Underdetermined Convolutive Source Separation Using Two Dimensional Non-Negative Factorization Techniques

Ahmed Sattar Hadi Al Tmeme

BSc

MSc

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

March 2017

ABSTRACT

In this thesis the underdetermined audio source separation has been considered, that is, estimating the original audio sources from the observed mixture when the number of audio sources is greater than the number of channels. The separation has been carried out using two approaches; the blind audio source separation and the informed audio source separation. The blind audio source separation approach depends on the mixture signal only and it assumes that the separation has been accomplished without any prior information (or as little as possible) about the sources. The informed audio source separation uses the exemplar in addition to the mixture signal to emulate the targeted speech signal to be separated. Both approaches are based on the two dimensional factorization techniques that decompose the signal into two tensors that are convolved in both the temporal and spectral directions. Both approaches are applied on the convolutive mixture and the high-reverberant convolutive mixture which are more realistic than the instantaneous mixture.

In this work a novel algorithm based on the nonnegative matrix factor two dimensional deconvolution (NMF2D) with adaptive sparsity has been proposed to separate the audio sources that have been mixed in an underdetermined convolutive mixture. Additionally, a novel Gamma Exponential Process has been proposed for estimating the convolutive parameters and number of components of the NMF2D/ NTF2D, and to initialize the NMF2D parameters. In addition, the effects of different window length have been investigated to determine the best fit model that suit the characteristics of the audio signal. Furthermore, a novel algorithm, namely the fusion K models of full-rank weighted nonnegative tensor factor two dimensional deconvolution *(K-wNTF2D)* has been proposed. The *K*-wNTF2D is developed for its ability in modelling both the spectral and temporal changes, and the spatial covariance matrix that addresses the high reverberation problem. Variable sparsity that derived from the Gibbs distribution is optimized under the Itakura-Saito divergence and adapted into the *K*-wNTF2D model. The tensors of this algorithm have been initialized by a novel initialization method, namely the SVD two-dimensional deconvolution (SVD2D). Finally, two novel informed source separation algorithms, namely, the semi-exemplar based algorithm and the exemplar-based algorithm, have been proposed. These algorithms based on the NMF2D model and the proposed two dimensional nonnegative matrix partial co-factorization (2DNMPCF) model. The idea of incorporating the exemplar is to inform the proposed separation algorithms about the targeted signal to be separated by initializing its parameters and guide the proposed separation algorithms. The adaptive sparsity is derived for both

of the proposed algorithms. Also, a multistage of the proposed exemplar based algorithm has been proposed in order to further enhance the separation performance.

Results have shown that the proposed separation algorithms are very promising, more flexible, and offer an alternative model to the conventional methods.

ACKNOWLEDGEMENT

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

This thesis is not only the results of the hard work, determination, and continuous efforts, but also the encouragement 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 sincere gratitude to my supervisors Dr. Wai Lok Woo and Professor Satnam Dlay for their guidance, encouragement, and support. I have been inspired by their attention to detail, quest for excellence work, and desire for hard work, I am fortunate to have them as my supervisors. They have not only taught me source separation and factorization techniques but they nurture me in becoming an independent researcher.

Grateful thanks to my parents, brothers, sisters, wife, and my children Mohammed and Sama for their sacrifices to help me and for their endless love. Thank you very much.

Last, but not least, thanks to my home country Iraq, the Iraqi Cultural Attaché in London, Ministry of Higher Education and Scientific Research (MoHESR) in Iraq, University of Baghdad, and Al-Khwarizmi College of Engineering for supporting me during my study abroad.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

I	Number of channels
i	Channel index
J	Number of sources
j	Source index
A^{-1}	Matrix inverse
I_I	Identity matrix of size $I \times I$
A^H	Matrix Hermitian (Matrix conjugate transpose)
$det(A)$	Matrix determinant
$tr(A)$	Matrix trace
t	Time sample index
N	Number of time frames
n	Time frame index
F	Number of frequency bins
f	Frequency bin index
$\boldsymbol{x}(t)$	The multichannel mixture signal
$x_i(t)$	The mixture signal of the i^{th} channel
$x_{i,f,n}$	The STFT of $x_i(t)$
$c_{i,j}(t)$	The spatial image of the source signal for the j^{th} source and i^{th} channel
$\boldsymbol{c}_{j,f,n}$	The STFT of $c_{i,j}(t)$
$b_i(t)$	The additive noise of the i^{th} channel
$b_{i,f}$	The STFT of $b_i(t)$
$a_{i,j}(t)$	The finite-impulse response of some (causal) filter for the j^{th} source and i^{th} channel
$a_{i,j,f}$	The STFT of $a_{i,j}(t)$
$s_j(t)$	The j^{th} source signal
$s_{j,f,n}$	The STFT of $s_j(t)$
$v_{j,f,n} \in \mathbb{R}^+$	The variance of the j^{th} source.
$\boldsymbol{\Sigma}_{f,n}^{(x)} \in \mathbb{C}^{I \times I}$	The mixture covariance matrix.
$\sigma_{\underline{i},f,n}^{(x)} \in \mathbb{C}$	The scalar element of the mixture covariance matrix.

$\boldsymbol{\Sigma}_f^{(\boldsymbol{b})} \in \mathbb{C}^{I\times I}$	The time invariant noise covariance matrix.
$\sigma_{\underline{i},f}^{(b)} \in \mathbb{C}$	The scalar element of the time invariant noise covariance matrix.
$\boldsymbol{\Sigma}_{j,fn}^{(\boldsymbol{c})} \in \mathbb{C}^{I\times I}$	The covariance matrix of the j^{th} source image.
$\underline{\boldsymbol{\Sigma}}_{j,f,n}^{(\boldsymbol{c})} \in \mathbb{C}^{I^2}$	The vectorized covariance matrix of the j^{th} source image.
$\sigma_{\underline{\boldsymbol{i}}\, j,f,n}^{(c)} \in \mathbb{C}$	The scalar element of the covariance matrix of the j^{th} source image and i^{th} channel.
$\boldsymbol{\Sigma}_{j,f}^{(\boldsymbol{a})} \in \mathbb{C}^{I\times I}$	The time-invariant spatial covariance matrix of the j^{th} source
$\underline{\boldsymbol{\sigma}}_{j,f}^{(\boldsymbol{a})} \in \mathbb{C}^{I^2}$	The vectorized time-invariant spatial covariance matrix of the j^{th} source.
$\sigma_{\underline{\boldsymbol{i}}\, j,f}^{(a)} \in \mathbb{C}$	The scalar time-invariant spatial covariance matrix of the j^{th} source and i^{th} channel.
$\hat{p}_{j,fn}$	The source power spectrogram posterior estimates
$\boldsymbol{W} = \{w_{f,k}^{\tau,j}\}$	The frequency basis tensor
$\boldsymbol{H} = \{h_{k,n}^{\phi,j}\}$	The encoding tensor
$\boldsymbol{\Lambda} = \{\lambda_{k,n}^{\phi,j}\}$	The tensor that contains the sparsity terms
Q	The channel gain matrix
$\odot$	The Khatri-Rao product
$\langle A, B\rangle_{\{C\}\{D\}}$	The contract product
$\boldsymbol{\Theta}$	model parameters
$\mathcal{N}_c(\mu, \Sigma)$	The proper complex Gaussian distribution with the mean (μ), and the covariance matrix (Σ).
$\mathrm{Q}(\underline{\boldsymbol{h}})$	The Gibbs distribution

LIST OF ABBREVIATIONS/ACRONYMS

a	Scalar value
a	Vector
A	Matrix
A	Tensor
BSS	Blind Source Separation
ISS	Informed Source Separation
NMF	Nonnegative Matrix Factorization
NMD	Nonnegative Matrix Deconvolution
NMF2D	Nonnegative Matrix Factor Two Dimensional Deconvolution
NTF	Nonnegative Tensor Factorization
NTF2D	Nonnegative Tensor Factor Double Deconvolution
GEM-MU	Generalized Expectation Maximization algorithm with Multiplicative Update
K-wNTF2D	*fusion K models of full-rank weighted Nonnegative Tensor Factor Two Dimensional Deconvolution*
NMPCF	Nonnegative Matrix Partial Co-Factorization
2DNMPCF	Two Dimensional Nonnegative Matrix Partial Co-Factorization
SVD	Singular Value Decomposition
SVD2D	SVD Two-Dimensional Deconvolution
PCA	Principle Component Analysis
ICA	Independent Component Analysis
VQ	Vector Quantization
DOA	Direction Of Arrival
EDU	Euclidian Distance
KL	KullBack-Leibler
IS	Itakura-Saito
PARAFAC	Parallel Factor Analysis
MIDI	Musical Instrument Digital Interface
GIG	Generalized Inverse Gaussian Distribution
SDR	Signal-to-Distortion Ratio
SIR	Signal-to-Interference Ratio

SAR	Signal-to-Artifact Ratio
SiSEC	Signal Separation Evaluation Campaign
GSMM	Gaussian Scaled Mixture Model

LIST OF PUBLICATIONS

- Ahmed Al Tmeme, W.L. Woo, S.S. Dlay, and B. Gao, "Underdetermined reverberant acoustic source separation using weighted full-rank nonnegative tensor models," *The Journal of the Acoustical Society of America*, **138**(6), 3411-3426, (2015).

- A. Al-Tmeme, W.L. Woo, S.S. Dlay, and B. Gao, "Underdetermined Convolutive Source Separation using GEM-MU with Variational Approximated Optimum Model Order NMF2D," *IEEE/ACM Transactions on Audio Speech and Language Processing,* vol. 25, no. 1, pp. 35-49, 2017.

- Ahmed Al Tmeme, W.L. Woo, and S.S. Dlay, "Underdetermined Reverberant Source Separation Using Optimized Complex Sparse Nonnegative Tensor 2D Deconvolution with Gamma Exponential Process," submitted to *IEEE Transactions on Audio Speech and Language Processing* (as under review).

- Ahmed Al Tmeme, W.L. Woo, and S.S. Dlay, "Single Channel Informed Audio Separation using Pseudo-Stereo and Guided Two-Dimensional Nonnegative Matrix Factorization", submitted to *IEEE Transactions on Audio Speech and Language Processing* (as under review).

CHAPTER 1

INTRODUCTION

In this chapter, the motivation behind considering the more realistic cases of the audio source separation will be presented. This included the convolutive blind audio source separation and the high-reverberant convolutive blind audio source separation instead of the instantaneous one. Also, the motivation behind going from blind to informed audio source separation will be presented too. Then, the big picture of audio source separation system that has been achieved in this thesis will be demonstrated in order to give a clear view of work done. After that the objectives and contributions will be drawn. Finally, by the end of this chapter, the outline of the thesis is presented chapter by chapter.

1.1 Motivation

Since more than two decades the researchers working on making the machine to have the same ability of the human to listen and distinguish between different sound sources. Although, the great efforts of the researchers, it is still an open problem, even, it is an ill-posed problem if they tried to solve it without any prior information about the sources. Therefore, to make it soluble many researchers considered that the sources have been mixed instantaneously, i.e., they neglect the reverberation from the surrounded environment which is unrealistic as the sounds reflected from the wall or/and the background noise cannot be avoided. Therefore, in this thesis the convolutive mixture (that considers the refection of the sound) will be considered instead of the instantaneous one. In spite that the convolutive blind source separation (BSS) performs well in low reverberation environment, their performance will drop sharply in high reverberation environment. So, in this thesis we go further by considering the high-reverberant blind audio source separation that simulates the real world environment. Even when the blind audio source separation modelled to consider the high reverberation it did not achieved the required performances that can challenge the human ability in sound sources separation. Therefore, researchers have sought an aid from an external source in addition to the mixture signal, and they opted to go from blind to informed audio source separation in order to achieve higher performance that the BSS cannot reach. Consequently,

the informed source separation will be considered as one of the challenges to be tackle in this thesis.

1.2 Big Picture of Audio Source Separation

In this section the terminology that used in the source separation will be explained and the source separation system will be demonstrated.

The idea behind the audio source separation is to extract the audio sources (such as the music and speech signals) from their mixtures (the observer of the sources where mostly assumed the sources have mixed instantaneously by adding them, or convolutively by considering their reverberation). This separation needs a system that is able to perform many processes; such as estimating the number of sources, estimating the required number of frequency basis and convolutive parameters to be assigned to each source, applying separation algorithms, and reconstructing the sources. Figure 1.1 shows the Big Picture of the proposed audio sources separation system. In which all the sources considered to be mixed convolutively, then the number of frequency basis and convolutive parameters will be estimated by using the proposed Gamma-Exponential Process (see Chapter 3). After that the parameters of the separation algorithms will be initialized by the proposed SVD two-dimensional deconvolution (SVD2D) initialization algorithm (see Chapter 4), and the sparsity (the penalty on the activation matrix that ensures only a few units (out of a large population) will be active at the same time. The sparsity can be added as a constraint to the cost function [1]) will be estimated by the proposed variable sparsity algorithm (see Chapter 4). After estimating the required parameters the separation will be carried out by the proposed convolutive blind source separation algorithm or by the proposed high-reverberant (Full-Rank) blind source separation algorithm (see Chapter 3 and 4, respectively). Finally, the sources will be reconstructed by Wiener filter that works as follows

$$\hat{s}_{j,fn} = \frac{\hat{p}_{j,fn}}{\sum_j \hat{p}_{j,fn}} x_{fn} \tag{1.1}$$

where $\hat{s}_{j,fn}$ is the estimated source, x_{fn} is the mixture, $\hat{p}_{j,fn}$ is the estimated power of the j^{th} source, and j is the source index.

The above scenario gives the general overview of the proposed blind source separation algorithms in the thesis.

For the proposed informed source separation system in addition to the mixture there will be an exemplar that emulates the targeted signal to be separated. The idea of adding the exemplar is to inform the proposed separation algorithms about the targeted signal to be separated by initializing its parameters and guide the proposed separation algorithms (see Chapter 5).

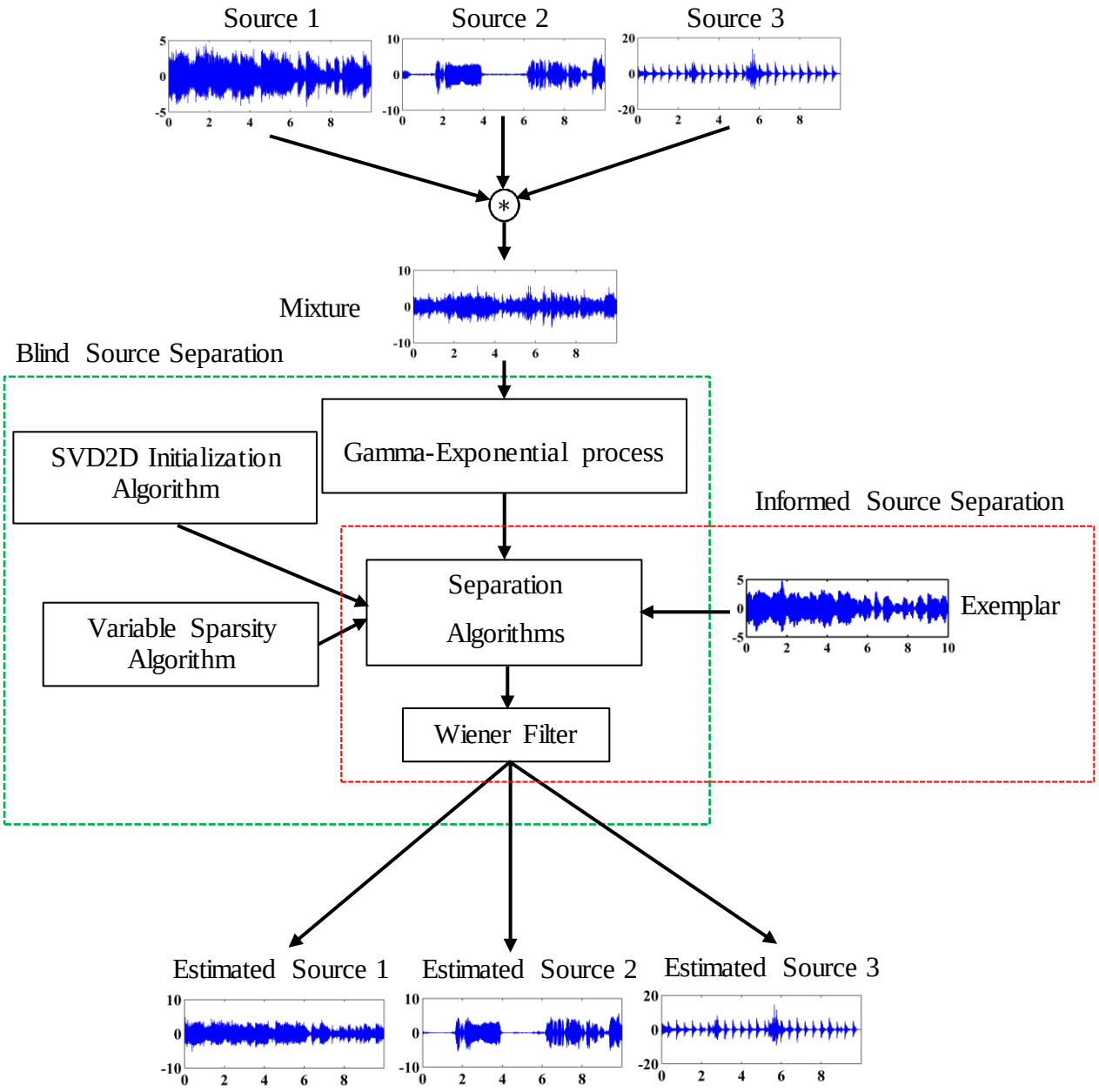


Figure 1.1 Proposed audio separation system

1.3 Aims and Objectives of Thesis

The aims of the thesis are to investigate and develop efficient algorithms for the underdetermined blind and informed audio source separation that mixed in convolutive mixture with low-reverberation environment (convolutive mixture) and high-reverberation environment (high-reverberant (Full-Rank) convolutive mixture). Three novel algorithms have been proposed to tackle these aims.

The objectives of this study are

1. To develop novel algorithms for the underdetermined audio source separation to tackle real world mixing scenario.

2. To exercise control over the parameters which affect the separation performance such as the initialization, number of frequency basis and convolutive parameters, cost functions, and windows length.

3. Develop background theories that further pave the understanding of the audio source separation and develop the mathematical derivations that verified the proposed algorithms.

4. Compare and analysis the performance of the proposed algorithms with the existing algorithms in order to show the efficiency of the proposed algorithms.

1.4 Contributions

The contribution of this thesis for the underdetermined convolutive blind audio source separation, underdetermined high-reverberant blind audio source separation, and underdetermined informed audio source separation can be summarised as follows

1. A novel unsupervised algorithm that based on the nonnegative matrix factor two dimensional deconvolution (NMF2D) with adaptive sparsity is proposed. This algorithm is proposed to blindly separate audio sources which have been mixed in underdetermined convolutive mixture.

2. A novel algorithm, namely *the fusion K models of full-rank weighted nonnegative tensor factor two dimensional deconvolution (K-wNTF2D)* is proposed to blindly separate audio sources which have been mixed in underdetermined high-reverberant mixture.

3. Two novel underdetermined informed audio source separation algorithms, namely, the semi-exemplar based algorithm and the exemplar-based algorithm, are proposed. The semi-exemplar based algorithm and the exemplar-based algorithm are based on the NMF2D model and the proposed two dimensional nonnegative matrix partial co-factorization (2DNMPCF) model, respectively. The proposed 2DNMPCF model factorizes both the mixture and the exemplar at the same time, and it is more powerful than the nonnegative matrix partial co-factorization (NMPCF) model. Also, a pseudo stereo channel is adapted in both algorithms in order to enhance the separation performance. Furthermore, the adaptive sparsity is derived for both of the proposed algorithms in order to adapt each sparse parameter for every temporal code in the 2DNMPCF and NMF2D. Finally, a multistage of the proposed exemplar based algorithm is proposed in order to enhance the separation performance.

4. A novel Gamma Exponential Process is proposed for estimating the convolutive parameters and number of components of the NMF2D, which is an essential step in audio source separation that based on the NMF2D or the nonnegative tensor factor double deconvolution **(**NTF2D) models. Also the proposed algorithm is used to initialize the NMF2D parameters.

5. A novel initialization method, the SVD2D is proposed to initialize the parameters in the NMF2D or the NTF2D. Initialization is the keystone of the audio source separation as a random initialization can lead to converge to local minima or even diverge.

6. A set of variable sparsity parameters derived from Gibbs distribution and optimized under the Itakura-Saito divergence has been encoded into the *K*-wNTF2D model. This optimizes each sub-model in *K*-wNTF2D with the required sparsity in order to model the time-varying variances of the sources in the spectrogram.

7. For faster convergence the proposed algorithms adapted under the hybrid framework that combines the generalized expectation maximization algorithm with the multiplicative update rule (GEM-MU).

8. Finally, the effects of different windows length have been investigated to best fit the model and the characteristics of the audio signal.

1.5 Thesis Outline

The prime focus of this thesis is the unsupervised underdetermined algorithms for audio source separation. Three chapters of this thesis are dedicated for the main contributions of the proposed works, while the first chapter is an introductory to the thesis followed by an overview chapter. Finally, the last chapter draws the conclusions and suggests a future works. A more details of the thesis outlines are given bellow

Chapter 2 provides an overview of the recent audio source separation (blind and informed) algorithms that based on factorization techniques, such as the nonnegative matrix factorization (NMF) and its extension the NMF2D, the nonnegative tensor factorization (NTF), and the NTF2D. Furthermore, it discusses the parameters which affect the separation, such as the cost function, the initialization, window's length, and number of components and convolutive parameters.

In Chapter 3 a novel unsupervised machine learning algorithm based on the NMF2D with adaptive sparsity is proposed. The proposed algorithm adapted under the GEM-MU hybrid framework. This chapter also proposes a method to optimize the number of components and convolutive parameters in the NMF2D by using the Gamma-Exponential process as the observation-latent model. In addition, it is also shown that the proposed Gamma-Exponential process can be used to initialize the NMF2D parameters. Finally, the chapter investigates the issue and advantages of using different window length with different number of convolutive parameters. Simulation results for the synthetic convolutive mixtures and live recordings are carried out in the end of this chapter.

Chapter 4 proposed the *K*-wNTF2D model. The derivation of the algorithm and the development of proposed full-rank *K*-wNTF2D are shown in this chapter. The algorithm also encodes a set of variable sparsity parameters derived from Gibbs distribution into the *K*-wNTF2D model. In addition, a novel initialization method, the SVD2D is proposed to initialize the parameters in the *K*-wNTF2D. Experimental results on the underdetermined reverberant mixing environment have been accomplished at the end of this chapter.

In Chapter 5 two novel algorithms for the underdetermined informed source separation, namely the semi-exemplar based algorithm and the exemplar-based algorithm are proposed. Also the 2DNMPCF model that simultaneously factorizes the mixture and the exemplar is proposed too. The derivation of the adaptive sparsity and the adaptation of the pseudo stereo channel for both of the proposed algorithms are shown in this chapter. Furthermore, a multistage of the proposed exemplar based algorithm is proposed. Finally, comparisons with other algorithms are presented at the end of this chapter.

Finally, Chapter 6 draws the conclusions of this thesis and suggests new avenues for the future work.

CHAPTER 2

OVERVIEW OF AUDIO SOURCE SEPARATION

In this chapter an overview of audio source separation that is based on factorization techniques, such as the NMF, the NMF2D, the NTF, and the NTF2D will be provided. Furthermore, the parameters which effect on the separation, such as the cost function, the initialization, window's length, and number of components and convolutive parameters will be discussed. Finally, the informed source separation will be reviewed.

2.1 What is Audio Source Separation

Source separation (SS) has attracted much research attention in recent years, where great deal of work has been undertaken to solve this problem [2-18]. SS is an acronym referring to estimating the sources from their mixtures, and if there is no information about the sources, then the separation will be achieved blindly, and the technique is called blind source separation (BSS) [19], while if there is additional information about the sources then the technique will called informed source separation (ISS) [20]. Until now SS is an open problem as it does not have the same ability of humans to listen and distinguish between different sources.

Audio source separation can be implemented by using supervised methods or unsupervised methods. The supervised methods have two phases, the training phase and the separation phase. In the training phase the model parameters are trained on the sources (either all of the sources or part of them). Then in the separation phase the separation of the sources will be accomplished by using these trained model parameters. The current trend in supervised methods is the deep neural network (DNN) [21-24] that model the nonlinear relationship between the trained parameters of the targeted speech signal and the mixture signal.

Unsupervised separation methods do not require any training [2, 5-11, 13, 14, 16-18] where the separation of the mixtures accomplished by depending on the mixture signal only. However in this type of source separation prior information is needed in advance before the separation can be carried out, such as the number of sources, and how the sources are mixed.

As the unsupervised source separation will be used in this thesis then more details about it will be given. There are two types of mixtures in source separation; the nonlinear mixture [18] and the linear mixture [2, 5-14, 16, 17]. In the non-linear mixture the mixture signal is constructed from nonlinear combination of the source signals, and it can be expressed as follows

$$x_i(t) = \sum_{j=1}^{J} f_i\left(s_j(t)\right) \tag{2.1}$$

where $x_i(t)$ is the mixture signal, $i = 1, 2, .., I$, I is the total number of channels, $t = 1, 2, \dots, T$, t is the time frame index, $s_j(t)$ is the source signal, $j = 1, 2, \dots, J$, J is the total number of sources, and $f_i(.)$ is the nonlinear mixing process.

In the case of linear mixture, the mixture signal is constructed from the linear combination of source signals. The linear source separation can be classified according to the mixing operation to instantaneous source separation [2, 11, 17] and convolutive source separation [5-10, 13, 14, 16]. In the instantaneous case the mixture is constructed from the sources by adding them without considering the reverberation of the sources due to the surrounding environment, i.e., considering the direct path of the sources and neglecting their echoes, which is unrealistic as it is impossible to avoid the echo of the signal. The instantaneous mixture can be expressed as follows

$$x_i(t) = \sum_{j=1}^{J} a_{ij} s_j(t) \tag{2.2}$$

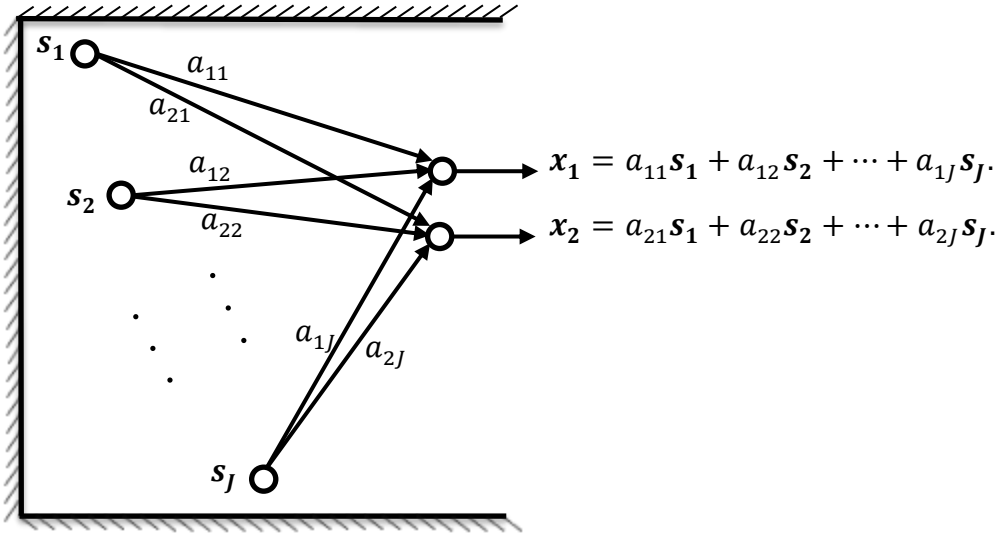
where a_{ij} is mixing filter associated with the sound propagate from source j to channel i.

While in the convolutive mixture the reverberation of the sources due to the surrounding environment are considered by modelling the direct path and the echoes of the sources. The mixture signal of the convolutive can be expressed as follows

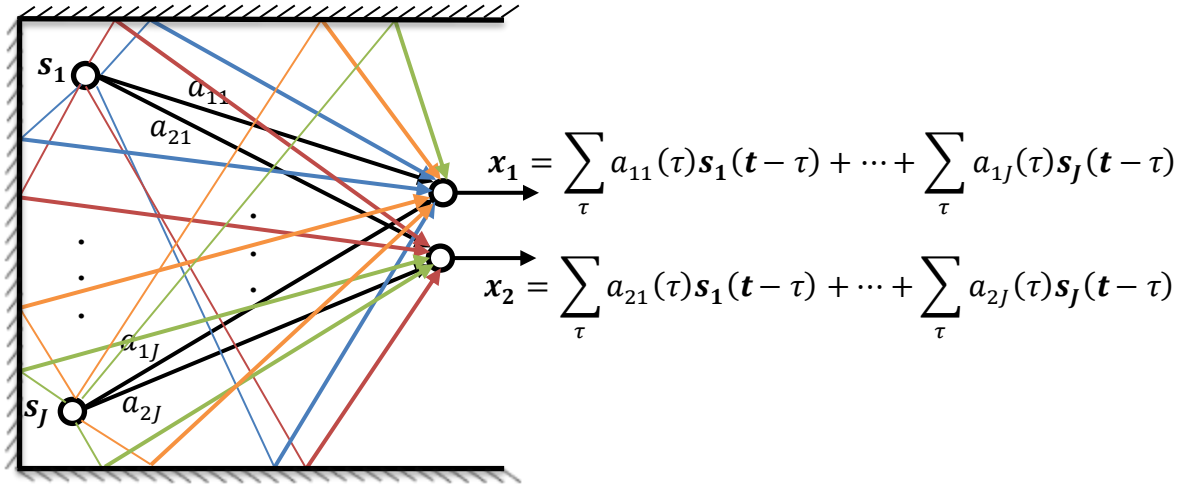
$$x_i(t) = \sum_{j=1}^{J} \sum_{\tau=0}^{L-1} a_{ij}(\tau) s_j(t-\tau) \tag{2.3}$$

where $a(\tau)$ is the finite-impulse response of some (causal) filter, and L is the filter length.

Figure 2.1 shows the difference between the instantaneous and convolutive mixture.



(a) Instantaneous mixture.



(b) Convolutive mixture.

Figure 2.1: Difference between instantaneous and convolutive mixtures.

Furthermore, by depending on the number of sources J and the number of channels I the following cases can be hold:

a. If $I = 1$; then it is a single channel case [2].

b. If $1 < I < J$; then it is the underdetermined case [5-11, 13, 14, 16].

c. If $I > J$; then it is the Over-determined case [17].

d. If $I = J$; then it is the determined case [17].

In addition to the above classifications, the NMF2D [25-32] can be appended to them. These methods consider a single channel (except [30] where it consider a stereo channel) with linear instantaneous mixture. The instantaneous mixture is not realistic as it does not consider the reverberation of the audio sources, therefore, in this thesis a more realistic case, which is the convolutive mixture will be considered.

In the following sections source separation that is based on factorization techniques such as the NMF, the NMF2D, the NTF, and the NTF2D will be explained in details.

2.2 Nonnegative Matrix Factorization (NMF)

NMF [33-36] is dimension reduction technique that is applied to the nonnegative data where it used as a part representation of the data instead of holistic representation as in principle component analysis (PCA), independent component analysis (ICA), and singular value decomposition (SVD). It is part representation instead of holistic representation, because, NMF represent each part of the data by the basis matrix and its corresponding distribution (encoding) matrix, e.g., [35] showed that each part of the face can be represented by the NMF while it is not possible to do that if the PCA or the vector quantization (VQ) is used, because they represent the whole face and not part of it. As the NMF works on positive data only; then there will be no cancellation in the data if it contains positive and negative values (i.e., it does not result in subtraction of any of the nonnegative data), as in PCA, ICA, and SVD. This feature attracts many researchers and makes it well known in the audio source separation community, due to the nature of sound signal (where the sounds from different sources are combined with each other and not cancel each other) that match the NMF feature. A great deals of research have been undertaken under the umbrella of the NMF in many applications, such as bioinformatics [37], digital watermark [38], image processing [39, 40], facial

recognition [41], audiovisual document structuring [42], speech enhancement [43], audio inpainting [44, 45], audio declipping [46, 47], direction of arrival (DOA) estimation [48], blind source separation [5, 6, 8, 13, 14, 25, 49-52], and the informed source separation [15, 53-56]. The other feature of the NMF which is the dimensionality reduction will be explained after understanding how the NMF works. A comparison between the NMF and the other factorization technique can be found in [57]. While, a comprehensive review about the NMF can be found in [58].

NMF can be summarized as follows, if X is an $F \times N$ data matrix with nonnegative entries, then NMF can decompose it as follows

$$\left|\hat{x}_{fn}\right| \approx \sum_{k=1}^{K} w_{f,k} h_{k,n} \tag{2.4a}$$

and in matrix form

$$\left|\hat{X}\right| \approx WH \tag{2.4b}$$

where $W = \{w_{f,k}\} \in \mathbb{R}^{+}$ is nonnegative matrix of dimension $F \times K$ that contains the basis of the data, and $H = \{h_{k,n}\} \in \mathbb{R}^{+}$ is nonnegative matrix of dimension $K \times N$ that contains the distribution of the basis in W matrix, K is the number of the basis (latent components) and it usually selected less than F and N, in order to achieve the decompositions, where $F \times K + K \times N \ll F \times N$, therefore, NMF considered as a dimensionality reduction technique.

From an audio point of view, the columns of W represent the frequency basis and their corresponding rows in H represent the time representation of these frequency basis, K represents the number of frequency basis, F is the number of frequency bins, and N is the number of time frames.

The factorization of eqn. (2.4) can be achieved by optimization method, as follows:

$$\min_{W,H \geq 0} \; D(|X||WH) \tag{2.5}$$

where $D(|X||WH)$ is the divergence between the mixture signal $|X|$, and the estimated (or approximated) signal WH in order to measure error between these two signals, this divergence can be expressed as follows

$$D(|X||WH) = \sum_{f=1}^{F}\sum_{n=1}^{N} d\left(|x_{fn}|\,\middle\|\, \sum_k w_{f,k}h_{k,n}\right) \tag{2.6}$$

The cost function in eqn. (2.6) can be solved by using Euclidian distance (EDU) [34], KullBack-Leibler (KL) [36], Itakura-Saito (IS) [49], β-divergence [59, 60], α-divergence [61], γ-divergence [62], Csiszàr's φ-divergence [63], Bregman divergence [64], and α-β-divergence [65].

The most common cost functions are the EUC, KL, and IS which are derived from the β-divergence which is a family of cost functions that tuned by β parameter as follows

$$D(x|y)_\beta \overset{def}{=} \begin{cases} \dfrac{1}{\beta(\beta-1)}\left((x)^\beta + (\beta-1)(y)^\beta - \beta x(y)^{\beta-1}\right) & \beta \in \mathbb{R}\backslash\{0,1\} \\ x log\dfrac{x}{y} + (y-x) & \beta = 1 \\ \dfrac{x}{y} - log\dfrac{x}{y} - 1 & \beta = 0 \end{cases} \tag{2.7}$$

If β =2 this will leads to the EUC distance, if β =1 this will leads to the KL divergence, and if β =0 this will leads to the IS divergence, and can be expressed as follows

$$D_{EUC}(|X||WH) = \frac{1}{2}\sum_{fn}\left(|x_{fn}| - \sum_k w_{f,k}h_{k,n}\right)^2 \tag{2.8}$$

$$D_{KL}(|X||WH) = \sum_{fn}\left(|x_{fn}| log\frac{|x_{fn}|}{\sum_k w_{f,k}h_{k,n}} - |x_{fn}| + \sum_k w_{f,k}h_{k,n}\right) \tag{2.9}$$

$$D_{IS}(|X||WH) = \sum_{fn}\left(\frac{|x_{fn}|}{\sum_k w_{f,k}h_{k,n}} - log\frac{|x_{fn}|}{\sum_k w_{f,k}h_{k,n}} - 1\right) \tag{2.10}$$

The multiplicative update (MU) rule for the above cost functions can be derived by using the gradient descent approach [36, 66], as follows

$$\theta \leftarrow \theta . \frac{[\nabla f(\theta)]_-}{[\nabla f(\theta)]_+} \tag{2.11}$$

where $\nabla f(\theta) = [\nabla f(\theta)]_+ - [\nabla f(\theta)]_-$. By applying eqn. (2.11) to eqn. (2.8), eqn. (2.9) and eqn. (2.10) the following update rules can be obtained:

For EUC

$$W \leftarrow W . \frac{|X| H^T}{WHH^T} \tag{2.12a}$$

$$H \leftarrow H . \frac{W^T |X|}{W^T WH} \tag{2.12b}$$

For KL

$$W \leftarrow W . \frac{(|X| ./ WH) H^T}{\mathbf{1} H^T} \tag{2.13a}$$

$$H \leftarrow H . \frac{(|X| ./ WH) W^T}{\mathbf{1} W^T} \tag{2.13b}$$

For IS

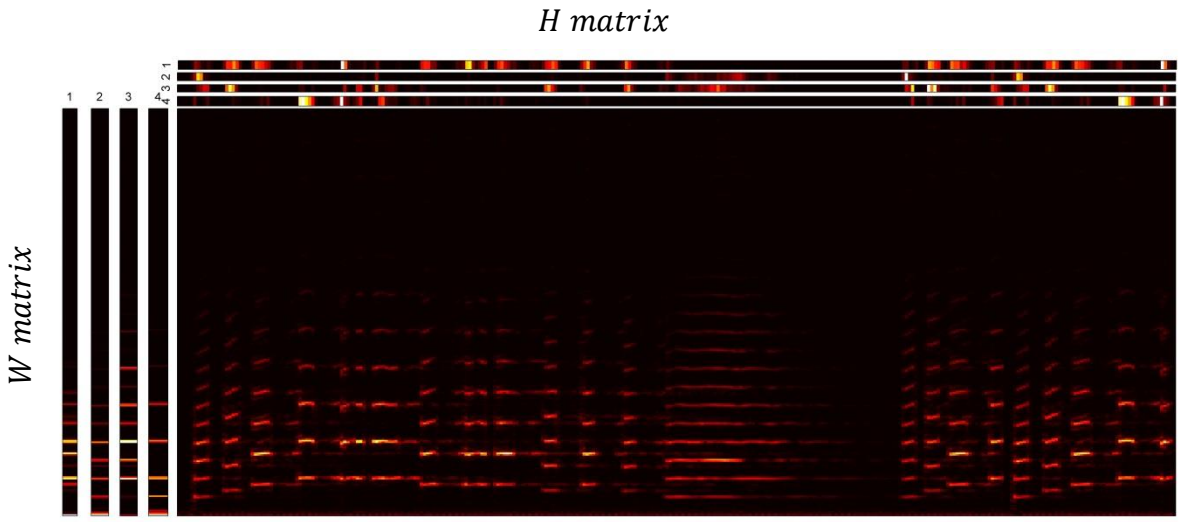
$$W \leftarrow W . \frac{(|X| ./ (WH)^{.2}) H^T}{H^T ./ WH} \tag{2.14a}$$

$$H \leftarrow H . \frac{(|X| ./ (WH)^{.2}) W^T}{W^T ./ WH} \tag{2.14b}$$

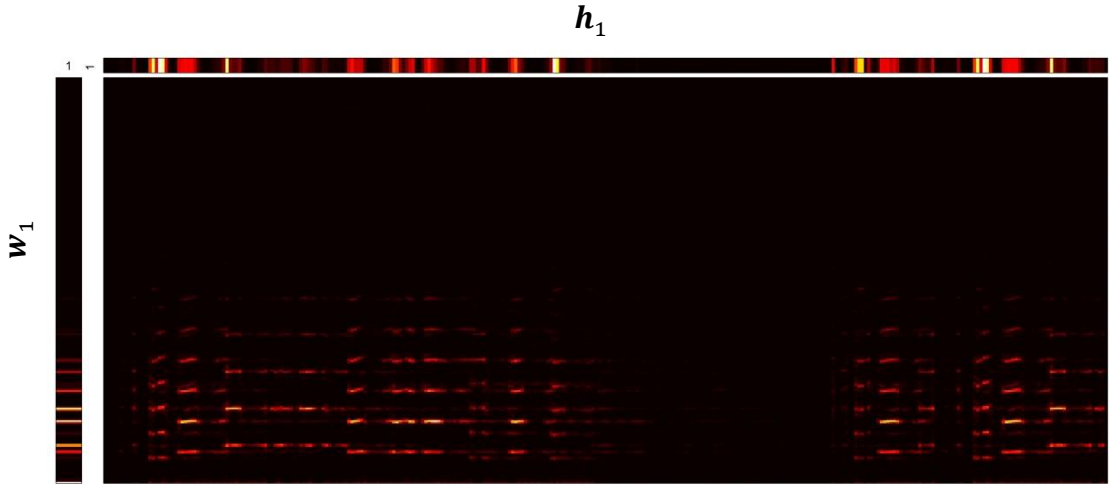
where '.' and './' are the element wise multiplication and division, respectively. A more details about the NMF and its algorithms can be found in [67].

Figure 2.2 below shows the spectrogram (visual representation of the signal in the short time Fourier transform (STFT) in which the x-axis represents the time frame and the y-axis represents the frequency bins) of trumpet signal with its corresponding W and H matrices. Figure 2.2a shows the whole spectrogram of the signal and how the W matrix represents the four frequency basis (i.e., the W matrix has four columns) and how the H matrix (that has four rows, i.e., one row for each

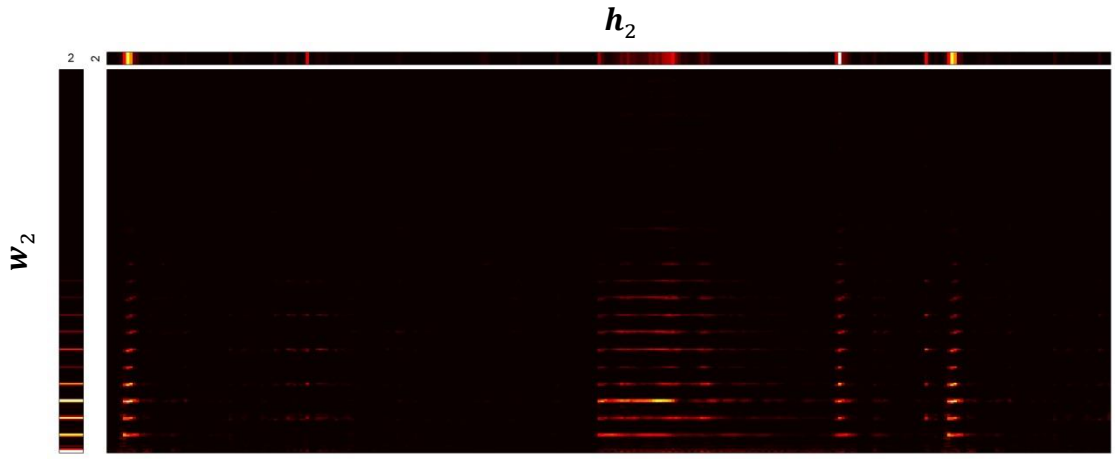
column in the W matrix) distributes them. While Figure 2.2b to Figure 2.2e show the spectrogram of each component and how the W and H matrices construct it. It can be seen that the whole spectrogram (Figure 2.2a) can be reconstructed by adding the spectrograms of Figure 2.2b - Figure 2.2e.



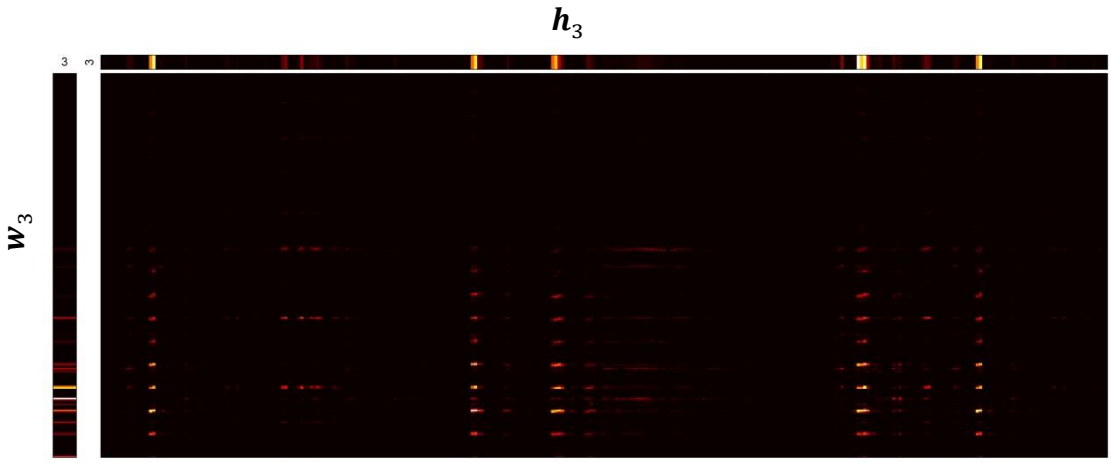
(a) Spectrogram of all components (whole signal).



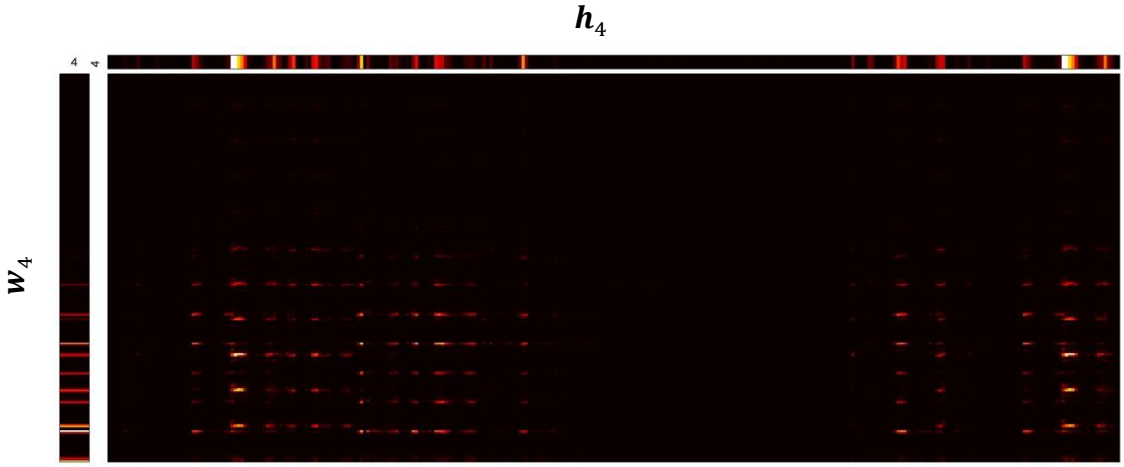
(b) Spectrogram of component 1.



(c) Spectrogram of component 2.



(d) Spectrogram of component 3.



(e) Spectrogram of component 4.

Figure 2.2 How the W and H matrices factorize the signal in the NMF.

2.3 Nonnegative Matrix Factor 2D Deconvolution (NMF2D)

Smaragdis [68] extended the NMF to the nonnegative matrix deconvolution (NMD), where he considered that each frequency base has a sequence of spectra that convolute with its corresponding temporal code, i.e., each column in the W matrix has a sequence of different spectra that convolute with the row of H matrix, and its model can be expressed as follows

$$|\hat{x}_{fn}| \approx \sum_{k=1}^{K} \sum_{\tau=0}^{\tau_{max}} w_{f,k}^{\tau} \overrightarrow{h}_{n,k}^{\tau} \tag{2.15}$$

where the arrow sign in $\overrightarrow{H}_{n,k}^{\tau}$ denotes the right shift operator which moves each element in the matrix by τ column to the right, and τ_{max} is the maximum number of the spectra for each frequency base, thus the W matrix will have τ_{max} columns for each frequency base. The applications of the NMFD can be found in [69-72].

After that paper several developments have been taking place such as the nonnegative matrix factor 2D deconvolution (NMF2D) [25], where it considered both the temporal structure and pitch change that occur when a musical instrument plays different notes. NMF2D considered that each frequency base has a sequence of spectra (represented by τ, see eqn. (2.16)) with its corresponding sequence of temporal code (represented by ϕ, see eqn. (2.16)). τ and ϕ called the convolutive parameters. The NMF2D model can be expressed as follows

$$|\hat{x}_{fn}| \approx \sum_{j=1}^{J} \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} \overset{\downarrow\phi}{w}_{f,j}^{\tau} \overrightarrow{h}_{j,n}^{\phi\,\tau} \tag{2.16}$$

where $w_{f,j}^{\tau}$ represents the spectral basis of the j^{th} source and $h_{j,n}^{\phi}$ represents the temporal code for each spectral basis element of the j^{th} source, for $f = 1, \ldots, F, n = 1, \ldots, N, and\ j = 1, \ldots, J$. The terms τ_{max} and ϕ_{max} are the maximum number of the convoltuive parameters τ and ϕ, respectively. In eqn. (2.16), the superscript upper arrow sign in $\overset{\downarrow\phi}{w}_{f,j}^{\tau}$ denotes downward shift operator which moves each element in the matrix by ϕ row down. At the same time, the arrow sign in $\overrightarrow{h}_{j,n}^{\phi\,\tau}$ denotes the right shift operator which moves each element in the matrix by τ column to the right. The NMF2D considered one frequency basis for each source, i.e., $K = J$.

The EUC [25], KL [25], and IS [27] cost functions of the NMF2D can be expressed as follows

$$D_{\text{EUC}}\left(|X|\middle|\sum_{\tau,\phi} {}^{\downarrow\phi}_{W^{\tau}}\ {}^{\rightarrow\tau}_{H^{\phi}}\right)=\frac{1}{2}\sum_{fn}\left(|x_{fn}|-\sum_{j,\tau,\phi} {}^{\downarrow\phi}_{w^{\tau}_{f,j}}\ {}^{\rightarrow\tau}_{h^{\phi}_{j,n}}\right)^{2} \tag{2.17}$$

$$D_{KL}\left(|X|\middle|\sum_{\tau,\phi} {}^{\downarrow\phi}_{W^{\tau}}\ {}^{\rightarrow\tau}_{H^{\phi}}\right)=\sum_{fn}\left(|x_{fn}|log\frac{|x_{fn}|}{\sum_{j,\tau,\phi} {}^{\downarrow\phi}_{w^{\tau}_{f,j}}\ {}^{\rightarrow\tau}_{h^{\phi}_{j,n}}}-|x_{fn}|+\sum_{j,\tau,\phi} {}^{\downarrow\phi}_{w^{\tau}_{f,j}}\ {}^{\rightarrow\tau}_{h^{\phi}_{j,n}}\right) \tag{2.18}$$

$$D_{IS}\left(|X|\middle|\sum_{\tau,\phi} {}^{\downarrow\phi}_{W^{\tau}}\ {}^{\rightarrow\tau}_{H^{\phi}}\right)=\sum_{fn}\left(\frac{|x_{fn}|}{\sum_{j,\tau,\phi} {}^{\downarrow\phi}_{w^{\tau}_{f,j}}\ {}^{\rightarrow\tau}_{h^{\phi}_{j,n}}}-log\frac{|x_{fn}|}{\sum_{j,\tau,\phi} {}^{\downarrow\phi}_{w^{\tau}_{f,j}}\ {}^{\rightarrow\tau}_{h^{\phi}_{j,n}}}-1\right) \tag{2.19}$$

by applying eqn. (2.11) to eqn. (2.17), eqn. (2.18) and eqn. (2.19) the following update rules can be obtained:

For EUC [25]

$$W^{\tau}\leftarrow W^{\tau}.\frac{\sum_{\phi} {}^{\uparrow\phi}_{|X|}\ {}^{\rightarrow\tau\ T}_{H^{\phi}}}{\sum_{\phi} {}^{\uparrow\phi}_{\Lambda}\ {}^{\rightarrow\tau\ T}_{H^{\phi}}} \tag{2.20a}$$

$$H^{\phi}\leftarrow H^{\phi}.\frac{\sum_{\tau} {}^{\downarrow\phi^{T}}_{W^{\tau}}\ {}^{\leftarrow\tau}_{|X|}}{\sum_{\tau} {}^{\downarrow\phi^{T}}_{W^{\tau}}\ {}^{\leftarrow\tau}_{\Lambda}} \tag{2.20b}$$

For KL [25]

$$W^{\tau}\leftarrow W^{\tau}.\frac{\sum_{\phi} {}^{\uparrow\phi}_{\left(\frac{|X|}{\Lambda}\right)}\ {}^{\rightarrow\tau\ T}_{H^{\phi}}}{\sum_{\phi}\mathbf{1}.\ {}^{\rightarrow\tau\ T}_{H^{\phi}}} \tag{2.21a}$$

$$H^{\phi}\leftarrow H^{\phi}.\frac{\sum_{\tau} {}^{\downarrow\phi^{T}}_{W^{\tau}}\ {}^{\leftarrow\tau}_{\left(\frac{|X|}{\Lambda}\right)}}{\sum_{\tau} {}^{\downarrow\phi^{T}}_{W^{\tau}}.\mathbf{1}} \tag{2.21b}$$

For IS [27]

$$W^{\tau} \leftarrow W^{\tau}.\frac{\sum_{\phi}\left(\left(\overset{\uparrow\phi}{\Lambda}\right)^{\cdot -2}\cdot \overset{\uparrow\phi}{|X|}\right)\overset{\rightarrow\tau}{H^{\phi}}^{T}}{\sum_{\phi}\left(\left(\overset{\uparrow\phi}{\Lambda}\right)^{\cdot -1}\right)\overset{\rightarrow\tau}{H^{\phi}}^{T}} \tag{2.22a}$$

$$H^{\phi} \leftarrow H^{\phi}.\frac{\sum_{\tau}\overset{\downarrow\phi}{W^{\tau}}^{T}\left(\left(\overset{\leftarrow\tau}{\Lambda}\right)^{\cdot -2}\cdot \overset{\leftarrow\tau}{|X|}\right)}{\sum_{\tau}\left(\overset{\downarrow\phi}{W^{\tau}}^{T}\left(\overset{\leftarrow\tau}{\Lambda}\right)^{\cdot -1}\right)} \tag{2.22b}$$

where $\Lambda = \sum_{\tau,\phi}\overset{\downarrow\phi}{W^{\tau}}\overset{\rightarrow\tau}{H^{\phi}}$.

