum_{f,n}\tilde{u}_{j',f}h^{f-f'}_{j',n-\tau'} \\ &= -\sum_{\phi,n}\left|y_{f'+\phi,n}\right|^2 \tilde{z}_{f'+\phi,n}{}^{-2}\tilde{u}_{j',f'+\phi}h^{\phi}_{j',n-\tau'} + \sum_{\phi,n}\tilde{z}_{f'+\phi,n}{}^{-1}\tilde{u}_{j',f'+\phi}h^{\phi}_{j',n-\tau'}\end{aligned} \tag{5.12}$$

Similarly,

$$\begin{aligned}\frac{\partial \tilde{C}_{IS}}{\partial h^{\phi'}_{j',n'}} &= -\sum_{f,n}\left|y_{f,n}\right|^2 \tilde{z}_{f,n}{}^{-2}\sum_{f,n}\tilde{u}_{j',f}\tilde{w}^{n-n'}_{f-\phi',j'} + \sum_{f,n}\tilde{z}_{f,n}{}^{-1}\sum_{f,n}\tilde{u}_{j',f}\tilde{w}^{n-n'}_{f-\phi',j'} + \lambda^{\phi'}_{j',n'} \\ &= -\sum_{f,\tau}\left|y_{f,n'+\tau}\right|^2 \tilde{z}_{f,n'+\tau}{}^{-2}\tilde{u}_{j',f}\tilde{w}^{\tau}_{f-\phi',j'} + \sum_{f,\tau}\tilde{z}_{f,n'+\tau}{}^{-1}\tilde{u}_{j',f}\tilde{w}^{\tau}_{f-\phi',j'} + \lambda^{\phi'}_{j',n'}\end{aligned} \tag{5.13}$$

For each of individual component, standard gradient descent method is applied with

$$u_{j,f} \leftarrow \tilde{u}_{j,f} - \eta_u \frac{\partial \tilde{C}_{IS}}{\partial u_{j,f}}, \quad w^{\tau'}_{f',j'} \leftarrow \tilde{w}^{\tau'}_{f',j'} - \eta_w \frac{\partial \tilde{C}_{IS}}{\partial w^{\tau'}_{f',j'}} \quad \text{and} \quad h^{\phi'}_{j',n'} \leftarrow h^{\phi'}_{j',n'} - \eta_h \frac{\partial \tilde{C}_{IS}}{\partial h^{\phi'}_{j',n'}} \tag{5.14}$$

where η_u, η_w, and η_h are the positive learning rate. Based on [84], the positive learning rate can be set to the followings:

$$\begin{aligned}&\eta_u = \frac{\tilde{u}_{j,f}}{\sum_{n}\tilde{z}_{fn}{}^{-1}\sum_{\phi,\tau}\tilde{w}^{\tau}_{f-\phi,j}h^{\phi}_{j,n-\tau}}, \quad \eta_w = \frac{\tilde{w}^{\tau'}_{f',j'}}{\sum_{\phi,n}\tilde{z}_{f'+\phi,n}{}^{-1}\tilde{u}_{j',f'+\phi}h^{\phi}_{j',n-\tau'}} \\ &\text{and} \quad \eta_h = \frac{h^{\phi'}_{j',n'}}{\sum_{f,\tau}\tilde{z}_{f,n'+\tau}{}^{-1}\tilde{u}_{j',f}\tilde{w}^{\tau}_{f-\phi',j'} + \lambda^{\phi'}_{j',n'}}\end{aligned} \tag{5.15}$$

Using (5.14) and (5.15), the multiplicative update (MU) rules are obtained where for $u_{j,f}$, the update is given by

$$
\begin{aligned}
u_{j,f} &\leftarrow \tilde{u}_{j,f} - \frac{\tilde{u}_{j,f}\left(-\sum_n \left|y_{f,n}\right|^2 \tilde{z}_{f,n}^{-2}\sum_{\phi,\tau}\tilde{w}_{f-\phi,j}^{\tau}h_{j,n-\tau}^{\phi} + \sum_n \tilde{z}_{f,n}^{-1}\sum_{\phi,\tau}\tilde{w}_{f-\phi,j}^{\tau}h_{j,n-\tau}^{\phi}\right)}{\sum_n \tilde{z}_{f,n}^{-1}\sum_{\phi,\tau}\tilde{w}_{f-\phi,j}^{\tau}h_{j,n-\tau}^{\phi}} \\
&= \tilde{u}_{j,f}\left(\frac{\sum_n \tilde{z}_{f,n}^{-1}\sum_{\phi,\tau}\tilde{w}_{f-\phi,j}^{\tau}h_{j,n-\tau}^{\phi} + \sum_n \left|y_{f,n}\right|^2 \tilde{z}_{f,n}^{-2}\sum_{\phi,\tau}\tilde{w}_{f-\phi,j}^{\tau}h_{j,n-\tau}^{\phi} - \sum_n \tilde{z}_{f,n}^{-1}\sum_{\phi,\tau}\tilde{w}_{f-\phi,j}^{\tau}h_{j,n-\tau}^{\phi}}{\sum_n \tilde{z}_{f,n}^{-1}\sum_{\phi,\tau}\tilde{w}_{f-\phi,j}^{\tau}h_{j,n-\tau}^{\phi}}\right) \\
&= \tilde{u}_{j,f}\left(\frac{\sum_n \left|y_{f,n}\right|^2 \tilde{z}_{f,n}^{-2}\sum_{\phi,\tau}\tilde{w}_{f-\phi,j}^{\tau}h_{j,n-\tau}^{\phi}}{\sum_n \tilde{z}_{f,n}^{-1}\sum_{\phi,\tau}\tilde{w}_{f-\phi,j}^{\tau}h_{j,n-\tau}^{\phi}}\right)
\end{aligned}
\tag{5.16}
$$

Similarly, the MU rules for $w_{f,j}^{\tau}$ and $h_{j,n}^{\phi}$ respectively gives

$$
\begin{aligned}
w_{f',j'}^{\tau'} &\leftarrow \tilde{w}_{f',j'}^{\tau'} - \frac{\tilde{w}_{f',j'}^{\tau'}\left(-\sum_{\phi,n}\left|y_{f'+\phi.n}\right|^2 \tilde{z}_{f'+\phi,n}^{-2}\tilde{u}_{j',f'+\phi}h_{j',n-\tau'}^{\phi} + \sum_{\phi,n}\tilde{z}_{f'+\phi,n}^{-1}\tilde{u}_{j',f'+\phi}h_{j',n-\tau'}^{\phi}\right)}{\sum_{\phi,n}\tilde{z}_{f'+\phi,n}^{-1}\tilde{u}_{j',f'+\phi}h_{j',n-\tau'}^{\phi}} \\
&= \tilde{w}_{f',j'}^{\tau'}\left(\frac{\sum_{\phi,n}\tilde{z}_{f'+\phi,n}^{-1}\tilde{u}_{j',f'+\phi}h_{j',n-\tau'}^{\phi} + \sum_{\phi,n}\left|y_{f'+\phi.n}\right|^2 \tilde{z}_{f'+\phi,n}^{-2}\tilde{u}_{j',f'+\phi}h_{j',n-\tau'}^{\phi} - \sum_{\phi,n}\tilde{z}_{f'+\phi,n}^{-1}\tilde{u}_{j',f'+\phi}h_{j',n-\tau'}^{\phi}}{\sum_{\phi,n}\tilde{z}_{f'+\phi,n}^{-1}\tilde{u}_{j',f'+\phi}h_{j',n-\tau'}^{\phi}}\right) \\
&= \tilde{w}_{f',j'}^{\tau'}\left(\frac{\sum_{\phi,n}\left|y_{f'+\phi,n}\right|^2 \tilde{z}_{f'+\phi,n}^{-2}\tilde{u}_{j',f'+\phi}h_{j',n-\tau'}^{\phi}}{\sum_{\phi,n}\tilde{z}_{f'+\phi,n}^{-1}\tilde{u}_{j',f'+\phi}h_{j',n-\tau'}^{\phi}}\right)
\end{aligned}
\tag{5.17}
$$

and as for $h_{j,n}^{\phi}$, the update is given by

$$
\begin{aligned}
h_{j',n'}^{\phi'} &\leftarrow h_{j',n'}^{\phi'} - \frac{h_{j',n'}^{\phi'}\left(-\sum_{f,\tau}\left|y_{f,n'+\tau}\right|^2 \tilde{z}_{f,n'+\tau}^{-2}\tilde{u}_{j',f}\tilde{w}_{f-\phi',j'}^{\tau} + \sum_{f,\tau}\tilde{z}_{f,n'+\tau}^{-1}\tilde{u}_{j',f}\tilde{w}_{f-\phi',j'}^{\tau} + \sum_{j,n,\phi}\lambda_{j',n'}^{\phi'}\right)}{\sum_{f,\tau}\tilde{z}_{f,n'+\tau}^{-1}\tilde{u}_{j',f}\tilde{w}_{f-\phi',j'}^{\tau} + \lambda_{j',n'}^{\phi'}} \\
&= h_{j',n'}^{\phi'}\left(\frac{\sum_{f,\tau}\tilde{z}_{f,n'+\tau}^{-1}\tilde{u}_{j',f}\tilde{w}_{f-\phi',j'}^{\tau} + \lambda_{j',n'}^{\phi'} + \sum_{f,\tau}\left|y_{f,n'+\tau}\right|^2 \tilde{z}_{f,n'+\tau}^{-2}\tilde{u}_{j',f}\tilde{w}_{f-\phi',j'}^{\tau} - \sum_{f,\tau}\tilde{z}_{f,n'+\tau}^{-1}\tilde{u}_{j',f}\tilde{w}_{f-\phi',j'}^{\tau} - \sum_{j,n,\phi}\lambda_{j',n'}^{\phi'}}{\sum_{f,\tau}\tilde{z}_{f,n'+\tau}^{-1}\tilde{u}_{j',f}\tilde{w}_{f-\phi',j'}^{\tau} + \lambda_{j',n'}^{\phi'}}\right) \\
&= h_{j',n'}^{\phi'}\left(\frac{\sum_{f,\tau}\left|y_{f,n'+\tau}\right|^2 \tilde{z}_{f,n'+\tau}^{-2}\tilde{u}_{j',f}\tilde{w}_{f-\phi',j'}^{\tau}}{\sum_{f,\tau}\tilde{z}_{f,n'+\tau}^{-1}\tilde{u}_{j',f}\tilde{w}_{f-\phi',j'}^{\tau} + \lambda_{j',n'}^{\phi'}}\right)
\end{aligned}
\tag{5.18}
$$

For the sparsity term, the update is obtained by solving $\frac{\partial \tilde{C}_{IS}}{\partial \lambda_{j',n'}^{\phi'}} = 0$ which leads to

$$
\begin{aligned}
\frac{\partial \tilde{C}_{IS}}{\partial \lambda_{j',n'}^{\phi'}} &= \frac{\partial\left(\sum_f\sum_n\left(\frac{\left|y_{f,n}\right|^2}{\tilde{z}_{f,n}} - \log\frac{\left|y_{f,n}\right|^2}{\tilde{z}_{f,n}} - 1\right) + \sum_{j,n,\phi} h_{j,n}^{\phi}\lambda_{j,n}^{\phi} - \sum_{j,n,\phi}\log\lambda_{j,n}^{\phi}\right)}{\partial \lambda_{j',n'}^{\phi'}} \\
&= h_{j',n'}^{\phi'} - \frac{1}{\lambda_{j',n'}^{\phi'}}
\end{aligned}
\tag{5.19}
$$

Therefore, the solution for $\lambda_{j',n'}^{\phi'}$ is given by

$$
\lambda_{j',n'}^{\phi'} = \frac{1}{h_{j',n'}^{\phi'}} \tag{5.20}
$$

To accommodate for adaptive tracking of $\lambda_{j',n'}^{\phi'}$, we may modify (5.20) to

$$\lambda_{j',n}^{\phi'}(n_{iter}) = \alpha\lambda_{j',n'}^{\phi'}(n_{iter}-1) + \frac{(1-\alpha)}{h_{j',n'}^{\phi'} + \varepsilon} \tag{5.21}$$

where n_{iter} denotes iteration number, α is a constant and ε is the threshold to prevent division by zero. If $\alpha = 0$, then (5.21) reduces to (5.20). If $\alpha = 1$, this then corresponds to constant sparsity. Thus, $\alpha \in [0,1]$ and it is found that $\alpha = 0.9$ yields the best performance. The multiplicative updating rules for (5.16), (5.17), (5.18) and (5.21) can be written in matrix notation as

$$\mathbf{u}_j \leftarrow \tilde{\mathbf{u}}_j . \frac{\left(\tilde{\mathbf{Z}}^{-2} . |\mathbf{Y}|^{.2} . \tilde{\mathbf{P}}_j\right)\mathbf{1}_{N\times 1} + \tilde{\mathbf{u}}_j \mathrm{diag}\left(\mathbf{1}\left(\mathbf{1}_{N\times 1} . \tilde{\mathbf{u}}_j\right)\right)}{\left(\tilde{\mathbf{Z}}^{-1} . \tilde{\mathbf{P}}_j\right)\mathbf{1}_{N\times 1} + \tilde{\mathbf{u}}_j \mathrm{diag}\left(\mathbf{1}\left(\left(\tilde{\mathbf{Z}}^{-2} . |\mathbf{Y}|^{.2} . \tilde{\mathbf{P}}_j\right)\mathbf{1}_{N\times 1} . \tilde{\mathbf{u}}_j\right)\right)} \tag{5.22}$$

where $\tilde{\mathbf{P}}_j = \sum_{\tau,\phi} \overset{\downarrow\phi}{\tilde{\mathbf{W}}}_j^{\tau} \overset{\rightarrow\tau}{\mathbf{H}}_j^{\phi}$ and vector $\mathbf{1}_{N\times 1}$ is a *N*-vector of ones. For **W** update, it is written as

$$\mathbf{W}_j^{\tau} \leftarrow \tilde{\mathbf{W}}_j^{\tau} . \left(\frac{\sum_{\phi} \mathrm{diag}(\overset{\uparrow\phi}{\tilde{\mathbf{u}}}_j) \left(\left(\overset{\uparrow\phi}{\tilde{\mathbf{Z}}} \right)^{-2} . |\mathbf{Y}|^{.2} \right) \mathbf{H}_j^{\phi\ \mathrm{T}} + \tilde{\mathbf{W}}_j^{\tau}\, \mathrm{diag}\left(\sum_{\tau} \mathbf{1} \left(\mathrm{diag}(\overset{\uparrow\phi}{\tilde{\mathbf{u}}}_j) \left(\overset{\uparrow\phi}{\tilde{\mathbf{Z}}} \right)^{-1} \mathbf{H}_j^{\phi\ \mathrm{T}} \right) . \tilde{\mathbf{W}}_j^{\tau} \right)}{\sum_{\phi} \mathrm{diag}(\overset{\uparrow\phi}{\tilde{\mathbf{u}}}_j) \left(\overset{\uparrow\phi}{\tilde{\mathbf{Z}}} \right)^{-1} \mathbf{H}_j^{\phi\ \mathrm{T}} + \tilde{\mathbf{W}}_j^{\tau}\, \mathrm{diag}\left(\sum_{\tau} \mathbf{1} \left(\mathrm{diag}(\overset{\uparrow\phi}{\tilde{\mathbf{u}}}_j) \left(\left(\overset{\uparrow\phi}{\tilde{\mathbf{Z}}} \right)^{-2} . |\mathbf{Y}|^{.2} \right) \mathbf{H}_j^{\phi\ \mathrm{T}} \right) . \tilde{\mathbf{W}}_j^{\tau} \right)} \right) \tag{5.23}$$