Figure 2.3 below shows how the NMF2D factorizes the trumpet signal for $\tau = \{0, 1, 2, 3\}$ and $\phi = \{0, 1, 2\}$. The figure shows that the same result of NMF has been obtained with only one frequency base but with different convolutive parameters. It can be seen from Figure 2.2 and Figure 2.3 that both the NMF and NMF2D give the same factorization results, as it has been applied to factorize one source only and not a real mixture, i.e., the idea here is to show how the NMF and NMF2D work. The comparison between the NMF and NMF2D will be shown in the following chapters.

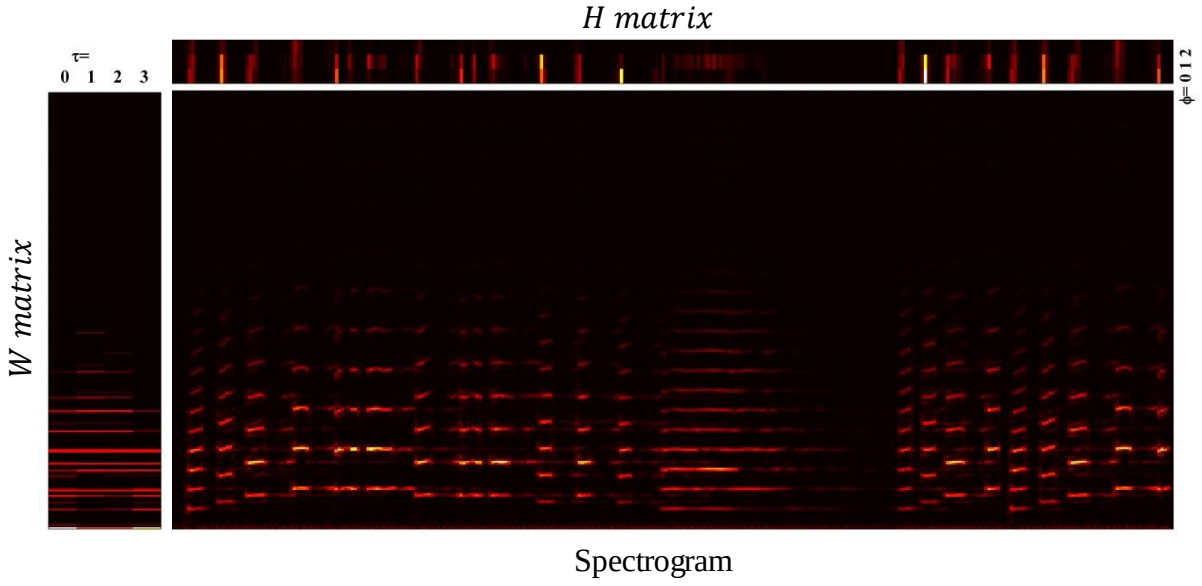


Figure 2.3: How the W and H matrices factorize the signal in the NMF2D.

After Schmidt et al. [25], the primary focus for the researchers was to extend the NMF2D by considering more constraints to be added to the cost functions and following the same procedure in deriving W and H, by using the multiplicative gradient descent approach [26-29, 31, 32]. The NMF2D is more powerful than the NMF in representing complex musical instruments due to its ability in controlling the pitch and temporal change through τ and ϕ, for the specific mixture of musical sources, where some sources have a high pitch but low temporal, and vice versa. If the NMF is considered for these sources, then an equal amount of components will be given for the mixture, which will lead to overfit, or underfit model. However, if τ_{max} and ϕ_{max} are chosen more than the actual requirement, then they will break the structure of the audio signal, i.e., $w_{f,j}^{\tau}$ and $h_{j,n}^{\phi}$ will be shifted more than the actual requirement. This will generate undesirable spurious artefacts to the audio signal and subsequently leads to interference. Therefore, the Gamma-Exponential process to estimate the convolutive parameters of the NMF2D will be proposed.

2.4 Nonnegative Tensor Factorization (NTF)

Nonnegative tensor factorization (or sometimes called Nonnegative tensor decomposition) has many applications in signal processing including source separation [73, 74]. The NTF extend the NMF to model the stereo channel [75, 76] instead of the single channel. Thus it extends the mixture signal from two dimensional matrix to three dimensional tensor $\hat{x}_{i,fn}$, where i is the channel index, and $i = 1,2$. This three dimensional tensor signal has been realized by invoking a channel gain $q_{k,i}$ for the components of each channel. Therefore, $\hat{x}_{i,fn}$ can be expressed as follows

$$\left|\hat{x}_{i,fn}\right| = \sum_{k=1}^{K} q_{k,i} w_{f,k} h_{k,n} \tag{2.23}$$

This model is equivalent to the parallel factor analysis (PARAFAC) [77] with nonnegative constrained.

The factorization of eqn. (2.23) can be achieved by optimization method, as follows:

$$\min_{Q,W,H\geq 0} D\left(|\boldsymbol{X}|\middle|\hat{\boldsymbol{X}}\right) \tag{2.24}$$

The divergence between the mixture signal and the estimated (or approximated) signal can be expressed by β-divergence as follows

$$D(|\boldsymbol{X}||\hat{\boldsymbol{X}}) = \sum_{i=1}^{I}\sum_{f=1}^{F}\sum_{n=1}^{N} d_{\beta}\left(|x_{i,fn}||\sum_{k} q_{k,i} w_{f,k} h_{k,n}\right) \tag{2.25}$$

By applying eqn. (2.11) to eqn. (2.25) the following update rules [73] will be obtained:

$$Q \leftarrow Q.\frac{\langle|\boldsymbol{X}|.\hat{\boldsymbol{X}}^{(\beta-2)}, W \circ H\rangle_{\{2,3\}\{1,2\}}}{\langle\hat{\boldsymbol{X}}^{(\beta-1)}, W \circ H\rangle_{\{2,3\}\{1,2\}}} \tag{2.26a}$$

$$W \leftarrow W.\frac{\langle|\boldsymbol{X}|.\hat{\boldsymbol{X}}^{(\beta-2)}, Q \circ H\rangle_{\{1,3\}\{1,2\}}}{\langle\hat{\boldsymbol{X}}^{(\beta-1)}, Q \circ H\rangle_{\{1,3\}\{1,2\}}} \tag{2.26b}$$

$$H \leftarrow H.\frac{\langle|\boldsymbol{X}|.\hat{\boldsymbol{X}}^{(\beta-2)}, Q \circ W\rangle_{\{1,2\}\{1,2\}}}{\langle\hat{\boldsymbol{X}}^{(\beta-1)}, Q \circ W\rangle_{\{1,\}\{1,2\}}} \tag{2.26.c}$$

where $W \circ H$ is $F \times N \times K$ tensor, $Q \circ H$ is $I \times N \times K$ tensor, $Q \circ W$ is $I \times F \times K$ tensor, and $\langle A, B\rangle_{\{C\}\{D\}}$ is contract product [67] that determined which slices of the tensor have to be multiplied.

If β =2 this will leads to the EUC distance, if β =1 this will leads to the KL divergence, and if β =0 this will leads to the IS divergence.

2.5 Nonnegative Tensor Factor Double Deconvolution (NTF2D)

It is an extension for the NMF2D to deal with multichannel (stereo channel) [30] instead of single channel, by invoking a channel gain $q_{j,i}$ and model it with the 2-D convolutive PARAFAC. The mixture signal $\hat{x}_{i,fn}$ can be expressed as follows

$$|\hat{x}_{i,fn}| = \sum_{j=1}^{J}\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} q_{j,i} \overset{\downarrow\phi}{w}{}^{\tau}_{f,j} \overset{\rightarrow\tau}{h}{}^{\phi}_{j,n} \tag{2.27}$$

The NTF2D considered one frequency base (one component) for each source, i.e., $K = J$, therefore $q_{j,i}$ is the gain between the sources and the channels and not the gain between the components of each source and the channels.

According to [30] two divergences the EUC distance and the KL divergence have been applied which lead to the following updates

For EUC

$$Q \leftarrow Q.\frac{|X|_{(1)}.Z + Qdiag\left(\mathbf{1}\left((QZ^TZ).Q\right)\right)}{QZ^TZ + Qdiag\left(\mathbf{1}\left(\left(|X|_{(1)}.Z\right).Q\right)\right)} \tag{2.28a}$$

$$W^\tau \leftarrow W^\tau.\frac{\sum_\phi \overset{\uparrow\phi}{|X|}_{(2)}\left(\overset{\downarrow\tau}{H^\phi}\odot Q\right) + W^\tau diag\left(\mathbf{1}\sum_\tau\left(\overset{\uparrow\phi}{\hat{X}}_{(2)}\left(\overset{\downarrow\tau}{H^\phi}\odot Q\right)\right).W^\tau\right)}{\sum_\phi \overset{\uparrow\phi}{\hat{X}}_{(2)}\left(\overset{\downarrow\tau}{H^\phi}\odot Q\right) + W^\tau diag\left(\mathbf{1}\sum_\tau\left(\overset{\uparrow\phi}{|X|}_{(2)}\left(\overset{\downarrow\tau}{H^\phi}\odot Q\right)\right).W^\tau\right)} \tag{2.28b}$$

$$H^\phi \leftarrow H^\phi.\frac{\sum_\tau \overset{\uparrow\tau}{|X|}_{(3)}\left(\overset{\downarrow\phi}{W^\tau}\odot Q\right)}{\sum_\tau \overset{\uparrow\tau}{\hat{X}}_{(3)}\left(\overset{\downarrow\phi}{W^\tau}\odot Q\right)} \tag{2.28c}$$

For KL

$$Q \leftarrow Q.\frac{\frac{|X|_{(1)}}{QZ^T} + Qdiag\left(\mathbf{1}.\left((\mathbf{1}Z).Q\right)\right)}{\mathbf{1}Z + Qdiag\left(\mathbf{1}.\left(\frac{|X|_{(1)}}{QZ^T}.Z.Q\right)\right)} \tag{2.29a}$$

$$W^{\tau} \leftarrow W^{\tau}.\frac{\sum_{\phi}\left(\frac{\overset{\uparrow\phi}{|X|}_{(2)}}{\hat{X}_{(2)}}\right)\left(\overset{\downarrow\tau}{H^{\phi}}\odot Q\right)+W^{\tau}diag\left(\mathbf{1}.\sum_{\tau}\left(\mathbf{1}\left(\overset{\downarrow\tau}{H^{\phi}}\odot Q\right)\right).W^{\tau}\right)}{\sum_{\phi}\mathbf{1}\overset{\downarrow\tau}{H^{\phi}}+W^{\tau}diag\left(\mathbf{1}\sum_{\tau}\left(\left(\frac{\overset{\uparrow\phi}{|X|}_{(2)}}{\hat{X}_{(2)}}\right)\left(\overset{\downarrow\tau}{H^{\phi}}\odot Q\right)\right).W^{\tau}\right)} \quad (2.29b)$$

$$H^{\phi} \leftarrow H^{\phi}.\frac{\sum_{\tau}\left(\frac{\overset{\uparrow\tau}{|X|}_{(3)}}{\hat{X}_{(3)}}\right)\left(\overset{\downarrow\phi}{W^{\tau}}\odot Q\right)}{\sum_{\tau}\mathbf{1}\left(\overset{\downarrow\phi}{W^{\tau}}\odot Q\right)} \quad (2.29c)$$

where $\odot$ is the Khatri-Rao product[1], $Z \in \boldsymbol{Z}^{(F.N)\times J}=\left(\sum_{\tau}\sum_{\phi}\overset{\downarrow\phi}{W^{\tau}}\odot\overset{\downarrow\tau}{H^{\phi}}\right)^{T}$, $X_{(1)} \in \boldsymbol{X}^{I\times(F.N)}$, $X_{(2)} \in \boldsymbol{X}^{F\times(I.N)}$, $\hat{X}_{(2)}=\sum_{\tau}\sum_{\phi}\overset{\downarrow\phi}{W^{\tau}}\left(\overset{\downarrow\tau}{H^{\phi}}\odot Q\right)^{T}$, $X_{(3)} \in \boldsymbol{X}^{N\times(I.F)}$, and $\hat{X}_{(3)}=\sum_{\tau}\sum_{\phi}\left(\overset{\downarrow\tau}{H^{\phi}}\right)^{T}\left(\overset{\downarrow\phi}{W^{\tau}}\odot Q\right)^{T}$.

2.6 Informed Source Separation

In informed source separation [20] an additional information about the sources (or even the sources themselves) in addition to the mixture are usually provided to the separation algorithm. This additional information is provided in order to improve the separation performance and to reach the separation quality that the blind source separation cannot reach.

The informed source separation use the side information to provide the extra information, and accordingly it can be classified as follows

1 The Khatri-Rao product can be defined as follows: Let A and B be a matrices of dimensions $F \times K$ and $N \times K$, respectively, then $A\odot B=\left[vec\left(\boldsymbol{a_1}\boldsymbol{b_1^T}\right)\ldots vec\left(\boldsymbol{a_K}\boldsymbol{b_K^T}\right)\right]_{FN\times K}=\left[\left(\boldsymbol{a_1}\otimes\boldsymbol{b_1}\right)\ldots\left(\boldsymbol{a_K}\otimes\boldsymbol{b_K}\right)\right]_{FN\times K}$, where $\otimes$ is the Kronecker product.

2.6.1 Score Informed Source Separation

In this method the parameters of the separation algorithm are initialized by depending on the side information that are available from the Musical Instrument Digital Interface (MIDI) files (sometimes they are called musical scores), such as the onset time, pitch, and duration of the musical notes [53, 78]. An overview of the score informed source separation can be found in [79]. Furthermore, similar to this idea the user can manually set or rest the $\boldsymbol{H}$ matrix in the NMF model [15, 80].

2.6.2 Exemplar-Based Source Separation

Here the informed source separation targeted a specific source in the mixture by providing another source that is similar to the one to be separated. Such as the user mimic the targeted source by singing [54], by humming [81], or by dubs the dialog in films [82]. Furthermore, using an additional audio references as a side information such as using the multitrack cover version of the same song [56, 83-85] or using several international versions of the same movie [55]. Additionally the text can be used as side information to mimic the targeted speech signal [86].

In both the score informed source separation and exemplar-based informed source separation there is a need for a synthesizer to convert them to music. In the score informed source separation an MIDI synthesizer or a user is usually used to convert the scores to music in order to use them as side information with the audio mixture. Similarly, the Exemplar-based informed source separation (especially the text based one) use a speech synthesizer or a user to convert the texts to music.

2.6.3 Coding Based Informed Source Separation

It is two stages scenario that contains the encoding stage and the decoding stage. At the encoding stage all the sources are available in addition to the mixture in order to generate a side information that can be transmitted with the mixture or be embedded in the mixture [87], and will be used in the decoder stage to separate the sources [88-91]. Ozerov et al. [91] show that the coding based informed source separation can outperforms the oracle estimation, if the required bitrate provided. It is bitrate vs quality of separation in this type of informed source separation, as it takes advantage from both source coding and source separation.

Among these types of informed source separation, the exemplar based informed source separation has been pursued in this thesis as the MIDI files are not always available in the case of the score informed source separation. Also, the coding based informed source separation did not progress far as it investigates the quality of separation achieved in terms of the available bitrate, and therefore it is far from the scope covered of this thesis; however it is very prompting future work if it can be proven that the NMF2D can achieve better performance and lower bitrate than the NMF.

2.7 Parameters Effecting The NMF/NMF2D

There are many parameters that effects on the NMF such as the cost function, initialization, number of components, and window's size. For the NMF2D the convolutive parameters can be added to these parameters. These parameters can be explained as follows

2.7.1 Cost function: The cost function with Itakura-Saito divergence will be considered in this thesis, due its scale invariant properties [49], which is important because it will deals with the low and high energy components equally, compared with the EUC distance and KL divergence where both methods consider the high energy components and suppress the low one.

2.7.2 Initialization: The initialization is an essential part for the separation because the NMF2D/ NMF are very sensitive to the initialization, where it can lead to convergence to unwanted local minima, while good initialization can lead to faster convergence to the desired solution [92]. A novel initialization method will be proposed to initialize the parameters of the proposed algorithms.

2.7.3 Number of Components and Convolutive Parameters: If number of components (number of frequency basis) is selected lower than the required value then the model will not fit, while if it selected larger than the required value then an overfitting will occur. For the convolutive parameters the wrong selection can destroy the structure of the audio signal. Therefore, selecting the number of components and convolutive parameter is an important factor in the separation, which will be addressed by proposing a novel method that enables selection of all parameters automatically.

2.7.4 Window Length: The spectrograms of the musical instruments act differently under different windows length, where pitched instruments are smooth and continue in temporal direction and discrete in spectral direction, and the opposite for the percussive instruments [93].

Therefore, different windows length will be considered in order to enhance the performance of the separation.

In this thesis all the parameters those effects on the separation will be tackled, by considering the IS divergence as a cost function, proposing a novel initialization method for the proposed separation algorithms, suggesting a novel method for estimating the number of components and the convolutive parameters, and considering the effects of the window's length.

In the proposed separation methods the GEM-MU algorithm [80] will be considered. The GEM-MU algorithm is a hybrid-model that combines both the Expectation-Maximization model and Multiplicative Update rule, it will be explained in the following chapters. Also, the NMF2D has been applied directly on the statistics (e.g., the spectral covariance matrix) instead of the data itself (i.e., the mixed signal or its spectrogram). Hence the domain of interest is required to match with the statistical quantity to be decomposed rather than the data domain. Data domain such as the log-frequency spectrogram is intrinsically a nonlinear transform. The GEM-MU algorithm is developed based on the linear model in the linear spectrogram and as such, the linearity structure will not be preserved in the log-frequency domain. In particular, the NMF2D is used in the M-step of the GEM-MU algorithm which normally is based on the statistics from the E-step of the GEM-MU algorithm. Hence the log-frequency will violate the linearity structure of the statistics from the E-step of the GEM-MU algorithm, and this will leads to breaking the audio structure (signature) of the signal. Therefore as the proposed decomposition does not work directly on the data, it is not necessary to transform the data to the log-frequency domain. Furthermore, the log-frequency will lose information when resynthesizing the estimated sources as any mapping back from log-frequency to linear frequency is only an approximate mapping.

2.8 Summary

In this chapter the blind source separation and the informed source separation have been reviewed. Also, the audio source separation that is based on factorization techniques such as the NMF, NMF2D, NTF, and NTF2D have been discussed. It has been shown that the NMF2D is more flexible than the NMF as it has the ability to control the pitch and temporal changes. Furthermore, parameters that effect on the separation have been highlighted and these will be tackled in Chapters three, four, and five.

CHAPTER 3

BLIND SOURCE SEPARATION USING GAMMA EXPONENTIAL PROCESS AND TWO DIMENSIONAL MATRIX FACTORIZATION TECHNIQUES

In this chapter a novel underdetermined blind source separation algorithm based on the NMF2D with adaptive sparsity[2] will be proposed. The proposed algorithm is adapted in an unsupervised manner under the GEM-MU hybrid framework [80]. As the number of parameters in the NMF2D grows exponentially as the number of frequency basis increases linearly, the issues of model order fitness, initialization and parameters estimation become ever more critical. Furthermore, a novel method that uses the Gamma-Exponential process as an observation-latent model will be proposed to optimize the convolutive parameters and number of components in the NMF2D. Additionally, it is also shown that the parameters of the NMF2D can be initialized by the proposed Gamma-Exponential process. In addition, the issue and advantages of using different window length with different number of convolutive parameters will be investigated in this chapter. Finally, the effectiveness of the proposed algorithm will be verified through the experimental results on the synthetic convolutive mixtures and live recordings mixtures.

The chapter is organized as follows: The proposed model will be introduced in Section 3.1. Section 3.2 is dedicated for the details of the source model. The development of GEM-MU algorithm to work with the NMF2D and with adaptive sparsity will be presented in Section 3.3. In Section 3.4 the Gamma-Exponential process will be proposed for estimating the number of components and convolutive parameters, and initializing the NMF2D. Section 3.5 will discuss the influence of the windows length on the separation. Experimental results will be shown in Section 3.6. Finally, Section 3.7 draws the conclusions.

[2] The sparsity is the penalty on the activation matrix that ensures only a few units (out of a large population) will be active at the same time. The sparsity can be added as a constraint to the cost function [1]P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457-1469, Nov, 2004..

3.1 Introduction

As most research on NMF2D has been limited to instantaneous mixture [25-29, 31, 32] and as the number of sources in most cases is greater than the number of channels, then, in this chapter the case of the underdetermined channel with convolutive mixture will be considered. The proposed NMF2D with adaptive sparsity instead of uniform sparsity will be developed within the framework of the GEM-MU algorithm [80]. Furthermore, the factors that effect on the NMF2D such as the cost function, initialization, windows length, and convolutive parameters will be controlled. The cost function with Itakura-Saito divergence will be considered in this chapter due its advantage of scale invariance properties [49]. This is important because source separation requires us to deal with the low and high energy components equally. Compared with the Euclidian distance (EDU) distance and Kullback-Leibler (KL) divergence, both methods favor the high energy components but suppress the low energy ones. Furthermore, as each musical instrument has its own characteristics in terms of the spectral and temporal features e.g., drum instrument has a high pitch with low temporal note while the opposite is true for the piano; then different windows length will be considered in the separation. To understand the effects of the convolutive parameters on the separation performance, the NMF2D will briefly described. Let $C(n,m)$ be a data matrix of size $N \times M$ with nonnegative entries, then $C(n,m)$ is approximated with two nonnegative tensors $A(n,k,\tau)$ and $B(k,m,\phi)$ as $C(n,m) \approx \sum_{k=0}^{K} \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} A(n-\phi,k,\tau) B(k,m-\tau,\phi)$. The terms K, τ_{max} and ϕ_{max} are the number of components, and the maximum number of the convolutive parameters τ and ϕ, respectively. If τ_{max} and ϕ_{max} are chosen more than the actual requirement, then they will break the structure of the audio signal, i.e., $A(n,k,\tau)$ and $B(k,m,\phi)$ will be shifted more than the actual requirement. This will generate undesirable spurious artefacts to the audio signal and subsequently leads to interference. Therefore, in this chapter a novel method will be proposed to estimate the convolutive parameter. Another dimension for consideration is initialization which is an essential part for the NMF and NMF2D. Good initialization of the model parameters will lead to faster convergence to the desired solution. Therefore, the spectral and temporal tensors of the proposed Gamma-Exponential process will be used to initialize the spectral and temporal tensors of the proposed NMF2D model.

3.2 Source Model

Consider the underdetermined channel with convolutive mixture, namely:

$$\tilde{x}_i(t) = \sum_{j=1}^{J} \sum_{\tau=0}^{L-1} \tilde{a}_{ij}(\tau) \tilde{s}_j(t-\tau) + \tilde{b}_i(t) \tag{3.1}$$

where $\tilde{x}_i(t)(i = 1, \dots, I, t = 1, \dots, T)$ is the sampled mixture signal and I is the number of channels, $\tilde{s}_j\,(j = 1, \dots, J)$ is the source signal and J is the number of sources, $\tilde{a}_{ij}(\tau)$ is the finite-impulse response of some (causal) filter, L is the filter length, and $\tilde{b}_i(t)$ is some additive noise. By assuming that the mixing channel is time-invariant then the short-time Fourier transform (STFT) of eqn. (3.1) can be expressed as

$$x_{i,fn} = \sum_{j=1}^{J} a_{ij,f} s_{j,fn} + b_{i,fn} \tag{3.2a}$$

and in matrix form

$$X_f = A_f S_f + B_f \tag{3.2b}$$

where $X_f = \left[x_{i,fn}\right]_f \in \mathbb{C}^{I \times N}$, $A_f = \left[a_{ij,f}\right]_f \in \mathbb{C}^{I \times J}$, $S_f = \left[s_{j,fn}\right]_f \in \mathbb{C}^{J \times N}$, and $B_f = \left[b_{i,fn}\right]_f \in \mathbb{C}^{I \times N}$ and $f = 1, \dots, F$ is the index of a frequency bin. As the NMF2D with multiple frequency basis will be considered as the spectral variance model in this chapter instead of the NMF spectral model [49], then each source in the STFT can be expressed by K_j complex-valued latent components, i.e.,

$$s_{j,fn} = \sum_{k=1}^{K_j} c_{k,j,fn} \tag{3.3}$$

and can be modeled as realization of proper complex zero-mean variables:

$$\begin{aligned} c_{k,j,fn} &\sim \mathcal{N}_c\left(0, \sigma^2_{k,j,fn}\right) \\ &= \mathcal{N}_c\left(0, \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} w^{\tau,j}_{f-\phi,k} h^{\phi,j}_{k,n-\tau}\right) \end{aligned} \tag{3.4}$$

where $\mathcal{N}_c(\mu, \Sigma)$ is the proper complex Gaussian distribution [94], $w_{f,k}^{\tau,j}$ represents the spectral basis of the j^{th} source, and $h_{k,n}^{\phi,j}$ represents the temporal code for each spectral basis element of the j^{th} source, for $f = 1, \ldots, F, n = 1, \ldots, N, j = 1, \ldots, J,$ and $k = 1, \ldots K_j$. The noise $b_{i,fn}$ is assumed to be stationary and spatially uncorrelated, i.e.

$$b_{i,fn} \sim \mathcal{N}_c\left(0, \sigma_{i,f}^2\right) \tag{3.5a}$$

and

$$\Sigma_{b,f} = diag\left[\sigma_{i,f}^2\right] \tag{3.5b}$$

The parameters $\boldsymbol{A}, \boldsymbol{\Sigma}_b, \boldsymbol{\Lambda}, \boldsymbol{C} = \{c_{k,j,fn}\}, \boldsymbol{W} = \{w_{f,k}^{\tau,j}\}, \boldsymbol{H} = \{h_{k,n}^{\phi,j}\}$ will be estimated via the posterior probability

$$P(\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H} | \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\Sigma}_b, \boldsymbol{\Lambda}) = \frac{P(\boldsymbol{X}|\boldsymbol{C}, \boldsymbol{A}, \boldsymbol{\Sigma}_b) P(\boldsymbol{C}|\boldsymbol{W}, \boldsymbol{H}) P(\boldsymbol{W}, \boldsymbol{H}|\boldsymbol{\Lambda})}{P(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{\Sigma}_b)} \tag{3.6}$$

and their minus log-posterior is given by

$$-\log P(\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H} | \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\Sigma}_b, \boldsymbol{\Lambda}) = -\log P(\boldsymbol{X}|\boldsymbol{C}, \boldsymbol{A}, \boldsymbol{\Sigma}_b) - \log P(\boldsymbol{C}|\boldsymbol{W}, \boldsymbol{H}) - \log P(\boldsymbol{W}, \boldsymbol{H}|\boldsymbol{\Lambda}) + const \tag{3.7}$$

where $\boldsymbol{\Lambda} = \{\lambda_{k,n}^{\phi,j}\}$ is a tensor that contains the sparsity terms.

3.3 Proposed Estimation Algorithm

The GEM-MU [80] combines both the expectation maximization (EM) algorithm and the multiplicative update (MU) algorithm. The source power spectrogram posterior estimates ($\hat{p}_{j,fn}$) (see eqn. (3.12)), the mixing parameter, and the noise covariance will be estimated in the E-step of the EM algorithm, while $\boldsymbol{W}$ and $\boldsymbol{H}$ will be estimated in the M-step of the EM algorithm by using the MU algorithm with adaptive sparsity NMF2D.

3.3.1 E-Step: Conditional Expectations of Natural Statistics

The log-likelihood in the right hand side of eqn. (3.7) can be expressed as

$$-\log P(\boldsymbol{X}|\boldsymbol{C},\boldsymbol{A},\boldsymbol{\Sigma}_b) = \sum_{fn}\left(\boldsymbol{x}_{fn} - A_f\boldsymbol{s}_{fn}\right)^H \Sigma_{b,f}^{-1}\left(\boldsymbol{x}_{fn} - A_f\boldsymbol{s}_{fn}\right) + \sum_{fn} log\, det\, \Sigma_{b,f}$$

$$= N\sum_f tr\{\Sigma_{b,f}^{-1}R_{XX,f}\} - N\sum_f tr\{A_f^H\Sigma_{b,f}^{-1}R_{XS,f}\}$$

$$-N\sum_f tr\left\{\Sigma_{b,f}^{-1}A_f\left(R_{XS,f}\right)^H\right\} + N\sum_f tr\{A_f^H\Sigma_{b,f}^{-1}A_fR_{SS,f}\}$$

$$+\sum_{fn} log\, det\, \Sigma_{b,f} \tag{3.8}$$

where the superscript H is the Hermitian transpose. The correlation matrices are given by

$$R_{XX,f} = \frac{1}{N}\sum_n \boldsymbol{x}_{fn}\boldsymbol{x}_{fn}^H \tag{3.9}$$

$$R_{SS,f} = \frac{1}{N}\sum_n \boldsymbol{s}_{fn}\boldsymbol{s}_{fn}^H \tag{3.10}$$

and the cross-correlation matrix is given by

$$R_{XS,f} = \frac{1}{N}\sum_n \boldsymbol{x}_{fn}\boldsymbol{s}_{fn}^H \tag{3.11}$$

The source power spectrogram posterior estimates [80] is as follows

$$\hat{p}_{j,fn} = \hat{R}_{SS,fn}(j,j) \tag{3.12}$$

where

$$\hat{R}_{SS,fn} = E\left[\boldsymbol{s}_{fn}\right]E\left[\boldsymbol{s}_{fn}^H\right] + \hat{\Sigma}_{s,fn}$$

$$= \hat{\boldsymbol{s}}_{fn}\hat{\boldsymbol{s}}_{fn}^H + \hat{\Sigma}_{s,fn} \tag{3.13}$$

$$\hat{\boldsymbol{s}}_{fn} = \Sigma_{s,fn}A_f^H\Sigma_{x,fn}^{-1}\boldsymbol{x}_{fn} \tag{3.14}$$

$$\hat{\Sigma}_{s,fn} = \left(I_J - \Sigma_{s,fn}A_f^H\Sigma_{x,fn}^{-1}A_f\right)\Sigma_{s,fn} \tag{3.15}$$

$$\Sigma_{X,fn} = A_f \Sigma_{s,fn} A_f^H + \Sigma_{b,f} \tag{3.16}$$

$$\Sigma_{s,fn} = diag\left(\left[\sum_{k=1}^{K_j} \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}\right]_j\right) \tag{3.17}$$

Detailed derivation of eqn. (3.14) and eqn. (3.15) can be found in [14].

3.3.2 M- Step: Update of Parameters

A_f and $\Sigma_{b,f}$, will be estimated as follows

$$\frac{\partial}{\partial A_f} \log P(\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H} | \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\Sigma}_b) = 0 \tag{3.18}$$

which leads to

$$A_f = \hat{R}_{XS,f} \hat{R}_{SS,f}^{-1} \tag{3.19}$$

Similarly,

$$\frac{\partial}{\partial \Sigma_{b,f}^{-1}} \log P(\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H} | \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\Sigma}_b) = 0 \tag{3.20}$$

which leads to

$$\Sigma_{b,f} = diag\left(\hat{R}_{XX,f} - \hat{R}_{XS,f} \hat{R}_{SS,f}^{-1} \hat{R}_{XS,f}^H\right) \tag{3.21}$$

where

$$\hat{R}_{XX,f} = R_{XX,f} \tag{3.22}$$

$$\hat{R}_{XS,f} = \frac{1}{N} \sum_n \boldsymbol{x}_{fn} E\left[\boldsymbol{s}_{fn}^H\right]$$

$$= \frac{1}{N} \sum_n \boldsymbol{x}_{fn} \hat{\boldsymbol{s}}_{fn}^H \tag{3.23}$$

$$\hat{R}_{SS,f} = \frac{1}{N}\sum_{n} \hat{R}_{SS,fn} \tag{3.24}$$

As $\hat{p}_{j,fn}$ is estimated from the E-step, then the second term in the right hand side of eqn. (3.7) can be written in term of $\hat{p}_{j,fn}$ and expressed with Itakura-Saito divergence as

$$-\log P\left(\hat{\boldsymbol{P}}|\boldsymbol{W},\boldsymbol{H}\right) = \sum_{j,f,n} D_{IS}\left(\hat{p}_{j,fn}\middle|\sum_k \sum_\tau \sum_\phi w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}\right) \tag{3.25}$$

where $\hat{\boldsymbol{P}} = \left\{\hat{p}_{j,fn}\right\}_{j,fn}$. The third term in the right hand side of eqn. (3.7) is the prior information on $\boldsymbol{W}$ and $\boldsymbol{H}$. The prior over $\boldsymbol{W}$ is flat where each column is assumed to be factor-wise normalized to unit length i.e. $p(W) = \prod_j \delta\left(\left\|W^j\right\|_2 - 1\right)$. Each element of $\boldsymbol{H}$ has independent decay parameter $\lambda_{k,n}^{\phi,j}$ with exponential distribution:

$$\begin{aligned} p(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{\Lambda}) &= \prod_j \delta\left(\left\|W^j\right\|_2 - 1\right) + \prod_{j,k} p\left(H_k^j|\Lambda_k^j\right) \\ &= \prod_j \delta\left(\left\|W^j\right\|_2 - 1\right) + \prod_j \prod_k \prod_n \prod_\phi p\left(h_{k,n}^{\phi,j}|\lambda_{k,n}^{\phi,j}\right) \\ &= \prod_j \delta\left(\left\|W^j\right\|_2 - 1\right) + \prod_j \prod_k \prod_n \prod_\phi \lambda_{k,n}^{\phi,j} exp\left(-\lambda_{k,n}^{\phi,j} h_{k,n}^{\phi,j}\right) \end{aligned} \tag{3.26}$$

The negative log-likelihood for prior on $\boldsymbol{W}$ and $\boldsymbol{H}$ is derived such as

$$\begin{aligned} -\log p(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{\Lambda}) &= -\log\left(\prod_j \delta\left(\left\|\boldsymbol{W}^j\right\|_2 - 1\right)\right) - \log\left(\prod_j \prod_k \prod_n \prod_\phi \lambda_{k,n}^{\phi,j} \exp\left(-\lambda_{k,n}^{\phi,j} h_{k,n}^{\phi,j}\right)\right) \\ &= -\sum_j \log\delta\left(\left\|\boldsymbol{W}^j\right\|_2 - 1\right) + \sum_j \sum_k \sum_n \sum_\phi \left(\lambda_{k,n}^{\phi,j} h_{k,n}^{\phi,j} - \log\lambda_{k,n}^{\phi,j}\right) \end{aligned} \tag{3.27}$$

The first term on the right hand side of eqn. (3.27) can be satisfied by explicitly normalizing each spectral dictionary to unity i.e. $w_{f,k}^{\tau,j} = w_{f,k}^{\tau,j} / \sqrt{\sum_{f,\tau,k}\left(w_{f,k}^{\tau,j}\right)^2}$. Thus, only the second term remains i.e. $-\log p(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{\Lambda}) = \sum_j \sum_k \sum_n \sum_\phi \left(\lambda_{k,n}^{\phi,j} h_{k,n}^{\phi,j} - \log\lambda_{k,n}^{\phi,j}\right)$. Adding eqn. (3.27) to IS divergence derived in eqn. (3.25), will leads to the following

$$-\log P(\boldsymbol{C}|\boldsymbol{W},\boldsymbol{H}) - \log P(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{\Lambda})$$

$$= \sum_{j,f,n} D_{IS}\left(\hat{p}_{j,fn} \middle| \sum_k \sum_\tau \sum_\phi w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}\right) + \sum_{j,k,n,\phi} \left(\lambda_{k,n}^{\phi,j} h_{k,n}^{\phi,j} - log\lambda_{k,n}^{\phi,j}\right)$$

$$= \sum_{j,k,f,n} \left(\frac{\hat{p}_{j,fn}}{\sum_{\tau,\phi}\left(w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}\right)} - \log \frac{\hat{p}_{j,fn}}{\sum_{\tau,\phi}\left(w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}\right)} - 1 \right)$$

$$+ \sum_{j,k,n,\phi} \lambda_{k,n}^{\phi,j} h_{k,n}^{\phi,j} - \sum_{j,k,n,\phi} log\lambda_{k,n}^{\phi,j} \tag{3.28}$$

Let

$$v_{j,fn} = \sum_k \sum_\tau \sum_\emptyset \left(w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}\right) \tag{3.29}$$

then the derivatives of individual component for proposed model with respect to $w_{f,k}^{\tau,j}$ and $h_{k,n}^{\phi,j}$ can be derived as:

$$\frac{\partial}{\partial w_{f',k'}^{\tau',j'}} \log P(\boldsymbol{C},\boldsymbol{W},\boldsymbol{H}|\boldsymbol{X},\boldsymbol{A},\boldsymbol{\Sigma}_b)$$

$$= -\sum_{\phi,n} \hat{p}_{j',f'+\phi,n} v_{j',f'+\phi,n}^{-2} h_{k',n-\tau'}^{\phi,j'} + \sum_{\phi,n} v_{j',f'+\phi,n}^{-1} h_{k',n-\tau'}^{\phi,j'} \tag{3.30}$$

Similarly,

$$\frac{\partial}{\partial h_{k',n'}^{\phi',j'}} \log P(\boldsymbol{C},\boldsymbol{W},\boldsymbol{H}|\boldsymbol{X},\boldsymbol{A},\boldsymbol{\Sigma}_b) = -\sum_{f,\tau} \hat{p}_{j',f,n'+\tau} v_{j',f,n'+\tau}^{-2} w_{f-\phi',k'}^{\tau,j'} + \sum_{f,\tau} v_{j',f,n'+\tau}^{-1} w_{f-\phi',k'}^{\tau,j'} + \lambda_{k',n'}^{\phi',j'} \tag{3.31}$$

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

$$w_{f',k'}^{\tau',j'} \leftarrow w_{f',k'}^{\tau',j'} - \eta_w \frac{\partial C_{IS}}{\partial w_{f',k'}^{\tau',j'}} \tag{3.32}$$

and

$$h_{k',n'}^{\phi',j'} \leftarrow h_{k',n'}^{\phi',j'} - \eta_h \frac{\partial C_{IS}}{\partial h_{k',n'}^{\phi',j'}} \tag{3.33}$$

where η_w and η_h are the positive learning rate. Based on [35], the positive learning rate can be set as

$$\eta_w = \frac{w_{f',k'}^{\tau',j'}}{\sum_{\phi,n} v_{j',f'+\phi,n}^{-1} h_{k',n-\tau'}^{\phi,j'}} \tag{3.34}$$

and

$$\eta_h = \frac{h_{k',n'}^{\phi',j'}}{\sum_{f,\tau} v_{j',f,n'+\tau}^{-1} w_{f-\phi',k'}^{\tau,j'} + \lambda_{k',n'}^{\phi',j'}} \tag{3.35}$$

The MU rules for $w_{f,k}^{\tau,j}$ is given by

$$w_{f',k'}^{\tau',j'} \leftarrow w_{f',k'}^{\tau',j'} - \frac{w_{f',k'}^{\tau',j'} \left(- \sum_{\phi,n} \hat{p}_{j',f'+\phi,n} v_{j',f'+\phi,n}^{-2} h_{k',n-\tau'}^{\phi,j'} + \sum_{\phi,n} v_{j',f'+\phi,n}^{-1} h_{k',n-\tau'}^{\phi,j'} \right)}{\sum_{\phi,n} v_{j',f'+\phi,n}^{-1} h_{k',n-\tau'}^{\phi,j'}}$$

$$w_{f',k'}^{\tau',j'} \leftarrow w_{f',k'}^{\tau',j'} \left(\frac{\sum_{\phi,n} \hat{p}_{j',f'+\phi,n} v_{j',f'+\phi,n}^{-2} h_{k',n-\tau'}^{\phi,j'}}{\sum_{\phi,n} v_{j',f'+\phi,n}^{-1} h_{k',n-\tau'}^{\phi,j'}} \right) \tag{3.36}$$

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

$$h_{k',n'}^{\phi',j'} \leftarrow h_{k',n'}^{\phi',j'} - \frac{h_{k',n'}^{\phi',j'} \left(- \sum_{f,\tau} \hat{p}_{j',f,n'+\tau} v_{j',f,n'+\tau}^{-2} w_{f-\phi',k'}^{\tau,j'} + \sum_{f,\tau} v_{j',f,n'+\tau}^{-1} w_{f-\phi',k'}^{\tau,j'} + \lambda_{k',n'}^{\phi',j'} \right)}{\sum_{f,\tau} v_{j',f,n'+\tau}^{-1} w_{f-\phi',k'}^{\tau,j'} + \lambda_{k',n'}^{\phi',j'}}$$

$$h_{k',n'}^{\phi',j'} \leftarrow h_{k',n'}^{\phi',j'} \left(\frac{\sum_{f,\tau} \hat{p}_{j',f,n'+\tau} v_{j',f,n'+\tau}^{-2} w_{f-\phi',k'}^{\tau,j'}}{\sum_{f,\tau} v_{j',f,n'+\tau}^{-1} w_{f-\phi',k'}^{\tau,j'} + \lambda_{k',n'}^{\phi',j'}} \right) \tag{3.37}$$

For the sparsity term, the update is obtained by solving $\frac{\partial}{\partial \lambda_{k',n'}^{\phi',j'}} \log P(\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H} | \boldsymbol{X}, \boldsymbol{A}, \boldsymbol{\Sigma}_b) = 0$ which leads to

$$\frac{\partial}{\partial \lambda_{k',n'}^{\phi',j'}} \log P(\boldsymbol{C},\boldsymbol{W},\boldsymbol{H}\,|\boldsymbol{X},\boldsymbol{A},\boldsymbol{\Sigma}_b)$$

$$= \frac{\partial \left(\sum_{fn} \left(\frac{\hat{p}_{j,fn}}{v_{j,f,n}} - log \frac{\hat{p}_{j,fn}}{v_{j,f,n}} - 1 \right) + \sum_{n,\phi} h_{k,n}^{\phi,j} \lambda_{k,n}^{\phi} - \sum_{n,\phi} log \lambda_{k,n}^{\phi,j} \right)}{\partial \lambda_{j',n'}^{\phi'}}$$

$$= h_{k',n'}^{\phi',j'} - \frac{1}{\lambda_{k',n'}^{\phi',j'}} \tag{3.38}$$

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

$$\lambda_{k',n'}^{\phi',j'} = \frac{1}{h_{k',n'}^{\phi',j'}} \tag{3.39}$$

3.3.3 Components Reconstruction

The estimated sources ($\hat{\boldsymbol{s}}_{fn}$) can be reconstructed by using Wiener filtering $\left(\Sigma_{s,fn} A_f^H \Sigma_{x,fn}^{-1}\right)$ as in eqn. (3.14), and due to the linearity of the STFT, the inverse-STFT (with dual synthesis window [95]) can be used to transform it to the time domain.

3.4 Estimating The Number Of Components And Number Of Convolutive Parameters In NMF2D

3.4.1 Variational Bayesian Formulation

The determination of the number of components in NMF has been previously investigated in [96] by means of nonparametric statistical fit. However, the method cannot be directly applied to the NMF2D model as the number of convolutive parameters and number of components will be lumped together. Thus the method in [96] will estimate an overfit model. In this work a constrained Gamma-Exponential process to estimate the convolutive parameters and the number of components of the NMF2D will be proposed. The proposed Gamma-Exponential process introduces a hidden tensor of nonnegative values $\theta_k^{\tau,\phi}$ that weight each element of the factor model ($\sum_{j,k,\tau,\phi} \theta_k^{\tau,\phi} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}$) such that the number of components and convolutive parameters are

inferred automatically based on the mixture power spectrogram p_{fn}^{x} which is estimated from the observations as $|x_{i,fn}|^2$. The model order k, τ, and ϕ are assigned to a large integer values (ideally infinity) and the proposed model will retain a finite number of each subset corresponding to the active elements in θ. To the best of our knowledge, this is the first proposed method on the NMF2D to estimate the number of convolutive parameters of the NMF2D model.

The generative process of the mixture power spectrogram is assumed to follow the Gamma-Exponential process as follows:

$$p_{fn}^{x} \sim Exponential\left(\sum\nolimits_{j,k,\tau,\phi} \theta_k^{\tau,\phi}\, w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}\right) \tag{3.40}$$

$$w_{f,k}^{\tau,j} \sim Gamma\left(a_k^{\tau,j}, a_k^{\tau,j}\right) \tag{3.41}$$

$$h_{k,n}^{\phi,j} \sim Gamma\left(b_k^{\phi,j}, b_k^{\phi,j}\right) \tag{3.42}$$

$$\theta_k^{\tau,\phi} \sim Gamma\left(\frac{\alpha_k^{\tau,\phi}}{L+\phi_{max}+\tau_{max}}, \alpha_k^{\tau,\phi} c\right) \tag{3.43}$$

where L is the truncation level, k number of components, α, a, and b are the shape parameters, and c is the inverse shape parameter $c = \frac{1}{\bar{x}}$, where $\bar{x}$ is the empirical mean of p_{fn}^{x}. The empirical mean of p_{fn}^{x} can be expressed as follows

$$\mathbb{E}_p\left[p_{fn}^{x}\right] = \sum\nolimits_{j,k,\tau,\phi} \mathbb{E}_p\left[\theta_k^{\tau,\phi}\right] \mathbb{E}_p\left[w_{f-\phi,k}^{\tau,j}\right] \mathbb{E}_p\left[h_{k,n-\tau}^{\phi,j}\right]$$

$$= \frac{1}{c} \tag{3.44}$$

The posterior distribution of parameters $\Omega = \left\{\left\{\theta_k^{\tau,\phi}\right\}, \left\{w_{f,k}^{\tau,j}\right\}, \left\{h_{k,n}^{\phi,j}\right\}\right\}$ is approximated by resorting to the generalized inverse Gaussian (GIG) distribution, the statistical properties of the GIG can be found in [97]. The PDF of the GIG distribution is

$$GIG(y;\gamma,\rho,\beta) = \frac{y^{\gamma-1} \exp\left(-\rho y - \frac{\beta}{y}\right)\left(\frac{\rho}{\beta}\right)^{\frac{\gamma}{2}}}{2\mathcal{K}_\gamma\left(2\sqrt{\rho\beta}\right)} \tag{3.45}$$

where $\mathcal{K}_\gamma(\cdot)$ is the modified Bessel function of the second kind and $y \geq 0, \rho \geq 0$, and $\beta \geq 0$. p_{fn}^x can be shown to be lower bounded by

$$\begin{aligned}\log p\left(p_{fn}^x \middle| \alpha_k^{\tau,\phi}, a_k^{\tau,j}, b_k^{\phi,j}, c\right) \geq\ & \mathbb{E}_q\left[\log p\left(p_{fn}^x \middle| w_{f,k}^{\tau,j}, h_{k,n}^{\phi,j}, \theta_k^{\tau,\phi}\right)\right] \\ & + \mathbb{E}_q\left[\log p\left(w_{f,k}^{\tau,j} \middle| a_k^{\tau,j}\right)\right] - \mathbb{E}_q\left[\log p\left(w_{f,k}^{\tau,j}\right)\right] \\ & + \mathbb{E}_q\left[\log p\left(h_{k,n}^{\phi,j} \middle| b_k^{\phi,j}\right)\right] - \mathbb{E}_q\left[\log p\left(h_{k,n}^{\phi,j}\right)\right] \\ & + \mathbb{E}_q\left[\log p\left(\theta_k^{\tau,\phi} \middle| \alpha_k^{\tau,\phi}, c\right)\right] - \mathbb{E}_q\left[\log p\left(\theta_k^{\tau,\phi}\right)\right] \quad (3.46)\end{aligned}$$

The likelihood term in eqn. (3.46) can be solved as follows

$$\begin{aligned}\mathbb{E}_q\left[\log p\left(p_{fn}^x \middle| w_{f,k}^{\tau,j}, h_{k,n}^{\phi,j}, \theta_k^{\tau,\phi}\right)\right] \geq & -\sum_{f,n}\sum_k p_{fn}^x \left(\varphi_{f,n,k}^{\tau,\phi}\right)^2 \mathbb{E}_q\left[\frac{1}{\sum_{j,k,\tau,\phi} \theta_k^{\tau,\phi} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}}\right] \\ & - \log\left(\omega_{f,n}\right) + 1 - \frac{1}{\omega_{f,n}} \mathbb{E}_q\left[\sum\nolimits_{j,k,\tau,\phi} \theta_k^{\tau,\phi} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}\right] \quad (3.47)\end{aligned}$$

where

$$\varphi_{f,n,k}^{\tau,\phi} \propto \mathbb{E}_q\left[\frac{1}{\sum_j \theta_k^{\tau,\phi} w_{f,k}^{\tau,j} h_{k,n}^{\phi,j}}\right]^{-1} \quad (3.48)$$

and

$$\omega_{f,n} = \mathbb{E}_q\left[\sum\nolimits_{j,k,\tau,\phi} \theta_k^{\tau,\phi} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}\right] \quad (3.49)$$

The rest of eqn. (3.46) can be approximate by the generalized inverse Gaussian distribution (GIG)

$$q\left(w_{f,k}^{\tau,j}\right) = GIG\left(\gamma_{w,f,k}^{\tau,j}, \rho_{w,f,k}^{\tau,j}, \beta_{w,f,k}^{\tau,j}\right) \quad (3.50)$$

$$q\left(h_{k,n}^{\phi,j}\right) = GIG\left(\gamma_{h,k,n}^{\phi,j}, \rho_{h,k,n}^{\phi,j}, \beta_{h,k,n}^{\phi,j}\right) \quad (3.51)$$

$$q\left(\theta_k^{\tau,\phi}\right) = GIG\left(\gamma_{\theta,k}^{\tau,\phi}, \rho_{\theta,k}^{\tau,\phi}, \beta_{\theta,k}^{\tau,\phi}\right) \quad (3.52)$$

where

$$\gamma_{w,f,k}^{\tau,j} = a_k^{\tau,j} \tag{3.53a}$$

$$\rho_{w,f,k}^{\tau,j} = a_k^{\tau,j} + \mathbb{E}_q\left[\theta_k^{\tau,\phi}\right] \sum\nolimits_{n,\phi} \frac{\mathbb{E}_q\left[h_{k,n-\tau}^{\phi,j}\right]}{\omega_{f,n}} \tag{3.53b}$$

$$\beta_{w,f,k}^{\tau,j} = \mathbb{E}_q\left[\frac{1}{\theta_k^{\tau,\phi}}\right] \sum\nolimits_{n,\phi} p_{fn}^{x} \varphi_{f,n,k}^{\tau,\phi^2} \mathbb{E}_q\left[\frac{1}{h_{k,n-\tau}^{\phi,j}}\right] \tag{3.53c}$$

$$\gamma_{h,k,n}^{\phi,j} = b_k^{\phi,j} \tag{3.54a}$$

$$\rho_{h,k,n}^{\phi,j} = b_k^{\phi,j} + \mathbb{E}_q\left[\theta_k^{\tau,\phi}\right] \sum\nolimits_{f,\tau} \frac{\mathbb{E}_q\left[w_{f-\phi,k}^{\tau,j}\right]}{\omega_{f,n}} \tag{3.54b}$$

$$\beta_{h,k,n}^{\phi,j} = \mathbb{E}_q\left[\frac{1}{\theta_k^{\tau,\phi}}\right] \sum\nolimits_{f,\tau} p_{fn}^{x} \varphi_{f,n,k}^{\tau,\phi^2} \mathbb{E}_q\left[\frac{1}{w_{f-\phi,k}^{\tau,j}}\right] \tag{3.54c}$$

$$\gamma_{\theta,k}^{\tau,\phi} = \frac{\alpha_k^{\tau,\phi}}{L + \phi_{max} + \tau_{max}} \tag{3.55a}$$

$$\rho_{\theta,k}^{\tau,\phi} = \alpha_k^{\tau,\phi} c + \frac{\mathbb{E}_q\left[\sum_j w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}\right]}{\omega_{f,n}} \tag{3.55b}$$

$$\beta_{\theta,k}^{\tau,\phi} = p_{fn} \varphi_{k,f,n}^{2} \mathbb{E}_q\left[\frac{1}{\sum_j w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}}\right] \tag{3.55c}$$

Finally, expectation over $\mathrm{q}(\Omega)$ can be computed by

$$\mathbb{E}_q[y] = \frac{\mathcal{K}_{\gamma+1}\left(2\sqrt{\rho\beta}\right)\sqrt{\beta}}{\mathcal{K}_{\gamma}\left(2\sqrt{\rho\beta}\right)\sqrt{\rho}} \tag{3.56}$$

$$\mathbb{E}_q\left[\frac{1}{y}\right] = \frac{\mathcal{K}_{\gamma-1}\left(2\sqrt{\rho\beta}\right)\sqrt{\rho}}{\mathcal{K}_{\gamma}\left(2\sqrt{\rho\beta}\right)\sqrt{\beta}} \tag{3.57}$$

3.4.2 Initialization

The initialization is an essential part for the separation since the NMF2D and its variants are very sensitive to the initialization. In this work, the Gamma-Exponential process will be proposed to initialization the spectral and temporal tensors of the NMF2D as follows

$$w_{f,k}^{\tau,j(initial)} = \frac{\sqrt{\beta_{w,f,k}^{\tau,j}/\rho_{w,f,k}^{\tau,j}}\,\mathcal{K}_{\gamma_{w,f,k}^{\tau,j}+1}\left(2\sqrt{\rho_{w,f,k}^{\tau,j}\beta_{w,f,k}^{\tau,j}}\right)}{\mathcal{K}_{\gamma_{w,f,k}^{\tau,j}}\left(2\sqrt{\rho_{w,f,k}^{\tau,j}\beta_{w,f,k}^{\tau,j}}\right)} \tag{3.58a}$$

$$h_{k,n}^{\phi,j(initial)} = \frac{\sqrt{\beta_{h,k,n}^{\phi,j}/\rho_{h,k,n}^{\phi,j}}\,\mathcal{K}_{\gamma_{h,k,n}^{\phi,j}+1}\left(2\sqrt{\rho_{h,k,n}^{\phi,j}\beta_{h,k,n}^{\phi,j}}\right)}{\mathcal{K}_{\gamma_{h,k,n}^{\phi,j}}\left(2\sqrt{\rho_{h,k,n}^{\phi,j}\beta_{h,k,n}^{\phi,j}}\right)} \tag{3.58b}$$

for the convolutive parameters and number of components that obtained from the Gamma-Exponential process.

Table 3.1 summarizes the main steps of the proposed algorithm.

Table 3.1: Proposed algorithm

1. Estimate the number of components and convolutive parameters by using the proposed Gamma-Exponential process in eqns. (3.53)-(3.55) and compute $\mathbb{E}_q\left[\theta_k^{\tau,\phi}\right]$.
2. Initialize $w_{f,k}^{\tau,j}$ and $h_{k,n}^{\phi,j}$ with the proposed Gamma-Exponential process spectral and temporal tensors in eqn. (3.58a) and eqn. (3.58b), and initialize $\lambda_{k,n}^{\phi,j}$ with positive value.

3. E-step: compute $\hat{p}_{jfn}$ eqn. (3.12).

4. M-step: compute $A_f, \Sigma_{b,f}, w_{f,k}^{\tau,j}, h_{k,n}^{\phi,j}$ and $\lambda_{k,n}^{\phi,j}$ using eqn. (3.19), eqn. (3.21), eqn. (3.36), eqn. (3.37), and eqn. (3.39).

5. Normalize $w_{f,k}^{\tau,j} = w_{f,k}^{\tau,j} / \sqrt{\sum_{f,k,\tau} \left(w_{f,k}^{\tau,j}\right)^2}$

6. Repeat E- and M-steps, and the normalization until convergence is achieved where rate of cost change is below a prescribed threshold, ψ.

7. Take inverse STFT with dual synthetic window to $\hat{\boldsymbol{s}}_{fn}$.

3.5 Window Length

The power spectrogram of the pitched and percussive instruments has different characteristics. Pitched instruments are smooth and continue in temporal direction and discrete in spectral direction, and the opposite for the percussive instruments. Therefore, short and long windows will be used for the percussive and pitched instruments, respectively, in order to match their spectral-temporal characteristics. The impending challenge is in the singing voice which acts like a pitched instrument but with more fluctuations. Therefore it is difficult to separate the singing voice when it accompanied with pitched instrument since they share similar characteristics. As the singing voice acts like percussive instrument in long window (and as pitched instrument in short window), then the advantage of this characteristic will be considered, where a long window will be used when the singing voice accompanied with pitched instruments, in order to distinguish between them.

3.6 Results and Discussions

The proposed algorithm will be compared with the standalone EM and MU based algorithms [8], GEM–MU based NTF [80] with adaptive sparsity and proposed initialization, and the GEM-MU based NMF (by setting the convolutive parameters of the proposed algorithm to zero $\tau_{max} = 0$ and $\phi_{max} = 0$) with adaptive sparsity and proposed initialization.

3.6.1 Effects of Sparsity

First of all, the effect of the sparsity on the separation performance will be investigated by comparing between the uniform sparsity and the adaptive sparsity. An experiment has been ran for different values of the uniform sparsity and for the adaptive sparsity, for three sources that convolutively mixed in stereo mixture that has 1m space between its microphones, 130 ms reverberation time, and with 16 kHz sampling frequency. The following parameters were set for the proposed algorithm; $K_j = 5$ components per source, $\tau = \{0, 1, 2, 3, 4\}$, and $\phi = \{0, 1\}$. Furthermore, in order to focus on the sparsity effects only, an oracle initialization (where the input parameters are known) has been used. Figure 3.1 shows the average signal-to-distortion ratio (SDR) [98] w.r.t different values of sparsity. The SDR shows a total separation performance that includes a degree of separation and absence of nonlinear distortion. It is clear from Figure 3.1 that the adaptive sparsity gives the highest SDR as it has a specific sparsity value for each element of $\boldsymbol{H}$ instead of constant value for the entire elements of $\boldsymbol{H}$ as in the uniform sparsity. Furthermore, the spectrogram of one of the estimated source for adaptive sparsity, over uniform sparsity, and the under uniform sparsity is shown in Figure 3.2. It is clear from Figure 3.2 that the over sparsity eliminates many spectra from the estimated source, as it assigned a lot of zero values in the $\boldsymbol{H}$ matrix. While the under uniform sparsity has many unwanted spectra, as there are many of unwanted elements in the $\boldsymbol{H}$ matrix. While, the adaptive sparsity address them correctly, as it specified a specific value for each element of the H matrix, as in eqn. (3.39).

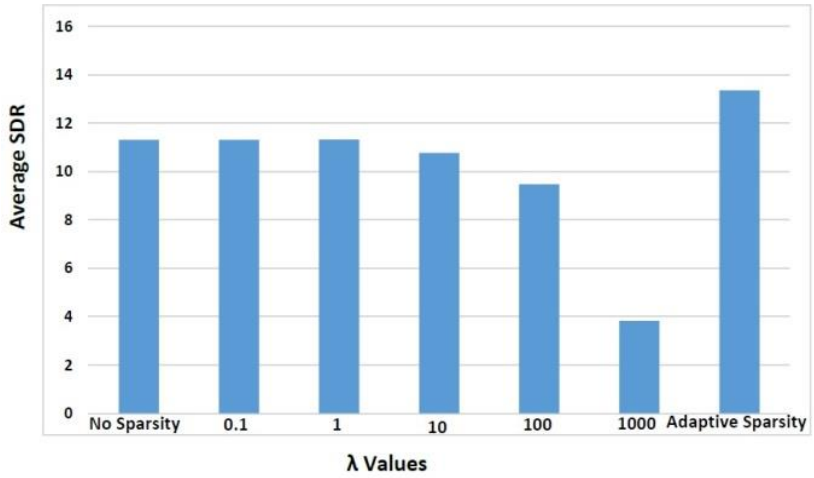


Figure 3.1: Average SDR w.r.t different sparsity values.

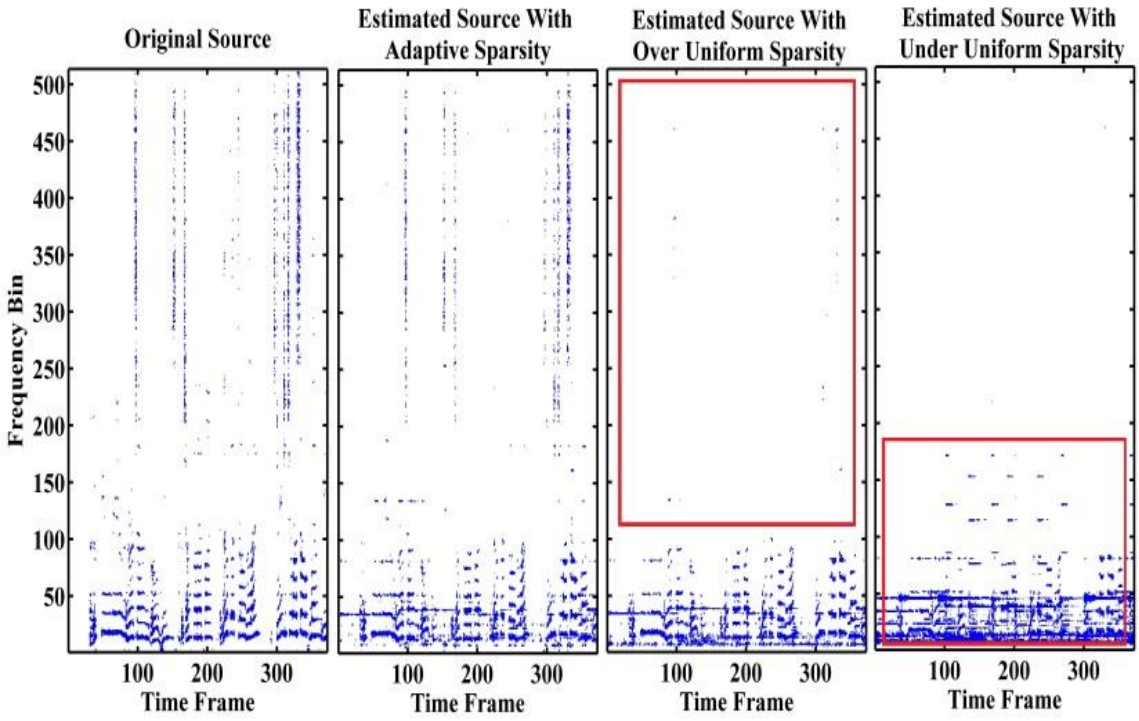


Figure 3.2: Effects of sparsity on the estimated source.

3.6.2 Evaluation

To evaluate the proposed algorithm the performance will be measured by using the SDR [98] which measures an overall sound quality of the source separation where it combines the signal-to-interference ratio (SIR), source image-to-spatial distortion ratio (ISR), and the signal-to-artifact ratio (SAR) into one measurement. MATLAB codes for this evaluation procedure can be found in [99].

3.6.3 Datasets

As our results will be compared with the MU and EM algorithms of [8], then the same datasets of this paper which match with the dataset dev2 of SiSEC'08 "underdetermined speech and music mixtures" will be considered, as follows

1. **Synthetic Convolutive Dataset:** This dataset consist of two groups. The wdrum group which consists of three percussive instruments and the ndrum group which consists of three non-percussive instruments.

2. **Live Recording (Convolutive) Dataset:** This dataset is more complicated than the Synthetic convolutive case as it contains different musical instruments with vocal signal. It consists of two groups the wdrum group which consists of vocal and musical instrument with drum, and the ndrum group which consists of vocal and musical instruments without drum.

All the mixtures were 10s long, and sampled at 16 kHz. Also, they have 130 ms of reverberation time with 1m space between their microphones. Different windows length will be used in the STFT with 50% overlaps. The STFT MATLAB code is available from [99].

3.6.4 Results of the Synthetic Convolutive Dataset:

1. **wdrum Case:** As all the musical instruments are percussive that have short temporal then the STFT with window length of 512-sample was selected. Firstly the effect of the proposed Gamma-Exponential process in estimating the number of components and the convolutive parameters will be investigated. The bounds of the proposed Gamma-Exponential process set as

follows: $\tau = \{0, 1, 2, \ldots, 10\}$, $\phi = \{0,1,2,\ldots,10\}$, and $K = 20$. The results of the proposed Gamma-Exponential process are shown in Figures 3.3 and 3.4. The number of active components in the NMF2D as estimated according to the hidden latent variable in eqn. (3.40) is given by

$$\mathbb{E}_q[\theta_k] = \frac{1}{(\tau_{max}+1)(\phi_{max}+1)} \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} \mathbb{E}_q\left[\theta_k^{\tau,\phi}\right]$$

where

$$\mathbb{E}_q\left[\theta_k^{\tau,\phi}\right] = \frac{\sqrt{\beta_{\theta,k}^{\tau,\phi}/\rho_{\theta,k}^{\tau,\phi}}\,\mathcal{K}_{\gamma_{\theta,k}^{\tau,\phi}+1}\left(2\sqrt{\rho_{\theta,k}^{\tau,\phi}\beta_{\theta,k}^{\tau,\phi}}\right)}{\mathcal{K}_{\gamma_{\theta,k}^{\tau,\phi}}\left(2\sqrt{\rho_{\theta,k}^{\tau,\phi}\beta_{\theta,k}^{\tau,\phi}}\right)}$$

In above, a uniform distribution for both $q(\tau)$ and $q(\phi)$ is assumed. The active component can be defined as

$$k_* = \underset{k}{arg}\left\{\frac{\mathbb{E}_q[\theta_k]}{\sum_{k=1}^{K}\mathbb{E}_q[\theta_k]} \geq \varepsilon\right\}$$

where ε is a small constant which can be set as 0.1. $\mathbb{E}_q[\theta_k]$ is treated as a histogram and the active components are selected as those that exceeds 10% of the overall sum. Figure 3.3 shows the values of $\mathbb{E}_q[\theta_k]$ for $k = 1, \ldots, 20$ which are predominantly zero except for $k = 3, 8, 11$ and 20 whose $\mathbb{E}_q[\theta_k]$ values are 1.46, 0.07, 2.1 and 3.23, respectively. The term $\sum_{k=1}^{K}\mathbb{E}_q[\theta_k]$ has been calculated to be 6.86 and thus, the active components are only $k_* = 3, 11$ and 20. Let $K_* = \#\, k_*$, that is, the number of active components e.g. in Figure 3.3 this corresponds to $K_* = 3$. Since there are $J = 3$ sources, then $K_j = K_*/J = 1$ for $j = 1, 2, 3$. In addition, for each k_* active component, the distribution for (τ, ϕ) has been determined which is given by $\mathbb{E}_q\left[\theta_{k=k_*}^{\tau,\phi}\right]$. These are shown in Figure 3.4. The optimum model for (τ, ϕ) is selected by treating each $\mathbb{E}_q\left[\theta_{k=K}^{\tau,\phi}\right]$ for various values

of (τ, ϕ) as a histogram. Thus the optimum model for (τ, ϕ) is given by the average of non-zero components:

$$\hat{\tau}_{max,k_*} = \frac{\sum_{l=0}^{\phi_{max}} F_l^{(\tau)}}{\#\left(F_l^{(\tau)} \neq 0, \forall l\right)} - 1$$

$$\hat{\phi}_{max,k_*} = \frac{\sum_{l=0}^{\tau_{max}} F_l^{(\phi)}}{\#\left(F_l^{(\phi)} \neq 0, \forall l\right)} - 1$$

where

$$F_l^{(\tau)} = \#component\ in\ \frac{\mathbb{E}_q\left[\theta_{k=k_*}^{\tau,\phi=l}\right]}{\sum_\tau \mathbb{E}_q\left[\theta_{k=k_*}^{\tau,\phi=l}\right]} \geq \varepsilon$$

$$F_l^{(\phi)} = \#component\ in\ \frac{\mathbb{E}_q\left[\theta_{k=k_*}^{\tau=l,\phi}\right]}{\sum_\phi \mathbb{E}_q\left[\theta_{k=k_*}^{\tau=l,\phi}\right]} \geq \varepsilon$$

The term $F_l^{(\tau)}$ counts the number of τ components in the normalized $\mathbb{E}_q\left[\theta_k^{\tau,\phi=l}\right]$ that exceeds ε, and $\#\left(F_l^{(\tau)} \neq 0, \forall l\right)$ counts the number of entries in $F_l^{(\tau)}$ that is non-zero. The same interpretation is applied to $F_l^{(\phi)}$ and $\#\left(F_l^{(\phi)} \neq 0, \forall l\right)$ for determining the model order ϕ_{max}. From Figure 3.4, it can be calculated that $\hat{\tau}_{max,k_*} = 4$ and $\hat{\phi}_{max,k_*} = 10$ for all k_*, then $\hat{\tau}_{max} = \frac{\sum_{k_*} \hat{\tau}_{max,k_*}}{K_*} = 4$, and $\hat{\phi}_{max}\ \frac{\sum_{k_*} \hat{\phi}_{max,k_*}}{K_*} = 10$. Thus, the optimum model order for the NMF2D model in eqn. (3.4) is given by $K_j = 1, \hat{\tau}_{max} = 4$ and $\hat{\phi}_{max} = 10$.

For the current values of the convolutive parameters ($\tau_{max} = 4$ and $\phi_{max} = 10$) the tensors of proposed Gamma-Exponential process eqn. (3.58a) and eqn. (3.58b) will be used to initialize the proposed GEM-MU based NMF2D algorithm, and its SDRs are tabulated in Table 3.2. It can be seen from Table 3.2 that the SDRs of the proposed GEM-MU based NMF2D is better than all other algorithms. Thus by using the proposed Gamma-Exponential process, the number of components and convolutive parameters can be estimated, and the proposed algorithm can be initialized.

Furthermore, despite it is not straight forward to compare the proposed Gamma-exponential process with other methods as it is for the best of our knowledge is the first method to estimate the convolutive parameters in the NMF2D. However we proposed to compare with the mesh method that compute the SDR for each single selection of the convolutive parameter (for $\tau = \{0, 1, \ldots, 10\}$

and $\phi = \{0,1,\ldots,20\}$) and check the convolutive parameters that give the highest SDR. This method is time consuming and unrealistic as it required the original sources to compute their SDRs. We applied it on the above case of synthetic convolutive with drum, as shown in Figure 3.5. The figure shows the results of the mesh method of running the NMF2D algorithm for every possible case of τ and ϕ. In total, there are $11 \times 21 = 231$ possible model order. The highest SDR is obtained at SDR = 4.08 dB with $\tau_{max} = 9$ and $\phi_{max} = 10$. There is 0.06 dB difference between the SDR of the Mesh method and the SDR of the Gamma-Exponential process, which is acceptable difference in comparison with the time required to find the model order using the mesh method.

Finally, the cost function versus iteration number is plotted in Figure 3.6 (a large constant value has been added to the curve to ensure positivity). Figure 3.6 shows that the cost function has been converged. Finally the waveforms of the estimated sources are shown in Figure 3.7.

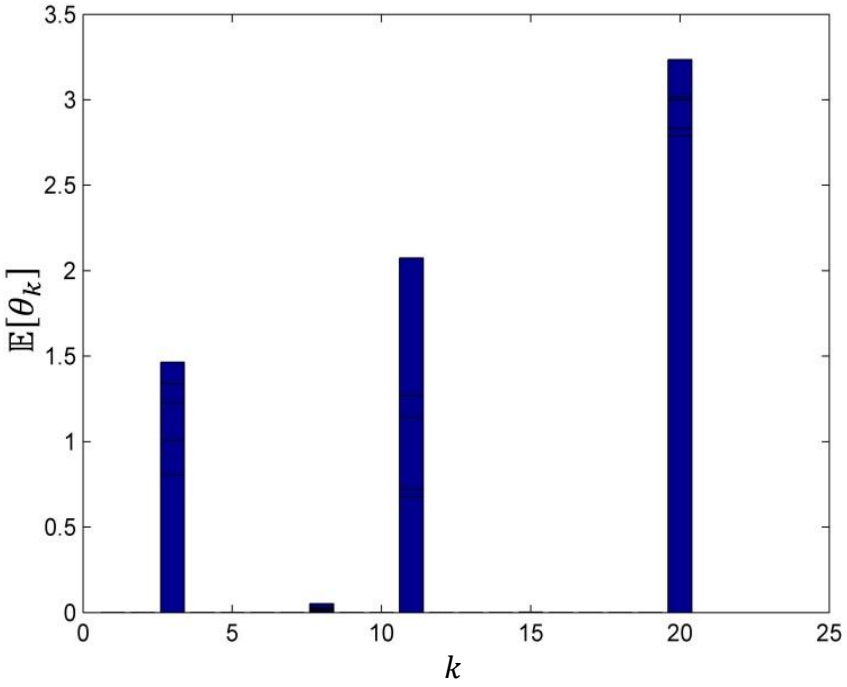


Figure 3.3: Number of components by using Ga-Exp.

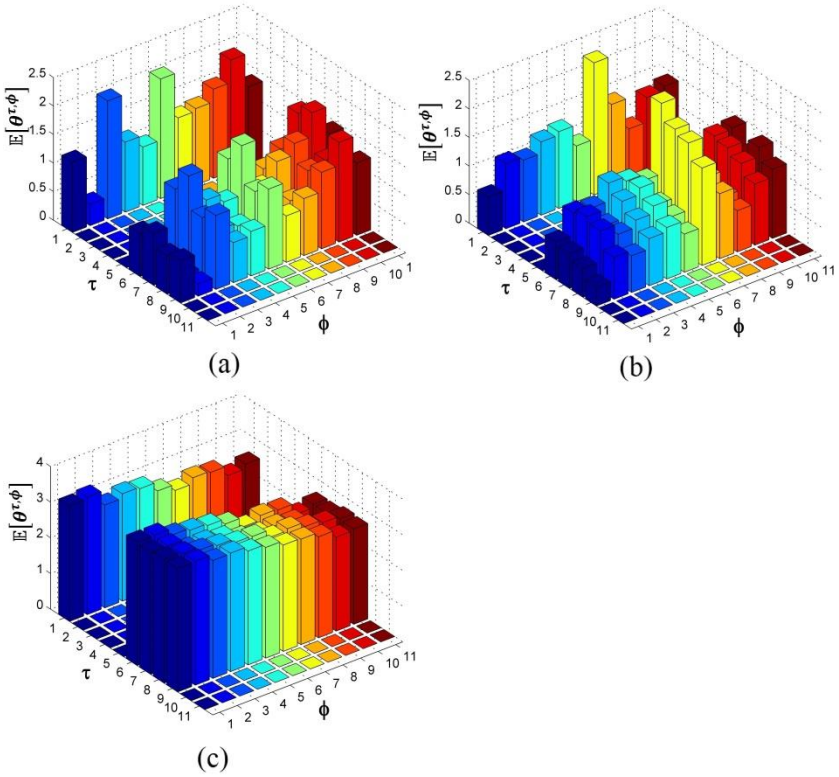


Figure 3.4: Convolutive parameters corresponding to each component by using Ga-Exp.

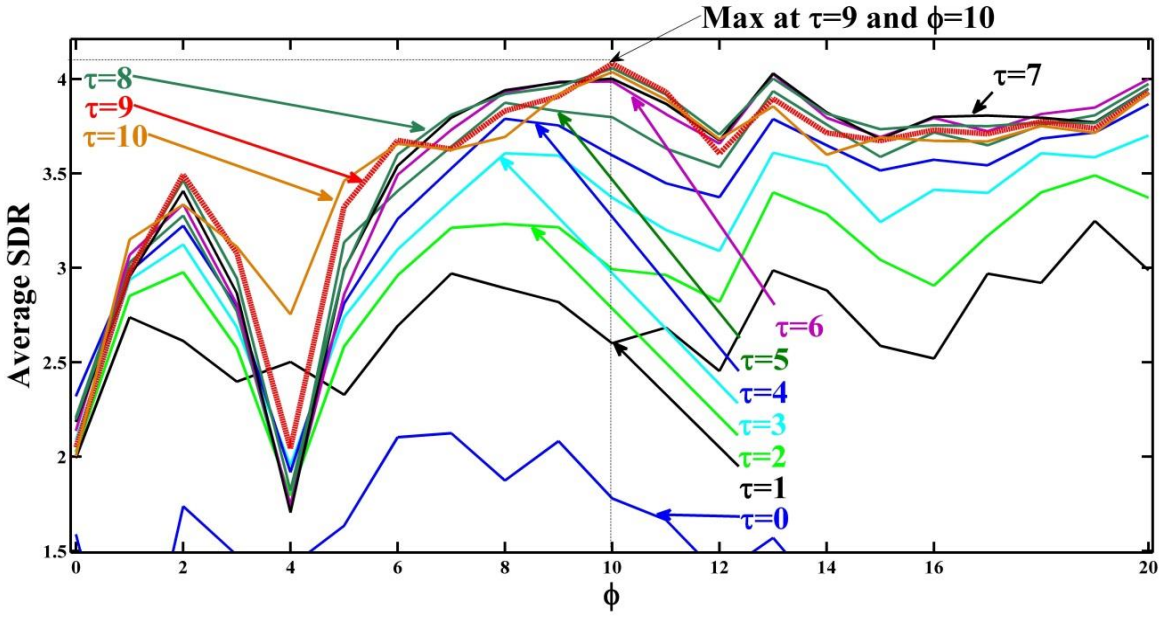


Figure 3.5: Average SDR w.r.t the convolutive parameters.

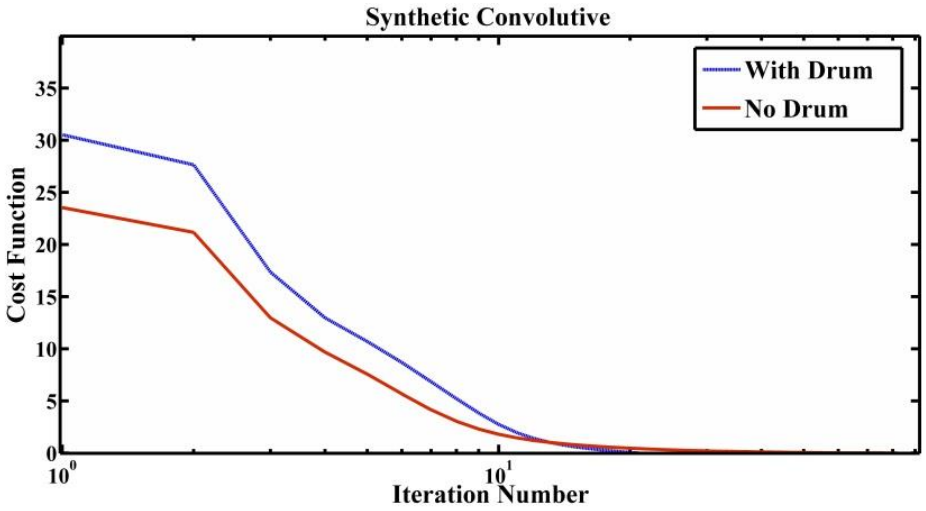


Figure 3.6: Convergence of the cost functions.

Table 3.2: Convolutive mixture with drum (wdrurm).

Algorithm	**Parameters**	**SDRs**			**Avrg SDR**
		S_1	**S_2**	**S_3**	
EM NMF	Window=512	6.89	-4.83	1.75	1.27
MU NMF	Window=512	5.10	-9.87	2.46	-0.77
GEM-MU NTF	Window=512	6.18	-1.32	3.00	2.62
GEM-MU NMF	Window=512 $K_j = 1$	5.54	-0.28	1.21	2.16
Proposed GEM-MU NMF2D With Mesh Method	Window=512 $\hat{\tau}_{max} = 9$ $\hat{\phi}_{max} = 10$ $K_j = 1$	**8.42**	-0.46	**4.27**	**4.08**
Proposed GEM-MU NMF2D With Ga-Exp	Window=512 $\hat{\tau}_{max} = 4$ $\hat{\phi}_{max} = 10$ $K_j = 1$	7.99	**0.22**	3.86	4.02

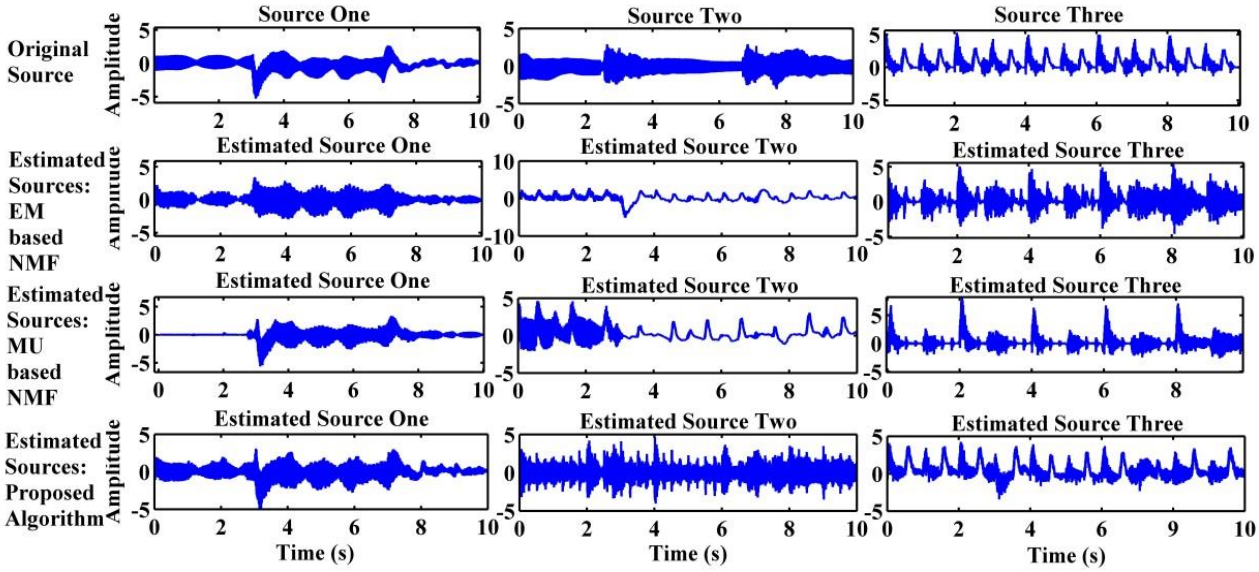


Figure 3.7: Waveforms of the estimated sources for drum case.

2. **ndrum Case:** Since all the musical instruments were pitched (non-percussive) and had long temporal characteristics then the STFT with window length of 2048-sample was selected. By following the same procedure of the wdrum case, the number of components and convolutive parameters are selected from Figure 3.8 and Figure 3.9, respectively. From Figure 3.8, it is calculated that $K_* = 5$ and since there are 3 sources, one may consider partitioning this into $K_j = 2$. Also from Figure 3.9, the convolutive model order are determined as follows $\hat{\tau}_{max} = 5$, and $\hat{\phi}_{max} = 10$. For the mesh method the highest SDR (which is equal to 3.41 dB) is obtained from $\tau_{max} = 8$ and $\phi_{max} = 9$ as shown in Figure 3.10. The cost function and the waveforms of the estimated sources are shown in Figure 3.6 and Figure 3.11, respectively. Furthermore, all the results are tabulated in Table 3.3. It can be seen from this table that the average SDRs of the proposed algorithm with window 2048-sample are better than the rest the algorithms. Also, it can be seen that there is 0.1 dB difference between the SDR of the Mesh method and the SDR of the Gamma-Exponential process, again the 0.1 dB is an acceptable difference.

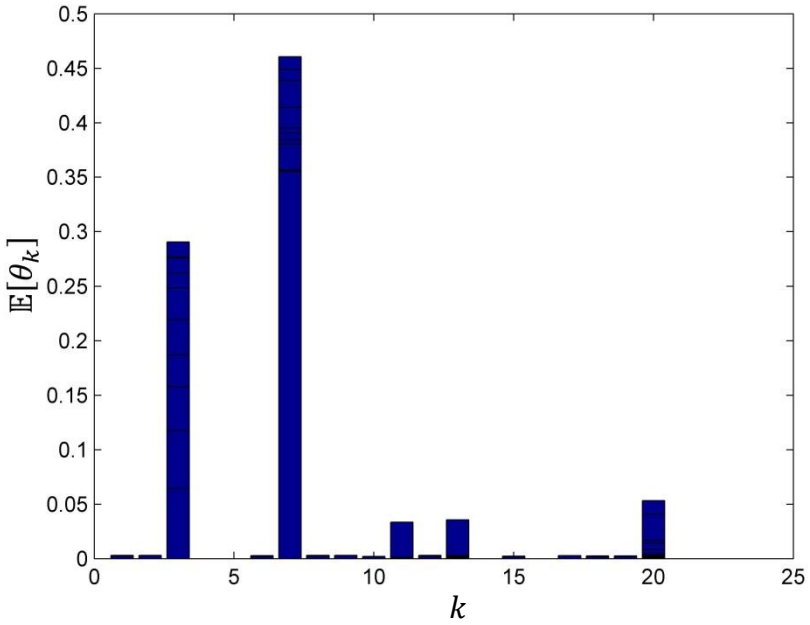


Figure 3.8: Number of components by using Ga-Exp.

Table 3.3: Synthetic convolutive without drum (ndrum).

Algorithm	**Parameters**	**SDRs**			**Avrg SDR**
		S_1	**S_2**	**S_3**	
EM NMF	Window=2048	**4.18**	1.02	-1.8	1.10
MU NMF	Window=2048	2.89	1.04	-2.09	0.61
GEM-MU NTF	Window=2048	2.93	3.09	1.57	2.53
GEM-MU NMF	Window=2048 $K_j = 2$	2.98	2.57	1.15	2.23
Proposed GEM-MU NMF2D With Mesh Method	Window=2048 $\hat{\tau}_{max} = 8$ $\hat{\phi}_{max} = 9$ $K_j = 2$	1.63	**3.39**	**5.21**	**3.41**
Proposed GEM-MU NMF2D With Ga-Exp	Window=2048 $\hat{\tau}_{max} = 5$ $\hat{\phi}_{max} = 10$ $K_j = 2$	1.85	3.33	4.75	3.31

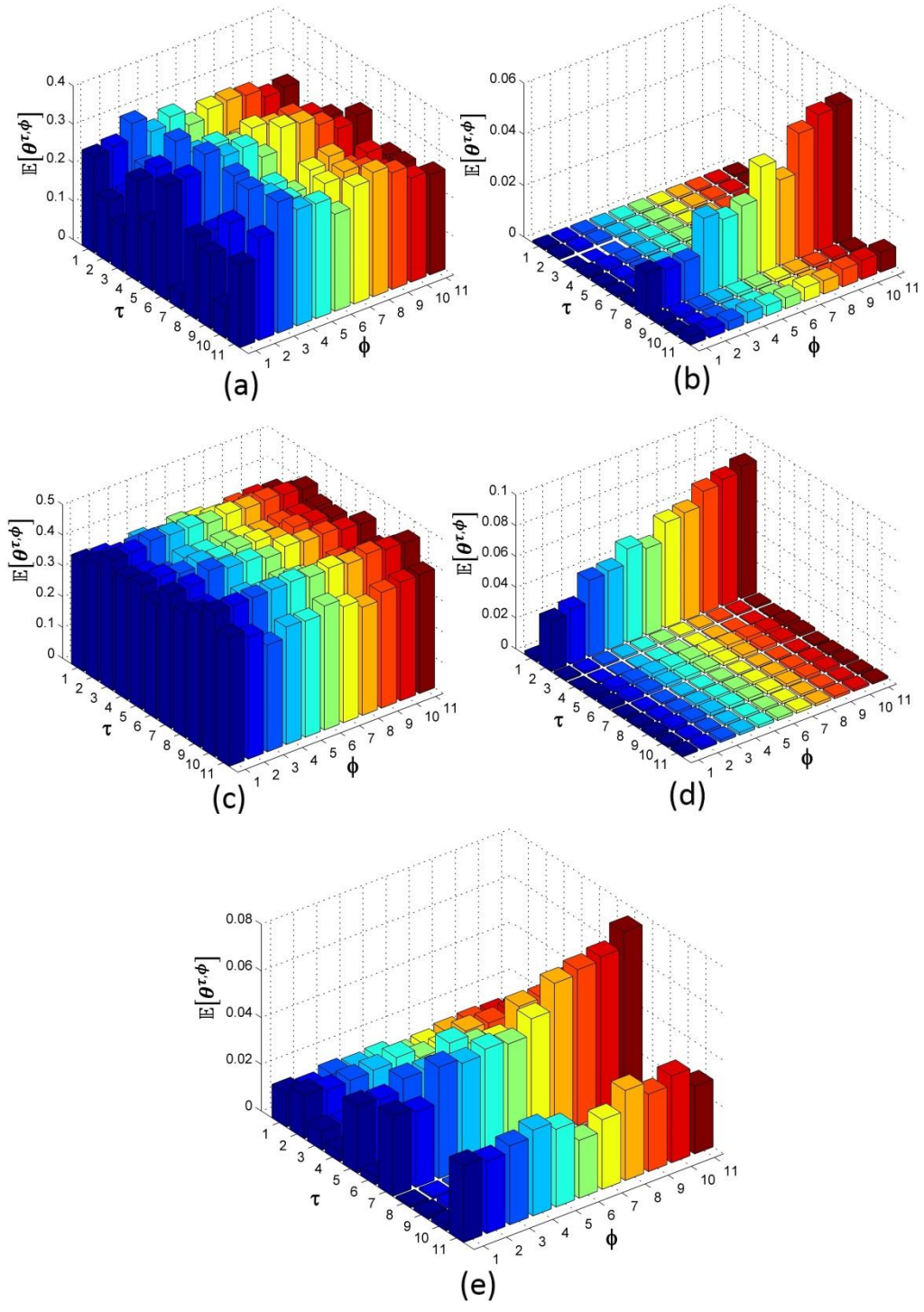


Figure 3.9: Convolutive parameters corresponding to each component by using Ga-Exp.

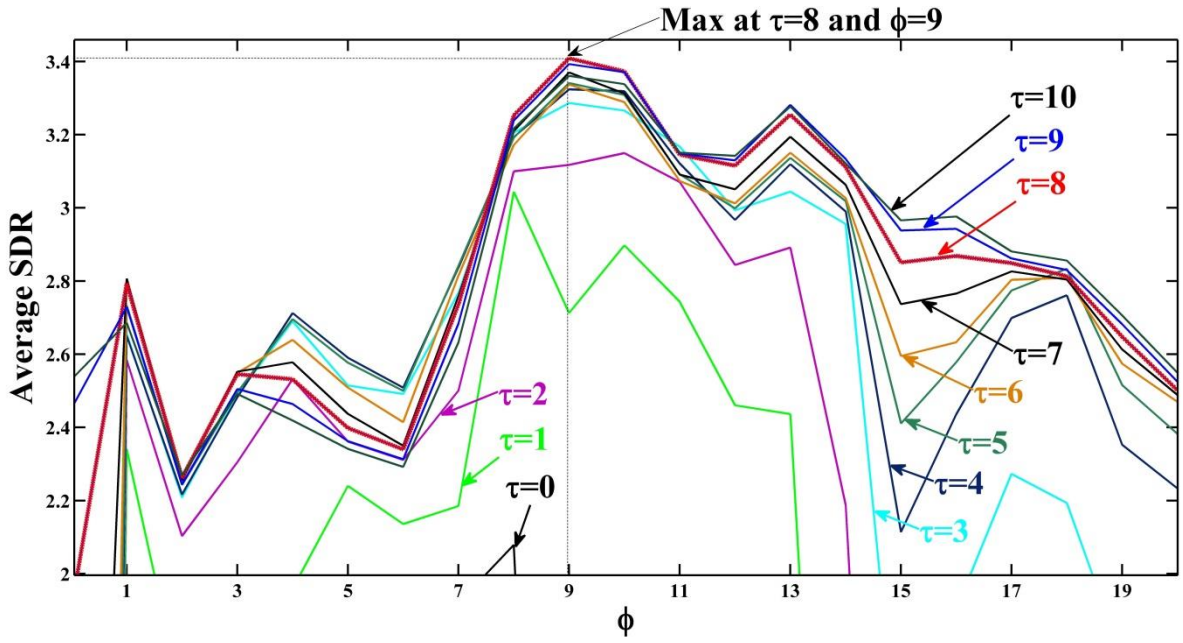


Figure 3.10: Average SDR w.r.t the convolutive parameters.

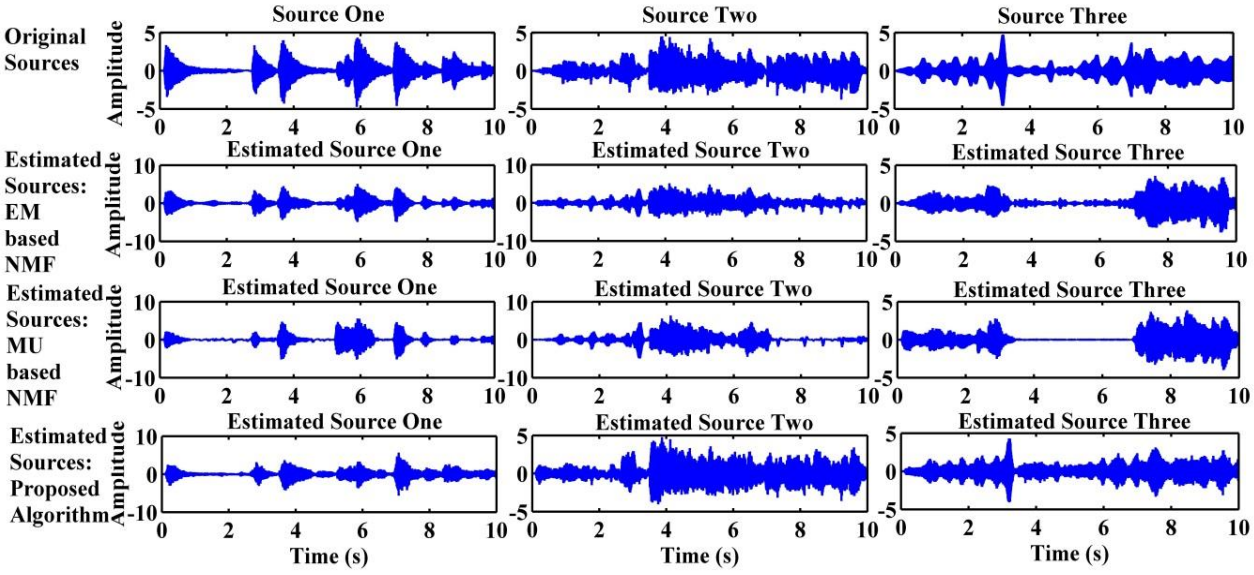


Figure 3.11: Waveforms of the estimated sources for no drum case.

3.6.5 Results of the Live Recording (Convolutive) Dataset:

1. wdrum Case: By following the same procedure of the previous sections, window length of 2048-sample was selected for the STFT, the number of components and convolutive parameters were selected from Figure 3.12 and Figure 3.13, respectively, where it is clear from these figures that $K_j = 3$, $\hat{\tau}_{max} = 1$, and $\hat{\phi}_{max} = 3$. For the mesh method the highest SDR (which is equal to 7.96 dB) is obtained from $\tau_{max} = 1$ and $\phi_{max} = 1$ as shown in Figure 3.14. Figure 3.15 shows the convergence of the cost function w.r.t the iteration number. Additionally, all the results are tabulated in Table 3.4. It is clear from Table 3.4, that the SDRs of the proposed algorithm are the best. Also, it can be seen that there is 0.18 dB difference between the SDR of the Mesh method and the SDR of the Gamma-Exponential process, which is an acceptable difference. Finally the waveforms of the estimated sources are shown in Figure 3.16.

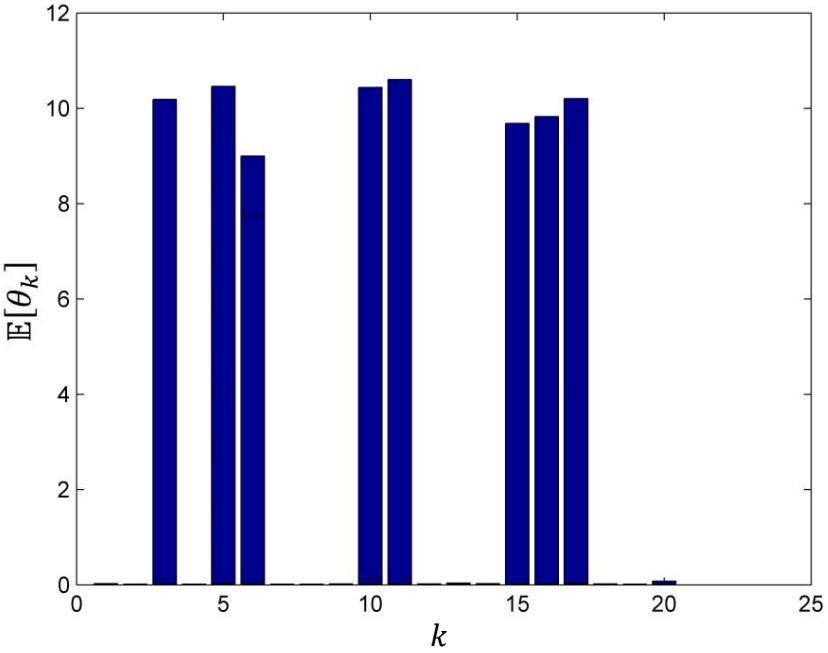


Figure 3.12: Number of components by using Ga-Exp.

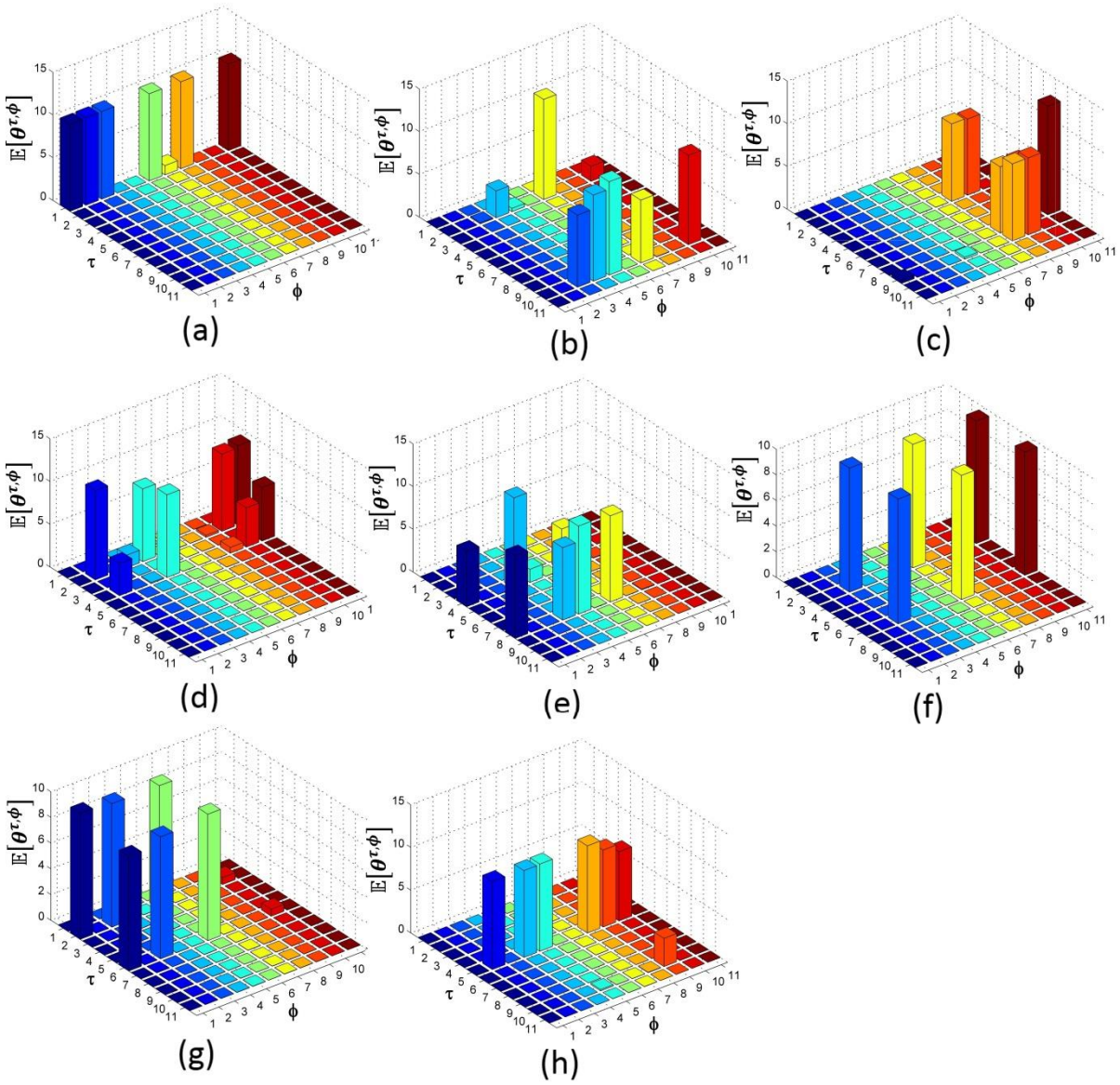


Figure 3.13: Convolutive parameters corresponding to each component by using Ga-Exp.

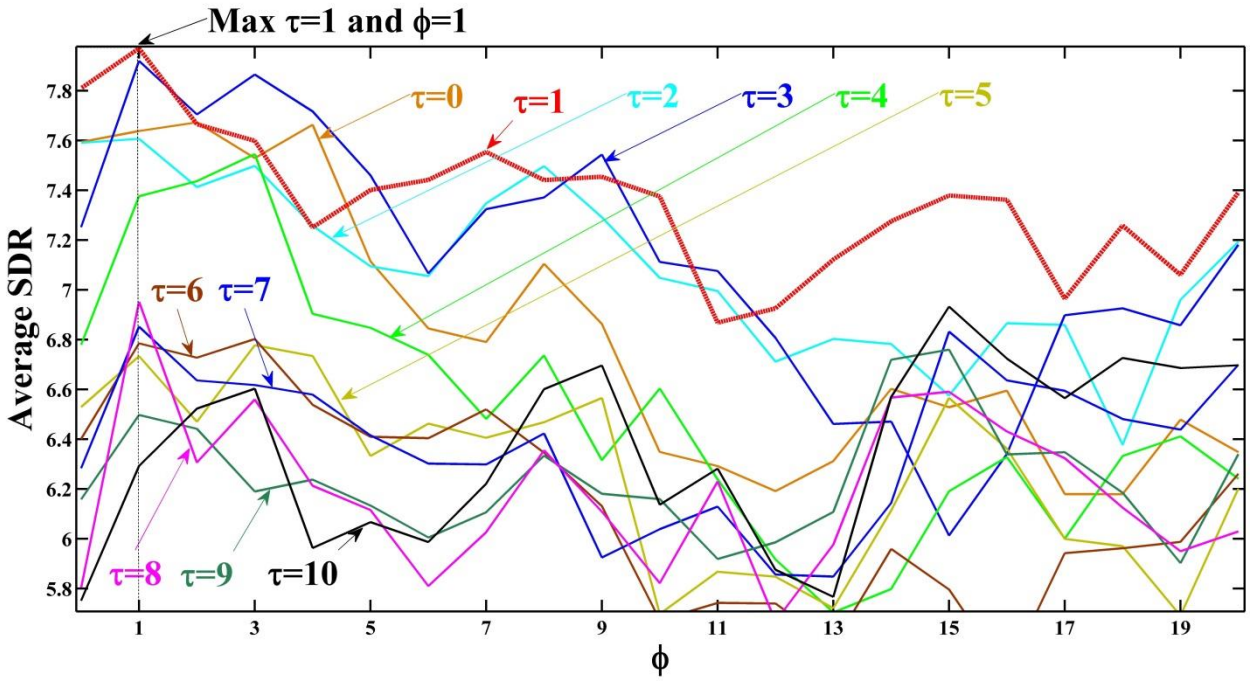


Figure 3.14: Average SDR w.r.t the convolutive parameters.