And similarly for **H,**

$$\mathbf{H}_j^{\phi} \leftarrow \mathbf{H}_j^{\phi}.\frac{\sum_{\tau}\left(\mathrm{diag}(\tilde{\mathbf{u}}_j)\,\tilde{\tilde{\mathbf{W}}}_j^{\tau}\right)^{\mathrm{T}}\left(\left(\tilde{\mathbf{Z}}^{\leftarrow\tau}\right)^{-2}.\left|\mathbf{Y}^{\leftarrow\tau}\right|^{2}\right)}{\sum_{\tau}\left(\mathrm{diag}(\tilde{\mathbf{u}}_j)\,\tilde{\tilde{\mathbf{W}}}_j^{\tau}\right)^{\mathrm{T}}\left(\tilde{\mathbf{Z}}^{\leftarrow\tau}\right)^{-1}+\boldsymbol{\Lambda}_j^{\phi}} \qquad (5.24)$$

.

As for sparsity parameter, $\boldsymbol{\Lambda}$ update is expressed as

$$\boldsymbol{\Lambda}_j^{\phi} \leftarrow \alpha\boldsymbol{\Lambda}_j^{\phi} + \frac{(1-\alpha)}{\mathbf{H}_j^{\phi}+\varepsilon} \qquad (5.25)$$

where the division operation is element-wise. Note that α parameter is selected after conducting the Monte-Carlo experiment over 100 independent realisations of each mixture. Table 5.1 presents the main steps of the proposed algorithm for blind separation in convolutive mixture. The stopping criterion is given by $\left(\tilde{C}_{IS}(n_{iter}-1)-\tilde{C}_{IS}(n_{iter})\right)/\tilde{C}_{IS}(n_{iter})<\psi$ where $\psi=10^{-6}$ is the threshold for ascertaining the convergence.

Table 5.1: Proposed FC-SNMF2D algorithm

1. Initialise **U**, **W** and **H** with nonnegative random values.

2. Normalise $\tilde{w}^{\tau}_{f,j} = \dfrac{w^{\tau}_{f,j}}{\sqrt{\sum_{\tau,f}(w^{\tau}_{f,j})^2}}$ and $\tilde{u}_{j,f} = \dfrac{u_{j,f}}{\sqrt{\sum_{f}(u_{j,f})^2}}$

3. Compute $\tilde{\mathbf{P}}_j = \sum_{\tau,\phi} \overset{\downarrow\phi}{\tilde{\mathbf{W}}}{}^{\tau}_{j} \overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi}_{j}$ and $\tilde{\mathbf{Z}} = \sum_{j,\tau,\phi} \tilde{\mathbf{U}}_j \overset{\downarrow\phi}{\tilde{\mathbf{W}}}{}^{\tau}_{j} \overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi}_{j}$

4. Update $\mathbf{u}_j \leftarrow \tilde{\mathbf{u}}_j . \dfrac{\left(\tilde{\mathbf{Z}}^{-2}.|\mathbf{Y}|^{.2}.\tilde{\mathbf{P}}_j\right)\mathbf{1}_{N\times 1} + \tilde{\mathbf{u}}_j \mathrm{diag}\left(\mathbf{1}\left(\mathbf{1}_{N\times 1}.\tilde{\mathbf{u}}_j\right)\right)}{\left(\tilde{\mathbf{Z}}^{-1}.\tilde{\mathbf{P}}_j\right)\mathbf{1}_{N\times 1} + \tilde{\mathbf{u}}_j \mathrm{diag}\left(\mathbf{1}\left(\left(\tilde{\mathbf{Z}}^{-2}.|\mathbf{Y}|^{.2}.\tilde{\mathbf{P}}_j\right)\mathbf{1}_{N\times 1}.\tilde{\mathbf{u}}_j\right)\right)}$

5. Compute $\tilde{\mathbf{Z}} = \sum_{j,\tau,\phi} \tilde{\mathbf{U}}_j \overset{\downarrow\phi}{\tilde{\mathbf{W}}}{}^{\tau}_{j} \overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi}_{j}$

6. Assign $\boldsymbol{\Lambda}^{\phi}_{j} \leftarrow \alpha \boldsymbol{\Lambda}^{\phi}_{j} + \dfrac{(1-\alpha)}{\mathbf{H}^{\phi}_{j} + \varepsilon}$

7. Update $\mathbf{H}^{\phi}_{j} \leftarrow \mathbf{H}^{\phi}_{j} . \dfrac{\sum_{\tau}\left(\mathrm{diag}(\tilde{\mathbf{u}}_j)\overset{\downarrow\phi}{\tilde{\mathbf{W}}}{}^{\tau}_{j}\right)^{\mathrm{T}}\left(\left(\overset{\leftarrow\tau}{\tilde{\mathbf{Z}}}\right)^{-2}.|\mathbf{Y}|^{.2}\right)}{\sum_{\tau}\left(\mathrm{diag}(\tilde{\mathbf{u}}_j)\overset{\downarrow\phi}{\tilde{\mathbf{W}}}{}^{\tau}_{j}\right)^{\mathrm{T}}\left(\overset{\leftarrow\tau}{\tilde{\mathbf{Z}}}\right)^{-1} + \boldsymbol{\Lambda}^{\phi}_{j}}$

8. Compute $\tilde{\mathbf{Z}} = \sum_{j,\tau,\phi} \tilde{\mathbf{U}}_j \overset{\downarrow\phi}{\tilde{\mathbf{W}}}{}^{\tau}_{j} \overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi}_{j}$

9. Update $\mathbf{W}^{\tau}_{j} \leftarrow \tilde{\mathbf{W}}^{\tau}_{j} . \left(\dfrac{\sum_{\phi}\mathrm{diag}(\overset{\uparrow\phi}{\tilde{\mathbf{u}}}_j)\left(\left(\overset{\uparrow\phi}{\tilde{\mathbf{Z}}}\right)^{-2}.|\mathbf{Y}|^{.2}\right)\overset{\rightarrow\tau\ \mathrm{T}}{\mathbf{H}}{}^{\phi}_{j} + \tilde{\mathbf{W}}^{\tau}_{j}\mathrm{diag}\left(\sum_{\tau}\mathbf{1}\left(\mathrm{diag}(\overset{\uparrow\phi}{\tilde{\mathbf{u}}}_j)\left(\overset{\uparrow\phi}{\tilde{\mathbf{Z}}}\right)^{-2}.|\mathbf{Y}|^{.2}\right)\overset{\rightarrow\tau\ \mathrm{T}}{\mathbf{H}}{}^{\phi}_{j}.\tilde{\mathbf{W}}^{\tau}_{j}\right)}{\sum_{\phi}\mathrm{diag}(\overset{\uparrow\phi}{\tilde{\mathbf{u}}}_j)\left(\overset{\uparrow\phi}{\tilde{\mathbf{Z}}}\right)^{-1}\overset{\rightarrow\tau\ \mathrm{T}}{\mathbf{H}}{}^{\phi}_{j} + \tilde{\mathbf{W}}^{\tau}_{j}\mathrm{diag}\left(\sum_{\tau}\mathbf{1}\left(\mathrm{diag}(\overset{\uparrow\phi}{\tilde{\mathbf{u}}}_j)\left(\left(\overset{\uparrow\phi}{\tilde{\mathbf{Z}}}\right)^{-2}.|\mathbf{Y}|^{.2}\right)\overset{\rightarrow\tau\ \mathrm{T}}{\mathbf{H}}{}^{\phi}_{j}\right).\tilde{\mathbf{W}}^{\tau}_{j}\right)} \right)$

10. Compute $\tilde{\mathbf{Z}} = \sum_{j,\tau,\phi} \tilde{\mathbf{U}}_j \overset{\downarrow\phi}{\tilde{\mathbf{W}}}{}^{\tau}_{j} \overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi}_{j}$

11. Repeat from steps 2 till 10 until convergence is achieved where rate of cost change is below a prescribed threshold, ψ

5.2.4 Reconstruction of the separated source images

For the proposed method, the estimated sources images are defined as $\left|\hat{x}_{j,f,n}^{im}\right|^2 = \sum_j \sum_\tau \sum_\phi \tilde{u}_{j,f} \tilde{w}_{f-\phi,j}^{\tau} h_{j,n-\tau}^{\phi}$. By using wiener mask e.g. $m_{j,f,n} = \left|\hat{x}_{j,f,n}^{im}\right|^2 \Big/ \sum_j \left|\hat{x}_{j,f,n}^{im}\right|^2$, the estimated sources images is computed through

$$\left|\hat{x}_{j,f,n}^{im}\right|^2 = m_{j,f,n} \left|y_{f,n}\right|^2 \tag{5.26}$$

Then, the time domain estimated signal $\hat{x}_j(t)$ is obtained by doing the inverse mapping of log-frequency axis to the original frequency axis and then followed by inverse STFT.

5.3 Results and Analysis

5.3.1 Experiment set-up

The proposed method was tested on recorded audio signals. The objective was to separate the single channel mixture in convolutive mixture. Several experiments have been conducted under different conditions to investigate the algorithm performance. All computations and analysis are performed using a PC with Intel core i3 M380 @ 2.53GHz and 3GB RAM. The experiments consist of two audio sources namely piano and trumpet sound. The mixture is approximately 5s long and sampled at 16kHz. In this experiment, STFT using 2048-point Hamming window with 50% overlap was used and the frequency axis of the obtained spectrogram was logarithmically scaled

and grouped into 175 frequency bins in the range of 50Hz to 8kHz (given $f_s = 16\text{kHz}$) with 24 bins per octave and the bandwidth follows the constant-Q rule [122]. Synthetic convolutive mixture of the sources is produced using the Room Simulation (Roomsim) toolbox [124]. We used the same setting as in section 4.3.2 in previous chapter for the Roomsim simulation where we simulate omnidirectional microphone in a room of dimension 4.45m x 3.00m x 3.00m. The distance between the sources and the microphone is 2m and are located 1.2m from the floor. Please refer Figure 4.6 for the plot of impulse response for source 1 and source 2. The reverberation time (RT60) was calculated to be 0.35s for the 1000Hz band.

For audio separation, after conducting the Monte-Carlo experiments over 100 independent realisations of the mixture, the parameters of the convolutive factors selected such that $\tau = \{0,...,7\}$ and $\phi = \{0,...,31\}$. This is the best reasonable parameter setting to represent the temporal code and spectral basis in the factorization for most of music signals since audio signal have higher variability and require higher number of τ shift and ϕ shift to capture the temporal dependency of the frequency pattern in audio signal. For performance evaluation, the signal-to-distortion ratio (SDR), source-to-artifacts ratio (SAR) and source-to-interference ratio (SIR) will be used to measures an overall sound quality of the source separation. The MATLAB implementation of these measures can be found in [109, 110]. Figure 5.1 shows the time and TF domains of the source images of piano and trumpet as well as its convolutive mixture. The piano and trumpet play a different short melodic with many different distinct note. This is a challenging task for single channel separation since

both piano and trumpet overlap in time while the piano notes are scattered and interspersed between frequencies with the trumpet notes.