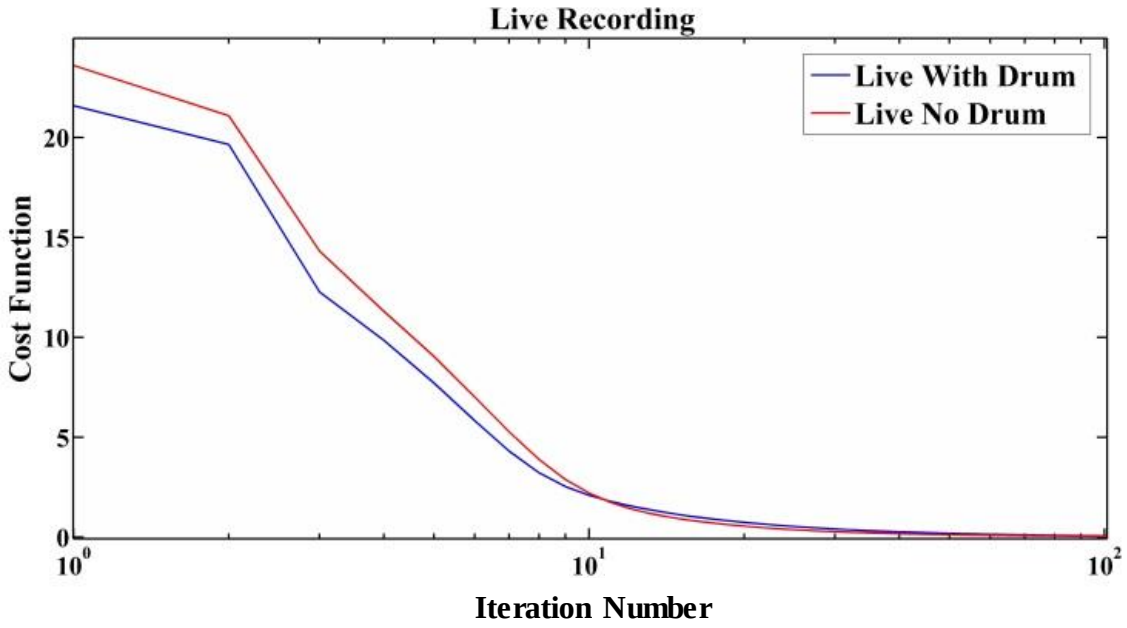


Figure 3.15: Convergence of cost functions.

Table 3.4: Live recording with drum (wdrum).

Algorithm	Parameters	SDRs			Avrg SDR
		S_1	S_2	S_3	
EM NMF	Window=2048	4.96	5.55	8.03	6.18
MU NMF	Window=2048	4.19	4.50	7.58	5.42
GEM-MU NTF	Window=2048	5.89	7.90	7.68	7.16
GEM-MU NMF	Window=2048, $K_j = 3$	5.99	7.74	7.58	7.10
Proposed GEM-MU NMF2D With Mesh Method	Window=2048 $\hat{\tau}_{max} = 1$ $\hat{\phi}_{max} = 1$ $K_j = 3$	**6.77**	8.65	**8.47**	**7.96**
Proposed GEM-MU NMF2D With Ga-Exp	Window=2048, $\hat{\tau}_{max} = 1$, $\hat{\phi}_{max} = 3$, $K_j = 3$	6.58	8.65	8.12	7.78

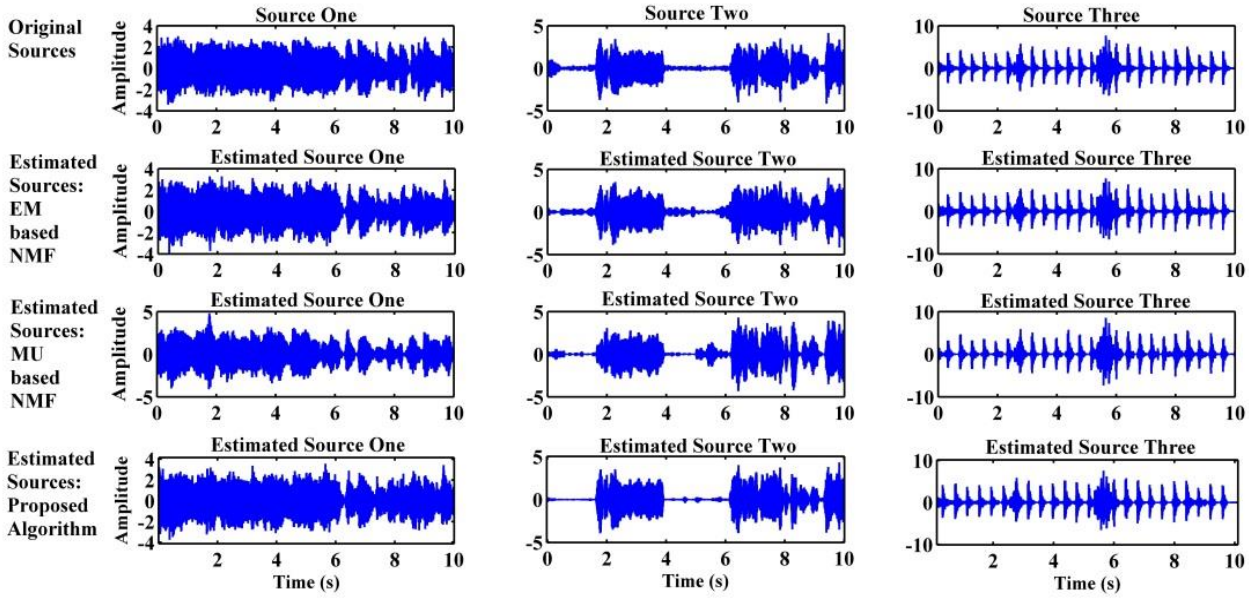


Figure 3.16: Waveforms of the estimated sources for the live recording with drum case.

2. **ndrum Case:** Since this dataset contains pitched musical instruments and vocal, and as the vocal sound acts like percussive instrument in long window, then a long window of 4096-sample is selected for the STFT. The number of components and convolutive parameters were selected from Figure 3.17 and Figure 3.18, respectively, where it is clear from these Figures that $K_j = 5$, $\hat{\tau}_{max} = 1$ and $\hat{\phi}_{max} = 7$. For the mesh method the highest SDR (which is equal to 5.16 dB) is obtained from $\tau_{max} = 2$ and $\phi_{max} = 9$ as shown in Figure 3.19. The cost function with respect to the iteration number is shown in Figure 3.15. All the result has been tabulated in Table 3.5. Also, it can be seen that there is 0.55 dB difference between the SDR of the Mesh method and the SDR of the Gamma-Exponential process. Finally, the waveforms of the estimated sources are shown in Figure 3.20.

It can be seen from Tables 3.2 to 3.5 that the SDR of the proposed algorithm based on Gamma-Exponential process on average is 0.22 dB less than the SDR of the proposed algorithm that based on the mesh method. The 0.22 dB is acceptable difference in comparison with the time required to find the mesh method convolutive parameters.

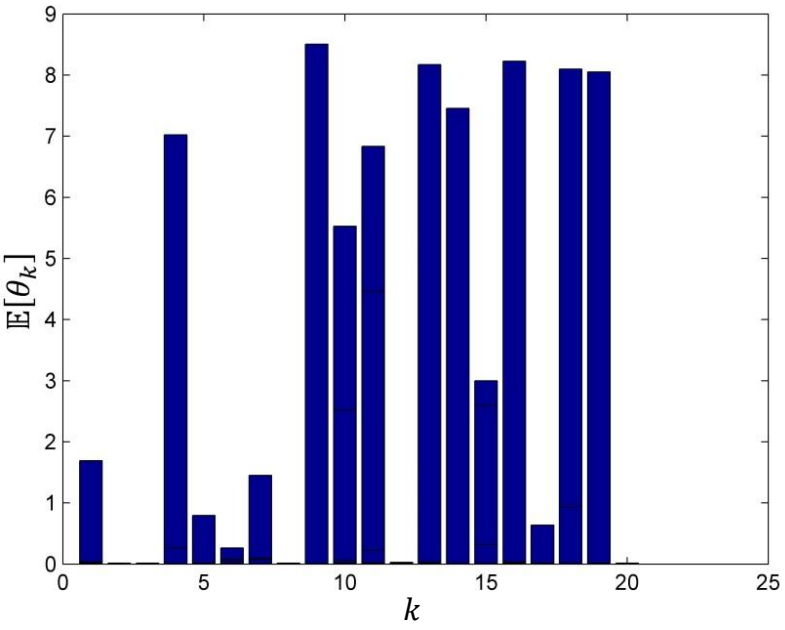


Figure 3.17: Number of components by using Ga-Exp.

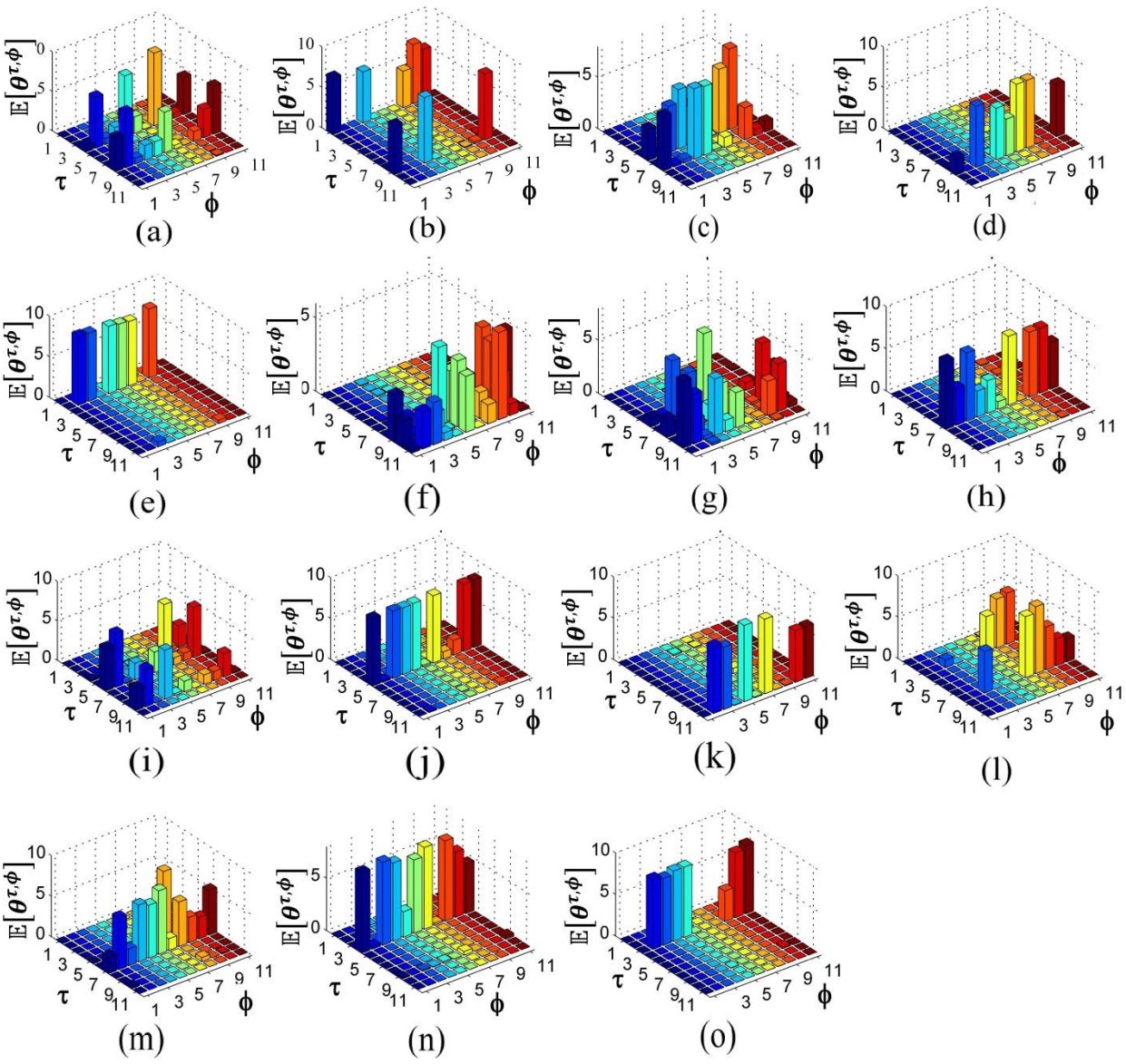


Figure 3.18: Convolutive parameters corresponding to each component by using Ga-Exp.

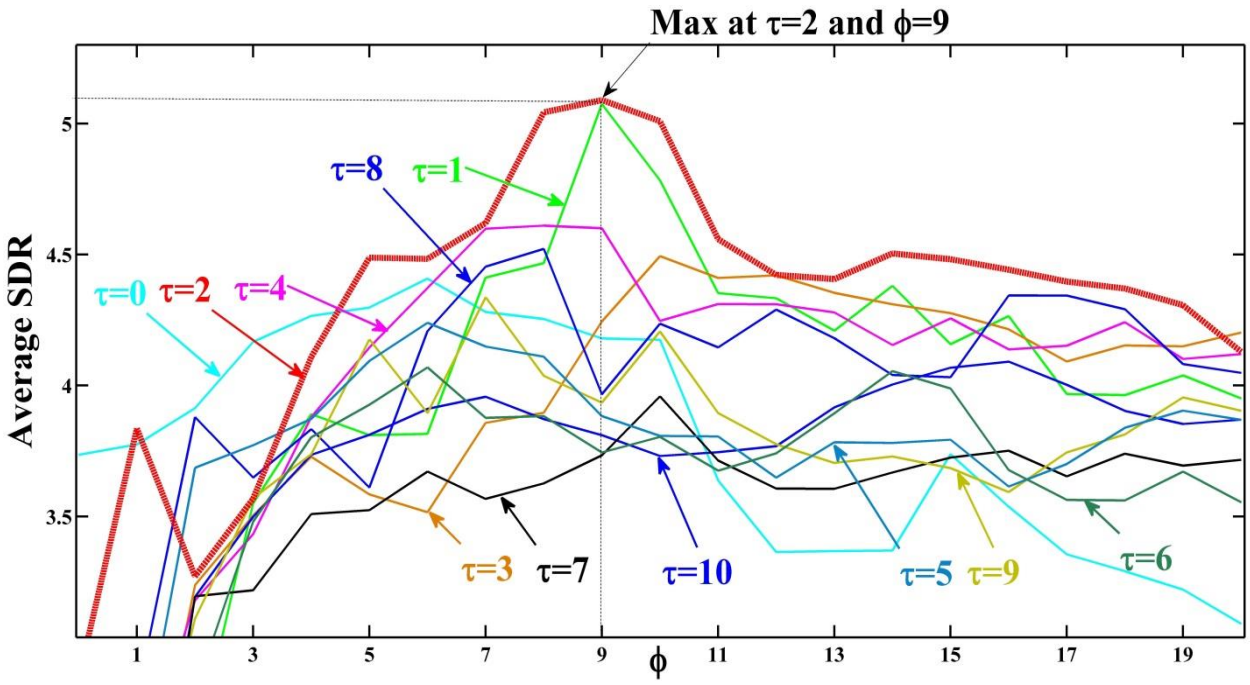


Figure 3.19: Average SDR w.r.t the convolutive parameters.

Table 3.5: Live recording without drum (ndrum).

Algorithm	**Parameters**	**SDRs**			**Avrg SDR**
		S_1	**S_2**	**S_3**	
EM NMF	Window=4096	6.02	1.68	-0.91	2.26
MU NMF	Window=4096	4.27	0.05	-3.14	0.39
GEM-MU NTF	Window=4096	7.71	3.60	-0.40	3.64
GEM-MU NMF	Window=4096, $K_j = 5$	6.80	2.10	-0.24	2.89
Proposed GEM-MU NMF2D With Mesh Method	Window=4096 $\hat{\tau}_{max} = 2$ $\hat{\phi}_{max} = 9$ $K_j = 5$	**9.28**	**5.75**	**0.44**	**5.16**
GEM-MU NMF2D with Proposed Ga-Exp	Window=4096, $\hat{\tau}_{max} = 1$, $\hat{\phi}_{max} = 7$, $K_j = 5$	8.93	4.83	0.08	4.61

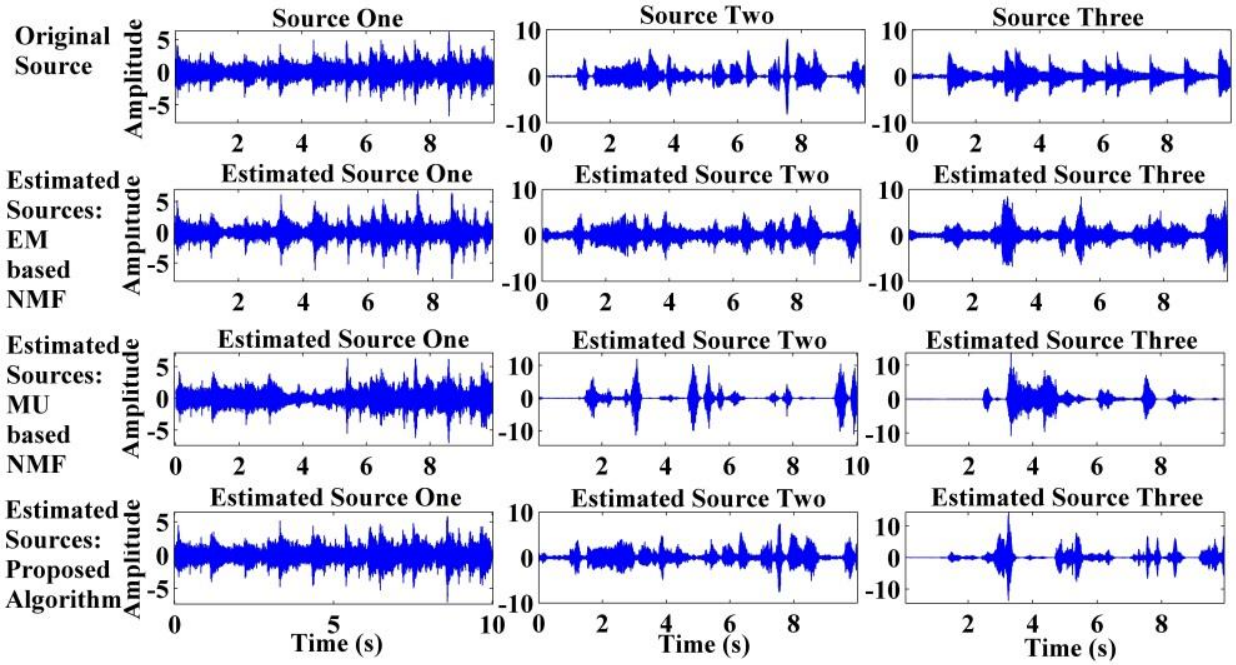


Figure 3.20: Waveforms of the estimated sources for the live recording no drum case.

3.7 Summary

In this chapter the NMF2D has been proposed to develop a machine learning solution for separating the underdetermined convolutive mixture in unsupervised manner and with adaptive sparsity instead of the constant uniform sparsity. For faster convergence the proposed algorithm has been adapted in the GEM-MU algorithm. Also in this chapter a new approach to efficiently initialize the NMF2D has been proposed. Furthermore, the number of components and convolutive parameters of the NMF2D have been estimated by the proposed Gamma-Exponential process. Additionally, this chapter has shown that the window length used in the STFT can be used to match the characteristics of the audio signals. If the mixture contains sources that exhibit pitch-like characteristics, a long-time processing window will extract these sources more efficiently. Conversely, a short-time processing window is more suitable for percussive-like sources. Results have shown that the proposed algorithm is very promising, considerably more flexible and offers an alternative model to the EM- and MU-based NMF, or NTF.

CHAPTER 4
UNDERDETERMINED HIGH-REVERBERANT AUDIO SOURCE SEPARATION USING TWO DIMENSIONAL TENSOR FACTORIZATION TECHNIQUES

In this chapter, a novel algorithm that able to separate the audio sources that have been mixed in an underdetermined reverberant environment will be proposed. Namely, the fusion of K models of full-rank weighted nonnegative tensor factor 2D deconvolution (K-wNTF2D) will be proposed. This model will be adapted under the hybrid framework of the generalized expectation maximization and multiplicative update algorithms in unsupervised manner. In addition, the development and derivation of the proposed full-rank K-wNTF2D algorithm will be shown. Also, the variable sparsity parameters that derived from the Gibbs distribution will be encoded into the K-wNTF2D model in order to optimize each sub-model in K-wNTF2D with the required sparsity which in turn will model the time-varying variances of the sources in the spectrogram. Furthermore, the parameters of the K-wNTF2D will be initialized by the proposed initialization method. Experimental results showed the effectiveness of the proposed algorithm in separating the sources that have been mixed in underdetermined reverberant environment.

This chapter is organized as follows: The proposed K-wNTF2D model will be introduced in Section 4.1. The sources model will be presented in Section 4.2. Section 4.3 is dedicated to the derivation of variable sparsity and the adaptation of GEM-MU algorithm to work with the full-rank K-wNTF2D. The initialization strategy will be proposed in Section 4.4. Experimental results on the SiSEC'13 real datasets and comparison with recent methods will be discussed in Section 4.5. Finally, the conclusions will be drawn in Section 4.6.

4.1 Introduction

A certain set of assumptions are needed to solve the ill-posed problem of the blind source separation. One of these common assumptions in the BSS is the narrowband approximation, and to understand it, it should be known how the observed multichannel signal $\boldsymbol{x}(t)$ can be expressed in Short Time Fourier transform (STFT).The mixture $\boldsymbol{x}(t)$ can be expressed in time domain as

$$x_i(t) = \sum_{j=1}^{J} c_{i,j}(t) + b_i(t), \qquad i = 1,2,\dots I \tag{4.1}$$

where $x_i(t) \in \mathbb{R}, t = 1,\dots,T$ is the received signal from the i^{th} microphone, $c_{i,j}(t) \in \mathbb{R}$ is the spatial image of the source signal j and channel i, J is the number of sources, and $b_i(t) \in \mathbb{R}$ is some additive noise. The spatial image of the source $c_{i,j}(t)$ can be expressed as

$$c_{i,j}(t) = \sum_{\tau=0}^{L-1} a_{i,j}(\tau) s_j(t-\tau) \tag{4.2}$$

where $a_{i,j}(t) \in \mathbb{R}$ is the finite-impulse response of some (causal) filter, L is the filter length, and $s_j(t) \in \mathbb{R}$ is the original source signal.

By substituting eqn. (4.2) into eqn. (4.1), and assuming that the mixing channel is time-invariant then, the STFT of eqn. (4.1) becomes

$$x_{i,f,n} = \sum_{j=1}^{J} a_{i,j,f} s_{j,f,n} + b_{i,f} \tag{4.3a}$$

or in vector form

$$\boldsymbol{x}_{f,n} = \sum_{j=1}^{J} \boldsymbol{a}_{j,f} s_{j,f,n} + \boldsymbol{b}_{f,n} \tag{4.3b}$$

where $\boldsymbol{x}_{f,n} = [x_{1,f,n} \quad \cdots \quad x_{I,f,n}]^H$, $\boldsymbol{a}_{j,f} = [a_{1,j,f} \quad \cdots \quad a_{I,j,f}]^H$, and $x_{i,f,n}$, $a_{i,j,f}$, $s_{j,f,n}$, $b_{i,f,n}$ are the complex-valued STFT of $x_i(t)$, $a_{i,j}(t)$, $s_j(t)$, and $b_i(t)$, respectively. The term $f = 1,2,\dots,F$ is the frequency bin index, and $n = 1,2,\dots,N$ is the time frame index. Thus, the convolutive mixture in eqn. (4.2) is approximated by the narrowband approximation to an instantaneous mixture, where it is assumed that L is shorter than the STFT window size [100]. According to this assumption the covariance matrix of $c_{i,j,f,n}$ (the complex-valued STFT of $c_{i,j}(t)$) defined as

$$\Sigma_{j,f,n}^{(\boldsymbol{c})} = E\left[\boldsymbol{c}_{j,f,n}\boldsymbol{c}_{j,f,n}^H\right] \tag{4.4a}$$

and can be expressed as

$$\Sigma_{j,f,n}^{(c)} = \Sigma_{j,f}^{(a)} v_{j,f,n} \tag{4.4b}$$

or its scalar form as

$$\sigma_{\underline{i}\,j,f,n}^{(c)} = \sigma_{\underline{i}\,j,f}^{(a)} v_{j,f,n} \tag{4.4c}$$

where $\underline{i}$ is the index that represents the column vectorization of a $I \times I$ matrix i.e. $\underline{i} = \{(1,1),(2,1),\ldots,(I,1),(1,2),(2,2),\ldots,(I,I)\} \in \mathbb{R}^{I^2}$, $\Sigma_{j,fn}^{(c)} \in \mathbb{C}^{I \times I}$ is the covariance matrix of the j^{th} source image, $\Sigma_{j,f}^{(a)} \in \mathbb{C}^{I \times I}$ is the time-invariant spatial covariance matrix of the j^{th} source, and $v_{j,f,n} \in \mathbb{R}^{+}$ is the source variance. Therefore, in the case of high-reverberant environment where L is greater than the STFT window size, this assumption will not work. To resolve this issue, Duong *et al.* [9] propose a full-rank spatial covariance matrix (which models the spatial position of the sources as well as their spatial spread) in place of the conventional rank-1 matrix formed from $\Sigma_{j,f}^{(a)} = \boldsymbol{a}_{j,f} \boldsymbol{a}_{j,f}^{H}$. They showed that their results are better than the rank-1 method. Arberet *et al.* [16] take advantages of the full-rank spatial covariance matrix to model the mixing process, and used the NMF to model the source variance. They showed that their results are better than Doung *et al.* under the oracle initialization where both $v_{j,f,n}$ and $\Sigma_{j,f}^{(a)}$ are initialized from the original sources.

However, for a more realistic case, it is not always possible to adapt the oracle initialization approach. In addition, the NMF is practically too simplistic and does not efficiently model more complex sources such as polyphonic music. Therefore, a more powerful source variance representation should be used instead of the NMF (based on Arberet *et al.* [16]). One possible representation is the NMF2D [25], which has a set of convolutive parameters (τ and ϕ) that are convolved in both time and frequency directions by a time-pitch weighted matrix. A set of K number of frequency basis is used instead of the single frequency basis to model the j^{th} source variance which results in

$$v_{j,f,n} = \sum_{k=1}^{K} \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j} \tag{4.5}$$

where K is the number of components or frequency basis assigned to the j^{th} source. The terms τ_{max} and ϕ_{max} are the maximum number of the convolutive parameters τ and ϕ respectively. $w_{f,k}^{\tau,j}$

represents the k^{th} spectral basis of the j^{th} source, and $h_{k,n}^{\phi,j}$ represents the k^{th} temporal code for each spectral basis element of the j^{th} source, for $f = 1, \dots, F, n = 1, \dots, N,$ and $j = 1, \dots, J$. With eqn. (4.5), the covariance matrix in eqn. (4.4) can now be expressed as

$$\Sigma_{j,f,n}^{(\boldsymbol{c})} = \sum_{k=1}^{K} \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} \Sigma_{j,f}^{(\boldsymbol{a})} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j} \tag{4.6a}$$

and its scalar form as

$$\sigma_{\underline{i}\,j,f,n}^{(c)} = \sum_{k=1}^{K} \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} \sigma_{\underline{i}\,j,f}^{(a)} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j} \tag{4.6b}$$

The full-rank "*mixture covariance matrix*" of $\boldsymbol{x}_{f,n}$ in eqn. (4.3*b*) is defined as

$$\Sigma_{f,n}^{(\boldsymbol{x})} = E\left[\boldsymbol{x}_{f,n} \boldsymbol{x}_{f,n}^{H}\right]$$

$$= \sum_{j=1}^{J} \Sigma_{j,f,n}^{(\boldsymbol{c})} + \Sigma_{f}^{(\boldsymbol{b})} \tag{4.7a}$$

where $\Sigma_{f}^{(\boldsymbol{b})}$ is the time invariant noise covariance matrix. Using eqn. (4.6a), $\Sigma_{f,n}^{(\boldsymbol{x})}$ can be expressed as

$$\Sigma_{f,n}^{(\boldsymbol{x})} = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} \Sigma_{j,f}^{(\boldsymbol{a})} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j} + \Sigma_{f}^{(\boldsymbol{b})} \tag{4.7b}$$

The scalar form of $\Sigma_{f,n}^{(\boldsymbol{x})}$ can be expressed as

$$\sigma_{\underline{i},f,n}^{(x)} = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} \sigma_{\underline{i},j,f}^{(a)} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j} + \sigma_{\underline{i},f}^{(b)} \tag{4.7c}$$

Of special note is that eqn. (4.7b) represents a non-negative tensor factorization of the *mixture covariance matrix* (arranged as a 3-dimensional tensor) into a product of spatial covariance matrix (arranged as a 3-dimensional tensor), spectral basis and temporal codes (the latter two estimate the

source image variances). Since eqn. (4.7b) is a combination of K models of weighted NTF2D, it will be termed as the "K-wNTF2D"[3].

The full-rank K-wNTF2D will be optimized using the GEM-MU algorithm [80] which provides a probabilistic platform for joint estimation of the sources and the parameters as well as preserving the non-negativity constraints of the model. In addition, the GEM-MU algorithm accelerates the convergence speed of the parameters update. Concurrently, the variable sparsity will be encoded into the K-wNTF2D instead of using some heuristics approaches to fix them to a constant value. The variable sparsity will be developed based on the Gibbs distribution framework and optimized under the Itakura-Saito divergence. This will be contrasted with the uniform sparsity which assigns a fixed sparsity over all the elements of $\boldsymbol{H} = \{h_{k,n}^{\phi,j}\}$. Since the acoustic sources such as speech changes dynamically over time, uniform sparsity will lead to either over-sparseness (resulting in too many elements of $\boldsymbol{H}$ set to zero), or under-sparseness (a lot of ineffective elements in $\boldsymbol{H}$). The proposed variable sparsity relieves this problem by optimizing the sparsity for each individual elements of $\boldsymbol{H}$ through learning from the data.

The Itakura-Saito (IS) divergence will be considered in this chapter due to its scale invariant property [49]. Compared with the Least Square (LS) distance and Kullback-Leibler (KL) divergence cost functions, IS divergence deals with both low and high energy components with equal emphasis. Since both speech and music signals have large magnitude dynamic ranges, IS divergence provides a faithful measure between the observed data and the output generated from the adapted K-wNTF2D model. Also initialization strategy for the NMF2D will be considered. Since poor initialization can lead to converge to unwanted local minima, a novel initialization method will be developed to initialize the K-wNTF2D. For ease of understanding, a high-level presentation of the proposed algorithm is shown in Figure 4.1.

[3] By definition, a 3-dimensional NTF is given by $V_{i,f,n} = \sum_j a_{i,j} b_{f,j} c_{j,n}$. This can be extended to NTF2D by introducing the convolutive parameters as $V_{i,f,n} = \sum_j \sum_\tau \sum_\phi a_{i,j} b_{f-\phi,j}^{\tau} c_{j,n-\tau}^{\phi}$. We can further extend the NTF2D by introducing a dependence of $a_{i,j}$ with respect to one of the dimension say f i.e. $a_{i,j}(f)$. In this case, we replace $a_{i,j}$ with $a_{i,j,f}$ so that $V_{i,f,n} = \sum_j \sum_\tau \sum_\phi a_{i,j,f} b_{f-\phi,j}^{\tau} c_{j,n-\tau}^{\phi}$. This coupling allows us to weight the NTF2D as a function of f. We term this as the weighted NTF2D (wNTF2D). Finally, we introduce a fusion of K models of weighted NTF2D resulting to $V_{i,f,n} = \sum_{k=1}^{K} \sum_j \sum_\tau \sum_\phi a_{i,j,f} b_{f-\phi,j}^{\tau,k} c_{j,n-\tau}^{\phi,k}$, which we term it as the "K-wNTF2D".

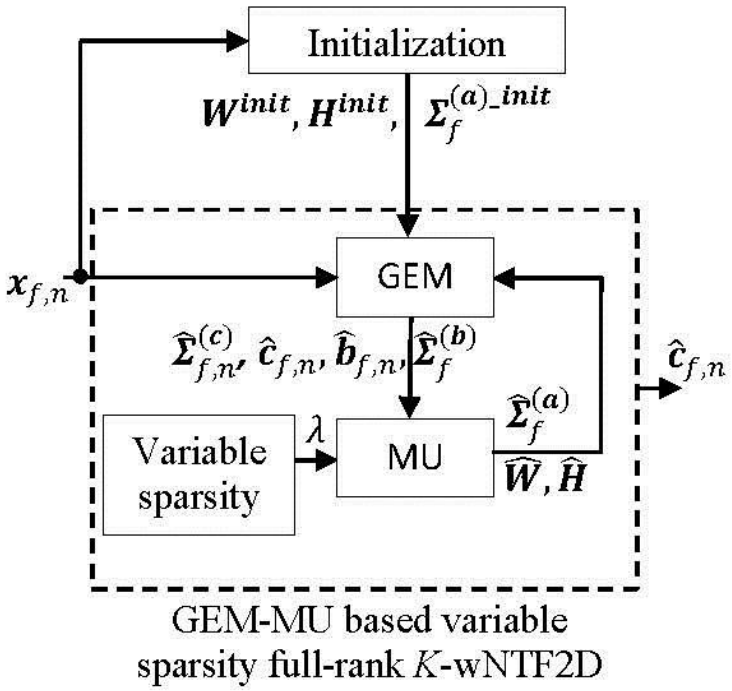


Figure 4.1: High level presentation of the proposed algorithm.

4.2 Source Model

The spatial image of the sources can be modeled as realization of zero-mean proper complex distribution

$$\boldsymbol{c}_{j,f,n} \sim \mathcal{N}_c\left(0, \Sigma_{j,f,n}^{(\boldsymbol{c})}\right) \tag{4.8}$$

and its probability density function (pdf) can be expressed as

$$\mathcal{N}_c\left(0, \Sigma_{j,f,n}^{(\boldsymbol{c})}\right) \triangleq \frac{1}{det\left(\pi\Sigma_{j,f,n}^{(\boldsymbol{c})}\right)} e^{-\left(\boldsymbol{c}_{j,f,n}^{H} \Sigma_{j,f,n}^{(\boldsymbol{c})^{-1}} \boldsymbol{c}_{j,f,n}\right)} \tag{4.9}$$

Substituting eqn. (4.6a) into eqn. (4.8) will results in the following

$$\boldsymbol{c}_{j,f,n} \sim \mathcal{N}_c\left(0, \Sigma_{j,f}^{(\boldsymbol{a})}\left(\sum_{k=1}^{K} \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}\right)\right) \tag{4.10}$$

The noise $\boldsymbol{b}_{f,n}$ in eqn. (4.3) is assumed to be time invariant, stationary and spatially uncorrelated, i.e.

$$\boldsymbol{b}_{f,n} \sim \mathcal{N}_c\left(0, \Sigma_f^{(\boldsymbol{b})}\right) \tag{4.11}$$

and its pdf can be expressed as

$$\mathcal{N}_c\left(0, \Sigma_f^{(\boldsymbol{b})}\right) \triangleq \frac{1}{det\left(\pi \Sigma_f^{(\boldsymbol{b})}\right)} e^{-\left(\boldsymbol{b}_{f,n}^H \Sigma_f^{(\boldsymbol{b})^{-1}} \boldsymbol{b}_{f,n}\right)} \tag{4.12}$$

4.3 Proposed Estimation Algorithm

The conditional expectation of the natural statistics will be estimated using the GEM algorithm, and the mixing parameter, $\boldsymbol{W} = \left\{w_{f,k}^{\tau,j}\right\}$, and $\boldsymbol{H} = \left\{h_{k,n}^{\phi,j}\right\}$ will be estimated in the M step using the MU algorithm. The model parameters are $\boldsymbol{\Theta} = \left\{\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\Sigma}^{(\boldsymbol{a})}, \boldsymbol{\Sigma}^{(\boldsymbol{b})}, \boldsymbol{\Lambda}\right\}$. To facilitate the estimation, the following posterior probability is formed:

$$P\left(\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H} \middle| \boldsymbol{X}, \boldsymbol{\Sigma}^{(\boldsymbol{a})}, \boldsymbol{\Sigma}^{(\boldsymbol{b})}, \boldsymbol{\Lambda}\right) = \frac{P\left(\boldsymbol{X} \middle| \boldsymbol{C}, \boldsymbol{\Sigma}^{(\boldsymbol{b})}\right) P\left(\boldsymbol{C} \middle| \boldsymbol{\Sigma}^{(\boldsymbol{a})}, \boldsymbol{W}, \boldsymbol{H}\right) P\left(\boldsymbol{W}, \boldsymbol{H} | \boldsymbol{\Lambda}\right)}{P\left(\boldsymbol{X} \middle| \boldsymbol{C}, \boldsymbol{\Sigma}^{(a)}, \boldsymbol{\Sigma}^{(b)}\right)} \tag{4.13}$$

and their minus log-posterior is

$$\begin{aligned} -\log P\left(\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H} \middle| \boldsymbol{X}, \boldsymbol{\Sigma}^{(\boldsymbol{a})}, \boldsymbol{\Sigma}^{(\boldsymbol{b})}, \boldsymbol{\Lambda}\right) = -\log P\left(\boldsymbol{X} \middle| \boldsymbol{C}, \boldsymbol{\Sigma}^{(\boldsymbol{b})}\right) - \log P\left(\boldsymbol{C} \middle| \boldsymbol{\Sigma}^{(\boldsymbol{a})}, \boldsymbol{W}, \boldsymbol{H}\right) - \\ \log P\left(\boldsymbol{W}, \boldsymbol{H} | \boldsymbol{\Lambda}\right) + const \end{aligned} \tag{4.14}$$

where $\boldsymbol{\Lambda} = \left\{\lambda_{k,n}^{\phi,j}\right\}$ is a tensor that contains the sparsity terms. The log- posterior will be computed by the GEM-MU based full-rank variable sparsity *K*-wNTF2D in the following sections.

4.3.1 E-Step: Conditional Expectations of Natural Statistics

Maximizing the log-likelihood in eqn. (4.14) is equivalent to minimizing

$$-\log P\left(\boldsymbol{X} \middle| \boldsymbol{C}, \boldsymbol{\Sigma}^{(\boldsymbol{b})}\right) = \left(\boldsymbol{x}_{f,n}^H \Sigma_{f,n}^{(\boldsymbol{x})^{-1}} \boldsymbol{x}_{f,n}\right) + \log\left(det\left(\pi \Sigma_{f,n}^{(\boldsymbol{x})}\right)\right) \tag{4.15}$$

The conditional expectation of the natural statistics $\hat{R}_{j,f,n}^{(c)}, \hat{R}_{f}^{(b)}, \hat{\Sigma}_{j,f,n}^{(c)}, \hat{\Sigma}_{f}^{(b)}, \hat{\boldsymbol{c}}_{j,f,n}$ and $\widehat{\boldsymbol{b}}_{f,n}$ are shown below:

$$\hat{R}_{j,f,n}^{(\boldsymbol{c})} = \hat{\boldsymbol{c}}_{j,f,n}\hat{\boldsymbol{c}}_{j,f,n}^{H} + \hat{\Sigma}_{j,f,n}^{(c)} \tag{4.16}$$

$$\hat{\Sigma}_{j,f,n}^{(c)} = \left(\mathrm{I} - \Sigma_{j,f,n}^{(c)}\Sigma_{f,n}^{(x)^{-1}}\right)\Sigma_{j,f,n}^{(c)} \tag{4.17}$$

$$\hat{\boldsymbol{c}}_{j,f,n} = \Sigma_{j,f,n}^{(c)}\Sigma_{f,n}^{(x)^{-1}}\boldsymbol{x}_{f,n} \tag{4.18}$$

$$\hat{R}_{f}^{(b)} = \widehat{\boldsymbol{b}}_{f,n}\widehat{\boldsymbol{b}}_{f,n}^{H} + \hat{\Sigma}_{f}^{(b)} \tag{4.19}$$

$$\hat{\Sigma}_{f}^{(b)} = \left(\mathrm{I} - \Sigma_{f}^{(b)}\Sigma_{f,n}^{(x)^{-1}}\right)\Sigma_{f}^{(b)} \tag{4.20}$$

$$\widehat{\boldsymbol{b}}_{f,n} = \Sigma_{f}^{(b)}\Sigma_{f,n}^{(x)^{-1}}\boldsymbol{x}_{f,n} \tag{4.21}$$

Appendix A is dedicated for the detailed derivation of eqns. (4.16) to (4.21).

4.3.2 M-Step: Update of Parameters

For clarification and simplification, $\hat{R}_{j,f,n}^{(c)}$ and $\Sigma_{j,f}^{(a)}$ will be vectorized to $I^2 \times 1$ vectors as follows:

$$\begin{aligned}\underline{\hat{\boldsymbol{r}}}_{j,f,n}^{(\boldsymbol{c})} &= \left\{\hat{r}_{\underline{i},j,f,n}^{(c)}\right\} \\ &= \left[\hat{r}_{1,1,j,f,n}^{(c)} \quad \hat{r}_{2,1,j,f,n}^{(c)} \quad \cdots \quad \hat{r}_{I,1,j,f,n}^{(c)} \quad \hat{r}_{1,2,j,f,n}^{(c)} \quad \cdots \quad \hat{r}_{I,I,j,f,n}^{(c)}\right]^{T}\end{aligned} \tag{4.22}$$

$$\begin{aligned}\underline{\boldsymbol{\sigma}}_{j,f}^{(\boldsymbol{a})} &= \left\{\sigma_{\underline{i}\,j,f}^{(a)}\right\} \\ &= \left[\sigma_{1,1,j,f,n}^{(a)} \quad \sigma_{2,1,j,f,n}^{(a)} \quad \cdots \quad \sigma_{I,1,j,f,n}^{(a)} \quad \sigma_{1,2,j,f,n}^{(a)} \quad \cdots \quad \sigma_{I,I,j,f,n}^{(a)}\right]^{T}\end{aligned} \tag{4.23}$$

Therefore, eqn. (4.6a) can be rewritten as follows:

$$\underline{\boldsymbol{\sigma}}_{j,f,n}^{(\boldsymbol{c})} = \sum_{k=1}^{K}\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}}\underline{\boldsymbol{\sigma}}_{j,f}^{(a)}w_{f-\phi,k}^{\tau,j}h_{k,n-\tau}^{\phi,j} \tag{4.24}$$

The second term in the right hand side of eqn. (4.14) can be expressed with IS divergence as

$$-\log P\left(\boldsymbol{C}\middle|\boldsymbol{\Sigma}^{(a)},\boldsymbol{W},\boldsymbol{H}\right) = \sum_{\underline{i},j,f,n} D_{IS}\left(\hat{r}^{(\boldsymbol{c})}_{\underline{i},j,f,n}\middle|\sum_k \sigma^{(\boldsymbol{a})}_{\underline{i},j,f}\left(\sum_\tau \sum_\phi w^{\tau,j}_{f-\phi,k} h^{\phi,j}_{k,n-\tau}\right)\right) \quad (4.25)$$

The third term in the right hand side of eqn. (4.14) is the prior information on $\boldsymbol{W}$ and $\boldsymbol{H}$. An improper prior is assumed for $\boldsymbol{W}$ and factor-wise normalized to unit length i.e. $p(\boldsymbol{W}) = \prod_j \delta\left(\left\|\boldsymbol{W}^j\right\|_2 - 1\right)$ where $\boldsymbol{W}^j = \left\{w^{\tau,j}_{f,k}\right\}$ is the spectral basis that belongs to the j^{th} source. Each element of $\boldsymbol{H}$ has independent decay parameter $\lambda^{\phi,j}_{k,n}$ with exponential distribution:

$$\begin{aligned}
-\log p(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{\Lambda}) &= -\log\left(\prod_j \delta\left(\left\|\boldsymbol{W}^j\right\|_2 - 1\right)\right) - \log\left(\prod_{j,k} p\left(H^j_k\middle|\Lambda^j_k\right)\right) \\
&= -\log\left(\prod_j \delta\left(\left\|\boldsymbol{W}^j\right\|_2 - 1\right)\right) - \log\left(\prod_j\prod_k\prod_n\prod_\phi \lambda^{\phi,j}_{k,n} exp\left(-\lambda^{\phi,j}_{k,n} h^{\phi,j}_{k,n}\right)\right) \\
&= -\sum_j \log\delta\left(\left\|\boldsymbol{W}^j\right\|_2 - 1\right) + \sum_j\sum_k\sum_n\sum_\phi\left(\lambda^{\phi,j}_{k,n} h^{\phi,j}_{k,n} - \log\lambda^{\phi,j}_{k,n}\right) \quad (4.26)
\end{aligned}$$

The first term on the right hand side of eqn. (4.26) can be satisfied by explicitly normalizing each spectral dictionary to unity i.e. $w^{\tau,j}_{f,k} = w^{\tau,j}_{f,k} \Big/ \sqrt{\sum_{f,\tau,k}\left(w^{\tau,j}_{f,k}\right)^2}$. Thus, only the second term remains i.e. $-\log p(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{\Lambda}) = \sum_j\sum_k\sum_n\sum_\phi\left(\lambda^{\phi,j}_{k,n} h^{\phi,j}_{k,n} - \log\lambda^{\phi,j}_{k,n}\right)$. Adding this to the IS divergence derived in eqn. (4.25), will leads to the following

$$\begin{aligned}
&-\log P\left(\boldsymbol{C}\middle|\boldsymbol{\Sigma}^{(a)},\boldsymbol{W},\boldsymbol{H}\right) - \log P(\boldsymbol{W},\boldsymbol{H}|\boldsymbol{\Lambda}) \\
&\quad= \sum_{\underline{i},j,k,f,n}\left(\hat{r}^{(c)}_{\underline{i},j,f,n}\sigma^{(\boldsymbol{a})^{-\mathbf{1}}}_{\underline{i},j,f} v^{-1}_{j,f,n} - \log\left(\hat{r}^{(c)}_{\underline{i},j,f,n}\sigma^{(\boldsymbol{a})^{-\mathbf{1}}}_{\underline{i},j,f} v^{-1}_{j,f,n}\right) - 1\right) \\
&\qquad+ \sum_{j,k,n,\phi}\lambda^{\phi,j}_{k,n} h^{\phi,j}_{k,n} - \sum_{j,k,n,\phi}\log\lambda^{\phi,j}_{k,n} \quad (4.27)
\end{aligned}$$

Thus the derivatives of eqn. (4.27) with respect to $\sigma^{(\boldsymbol{a})}_{\underline{i},j,f}$, $w^{\tau,j}_{f,k}$ and $h^{\phi,j}_{k,n}$ can be given as follows:

$$\frac{\partial}{\partial \sigma^{(\boldsymbol{a})}_{\underline{i}',j',f'}} \log P\left(\boldsymbol{C},\boldsymbol{W},\boldsymbol{H}\middle|\boldsymbol{X},\boldsymbol{\Sigma}^{(a)},\boldsymbol{\Sigma}^{(b)},\boldsymbol{\Lambda}\right) = -\sum_{n} \hat{r}^{(c)}_{\underline{i}',j',f',n} \sigma^{(a)^{-2}}_{\underline{i}',j',f'} v^{-1}_{j',f',n} + \sigma^{(a)^{-1}}_{\underline{i}',j',f'} \quad (4.28)$$

Similarly,

$$\frac{\partial}{\partial w^{\tau',j'}_{f',k'}} \log P\left(\boldsymbol{C},\boldsymbol{W},\boldsymbol{H}\middle|\boldsymbol{X},\boldsymbol{\Sigma}^{(a)},\boldsymbol{\Sigma}^{(b)},\boldsymbol{\Lambda}\right) = -\sum_{\underline{i},\phi,n} \hat{r}^{(c)}_{\underline{i},j',f'+\phi,n} \sigma^{(a)^{-1}}_{\underline{i},j',f'+\phi} v^{-2}_{j',f'+\phi,n} h^{\phi,j'}_{k',n-\tau'} + \sum_{\phi,n} v^{-1}_{j',f'+\phi,n} h^{\phi,j'}_{k',n-\tau'} \quad (4.29)$$

Likewise,

$$\frac{\partial}{\partial h^{\phi',j'}_{k',n'}} \log P\left(\boldsymbol{C},\boldsymbol{W},\boldsymbol{H}\middle|\boldsymbol{X},\boldsymbol{\Sigma}^{(a)},\boldsymbol{\Sigma}^{(b)},\boldsymbol{\Lambda}\right) = -\sum_{\underline{i},f,\tau} \hat{r}^{(c)}_{\underline{i},j',f,n'+\tau} \sigma^{(a)^{-1}}_{\underline{i},j',f} v^{-2}_{j',f,n'+\tau} w^{\tau,j'}_{f-\phi',k'} + \sum_{f,\tau} v^{-1}_{j',f,n'+\tau} w^{\tau,j'}_{f-\phi',k'} + \lambda^{\phi',j'}_{k',n'} \quad (4.30)$$

For each component, standard gradient descent method is applied with

$$\sigma^{(\boldsymbol{a})}_{\underline{i}',j',f'} \leftarrow \sigma^{(\boldsymbol{a})}_{\underline{i}',j',f'} - \eta_{\Sigma^{(a)}} \frac{\partial \log P\left(\boldsymbol{C},\boldsymbol{W},\boldsymbol{H}\middle|\boldsymbol{X},\boldsymbol{\Sigma}^{(a)},\boldsymbol{\Sigma}^{(b)},\boldsymbol{\Lambda}\right)}{\partial \sigma^{(\boldsymbol{a})}_{\underline{i}',j',f'}} \quad (4.31)$$

$$w^{\tau',j'}_{f',k'} \leftarrow w^{\tau',j'}_{f',k'} - \eta_{w} \frac{\partial \log P\left(\boldsymbol{C},\boldsymbol{W},\boldsymbol{H}\middle|\boldsymbol{X},\boldsymbol{\Sigma}^{(a)},\boldsymbol{\Sigma}^{(b)},\boldsymbol{\Lambda}\right)}{\partial w^{\tau',j'}_{f',k'}} \quad (4.32)$$

$$h^{\phi',j'}_{k',n'} \leftarrow h^{\phi',j'}_{k',n'} - \eta_{h} \frac{\partial \log P\left(\boldsymbol{C},\boldsymbol{W},\boldsymbol{H}\middle|\boldsymbol{X},\boldsymbol{\Sigma}^{(a)},\boldsymbol{\Sigma}^{(b)},\boldsymbol{\Lambda}\right)}{\partial h^{\phi',j'}_{k',n'}} \quad (4.33)$$

where $\eta_{\Sigma^{(a)}}, \eta_w$, and η_h are the positive learning rate, which can be set as

$$\eta_{\Sigma^{(a)}} = \frac{\sigma^{(\boldsymbol{a})}_{\underline{i}',j',f'}}{\sigma^{(a)^{-1}}_{\underline{i}',j',f'}} \quad (4.34)$$

$$\eta_{w} = \frac{w^{\tau',j'}_{f',k'}}{\sum_{\phi,n} v^{-1}_{j',f'+\phi,n} h^{\phi,j'}_{k',n-\tau'}} \quad (4.35)$$

$$\eta_h = \frac{h_{k',n'}^{\phi',j'}}{\sum_{f,\tau} v_{j',f,n'+\tau}^{-1} w_{f-\phi',k'}^{\tau,j'} + \lambda_{k',n'}^{\phi',j'}} \tag{4.36}$$

The MU rules for $\sigma_{\underline{i}',j',f'}^{(\boldsymbol{a})}, w_{f,k}^{\tau,j}$ and $h_{k,n}^{\phi,j}$ respectively gives

$$\sigma_{\underline{i}',j',f'}^{(\boldsymbol{a})} \leftarrow \frac{1}{N}\sum_{n=1}^{N} \frac{\hat{r}_{\underline{i}',j',f',n}^{(c)}}{v_{j',f',n}} \tag{4.37}$$

$$w_{f',k'}^{\tau',j'} \leftarrow w_{f',k'}^{\tau',j'} \left(\frac{\sum_{\underline{i},\phi,n} \hat{r}_{\underline{i},j',f'+\phi,n}^{(c)} \sigma_{\underline{i},j',f'+\phi}^{(a)^{-1}} v_{j',f'+\phi,n}^{-2} h_{k',n-\tau'}^{\phi,j'}}{\sum_{\phi,n} v_{j',f'+\phi,n}^{-1} h_{k',n-\tau'}^{\phi,j'}} \right) \tag{4.38}$$

$$h_{k',n'}^{\phi',j'} \leftarrow h_{k',n'}^{\phi',j'} \left(\frac{\sum_{\underline{i},f\tau} \hat{r}_{\underline{i},j',f,n'+\tau}^{(c)} \sigma_{\underline{i},j',f}^{(a)^{-1}} v_{j',f,n'+\tau}^{-2} w_{f-\phi',k'}^{\tau,j'}}{\sum_{f\tau} v_{j',f,n'+\tau}^{-1} w_{f-\phi',k'}^{\tau,j'} + \lambda_{k',n'}^{\phi',j'}} \right) \tag{4.39}$$