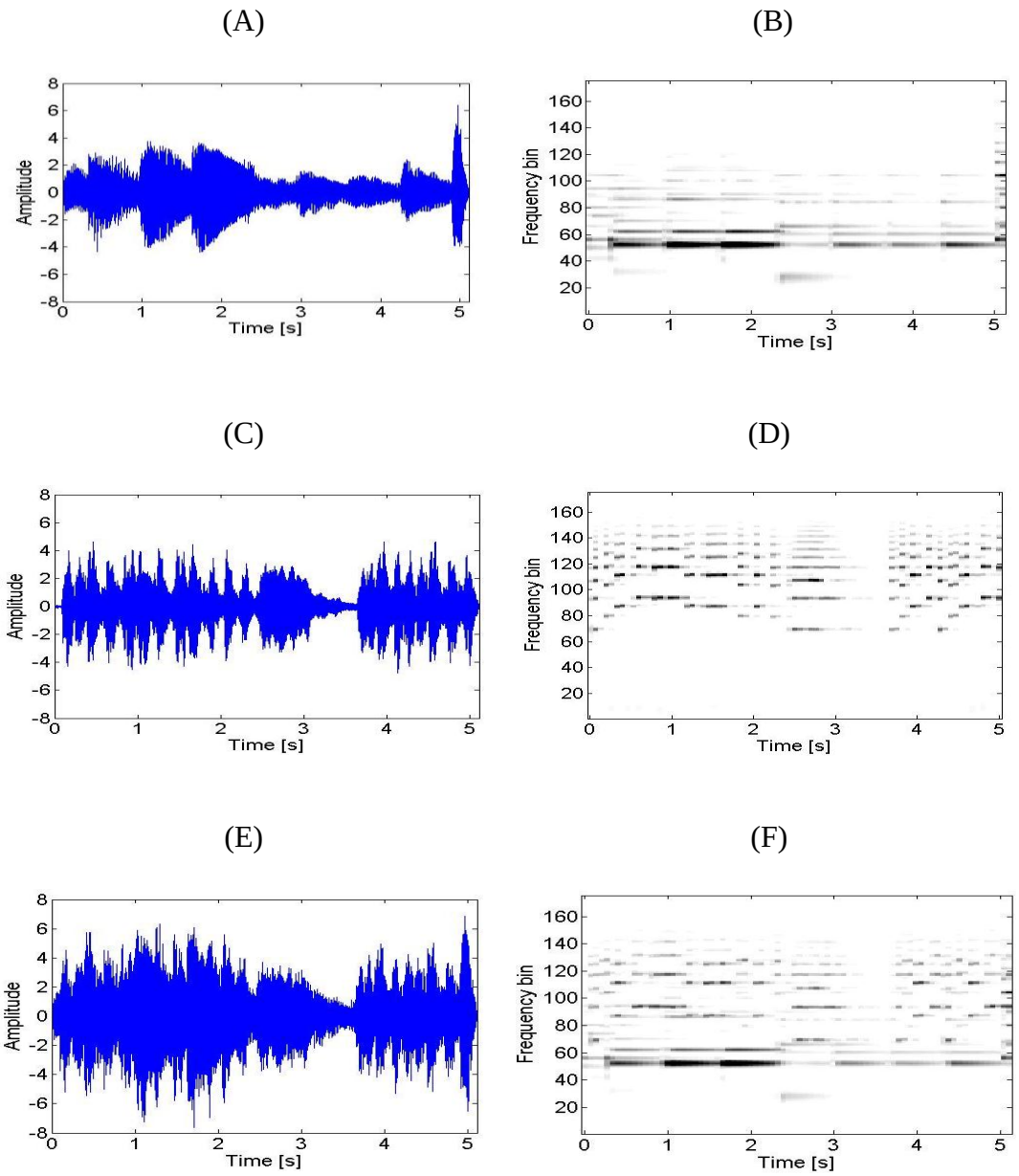


Figure 5.1: Time-domain representation and log-frequency spectrogram of piano (top panels), trumpet (middle panels) and mixed signals (bottom panels).

5.3.2 Evaluation of proposed algorithm

In this experiment, the proposed method is evaluated by comparing the performance under different sparsity regularity. We will also show the importance of adaptive behaviour of the sparsity parameter in reducing the ambiguity in the separated sources. The following three cases will be investigated:

Case (i): No sparseness, $\lambda_{j,n}^{\phi} = \lambda = 0$ for all j, n, ϕ.

Case (ii): Fixed and constant sparseness, $\lambda_{j,n}^{\phi} = \lambda = 0.5$ for all j, n, ϕ

Case (iii): Adaptive sparseness according to (5.25).

For each case, the same initial values of **U**, **W** and **H** are used which were obtained randomly.

5.3.2.1 Estimated spectral bases and temporal codes

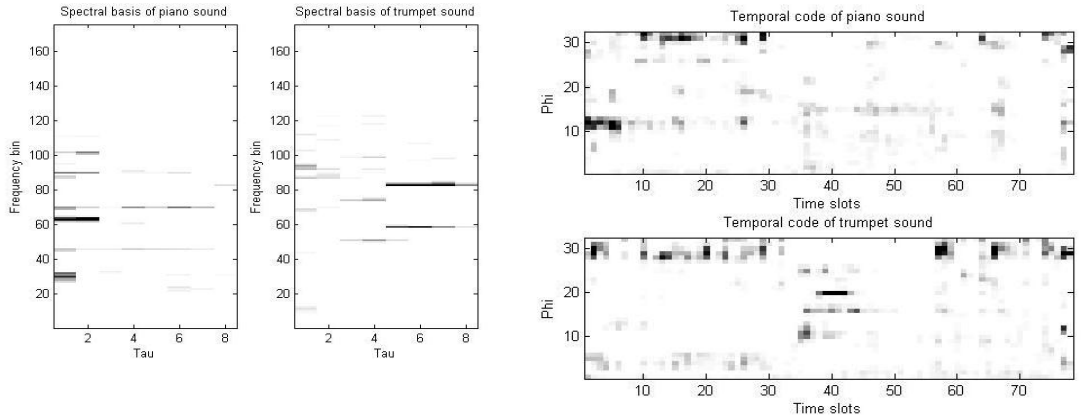
Figure 5.2 shows the matrix factorization results in term of spectral bases $\mathbf{W}_j^{\tau}$ and temporal codes $\mathbf{H}_j^{\phi}$ for case (i), (ii) and (iii), respectively. In Figure 5.2(A) reveals that the resulting factorizations is under-sparse since no sparsity is imposed. This is clearly shown by the spreading of the estimated temporal codes. Figure 5.2(B) reveals the over-sparse factorization where majority of the temporal codes have been discarded. In both cases, poor separation of the mixture has been resulted since the estimation of $\mathbf{W}_j^{\tau}$ and $\mathbf{H}_j^{\phi}$ are not optimal as evidenced by the spreading and discarding of important information in the temporal codes. On the other hand, Figure 5.2(C) shows the obtained factorization is optimally-sparse by using the proposed

adaptive sparsity parameter. Each temporal code was assigned with a sparsity parameter which is individually and adaptively tuned to produce the optimal results.

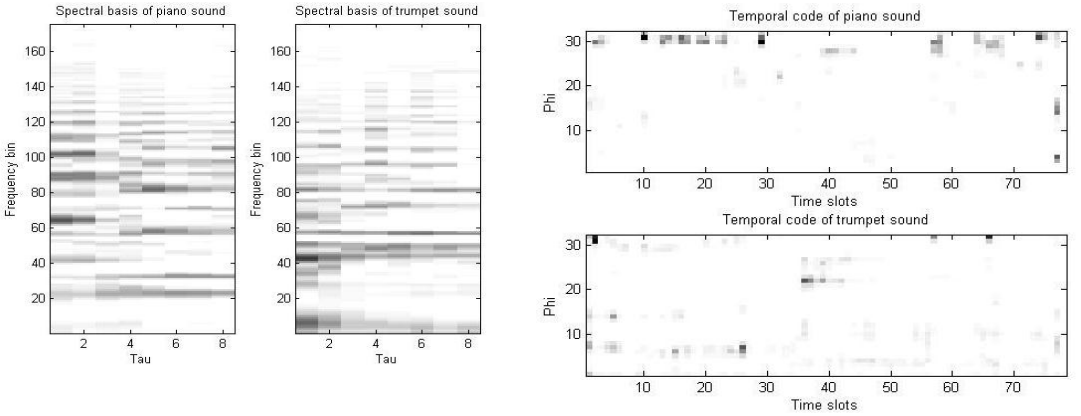
5.3.2.2 Source separation results

In this sub-section, the source separation performance of the convolutive mixture of piano and trumpet sound for each case will be shown. Figures 5.3 and 5.4 show the separation results in the log-frequency spectrogram and the time domain for case (i) to (iii), respectively. Compared with the original image sources in Figure 5.1, it is visually clear that the separation without the sparsity constraint has led to poor result since the factorization still contains the mixed signal (indicated by the box marked area) as in panels (A) and (B) for case (i). The estimation of parameters $\mathbf{W}_j^{\tau}$ and $\mathbf{H}_j^{\phi}$ is slightly coarse which has led to the ambiguity in the estimation of the source images. For case (ii), the result of factorization with constant $\lambda_{j,n}^{\phi}$ is shown in panel (C) and (D). There is still some small amount of mixed signal in the separated signals. This is due to the fact that the sparsity is only imposed uniformly on all the codes and they are not optimal. This situation raises a big issue which led the sparse factorization being either too sparse or not sparse enough for $\mathbf{H}^{\phi}$. As for case (iii) which is shown in panels (E) and (F), the proposed method has yielded the best performance with the source images almost fully recovered. In the proposed method, good separation result has been achieved because the sparsity on temporal code $\mathbf{H}_j^{\phi}$ is imposed element-wise and is adaptively tuned so that each individual code in $\mathbf{H}_j^{\phi}$ have an optimal sparse value.

(A)



(B)



(C)

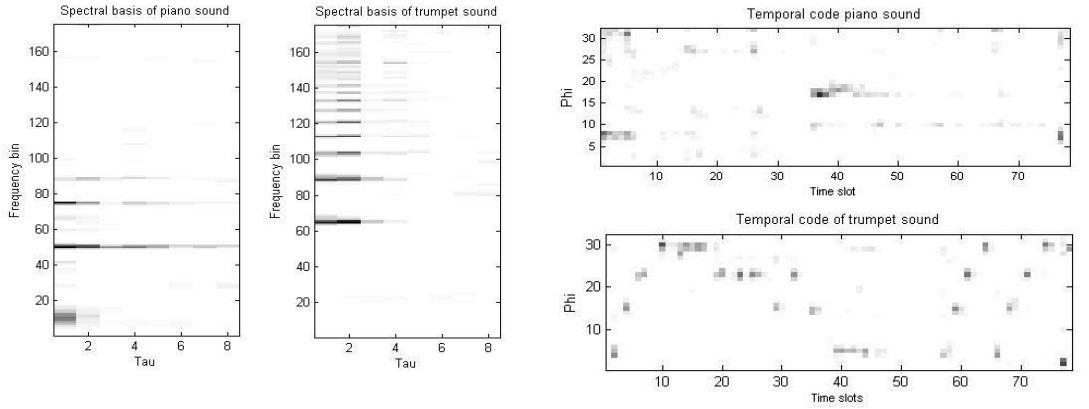


Figure 5.2: Estimated $\mathbf{W}_j^{\tau}$ and $\mathbf{H}_j^{\phi}$ for (A) case (i), (B) case (ii), and (C) case (iii).

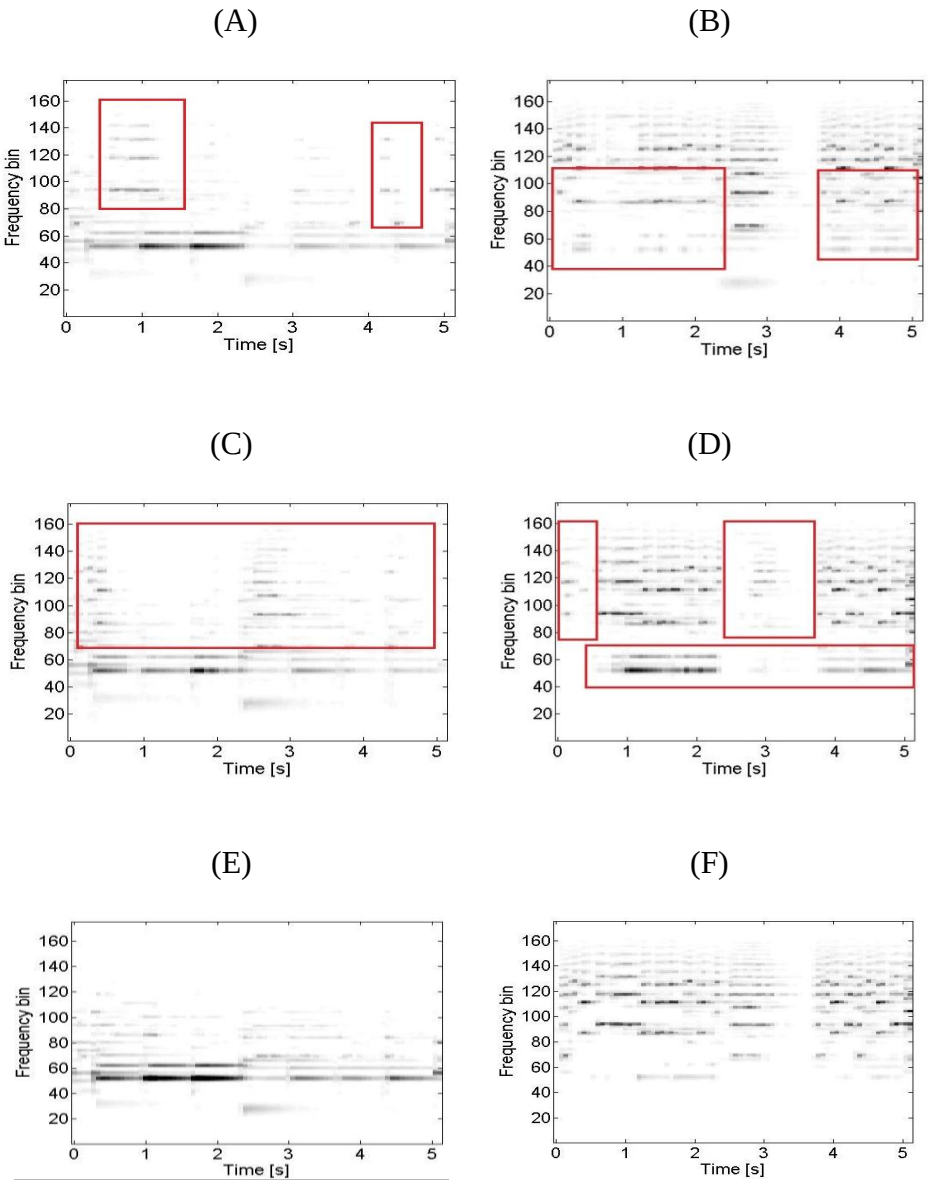


Figure 5.3: Separated signal in spectrogram. (A)-(B): piano and trumpet sound for case (i). (C)-(D): piano and trumpet sound for case (ii). (E)-(F): piano and trumpet sound for case (iii).

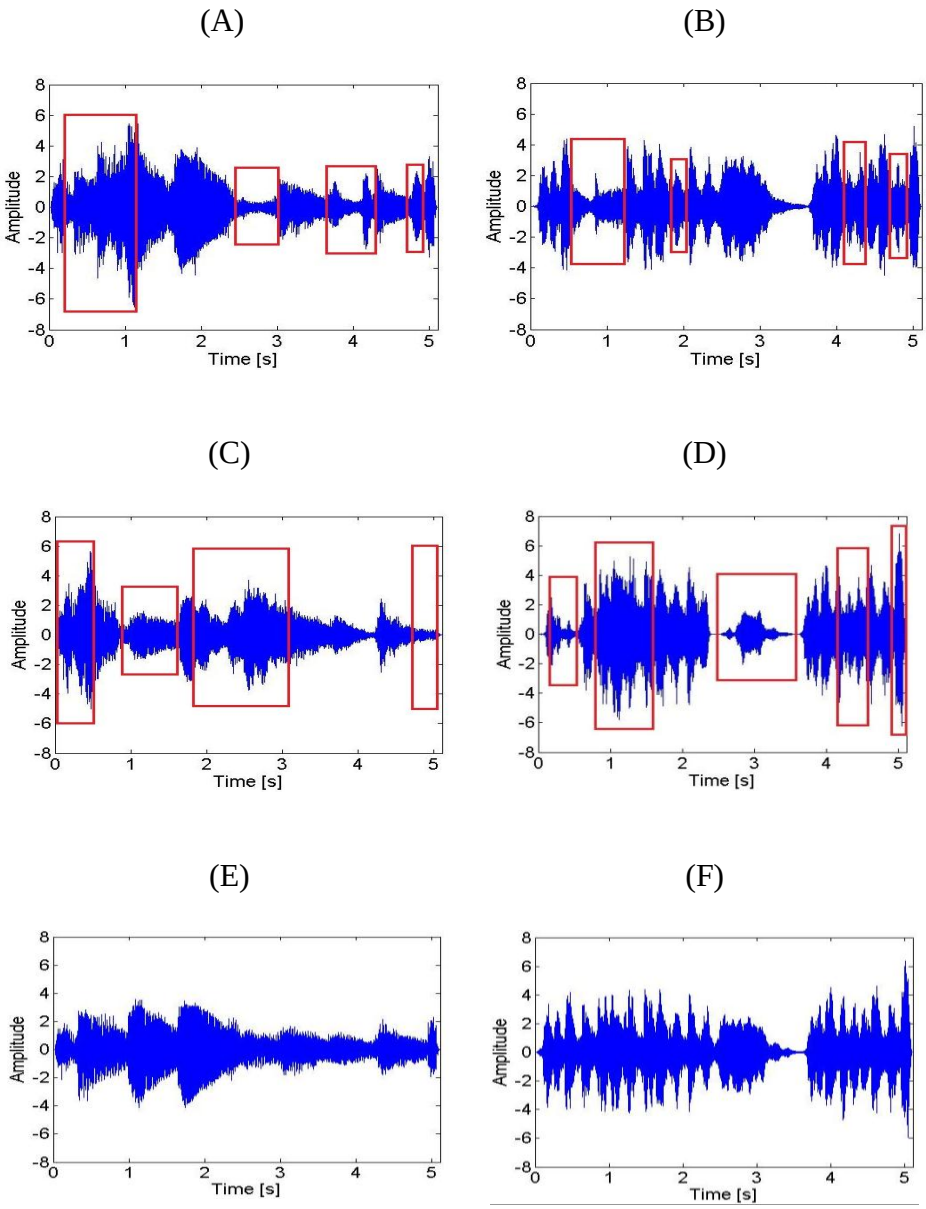


Figure 5.4: Separated signal in time domain (A)-(B): piano and trumpet sound for case (i). (C)-(D): piano and trumpet sound for case (ii). (E)-(F): piano and trumpet sound for case (iii).

Table 5.2: Performance comparison between different sparsity methods (dB)

Sparsity weight	Separated Piano			Separated trumpet		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
No sparsity	4.9	6.7	10.5	5.9	9.4	9.0
(Best) Constant sparsity	8.7	10.4	13.7	9.2	11.5	13.4
Adaptive sparsity	10.5	12.3	15.3	12.4	15.8	15.2

Table 5.2 shows the performance results in term of SDR, SIR and SAR. These results are averaged after conducting the Monte-Carlo experiment over 50 independent realisations of each mixture. The result shows that the proposed method with adaptive sparseness has yielded the best performance amongst all with SDR of 10.5 dB for the separated piano and 12.4 dB for the separated trumpet. This represents a 2.5dB per source improvement over the case of uniform constant sparsity. On the separate hand, when no sparsity is imposed onto the codes, the SDR result deteriorates as much as 6dB per source compared with the proposed adaptive sparsity method. From this result, it can be inferred that the sparsity constraints have significant effects on the separation performance. In addition, the results are ready to suggest that the performance of source separation had been undermined when the uniform constant sparsity scheme is used. On the contrary, improved performance can be obtained by allowing the sparsity parameters to be individually adapted for each element code.

Since Case (ii) represents the uniform constant sparsity, a question thus arises as to what is the best attainable sparsity value for $\lambda_{j,n}^{\phi}$ that gives the best separation performance. We investigate this by manually varying $\lambda_{j,n}^{\phi}$ over a range from 0 to 0.5 with every increment of 0.05. The obtained result has been plotted in Figure 5.5. As we increase the sparsity parameter in the algorithm, the performance also increases and it reaches a peak value when $\lambda_{j,n}^{\phi} = 0.2$ where the maximum average SDR value of 9 dB is obtained for each source. However, the SDR gradually decreases as $\lambda_{j,n}^{\phi}$ continues to increase until $\lambda_{j,n}^{\phi} = 0.4$ where the SDR is only 2dB. As soon as the sparsity parameter is further increased, the separation performance worsens due to the 'over-sparse' factorization where the spectral basis occurs too infrequently in the spectrogram. As a result, it cannot recover the source image due to the lack of information from the basis. In addition, it is practically difficult to select the appropriate value of sparsity parameter for matrix factorization to resolve the ambiguity. Thus, this points out the importance of imposing adaptive sparsity in each element code in order to obtain the optimal performance.

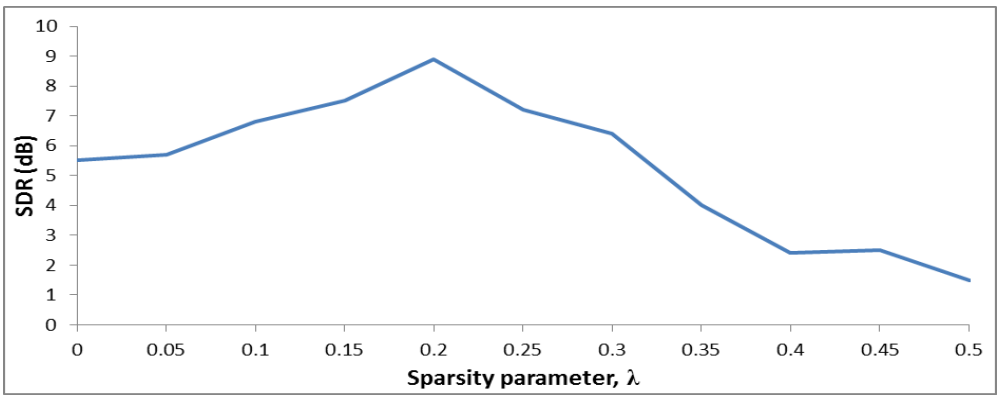


Figure 5.5: Separation result of the sparsity parameter with different constant value for $\lambda_{j,n}^{\phi}$.

5.3.2.3 Adaptive behavior of sparsity parameter

In this sub-section, experiment is carried out to demonstrate the adaptive behavior of the sparsity parameters. Several sparsity parameters have been selected to illustrate its adaptive behavior. Figure 5.6 shows the trajectory of four adaptive sparsity parameters $\lambda_{1,1}^{\phi=0}$, $\lambda_{1,6}^{\phi=0}$, $\lambda_{1,11}^{\phi=0}$ and $\lambda_{1,30}^{\phi=0}$ corresponding to their respective element codes. All sparsity parameters are initialized as $\lambda_{d,l}^{\phi} = 0.01$ for all d, l, ϕ. From Figure 5.6, it can be seen that even though the sparsity parameters started at the same initial condition, it is noted that the value of sparsity parameters are changing in order to adapt with the dynamics of the temporal code $\mathbf{H}_{j}^{\phi}$. This shows that each element code has its own sparseness. In addition, it is worth pointing out that in the case of piano and trumpet mixture the average SDR result rises up to 11.5dB, when $\lambda_{d,l}^{\phi}$ is adaptive (please refer to Table 5.2). This represents a 2.5dB per source improvement over the case of uniform constant sparsity. In percentage, this translates to an average improvement of 28% against the uniform constant sparsity. This is evident based on source separation performance as indicated in Table 5.2.

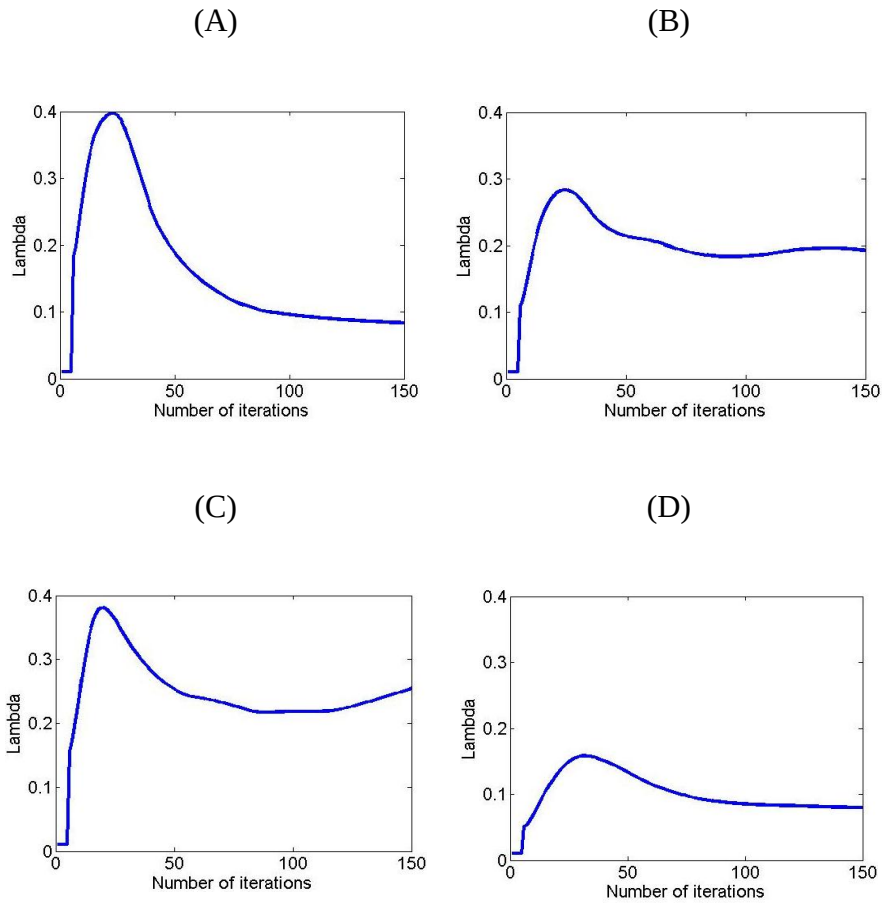


Figure 5.6: Trajectory of the sparsity parameters: (A) $\lambda_{1,1}^{\phi=0}$, (B) $\lambda_{1,6}^{\phi=0}$, (C) $\lambda_{1,11}^{\phi=0}$ and (D) $\lambda_{1,30}^{\phi=0}$.

5.3.2.4 Impact of convolutive mixing, U

In this sub-section, the impact of frequency mixing, **U** in proposed method is demonstrated. To observe this, we will evaluate the proposed algorithm where **U** is constant by simply set $\mathbf{U}_j = \mathbf{I}$ but the adaptation of $\mathbf{W}_j^{\tau}$, $\mathbf{H}_j^{\phi}$ and $\mathbf{\Lambda}_j^{\phi}$ still follows the proposed methodology. Figure 5.7 shows the separation result of the same convolutive mixture from previous experiment. By discounting the frequency

variation in the channels in the algorithm, the plots are clear to show that errors have accumulated in both separated sounds (highlighted by the red marked box) where some components have been attributed incorrectly. This is due to the assumption that **U** has the uniform value for all frequency which is not true for convolutive mixture. Consequently, this has led to misrepresentation of spectral basis and temporal in the separation. Comparing the separated signal in the plot of Figure 5.7(A)-(B) with the plot of proposed FC-SNMF2D in Figure 5.7(C)-(D), it is undoubtedly show the vital of constraining **U** in order to obtain the accurate result in separation.