4.3.3 Estimation of Variable Sparsity Using Gibbs Distribution

For the sparsity term, the update is obtained as follows:

$$\begin{aligned} \lambda &= \arg\max_{\lambda} \log \mathrm{P}\left(\boldsymbol{C}, \boldsymbol{W}, \boldsymbol{H} \middle| \boldsymbol{X}, \boldsymbol{\Sigma}^{(a)}, \boldsymbol{\Sigma}^{(b)}, \boldsymbol{\Lambda}\right) \\ &= \arg\max_{\lambda} \left(\log P\left(\boldsymbol{X} \middle| \boldsymbol{C}, \boldsymbol{\Sigma}^{(b)}\right) + \log P\left(\boldsymbol{C} \middle| \boldsymbol{\Sigma}^{(a)}, \boldsymbol{W}, \boldsymbol{H}\right) + \log P\left(\boldsymbol{W}, \boldsymbol{H} | \boldsymbol{\Lambda}\right) + const\right) \\ &= \arg\max_{\lambda} \log P\left(\boldsymbol{H} | \boldsymbol{\Lambda}\right) \end{aligned} \tag{4.40}$$

Solving $\frac{\partial}{\partial\lambda} \log P\left(\boldsymbol{H} | \boldsymbol{\Lambda}\right) = 0$ will lead to

$$\lambda_{k,n}^{\phi,j} = \frac{1}{h_{k,n}^{\phi,j}} \quad (\text{or in matrix form } \boldsymbol{\Lambda} = 1 \cdot / \boldsymbol{H}) \tag{4.41}$$

where "·/" represents element-wise division. However, as $\boldsymbol{H}$ can be partitioned into distinct subsets of positive value and zero value it will yield divergent updates for $h_{k,n}^{\phi,j} = 0$. Therefore, a better approximation to account for variability of $\boldsymbol{H}$ is required. To consider the variability of $\boldsymbol{H}$, it will be

casted in vector form and τ_{max} will be set to zero ($\tau_{max} = 0$). For any distribution $Q(\underline{\boldsymbol{h}})$ (that represents the lower bound to obtain the hidden variable $\underline{\boldsymbol{\lambda}}$), the log-likelihood function satisfies the following:

$$\log P\left(\underline{\boldsymbol{h}}|\underline{\boldsymbol{\lambda}}\right) = \log \int Q(\underline{\boldsymbol{h}}) \frac{P\left(\underline{\boldsymbol{h}}|\underline{\boldsymbol{\lambda}}\right)}{Q\left(\underline{\boldsymbol{h}}\right)} d\underline{\boldsymbol{h}} \tag{4.42}$$

where $\underline{\boldsymbol{h}} = \left[\mathrm{Vec}(\mathrm{H}^0)^{\mathrm{T}} \quad \mathrm{Vec}(\mathrm{H}^1)^{\mathrm{T}} \quad \cdots \quad \mathrm{Vec}\left(\mathrm{H}^{\Phi_{\max}}\right)^{\mathrm{T}}\right]^{\mathrm{T}}$, $\underline{\boldsymbol{\lambda}} = [\mathrm{Vec}(\lambda^0)^{\mathrm{T}} \quad \mathrm{Vec}(\lambda^1)^{\mathrm{T}} \quad \cdots \quad \mathrm{Vec}\left(\lambda^{\Phi_{\max}}\right)^{\mathrm{T}}\Big]^{\mathrm{T}}$, $\mathrm{Vec}(.)$ means column vectorization, and $\underline{h}$ and $\underline{\lambda}$ are vectors with dimension $\mathrm{D} \times 1$ where $\mathrm{D} = K \times N \times \Phi_{max}$. The elements of $\underline{\boldsymbol{h}}$ and $\underline{\boldsymbol{\lambda}}$ are denoted as h_p and λ_{p}, respectively, for $\mathrm{p} = 1, 2, \ldots, \mathrm{D}$. By using Jensen's inequality eqn. (4.42) becomes

$$\log P\left(\underline{\boldsymbol{h}}|\underline{\boldsymbol{\lambda}}\right) \geq \int Q\left(\underline{\boldsymbol{h}}\right) \log\left(\frac{P\left(\underline{\boldsymbol{h}}|\underline{\boldsymbol{\lambda}}\right)}{Q\left(\underline{\boldsymbol{h}}\right)}\right) d\underline{\boldsymbol{h}} \tag{4.43}$$

By substituting eqn. (4.43) into eqn. (4.40)

$$\underline{\boldsymbol{\lambda}} = \arg\max_{\underline{\boldsymbol{\lambda}}} \left(\int Q\left(\underline{\boldsymbol{h}}\right) \log P\left(\underline{\boldsymbol{h}}|\underline{\boldsymbol{\lambda}}\right) d\underline{\boldsymbol{h}} - \int Q\left(\underline{\boldsymbol{h}}\right) \log Q\left(\underline{\boldsymbol{h}}\right) d\underline{\boldsymbol{h}}\right)$$

$$= \arg\max_{\underline{\boldsymbol{\lambda}}} \int Q\left(\underline{\boldsymbol{h}}\right) \left(\log \lambda_p - \lambda_p h_p\right) d\underline{\boldsymbol{h}} \tag{4.44}$$

Eqn. (4.44) can be solved as follows:

$$\frac{\partial \int Q\left(\underline{\boldsymbol{h}}\right) \left(\log \lambda_p - \lambda_p\, h_p\right) d\underline{\boldsymbol{h}}}{\partial \lambda_p} = 0$$

$$\lambda_p = \frac{1}{\int h_p Q\left(\underline{\boldsymbol{h}}\right) d\underline{\boldsymbol{h}}}$$

$$= \frac{1}{E_{Q(\underline{\boldsymbol{h}})}[h_P]} \tag{4.45}$$

where $E_{Q(\underline{\boldsymbol{h}})}[h_P]$ is the expectation of h_P under the distribution $Q(\underline{\boldsymbol{h}})$. Eqn. (4.45) cannot be solved analytically therefore $Q\left(\underline{\boldsymbol{h}}\right)$ will be approximated with respect to the mode of distribution h_p. As

h_p can be partitioned into distinct subsets of positive value $(\underline{\boldsymbol{h}}_M)\ \forall_m \in M$ such that $h_m > 0$, and zero value $(\underline{\boldsymbol{h}}_L)\ \forall_l \in L$ such that $h_l = 0$, it follows that $Q(\underline{\boldsymbol{h}})$ can be partitioned as

$$F(\underline{h}) = \sum_{\underline{i},j,f,p} D_{IS}\left(\hat{r}^{(\boldsymbol{c})}_{\underline{i},j,f,p} \middle| \sigma^{(\boldsymbol{a})}_{\underline{i},j,f} v_{j,f,p}\right) + \sum_{p}\left(\lambda_p h_p - \log \lambda_p\right)$$

$$= \sum_{\underline{i},j,f,p} \left(\hat{r}^{(\boldsymbol{c})}_{\underline{i},j,f,p} \sigma^{(a)^{-1}}_{\underline{i},j,f} v^{-1}_{j,f,p} - log\left(\hat{r}^{(\boldsymbol{c})}_{\underline{i},j,f,p} \sigma^{(a)^{-1}}_{\underline{i},j,f} v^{-1}_{j,f,p}\right) - 1\right) + \sum_{p}\left(\lambda_p h_p - \log \lambda_p\right) \quad (4.46)$$

and by using the reverse Triangle Inequality [101], the following can be obtained

$$F(\underline{\boldsymbol{h}}) \geq \sum_{\underline{i},j,f,m} D_{IS}\left(\hat{r}^{(\boldsymbol{c})}_{\underline{i},j,f,m} \middle| \sigma^{(\boldsymbol{a})}_{\underline{i},j,f} v_{j,f,m}\right) + \sum_{m}(\lambda_m h_m - \log \lambda_m)$$
$$+ \sum_{\underline{i},j,f,l} D_{IS}\left(\hat{r}^{(\boldsymbol{c})}_{\underline{i},j,f,l} \middle| \sigma^{(\boldsymbol{a})}_{\underline{i},j,f} v_{j,f,l}\right) + \sum_{l}(\lambda_l h_l - \log \lambda_l)$$

$$F(\underline{\boldsymbol{h}}) \geq F(\underline{\boldsymbol{h}}_L) + F(\underline{\boldsymbol{h}}_M) \quad (4.47)$$

The distribution $\mathrm{Q}(\underline{\boldsymbol{h}})$ will be expressed by the Gibbs distribution [102], i.e.

$$Q(\underline{\boldsymbol{h}}) = \frac{1}{Z_h} exp[-F(\underline{\boldsymbol{h}})] \quad (4.48)$$

where $Z_h = \int exp[-F(\underline{\boldsymbol{h}})] d\underline{\boldsymbol{h}}$. Substituting eqn. (4.47) into eqn. (4.48) will leads to

$$Q(\underline{\boldsymbol{h}}) = \frac{1}{Z_h} exp[-F(\underline{\boldsymbol{h}}_L) - F(\underline{\boldsymbol{h}}_M)]$$

$$= \frac{1}{Z_L} exp[-F(\underline{\boldsymbol{h}}_L)] \frac{1}{Z_M} exp[-F(\underline{\boldsymbol{h}}_M)]$$

$$= Q_L(\underline{\boldsymbol{h}}_L) Q_M(\underline{\boldsymbol{h}}_M) \quad (4.49)$$

where $Z_L = \int exp[-F(\underline{\boldsymbol{h}}_L)]\, d\underline{\boldsymbol{h}}_L$ and $Z_M = \int exp[-F(\underline{\boldsymbol{h}}_M)] d\underline{\boldsymbol{h}}_M$.

The distribution $Q_M(\underline{\boldsymbol{h}}_M)$ is within the boundaries of this distribution and it leads to

$$E_{Q_M(\underline{\boldsymbol{h}}_M)}[h_P] = h_m \quad (4.50)$$

which is optimized in eqn. (4.39). While the distribution $Q_M(\underline{\boldsymbol{h}}_L)$ is on the boundaries of this distribution and it leads to

$$E_{Q_L(\underline{\boldsymbol{h}}_L)}[h_P] = u_l \tag{4.51}$$

where u_l is the variational parameter that model the distribution of $\underline{\boldsymbol{h}}_L$. Therefore, eqn. (4.45) is given by

$$\lambda_p = \begin{cases} \dfrac{1}{h_p} & \forall_p \in M \\ \dfrac{1}{u_p} & \forall_p \in L \end{cases} \tag{4.52}$$

The variational optimization [102] will be applied to derive the variational parameter u_l as follows. The parameter u_l is obtained by minimizing the Kullback-Leibler divergence between Q_L and $\hat{Q}_L$

$$u_l = arg\min_{u_l} \int \hat{Q}_L(\underline{\boldsymbol{h}}_L) \log \frac{\hat{Q}_L(\underline{\boldsymbol{h}}_L)}{Q_L(\underline{\boldsymbol{h}}_L)}\, d\underline{\boldsymbol{h}}_L \tag{4.53}$$

The distribution $Q_L(\underline{\boldsymbol{h}}_L)$ in eqn. (4.53) will be approximated by considering the Taylor expansion about the updated $h_l = h$ (given by eqn. (4.39)):

$$Q_L(\underline{\boldsymbol{h}}_L \geq 0) \propto exp\left\{ -\sum_{l\in L} \left(\left(\frac{\partial F(h_l)}{\partial h_l} \right)\Big|_{h_l=h} \right) h_l - \frac{1}{2} \sum_{l\in L} \left(\left(\frac{\partial^2 F(h_l)}{\partial h_l^2} \right)\Big|_{h_l=h} \right) h_l^2 \right\}$$

$$Q_L(\underline{\boldsymbol{h}}_L \geq 0) \propto exp\left\{ \begin{array}{l} \displaystyle\sum_{\underline{i}jkfl\phi} \left(\hat{r}^{(c)}_{\underline{i},j,f,l} \sigma^{(a)^{-1}}_{\underline{i},j,f} v^{-2}_{j,f,l} w^{j}_{f-\phi,k} - v^{-1}_{j,f,l} w^{j}_{f-\phi,k} - \lambda_l \right) h_l \\ \displaystyle +\frac{1}{2} \sum_{\underline{i}jkfl\phi} \left(-2\left(w^{j}_{f-\phi,k}\right)^2 \left(\hat{r}^{(c)}_{\underline{i},j,f,l} \sigma^{(a)^{-1}}_{\underline{i},j,f} v^{-3}_{j,f,l} \right) + \left(w^{j}_{f-\phi,k}\right)^2 v^{-2}_{j,f,l} \right) h_l^2 \end{array} \right\} \tag{4.54}$$

The variational approximation of $\hat{Q}_L(\underline{\boldsymbol{h}}_L)$ will be considered by the exponential distribution

$$\hat{Q}_L(\underline{\boldsymbol{h}}_L \geq 0) = \prod_{l\in L} \frac{1}{u_l} exp\left(-\frac{h_l}{u_l} \right) \tag{4.55}$$

where

$$\int \hat{Q}_L(\underline{\boldsymbol{h}}_L)[log\,\hat{Q}_L(\underline{\boldsymbol{h}}_L)]d\underline{\boldsymbol{h}}_L = \sum_{l\in L}\int_0^{\infty}\frac{1}{u_l}exp\left(-\frac{h_l}{u_l}\right)\left(-log\,u_l-\frac{h_l}{u_l}\right)dh_l$$

$$= -\sum_{l\in L} log\,u_l + 1 \tag{4.56}$$

and

$$\int \hat{Q}_L(\underline{\boldsymbol{h}}_L) log Q_L(\underline{\boldsymbol{h}}_L)\; d\underline{\boldsymbol{h}}_L$$

$$= \int \hat{Q}_L(\underline{\boldsymbol{h}}_L)\left(\sum_{\underline{i},j,k,f,l,\phi}\left(\hat{r}^{(\boldsymbol{c})}_{\underline{i},j,f,l}\sigma^{(a)^{-1}}_{\underline{i},j,f}v^{-2}_{j,f,l}w^{j}_{f-\phi,k} - v^{-1}_{j,f,l}w^{j}_{f-\phi,k} - \lambda_l\right)h_l\right.$$

$$\left. +\frac{1}{2}\sum_{\underline{i},j,k,f,l,\phi}\left(-2\left(w^{j}_{f-\phi,k}\right)^2\left(\hat{r}^{(\boldsymbol{c})}_{\underline{i},j,f,l}\sigma^{(a)^{-1}}_{\underline{i},j,f}v^{-3}_{j,f,l}\right)+\left(w^{j}_{f-\phi,k}\right)^2 v^{-2}_{j,f,l}\right)h_l^2\right)d\underline{\boldsymbol{h}}_L$$

$$= E_{\hat{Q}_L(\underline{\boldsymbol{h}}_L)}\left[\sum_{\underline{i},j,k,f,l,\phi}\left(\hat{r}^{(\boldsymbol{c})}_{\underline{i},j,f,l}\sigma^{(a)^{-1}}_{\underline{i},j,f}v^{-2}_{j,f,l}w^{j}_{f-\phi,k} - v^{-1}_{j,f,l}w^{j}_{f-\phi,k} - \lambda_l\right)h_l\right.$$

$$\left. +\frac{1}{2}\sum_{\underline{i},j,k,f,l,\phi}\left(-2\left(w^{j}_{f-\phi,k}\right)^2\left(\hat{r}^{(\boldsymbol{c})}_{\underline{i},j,f,l}\sigma^{(a)^{-1}}_{\underline{i},j,f}v^{-3}_{j,f,l}\right)+\left(w^{j}_{f-\phi,k}\right)^2 v^{-2}_{j,f,l}\right)h_l^2\right] \tag{4.57}$$

where $E_{\hat{Q}_L(\underline{\boldsymbol{h}}_L)}$ is the expectation under the posterior $\hat{Q}_L(\underline{\boldsymbol{h}}_L)$

$$\int \hat{Q}_L(\underline{\boldsymbol{h}}_L) log Q_L(\underline{\boldsymbol{h}}_L)\; d\underline{\boldsymbol{h}}_L$$

$$= \left(\sum_{\underline{i},j,k,f,l,\phi}\left(\hat{r}^{(\boldsymbol{c})}_{\underline{i},j,f,l}\sigma^{(a)^{-1}}_{\underline{i},j,f}v^{-2}_{j,f,l}w^{j}_{f-\phi,k} - v^{-1}_{j,f,l}w^{j}_{f-\phi,k} - \lambda_l\right)\right)E_{\hat{Q}_L(\underline{\boldsymbol{h}}_L)}[h_l]$$

$$+\frac{1}{2}\sum_{\underline{i},j,k,f,l,\phi}\left(-2\left(w^{j}_{f-\phi,k}\right)^2\left(\hat{r}^{(\boldsymbol{c})}_{\underline{i},j,f,l}\sigma^{(a)^{-1}}_{\underline{i},j,f}v^{-3}_{j,f,l}\right)+\left(w^{j}_{f-\phi,k}\right)^2 v^{-2}_{j,f,l}\right)E_{\hat{Q}_L(\underline{\boldsymbol{h}}_L)}[h_l^2]$$

$$= \sum_{\underline{i},j,k,f,l,\phi}\left(\hat{r}^{(\boldsymbol{c})}_{\underline{i},j,f,l}\sigma^{(a)^{-1}}_{\underline{i},j,f}v^{-2}_{j,f,l}w^{j}_{f-\phi,k} - v^{-1}_{j,f,l}w^{j}_{f-\phi,k} - \lambda_l\right)u_l$$

$$+\frac{1}{2}\sum_{\underline{i},j,k,f,l,\phi}\left(-2\left(w_{f-\phi,k}^{j}\right)^{2}\left(\hat{r}_{\underline{i},j,f,l}^{(c)}\sigma_{\underline{i},j,f}^{(a)^{-1}}v_{j,f,l}^{-3}\right)+\left(w_{f-\phi,k}^{j}\right)^{2}v_{j,f,l}^{-2}\right)u_{l}{}^{2} \tag{4.58}$$

Thus

$$u_{l}\leftarrow arg\min_{u_{l}}\left(-\sum_{l\in L}logu_{l}+1+\sum_{\underline{i},j,k,f,l,\phi}\left(-\hat{r}_{\underline{i},j,f,l}^{(c)}\sigma_{\underline{i},j,f}^{(a)^{-1}}v_{j,f,l}^{-2}w_{f-\phi,k}^{j}+v_{j,f,l}^{-1}w_{f-\phi,k}^{j}+\lambda_{l}\right)u_{l}\right.$$
$$\left.+\frac{1}{2}\sum_{\underline{i},j,k,f,l,\phi}\left(2\left(w_{f-\phi,k}^{j}\right)^{2}\left(\hat{r}_{\underline{i},j,f,l}^{(c)}\sigma_{\underline{i},j,f}^{(a)^{-1}}v_{j,f,l}^{-3}\right)-\left(w_{f-\phi,k}^{j}\right)^{2}v_{j,f,l}^{-2}\right)u_{l}{}^{2}\right) \tag{4.59}$$

Let

$$b_{l}=\sum_{\underline{i},j,k,f,\phi}\left(-\hat{r}_{\underline{i},j,f,l}^{(c)}\sigma_{\underline{i},j,f}^{(a)^{-1}}v_{j,f,l}^{-2}w_{f-\phi,k}^{j}+v_{j,f,l}^{-1}w_{f-\phi,k}^{j}+\lambda_{l}\right) \tag{4.60}$$

and

$$\Theta_{l}=\sum_{\underline{i},j,k,f,\phi}\left(2\left(w_{f-\phi,k}^{j}\right)^{2}\left(\hat{r}_{\underline{i},j,f,l}^{(c)}\sigma_{\underline{i},j,f}^{(a)^{-1}}v_{j,f,l}^{-3}\right)-\left(w_{f-\phi,k}^{j}\right)^{2}v_{j,f,l}^{-2}\right) \tag{4.61}$$

Then it will leads to

$$u_{l}\leftarrow arg\min_{u_{l}}\left(\underline{\boldsymbol{b}}_{L}^{H}\underline{\boldsymbol{u}}+\frac{1}{2}\underline{\boldsymbol{u}}^{H}\widetilde{\boldsymbol{\Theta}}\underline{\boldsymbol{u}}-\sum_{l\in L}logu_{l}\right) \tag{4.62}$$

where $\widetilde{\boldsymbol{\Theta}}=diag(\Theta_{l})$. By using the nonnegative quadratic programming (NQP) [103]

$$G\left(\underline{\boldsymbol{u}},\underline{\tilde{\boldsymbol{u}}}\right)=\underline{\boldsymbol{b}}_{L}^{H}\underline{\boldsymbol{u}}+\frac{1}{2}\sum_{l\in L}\frac{\left(\widetilde{\boldsymbol{\Theta}}\underline{\tilde{\boldsymbol{u}}}\right)_{l}}{\tilde{u}_{l}}u_{l}^{2}-\sum_{l\in L}log\,u_{l} \tag{4.63}$$

Taking the derivative of $\mathrm{G}(\mathbf{u},\tilde{\mathbf{u}})$ in eqn. (4.63) with respect to u_{l} and setting it to zero yields

$$\frac{\left(\widetilde{\boldsymbol{\Theta}}\underline{\tilde{\boldsymbol{u}}}\right)_{l}}{\tilde{u}_{l}}u_{l}^{2}+b_{l}u_{l}-1=0 \tag{4.64}$$

which can be solved as follows

$$u_l \leftarrow u_l \frac{-b_l + \sqrt{b_l^2 + 4\frac{\left(\tilde{\boldsymbol{\Theta}}\underline{\boldsymbol{u}}\right)_l}{u_l}}}{2\left(\tilde{\boldsymbol{\Theta}}\underline{\boldsymbol{u}}\right)_l} \tag{4.65}$$

given that only the positive solution of eqn. (4.65) has been considered as we deal with nonnegative values only.

4.3.4 Components Reconstruction

The estimated STFT source spatial image $\hat{\boldsymbol{c}}_{j,f,n}$ can be reconstructed by using the multichannel Wiener filter that obtained by the minimum mean square error (MMSE) as in eqn. (4.18)

$$\hat{\boldsymbol{c}}_{j,f,n} = \sum_{k=1}^{K} \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} \Sigma_{j,f}^{(\boldsymbol{a})} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j} \Sigma_{f,n}^{(\boldsymbol{x})^{-1}} \boldsymbol{x}_{f,n} \tag{4.66}$$

The multichannel Wiener filter takes all the source spatial image components instead of the dominant one, as in the binary masking. Due to the linearity of the STFT, the inverse-STFT (with dual synthesis window [95]) can be used to transfer the source spatial image to time domain.

4.4 Initialization

The initialization is an essential part for the separation since the NMF and its variants are very sensitive to the initialization. One way to initialize the NMF is by using the SVD [104]. In this chapter, a new variant of SVD specially cater to initialize each *K*-wNTF2D sub-model will be proposed. It will be termed as the SVD two-dimensional deconvolution (SVD2D) and described as follows: Firstly, decompose the mixture X into K largest singular triplets, $X = \sum_{k=1}^{K} q_k C_k$, where q_k is the nonzero singular values of X, $C_k = \boldsymbol{u}_k \boldsymbol{v}_k^T$, and $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$ are the corresponding left and right singular vectors of X. Secondly, compute the SVD of C_k^+ (after decompose C_k into positive and negative components $C_k = C_k^+ - C_k^-$) in order to find the largest singular triplets. Let $W^i = \left\{w_{f,k}^{\tau=i,j}\right\}$ and $H^i = \left\{h_{k,n}^{\phi=i,j}\right\}$ represent fixing the i^{th}-slice of $\boldsymbol{W}$ and $\boldsymbol{H}$, respectively, i.e. setting $\tau = i$ in $W^{\boldsymbol{i}}$ and $\phi = i$ in H^i. The first column and row in W^0 and H^0 will be initialized by

using the largest singular triplet of X, and the rest by using the singular triplets of C_k^+. After initializing W^0 and H^0, the rest will be initialized in similar way.

Start $i = 1$, do the following:

Step 1: Compute $y_{f,n}^i = \sum_{j,k} w_{f-i,k}^{i-1,j} h_{k,n-i}^{i-1,j}$

Step 2: Apply SVD on Y^i to obtain $\sum_{k=1}^{K} q_{i,k} C_{i,k}$ where $C_{i,k} = \boldsymbol{u}_{i,k} \boldsymbol{v}_{i,k}^T$

Step 3: Apply SVD on $C_{i,k}^+$ to obtain $\sum_{l=1}^{L_{i,k}} q_{i,k,l} C_{i,k,l}$ where $C_{i,k,l} = \boldsymbol{u}_{i,k,l} \boldsymbol{v}_{i,k,l}^T$

Step 4: $W^i = [q_{i,1}\boldsymbol{u}_{i,1} \quad q_{i,2}q_{i,2,1}\boldsymbol{u}_{i,2,1} \quad q_{i,3}q_{i,3,1}\boldsymbol{u}_{i,3,1} \quad \cdots \quad q_{i,\mathrm{K}}q_{i,K,1}\boldsymbol{u}_{i,K,1}]$

and $H^i = [\boldsymbol{v}_{i,1} \quad \boldsymbol{v}_{i,2,1} \quad \boldsymbol{v}_{i,3,1} \quad \cdots \quad \boldsymbol{v}_{i,K,1}]^{\boldsymbol{T}}$

Step 5: $i \leftarrow i + 1$, repeat Steps 1 – 4

Stop when $i = max(\tau_{max} - 1, \phi_{max} - 1)$

The full-rank spatial covariance matrix will be initialized by using the hierarchical clustering. One simple method is to adopt Duong et al. [9].

Table 4.1 summarizes the main step of the proposed *K*-wNTF2D algorithm.

Table 4.1

Proposed algorithm *K-wNTF2D*

1. Initialize $\boldsymbol{W} = \{w_{f,k}^{\tau,j}\}$ and $\boldsymbol{H} = \{h_{k,n}^{\phi,j}\}$ with the proposed initialization method, $\Sigma_{j,f}^{(a)}$ with the hierarchical clustering approach, $\Sigma_f^{(b)}$ with random nonnegative diagonal matrix, and λ_p with a positive value.

2. E-step:

$$\hat{\Sigma}_{j,f,n}^{(c)} = \left(\mathrm{I} - \Sigma_{j,f,n}^{(c)} \Sigma_{f,n}^{(x)^{-1}}\right) \Sigma_{j,f,n}^{(c)}$$

$$\hat{R}_{j,f,n}^{(c)} = \hat{\boldsymbol{c}}_{j,f,n} \hat{\boldsymbol{c}}_{j,f,n}^{H} + \hat{\Sigma}_{j,f,n}^{(c)}$$

$$\hat{R}_{f}^{(b)} = \hat{\boldsymbol{b}}_{f,n} \hat{\boldsymbol{b}}_{f,n}^{H} + \left(\mathrm{I} - \Sigma_{f}^{(b)} \Sigma_{f,n}^{(x)^{-1}}\right) \Sigma_{f}^{(b)}$$

$$\hat{\boldsymbol{c}}_{j,f,n} = \Sigma_{j,f,n}^{(c)} \Sigma_{f,n}^{(x)^{-1}} \boldsymbol{x}_{f,n}$$

$$\hat{\boldsymbol{b}}_{f,n} = \Sigma_{f}^{(b)} \Sigma_{f,n}^{(x)^{-1}} \boldsymbol{x}_{f,n}$$

$$\Sigma_{f,n}^{(x)} = \sum_{j=1}^{J} \Sigma_{j,f,n}^{(c)} + \Sigma_{f}^{(b)}$$

$$\Sigma_{j,f,n}^{(c)} = v_{j,fn} \Sigma_{j,f}^{(a)}$$

$$v_{j,fn} = \sum_{k} \sum_{\tau} \sum_{\emptyset} \left(w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j}\right)$$

3. M-step:

$$\sigma_{\underline{i}',j',f'}^{(\boldsymbol{a})} \leftarrow \frac{1}{N} \sum_{n=1}^{N} \frac{\hat{r}_{\underline{i}',j',f',n}^{(c)}}{v_{j',f',n}}$$

$$w_{f',k'}^{\tau',j'} \leftarrow w_{f',k'}^{\tau',j'} \left(\frac{\sum_{\underline{i},\phi,n} \hat{r}_{\underline{i},j',f'+\phi,n}^{(c)} \sigma_{\underline{i},j',f'+\phi}^{(a)^{-1}} v_{j',f'+\phi,n}^{-2} h_{k',n-\tau'}^{\phi,j'}}{\sum_{\phi,n} v_{j',f'+\phi,n}^{-1} h_{k',n-\tau'}^{\phi,j'}} \right)$$

$$h_{k',n'}^{\phi',j'} \leftarrow h_{k',n'}^{\phi',j'}\left(\frac{\sum_{\underline{i},f\tau}\hat{r}_{\underline{i},j',f,n'+\tau}^{(c)}\sigma_{\underline{i},j',f}^{(a)^{-1}}v_{j',f,n'+\tau}^{-2}w_{f-\phi',k'}^{\tau,j'}}{\sum_{f\tau}v_{j',f,n'+\tau}^{-1}w_{f-\phi',k'}^{\tau,j'}+\lambda_{k',n'}^{\phi',j'}}\right)$$

$$\lambda_p = \begin{cases}\frac{1}{h_p} & \forall_p \in M \\ \frac{1}{u_p} & \forall_p \in L\end{cases}$$

$$u_p \leftarrow u_p \frac{-b_p + \sqrt{b_p^2 + 4\frac{\left(\tilde{\boldsymbol{\Theta}}\underline{\boldsymbol{u}}\right)_p}{u_p}}}{2\left(\tilde{\boldsymbol{\Theta}}\underline{\boldsymbol{u}}\right)_p}$$

$$b_p = \sum_{\underline{i},j,k,f,\phi}\left(-\hat{r}_{\underline{i},j,f,l}^{(\boldsymbol{c})}\sigma_{\underline{i},j,f}^{(a)^{-1}}v_{j,f,p}^{-2}w_{f-\phi,k}^{j} + v_{j,f,p}^{-1}w_{f-\phi,k}^{j} + \lambda_p\right)$$

$$\tilde{\boldsymbol{\Theta}} = diag(\Theta_l)$$

$$\Theta_l = \sum_{\underline{i},j,k,f,\phi}\left(2\left(w_{f-\phi,k}^{j}\right)^2\left(\hat{r}_{\underline{i},j,f,l}^{(\boldsymbol{c})}\sigma_{\underline{i},j,f}^{(a)^{-1}}v_{j,f,l}^{-3}\right) - \left(w_{f-\phi,k}^{j}\right)^2 v_{j,f,l}^{-2}\right)$$

4. **Normalize** $w_{f,k}^{\tau,j} = \dfrac{w_{f,k}^{\tau,j}}{\sqrt{\sum_{f,k,\tau}\left(w_{f,k}^{\tau,j}\right)^2}}$

5. Repeat E-step, M-step, and the normalization until convergence is achieved where rate of cost change is below a prescribed threshold, ψ.

6. Take inverse STFT with dual synthetic window to estimate $c_{i,j}(t)$.

4.5 Results and Discussions

4.5.1 Dataset

The following two datasets will be used in the experiments.

4.5.1.1 Dataset 1: This dataset is identical to the one used in the full-rank NMF of Arberet *et al.* algorithm [16]. This dataset consist of four groups depending on the distance between their microphones and the reverberation time (RT_{60}, which is the time taken by late echoes to decay by 60 dB). These are the 5 cm apart with 130 ms reverberation time group, 5 cm and 250 ms group, 1 m and 130 ms group, and 1 m 250 ms group. Each group consists of ten stereo mixtures, and each mixture has a length of 10 seconds, sampled at 16 kHz, and generated from three convolutive sources.

4.5.1.2 Dataset 2: This is an underdetermined speech and music mixtures development dataset of SiSEC 2013 [99]. This dataset consist of two groups. The first group is the live recording music group, which consists of dev1 and dev2 datasets, where each dataset has the with drum (wdrum) group; which consists of vocal and musical instrument with drum, and the without drum (nodrum) group; which consists of vocal and musical instruments without drum. The sources of this group are mixed in stereo mixture that has 1 m or 5 cm space between its microphones, and 250 ms reverberation time. The second group of this dataset is a simulated recording speech group, which consists of dev3 dataset, this dataset contains four females (female4) and four males (males4) that mixed in stereo mixture, with 5 cm or 50 cm distance between its microphones, and has a reverberation time of 130 ms or 380 ms. dev3 has three channels (left, right, and mono) and it has been reduced to two channels (left and right). Additionally, each mixture has duration of 10 s and sampled at 16 kHz.

4.5.2 Effects of Variable Sparsity versus Uniform Sparsity

In this subsection, the effects of the sparsity on the separation performance will be shown by considering a fixed uniform sparsity; $\lambda_{k,n}^{\phi,j} = \lambda = c$ all over the elements of $\boldsymbol{H}$, and the variable sparsity $\lambda_{k,n}^{\phi,j}$ for each element of $\boldsymbol{H}$. The fixed uniform sparsity is commonly used throughout the literature of matrix factorization. Each experiment will be run for different values of sparsity for the three sources that convolutively mixed in the stereo mixture that has 1 m space between its

microphones, 130 ms reverberation time, and with 16 kHz sampling frequency. The following parameters are set for the proposed algorithm: $K = 5$, $\tau_{max} = 10$, and $\phi_{max} = 1$. In order to focus on the sparsity effects only, an oracle initialization has been used.

Figure 4.2 shows the average SDR performance with respects to different values of sparsity. It is clear from Figure 4.2 that the variable sparsity gives the highest SDR performance. This is attributed to the fact that the proposed algorithm has a specific sparsity value for each element of $\boldsymbol{H}$, instead of constant value for the entire elements of $\boldsymbol{H}$ as in the case of uniform sparsity. It is seen that for variable sparsity, the average SDR is 4.5 dB higher than the best uniform sparsity (the value of constant λ that results in the highest SDR) $\lambda = 10$. Additionally, as the sparsity value increases (leading to over-sparseness) the SDR begins to decrease since many elements in $\boldsymbol{H}$ become very small and tends to zero. This resulted in switching off several parts of the spectrum in the estimated sources, as shown in Figure 4.3. In particular, the figure shows the spectrogram of one of the estimated sources for the case of variable sparsity, over-sparse, and the best uniform sparsity. It is visually apparent from the figure that the over-sparse and the best uniform sparsity have not fully recovered the original source. Many portions of the spectrum have been removed from the estimated source. While, the result from the variable sparsity has seen almost full recovery the original source, as it has been optimally tuned by the degree of sparseness over all the elements of $\boldsymbol{H}$.

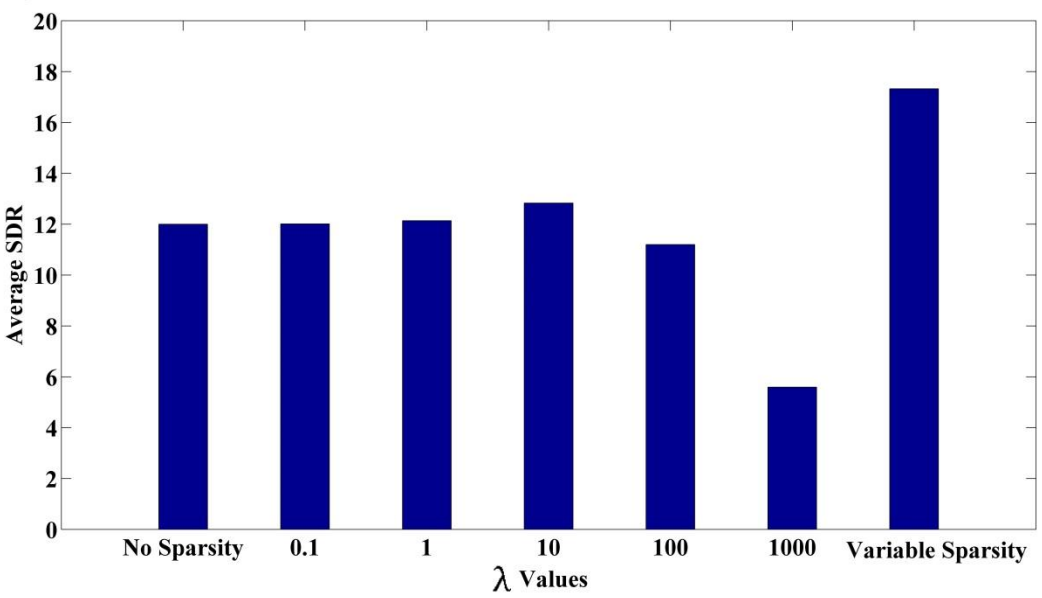


Figure 4.2: Average SDR w.r.t different sparsity values.

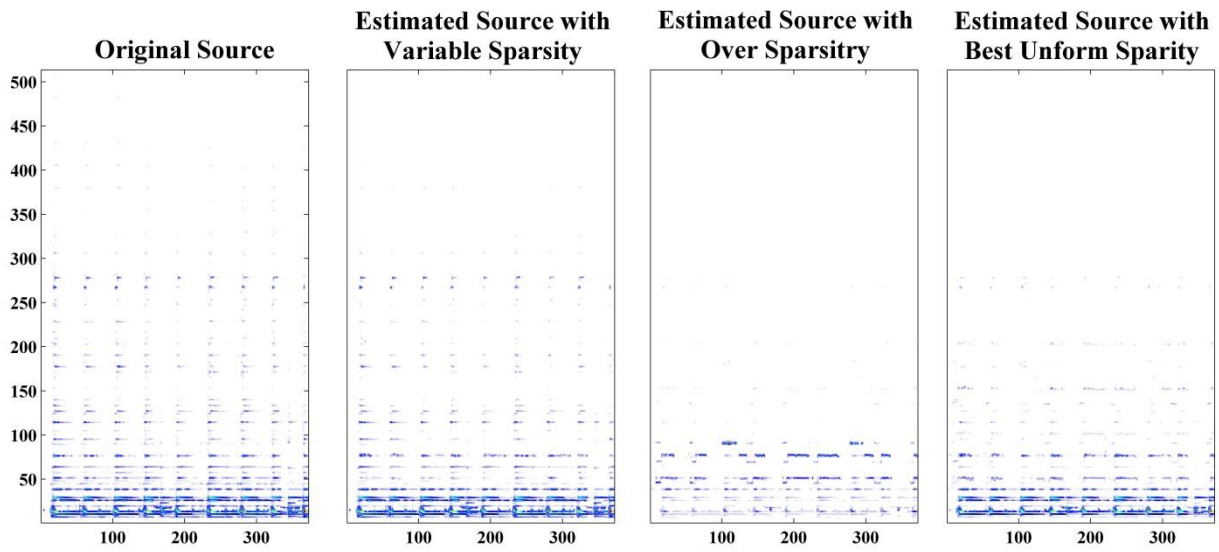


Figure 4.3: The effects of sparsity on the estimated source.

4.5.3 Separation Results

4.5.3.1 Results of Dataset 1: first of all the STFT window length was set to 1024 with 50% overlaps, 5 components per source were set for the full-rank NMF algorithm [16], 1 and 5 components per source were set for the proposed full-rank variable sparsity *K*-wNTF2D algorithm, different convolutive parameters were set for the proposed algorithm as tabulated in Table 4.2, and 50 iterations was set for both algorithms. Finally, for matter of comparison, the same initialization that used in Arberet *et al.* algorithm will be considered, where oracle initialization has been used to initialize $v_{j,f,n}$ and $\Sigma_{j,f}^{(\boldsymbol{a})}$.

To show the convergence of the proposed algorithm, the average cost functions (eqn. (4.14)) of the ten mixtures with different conditions (low and high reverberations time, and short and long distance between the microphones) are shown in Figure 4.4. It is noted that the speed of convergence (as measured by the gradient of the cost function) is fastest for the short microphone distance with low reverberation. As the microphone distance becomes larger and the level of reverberation increases, the speed tends to slow down. Nonetheless, all cost functions have converged to the steady state in less than 50 iterations.

Table 4.2

Convolutive parameters for mixtures 1 to 10

Mixture	τ_{max}	ϕ_{max}
1	1	1
2	2	1
3	2	1
4	3	1
5	3	1
6	4	1
7	4	1
8	8	1
9	10	1
10	10	1

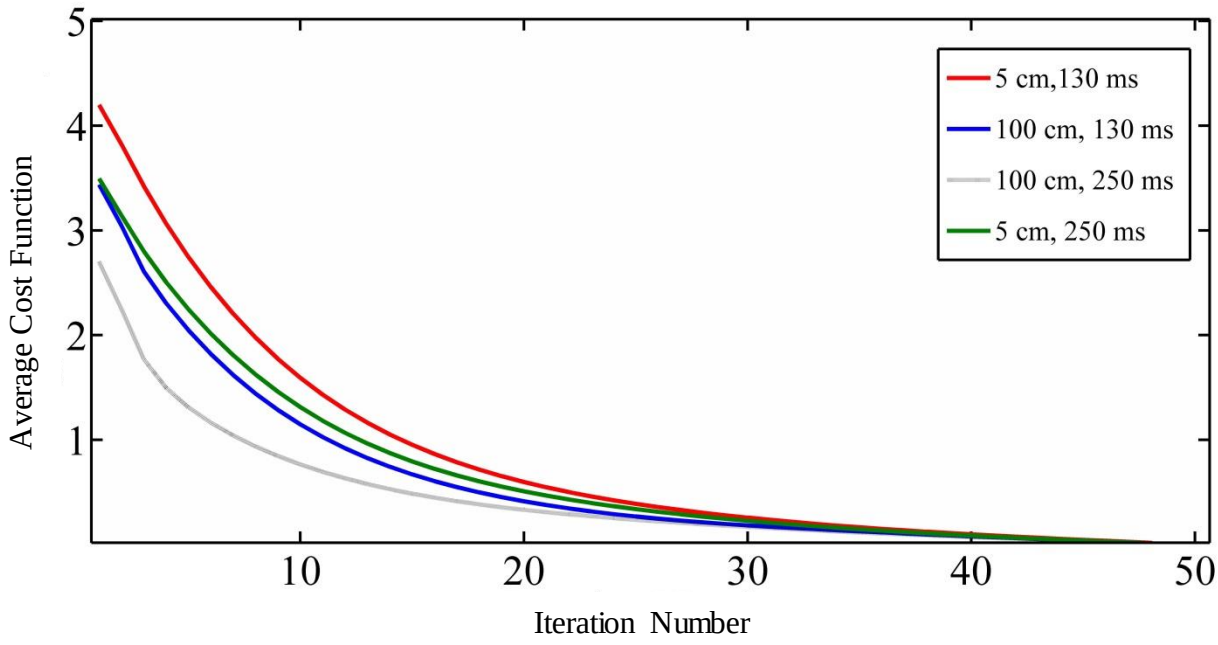


Figure 4.4: Average cost function for different conditions.

Furthermore, the SDRs of the full-rank NMF and the proposed algorithm are tabulated in Table 4.3. The table indicates that the proposed algorithm has better performance than the full-rank NMF since it has more powerful representation (using the *K*-wNTF2D), as well as the variable sparsity over all the elements of $\boldsymbol{H}$. The results for all the conditions can be summarized as follows: An achievement of 1.2 dB more for the low reverberation group, and at least 1 dB more on average for the high reverberations group. This is complemented by Figure 4.5. It shows that high SDR performance has been achieved for the 130ms reverberation for both 100cm and 5cm microphone separation. This case corresponds to the low reverberation environment. For the case of high reverberation, the proposed algorithm performs better with shorter microphone distance. As the distance between the microphones decreases, the signal at each microphone becomes more correlated with each other and therefore the channel covariance matrix $\Sigma_{j,f}^{(\boldsymbol{a})}$ tends to have some specific structure and hence reinforces the requirement of full-rank condition. While, as the separation between the microphones increases, the signal at each microphone becomes less correlated with each other. The effect is that each channel behaves independently and the channel

Table 4.3

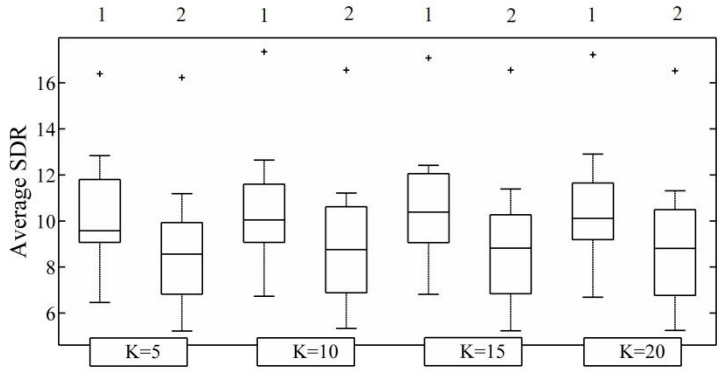
Average SDRs of the 10 mixtures with different conditions for the full-rank NMF and the proposed algorithm

Reverberation Time (ms)	**130**		**250**	
Microphone Distance (cm)	**5**	**100**	**5**	**100**
SDR of Full-Rank NMF K=5	9.1	10.2	8.8	9.6
SDR of the proposed algorithm K=1	6.6	7.8	6.5	7.3
SDR of the proposed algorithm K=5	10.3	11.4	9.8	10.4

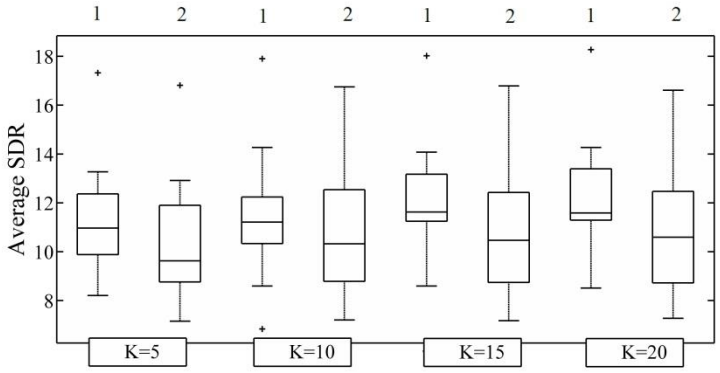
covariance matrix $\Sigma_{j,f}^{(\boldsymbol{a})}$ can be modelled by rank-1 structure. Thus as the separation between microphones become progressively small, this induces a complex structure to the channel

covariance which will benefit from the full-rank estimation procedure in the proposed algorithm. This is a clear indication that the proposed algorithm has outperformed the NMF for both the low and high reverberation time. In addition, to show the effects of the number of components on the proposed algorithm in comparison with the full-rank NMF the SDR of both algorithms for $K = 5, 10, 15$ and 20 have been also plotted in Figure 4.5. It shows the box plot for the ten mixtures with their median, maximum, and minimum SDR values for all the conditions. From the plot, it can be seen that the proposed algorithm gives higher median value in comparison with the full-rank NMF, for all the components under the different conditions, as the proposed algorithm is modeled to address the change in the time and frequency directions through the convolutive parameters (i.e. τ and ϕ) of the *K*-wNTF2D.

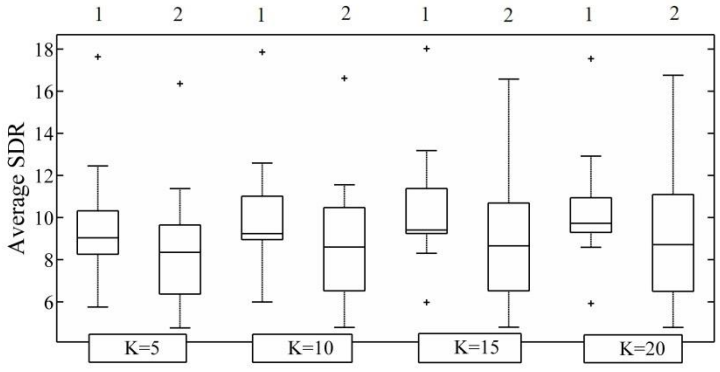
The spectrogram of one of the original sources, and its estimate by using the full-rank NMF and the full-rank variable sparsity *K*-wNTF2D are shown in Figure 4.6(a), (b), and (c), respectively. These figures show that the full-rank variable sparsity *K*-wNTF2D has successfully detected the pitch change of the source (as shown in the high frequency of its spectrogram), due to its two-dimensional deconvolution while the full-rank NMF failed to detect these changes. Furthermore, in order to show that $\boldsymbol{W}$ and $\boldsymbol{H}$ of the full-rank variable sparsity *K*-wNTF2D contain more information than those of the NMF, one component of the $\boldsymbol{W}$ and $\boldsymbol{H}$ matrices and its corresponding spectrogram for both the NMF and the full-rank variable sparsity *K*-wNTF2D are plotted in Figure 4.6(d) and 4. Figure 4.6(e), respectively. This indicates that both $\boldsymbol{W}$ and $\boldsymbol{H}$ have modelled the sources quite accurately. It is seen that $\boldsymbol{W}$ has successfully modelled the frequencies of the source especially in the high frequency region and $\boldsymbol{H}$ has shown a correct distribution in the time domain. On the separate hand, $\boldsymbol{W}$ and $\boldsymbol{H}$ of the NMF contain very little or virtually null information for these frequencies and their corresponding positions. Finally, Figure 4.7 shows another set of spectrograms which emphasize that the proposed full-rank variable sparsity *K*-wNTF2D algorithm has estimated the sources correctly in comparison with the full-rank NMF. The proposed algorithm has correctly detected the required number of frequency basis as well as their pitch change since the model has multiple frequency basis that convolve with the time–pitched weighted matrix in both time and frequency directions. While, the NMF fails to detect the required number of frequency basis since it contains too many unwanted frequency basis. In addition, it fails to detect the high frequency pitch change.



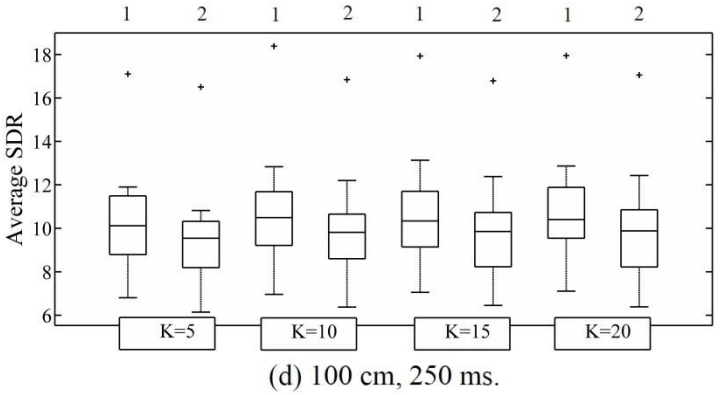
(a) 5 cm, 130 ms.



(b) 100 cm, 130 ms.



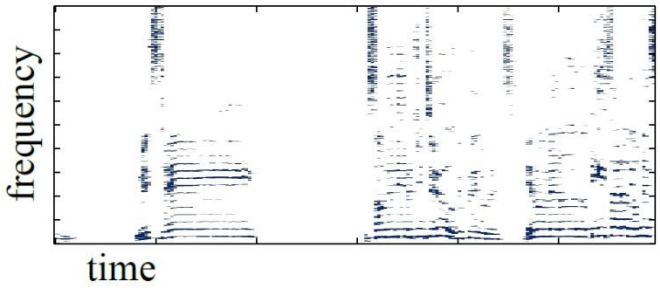
(c) 5 cm, 250 ms.



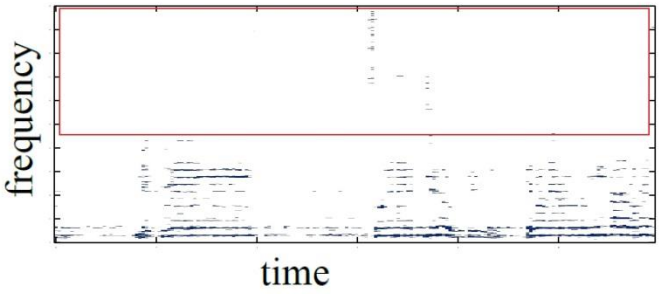
(d) 100 cm, 250 ms.

Figure 4.5: Box plot of the proposed algorithm (1) and the full rank NMF (2) with different components and different conditions.