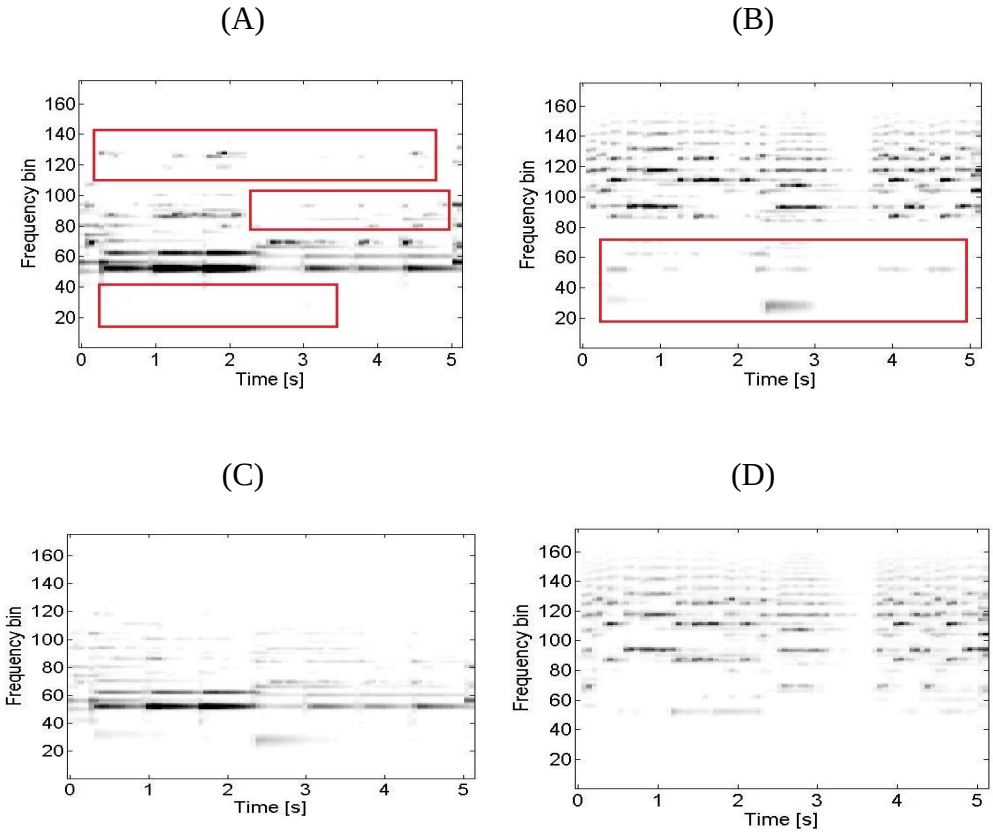


Figure 5.7: Separated piano and trumpet sound, respectively in TF domain using (A)-(B) Fixed $\mathbf{U}_j = \mathbf{I}$ (C)-(D) Proposed FC-SNMF2D

In Table 5.3, through updating the frequency mixing parameter the SDR performance of the algorithm has improved by 2.4 dB for separated piano and 3.5 dB for separated trumpet. This improvement indicates that for convolutive mixture, it is crucial for frequency distribution of the spectral basis be constrained through **U** to avoid the distortion in the decomposition. These results are averaged after conducting the Monte-Carlo experiment over 50 independent realisations of each mixture.

Table 5.3: Impact of **U** on separation performance

Method	**Separated piano**			**Separated trumpet**		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
Proposed Algorithm with $\mathbf{U}_j = \mathbf{I}$	8.1	9.6	13.7	8.9	11.1	13.2
Proposed Algorithm FC-SNMF2D	10.5	12.3	15.3	12.4	15.8	15.2

5.3.2.5 Convergence behaviour

In this sub-section, the convergence behaviour of the cost function in the proposed method is demonstrated. In this experiment, the algorithm was run for 1000 iterations from ten different random initialisations. Figure 5.8 shows the evolution of the cost function along the 1000 iterations. It can be seen that the cost function values decrease steadily and converged after 500 iterations. As for computational loads, the MATLAB implementation of the proposed algorithm takes about 4 minutes per 1000 iterations for this particular experiment.

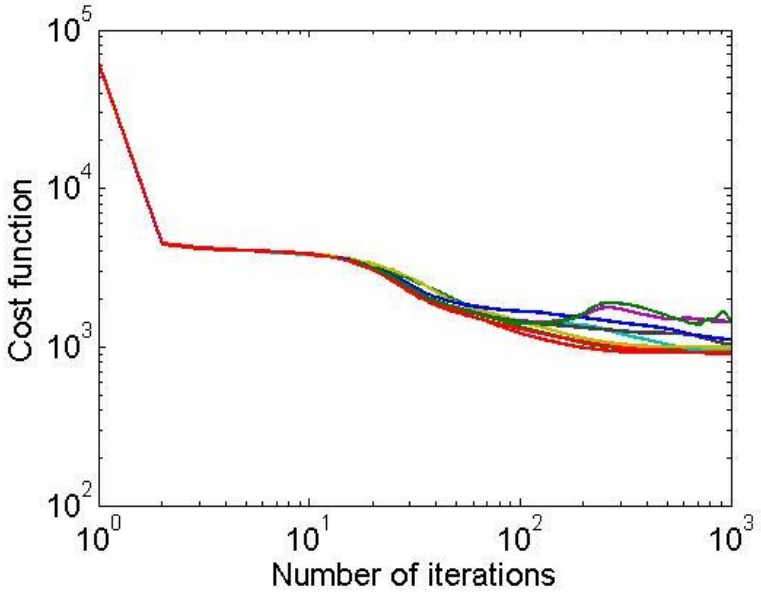


Figure 5.8: Evolution in log-log scale of the cost functions along the 1000 iterations of all 10 runs of the proposed algorithm.

5.3.3 Comparison between different cost function

In this sub-section, experiments using different cost functions are conducted to evaluate the efficiency of the proposed algorithm. Here we consider the Least Square (LS) and Kullback-Leibler (KL) divergence based on [126] and [127]. After conducting the Monte Carlo experiment over 50 independent realisations of each mixture, averaged separation results of FC-SNMF2D based on LS, KL and IS cost function is shown in Table 5.4. From Table 5.4, IS divergence outperforms those of LS distance and KL divergence with an average SDR of 3.4dB and 2.2dB, respectively. This is supported by the circumstance that the IS divergence embraces a desirable property of scale invariant which enables a more accurate representation of the factorization where lower power components cost as much as the higher power components. On the contrary, the Kullback-Leibler (KL) divergence and Least Square

(LS) distance whose estimation of lower power components is often ignored in favour of the higher power components. This leads to mixing ambiguities especially for lower power components in which case when they are incorporated together leads to major loss of spectral-temporal information of the sources.

Table 5.4: Performance comparison between different cost functions

Algorithms	**Separated piano**			**Separated trumpet**		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
LS FC-SNMF2D	8.0	8.9	9.3	8.1	9.2	9.5
KL FC-SNMF2D	9.1	11.4	13.3	9.4	11.7	13.9
IS FC-SNMF2D	10.5	12.3	15.3	12.4	15.8	15.2

5.3.4 Comparison with NMF-based method in convolutive mixture

In this experiment, comparison of the proposed method with the recent NMF-based method is carried out. The NMF method uses multiple components by grouping the individual spectral basis to estimate the image sources. We will compare the proposed method with the algorithm proposed in [117] and NMF with Temporal Continuity and Sparseness Criteria [43] (NMF-TCS). The optimisation is based on multiplicative update rule and convergence is set such that the rate of cost change is below $\psi = 10^{-6}$. In this NMF-based grouping method, choosing the number of components per source is crucial since different type of sources required different

number of component in order to perform optimally. Currently, there is no reliable NMF method for automatic estimation of the number of components and normally, this has to set manually. In order to obtain the baseline comparison of each method, all NMF algorithms are tested by factorizing the mixture signal into $J = 2,4,...,10$ components. Since more than two components are used and the tested methods are blind, there is no information to tell which component belongs to which source. Thus, we utilize the clustering method proposed in [43] where the original sources are used as reference to create component clusters for each source. After conducting a Monte-Carlo experiment of 100 independent trials, the number of components per source for both source 1 and source 2 that produces reliable separation has been determined to be 8. As for NMF-TCS, the temporal continuity α is chosen as [0,1,10,100,1000], sparseness weight β is chosen as [0,1,10,100,1000]. The best separation result is retained for comparison.

In Table 5.5, the proposed method shows a superior performance and outperformed the NMF [117] and NMF-TCS [43] methods by an average of 6.1 dB and 3.8 dB, respectively for both sources. This is because the spectral basis obtained by the NMF-based grouping method is still not adequate to capture the temporal dependency of the continuous frequency patterns within the signal. This has led to the ambiguity in each separated sources and contributed to poorer performance. Whereas in our proposed algorithm, the temporal information is considered by take into account the relative position of each spectrum using two dimensional factors of τ and ϕ. In addition, the adaptive sparseness imposed in the proposed algorithm reduced the component ambiguity which will result in significant high SDR performance. Note

that these results are averaged after conducting the Monte Carlo experiment over 50 independent realisations of each mixture.

Table 5.5: Performance comparison between different methods

Method	**Separated piano**			**Separated trumpet**		
	SDR	**SIR**	**SAR**	**SDR**	**SIR**	**SAR**
NMF [117]	5.8	11.2	6.5	5.0	7.7	9.2
NMF-TCS	7.9	9.1	9.0	7.4	8.3	8.5
FC-SNMF2D	10.5	12.3	15.3	12.4	15.8	15.2

5.3.5 Experiment using a live recorded sound

In this section, a live-recording of audio signals mixture is used as the mixture dataset. The setting employed for the live-recorded mixture corresponds very closely to the synthetic convolutive mixture by Roomsim as mention in section experiment set up. The instruments are played simultaneously through loudspeakers in a room and recorded using a passive microphone. The live-recorded mixture is sampled at 16 kHz. Figures 5.9 and 5.10 show the separation results in the time domain for piano-trumpet mixture and trumpet-drum mixture, respectively. From Figure 5.9, it can be seen that both separated piano and trumpet resemble the original signals. The continuous pattern of trumpet sound has been well separated with little mixed signal contaminating the

separated discrete piano sound. Similarly in Figure 5.10, our proposed method has also successfully separated the trumpet and drum sound. It can be visually seen that even though there is some slight portion of information discarded in both recovered trumpet and drum, the separated signals still preserve a good resemblance to the original signals.

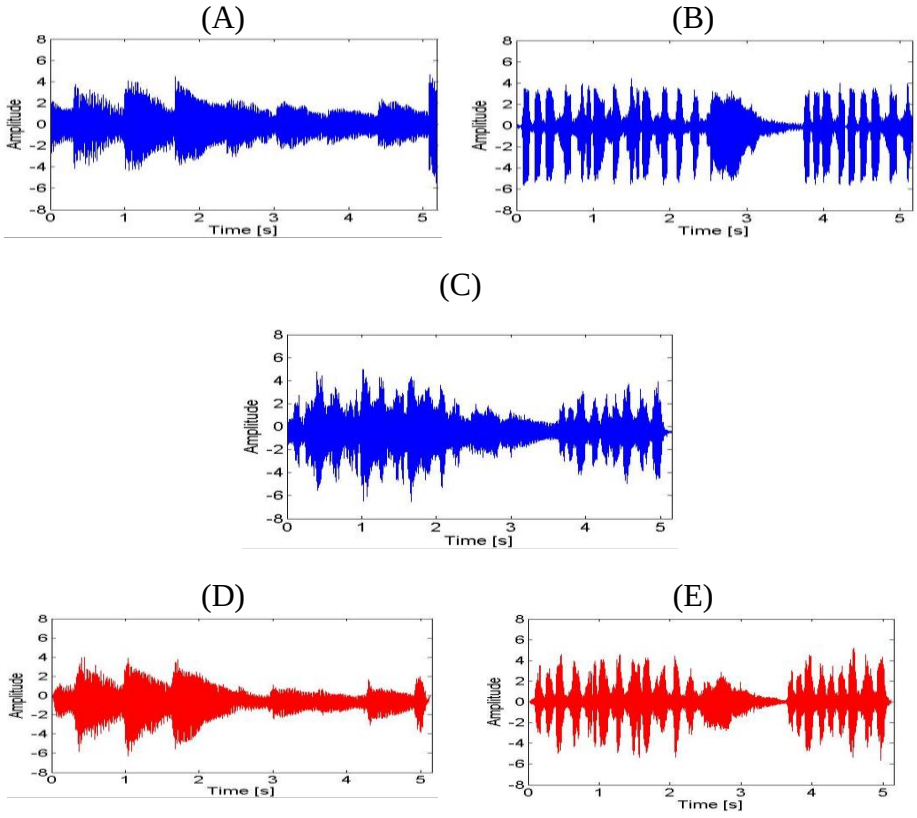


Figure 5.9: Separation result in time domain. (A)-(B): Original piano and trumpet. (C): Live recorded mixture of piano and trumpet sound. (D)-(E): Separated piano and trumpet.

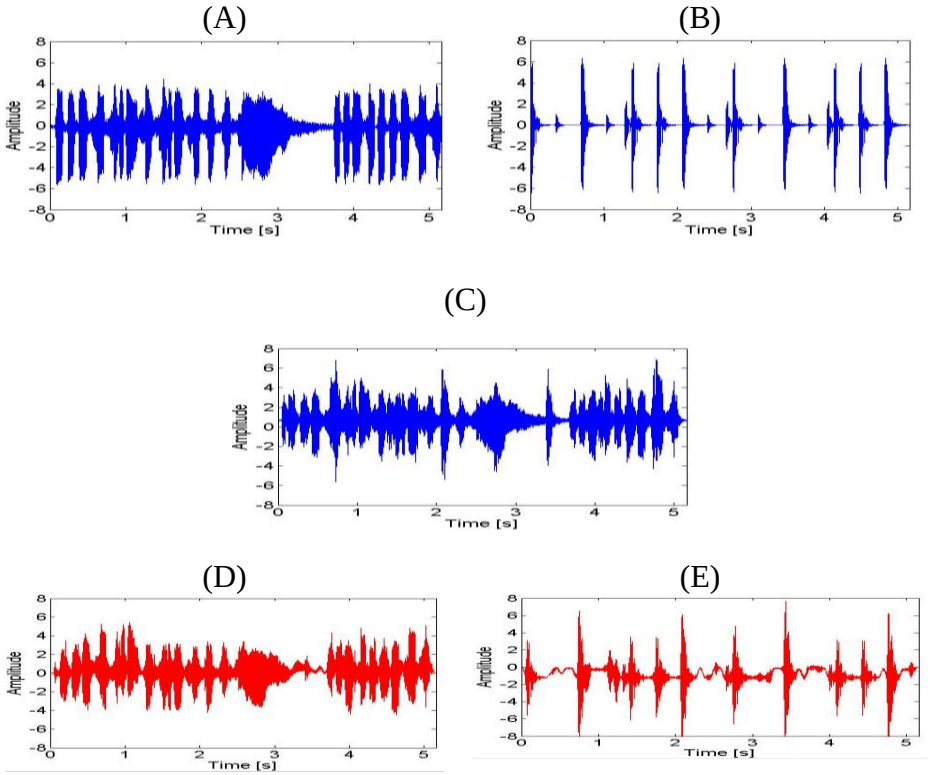


Figure 5.10: Separation result in time domain. (A)-(B): Original trumpet and drum. (C): Live recorded mixture of trumpet and drum sound. (D)-(E): Separated trumpet and drum.

Table 5.6 gives the SDR results for the live-recorded audio mixture between piano, trumpet, drum and flute. Our proposed method yields a very good separation performance especially for piano-trumpet mixture with an average SDR of 6.2 dB for both sources. On the other hand, for piano-drum mixture the achieved average SDR is 3 dB for both sources. The reason is because both piano and drum occupy the lower part of the log-frequency spectrogram and the possibility of signal overlapping each other is very high which has affected the separation of the sources. This pattern also happens for trumpet-flute mixture where both trumpet and flute occupy the upper part of log-frequency spectrogram. An average SDR of 4.6 dB has been achieved. On the overall, the obtained results is be considered good as external factors such as noise

from the environment e.g. background noise and street noise and possibly also the nonlinearity of the microphone were present during the live recording of the audio mixtures. It should be noted the SDR results are lower than the synthetic convolutive mixture using Roomsim because we are comparing with the source signals instead of source images.

Table 5.6: Source separation performances for various types of live-recorded audio mixture in terms of SDR (dB)

Mixture	**Separated source 1**	**Separated source 2**
Piano and trumpet	Piano	Trumpet
	5.0	7.4
Piano and flute	Piano	Flute
	4.7	6.3
Piano and drum	Piano	Drum
	3.2	2.9
Trumpet and flute	Trumpet	Flute
	4.6	4.5
Trumpet and drum	Trumpet	Drum
	5.5	4.9
Flute and drum	Flute	Drum
	6.0	4.1

5.3.6 Experiment on professionally produced music recordings

In this experiment, the proposed method is tested on two professionally produced music recordings of well-known songs namely “Make you feel my love” by Adele, and “You raise me up” by Kenny G. The music consists of two excerpts of length

approximately 20s on mono channel and resampled to 16 kHz. The "Make you feel my love" song consist of female vocal and piano sound while "You raise me up" is an instrumental music consist of saxophone and piano sound. The factors of τ and ϕ shifts are set to have $\tau_{\max} = 8$ and $\phi_{\max} = 32$. Since the original source spatial images are not available for this experiment, the separation performance is assessed perceptually and informally by analysing the log-frequency spectrogram of the estimated source images and listening to the separated sound. This task was a tough task since the instruments play many different notes in the recording. In addition, the blind separation of the vocal is considered very challenging since it usually consists of both low and high frequency which will easily overlap with the instrument sound in the mixture. Figure 5.11 shows the separation results in term of log-frequency spectrogram for song "Make you feel my love". It can be clearly seen that the female vocal and the piano sound have been well separated. This is evidenced from Figure 5.11(B) where it shows the three sentences of lyrics of the female vocal singing with vibration and at high pitch occupying the upper part of the log-frequency spectrogram while in Figure 5.11(C), it clearly shows the sequence of piano notes which is characterised by the discrete nature of sound. As for Figure 5.12, it can be clearly visible that the saxophone and piano sound has been well separated. The high pitch of continuous saxophone sound is shown in the Figure 5.12(B) while the notes of the piano are evidently present in Figure 5.12(C). In the overall, our proposed method has successfully separated the professionally produced music recordings and gives a perceptually pleasant listening experience.

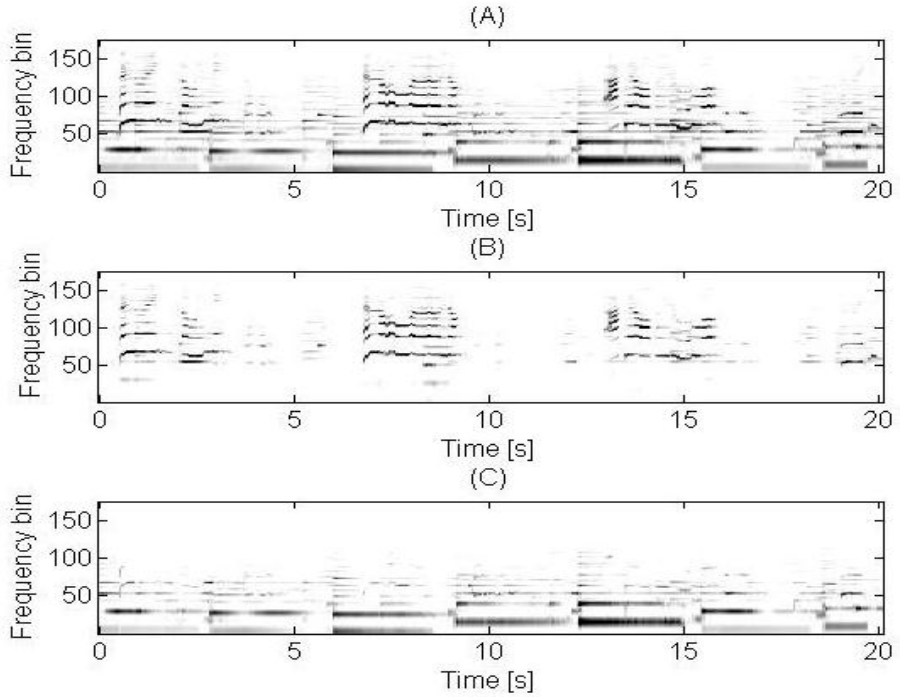


Figure 5.11: Separation result in spectrogram for song "Make you feel my love" by Adele. (A) music recording (B) estimated female vocal (C) estimated piano sound.

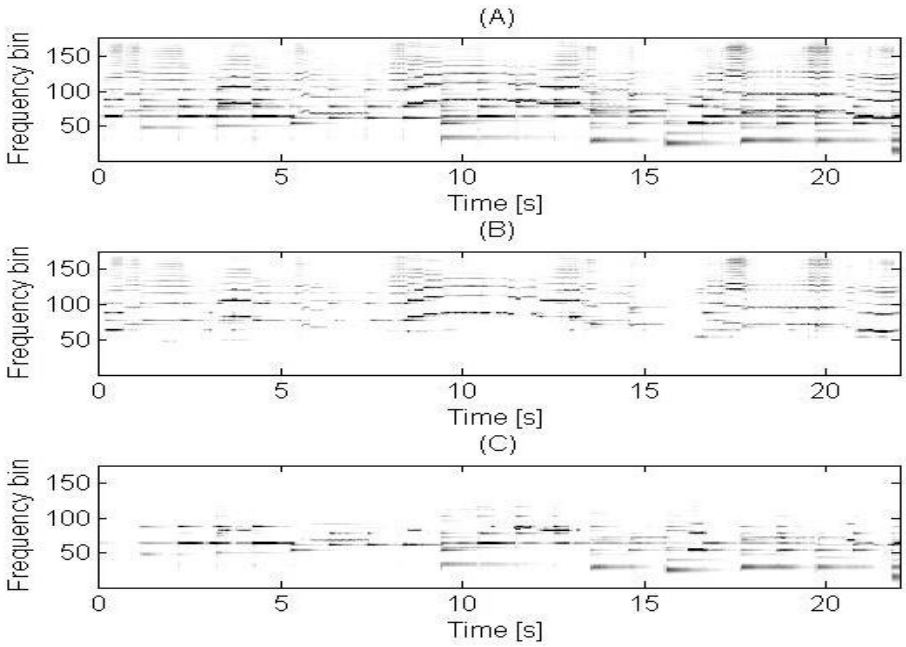


Figure 5.12: Separation result in spectrogram for song "You raised me up" by Kenny G. (A) music recording (B) estimated saxophone sound (C) estimated piano sound.

5.4 Summary

A novel solution to separate convolutively mixed sources in a single channel recording has been presented. The proposed FC-SNMF2D is developed under the probabilistic framework which enables adaptive sparseness to be incorporated in the solution. The adaptive sparseness has resulted in the desired degree of sparsity for the decomposition. It has also been shown that significant performance improvement has been achieved compared with the conventional methods of fixed sparsity. Experiment using live-recorded mixtures containing music sounds in real environment has been conducted to further substantiate the separation capability of the proposed method. In addition, our method has proven to be exceptional in blind separation of professionally produced music recording. There are at least three significant advantages of proposed method: Firstly, the proposed method considers the convolutive mixing model which signifies more accurate representation of the real environment. Secondly, the sparsity term is adaptively regulated to find the anticipated decomposition. Finally, the method is computationally efficient.

CHAPTER 6

CONCLUSION AND FUTURE WORKS

This chapter will summarise the single channel source separation (SCSS) methods and lists the contributions of this thesis. Nevertheless, there still remain open questions to be addressed in the future. Thus, we raise possible research directions towards achieving more efficient SCSS methods. The work in this thesis has fulfilled.

6.1 Summary and Contributions

In this thesis, the aims and objectives of the research work set out in Chapter 1 have been fulfilled. In Chapter 2, a literature review of SCSS methods in linear instantaneous mixture was presented. A SCSS general framework and task is explained and SCSS methods that aim to estimate the original sources accurately through various approaches were organised and summarised into unifying context. However, the practicality of current linear methods is undermined by the linear assumption adopted at the theoretical level which therefore limits the applications in reality. Hence, this requires the development of reliable solutions for SCSS that considered the problem such as nonlinearity and signal reverberation of the observed signals. This thesis also has summarised other unresolved challenges and problems in

current SCSS method in Chapter 2 which motivate this thesis to develop new strategies in providing effective and accurate solutions of SCSS problems.

In Chapter 3, a new two stage approach has been presented for solving SCSS problem in post-nonlinear instantaneous mixture. The post-nonlinear mixture model is popular not only due to its simplicity in analysis, but also widely applicable. The proposed technique combines the Gaussianization transform and the time-domain maximum likelihood separation algorithm. Using the Gaussianization transform, the nonlinearly distorted observed mixture was inverted so that the mixture can be efficiently be separated by the linear separation algorithm. From the carried out experiments, Gaussianization transform performed very well in recovering the loss of signal information due to the nonlinearity. In addition, the proposed method yields significant performance in post-nonlinear mixture compared with the linear algorithm.

In Chapter 4, a new framework of FCNMF2D has been presented for solving SCSS problem in convolutive mixture. The proposed model contemplates the convolutive mixing model which implies more accurate representation of the actual environment. Two inference techniques have been proposed: variant of EM algorithm which maximises the joint log-likelihood called Quasi-EM FCNMF2D and MU rules for the maximisation of individual log-likelihood called MU FCNMF2D. These proposed methods are unsupervised which required no training data. In addition, IS divergence used as a cost function holds the desirable property of scale invariant that enables low energy components in the log spectrogram bear the same relative importance as the high energy ones. Experimental results show that significant performance improvement has been achieved by updating the frequency mixing parameter for convolutive mixture. In addition, separation performance on feature

extraction and blind audio source separation has proven to be exceptional especially for Quasi-EM FCNMF2D algorithm.

In Chapter 5, a novel framework of FC-SNMF2D has been presented to separate convolutively mixed sources in a single channel recording. The impetus behind this is that there are still ambiguities between the factors in the FCNMF2D solution. Hence, it is necessary to impose sparseness to give unique representations which will improve the separation performance. The proposed FC-SNMF2D is developed under the probabilistic framework which enables adaptive sparseness to be incorporated in the solution. The regularization term is adaptively tuned to yield desired degree of sparsity thus enabling the spectral basis and temporal codes of non-stationary audio signals to be separated more efficiently. From the experiments, it has been shown that significant performance improvement has been achieved compared with the conventional methods of fixed sparsity. In addition, simulations of live-recorded mixtures and professionally produced music recording have been carried out to verify the effectiveness of the proposed algorithm and result shows an exceptional separation performance has been obtained. Table 6.1 summarise the proposed method in this thesis.

Table 6.1: Summary of proposed SCSS methods

Case	Method	Category	Signal representation	Update method
Post-nonlinear instantaneous	Two stage approach of Gausianization and ML SCSS approach.	Supervised SCSS	Time domain	ML and MAP
Linear convolutive	Quasi-EM FCNMF2D	Unsupervised SCSS	TF domain (Log-frequency spectrogram)	Quasi-EM
	MU FCNMF2D			MU
	MU FC-SNMF2D			MU

6.2 Future works

6.2.1 Development of nonlinear SCSS in convolutive mixture

In the future work, a new SCSS solution is needed for the recovery of convolutive mixed and post-nonlinear distorted source to the practical level. So far, there is no method proposed to solve this post-nonlinear convolutive SCSS problem. Taking noise into consideration, the observation of post-nonlinear convolutive mixing model can be expressed as:

$$y(t) = \sum_{j=1}^{J} \sum_{\rho=0}^{L-1} f\left(a_j(\rho) x_j(t-\rho) + e(t)\right) \tag{6.1}$$

where $a_j(\rho)$ is the finite-impulse response (FIR) of some causal filters and $e(t)$ is some additive noise. $f(.)$ is an invertible nonlinear function. Thus, the power TF representation of matrix representation is given by $|\mathbf{Y}|^{.2} = \sum_{j=1}^{J} \mathbf{f}\left(|\mathbf{A}_j|^{.2} |\mathbf{X}_j|^{.2} + |\mathbf{E}|^{.2}\right)$. The aim of the developed SCSS method is to compensate the nonlinear distortion, **f**, convolutive mixing model, $\mathbf{A}_j$ and the sources $|\mathbf{X}_j|^{.2}$.