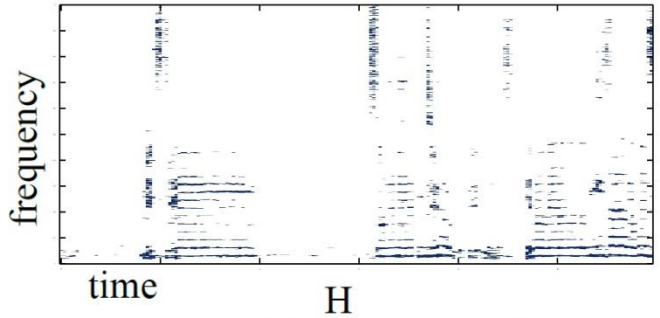
(a) spectrogram of the original source.



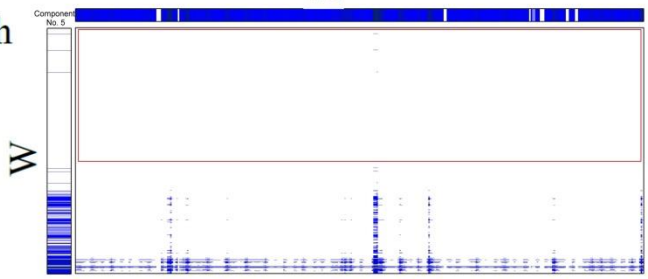
(b) spectrogram of the estimated source by using the full rank NMF.



(c) spectrogram of the estimated source by using the full rank *K*-wNTF2D.



(d) one component of W, and H, with their corresponding spectrogram for the full rank NMF.



(e) one component of W, and H, with their corresponding spectrogram for the full rank *K*-wNTF2D.

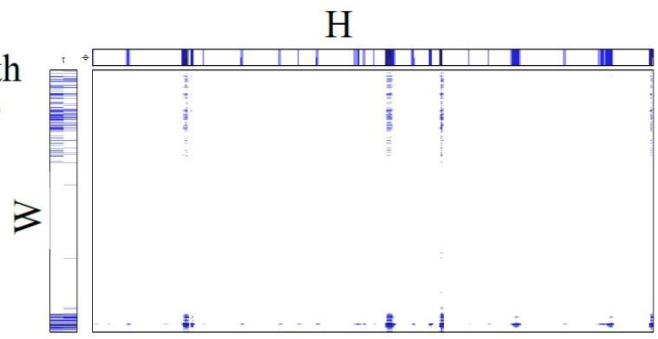


Figure 4.6: Comparison between the spectrogram of the Full Rank NMF and the proposed Full Rank *K*-wNTF2D.

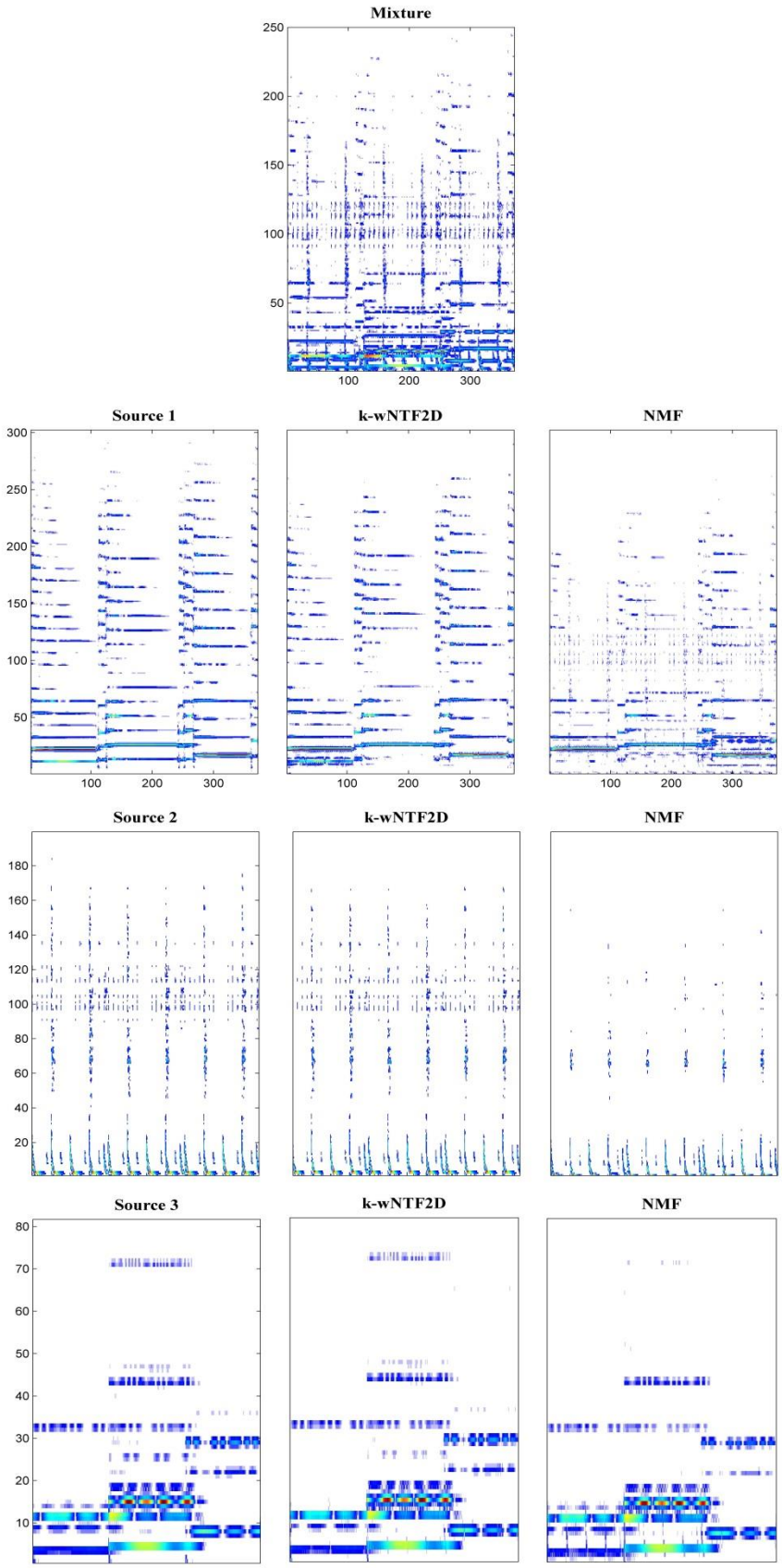


Figure 4.7: Spectrogram of the original and estimated sources by using the proposed Full Rank *K*-wNTF2D algorithm and the Full Rank NMF algorithm.

4.5.3.2 Results of Dataset 2: In this section, the proposed algorithm will be compared with Adiloglu *et al.* algorithm from the SiSEC'13 evaluation campaign for the tasks of underdetermined speech and music mixtures [105]; that used fully Bayesian source separation algorithm based on variational inference method [106], with the multi-level NMF model [52] as a source variance, and the time difference of arrival (TDOA) as an initialization method [107]. In the proposed algorithm, a different number of components and different convolutive parameters are set for each dataset, as tabulated in Tables 4.4, 4.5, 4.6, and 4.7. The STFT window length is set to 2048 with, 50% overlaps. The proposed initialization has been blindly initialized $v_{j,f,n}$ and $\Sigma_{j,f}^{(\boldsymbol{a})}$, respectively.

The average cost functions are shown in Figure 4.8. The figure indicates that all the cost functions converged to a low value within 10 iterations while Adiloglu *et al.* algorithm required about 250 iterations. Furthermore, it can be seen that the SDRs of the proposed algorithm for the music group (Table 4.4 and 4.5) on average is higher than the Adiloglu *et al.* algorithm. For clarity of comparison, the results are summarized as follows: An improvement of 2.65 dB is achieved for the 5 cm distance and 250 ms reverberation time datasets, and 2.6 dB for the 100 cm, 250 ms datasets. For the speech group (Table 4.6 and 4.7) on average an improvement of 2.5 dB is achieved for the 5 cm, 380 ms datasets, and 1.8 dB for the 50 cm, 380 ms datasets. Finally, an improvement of 0.3 dB is achieved for the 5 cm, 130 ms datasets, and approximately equal for the 50 cm, 130 ms datasets. From above, it can be concluded that the proposed algorithm outperforms Adiloglu *et al.* algorithm, especially for the case of high reverberation time. This is attributed to the proposed algorithm's ability to model the full-rank spatial covariance matrix (that modeled the spatial position and spread of the sources) instead of rank-1. Finally, Figure 4.9 shows the spectrogram of the estimated sources. It has indicated that the proposed algorithm has successfully estimated the sources to a high degree of accuracy. In particular, it is evident that all the low and high frequency components as well as the time-frequency patterns have been preserved in the estimated sources.

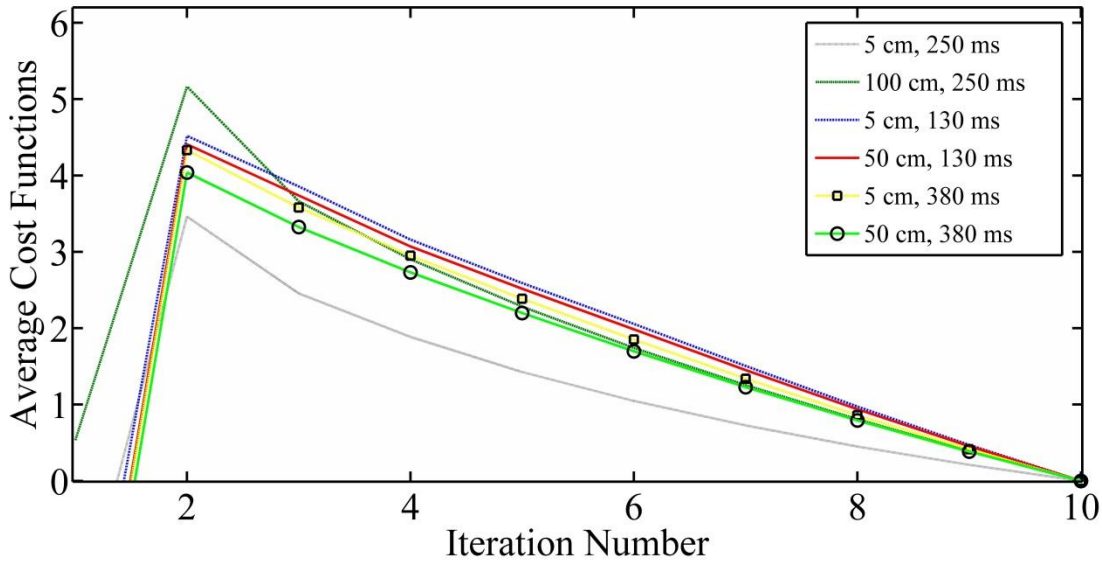


Figure 4.8: Average cost function for different conditions.

Table 4.4

SDRs of Adiloglu *et al.* and the proposed algorithm for dev. 1

SiSEC 2013: Dev. 1			**ndrums**		**wdrums**	
Reverberation Time (ms)			**250**		**250**	
Microphone Distance (cm)			**5**	**100**	**5**	**100**
Adiloglu *et al.* algorithm	SDR	s_1	-5.5	-0.6	7.0	2.4
		s_2	-1.2	-0.0	-0.1	3.0
		s_3	3.7	0.6	-0.5	-11.1
		Avg	-1.0	0.0	2.1	-1.9
GEM–MU based Variable Sparsity NTF $\tau_{max} = 0$ $\phi_{max} = 0$	K		**3**		**20**	
	SDR	s_1	0.5	2.1	5.7	6.7
		s_2	0.8	1.2	0.3	-1.1
		s_3	0.8	2.6	-0.8	0.1
		Avg	0.7	2.0	1.7	1.9
Proposed algorithm $\tau_{max} = 13$ $\phi_{max} = 2$	K		**3**		**20**	
	SDR	s_1	2.3	1.4	7.6	8.2
		s_2	0.9	2.6	0.9	0.5
		s_3	0.7	4.2	0.7	-0.1
		Avg	**1.3**	**2.7**	**3.1**	**2.9**

Table 4.5

SDRs of Adiloglu *et al.* and the proposed algorithm for dev. 2

SiSEC 2013: Dev. 2			**ndrums**		**wdrums**	
Reverberation Time (ms)			**250**		**250**	
Microphone Distance (cm)			**5**	**100**	**5**	**100**
Adiloglu *et al.* algorithm	*SDR*	s_1	1.8	4.7	3.7	4.8
		s_2	2.7	2.0	3.7	2.0
		s_3	-11.7	-3.9	3.7	2.7
		Avg	-2.4	0.9	3.7	3.2
GEM–MU based Variable Sparsity NTF	τ_{max}		**0**			
	ϕ_{max}		**0**			
	K		3		3	7
	SDR	s_1	9.6	6.7	1.0	1.9
		s_2	0.4	1.6	2.6	1.6
		s_3	-2.0	0.0	1.4	3.1
		Avg	2.7	2.8	2.7	2.2
Proposed algorithm	τ_{max}		**2**		**3**	
	ϕ_{max}		**2**		**9**	
	K		**3**		**3**	**7**
	SDR	s_1	10.5	7.6	3.5	2.9
		s_2	1.4	2.3	4.2	2.2
		s_3	0.8	0.7	5.4	4.6
		Avg	**4.2**	**3.5**	**4.4**	**3.2**

Table 4.6

SDRs of Adiloglu *et al.* and the proposed algorithm of dev. 3, for 5 cm, 380 ms case, and 50 cm, 380 ms case

SiSEC 2013: Dev. 3			**male4**		**female4**	
Reverberation Time (ms)			**380**		**380**	
Microphone Distance (cm)			**5**	**50**	**5**	**50**
Adiloglu *et al.* algorithm	SDR	$\boldsymbol{s_1}$	0.4	-1.7	0.2	-0.2
		$\boldsymbol{s_2}$	-2.6	-0.9	0.2	-1.0
		$\boldsymbol{s_3}$	-2.1	0.8	-3.1	-2.4
		$\boldsymbol{s_4}$	0.0	-0.4	-2.8	0.1
		Avg	-1.1	-0.6	-1.4	-0.9
GEM–MU based Variable Sparsity NTF $\tau_{max} = 0$ $\phi_{max} = 0$ $K = 10$	SDR	$\boldsymbol{s_1}$	0.7	0.2	0.3	0.3
		$\boldsymbol{s_2}$	0.8	0.6	0.8	0.4
		$\boldsymbol{s_3}$	0.2	1.1	-0.9	0.2
		$\boldsymbol{s_4}$	1.1	-0.1	0.2	0.5
		Avg	0.7	0.5	0.1	0.4
Proposed algorithm $\tau_{max} = 10$ $\phi_{max} = 20$ $K = 10$	SDR	$\boldsymbol{s_1}$	1.3	0.6	1.9	0.8
		$\boldsymbol{s_2}$	1.2	1.1	0.8	0.7
		$\boldsymbol{s_3}$	1.3	1.8	1.3	0.1
		$\boldsymbol{s_4}$	1.3	0.7	0.9	1.8
		Avg	**1.3**	**1.1**	**1.2**	**0.9**

Table 4.7

SDRs of Adiloglu *et al.* and the proposed algorithm of dev. 3, for 5 cm, 130 ms case, and 50 cm, 130 ms case

SiSEC 2013: Dev. 3			**male4**		**female4**	
Reverberation Time (ms)			**130**		**130**	
Microphone Distance (cm)			**5**	**50**	**5**	**50**
Adiloglu *et al.* algorithm	SDR	$\boldsymbol{s_1}$	-2.6	-2.1	-0.0	-1.2
		$\boldsymbol{s_2}$	-0.2	2.6	-0.9	0.6
		$\boldsymbol{s_3}$	1.5	0.8	0.4	1.4
		$\boldsymbol{s_4}$	5.2	3.9	4.1	4.4
		Avg	1.0	1.3	0.9	1.3
GEM–MU based Variable Sparsity NTF K=10	τ_{max}		0			
	ϕ_{max}		0			
	SDR	$\boldsymbol{s_1}$	0.5	-0.5	-0.3	-2.8
		$\boldsymbol{s_2}$	-0.7	0.7	1.3	0.1
		$\boldsymbol{s_3}$	0.6	0.4	0.3	1.4
		$\boldsymbol{s_4}$	1.0	-0.8	1.0	0.9
		Avg	0.4	-0.0	0.6	-0.1
Proposed algorithm K=10	τ_{max}		**10**			
	ϕ_{max}		**50**		**60**	
	SDR	$\boldsymbol{s_1}$	1.2	0.5	1.5	0.8
		$\boldsymbol{s_2}$	1.1	2.6	1.6	0.9
		$\boldsymbol{s_3}$	1.4	0.9	1.0	2.7
		$\boldsymbol{s_4}$	1.2	1.2	1.1	0.8
		Avg	**1.2**	**1.3**	**1.2**	**1.3**

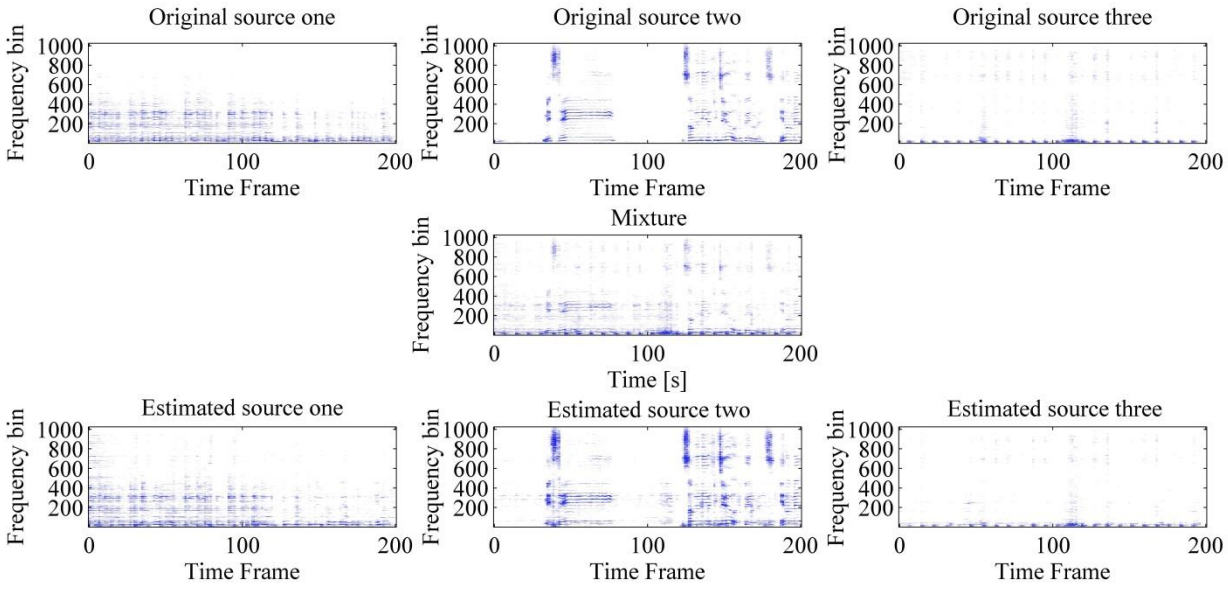


Figure 4.9: Spectrogram of one of the mixtures and its original and estimated sources.

4.6 Summary

In this chapter, the K models of the weighted NTF2D have been combined with the variable sparsity to propose a novel algorithm for the underdetermined multichannel audio source separation. It has been shown that using the GEM-MU algorithm as a platform for the proposed algorithm enabled the joint estimation of the parameters and sources, and preserving the non-negativity constraints of the proposed model. Also, a tractable approach that adapts each sparse parameter for every temporal code in the NTF2D has been provided through the variable sparse parameters that derived from the Gibbs distribution. Furthermore, the NTF2D has been efficiently initialized by the proposed initialization approach. The full-rank NMF and NTF algorithms, and a recent algorithm based on variational inference multi-level NMF model with TDOA initialization have been outperformed by the proposed algorithm. Additionally, it has been shown that using the full-rank spatial covariance matrix instead of rank 1 has enabled the proposed algorithm to maintain its high level performance in high reverberation environment. Finally, the proposed algorithm fast converged to the steady state in less than 10 iterations.

CHAPTER 5
INFORMED SOURCE SEPARATION BASED TWO DIMENSIONAL MATRIX FACTORIZATION TECHNIQUES

In this chapter two algorithms are proposed for informed source separation, i.e., an exemplar-based algorithm and a semi-exemplar based algorithm. The semi-exemplar based algorithm takes advantage of the NMF2D to describe the temporal and spectral changes, and the number of spectral components of targeted speech signal. The description is carried out indirectly by the exemplar: Firstly the exemplar is used to emulate the targeted speech signal, then the parameters of the NMF2D are optimized inline with the exemplar, finally these parameters are used to separate the targeted speech signal. Additionally the spectral and temporal tensors generated from the exemplar will be used to initialize the tensors of the targeted speech signals. In the full exemplar-based algorithm, the separating algorithm describes the targeted speech signal in the same way as in the semi-exemplar based algorithm; however the separation is carried out based on the two dimensional nonnegative matrix partial co-factorization (2DNMPCF) that jointly factorizes the exemplar's and the mixture's spectrogram. In addition, the chapter proposes an artificial stereo channel. It introduces diversity to the mixing channel by augmenting the dimensionality of the mixing matrix, increases its matrix rank and thus reduces the ambiguity associated with estimating several sources given only a single observation of the mixture signal. The proposed algorithms with artificial stereo channel have been adapted under the hybrid framework that combines the generalized EM algorithm with multiplicative update. The algorithms lead to fast and stable convergence, and ensure the non-negativity constraints are satisfied. Additionally, the adaptive sparsity is imposed on each sparse parameter in the 2DNMPCF. Experimental results have shown the effectiveness of the proposed algorithms in comparison with other algorithms.

This chapter is organized as follows: Section 5.1 introduces the proposed model. Section 5.2 is dedicated for the problem formulation, where the mixture model with pseudo-stereo channel and the maximum a Posterior probability (MAP) model will be formulated. The proposed semi-exemplar based algorithm and the proposed exemplar based algorithm will be explained and derived in Section 5.3. Experimental results and discussions of these results will be shown in Section 5.4. Section 5.5 proposed a multistage of the exemplar based algorithm. Finally, Section 5.6 draws the conclusions.

5.1 Introduction

Blind source separation (BSS) [19, 67, 108] is ill-posed problem that cannot be solved totally blind, i.e., a certain assumptions has to be made to solve them, e.g. the number of sources, how the sources are mixed, the location of the sources with respect to the microphones, and the channel type. However, even with these assumptions the BSS did not fully achieve the required performance. Therefore, researchers moved from blind to informed audio source separation in order to achieve higher performance that the BSS cannot reach, where, researchers seek an aid from an external source in addition to the mixture signal as side information to enhance the separation performance. Such as the user mimic the targeted signal in the mixture by singing [54], by humming [81], or by dubs the dialog in films [82] in order to separate the targeted signal. Another examples is by using additional audio references such as using the multitrack cover version of the same song [56, 83-85] or using several international versions of the same movie [55]. Additionally using the text to mimic the targeted speech signal [86].

In this chapter, exemplar signal from the text associated with the mixture will be generated by using a speech synthesizer or human speakers. The approach is essentially belonging to the category of text informed source separation [86]. The text informed source separation [86] used the NMPCF [109, 110] based on excitation-filter channel speech model and the structural Gaussian Scaled Mixture Model (GSMM). In the current chapter, two algorithms will be proposed the exemplar-based algorithm and the semi-exemplar based algorithm. In the exemplar-based algorithm the exemplar will be used to optimize the parameters and initialize the tensors of the proposed 2DNMPCF that will carry out the separation. The proposed 2DNMPCF will be used as it has the ability to describe the pitch and temporal changes of the signal through ϕ and τ, in addition to the frequency basis (as in NMPCF) through K. Therefore, the proposed 2DNMPCF is more powerful than the NMPCF. The idea of using the co-factorization technique is to simultaneously factorize the mixture and the exemplar signals in order to guide the separation. In the case of the semi-exemplar based algorithm, the exemplar will be used to optimize the parameters and initialize the tensors of the NMF2D [25] which alone will be used to carry out the separation. The difference between the semi-exemplar based algorithm and the exemplar based algorithm is that the former algorithm will guide the separation for the first iteration only (i.e., to give the correct start) by initializing its tensors through the exemplar signal, while in the latter algorithm the exemplar signal is used to initialize as well as to guide the separation process for every iteration via the 2DNMPCF until it converges to the steady state. For faster convergence both algorithms are adapted under the

GEM-MU model [80]. Furthermore, the adaptive sparsity will be optimized in the proposed algorithms instead of the uniform fixed sparsity. As the speech source changes rapidly over time, then assigning a uniform fixed sparsity will leads to either too many ineffective temporal codes (under-sparseness), or too many temporal codes set to zero (over-sparseness), while the adaptive sparsity will optimize the sparsity for each individual temporal code. Finally for better performance the pseudo-stereo channel [111] will be adapted using the GEM-MU algorithm. The pseudo-stereo mixture allows us to extract the temporal feature of the mixed signal to enable the estimation of the mixing process and relieving the ill-posed problem of single-channel source separation. The single-channel source separation is a highly underdetermined problem where only a single channel recording is available to estimate more than one source signals. Hence, given only the mixed signal, potentially innumerable number of solutions exists for the source signals. Thus the pseudo-stereo channel creates an artificial mixed signal to increase the dimensionality of the mixing matrix and reduce the ambiguity in estimating source signals. Figure 5.1 shows the high level presentation of the proposed algorithms (see Section 5.3 for the details).

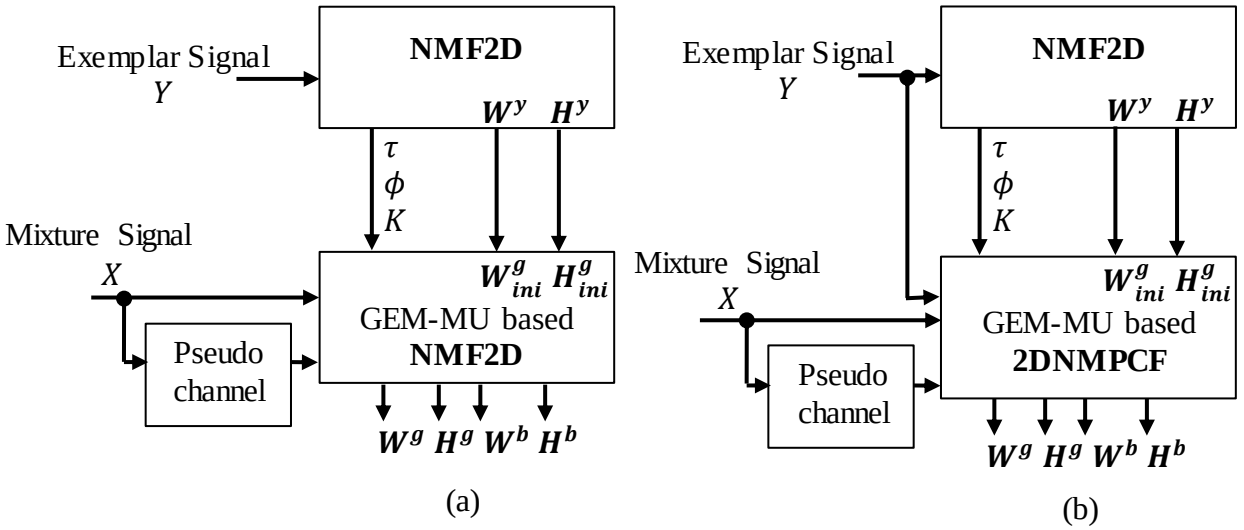


Figure 5.1: High level presentation of (a) the semi-exemplar based algorithm, and (b) the exemplar based algorithm.

5.2 Problem Formulation

5.2.1 Pseudo-Stereo Channel

Consider the underdetermined single channel mixture, namely:

$$\tilde{x}_1(t) = \tilde{g}(t) + \tilde{b}(t) + \tilde{n}(t) \tag{5.1a}$$

$$= \sum_{\tau=0}^{L-1} \tilde{a}_1(\tau)\tilde{g}(t-\tau) + \sum_{\tau=0}^{L-1} \tilde{a}_2(\tau)\tilde{b}(t-\tau) + \tilde{n}(t) \tag{5.1b}$$

where $\tilde{x}_1(t)$ is the sampled mixture signal, $\tilde{g}(t)$ the sampled speech signal, $\tilde{b}(t)$ is the sampled background signal (which will take as either a music or effects (fx)), $\tilde{n}(t)$ is some additive noise, for $(t = 1, \ldots, T)$, $\tilde{a}_1(\tau)$ and $\tilde{a}_2(\tau)$ are the finite-impulse response of some (causal) filters.

As the separation performance is enhanced when the number of channels is greater than or equal to the number of sources, and as the single channel is heavily underdetermined which creates ambiguity when estimating the sources, then a pseudo-stereo channel model [111] will be formulated in this chapter, i.e., a pseudo microphone (virtual channel) will be formulated by weighting and time-shifting the single channel mixture $\tilde{x}_1(\mathrm{t})$ as follows

$$\tilde{x}_2(t) = \frac{\tilde{x}_1(t) + \gamma\tilde{x}_1(t-\delta)}{1+|\gamma|} \tag{5.2}$$

where $\gamma \in \mathcal{R}$ is the weight parameter, and δ is the time delay between $\tilde{x}_2$ and $\tilde{x}_1$. The mixed signals in eqn. (5.1a) and eqn. (5.2) are termed as "pseudo-stereo" since they have an artificial resemblance of a stereo signal. Substituting eqn. (5.1b) in eqn. (5.2) will leads to

$$\tilde{x}_2(t) = \frac{1}{1+|\gamma|}\left(\sum_{\tau=0}^{L-1} \tilde{a}_1(\tau)\big(\tilde{g}(t-\tau) + \gamma\tilde{g}(t-\tau-\delta)\big) + \sum_{\tau=0}^{L-1} \tilde{a}_2(\tau)\left(\tilde{b}(t-\tau) + \gamma\tilde{b}(t-\tau-\delta)\right) + (\tilde{n}(t) + \gamma\tilde{n}(t-\delta))\right) \tag{5.3}$$

By assuming that the mixing channel is time-invariant and by considering the narrowband approximation then the short-time Fourier transform (STFT) of eqn. (5.1b) can be expressed as

$$x_{1,f,n} = a_{1,f}g_{f,n} + a_{2,f}b_{f,n} + n_{f,n} \tag{5.4}$$

Similarly, the STFT of eqn. (5.3)

$$x_{2,f,n} = \left(a_{1,f}\frac{1}{1+|\gamma|}\left(g_{f,n} + \gamma g_{f,n}e^{-iw\delta}\right) + a_{2,f}\frac{1}{1+|\gamma|}\left(b_{f,n} + \gamma b_{f,n}e^{-iw\delta}\right)\right.$$

$$+\frac{1}{1+|\gamma|}\left(n_{f,n}+\gamma e^{-iw\delta}n_{f,n}\right)\Bigg)$$

$$=\Bigg(a_{1,f}\frac{1}{1+|\gamma|}g_{f,n}\left(1+\gamma e^{-iw\delta}\right)+a_{2,f}\frac{1}{1+|\gamma|}b_{f,n}\left(1+\gamma e^{-iw\delta}\right)$$

$$+\frac{1}{1+|\gamma|}n_{f,n}\left(1+\gamma e^{-iw\delta}\right)\Bigg)$$

$$=\left(a_{1,f}\frac{1+\gamma e^{-iw\delta}}{1+|\gamma|}g_{f,n}+a_{2,f}\frac{1+\gamma e^{-iw\delta}}{1+|\gamma|}b_{f,n}+n_{f,n}\frac{1+\gamma e^{-iw\delta}}{1+|\gamma|}\right) \quad (5.5a)$$

Let $\acute{a}_{1,f}=a_{1,f}\frac{1+\gamma e^{-iw\delta}}{1+|\gamma|}$, $\acute{a}_{2,f}=a_{2,f}\frac{1+\gamma e^{-iw\delta}}{1+|\gamma|}$, and $\acute{n}_{f,n}=n_{f,n}\frac{1+\gamma e^{-iw\delta}}{1+|\gamma|}$, then eqn. (5.5a) can be rewritten as follows

$$x_{2,f,n}=\acute{a}_{1,f}g_{f,n}+\acute{a}_{2,f}b_{f,n}+\acute{n}_{f,n} \quad (5.5b)$$

Eqn. (5.4) and eqn. (5.5b) can be written in matrix form

$$X_f=A_fS_f+N_f \quad (5.6)$$

where

$X_f=\left[x_{i,f,n}\right]_f=\begin{bmatrix}x_{1,f,1} & x_{1,f,2}\dots & x_{1,f,N}\\ x_{2,f,1} & x_{2,f,2}\dots & x_{2,f,N}\end{bmatrix}_f\in\mathbb{C}^{2\times N}$, $i=1,2$ is the channel index, $f=1,\dots,F$ is the frequency bin index,

$A_f=\begin{bmatrix}a_{1,f} & a_{2,f}\\ \acute{a}_{1,f} & \acute{a}_{2,f}\end{bmatrix}_f\in\mathbb{C}^{2\times 2}$,

$S_f=\left[s_{j,f,n}\right]_f=\begin{bmatrix}g_{f,1} & g_{f,2}\dots & g_{f,N}\\ b_{f,1} & b_{f,2}\dots & b_{f,N}\end{bmatrix}_f\in\mathbb{C}^{2\times N}$, $j=1,2$ is the source index, and

$N_f=\left[\bar{n}_{i,f,n}\right]_f=\begin{bmatrix}n_{f,1} & n_{f,2}\dots & n_{f,N}\\ \acute{n}_{f,1} & \acute{n}_{f,2}\dots & \acute{n}_{f,N}\end{bmatrix}_f\in\mathbb{C}^{2\times N}$

The NMF2D has the ability to specify the temporal and spectral changes of the targeted speech signal through its convolutive parameters (τ and ϕ), and the number of frequency basis (K) of the targeted speech signal. If the NMF used here it will be able to describe the number of frequency

basis only. Therefore, the NMF2D with multiple components will be considered as the spectral variance model instead of the NMF spectral model [49]. Thus, each source in the STFT can be expressed by K complex-valued latent components, i.e., $g_{f,n} = \sum_{k=1}^{K_g} c_{k,f,n}^{g}$, and $b_{f,n} = \sum_{k=1}^{K_b} c_{k,f,n}^{b}$ and can be modeled as realization of proper complex zero-mean variables:

$$c_{k,f,n}^{g} \sim \mathcal{N}_c\left(0, \sigma_{k,f,n}^{g^2}\right) = \mathcal{N}_c\left(0, \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k,\tau}^{g} h_{k,n-\tau,\phi}^{g}\right) \tag{5.7a}$$

$$c_{k,f,n}^{b} \sim \mathcal{N}_c\left(0, \sigma_{k,f,n}^{b^2}\right) = \mathcal{N}_c\left(0, \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k,\tau}^{b} h_{k,n-\tau,\phi}^{b}\right) \tag{5.7b}$$

where $\mathcal{N}_c(\mu, \Sigma)$ is proper complex Gaussian distribution [94], $w_{f,k,\tau}^{g}$ and $w_{f,k,\tau}^{b}$ represent the spectral basis of the speech and background sources, respectively, and $h_{k,n,\phi}^{g}$ and $h_{k,n,\phi}^{b}$ represent the temporal code for each spectral basis element of the speech and background sources, respectively, for $f = 1, \dots, F, n = 1, \dots, N, k = 1, \dots K$.

5.2.2 Maximum A Posterior Probability (MAP) Estimation

The maximum *a posteriori* (MAP) probability will be used as the criterion for optimization. The noise $\bar{n}_{i,fn}$ is assumed to be stationary and spatially uncorrelated, i.e.

$$\bar{n}_{i,fn} \sim \mathcal{N}_c\left(0, \sigma_{i,f}^{\bar{n}\,2}\right), \quad \text{and} \quad \Sigma_{\bar{n},f} = diag\left[\sigma_{i,f}^{\bar{n}\,2}\right] \tag{5.8}$$

Let $\boldsymbol{C} = \left\{\left\{c_{k,f,n}^{g}\right\}, \left\{c_{k,f,n}^{b}\right\}\right\}_{k,f,n}$ be the latent variables, and $\boldsymbol{\theta} = \{\boldsymbol{A}, \boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}_{\bar{\boldsymbol{n}}}\}$ as the parameters of the model where $\boldsymbol{W} = \{\boldsymbol{W}^{g}, \boldsymbol{W}^{b}\}$, $\boldsymbol{H} = \{\boldsymbol{H}^{g}, \boldsymbol{H}^{b}\}$, $\boldsymbol{\Lambda} = \{\boldsymbol{\Lambda}^{g}, \boldsymbol{\Lambda}^{b}\}$ with $\boldsymbol{W}^{g} = \left\{w_{f,k,\tau}^{g}\right\}_{f,k,\tau}$, $\boldsymbol{W}^{b} = \left\{w_{f,k,\tau}^{b}\right\}_{f,k,\tau}$, $\boldsymbol{H}^{g} = \left\{h_{k,n,\phi}^{g}\right\}_{k,n,\phi}$, $\boldsymbol{H}^{b} = \left\{h_{k,n,\phi}^{b}\right\}_{k,n,\phi}$, $\boldsymbol{\Lambda}^{g} = \left\{\lambda_{k,n,\phi}^{g}\right\}_{k,n,\phi}$, $\boldsymbol{\Lambda}^{b} = \left\{\lambda_{k,n,\phi}^{b}\right\}_{k,n,\phi}$. The tensor $\boldsymbol{\Lambda}$ contains the sparsity terms for $\boldsymbol{H}$. The estimation of model parameters and latent variables will proceed via the posterior probability:

$$\widehat{\boldsymbol{\theta}}_{MAP} = arg\max_{\theta}\ log\, p(\boldsymbol{\theta}|\boldsymbol{X})$$

where

$$log\, p(\boldsymbol{\theta}|\boldsymbol{X}) \geq \int Q(\boldsymbol{C})\, log\left[\frac{p(\boldsymbol{C},\boldsymbol{\theta}|\boldsymbol{X})}{Q(\boldsymbol{C})}\right] d\mathbf{C} \tag{5.9}$$

for any distribution $Q(\boldsymbol{C})$. Defining $F(Q(\boldsymbol{C}),\boldsymbol{\theta}) = \int Q(\boldsymbol{C})\, log\left[\frac{p(\boldsymbol{C},\boldsymbol{\theta}|\boldsymbol{X})}{Q(\boldsymbol{C})}\right] d\boldsymbol{C}$, then the E-step consists of determining $Q(\boldsymbol{C})$ that maximizes $F(Q(\boldsymbol{C}),\boldsymbol{\theta})$ where the optimal $Q(\boldsymbol{C})$ is given by $Q^*(\boldsymbol{C}) = P(\boldsymbol{C}|\boldsymbol{X},\boldsymbol{\theta}')$ for the current model $\boldsymbol{\theta}'$. The M-step consists of maximizing $F(Q^*(\boldsymbol{C}),\boldsymbol{\theta})$ with respect to the model $\boldsymbol{\theta}$ when $Q(\boldsymbol{C})$ is fixed at $Q^*(\boldsymbol{C})$ i.e. $\boldsymbol{\theta}^* = \underset{\theta}{arg\,max} \int Q^*(\boldsymbol{C})\, log\, p(\boldsymbol{C},\boldsymbol{\theta}|\boldsymbol{X})\, d\boldsymbol{C}$. The posterior probability is given by

$$p(\boldsymbol{C},\boldsymbol{\theta}|\boldsymbol{X}) = \frac{p(\boldsymbol{X},\boldsymbol{C}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{X})} \propto p(\boldsymbol{X}|\boldsymbol{C},\boldsymbol{\theta})p(\boldsymbol{C}|\boldsymbol{\theta})P(\boldsymbol{\theta}) \tag{5.10}$$

5.3 Proposed Exemplar and Semi-Exemplar Algorithms

The GEM-MU [80] will be used as the platform for deriving the proposed algorithms. The source power spectrogram posterior estimates ($\hat{p}_{j,fn}$) (see eqn. (5.12)), the mixing parameters, and the noise covariance will be estimated in the E-step of the EM algorithm, while $\boldsymbol{W}$ and $\boldsymbol{H}$ will be estimated in the M-step of the EM algorithm by using the MU algorithm with adaptive sparsity NMF2D.

First of all, the common part between the two proposed algorithms will be derived, and then each one will be derived separately.

5.3.1 E-Step: Conditional Expectations of Natural Statistics

In the E-step, the complete data $\{\boldsymbol{X},\boldsymbol{C}\}$ and its pdfs $p(\boldsymbol{X},\boldsymbol{C}|\boldsymbol{\theta})$ form an exponential family. The complete data log-likelihood is given by

$$\begin{aligned} -log\, p(\boldsymbol{C},\boldsymbol{\theta}|\boldsymbol{X}) &= -log\, p(\boldsymbol{X}|\boldsymbol{C},\boldsymbol{\theta}) - log\, p(\boldsymbol{C}|\boldsymbol{\theta}) - log\, P(\boldsymbol{\theta}) \\ &\overset{c}{=} \sum_{f,n}\left[log\left|\Sigma_{\bar{n},f}\right| + \left(\boldsymbol{x}_{fn} - A_f\boldsymbol{s}_{fn}\right)^H \Sigma_{\bar{n},f}^{-1}\left(\boldsymbol{x}_{fn} - A_f\boldsymbol{s}_{fn}\right)\right] \\ &\quad + \sum_{k=1}^{K_g}\sum_{f,n}\left[log\left(\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k,\tau}^{g} h_{k,n-\tau,\phi}^{g}\right) + \frac{\left|c_{k,f,n}^{g}\right|^2}{\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k,\tau}^{g} h_{k,n-\tau,\phi}^{g}}\right] \end{aligned}$$

$$+\sum_{k=1}^{K_b}\sum_{f,n}\left[log\left(\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k,\tau}^{b} h_{k,n-\tau,\phi}^{b}\right)+\frac{\left|c_{k,f,n}^{b}\right|^2}{\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k,\tau}^{b} h_{k,n-\tau,\phi}^{b}}\right]$$

$$-log\,p(A_f)-log\,p(\Sigma_{\bar{n},f})-log\,p(\boldsymbol{W})-log\,p(\boldsymbol{H}|\boldsymbol{\Lambda})$$

$$=\sum_{f,n}\left[log|\Sigma_{\bar{n},f}|+\sum_{k=1}^{K_g} log\left(\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k,\tau}^{g} h_{k,n-\tau,\phi}^{g}\right)+\sum_{k=1}^{K_g}\frac{\left|c_{k,f,n}^{g}\right|^2}{\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k,\tau}^{g} h_{k,n-\tau,\phi}^{g}}\right.$$

$$\left.+\sum_{k=1}^{K_b} log\left(\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k,\tau}^{b} h_{k,n-\tau,\phi}^{b}\right)+\sum_{k=1}^{K_b}\frac{\left|c_{k,f,n}^{b}\right|^2}{\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k,\tau}^{b} h_{k,n-\tau,\phi}^{b}}\right]$$

$$+N\sum_{f} tr\left[\Sigma_{\bar{n},f}^{-1}R_{\boldsymbol{xx},f}-\Sigma_{\bar{n},f}^{-1}A_f R_{\boldsymbol{xs},f}^{H}-\Sigma_{\bar{n},f}^{-1}R_{\boldsymbol{xs},f}A_f^{H}+\Sigma_{\bar{n},f}^{-1}A_f R_{\boldsymbol{ss},f}A_f^{H}\right]-log\,p(A_f)$$

$$-log\,p(\Sigma_{\bar{n},f})-log\,p(\boldsymbol{W})-log\,p(\boldsymbol{H}|\boldsymbol{\Lambda}) \tag{5.11}$$

where the superscript H is the Hermitian transpose, $R_{\boldsymbol{xx},f}=\frac{1}{N}\sum_n \boldsymbol{x}_{fn}\boldsymbol{x}_{fn}^H$, $R_{\boldsymbol{ss},f}=\frac{1}{N}\sum_n \boldsymbol{s}_{fn}\boldsymbol{s}_{fn}^H$ and $R_{\boldsymbol{xs},f}=\frac{1}{N}\sum_n \boldsymbol{x}_{fn}\boldsymbol{s}_{fn}^H$. The conditional expectations $\hat{R}_{\boldsymbol{xx},f}=R_{\boldsymbol{xx},f}=\frac{1}{N}\sum_n \boldsymbol{x}_{fn}\boldsymbol{x}_{fn}^H$, $\hat{R}_{\boldsymbol{xs},f}=\frac{1}{N}\sum_n \boldsymbol{x}_{fn}\hat{\boldsymbol{s}}_{fn}^H$, $\hat{R}_{\boldsymbol{ss},f}=\frac{1}{N}\sum_n \hat{\boldsymbol{s}}_{fn}\hat{\boldsymbol{s}}_{fn}^H+\hat{\Sigma}_{\boldsymbol{s},fn}$. The source power spectrogram posterior estimates [80] is given by

$$\hat{p}_{j,fn}=\hat{R}_{SS,fn}(j,j) \tag{5.12}$$

where

$$\hat{\boldsymbol{s}}_{fn}=\Sigma_{s,fn}A_f^H\Sigma_{x,fn}^{-1}\boldsymbol{x}_{fn} \tag{5.13}$$

$$\Sigma_{X,fn}=A_f\Sigma_{s,fn}A_f^H+\Sigma_{\bar{n},f} \tag{5.14}$$

$$\hat{\Sigma}_{s,fn}=\left(I_2-\Sigma_{s,fn}A_f^H\Sigma_{x,fn}^{-1}A_f\right)\Sigma_{s,fn} \tag{5.15}$$

$$\Sigma_{s,fn}=\begin{bmatrix}\sum_{k=1}^{K}\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k,\tau}^{g} h_{k,n-\tau,\phi}^{g} & 0\\ 0 & \sum_{k=1}^{K}\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k,\tau}^{b} h_{k,n-\tau,\phi}^{b}\end{bmatrix}_{f,n} \tag{5.16}$$

Detailed derivation of eqn. (5.13) and eqn. (5.15) can be found in [14].

5.3.2 M Step: Update Of Parameters

To find A_f, we set

$$\frac{\partial}{\partial A_f}\langle log\, p(\boldsymbol{X}|\boldsymbol{C},\boldsymbol{\theta}) + log\, p(A_f)\rangle_{P(\boldsymbol{C}|\boldsymbol{X},\boldsymbol{\theta}')} = 0$$

$$\Rightarrow -\Sigma_{\bar{n},f}^{-1}\langle R_{xs,f}\rangle + \Sigma_{\bar{n},f}^{-1} A_f \langle R_{ss,f}\rangle + \varphi(A_f) = 0 \tag{5.17}$$

where $\varphi(A_f) = \partial\, log\, p(A_f)/\partial A_f$. In the case of $P(A_f)$ is a uniform distribution, then eqn. (5.17) leads to a simple closed form expression:

$$A_f = \hat{R}_{xs,f}\hat{R}_{ss,f}^{-1} \tag{5.18}$$

Similarly, for $\Sigma_{\bar{n},f}$ we have

$$\frac{\partial}{\partial \Sigma_{\bar{n},f}^{-1}}\langle log\, p(\boldsymbol{X}|\boldsymbol{C},\boldsymbol{\theta}) + log\, p(\Sigma_{\bar{n},f})\rangle_{P(\boldsymbol{C}|\boldsymbol{X},\boldsymbol{\theta}')} = 0$$

$$\Rightarrow -\Sigma_{\bar{n},f} + R_{xx,f} - A_f\langle R_{xs,f}^H\rangle - \langle R_{xs,f}\rangle A_f^H + A_f\langle R_{ss,f}\rangle A_f^H + \varphi(\Sigma_{\bar{n},f}) = 0 \tag{5.19}$$

where $\varphi(\Sigma_{\bar{n},f}) = \partial \log p(\Sigma_{\bar{n},f})/\partial \Sigma_{\bar{n},f}^{-1}$. When $P(\Sigma_{\bar{n},f})$ assumes a uniform distribution, then eqn. (5.19) leads to

$$\Sigma_{\bar{n},f} = diag\left(R_{\boldsymbol{xx},f} - A_f\hat{R}_{\boldsymbol{xs},f}^H - \hat{R}_{\boldsymbol{xs},f}A_f^H + A_f\hat{R}_{\boldsymbol{ss},f}A_f^H\right) \tag{5.20}$$

Various models exist to model the prior distribution $p(A_f)$ and $p(\Sigma_{\bar{n},f})$ which can be incorporated into the above estimation; however, uniform prior distribution results in computational stable and ease of implementation.

The determination of $\boldsymbol{W}$ and $\boldsymbol{H}$ will follow the multiplicative update rule. At this point it can be distinguished between the two proposed algorithms, and show how the targeted speech signal will be described through the exemplar signal.

5.3.3 Exemplar Based Algorithm

In this algorithm the exemplar signal will be used to initialize the targeted speech signal (see eqn. (5.35) and eqn. (5.36) of Section 5.3.7), and guide separation through matrix co-partial factorization. The NMPCF simultaneously decompose the targeted signal and the side information

and drive them to partially share the common frequency basis in order to enable the side information to guide the separation of the targeted signal [86, 109, 110]. In this chapter, the 2DNMPCF is proposed which is a two-dimensional deconvolution of the NMPCF. The 2DNMPCF shares not only the frequency basis as in the NMPCF but also the convolutive parameters (τ and ϕ) in order to describe the temporal and spectral changes of the targeted speech signal, and therefore renders it more distinguishable and hence more separable than the other sources in the mixture.