6.2.2 Development of EM based FC-SNMF2D

In Quasi-EM FCNMF2D, the convergence of to a stationary point of $d_{IS}\left(\mathbf{V}_k^{'} \middle| \sum_{\tau,\phi}\mathbf{u}_k\mathbf{w}_k^{\tau}\mathbf{h}_k^{\phi}\right)$ is granted by property of Quasi-EM. Nevertheless, it can converge only to a point in the interior domain of the parameter space which leads Quasi-EM algorithm prohibits zeros in the factors i.e. $\mathbf{W}^{\tau}$ and $\mathbf{H}^{\phi}$ cannot take entries equal to zero. In particular, in order to minimise the cost function, if either $w_{f,j}^{\tau}$ or $h_{j,n}^{\phi}$ is zero then the resulting cost function becomes infinite. On the contrary, this is not feature shared by MU FCNMF2D algorithm, which does not a priori exclude zeros coefficients in $\mathbf{W}^{\tau}$ and $\mathbf{H}^{\phi}$ (except for $\mathbf{Z}_{f,n}=0$ which would lead to a division by zero). Beacause zero coefficients are invariant under multiplicative updates, if the MU FCNMF2D algorithm attains a fixed point solution with zero entries, then it cannot be determined if the limit point is a stationary point. On the other hand, if the limit point does not take zero entries (i.e. belongs to the interior of the parameter space) then it is a stationary point, which may or may not be a local minimum. Thus, the Quasi-EM FCNMF2D is more reliable compare to MU FCNMF2D. In addition, it is necessary to impose an adaptive sparseness in the matrix factorization to resolve the ambiguities between factors. This has been verified in Chapter 5. Thus, the development of Quasi-EM FC-SNMF2D is essential to increase the accuracy of separation performance. Consider the generative model in (4.30), the EM algorithm works by formulating the conditional expectation of the negative log likelihood of $\mathbf{C}_k$ as:

$$p\left(\mathbf{u}_k, \mathbf{w}_k^{\tau}, \mathbf{h}_k^{\phi} \middle| \mathbf{C}_k, \boldsymbol{\lambda}_k^{\phi}\right) = \frac{p\left(\mathbf{C}_k \middle| \mathbf{u}_k, \mathbf{w}_k^{\tau}, \mathbf{h}_k^{\phi}\right) p\left(\mathbf{u}_k\right) p\left(\mathbf{w}_k^{\tau}\right) p\left(\mathbf{h}_k^{\phi} \middle| \boldsymbol{\lambda}_k^{\phi}\right)}{p\left(\mathbf{C}_k\right)} \tag{6.2}$$

where the denominator is a constant and it is assumed $\mathbf{w}_k^{\tau}$ and $\mathbf{h}_k^{\phi}$ are jointly independent so that EM algorithm can be presented as:

$$\begin{aligned}
&Q_k^{MAP}\left(\boldsymbol{\theta}_k \middle| \boldsymbol{\theta}'\right) \\
&\overset{c}{=} -\int_{\mathbf{C}_k} p\left(\mathbf{C}_k \middle| \mathbf{Y}, \boldsymbol{\theta}'\right) \log p\left(\mathbf{C}_k \middle| \boldsymbol{\theta}_k\right) d\mathbf{C}_k \\
&= -\int_{\mathbf{C}_k} p\left(\mathbf{C}_k \middle| \mathbf{Y}, \boldsymbol{\theta}'\right) \left[\log p\left(\mathbf{C}_k \middle| \boldsymbol{\theta}_k\right) + \log p\left(\mathbf{u}_k\right) + \log p\left(\mathbf{w}_k^{\tau}\right) + \log p\left(\mathbf{h}_k^{\phi} \middle| \boldsymbol{\lambda}_k^{\phi}\right)\right] d\mathbf{C}_k \\
&= -\int_{\mathbf{C}_k} p\left(\mathbf{C}_k \middle| \mathbf{Y}, \boldsymbol{\theta}'\right) \left[\log p\left(\mathbf{C}_k \middle| \boldsymbol{\theta}_k\right)\right] d\mathbf{C}_k + \log p\left(\mathbf{u}_k\right) + \log p\left(\mathbf{w}_k^{\tau}\right) + \log p\left(\mathbf{h}_k^{\phi} \middle| \boldsymbol{\lambda}_k^{\phi}\right)
\end{aligned} \tag{6.3}$$

where "$\overset{c}{=}$" denotes e quality up to a positive scale and constant. The prior distribution over $\mathbf{u}_k$ and $\mathbf{w}_k^{\tau}$ are flat where each column is assumed to be factor-wise normalised to unit length. The prior over $\mathbf{h}_k^{\phi}$ is assumed to be exponentially distributed with decay parameters of $\lambda_{j,n}^{\phi}$ for each element in $\mathbf{h}_k^{\phi}$. With this assumption, $\mathbf{u}_k$, $\mathbf{w}_k^{\tau}$ and $\mathbf{h}_k^{\phi}$ can be optimised by following the approach presented in Chapter 5.

REFERENCES

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

[2] A. Bronkhorst, "The cocktail party phenomenon: A review of research on speech intelligibility in multiple talker condition." *Acoustica*, vol. 86, pp. 117–128, 2000.

[3] A. Cichocki and S.I.Amari, *Adaptive Blind Signal and Image Processing Processing – Learning Algorithm and Applications*, John Wiley and Sons, 2003

[4] S. Amari, A. Hyvarinen, S. Lee, T. W. Lee and S. A. David, " Blind Signal Separation and Independent Component Analysis," *Neurocomputing*, vol. 49, pp. 1-5, 2002

[5] R. Vigario, V. Joursmaki, M. Hamalainen, R. Hari, and E. Oja, "Independent Component Analysis for identification of artifacts in magnetoencephalographic recordings," in *Advances in Neural Information Processing Systems*, vol. 10, pp. 229-235, 1998.

[6] A. Hyvarinen, "Survey on Independent Component Analysis," *Neural Computing Surveys*, vol. 1, pp. 94-128, 1999.

[7] J. F. Cardoso, "Source separation using higher order moments", in *Proceedings ICASSP*, pp. 2109-2112, Glasgow, 1989.

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

[9] E. Oja, J. Karhunen, L. Wang and R. Vigario, “Principal and independent components in neural networks,” in *Proc. VII Italian Workshop on Neural Networks WIRN*, Italy, 1995.

[10] C. Jutten and A. Taleb, “Source separation: from dusk till dawn,” in *Proc. 2nd Int. Workshop on Independent Component Analysis and Blind Source Separation (ICA2000),* Helsinki, Finland, pp. 12-26, 2000.

[11] M. Girolami, *Advances in Independent Component Analysis*, Springer-Verlag, 2000.

[12] S. Roberts and R. Everson, *Independent Component Analysis: Principles and Practice*, Cambridge Univ. Press, 2001.

[13] S. I. Amari and A. Cichocki, "Adaptive blind signal processing - Neural network approaches," *Proceedings of the IEEE,* vol. 86, pp. 2026-2048, 1998.

[14] C. Jutten and J. Karhunen, “Advances in Blind Source Separation (BSS) and Independent Component Analysis (ICA) for nonlinear mixtures”, *International Journal of Neural Systems*, vol. 14, no. 5, pp. 267-292, 2004.

[15] S. Harmeling, A. ziehe, B. Blankertz, and K. R. Muller, “ Nonlinear Blind Source Separation using Kernel Feature Spaces,” in *Proc. of Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001),* San Diego, USA, pp. 102-107, 2001.

[16] T. W. Lee, B. Koehler and R. Orglmeister, “Blind Source Separation of Nonlinear Mixing Models,” *Neural Networks for Signal Processing VII. IEEE Press*, pp. 406-415, 1997.

[17] C. Jutten, M. Babaie-Zadeh, and S. Hosseinin, "Three easy ways for separating nonlinear mixtures," *Signal Processing*, vol. 84 no. 2, pp. 217-229, 2009.

[18] M. Solazzi, R. Parisi and A. Uncini, " Blind source separation in nonlinear mixtures by adaptive spline neural networks," in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, USA, pp. 254-259, 2001.

[19] A. Hyvarinen, and P. Pajunen, "Nonlinear independent component analysis: Existence and uniqueness results," *IEEE Trans. on Neural networks*, vol. 12 no. 3, pp. 429-439, 1999.

[20] A. Taleb and C. Jutten, "Source separation in post-nonlinear mixtures," *IEEE Trans. on Signal Processing*, vol. 47, no. 10, pp. 2807-2820, 1999.

[21] A. Taleb and C. Jutten, "Batch algorithm for source separation in post-nonlinear mixtures," in *Proc. of First Int. Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pp.155-160, Aussois, France, 1999.

[22] H. H. Yang, S. I. Amari and A. Cichocki, "Information-theoretic approach to blind separation of sources in nonlionear mixture," *Signal Processing*, vol. 64, no. 3, pp. 291-300, 1998.

[23] A. Blin, S. Araki, and S. Makino, "Underdetermined blind separation of convolutive mixtures of speech using time-frequency mask and mixing matrix estimation," *IEICE Trans. Fundamentals*, vol. E88-A, no. 7, pp. 1693–1700, 2005.

[24] C. Fevotte and S. J. Godsill, "*A Bayesian approach for blind separation of sparse sources*," Technical Report, Cambridge University, Engineering Dept., January 2005.

[25] S. Winter, H. Sawada, and S. Makino, "On real and complex valued L1-norm minimization for overcomplete blind source separation," in *2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, 2005, pp. 86–89.

[26] V. D. Calhoun, T. Adali, L. K. Hansen, J. Larsen and J. J. pekar, "ICA of functional MRI data: An overview," in *4th Int. Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, April 2003.

[27] F. Acernese, A. Ciaramella, S. De Martino, R. De Rosa, M. Falanga and R. Tagliaferri, "Neural networks for blind source separation of Stromboli explosion quakes", *IEEE Transactions on Neural Networks*, vol. 14, pp. 167-175, 2003.

[28] M. Burghoff and P. Van Leeuwen, "Separation of fetal and maternal magnetocardiographic signals in twin pregnancy using independent component analysis (ICA)", in *Biomag 2004*, pp. 311-312, Boston, USA, Aug. 2004.

[29] N. Correa, T. Adali and V. D. Calhoun, "Performance of blind source separation algorithms for fMRI analysis using a group ICA method", *Magnetic Resonance Imaging*, vol. 25, no. 5, pp. 684-694, June 2007.

[30] J. V. Stone, J. Porrill, N. R. Porter and I. D. Wilkinson, "Spatio-temporal independent component analysis of event-related fMRI fata using skewed

probability density functions", *Neuroimage*, vol. 15, no. 2, pp. 407-421, Feb. 2002.

[31] J. Koikkalainen and J. Lotjonen, " Image segmentation with the combination of the PCA-and ICA-based modes of shape variation", in *IEEE International Symposium on Biomedical Imaging: Nano to Macro*, vol. 1, pp. 149-152, Apr. 2004.

[32] C. Beckmann and S. Smith, " Probability independent component analysis for functional magnetic resonance imaging", *IEEE Transactions on Medical Imaging*, vol. 23, pp. 137-152, 2004.

[33] A. D. Back and A. s. Weigend, "A first application of independent component analysis to extracting structure from stock returns*", International Journal of Neural Systems*, vol. 8, no. 4, pp. 474-484, 1997.

[34] A. Hyvarinen, P. O. Hoyer and M. Inki, "Topographic independent component analysis", *Neural Computation*, vol. 13, pp. 1527-1558, 2001.

[35] C. Liu and H. Wechsler, "Independent component analysis of gabor features for face recognition", *IEEE Trans. on Neural Networks*, vol. 14, pp. 919-928, 2003.

[36] U. Madhow, "Blind adaptive interference suppression for direct-sequence CDMA", *Proceedings of the IEEE*, vol. 86, no. 10, pp. 2049-2069, 1998.

[37] R. Cristescu, T. Ristaniemi, J. Joutsensalo and J. Karhunen, "Delay estimation in CDMA communications using a Fast ICA algorithm", In *Proc. Int. Workshop on Independent Component Analysis and Blind Signal Separation (ICA2000)*, pp. 105-110, Helsinki, Finland, 2000.

[38] C. L. Isbell and P. Viola, "Restructuring sparse high-dimensional data for effective retrieval," *Advances in Neural Information processing Systems,* vol. 11, The MIT Press, 1999.

[39] W. L. Woo and S. S. Dlay, "Neural network approach to blind separation mono-nonlinearly mixed sources," *IEEE Trans. On Circuits and System-1,* vol. 52, no. 6, pp. 1236-1247, 2005.

[40] W. L. Woo and L. C. Khor, "Blind restoration of nonlinearly mixed signals using multilayer polynomial neural network", *IEE Proc. on Vision, Image and Signal Processing*, vol. 151, no. 1, pp. 51-61, 2004.

[41] N. Mitianoudis and M.E. Davies, "Audio source separation of convolutive mixtures", *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 5, pp. 489-497, 2003.

[42] J.-F. Cardoso, J. Delabrouille, and G. Patanchon, "Independent component analysis of the cosmic microwave background," in *Fourth International Symposium on Independent Component Analysis and Blind Signal Separation*, pp. 1111–1116, Nara, Japan, Apr. 2003.

[43] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans on. Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066-1074, 2007.

[44] R. Quain Quiroga, L. Reddy, G. Kreiman, C. Koch and I. Fried, "Invariant visual representation by single-neurons in the human brain", *Nature*, vol. 435, pp.1102-1107, 2005.

[45] R. Quain Quiroga, Z. Nadasdy and Y. Ben-Shaul, "Unsupervised spike sorting with wavelets and superparamagnetic clustering", *Neural Computation*, vol. 16, pp. 1661-1687, 2004.

[46] Y. Ephraim, "Statistical model based speech enhancement systems," *IEEE Proc.*, vol. 80, no. 10, pp. 1526–1555, Oct. 1992.

[47] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. on Speech and Audio Processing*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[48] H. Sameti, H. Sheikzadeh, D. Li, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 5, pp. 445–455, Sep. 1998.

[49] D. Burshtein and S. Gannot, "Speech enhancement using a mixture maximum model," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 6, pp. 341–351, 2002.

[50] R. Martin, "Speech enhancement based on minimum square error estimation and supergaussian priors," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, 2005.

[51] A. M. Reddy and B. Raj, "A minimum mean squared error estimator for single channel speaker separation," in *Interspeech'04*, Oct. 2004, pp. 2445–2448.

[52] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proc. ICASSP'04*, May 2004, pp. 817–820.

[53] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "A joint probabilistic-deterministic approach using source-filter modelling of speech signals for

single channel speech separation," in *Proc. IEEE MSLP'06*, Maynooth, U.K., Sep. 2006, pp. 47–52.

[54] S. Roweis, "One microphone source separation," in *Proc. Neural Inf. Process. Syst.*, 2000, pp. 793–799.

[55] L. Benaroya and F. Bimbot, "Wiener based source separation with HMM/GMM using a single sensor," in *International Conference on Independent Component Analysis and Blind Signal Separation*, Apr 2003.

[56] M. J. Reyes-Gomez, D. Ellis, and N. Jojic, "Multiband audio modelling for single channel acoustic source separation," in *Proc. ICASSP'04*, May 2004, vol. 5, pp. 641–644.

[57] M. H. Radfa and R. M. Dansereau, "Single-channel speech separation using soft mask filtering", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no.6, pp. 2299-2310, 2007.

[58] A. Ozerov, P. Philippe, F. Bimbot and R. Gribonval, "Adaptation of Bayesian models for single channel source separation and its application to voice/music separation in popular songs", *IEEE Trans. on Audio, Speech and Language Processing, special issue on Blind Signal Processing for Speech and Audio Applications,*, vol. 15, no. 5, pp. 1564-1578, July 2007.

[59] A. Hyvarinen and P. Hoyer, "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces", *Neural Computation*, vol. 12, no. 7, pp. 1705-1720, 2000.

[60] E. Vincent and X. Rodet, "Music transcription with ISA and HMM", in *Proc. of the 5th International Symposium on Independent Component Analysis and Blind Signal Separation,* Granada, Spain, 2004.

[61] M. A. Casey, "Separation of mixed audio sources by independent subspace analysis," Merl - A Mitsubishi Electric Research Laboratory, Massachusetts, USA, Tech. Rep. TR-2001-31, Sep. 2001.

[62] Md. K. I. Molla and K. Hirose, "Single-Mixture Audio Source separation by subspace decomposition of Hilbert spectrum", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 893-900, March 2003.

[63] P. Li, Y. Guan, B. Xu, and W. Liu, "Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2014-2023, Nov. 2006.

[64] G. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation", *IEEE. Trans. on Neural Networks*, vol. 15, no. 5, pp. 1135-1150, Sept. 2004.

[65] M.S. Pedersen, D.L. Wang, J. Larsen and U. Kjems, "Two-microphone separation of speech mixtures", *IEEE. Trans. on Neural Networks*, vol. 19, no. 3, pp. 475-492, 2008.

[66] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking", *IEEE. Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830-1847, Jul. 2004.

[67] R. Weiss and D. Ellis, "Monaural speech separation using source-adapted models", in *Proc. IEEE Workshop on Apps. of Sig. Processing to Acous. and Audio WASPAA-07*, pp. 114-117, Mohonk NY, October 2007.

[68] S.H. Srinivasan and M. S. Kankanhalli, "Harmonicity and dynamics based audio separation", in *Proc. IEEE International Conference on Acoustics,*

Speech, and Signal Processing (ICASSP), vol. 5, pp. 640-643, Hong Kong, China, 2003.

[69] Li. Y, J. Woodruff and D.L. Wang, "Monaural musical sound separation based on pitch and common amplitude modulation", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, pp. 1361-1371, 2009.

[70] M. Cooke and D. P. W. Ellis, "The auditory organization of speech and other sources in listeners and computational models", *Speech Communication*, vol. 35(3-4), pp. 141 – 177, 2001.

[71] F. R. Bach and M. I. Jordan, "Blind one-microphone speech separation: A spectral learning approach", *Advances in Neural Information Processing Systems*, 2004.

[72] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Computation*, vol. 17, no. 9, pp. 1875–1902, 2005

[73] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values", *Environmetrics*, vol. 5, no. 2, pp. 111-126, 1994.

[74] N. Berin, R. Badeau and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 5480-5493, 2010.

[75] E. Vincent, N. Bertin and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18, no. 3, pp. 528-537, 2010.

[76] P. Smaragdis and J.C.Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2003, pp. 177–180.

[77] Y.C. Cho and S. Choi, "Nonnegative features of spectro-temporal sounds for classification", *Pattern Recognition Letters*, vol. 26, pp. 1327-1336, 2005.

[78] M.D. Plumbley, "Algorithms for non-negative independent component analysis", *IEEE Trans. on Neural Networks*, vol. 14, no. 3, pp. 534-543, May 2003.

[79] R. Zdunek and A. Cichocki, "Nonnegative matrix factorization with constrained second-order optimization", *Signal Processing*, vol. 87, no. 8, pp. 1904-1916, Aug. 2007.

[80] P. Sajda, S. Du, T. Brown, R. Stoyanova, D. Shungu, X. Mao and L. Parra, "Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain", *IEEE Trans. on Medical Imaging*, vol. 23, no. 12, pp. 1453-1465, 2004.

[81] Y. Li, A. Ngom, "A new Kernel non-negative matrix factorization and its application in microarray data analysis," in *Proc. of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB),* pp. 371-378, 9-12 May 2012.

[82] H. Kim and H. P. Denning, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, May 2007.

[83] S. Xie, Z. Yang, and Y. Fu, "Nonnegative matrix factorization applied to nonlinear speech and image cryptosystems," *IEEE Trans. on Circuits and Systems I*, vol. 55, no. 8, pp. 2356-2367, Sep 2008.

[84] D. Lee and H. Seung, "Learning the parts of objects by nonnegative matrix factorisation," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[85] D. Donoho and V. Stodden, "When does non-negative matrix factorisation give a correct decomposition into parts?" in *Proc of NIPS*, 2003, pp. 1141–1148

[86] P. Smaragdis, " Discovering auditory objects through non-negativity constraints." in *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, Jeju, Korea, 2004.

[87] C. Fevotte, N. Bertin and J.L. Durrie, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793-830, Mar. 2009.

[88] R. Zdunek, and A. Cichocki, "Nonnegative matrix factorization with constrained second-order optimization" *Signal Processing*, vol. 87, no. 8, pp. 1904-1916, August 2007.

[89] I. Biciu, N. Nikolaidis, and I. Pitas, "Nonnegative matrix factorization in polynomial feature space", *IEEE Trans. on Neural Network*, vol. 19, pp. 1090-1100, 2007.

[90] R. Kompass, "A generalized divergence measure for nonnegative matrix factorization", *Neural Computation*, vol. 19, no. 3, pp. 780-791, March 2007.

[91] A. Cichocki, R. Zdunek, and S.I. Amari, "Csiszar's divergences for non-negative matrix factorization: family of new algorithms," in *Proc. Intl. Conf. on*

Independent Component Analysis and Blind Signal Separation (ICABSS'06), Charleston, USA, March 2006, vol. 3889, pp. 32–39.

[92] D. FitzGerald, "Automatic drum transcription and source separation," Ph.D. thesis, Dublin Institute of Technology, Dublin, Ireland, 2004.

[93] P. Smaragdis, " Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs", in *Fifth International Conference on Independent Component Analysis, LNCS* 3195, pages 494–499, Granada, Spain, Sept. 22–24 2004. Springer-Verlag.

[94] M. N. Schmidt and M. Morup, "Nonnegative matrix factor 2-D deconvolution for blind single channel source separation", in *Proc. 6th International Conf. on Independent Component Analysis and Signal Separation (ICA '06)*, Charleston, USA, March 2006, pp. 700-707.

[95] T.F. Quatieri, D.A. Reynolds and G.C. O'Leary, "Estimation of handset nonlinearity with application to speaker recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 5, pp. 567-584, Sept. 2000.

[96] D.A. Reynolds, M.A. Zissman, T.F. Quatieri, G.C. O'Leary and B.A. Carlson, "The effects of telephone transmission degradation speaker recognition performance," in *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, Munich, Germany, Apr. 1997.

[97] T.F. Quatieri, D.A. Reynolds and G.C. O'Leary, "Estimation of handset nonlinearity with application to speaker recognition," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, vol. 2, pp. 745-748, Seattle, USA, May 1998.

[98] B.K. Meadows, T.H. Heath, J.D. Neff, E.A. Brown, D.W. Fogliatti, M. Gabbay, V. In, P. Hasler, S.P. Deweerth and W.L. Ditto, "Nonlinear antenna technology," *Proc. of IEEE*, vol. 90, no. 5, pp. 882-897, May 2002.

[99] R.J. Ram and R. Sporer et al., "Chaos in microwave antenna arrays," in *1996 IEEE MTT-S Int. Microwave Symp. Dig.*, San Francisco, CA: IEEE, 1996.

[100] J.J. Lynch and R.A. York, "A mode locked array of coupled phase locked loops," *IEEE Microwave Guided Wave Letter*, vol. 5, pp. 213-215, July 1995.

[101] G.J. Jang and T.W. Lee, "A maximum likelihood approach to single channel source separation," *Journal of Machine Learning Research*, vol. 4, pp. 1365-1392, 2003.

[102] B. Gao, W.L. Woo, and S.S. Dlay, "Single Channel Blind Source Separation using best characteristic basis," in *3rd International Conference of ICTTA*, 2008.

[103] G. Hu and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. on Neural Network*, vol. 15, no. 5, pp. 1135-1150, Sept. 2004.

[104] A. Taleb and C. Jutten, "Source separation in post-nonlinear mixtures," *IEEE Trans. on Signal Processing*, vol. 47, no. 10, pp. 2807-2820, 1999.

[105] A. Ziehe, M. Kawanabe, S. Harmeling and R. M. Kalus, "Blind separation of post-nonlinear mixtures using linearizing transformations and temporal decorrelation." *Journal of Machine Learning Research*, no. 4, pp. 1319-1338, 2003.

[106] S.Chen, and R.A.Gopinath, "Gaussianization", in *Proc. of NIPS*, Denver, USA, 2000.

[107] J. Sole-Casals, C. Jutten and D.T. Pham, "Fast approximation of nonlinearities for improving inversion algorithms of PNL mixtures and Wiener systems", *IEEE Trans. on Signal processing*, no. 85, pp. 1780-1786, 2005.

[108] A. Hyvärinen."Fast and Robust Fixed-Point Algorithms for Independent Component Analysis", *IEEE Transactions on Neural Networks*, no.10 vol.3, pp. 626-634, 1999

[109] C. Fevotte, R. Gribonval and E. Vincent, "*BSS EVAL Toolbox User Guide*", IRISA Technical Report 1706, Rennes, France, April 2005. http://www.irisa.fr/metiss/bsseval.

[110] "Signal Separation Evaluation Campaign (SiSEC 2008)," 2008. [Online]. Available: http://sisec.wiki.irisa.fr

[111] P. Sajda, S. Du, T. Brown, R. Stoyanova, D. Shungu, X. Mao, and L. Parra, "Non-negative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain," *IEEE Trans. on Medical Imaging*, vol. 23, no. 12, pp. 1453–1465, 2004.

[112] S. Xie, Z. Yang, and Y. Fu, "Nonnegative matrix factorization applied to nonlinear speech and image cryptosystems," *IEEE Trans. on Circuits and Systems I*, vol. 55, no. 8, pp. 2356-2367, Sep 2008.

[113] R. Zdunek, and A. Cichocki, "Nonnegative matrix factorization with constrained second-order optimization" *Signal Processing*, vol. 87, no. 8, pp. 1904-1916, August 2007.

[114] R. Schachtner, G. Poeppel, and E. W. Lang, "A Nonnegative Blind Source Separation Model for Binary Test Data", *IEEE Trans. on Circuits and Systems I*, vol. 57, no. 7, pp. 1439-1448, Jul. 2010.

[115] J. Taghia and J. Taghia, "One-channel audio source separation of convolutive mixture," *Advances in Computer and Information Sciences and Engineering*, pp. 202-206, 2008.

[116] L. Mark, D. Barry, D. Dorran and E. Coyle, "Single Channel Sound Source Separation combining Delay Estimation and the ADRess algorithm," in *IET Proc. of Signal and Systems Conference*, Ireland, pp. 288-292, 2008.

[117] A. Ozerov and C. Févotte," Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550-563, March 2010.

[118] C. Kyriakakis, "Fundamental and technological limitations of immersive audio systems," *Proceedings of the IEEE* , vol.86, no.5, pp.941-951, 1998.

[119] A.Mouchtaris, P. Reveliotis, C. Kyriakakis, "Inverse filter design for immersive audio rendering over loudspeakers," *IEEE Transactions on Multimedia* , vol.2, no.2, pp.77-87, Jun 2000.

[120] Y. Huang, J. Chen and J. Benesty , "Immersive audio schemes," *IEEE Signal Processing Magazine,* , vol.28, no.1, pp.20-32, Jan. 2011

[121] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method", in *Proc. 6th Int. Congress on Acoustics*, Tokyo, Japan, Aug. 1968.

[122] Judith C. Brown, "Calculation of a constant Q spectral transform", *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425-434, 1991.

[123] C. J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Transactions on Neural Networks* , vol.18, no.6, pp.1589-1596, Nov. 2007

[124] D. Campbell, Roomsim Toolbox. http://media.paisley.ac.uk/~campbell

[125] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech Separation by Humans and Machines*, pp.181-197, 2005.

[126] M. Morup and M. N. Schmidt, "Sparse nonnegative matrix factor 2-D deconvolution," Techical Report, Technical University of Denmark, Copenhagen, Denmark, 2006.

[127] M. N. Schmidt and M. Morup, "Sparse non-negative matrix factor 2-d deconvolution for automatic transcription of polyphonic music," Technical Report, Technical University of Denmark, 2006.

[128] J. Eggert and E. Korner," Sparse coding and NMF," in *IEEE Proc. of International Joint Conf. on Neural Networks*, vol. 4, pp. 2529-2533, July 2004.

[129] P.O. Hoyer," Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research,* vol. 5, pp. 1457-1469, 2004.