The second term in the right hand side of eqn. (5.10) can be expressed using the Itakura-Saito divergence with power spectrogram estimated from the E-step. The third term involves the parametrization of $\{\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\Lambda}\}$. Specifically, each element of H has independent decay parameter $\lambda^{j}_{k,n,\phi}$ with exponential distribution given by $p(\boldsymbol{H}^{j}|\boldsymbol{\Lambda}^{j}) = \prod_{k,n,\phi} p\left(h^{j}_{k,n,\phi}\middle|\lambda^{j}_{k,n,\phi}\right) = \prod_{k,n,\phi} \lambda^{j}_{k,n,\phi} \exp\left(-\lambda^{j}_{k,n,\phi} h^{j}_{k,n,\phi}\right)$. The prior over $\{\boldsymbol{W}^{j}\}$ is flat such that each spectral component is factor-wise normalized to unit length i.e. $p(\boldsymbol{W}^{j}) = \prod_{k} \delta\left(\left\|\boldsymbol{W}^{j}_{k}\right\|_{2} - 1\right)$ where $\left\|\boldsymbol{W}^{j}_{k}\right\|_{2} = \sqrt{\sum_{f,\tau}\left(w^{j}_{f,k,\tau}\right)^{2}}$. Thus, taking the conditional expectation of the negative logarithm of the second and third terms of eqn. (5.10) leads to

$$
\begin{aligned}
&-\langle \log p(\boldsymbol{C}|\boldsymbol{W}, \boldsymbol{H}) + \log p(\boldsymbol{W}) + \log p(\boldsymbol{H}|\boldsymbol{\Lambda}) \rangle_{P\left(\boldsymbol{C}|\boldsymbol{X}, \boldsymbol{\theta}'\right)} \\
&= \sum_{j}\left(\sum_{f,n} D_{IS}\left(\hat{p}_{j,fn}\middle|\sum_{k,\tau,\phi} w^{j}_{f-\phi,k,\tau} h^{j}_{k,n-\tau,\phi}\right) - \sum_{k} \log \delta\left(\left\|\boldsymbol{W}^{j}_{k}\right\|_{2} - 1\right)\right. \\
&\quad \left. + \sum_{k,n,\phi}\left(\lambda^{j}_{k,n-\tau,\phi} h^{j}_{k,n-\tau,\phi} - \log \lambda^{j}_{k,n-\tau,\phi}\right)\right) \qquad (5.21)
\end{aligned}
$$

where $j = \{g, b\}$ and $\hat{p}_{j,fn}$ is the j-th source power spectrogram estimated from (5.12). Thanks to the E-step, that permits direct access to the estimates of the target speech and background signals in order to estimate $\{\boldsymbol{W}^{g}, \boldsymbol{W}^{b}\}$ and $\{\boldsymbol{H}^{g}, \boldsymbol{H}^{b}\}$ rather than from the mixture signal which is noisy. Also the mixing gain will be able to estimate thanks to the pseudo-stereo channel which augments the dimensionality of the mixing matrix and increases its rank. The separation performance, however, can be weakened under the adverse conditions of low signal-to-interference ratio and the background signal shares some characteristics with the target speech. To alleviate these conditions,

a form of exemplar signal whose spectral and temporal characteristics resemble the target speech will be used. The exemplar signal can be derived from the text associated with the mixture and generated by using a speech synthesizer or human speakers. Let $\tilde{y}(t)$ be the sampled exemplar signal, $y_{f,n_y} \in \mathbb{C}^{1\times N_y}$ be the STFT of $\tilde{y}(t)$, and $p_{y,f,n_y} = |y_{f,n_y}|^2$ is the power spectrogram of the exemplar signal. It should be emphasized that N can differ from N_y due to the temporal mismatch between the exemplar signal and the mixture, since it is not practically feasible to emulate the exemplar to be an exact match to the targeted speech signal. These temporal mismatches between the exemplar and the targeted speech signals will result in mismatch between the activation tensors of the exemplar and the targeted speech. A synchronization matrix has been adopted to address this issue [112]. With the exemplar signal, a joint decomposition using the mixture and exemplar spectrograms have been developed to obtain improved estimates of the spectral basis tensor $\boldsymbol{W}$ and the temporal tensor $\boldsymbol{H}$. This is done allowing the exemplar signal to be factorized using similar model i.e., multiple components NMF2D $p_{y,f,n_y} \approx \sum_{k=1}^{K_g}\sum_{\tau=0}^{\tau_{max}}\sum_{\phi=0}^{\phi_{max}} w_{f-\phi,\mathrm{k},\tau}^{y} h_{k,n-\tau,\phi}^{y}$ and constraining the structures of spectral basis and temporal tensors between the target and exemplar signals. (5.21) augment with a weighted joint factorization of the exemplar spectrogram as follows:

$$\mathcal{J} = \underbrace{\sum_{j,f,n} D_{IS}\left(\hat{p}_{j,fn}\middle|\sum_{k,\tau,\phi} w_{f-\phi,k,\tau}^{j} h_{k,n-\tau,\phi}^{j}\right) + \eta D_{IS}\left(p_{y,fn_y}\middle|\sum_{k,\tau,\phi} w_{f-\phi,\mathrm{k},\tau}^{y} h_{k,n_y-\tau,\phi}^{y}\right)}_{\mathcal{J}_1} -$$

$$\underbrace{\sum_{j,k}\log\left(\delta\left(\left\|\boldsymbol{W}_k^j\right\|_2 - 1\right)\right)}_{\mathcal{J}_2} + \underbrace{\sum_{j,k,n,\phi}\left(\lambda_{k,n,\phi}^{j} h_{k,n,\phi}^{j} - \log\lambda_{k,n,\phi}^{j}\right)}_{\mathcal{J}_3} \tag{5.22}$$

In above, η is the scalar that weigh the importance of the exemplar signals in the factorization process. The term $\mathcal{J}_1$ represents the matrix factorization of the sources and exemplar spectrograms into the spectral basis and activation tensors, $\mathcal{J}_2$ denotes the regularization on the spectral basis, and $\mathcal{J}_3$ represents the sparseness of the activation. The regularization involving $\delta\left(\left\|\boldsymbol{W}_k^j\right\|_2 - 1\right)$ can be satisfied by explicitly normalizing each spectral dictionary to unity i.e. $w_{f,k,\tau}^{j} = w_{f,k,\tau}^{j} \Big/ \sqrt{\sum_{f,\tau}\left(w_{f,k,\tau}^{j}\right)^2}$. Using the definition of the Itakura-Saito divergence and by letting $v_{fn}^{g} = \sum_{k,\tau,\phi} w_{f-\phi,k,\tau}^{g} h_{k,n-\tau,\phi}^{g}$, $v_{fn}^{b} = \sum_{k,\tau,\phi} w_{f-\phi,k,\tau}^{b} h_{k,n-\tau,\phi}^{b}$ and $v_{fn}^{y} = \sum_{k,\tau,\phi} w_{f-\phi,k,\tau}^{y} h_{k,n-\tau,\phi}^{y}$, the above cost function reduces up to the constant terms to

$$\mathcal{J} \overset{c}{=} \sum_{f,n} \left(\hat{p}_{1,fn}\, v_{fn}^{g^{-1}} - \log\, v_{fn}^{g^{-1}} \right) + \sum_{k,n,\phi} \lambda_{k,n,\phi}^{g} h_{k,n,\phi}^{g} - \sum_{k,n,\phi} \log \lambda_{k,n,\phi}^{g}$$
$$+ \sum_{f,n} \left(\hat{p}_{2,fn}\, v_{fn}^{b^{-1}} - \log\, v_{fn}^{b^{-1}} \right) + \sum_{k,n,\phi} \lambda_{k,n,\phi}^{b} h_{k,n,\phi}^{b} - \sum_{k,n,\phi} \log \lambda_{k,n,\phi}^{b}$$
$$+ \sum_{f,n_y} \eta \left(p_{y,fn_y}\, v_{fn_y}^{y^{-1}} - \log\, v_{fn_y}^{y^{-1}} \right) \tag{5.23}$$

The MU approach will be used to estimate $\boldsymbol{W}^{g}$ and $\boldsymbol{H}^{g}$:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} \cdot \frac{[\nabla \mathcal{J}]_{-}}{[\nabla \mathcal{J}]_{+}} \tag{5.24}$$

where $\nabla \mathcal{J} = [\nabla \mathcal{J}]_{+} - [\nabla \mathcal{J}]_{-}$. This leads to

$$w_{f',k',\tau'}^{g}$$
$$\leftarrow w_{f',k',\tau'}^{g} \frac{\sum_{\phi,n} \hat{p}_{1,f'+\phi,n} v_{f'+\phi,n}^{g^{-2}} h_{k',n-\tau',\phi}^{g} + \eta \sum_{\phi,n_y} p_{y,f'+\phi,n_y} v_{f'+\phi,n_y}^{y^{-2}} h_{k',n_y-\tau',\phi}^{y}}{\sum_{\phi,n} v_{f'+\phi,n}^{g^{-1}} h_{k',n-\tau',\phi}^{g} + \eta \sum_{\phi,n_y} v_{f'+\phi,n_y}^{y^{-1}} h_{k',n_y-\tau',\phi}^{y}} \tag{5.25}$$

given that $\boldsymbol{W}^{y} = \boldsymbol{W}^{g}$ [86].

$$h_{k',n',\phi'}^{g}$$
$$\leftarrow h_{k',n',\phi'}^{g} \left(\frac{\sum_{f,\tau} \hat{p}_{1,f,n'+\tau} v_{f,n'+\tau}^{g^{-2}} w_{f-\phi',k',\tau}^{g} + \eta \sum_{f,\tau,n_y'} w_{f-\phi',k',\tau}^{y} p_{f,n_y'+\tau} v_{f,n_y'+\tau}^{y^{-2}} d_{n_y',n'}}{\sum_{f,\tau} v_{f,n'+\tau}^{g^{-1}} w_{f-\phi',k',\tau}^{g} + \lambda_{k',n',\phi'}^{g} + \eta \left(\sum_{f,\tau,n_y'} w_{f-\phi',k',\tau}^{y} v_{f,n_y'+\tau}^{y^{-1}} d_{n_y',n'} + \lambda_{k',n_y',\phi'}^{y} d_{n_y',n'} \right)} \right)$$

(5.26)

given that $\boldsymbol{H}^{y} = \boldsymbol{H}^{g} D^{T}$ [86], where D is the synchronization matrix [112] of dimension $N_y \times N$ to ameliorate the temporal mismatch between the exemplar and the mixture. For the sparsity term, the update is obtained by solving $\frac{\partial}{\partial \lambda_{k',n',\phi'}^{g}} \langle log\, p(\boldsymbol{C}, \boldsymbol{\theta} | \boldsymbol{X}) \rangle_{P(\boldsymbol{C}|\boldsymbol{X},\boldsymbol{\theta}')} = 0$ which leads to

$$\lambda_{k',n',\phi'}^{g} = \frac{1}{h_{k',n',\phi'}^{g}} \tag{5.27}$$

By following the same procedure as above, $\boldsymbol{W^b}$, $\boldsymbol{H^b}$ and $\lambda_{k,n,\phi}^{b}$ can be estimated as follows

$$w^{b}_{f',k',\tau'} \leftarrow w^{b}_{f',k',\tau'} \frac{\sum_{\phi,n} \hat{p}_{2,f'+\phi,n} v^{b^{-2}}_{f'+\phi,n} h^{b}_{k',n-\tau',\phi}}{\sum_{\phi,n} v^{b^{-1}}_{f'+\phi,n} h^{b}_{k',n-\tau',\phi}} \tag{5.28}$$

$$h^{b}_{k',n',\phi'} \leftarrow h^{b}_{k',n',\phi'} \left(\frac{\sum_{f,\tau} \hat{p}_{2,f,n'+\tau} v^{b^{-2}}_{f,n'+\tau} w^{b}_{f-\phi',k',\tau}}{\sum_{f,\tau} v^{b^{-1}}_{f,n'+\tau} w^{b}_{f-\phi',k',\tau} + \lambda^{b}_{k',n',\phi'}} \right) \tag{5.29}$$

$$\lambda^{b}_{k',n',\phi'} = \frac{1}{h^{b}_{k',n',\phi'}} \tag{5.30}$$

Similarly, $\boldsymbol{W}^{y}$, and $\boldsymbol{H}^{y}$ can be estimate as

$$w^{y}_{f',k',\tau'} \leftarrow w^{y}_{f',k',\tau'} \frac{\sum_{\phi,n_y} p_{y,f'+\phi,n_y} v^{y^{-2}}_{f'+\phi,n_y} h^{y}_{k',n_y-\tau',\phi}}{\sum_{\phi,n} v^{y^{-1}}_{f'+\phi,n_y} h^{y}_{k',n_y-\tau',\phi}} \tag{5.31}$$

$$h^{y}_{k',n'_y,\phi'} \leftarrow h^{y}_{k',n'_y,\phi'} \left(\frac{\sum_{f,\tau} p_{y,f,n'_y+\tau} v^{y^{-2}}_{f,n'_y+\tau} w^{y}_{f-\phi',k',\tau}}{\sum_{f,\tau} v^{y^{-1}}_{f,n'_y+\tau} w^{y}_{f-\phi',k',\tau}} \right) \tag{5.32}$$

5.3.4 Semi-Exemplar Based Algorithm

In this algorithm the exemplar will be used to initialize the targeted speech signal as in eqn. (5.35) and eqn. (5.36) of Section 5.3.7, then the MU rule will be used to updates the NMF2D tensors of both the speech and background signal. The tensors of the speech signal can be obtained by setting η to zero for eqn. (5.25) and eqn. (5.26) leading to

$$w^{g}_{f',k',\tau'} \leftarrow w^{g}_{f',k',\tau'} \frac{\sum_{\phi,n} \hat{p}_{1,f'+\phi,n} v^{g^{-2}}_{f'+\phi,n} h^{g}_{k',n-\tau',\phi}}{\sum_{\phi,n} v^{g^{-1}}_{f'+\phi,n} h^{g}_{k',n-\tau',\phi}} \tag{5.33}$$

$$h^{g}_{k',n',\phi'} \leftarrow h^{g}_{k',n',\phi'} \left(\frac{\sum_{f,\tau} \hat{p}_{1,f,n'+\tau} v^{g^{-2}}_{f,n'+\tau} w^{g}_{f-\phi',k',\tau}}{\sum_{f,\tau} v^{g^{-1}}_{f,n'+\tau} w^{g}_{f-\phi',k',\tau} + \lambda^{g}_{k',n',\phi'}} \right) \tag{5.34}$$

and the same sparsity in eqn. (5.27) will be used. The tensors of the background signal follow similarly and they are shown in eqn. (5.28) and eqn. (5.29), while the sparsity update in eqn. (5.30).

The semi-exemplar based algorithm use the exemplar signal to initialize the tensors of the NMF2D and, thus it depends on the exemplar to give the good start only. On the other hand, the exemplar based algorithm uses the exemplar signal not only to give the correct initialization but also to guide the whole algorithm through the 2DNMPCF which factorizes both exemplar and mixture signals at the same time. Therefore, the exemplar based algorithm recycles the use of signal $\tilde{y}(t)$ more than the semi-exemplar based algorithm.

5.3.5 Describing The Targeted Speech Signal By Using The Exemplar

The description of the targeted speech signal will be carried out indirectly by the exemplar signal and with the aid of the NMF2D that optimizes its parameters. The parameters of the NMF2D will be optimized by depending on the exemplar signal instead of the mixture. The exemplar is considered instead of the targeted speech signal as it is unavailable. The NMF2D is proposed due to its ability in descripting the temporal and spectral changes through the convolutive parameters (τ and ϕ), and specifying the required number of frequency basis K.

The determination of the model order for NMF2D will be realized using the exemplar signal y(t):

Step 1: Optimize $\boldsymbol{W^y}$, and $\boldsymbol{H^y}$ by using eqns. (5.31) and (5.32):

Step 2: Optimizing τ and ϕ:

1) Set $K = 1$
2) For $\tau_{max} = 1$ to T

   For $\phi_{max} = 1$ to Φ

   - Estimate $v_{fn}^{y} = \sum_{k,\tau,\phi} w_{f-\phi,k,\tau}^{y} h_{k,n-\tau,\phi}^{y}$
   - Estimate the signal-to-distortion ratio (SDR) [113] between the exemplar signal p_{y,fn_y} and its approximate v_{fn}^{y} in order to evaluated the factorization performance

Select the convolutive parameters (τ_{max}, ϕ_{max}) that give the highest SDR.

Step 3: Optimizing K:

1) For $K = 2$ to K_{max}

   - Estimate $v_{fn}^{y} = \sum_{k,\tau,\phi} w_{f-\phi,k,\tau}^{y} h_{k,n-\tau,\phi}^{y}$
   - Estimate the SDR between the exemplar signal p_{y,fn_y} and its approximate v_{fn}^{y}

Select K that give the highest SDR.

5.3.6 Components Reconstruction

The estimated sources ($\hat{\boldsymbol{s}}_{fn}$) can be reconstructed by using Wiener filtering $\left(\Sigma_{s,fn}A_f^H\Sigma_{x,fn}^{-1}\right)$ as in eqn. (5.13), and due to the linearity of the STFT, the inverse-STFT (with dual synthesis window [95]) can be used to transform it to the time domain.

5.3.7 Initialization

The initialization is an essential part for the separation since the NMF2D is very sensitive to the initialization. In this chapter, the spectral and temporal tensors of the proposed algorithms will be initialized by using the exemplar signal $\tilde{y}(t)$ which itself is decomposed into $w_{f,k,\tau}^{y}$ and $h_{k,n,\phi}^{y}$:

$$\left(w_{f,k,\tau}^{g}\right)_{ini} = w_{f,k,\tau}^{y} \tag{5.35}$$

$$\left(h_{k,n,\phi}^{g}\right)_{ini} = h_{k,n_Y,\phi}^{y}\, d_{n_y,n} \tag{5.36}$$

where $d_{n_y,n}$ is synchronization parameter [112]. For the background, $\left(w_{f,k,\tau}^{b}\right)_{ini}$ and $\left(h_{k,n,\phi}^{b}\right)_{ini}$ will be randomly initialized. Thus the mixture can be initialized as follows:

$$\left(w_{f,k,\tau}^{x}\right)_{ini} = \left[\left(w_{f,k,\tau}^{g}\right)_{ini} \quad \left(w_{f,k,\tau}^{b}\right)_{ini}\right] \tag{5.37}$$

$$\left(h_{k,n,\phi}^{x}\right)_{ini} = \begin{bmatrix}\left(h_{k,n,\phi}^{g}\right)_{ini} \\ \left(h_{k,n,\phi}^{b}\right)_{ini}\end{bmatrix} \tag{5.38}$$

Table 5.1 and 5.2 summarize the proposed algorithms.

Table 5.1: Proposed algorithm 1 (Semi-Exemplar)

1. Optimize the convolutive parameters and number of components based on the exemplar.

2. Initialize $w_{f,k,\tau}^{g}$ and $h_{k,n,\phi}^{g}$ based on the exemplar, $w_{f,k,\tau}^{b}$ and $h_{k,n,\phi}^{b}$ randomly.

3. Generate the pseudo channel $\tilde{x}_2(t)$ as in eqn. (5.2).

4. Apply the STFT on the mixture signal.

5. E-step: compute $\hat{p}_{jfn}$ and $\hat{s}_{fn}$ using eqns. (5.12) and (5.13), respectively.

6. M-step: compute $A_f, \Sigma_{\bar{n}}, w^g_{f,k,\tau}, h^g_{k,n,\phi}$, $\lambda^g_{k,n,\phi}, w^b_{f,k,\tau}, h^b_{k,n,\phi}$, and $\lambda^b_{k,n,\phi}$ using eqn. (5.18), eqn. (5.20), eqn. (5.33), eqn. (5.34), eqn. (5.27), eqn. (5.28), eqn. (5.29), and eqn. (5.30).

7. Normalize $w^x_{f,k,\tau} = w^x_{f,k,\tau} / \sqrt{\sum_{f,k,\tau} \left(w^x_{f,k,\tau}\right)^2}$

8. Repeat E- and M-steps, and the normalization until convergence is achieved i.e. rate of cost change is below a prescribed threshold, ψ.

9. Perform inverse STFT with dual synthetic window to estimate $\tilde{g}(t)$, and $\tilde{b}(t)$.

Table 5.2: Proposed algorithm 2 (Full-Exemplar)

1. Optimize the convolutive parameters and number of components based on the exemplar
2. Initialize $w^g_{f,k,\tau}$ and $h^g_{k,n,\phi}$ based on the exemplar, $w^b_{f,k,\tau}$ and $h^b_{k,n,\phi}$ randomly
3. Generate the pseudo channel $\tilde{x}_2(t)$ as in eqn. (5.2)
4. Apply the STFT on the mixture signal.
5. E-step: compute $\hat{p}_{jfn}$ and $\hat{s}_{fn}$ using eqns. (5.12) and (5.13), respectively.
6. M-step: compute $A_f, \Sigma_{\bar{n}}, w^y_{f,k,\tau}, h^y_{k,n_y,\phi}$, $w^g_{f,k,\tau}, h^g_{k,n,\phi}$, $\lambda^g_{k,n,\phi}, w^b_{f,k,\tau}, h^b_{k,n,\phi}$, and $\lambda^b_{k,n,\phi}$ using eqn. (5.18), eqn. (5.20), eqn. (5.31), eqn. (5.32), eqn. (5.25), eqn. (5.26), eqn. (5.27), eqn. (5.28), eqn. (5.29), and eqn. (5.30).
7. Normalize $w^x_{f,k,\tau} = w^x_{f,k,\tau} / \sqrt{\sum_{f,k,\tau} \left(w^x_{f,k,\tau}\right)^2}$
8. Repeat E- and M-steps, and the normalization until convergence is achieved i.e. rate of cost change is below a prescribed threshold, ψ.
9. **Perform** inverse STFT with dual synthetic window to estimate $\tilde{g}(t)$, and $\tilde{b}(t)$.

5.4 Results and Discussions

5.4.1 Dataset

The performance of the proposed algorithms was investigated and compared with recent state-of-art text informed source separation [86]. For fair comparison, the same datasets was used. These datasets are 10 speech mixtures that mixed with music (Speech + music) and with effect (Speech + Fx). So it resulted in 20 mixtures in total. For each mixture the speech is emulated by using 12 exemplars (synth Man, Synth Woman, TMT Man, TMT woman, and other 8 foreign speakers). Thus there were 240 experiments (generated from the 20 mixtures and the 12 exemplars for each mixture) for SNR of -5dB.

5.4.2 Evaluation

In order to evaluate the proposed algorithm the SDR [98] that combines both the source-to-interference ratio (SIR), source image-to-spatial distortion ratio (ISR), and the source-to-artefacts ratio (SAR) will be used to evaluate the estimated sources with respect to the original sources. The Matlab codes for this evaluation procedure can be found in [99].

5.4.3 Selections of η, δ, and γ

The contribution of the exemplar on the separation is weighted by η, so if $\eta = 0$ the exemplar will have little effect, while if its value increased the exemplar will have more influence. According to [86] the value of η can be found as follows

$$\eta = \eta_0 \frac{N}{N_y} \tag{5.39}$$

where N and N_y is the temporal length of the mixture and the exemplar, respectively, and for our case η_0 has been set to $\eta_0 = 0.5$.

The other parameter, which is the time-delay δ can be computed as follows [111]

$$\delta_{max} < \frac{f_s}{2f_{max}} \tag{5.40}$$

where f_s is the sampling frequency and f_{max} is the maximum frequency presented in the mixture. For the weighting parameter γ that determine the attenuation on the delayed mixture $\gamma\tilde{x}_1(t-\delta)$ (see eqn. (5.2)), it has been found that exists a range of γ with high SDR as shown Figure 5.2. The plot suggests that this range to be $0.1 \leq \gamma \leq 0.25$. In all our cases, we use $\gamma = 0.15$.

.

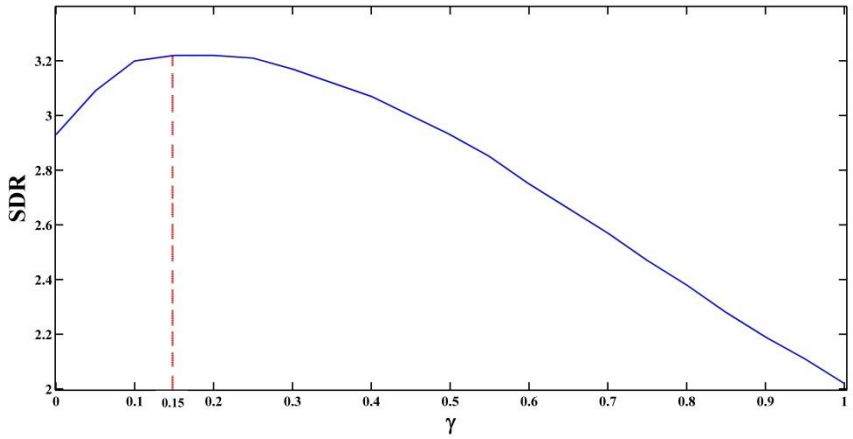
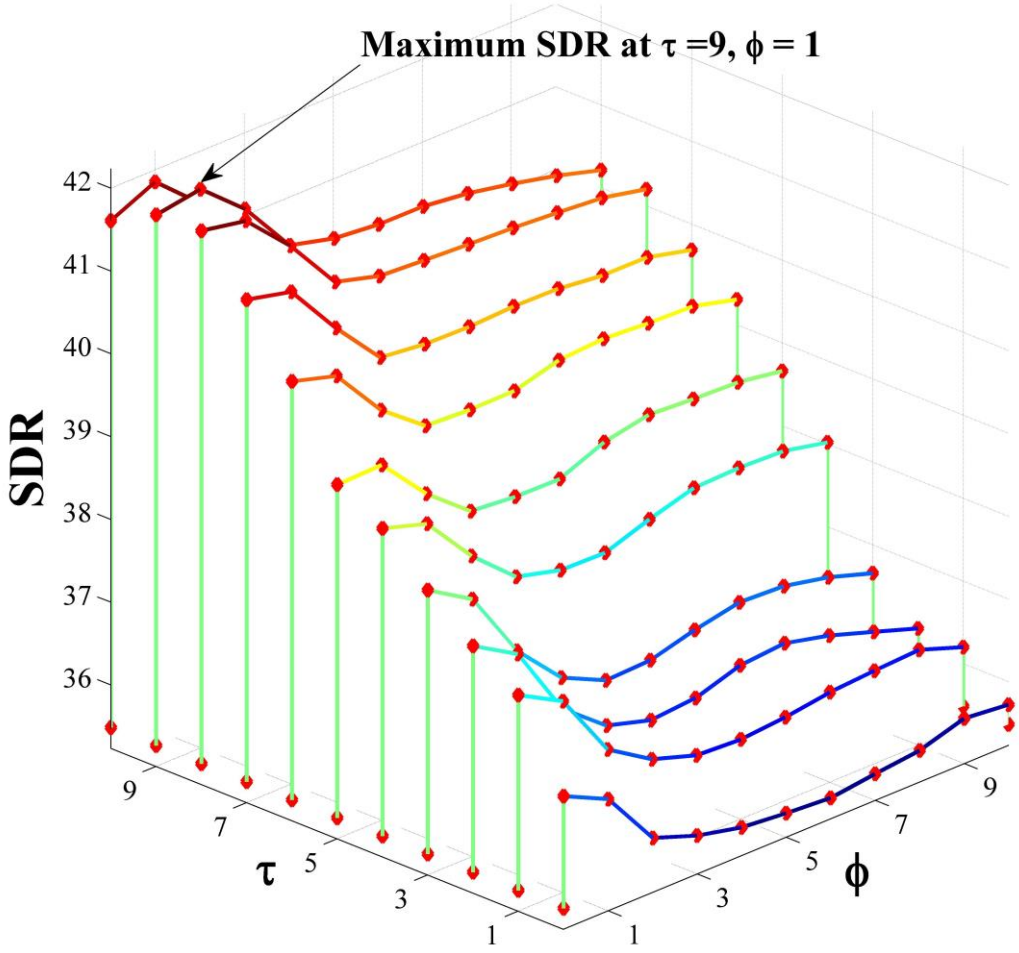


Figure 5.2. The SDRs w.r.t. different values of γ.

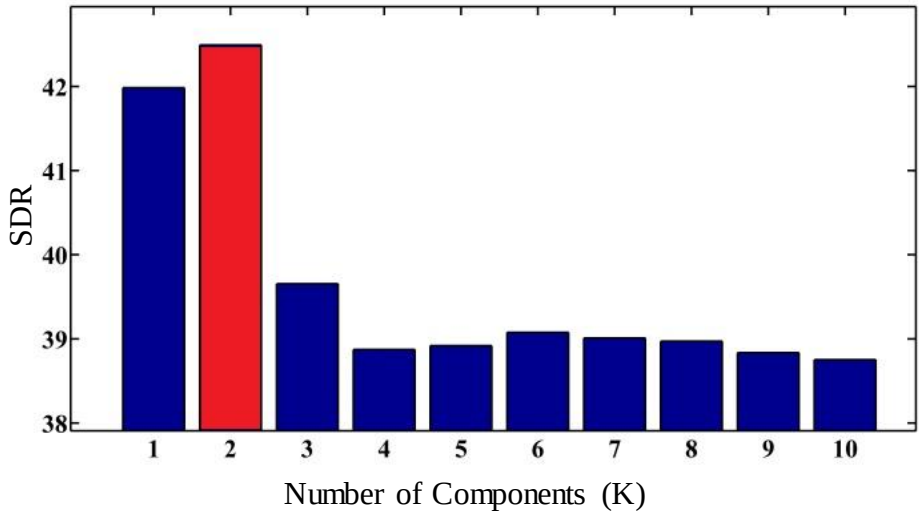
5.4.4 Optimization of $\tau, \phi, and\ K$

By following the procedure described in Section 5.3.5 (setting $T = 10$, $\Phi = 10$, and $K_{max} = 10$), the results for one exemplar are shown in Figure 5.3. Figure 5.3(a) shows that the best SDR is attained at $\tau = 9$ and $\phi = 1$. In addition, Figure 5.3(b) reveals that $K = 2$ results in the optimum number of components. By following the same procedure the parameters of the exemplars for mixture 1 to 10 are tabulated in Table 5.3. It can be seen from Table 5.3, that there is 120 different parameters $(\tau, \phi, and\ K)$. These 120 different parameters came from 120 different exemplars (each speech signal in the mixture is emulated by 12 exemplars, and as there are 10 mixtures, this results in 120 exemplars)[4]. Despite 12 exemplars emulate the same speech signal, they have different parameters because they derived from different speakers (native and on-native English speaker) and different genders, and as a result of these differences there will be a different parameters of the NMF2D that describe each exemplar.

[4] The 120 (Speech+Music) mixture group and the 120 (Speech+Effects) mixture group have the same speech signal.



(a)



(b)

Figure 5.3: The SDRs w.r.t. (a) τ and ϕ, (b) Number of components K.

Table 5.3

Optimizing the parameters of the exemplars for mixtures 1 to 10.

Exemplar	**1**	**2**	**3**	**4**	**5**	**6**
	τ, ϕ, k	τ, ϕ, k	τ, ϕ, k	τ, ϕ, k	τ, ϕ, k	τ, ϕ, k
Mix 1	10, 2, 1	10, 1,1	8, 1, 1	1, 1, 1	10, 0, 1	9, 0, 1
Mix 2	1, 0, 1	1, 0, 5	1, 0, 1	5, 1, 1	3, 0, 1	5, 0, 2
Mix 3	3, 0, 1	1, 0, 3	2, 0, 1	1, 0, 3	4, 0, 1	6, 0, 1
Mix 4	6, 0, 1	5, 0, 1	4, 0, 1	1, 0, 1	1, 0, 1	1, 0, 3
Mix 5	1, 0, 1	4, 0, 1	1, 0, 1	0, 1, 2	10, 0, 1	1, 0, 1
Mix 6	1, 0, 1	1, 0, 1	5, 0, 1	1, 0, 1	6, 0, 1	4, 0, 1
Mix 7	1, 0, 1	0, 1, 1	1, 0, 1	6, 0, 1	9, 0, 1	5, 1, 1
Mix 8	1, 0, 1	1, 0, 1	3, 0, 1	0, 1, 4	5, 0, 1	7, 0, 1
Mix 9	1, 0, 1	1, 0, 1	1, 0, 1	1, 0, 1	5, 0, 1	10, 0, 1
Mix 10	1, 0, 1	1, 1, 1	1, 0, 1	10, 1,1	1, 0, 1	5, 0, 1
Exemplar	**7**	**8**	**9**	**10**	**11**	**12**
	τ, ϕ, k	τ, ϕ, k	τ, ϕ, k	τ, ϕ, k	τ, ϕ, k	τ, ϕ, k
Mix 1	9, 0, 1	9, 0, 1	8, 0, 1	4, 0, 1	1, 0, 1	1, 0, 1
Mix 2	4, 0, 1	6, 0, 1	7, 0, 1	10, 0, 2	3, 0, 1	0, 1, 1
Mix 3	7, 0, 1	9, 0, 1	10, 0, 1	10, 0, 2	6, 0, 1	1, 0, 1
Mix 4	10, 0, 1	1, 0, 1	10, 0, 1	10, 0, 2	2, 0, 1	2, 0, 1
Mix 5	1, 0, 1	1, 0, 1	10, 0, 1	7, 1, 1	10, 0, 1	1, 0, 1
Mix 6	8, 1, 2	6, 0, 1	9, 0, 1	9, 1, 2	6, 0, 1	1, 0, 1
Mix 7	10, 0, 1	8, 0, 1	7, 0, 1	10, 0, 1	8, 0, 1	0, 1, 2
Mix 8	6, 0, 1	6, 0, 1	9, 0, 1	10, 0, 1	4, 0, 1	1, 0, 1
Mix 9	7, 0, 2	8, 0, 1	8, 0, 1	5, 0, 1	1, 0, 1	1, 0, 1
Mix 10	4, 0, 1	5, 0, 1	10, 0, 1	10, 0, 2	1, 0, 1	3, 0, 1

5.4.5 Results

The STFT windows length was set to 512 with 50% overlaps. To show the convergence of the proposed algorithms, the convergence of the cost functions eqn. (5.10) of both algorithms are shown in Figure 5.4. This plot is obtained for one mixture with twelve exemplars. It is noted that all trajectories have converged to the steady state in less than 50 iterations. The fast and stable convergence is attributed to the manner of how the GEM-MU algorithm adapts the model parameters and latent variables.

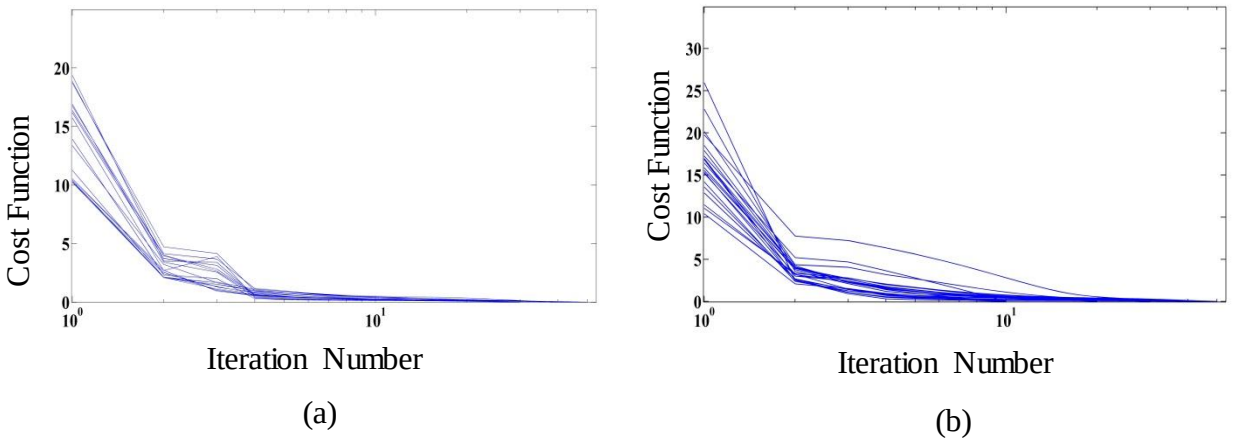


Figure 5.4: Cost function for (a) Semi-Exemplar based algorithm and (b) Exemplar based algorithm.

The proposed algorithms will be compared with the NMPCF model based on the excitation-filter channel speech model [86]. In this algorithm the variations between the speech example and the targeted speech in the mixture such as pitch variation, phonemes pronounced, recording conditions, and speaker's vocal tract are modelled by the excitation-filter channel speech model. The NMPCF jointly factorizes the spectrograms of the mixture and the exemplar that emulate the speech signal. Also, the proposed algorithms will be compared with the Gaussian Scaled Mixture Model (GSMM) [86] with constraints applied on the matrices of the excitation-filter channel speech model under the NMPCF model umbrella, in order to have a physical motivation, such as allowing one phoneme to be pronounced at a time and one fundamental frequency to be active at a time.

The SDRs of the NMPCF model based on the excitation-filter channel speech model [86], the structural GSMM algorithm [86], and the proposed algorithms are tabulated in Table 5.4. The Table indicates that the proposed algorithms have better performance than the NMPCF, which can

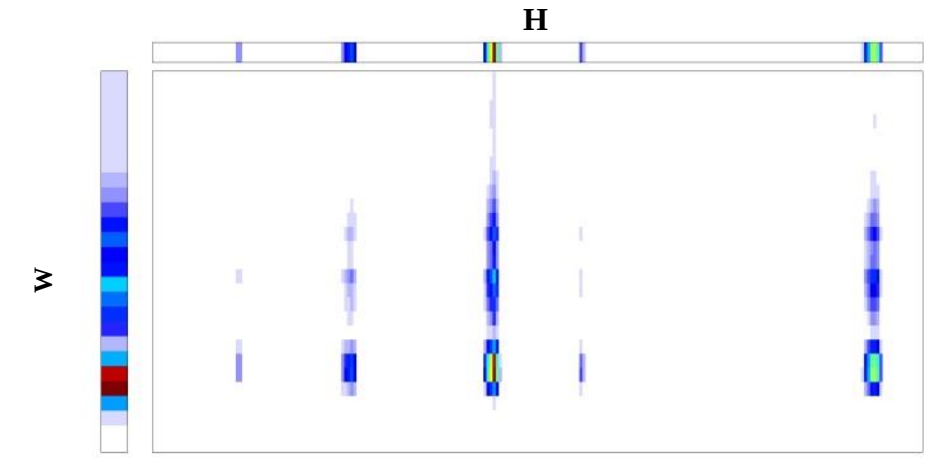
be summarized as follows: An achievement of 2.57 dB more for the speech and music group, and 1.89 dB more for the speech and effects group for the Semi-Exemplar based algorithm. For the Exemplar based algorithm an achievement of 3.12 dB more for the speech and music group, and 3.37 dB more for the speech and effects group. Furthermore, the Exemplar based algorithm achieved an improvement of 1.86 dB for the speech and effects group and 0.16 dB for the speech and music group, in comparison with the structural GSMM algorithm. On the other hand, the Semi-Exemplar based algorithm achieves 0.38 dB more for the speech and effects group, and 0.39 dB less for the speech and music group. Although the proposed semi-exemplar based algorithm is less dependent on the exemplar signal, its high performance is attributed to the pseudo-stereo channels.

Table 5.4

Average SDRs of the 10 mixtures with their different 12 exemplars for the NMPCF and the proposed algorithms.

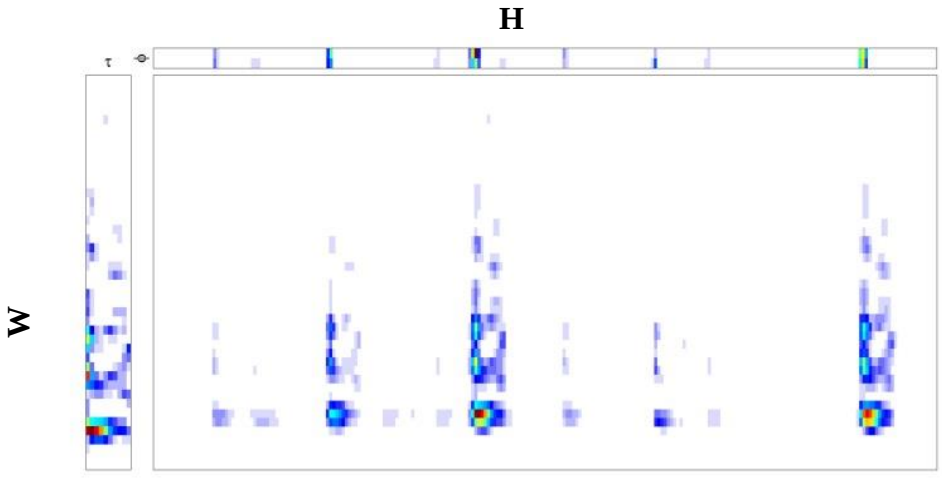
SNR= -5dB	SPEECH + Music	SPEECH + Fx
NMPCF	-0.74	0.67
Structural GSMM	2.22	2.18
Proposed Semi-Exemplar based algorithm	1.83	2.56
Proposed Exemplar based algorithm	**2.38**	**4.04**

The proposed algorithms have achieved higher results than the NMPCF since they have more powerful source representation the NMF2D and the 2DNMPCF, which address the change in the time and frequency directions through the parameters (τ and ϕ). To show the effects of these parameters, one component of the $\boldsymbol{W}$ and $\boldsymbol{H}$ tensors and its corresponding spectrogram for both the NMPCF and the proposed 2DNMPCF has been plotted in Figure 5.5(a) and Figure 5.5(b), respectively. Both plots show how $\boldsymbol{W}$ modelled the changes in the frequencies of the source and how $\boldsymbol{H}$ modelled the distribution in the time domain. On the separate hand, $\boldsymbol{W}$ and $\boldsymbol{H}$ of the NMPCF detect the frequency bases, however it was not able to address the frequency and the temporal changes.



Spectrogram

(a)



Spectrogram

(b)

Figure 5.5: One component of **W**, and **H** with their corresponding spectrogram for the (a) NMPCF and (b) 2DNMPCF.

Additionally, the spectrogram of the original speech, the exemplar, the mixture, and the estimated speech by using the proposed algorithms and the NMPCF are shown in Figure 5.6. These plots clearly show that the proposed algorithms have successfully detected the pitch and temporal change of the source, due to its two-dimensional deconvolution while NMPCF failed to detect these changes. Furthermore, Figure 5.7 shows the waveforms of these signals.

Finally, from Table 5.4 it can be seen that the Exemplar based algorithm achieved better separation results than the Semi-Exemplar based algorithm since the latter only uses the exemplar to initialize the tensors of the targeted speech signal. Thus the initialization will guide the algorithm for the first iteration and gives the correct start but it may get trapped in local minima or drifted away from the solution as the iterations increases. Although the Exemplar based algorithm has been given the identical start as the Semi-Exemplar based algorithm, its separation is guided by the 2DNMPCF which models both the exemplar and the targeted speech signal. To show this, the waveform of the original voice, exemplar, and the estimated voice by using these two algorithms are shown in Figure 5.8. The plot indicates that the exemplar based algorithm has successfully estimated the original source. This shows the importance and contribution of 2DNMPCF on the proposed algorithm.

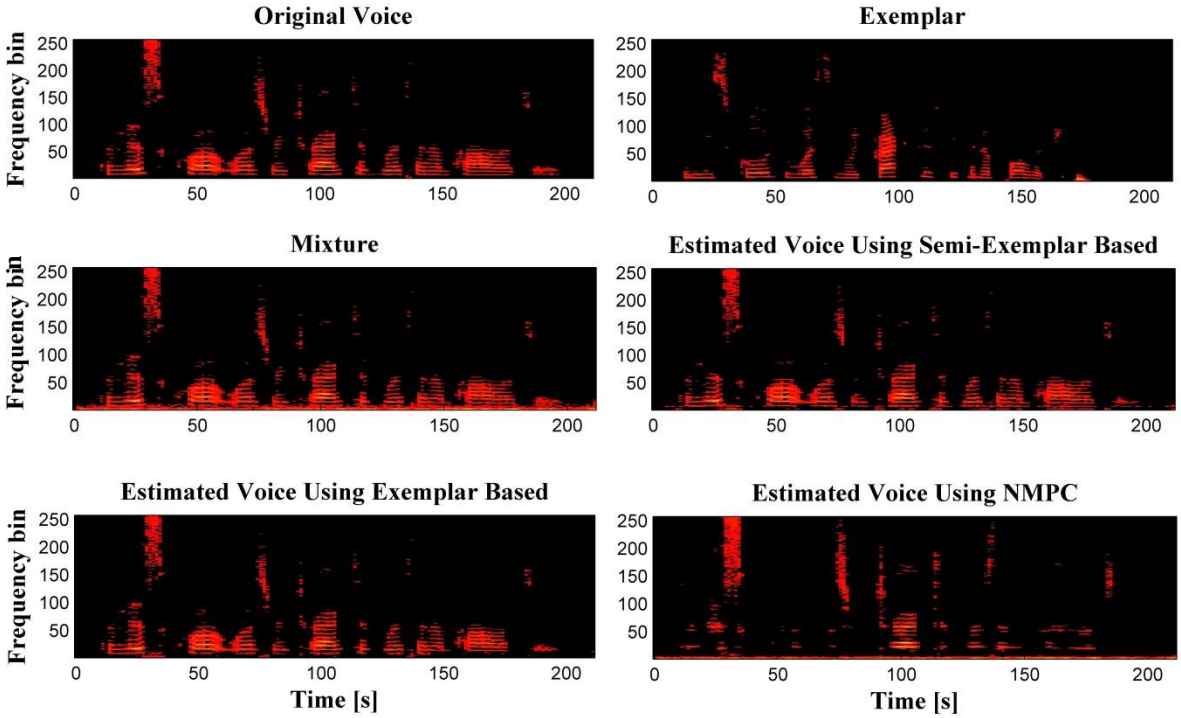


Figure 5.6: Spectrogram of the original speech, exemplar, and the estimated speech by using the proposed algorithms and the NMPCF algorithm.

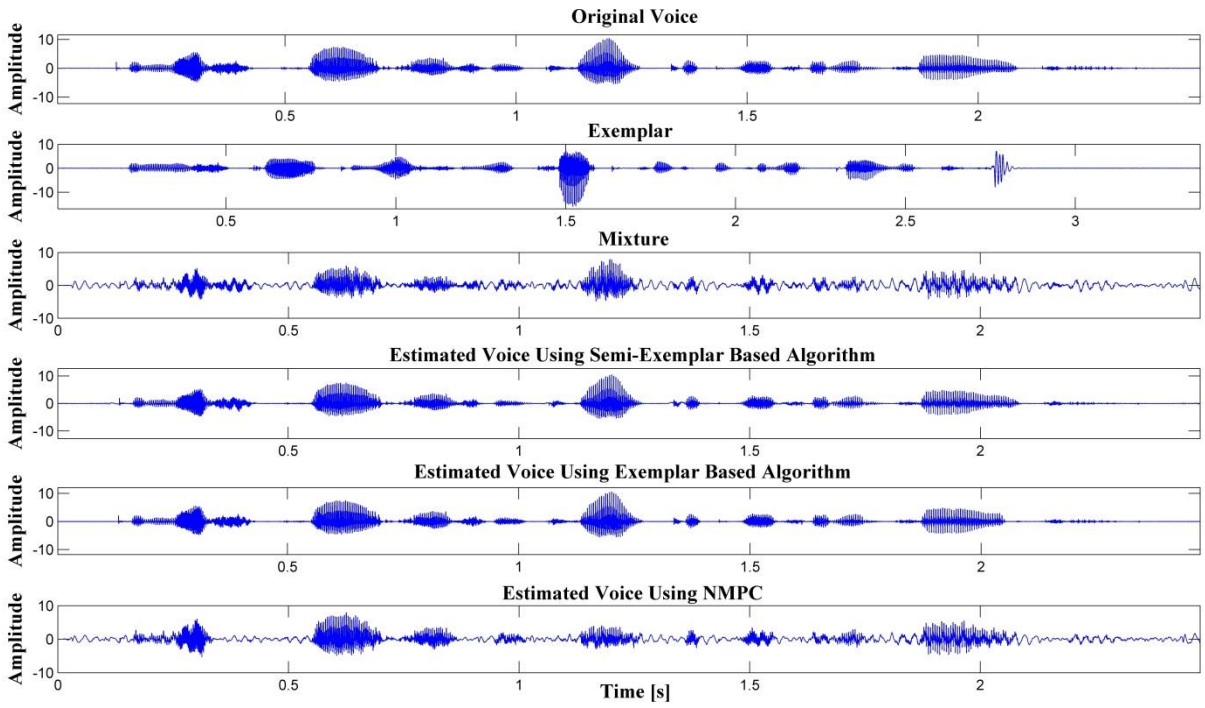


Figure 5.7: Waveform of the original speech, exemplar, and the estimated speech by using the proposed algorithms and the NMPCF algorithm.

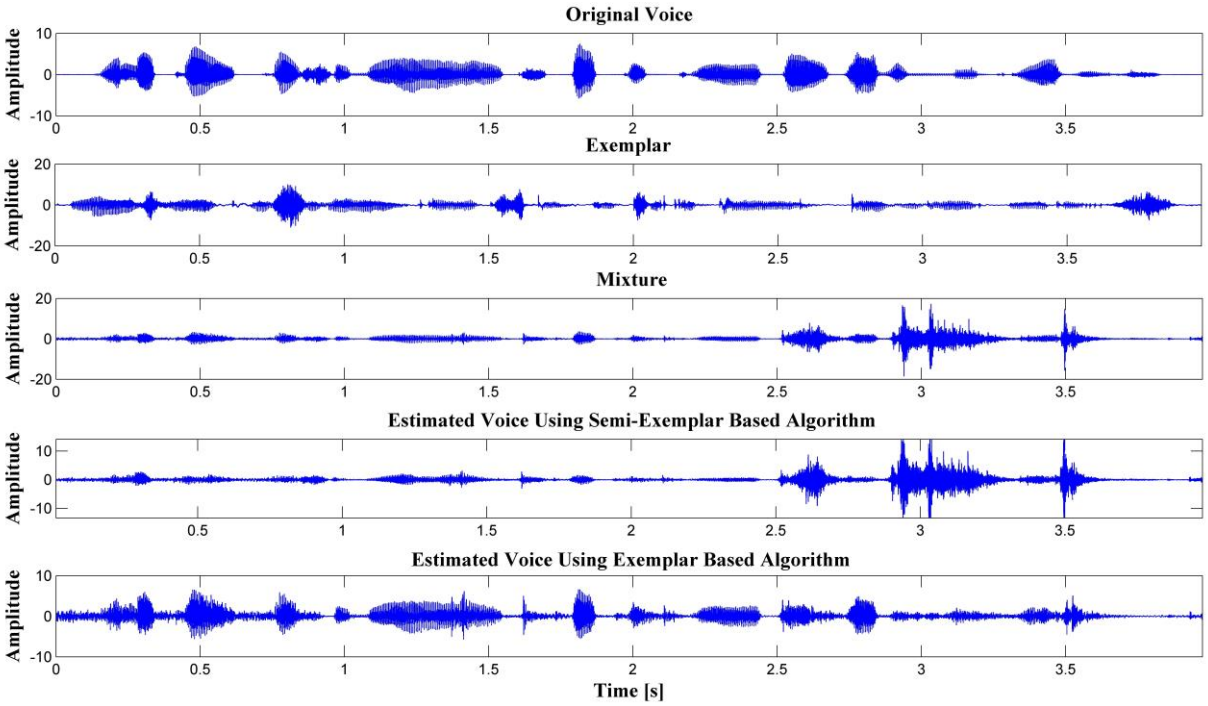


Figure 5.8: Waveform of the original speech, exemplar, and the estimated speech by using the proposed algorithms.

5.5 Multistage of The Exemplar Based Algorithm

This section will take advantage from the characteristics of the spectrograms under different windows length, where the length of the window has an effect on the separation performance [50, 93]. In the short window the spectrogram of the percussive musical instruments is continue in the spectral direction and discrete in the temporal direction, which act differently from the spectrogram of the speech and pitch musical instrument that continue in temporal direction and discrete in spectral direction. Therefore, the percussive instruments are distinct from the pitched instruments and speech signals which make them more separable under short window.

Furthermore, the speech signals act like percussive musical instruments in long window. Therefore, two stages of the proposed exemplar based algorithm with short (512-samples) and long (1024-samples) windows will be proposed, as shown in Figure 5.9. In the short window's stage the frequency basis of the percussive instruments will be removed (as it will act differently from the pitched and speech signals), while in the long window's stage the frequency basis of pitch musical instruments and the speech signal will be separated (as the speech signal act differently from pitched instruments).

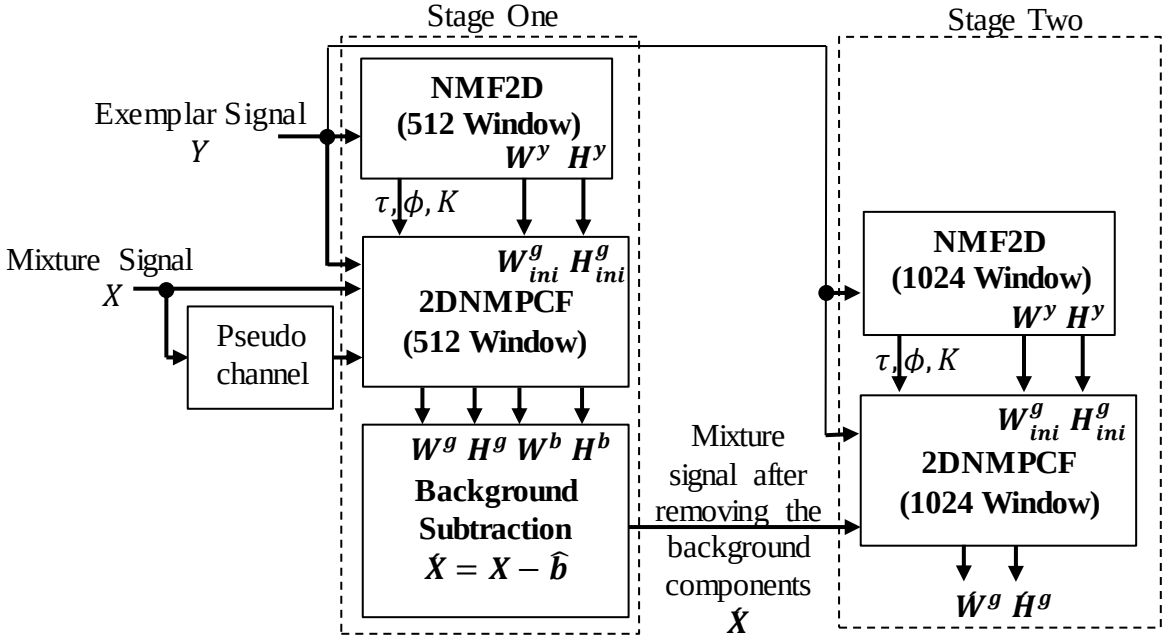


Figure 5.9: High level presentation of the multistage of the exemplar based algorithm.

The background subtraction is done in the spectral domain as follows

$$\boldsymbol{x}'_{fn} = \boldsymbol{x}_{fn} - \hat{\boldsymbol{b}}_{fn} \tag{5.41}$$

where $\hat{b}_{fn}$ is the estimated background signal as in eqn. (5.13), and $\boldsymbol{x}'_{fn}$ is the mixture signal after removing the estimated background signal. By following this procedure the results for the "speech and music"[5] group is tabulated in Table 5.5. It can be seen that the average SDR has been increased by 0.72 dB. Furthermore, the spectrogram of the mixture, the original speech and background sources, and their estimate for both stages are shown in Figure 5.10. It can be seen that the spectrograms of the percussive instruments has been separated clearly, as shown in the spectrograms of the estimated background of stage one, while the speech and pitched instruments still mixed together (they have the same characteristics in the short window) as shown in the spectrograms of the estimated speech signal of stage one. While, the spectrograms of the second stage show that the speech signal is clearer than the spectrogram of the first stage where a part (ideally all) of the spectrograms of the pitched instruments has been removed, as the speech signal act like percussive instruments in long window. Thus the second stage enhanced the separation performance by removing the pitched spectrograms that the first stage did not remove.

Table 5.5

Average SDRs of the 10 mixtures with their different 12 exemplars for the Multistage of the Exemplar based algorithm.

SNR= -5dB	SPEECH + Music First Stage Window=512	SPEECH + Music Second Stage Window=1024
Proposed Exemplar based algorithm	2.38	3.10

[5] This approach cannot be applied on "Speech+Effects" group as the effects have random spectral features which cannot be modeled with short or long window.

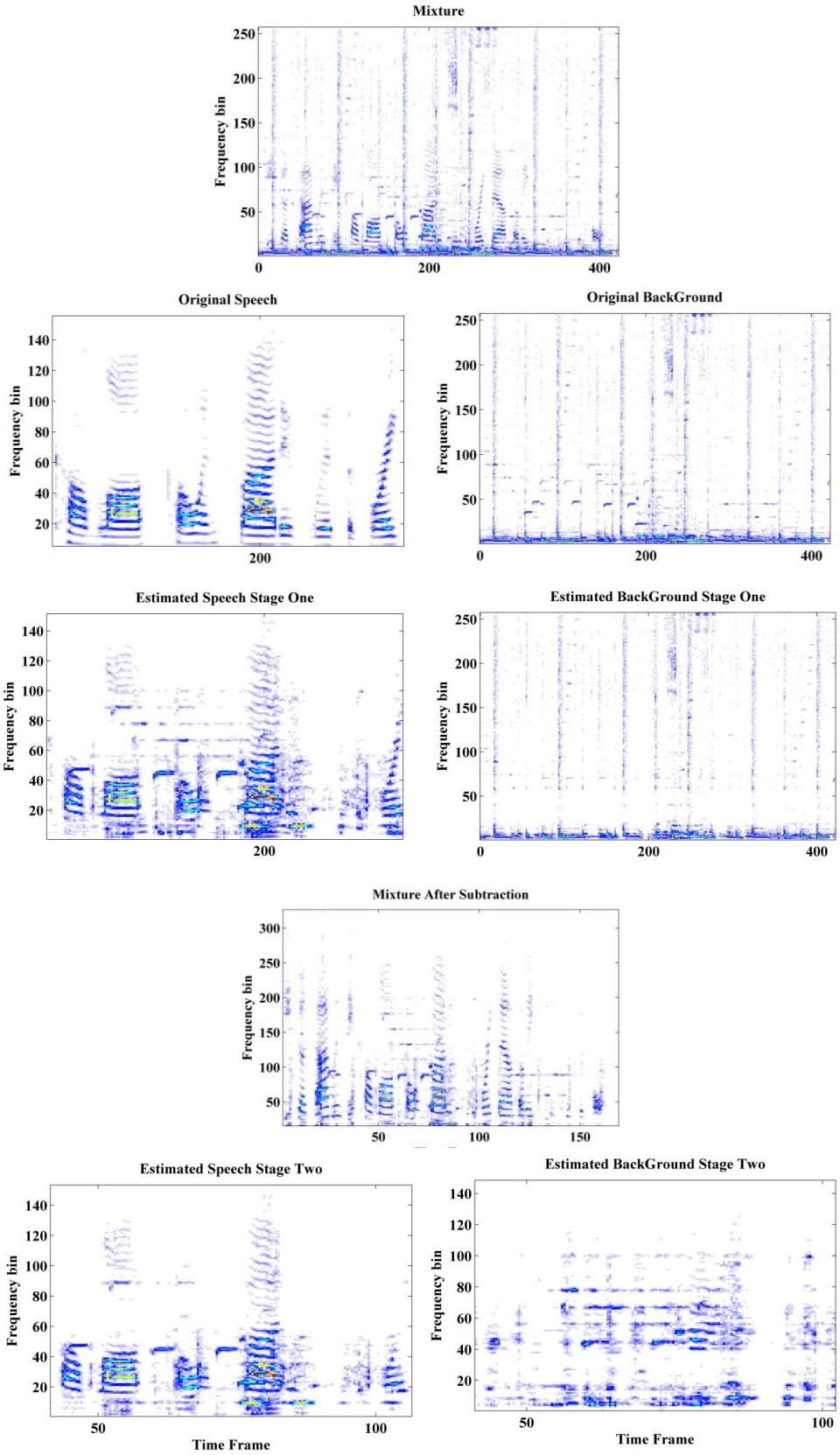


Figure 5.10: Spectrograms of the original speech and original background, and their estimate for the first and the second stage of the proposed multistage algorithm.

5.6 Summary

In this chapter two algorithms for the underdetermined informed source separation, namely the semi-exemplar based algorithm and the exemplar-based algorithm have been proposed. These algorithms are based on the two dimensional matrix factorization techniques, the NMF2D and the proposed 2DNMPCF. These two dimensional factorization techniques have the advantage of describing the targeted signal by describing the pitch and temporal changes of that signal, which cannot carry out by the NMF or NMPCF. For faster convergence and better performance both of the algorithms are modelled by the GEM-MU algorithm with pseudo-stereo channel and with adaptive sparsity. It has been shown that the proposed algorithm outperformed the NMPCF algorithm and the structural GSMM algorithm. Furthermore, it has been shown that by using a multistage of the proposed exemplar based algorithm the overall performance can be enhanced.

CHAPTER 6

CONCLUSIONS AND FUTURE WORKS

In this chapter the contributions will be summarized and the future work of the thesis will be discussed.

The work in this thesis has fulfilled the aims of the research set out in Chapter 1. The advantage of the two dimensional factorization techniques over the matrix factorization techniques paved the way for the development of four underdetermined separation algorithms as follows: Firstly, the NMF2D has been developed for the convolutive underdetermined mixture. The convolutive NMF2D achieved better performance than the NMF due to its ability in addressing both the temporal and spectral change of the signal. Secondly, the *K-wNTF2D* has been developed to address a more realistic case in blind source separation which is the high-reverberant convolutive underdetermined mixture. The *K-wNTF2D* maintains its high level performance in the high reverberation environment due to its ability in modeling both the spectral and temporal changes, and the spatial covariance matrix. Thirdly, the semi-exemplar based and the full exemplar based algorithms have been developed. These two algorithms have been dedicated to the informed source separation. Both algorithms achieved better performance than the conventional methods due to their ability in describing the exemplar which cannot be carried out by the conventional methods. The above algorithms have been modeled with adaptive/ variable sparsity in order to avoid the over or under sparseness. Finally, essential algorithms; the Gamma-exponential algorithm and the SVD2D algorithm have been developed to support the separation algorithms by estimating their model order to avoid over shifting and initialize them to prevent them from stuck in local minima or even diverge.

In Chapter 2, the theories which paved the understanding of the work that carried out in the rest of the chapters have been explained. The overview of the blind source separation and informed source separation was presented. The motivation behind considering the convolutive mixture instead of the instantaneous one, and the motivation behind tackling the high-reverberant mixture were discussed. Furthermore, the potential of going from blind to informed source separation was shown. Moreover, the factorization techniques such as the NMF, NMFD, NMF2D, NTF, and

NTF2D have been explained. Additionally, the parameters those effects on the separation performance have been highlighted and discussed.

In Chapter 3 the NMF2D with adaptive sparsity has been proposed to solve the underdetermined convolutive mixture. The impetus behind this is that the NMF2D is more powerful than the NMF due to its ability in addressing both the temporal and spectral change of the signal. Furthermore the adaptation of the adaptive sparsity in this model gives it the capability to control the degree of sparseness over the activation matrix of the NMF2D. Also the proposed Gamma-Exponential process algorithm ensured that a suitable number of frequency basis and convolutive parameters are assigned to each source in order to avoid over-shifting. Moreover, the proposed Gamma-Exponential process algorithm has been used to initialize the tensors of the NMF2D in order to avoid the separation algorithm to stick in local minima or to even diverge. Additionally it has been shown that using different windows length will match the characteristics of the sources to be estimated and this will leads to better representation for them. The significant improvements in the results in term of the SDRs showed that the proposed algorithm is better than the conventional methods that based on NMF or NTF and it is more flexible.

In Chapter 4 the *K*-wNTF2D with variable sparsity has been proposed to solve the more realistic case in blind source separation which is the underdetermined high-reverberant convolutive mixture. The motivation behind proposing the *K*-wNTF2D in this chapter is due to its ability in modeling both the spectral and temporal changes, and the spatial covariance matrix that address the high reverberation problem. Furthermore, the variable sparsity that derived from the Gibbs distribution has been integrated with this model in order to provide a tractable approach that adapts each sparse parameter for every temporal code in the *K*-wNTF2D model. Moreover, the SDV2D initialization method has been proposed in this chapter to initialize the tensors of the *K*-wNTF2D in order to avoid divergence or sticking in undesired minima. The experiments in this chapter showed that the proposed algorithm maintains its high level performance in the high reverberation environment, where it achieved higher performance than the Full-Rank NMF and the multi-level NMF.

In Chapter 5 two algorithms which based on the NMF2D and the proposed 2DNMPCF models have been proposed to solve the informed source separation, namely the semi-exemplar based algorithm and the exemplar-based algorithm. Both algorithms were depending on the provided exemplar that emulates the speech signal to be separated. The impetus behind using the NMF2D

and the proposed 2DNMPCF models here is due to their ability in describing the exemplar which cannot carry out by the NMF or NMPCF. Furthermore, to enhance the performance of the separation a multistage of the exemplar-based algorithm has been proposed. The first stage of the proposed multistage algorithm has been modeled with short window to remove the percussive sources, while the second stage has been modeled with long window in order to separate the speech signal from the pitched one. Throughout the experiments it has been shown that the proposed algorithms outperformed the NMPCF algorithm and the structural GSMM algorithm.

All the proposed algorithms in the thesis have been compared in terms of the type of audio source separation, cost function, update of the parameters, and sparsity, as shown in Table 6.1. In terms of the computational complexity, the *K-wNTF2D* requires the most computational resources compared to the Convolutive NMF2D, and the Exemplar-Based algorithm and the Semi-Exemplar based algorithm, as it avoided the narrowband assumption by considering the full-rank spatial covariance matrix instead of rank one matrix. The performance of proposed algorithms depends in some degrees on the prior information provided to them. In the Convolutive NMF2D and the *K-wNTF2D,* the prior information takes the form of how the specific source is modelled. In the Exemplar-Based algorithm and the Semi-Exemplar based algorithm, the prior information takes the form of how similar the exemplar is to the targeted signal to be separated.

Table 6.1: Summary of the proposed algorithms.

Algorithm	**Type of Audio Source Separation**	**Cost Function**	**Update**	**Sparsity**
Convolutive NMF2D	Blind	IS	GEM-MU	Adaptive
K-wNTF2D	Blind	IS	GEM-MU	Variable Sparsity
Exemplar Based Algorithm	Informed	IS	GEM-MU	Adaptive
Semi-Exemplar Based Algorithm	Informed	IS	GEM-MU	Adaptive
2DNMPCF	Informed	-	-	-
Gamma Exponential Process	Blind	-	-	-
SVD2D	Blind	-	-	-

6.1 Future Works

In this section some research areas in both the blind and informed source separation will be presented for future investigation with the goal of developing novel algorithms.

6.1.1 Harmonic and Percussive Source Separation

Recently many researches have been dedicated to separate the harmonic and percussive sources [114-121] due to the distinct characteristics of their spectrogram. Harmonic instruments are smooth and continue in temporal direction and discrete in spectral direction, and the opposite for the percussive instruments. Inspired by the harmonic and percussive source separation algorithm [114], an algorithm will be proposed for separating the harmonic and percussive musical instruments by customizing the two dimensional matrix factorization techniques to match the characteristics of these signals. Instead of using two stages with different windows to match the characteristics of these signals one stage only will be proposed but with different convolutive parameters for each source, such as using high τ and low ϕ for the harmonic signal and low τ and high ϕ for percussive signal, as shown in Figure 6.1. The idea is to use the proposed Gamma-Exponential process to estimate the convolutive parameters, and then use these parameters to model the mixture as follows

$$x_{fn} = \boldsymbol{a}_{1,f} p_{fn} + \boldsymbol{a}_{2,f} h_{fn} \tag{6.1}$$

where $a_{j,f}$ (j=1 or 2) is the time invariant mixing matrix, p_{fn} is the percussive signal and h_{fn} is the harmonic signal, which can be expressed by K complex-valued latent components, i.e.,

$$p_{fn} = \sum_{k=1}^{K} c_{k,fn}^{p} \tag{6.2a}$$

and

$$h_{fn} = \sum_{k=1}^{K_j} c_{k,fn}^{h} \tag{6.2b}$$

which can be modeled as realization of proper complex zero-mean variables:

$$c_{k,fn}^{p} = \mathcal{N}_c\left(0, \sum_{\tau=0}^{\tau_{min}} \sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k,\tau}^{p} h_{k,n-\tau,\phi}^{p}\right) \tag{6.3}$$

$$c_{k,fn}^{h} = \mathcal{N}_c\left(0, \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{min}} w_{f-\phi,k,\tau}^{h} h_{k,n-\tau,\phi}^{h}\right) \tag{6.4}$$

where $\mathcal{N}_c(\mu, \Sigma)$ is proper complex Gaussian distribution [94], $w_{f-\phi,k,\tau}^{p}$ and $w_{f-\phi,k,\tau}^{h}$ represent the spectral basis of the percussive and harmonic sources, respectively, and $h_{k,n-\tau,\phi}^{p}$ and $h_{k,n-\tau,\phi}^{h}$ represent the temporal code for each spectral basis element of the percussive source and harmonic sources, respectively, for $f = 1, \ldots, F, n = 1, \ldots, N,$ and $k = 1, \ldots K$. It can be seen that eqn. (6.3) has ϕ_{max} and τ_{min} in order to match the characteristics of the percussive instruments, while eqn. (6.4) has the opposite values of the convolutive parameters in order to match the characteristics of the harmonic instruments.

The parameters $A, C, \boldsymbol{W}$, and $\boldsymbol{H}$ will be estimated via the posterior probability

$$P(C, \boldsymbol{W}, \boldsymbol{H} | X, A) = \frac{P(X|C, A)P(C|\boldsymbol{W}, \boldsymbol{H})}{P(X|A)} \tag{6.5}$$

and their log-posterior is given by

$$\log P(C, \boldsymbol{W}, \boldsymbol{H} | X, A) = \log P(X|C, A) + \log P(C|\boldsymbol{W}, \boldsymbol{H}) + const \tag{6.6}$$

Finally the GEM-MU [80] algorithm can be applied to estimate the percussive and harmonic sources.

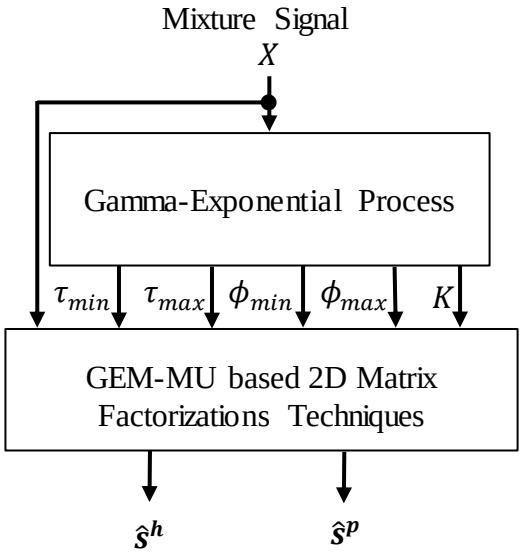


Figure 6.1: High level presentation of the harmonic and percussive source separation algorithm.

6.1.2 Coding Based Informed Source Separation

Ozerov et al. [91] proposed a coding based informed source separation system that based on the nonnegative tensor factorization (NTF) that able to reach any quality (in expense of bandwidth as in source coding) that the conventional methods cannot reach, as it take advantage from both the informed source separation and source coding. This system consists of two stages; the encoding stage and the decoding stage. In the encoding stage the side information that contains the sources, the sources model parameters (that represented by NTF model, i.e., Q, W, and H matrices), and the perceptual model parameters are encoded and transmitted with the mixture. In the decoding stage the sources are reconstructed by depending on the received mixture signal and the side information.

Inspired by this system the *K*-wNTF2D model can be used instead of the NTF as it is more powerful. Furthermore, the parameters which describe the sources (as described in Chapter 5) can be transmitted with the side information, i.e., transmitting the number of components and convolutive parameters. Thus the separation performance will be significantly improved but in the cost of the bitrate.

6.1.3 Complex Two Dimensional Matrix Factorization Techniques

It has been shown in [3] that incorporating the phase information into the NMF has the potential to increase the separation performance, therefore, a full rank complex 2D matrix factorization techniques will be proposed to model the spectral variance of the source, takes into account the phase information of the source spectral variance, and model the full rank of the spatial covariance matrix. Thus the issue of high-reverberant mixture will be addressed through the full rank spatial covariance matrix, and better separation performance will be achieved through the phase information. This model can be realized as follows

Let $\boldsymbol{x}(t)$ be the observed multichannel signal that can be expressed in time domain as

$$x_i(t) = \sum_{j=1}^{J} c_{i,j}(t) + b_i(t) \tag{6.7}$$

where $i = 1,2,\dots I,\ x_i(t) \in \mathbb{R}, \mathrm{t} = 1,\dots,\mathrm{T}$ is the receiving signal from the *i*-th microphone, $c_{i,j}(t) \in \mathbb{R}$ is the spatial image of the source signal *j* and channel *i*, *J* is the number of

sources, and $b_i(t) \in \mathbb{R}$ is some additive noise. The spatial image of the source $c_{i,j}(t)$ can be expressed as

$$c_{i,j}(t) = \sum_{\tau=0}^{L-1} a_{i,j}(\tau) s_j(t-\tau) \tag{6.8}$$

where $a_{i,j}(t) \in \mathbb{R}$ is the finite-impulse response of some (causal) filter, L is the filter length, and $s_j(t) \in \mathbb{R}$ is the original source signal.

By substituting eqn. (6.8) into eqn. (6.7), and assuming that the mixing channel is time-invariant then, the STFT of eqn. (6.7) becomes

$$x_{i,f,n} = \sum_{j=1}^{J} c_{i,j,f,n} + b_{i,f} \tag{6.9}$$

where $\boldsymbol{x}_{f,n} = [x_{1,f,n} \quad \cdots \quad x_{I,f,n}]^H$, and $x_{i,f,n}$, $c_{i,j,f,n}$, $b_{i,f,n}$ are the complex-valued STFT of $x_i(t)$, $c_{i,j}(t)$, and $b_i(t)$, respectively. The term $f = 1, 2, \ldots, F$ is the frequency bin index, and $n = 1, 2, \ldots, N$ is the time frame index. The spectral covariance matrix of $c_{i,j,f,n}$ defined as

$$\Sigma_{j,f,n}^{(c)} = E\left[\boldsymbol{c}_{j,f,n} \boldsymbol{c}_{j,f,n}^H\right] \tag{6.10a}$$

can be expressed as

$$\Sigma_{j,f,n}^{(c)} = \Sigma_{j,f}^{(a)} v_{j,f,n} \, e^{\sqrt{-1}\alpha_{j,f,n}} \tag{6.10b}$$

where $\Sigma_{j,fn}^{(c)} \in \mathbb{C}^{I \times I}$ is the spectral covariance matrix of the j-th source image, $\boldsymbol{\Sigma}_{j,f}^{(a)}$ is the time-invariant spatial covariance matrix of the j-th source, $v_{j,f,n} \in \mathbb{R}$ is the j-th source variance in the STFT domain, and $\alpha_{j,f,n} \in \mathbb{C}$ is the time-varying phase spectrum [122] to explicitly model the phase in $v_{j,f,n}$. The term $\sqrt{-1}$ is adopted to represent the imaginary component. The j-th source variance can be expressed as

$$v_{j,f,n} = \sum_{k=1}^{K} \sum_{\tau=0}^{\tau_{max}} \sum_{\phi=0}^{\phi_{max}} w_{f-\phi,k}^{\tau,j} h_{k,n-\tau}^{\phi,j} \tag{6.11}$$

where K is the number of components or frequency basis assigned to the j-th source, τ_{max} and ϕ_{max} are the maximum number of the convolutive parameters τ and ϕ respectively. $w_{f,k}^{\tau,j}$ represents the k-th spectral basis of the j-th source, and $h_{k,n}^{\phi,j}$ represents the k-th temporal code for each spectral basis element of the j-th source, for $f = 1, \dots, F, n = 1, \dots, N$, and $j = 1, \dots, J$.

The full-rank spectral covariance matrix of $\boldsymbol{x}_{f,n}$ in eqn. (6.9) can be expressed as

$$\begin{aligned} \Sigma_{f,n}^{(x)} &= E\left[\boldsymbol{x}_{f,n}\boldsymbol{x}_{f,n}^{H}\right] \\ &= \sum_{j=1}^{J} \Sigma_{j,f,n}^{(c)} + \Sigma_{f}^{(b)} \\ &= \sum_{j=1}^{J} \Sigma_{j,f}^{(a)} v_{j,f,n} e^{\sqrt{-1}\alpha_{j,f,n}} + \Sigma_{f}^{(b)} \end{aligned} \tag{6.12}$$

where $\Sigma_{f}^{(b)}$ is the time invariant noise covariance matrix.

The spatial image of the sources needs to be modeled as realization of complex distribution that consider the complex signals in order to optimize the model parameters $\Theta = \{\boldsymbol{W}, \boldsymbol{H}, \Sigma^{(a)}, \Sigma^{(b)}, \boldsymbol{\Lambda}, \alpha\}$. Then the separation can be carried out by using the GEM-MU algorithm.

Furthermore, the variable sparsity can be proposed to estimate the sparsity, Gamma-Exponential process algorithm can be proposed to estimate the number of components, convolutive parameters, and initialize its tensors.

6.1.4 Totally Blind Source Separation System

There is some prior information that needed in the blind source separation methods in order to carry out the separation. One of the prior information is the number of sources. Therefore, the blind source separation is not totally blind as prior information is required. The idea of this section is to develop a system that able to detect the number of sources in order to achieve a totally blind system, e.g., by using the Direction Estimation of Mixing matrix (DEMIX) algorithm [123]. This algorithm proposed for both instantaneous and anechoic systems, therefore, it needs to be developed to deal with convolutive mixture. To be able to deal with convolutive mixture it will

required from the algorithm to distinguish between the sources and their reverberation or an echo cancellation is required to be added to this system, which is not forward if the system is suggested for the high reverberation environment. If this system is developed then the whole picture of the source separation can be realized blindly, as there will be no need for prior information about the number of sources. The suggested system can be realized as in Figure 6.2 below.

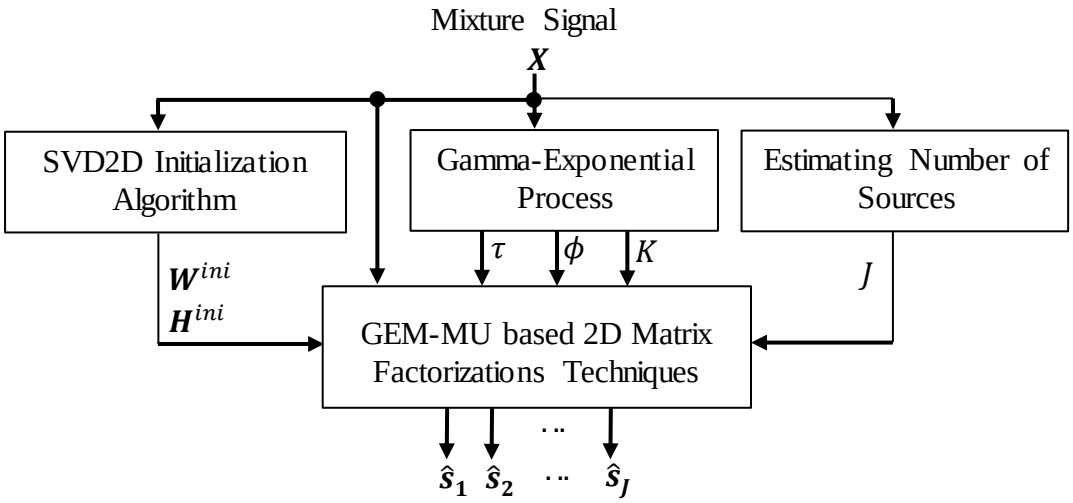


Figure 6.2: Suggested blind source separation system.

APPENDIX A

Derivation of the conditional expectation of the natural statistics

The *posterior* $P(\boldsymbol{c}_{j,f,n}|\boldsymbol{x}_{f,n})$ can be written as

$$P(\boldsymbol{c}_{j,f,n}|\boldsymbol{x}_{f,n}) = \frac{P(\boldsymbol{x}_{f,n}, \boldsymbol{c}_{j,f,n})}{P(\boldsymbol{x}_{f,n})}$$

$$= \frac{\left(\pi^{I+1} det\Sigma_{j,f,n}^{(joint)}\right)^{-1}}{\left(\pi^{I} det\Sigma_{f,n}^{(x)}\right)^{-1}} \frac{exp\left\{-\begin{bmatrix}\boldsymbol{x}_{f,n}\\ \boldsymbol{c}_{j,f,n}\end{bmatrix}^{H} \Sigma_{j,f,n}^{(joint)^{-1}} \begin{bmatrix}\boldsymbol{x}_{f,n}\\ \boldsymbol{c}_{j,f,n}\end{bmatrix}\right\}}{exp\left\{-\boldsymbol{x}_{f,n}^{H}\Sigma_{f,n}^{(x)^{-1}}\boldsymbol{x}_{f,n}\right\}}$$

$$= \left(\pi det\Gamma_{j,f,n}\right)^{-1} exp\{-\psi_{j,f,n}\} \tag{A.1}$$

where

$$\Gamma_{j,f,n} = \Sigma_{j,f,n}^{(c)} - \Sigma_{j,f,n}^{(xc)}\Sigma_{f,n}^{(x)^{-1}}\Sigma_{j,f,n}^{(cx)} \tag{A.2}$$

$$\Sigma_{j,f,n}^{(joint)^{-1}} = \begin{bmatrix}\left(\Sigma_{f,n}^{(x)} - \Sigma_{j,f,n}^{(xc)}\Sigma_{j,f,n}^{(c)^{-1}}\Sigma_{j,f,n}^{(cx)}\right)^{-1} & -\Sigma_{f,n}^{(x)^{-1}}\Sigma_{j,f,n}^{(xc)}\Gamma_{j,f,n}^{-1}\\ -\Sigma_{f,n}^{(x)^{-1}}\Sigma_{j,f,n}^{(cx)}\Gamma_{j,f,n}^{-1} & \Gamma_{j,f,n}^{-1}\end{bmatrix} \tag{A.3}$$

$$\psi_{j,f,n} = \begin{bmatrix}\boldsymbol{x}_{f,n}\\ \boldsymbol{c}_{j,f,n}\end{bmatrix}^{H} \Sigma_{j,f,n}^{(joint)^{-1}} \begin{bmatrix}\boldsymbol{x}_{f,n}\\ \boldsymbol{c}_{j,f,n}\end{bmatrix} - \boldsymbol{x}_{f,n}^{H}\Sigma_{f,n}^{(x)^{-1}}\boldsymbol{x}_{f,n}$$

$$= \left(\boldsymbol{c}_{j,f,n} - \Sigma_{j,f,n}^{(cx)}\Sigma_{f,n}^{(x)^{-1}}\boldsymbol{x}_{f,n}\right)^{H} \Gamma_{j,f,n}^{-1}\left(\boldsymbol{c}_{j,f,n} - \Sigma_{j,f,n}^{(cx)}\Sigma_{f,n}^{(x)^{-1}}\boldsymbol{x}_{f,n}\right) \tag{A.4}$$

$$\Sigma_{j,f,n}^{(xc)} = E\left[\boldsymbol{x}_{f,n}\boldsymbol{c}_{j,f,n}^{H}\right] = E\left[(\boldsymbol{c}_{j,f,n} + \boldsymbol{b}_{f,n})\boldsymbol{c}_{j,f,n}^{H}\right]$$

$$= E\left[\boldsymbol{c}_{j,f,n}\boldsymbol{c}_{j,f,n}^{H}\right] + E\left[\boldsymbol{b}_{f,n}\boldsymbol{c}_{j,f,n}^{H}\right] = \Sigma_{j,f,n}^{(c)} \tag{A.5}$$

where $E\left[\boldsymbol{b}_{f,n}\boldsymbol{c}_{j,f,n}^{H}\right] = \boldsymbol{0}$ as they are uncorrelated. Thus

$$P(\boldsymbol{c}_{j,f,n}|\boldsymbol{x}_{f,n}) = \left(\pi det\Gamma_{j,f,n}\right)^{-1} exp\left(\left(\boldsymbol{c}_{j,f,n} - \Sigma_{j,f,n}^{(c)}\Sigma_{f,n}^{(x)^{-1}}\boldsymbol{x}_{f,n}\right)^{H} \Gamma_{j,f,n}^{-1}\left(\boldsymbol{c}_{j,f,n} - \Sigma_{j,f,n}^{(c)}\Sigma_{f,n}^{(x)^{-1}}\boldsymbol{x}_{f,n}\right)\right) \tag{A.6}$$

Comparing eqn. (A.6) with eqn. (4.9), eqns. (4.16)-(4.18) will be obtained. By following the same procedure for the noise, eqns. (4.19)-(4.21) will be obtained.

REFERENCES

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

[2] J. L. Durrieu, G. Richard, B. David, and C. Fevotte, "Source/Filter Model for Unsupervised Main Melody Extraction From Polyphonic Audio Signals," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 3, pp. 564-575, Mar, 2010.

[3] B. J. King, and L. Atlas, "Single-Channel Source Separation Using Complex Matrix Factorization," *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, no. 8, pp. 2591-2597, Nov, 2011.

[4] B. King, and L. Atlas, "Single-Channel Source Separation Using Simplified-Training Complex Matrix Factorization," in 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2010, pp. 4206-4209.

[5] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel Extensions of Non-Negative Matrix Factorization With Complex-Valued Data," *IEEE Transactions on Audio Speech and Language Processing*, vol. 21, no. 5, pp. 971-982, May, 2013.

[6] K. Takeda, H. Kameoka, H. Sawada, S. Araki, S. Miyabe, T. Yamada, and S. Makino, "Underdetermined BSS With Multichannel Complex NMF Assuming W-Disjoint Orthogonality of Source," in IEEE Region 10 Conference Tencon, 2011, pp. 413-416.

[7] H. Sawada, S. Araki, and S. Makino, "Underdetermined Convolutive Blind Source Separation via Frequency Bin-Wise Clustering and Permutation Alignment," *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, no. 3, pp. 516-527, Mar, 2011.

[8] A. Ozerov, and C. Fevotte, "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, Mar, 2010.

[9] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model," *IEEE Transactions on Audio Speech and Language Processing*, vol. 18, no. 7, pp. 1830-1840, Sep, 2010.

[10] N. Q. K. Duong, E. Vincent, and R. Gribonval, "Under-Determined Reverberant Audio Source Separation Using Local Observed Covariance and Auditory-Motivated Time-Frequency Representation," in 9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10), 2010, pp. 73-80.

[11] S. Arberet, A. Ozerov, R. Gribonval, and F. Bimbot, “Blind Spectral-GMM Estimation for Underdetermined Instantaneous Audio Source Separation,” in Independent Component Analysis and Signal Separation, Proceedings, 2009, pp. 751-758.

[12] M. Parvaix, and L. Girin, “Informed Source Separation of Linear Instantaneous Under-Determined Audio Mixtures by Source Index Embedding,” *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, no. 6, pp. 1721-1733, Aug, 2011.

[13] R. Zdunek, “Convolutive Nonnegative Matrix Factorization with Markov Random Field Smoothing for Blind Unmixing of Multichannel Speech Recordings,” *Advances in Nonlinear Speech Processing*, vol. 7015, pp. 25-32, 2011.

[14] R. Zdunek, “Improved Convolutive and Under-Determined Blind Audio Source Separation with MRF Smoothing,” *Cognitive Computation*, vol. 5, no. 4, pp. 493-503, Dec, 2013.

[15] N. Q. K. Duong, A. Ozerov, and L. Chevallier, “Temporal annotation-based audio source separation using weighted nonnegative matrix factorization,” in 4th IEEE International Conference on Consumer Electronics - Berlin (ICCE-Berlin 2014), Berlin, Germany, 2014, pp. 220 - 224.

[16] S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, “Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation,” in Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on, 2010, pp. 1-4.

[17] J. L. Yao, X. N. Yang, J. D. Li, and Z. Li, “An MRC Based Over-determined Blind Source Separation Algorithm,” in 2010 IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), 2010, pp. 309-313.

[18] A. M. Darsono, G. Bin, W. L. Woo, and S. S. Dlay, “Nonlinear single channel source separation,” in Communication Systems Networks and Digital Signal Processing (CSNDSP), 2010 7th International Symposium on, 2010, pp. 507-511.

[19] Y. Xianchuan, H. Dan, and X. Jindong, "Blind Source Separation: Theory and Applications," JohnnWiley & Sons, 2014, p. 416.

[20] A. Liutkus, J. L. Durrieu, L. Daudet, and G. Richard, "An Overview of Informed Audio Source Separation," *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2013, pp. 1-4.

[21] X. L. Zhang, and D. Wang, "A Deep Ensemble Learning Method for Monaural Speech Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, no. 99, pp. 1-1, 2016.

[22] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint Optimization of Masks and Deep Recurrent Neural Networks for Monaural Source Separation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 23, no. 12, pp. 2136-2147, Dec, 2015.

[23] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-Voice Separation Using Deep Recurrent Neural Networks," in International Society for Music Information Retrieval Conference (ISMIR), 2014.

[24] H. Po-Sen, K. Minje, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, 2014, pp. 1562-1566.

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

[26] M. Morup, and M. N. Schmidt, *Sparse non-negative matrix factor 2-D deconvolution*, Tech. Rep Technical University of Denmark, Copenhagen, Denmark, 2006.

[27] B. Gao, W. L. Woo, and S. S. Dlay, "Unsupervised Single-Channel Separation of Nonstationary Signals Using Gammatone Filterbank and Itakura-Saito Nonnegative Matrix Two-Dimensional Factorizations," *IEEE Transactions on Circuits and Systems I-Regular Papers*, vol. 60, no. 3, pp. 662-675, Mar, 2013.

[28] B. Gao, W. L. Woo, and S. S. Dlay, "Variational Regularized 2-D Nonnegative Matrix Factorization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 5, pp. 703-716, May, 2012.

[29] B. Gao, W. L. Woo, and S. S. Dlay, "Nonnegative matrix factorization for single channel source separation " *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 989-1001, 2011.

[30] M. Mørup, and M. N. Schmidt, *Sparse Non-negative Tensor 2D Deconvolution (SNTF2D) for multi channel time-frequency analysis*, Technical Report, Technical University of Denmark, DTU, 2006.

[31] B. Gao, W. L. Woo, and C. Khor, "Cochleagram-based audio pattern separation using two-dimensional non-negative matrix factorization with automatic sparsity adaptation," *Journal of the Acoustical Society of America*, vol. 135, no. 3, pp. 1171-1185, Mar, 2014.

[32] B. Gao, W. L. Woo, and B. W. K. Ling, "Machine Learning Source Separation Using Maximum A Posteriori Nonnegative Matrix Factorization," *IEEE Transactions on Cybernetics*, vol. 44, no. 7, pp. 1169-1179, Jul, 2014.

[33] 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, 01, 1994.

[34] P. Paatero, "Least squares formulation of robust non-negative factor analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 37, no. 1, pp. 23-35, May, 1997.

[35] D. D. Lee, and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788-791, Oct 21, 1999.

[36] D. D. Lee, and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems 13*, vol. 13, pp. 556-562, 2001.

[37] Y. Li, and A. Ngom, "The non-negative matrix factorization toolbox for biological data mining," *Source Code for Biology and Medicine*, vol. 8, no. 1, pp. 1-15, 2013.

[38] Y. Gong, D. Cui, G. Yu, and J. Xiong, "An improved image watermarking algorithm based on NMF and DWT," in International Conference on Information and Network Security, ICINS 2014, pp. 6-11.

[39] R. Sandler, and M. Lindenbaum, "Nonnegative Matrix Factorization with Earth Mover's Distance Metric for Image Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1590-1602, Aug, 2011.

[40] M. Das Gupta, and J. Xiao, "Non-negative Matrix Factorization as a Feature Selection Tool for Maximum Margin Classifiers," in 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 2841–2848.

[41] A. Huang, "Face Recognition based on Non-Negative Matrix Factorization with Alpha Divergence," *International Journal of Advancements in Computing Technology*, vol. 4, no. 18, 2012.

[42] S. Essid, and C. Fevotte, "Smooth Nonnegative Matrix Factorization for Unsupervised Audiovisual Document Structuring," *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 415-425, Feb, 2013.

[43] T. D. Hien-Thanh, N. Quoc-Cuong, N. Cong-Phuong, T. Thanh-Huan, and Q. K. D. Ngoc, "Speech enhancement based on nonnegative matrix factorization with mixed group sparsity constraint," in Proceedings of the Sixth International Symposium on Information and Communication Technology, Hue City, Viet Nam, 2015, pp. 247-251.

[44] C. Bilen, A. Ozerov, and P. Perez, "Joint Audio Inpainting and Source Separation," *Latent Variable Analysis and Signal Separation: 12th International Conference, LVA/ICA 2015, Liberec, Czech Republic, August 25-28, 2015, Proceedings*, pp. 251-258, Cham: Springer International Publishing, 2015.

[45] C. Bilen, A. Ozerov, and P. Pérez, "Audio Inpainting, Source Separation, Audio Compression. All with a Unified Framework Based on NTF Model," in MissData 2015, Rennes, France, 2015.

[46] C. Bilen, A. Ozerov, and P. Pérez, "Audio declipping via nonnegative matrix factorization," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, United States, 2015, pp. 1 - 5.

[47] A. Ozerov, C. Bilen, and P. Pérez, "Multichannel audio declipping," in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'16), Shanghai, China, 2016.

[48] J. Traa, P. Smaragdis, N. D. Stein, and D. Wingate, "Directional NMF for joint source localization and separation," in 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015, pp. 1-5.

[49] C. Fevotte, N. Bertin, and J. L. Durrieu, "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *Neural Computation*, vol. 21, no. 3, pp. 793-830, Mar, 2009.

[50] B. King, C. Fevotte, and P. Smaragdis, "Optimal Cost Function and Magnitude Power for NMF-Based Speech Separation and Music Interpolation," in 2012 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2012, pp. 1 - 6.

[51] V. Y. F. Tan, and C. Fevotte, "Automatic Relevance Determination in Nonnegative Matrix Factorization with the beta-Divergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1592-1605, Jul, 2013.

[52] A. Ozerov, E. Vincent, and F. Bimbot, "A General Flexible Framework for the Handling of Prior Information in Audio Source Separation," *IEEE Transactions on Audio Speech and Language Processing*, vol. 20, no. 4, pp. 1118-1133, May, 2012.

[53] J. Fritsch, and M. D. Plumbley, "Score Informed Audio Source Separation Using Constrained Nonnegative Matrix Factorization and Score Synthesis," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 888-891.

[54] D. FitzGerald, "User Assisted Separation Using Tensor Factorisations," in 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), 2012, pp. 2412-2416.

[55] A. Liutkus, and P. Leveau, "Separation of Music+Effects sound track from several international versions of the same movie," in AES 128th Convention, London, United Kingdom, 2010.

[56] T. Gerber, M. Dutasta, L. Girin, and C. Févotte, "Professionally-produced music separation guided by covers," in International Society for Music Information Retrieval Conference (ISMIR 2012), Porto, Portugal, 2012, pp. 85-90.

[57] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional Models for Audio Processing," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125-144, Mar, 2015.

[58] Y. X. Wang, and Y. J. Zhang, "Nonnegative Matrix Factorization: A Comprehensive Review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336-1353, Jun, 2013.

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

[60] C. Fevotte, and J. Idier, "Algorithms for Nonnegative Matrix Factorization with the beta-Divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421-2456, Sep, 2011.

[61] A. Cichocki, H. Lee, Y. D. Kim, and S. Choi, "Non-negative matrix factorization with alpha-divergence," *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1433-1440, Jul 1, 2008.

[62] A. Cichocki, and S. Amari, "Families of Alpha- Beta- and Gamma- Divergences: Flexible and Robust Measures of Similarities," *Entropy*, vol. 12, no. 6, pp. 1532-1568, Jun, 2010.

[63] A. Cichocki, R. Zdunek, and S. Amari, "Csiszar's divergences for non-negative matrix factorization: Family of new algorithms," *Independent Component Analysis and Blind Signal Separation: 6th International Conference, ICA 2006, Charleston, SC, USA, March 5-8, 2006. Proceedings*, J. Rosca, D. Erdogmus, J. C. Príncipe *et al.*, eds., pp. 32-39, Berlin, Heidelberg: Springer Berlin Heidelberg, 2006.

[64] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705-1749, Oct, 2005.

[65] A. Cichocki, S. Cruces, and S. Amari, "Generalized Alpha-Beta Divergences and Their Application to Robust Nonnegative Matrix Factorization," *Entropy*, vol. 13, no. 1, pp. 134-170, Jan, 2011.

[66] A. Cichocki, and R. Zdunek, "Multilayer nonnegative matrix factorisation," *Electronics Letters*, vol. 42, no. 16, pp. 947-948, Aug 3, 2006.

[67] C. Andrzej, Z. Rafal, P. Anh Huy, and A. Shun-ichi, "Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation," John Wiley & Sons, 2009, p. p. 500.

[68] P. Smaragdis, "Non-negative matrix factor deconvolution; Extraction of multiple sound sources from monophonic inputs," *Independent Component Analysis and Blind Signal Separation*, vol. 3195, pp. 494-499, 2004.

[69] P. D. O'Grady, and B. A. Pearlmutter, "Convolutive non-negative matrix factorisation with a sparseness constraint," in 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing (MLSP 2006), Maynooth, Ireland, 2006, pp. 427-432.

[70] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, no. 1, pp. 1-12, Jan, 2007.

[71] P. D. O'Grady, and B. A. Pearlmutter, "Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint," *Neurocomputing*, vol. 72, no. 1-3, pp. 88-101, Dec, 2008.

[72] A. Hurmalainen, J. Gemmeke, and T. Virtanen, "Non-Negative Matrix Deconvolution in Noise Robust Speech Recognition," in 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2011, pp. 4588-4591.

[73] Y. Mitsufuji, and A. Roebel, "Sound Source Separation Based on Non-Negative Tensor Factorization Incorporating Spatial Cue as Prior Knowledge," in 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 71-75.

[74] Y. Mitsufuji, and A. Roebel, "On the use of a spatial cue as prior information for stereo sound source separation based on spatially weighted non-negative tensor factorization," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, pp. 1-9, Mar 28, 2014.

[75] A. Shashua, and T. Hazan, "Non-negative tensor factorization with applications to statistics and computer vision," in Proceedings of the 22nd international conference on Machine learning, 2005, pp. 792-799.

[76] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative tensor factorisation for sound source separation," in Irish Signals Syst. Conf. (ISSC), Dublin, Ireland, 2005, pp. 8–12.

[77] R. A. Harshman, *Foundations of the PARAFAC Procedure: Models and Conditions for an "explanatory" Multi-modal Factor Analysis*: University of California at Los Angeles, 1970.

[78] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, 2011, pp. 45-48.

[79] S. Ewert, B. Pardo, M. Muller, and M. D. Plumbley, "Score-Informed Source Separation for Musical Audio Recordings: An overview," *Signal Processing Magazine, IEEE*, vol. 31, no. 3, pp. 116-124, 2014.

[80] A. Ozerov, C. Fevotte, R. Blouet, and J. L. Durrieu, "Multichannel Nonnegative Tensor Factorization with Structured Constraints for User-Guided Audio Source Separation," in 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2011, pp. 257-260.

[81] P. Smaragdis, "User Guided Audio Selection from Complex Sound Mixtures," in Uist 2009: Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, 2009, pp. 89-92.

[82] R. Hennequin, J. J. Burred, S. Maller, and P. Leveau, "Speech-Guided Source Separation Using a Pitch-Adaptive Guide Signal Model," in 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014, pp. 6672 - 6676.

[83] N. Souviraa-Labastie, E. Vincent, and F. Bimbot, "Music separation guided by cover tracks: Designing the joint NMF model," in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on, 2015, pp. 484-488.

[84] N. Souviraa-Labastic, A. Olivero, E. Vincent, and F. Bimbot, "Multi-Channel Audio Source Separation Using Multiple Deformed References," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 23, no. 11, pp. 1775-1787, Nov, 2015.

[85] N. Souviraa-Labastie, A. Olivero, E. Vincent, and F. Bimbot, "Audio source separation using multiple deformed references," in Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European, 2014, pp. 311-315.

[86] L. Le Magoarou, A. Ozerov, and N. Q. K. Duong, "Text-Informed Audio Source Separation. Example-Based Approach Using Non-Negative Matrix Partial Co-Factorization," *Journal of Signal Processing Systems for Signal Image and Video Technology*, vol. 79, no. 2, pp. 117-131, May, 2015.

[87] A. Liutkus, J. Pinel, R. Badeau, L. Girin, and G. Richard, "Informed source separation through spectrogram coding and data embedding," *Signal Processing*, vol. 92, no. 8, pp. 1937-1949, Aug, 2012.

[88] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Informed Source Separation: Source Coding Meets Source Separation," in 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011, pp. 257-260.

[89] A. Liutkus, S. Gorlow, N. Sturmel, S. H. Zhang, L. Girin, R. Badeau, L. Daudet, S. Marchand, and G. Richard, "Informed Audio Source Separation: A Comparative Study," in 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), 2012, pp. 2397-2401.

[90] A. Liutkus, A. Ozerov, R. Badeau, and G. Richard, "Spatial Coding-Based Informed Source Separation," in 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), 2012, pp. 2407-2411.

[91] A. Ozerov, A. Liutkus, R. Badeau, and G. Richard, "Coding-Based Informed Source Separation: Nonnegative Tensor Factorization Approach," *IEEE Transactions on Audio Speech and Language Processing*, vol. 21, no. 8, pp. 1699-1712, Aug, 2013.

[92] G. Casalino, N. Del Buono, and C. Mencar, "Subtractive clustering for seeding non-negative matrix factorizations," *Information Sciences*, vol. 257, pp. 369-387, February, 2014.

[93] B. L. Zhu, W. Li, R. J. Li, and X. Y. Xue, "Multi-Stage Non-Negative Matrix Factorization for Monaural Singing Voice Separation," *IEEE Transactions on Audio Speech and Language Processing*, vol. 21, no. 10, pp. 2096-2107, Oct, 2013.

[94] F. D. Neeser, and J. L. Massey, "Proper Complex Random-Processes with Applications to Information-Theory," *IEEE Transactions on Information Theory*, vol. 39, no. 4, pp. 1293-1302, Jul, 1993.

[95] M. Goodwin, "The STFT, sinusoidal models, and speech modification," *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi and Y. Huang, eds., pp. 229 -258 New York: Springer (2008).

[96] M. Hoffman, D. Blei, and P. Cook, "Bayesian nonparametric matrix factorization for recorded music," in International Conference on Machine Learning (ICML), 2010, pp. 439-446.

[97] B. Jørgensen, *Statistical properties of the generalized inverse Gaussian distribution*, New York: Springer-Verlag, 1982.

[98] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results," in Independent Component Analysis and Signal Separation: 7th International Conference, ICA 2007, 2007, pp. 552-559.

[99] "Signal Separation Evaluation Campaign (SiSEC 2013)," 2013; https://sisec.wiki.irisa.fr/ (date last viewed 01/06/15).

[100] L. Parra, and C. Spence, "Convolutive blind separation of non-stationary sources," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 3, pp. 320-327, 2000.

[101] A. Abdullah, J. Moeller, and S. Venkatasubramanian, "Approximate Bregman near Neighbors in Sublinear Time: Beyond the Triangle Inequality," *International Journal of Computational Geometry & Applications*, vol. 23, no. 4-5, pp. 253-301, Aug-Oct, 2013.

[102] Y. Q. Lin, and D. D. Lee, "Bayesian regularization and nonnegative deconvolution for room impulse response estimation," *IEEE Transactions on Signal Processing*, vol. 54, no. 3, pp. 839-847, Mar, 2006.

[103] Y. Lin, "l1-norm sparse Bayesian learning: theory and applications," Ph.D. dissertation, University of Pennsylvania, 2008.

[104] C. Boutsidis, and E. Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350-1362, Apr, 2008.

[105] K. Adiloglu, H. Kayser, and L. Wang. "A variational inference based source separation approach for the separation of sources in underdetermined recording," (2013); http://www.onn.nii.ac.jp/sisec13/ evaluation_result/UND/submission/ob/Algorithm.pdf (date last viewed 01/06/15).

[106] K. Adiloglu, and E. Vincent, *"Variational Bayesian interference for source separation and robust feature extraction*, Tech. Rep. RT-0428, Inria, Augest 2012.

[107] C. Knapp, and G. C. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320-327, 1976.

[108] P. Comon, and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications*: Academic press, 2010.

[109] J. Yoo, M. Kim, K. Kang, and S. Choi, "Nonnegative Matrix Partial Co-Factorization for Drum Source Separation," in 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2010, pp. 1942-1945.

[110] M. Kim, J. Yoo, K. Kang, and S. Choi, "Nonnegative Matrix Partial Co-Factorization for Spectral and Temporal Drum Source Separation," *IEEE Journal of Selected Topics in Signal Processing,* vol. 5, no. 6, pp. 1192-1204, Oct, 2011.

[111] N. Tengtrairat, B. Gao, W. L. Woo, and S. S. Dlay, "Single-Channel Blind Separation Using Pseudo-Stereo Mixture and Complex 2-D Histogram," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 24, no. 11, pp. 1722-1735, Nov, 2013.

[112] A. Pedone, J. J. Burred, S. Maller, and P. Leveau, "Phoneme-level Text to Audio Synchronization on Speech Signals with Background Music," in 12th Annual Conference of the International Speech Communication Association 2011 (INTERSPEECH 2011), 2011, pp. 433–436.

[113] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio Speech and Language Processing,* vol. 14, no. 4, pp. 1462-1469, Jul, 2006.

[114] N. Q. K. Duong, H. Tachibana, E. Vincent, N. Ono, R. Gribonval, and S. Sagayama, "Multichannel Harmonic and Percussive Component Separation by Joint Modeling of Spatial and Spectral Continuity," in 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2011, pp. 205-208.

[115] F. J. Canadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, J. Carabias-Orti, and P. Cabanas-Molero, "Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints," *EURASIP Journal on Audio Speech and Music Processing*, Jul 11, 2014.

[116] H. Tachibana, N. Ono, H. Kameoka, and S. Sagayama, "Harmonic/Percussive Sound Separation Based on Anisotropic Smoothness of Spectrograms," *IEEE/ACM Transactions on Audio Speech and Language Processing,* vol. 22, no. 12, pp. 2059-2073, Dec, 2014.

[117] H. Tachibana, N. Ono, and S. Sagayama, "Singing Voice Enhancement in Monaural Music Signals Based on Two-stage Harmonic/Percussive Sound Separation on Multiple Resolution Spectrograms," *IEEE/ACM Transactions on Audio Speech and Language Processing,* vol. 22, no. 1, pp. 228-237, Jan, 2014.

[118] D. Fitzgerald, A. Liukus, Z. Rafii, B. Pardo, and L. Daudet, "Harmonic/percussive separation using Kernel Additive Modelling," in Irish Signals & Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communications Technologies (ISSC 2014/CIICT 2014). 25th IET, 2014, pp. 35-40.

[119] J. Driedger, M. Muller, and S. Ewert, "Improving Time-Scale Modification of Music Signals Using Harmonic-Percussive Separation," *IEEE Signal Processing Letters,* vol. 21, no. 1, pp. 105-109, Jan, 2014.

[120] W. Buyens, B. van Dijk, J. Wouters, and M. Moonen, "A Harmonic/Percussive Sound Separation Based Music Pre-Processing Scheme for Cochlear Implant Users," in 21st European Signal Processing Conference (EUSIPCO), Marrakech, 2013, pp. 1 - 5.

[121] H. Tachibana, H. Kameoka, N. Ono, and S. Sagayama, "Comparative Evaluations of Various Harmonic/Percussive Sound Separation Algorithms Based on Anisotropic Continuity of Spectrogram," in 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 465-468.

[122] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A New Sparse Representation for Acoustic Signals," in 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2009, pp. 3437-3440.

[123] S. Arberet, R. Gribonval, and F. Bimbot, "A Robust Method to Count and Locate Audio Sources in a Multichannel Underdetermined Mixture," *IEEE Transactions on Signal Processing,* vol. 58, no. 1, pp. 121-133, Jan, 2010.