BLIND SOURCE SEPARATION USING STATISTICAL NONNEGATIVE MATRIX FACTORIZATION

Phetcharat Parathai

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

Febuary 2015

ABSTRACT

Blind Source Separation (BSS) attempts to automatically extract and track a signal of interest in real world scenarios with other signals present. BSS addresses the problem of recovering the original signals from an observed mixture without relying on training knowledge. This research studied three novel approaches for solving the BSS problem based on the extensions of non-negative matrix factorization model and the sparsity regularization methods.

1) A framework of amalgamating pruning and Bayesian regularized cluster nonnegative tensor factorization with Itakura-Saito divergence for separating sources mixed in a stereo channel format: The sparse regularization term was adaptively tuned using a hierarchical Bayesian approach to yield the desired sparse decomposition. The modified Gaussian prior was formulated to express the correlation between different basis vectors. This algorithm automatically detected the optimal number of latent components of the individual source.

2) Factorization for single-channel BSS which decomposes an information-bearing matrix into complex of factor matrices that represent the spectral dictionary and temporal codes: A variational Bayesian approach was developed for computing the sparsity parameters for optimizing the matrix factorization. This approach combined the advantages of both complex matrix factorization (CMF) and variational L_1-sparse analysis.

3) An imitated-stereo mixture model developed by weighting and time-shifting the original single-channel mixture where source signals can be modelled by the AR processes. The proposed mixing mixture is analogous to a stereo signal created by two microphones with one being real and another virtual. The imitated-stereo mixture employed the nonnegative tensor factorization for separating the observed mixture. The separability analysis of the imitated-stereo mixture was derived using Wiener masking.

All algorithms were tested with real audio signals. Performance of source separation was assessed by measuring the distortion between original source and the estimated one according to the signal-to-distortion (SDR) ratio. The experimental results demonstrate that the proposed uninformed audio separation algorithms have surpassed among the conventional BSS methods; i.e. IS-cNTF, SNMF and CMF methods, with average SDR improvement in the ranges from 2.6dB to 6.4dB per source.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor Doctor Wai Lok Woo and Professor Satnam Dlay for their guidance, encouragement, support, and friendship. I have greatly appreciate with my supervisor Doctor Wai Lok Woo who has meticulously reviewed my work and provided helpful suggestions. He also taught me the basics of Blind Source Separation and provided useful techniques and materials for my research.

I would also like to thank my thesis examination committee members for their time, constructive criticism, and feedback. This thesis is much the better because of them.

I gratefully acknowledge Payap University for offering me the funding. Without the financial support from Payap University, this work would not have taken place.

I wish to convey my sincere thanks to my research colleague Asso. Prof. Dr. Bin Gao for stimulating discussions and for enlightening and guiding my research. I would also wish to thank my best friend Dr. Naruephorn Tengtrairat for her concern and good wishes. I have been blessed by her friendship and cheerfulness throughout my study.

I would like to give a special thanks to my husband Somchai Parathai for his personal support, love and great patience at all times. These have enabled me to complete this work.

I also want to thank my beloved family for their unconditional love and help at every stage of my life. I deeply miss my late father who would have loved being able to share in the joy of this achievement.

Above all, I owe it all to the omnipresent God who answered my prayers for a chance of pursuit a PhD, provided the strength when I was weaken and the wisdom needed to undertake this research and granted me everything I needed to complete this work. As always, a mere expression of thanks does not seem to suffice. Nevertheless, thank you so much, Dear Lord!

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

$\mathbf{Y}$	the output from the multiplication of the compressed matrix $\mathbf{W}$ weighted by the component of $\mathbf{H}$
$\mathbf{W} \underset{=}{\Leftrightarrow} \{w_{fk}\}$	the basis matrix
$\mathbf{H} \underset{=}{\Leftrightarrow} \{h_{kt_s}\}$	the weight matrix
$\mathbf{Q} \underset{=}{\Leftrightarrow} \{q_{ik}\}$	$q_{ik} = \lvert m_{ik} \rvert^2$ where $m_{ik} = a_{id}$
$\Re_+^{F\times T}$	Real positive number of rows F and columns T
$d = 1, \ldots, d_{max}$	the maximum number of sources
K_d	estimating the source which renders the better separation performance than a general $\mathbf{W}$
$i = 1, \ldots, I$	the channel number
$t_s = 1,2, \ldots, T_s$	The time slots
k	the basis component
K	arbitrary set to be smaller than F and T
j	an index of source
$V_{F\times T}$	a non-negative data matrix
Z	the number of basis
f	frequency bin
t	time frame
$\mathbf{Q} \circ \mathbf{H}$	the $I \times K \times T_s$ tensor with elements $q_{ik} h_{kt_s}$
$\mathbf{Q} \circ \mathbf{W}$	the $F \times I \times K$ tensor with elements $q_{ik} w_{fk}$
$\mathbf{W} \circ \mathbf{H}$	the $F \times K \times T_s$ tensor with elements $w_{fk} h_{kt_s}$
$P(f_t \vert z_t)$	a multinomial distribution of frequencies for spectral component z
$P_t(z_t)$.	a multinomial distribution of mixture weights at time frame t
$P_t(f_t)$	the normalized spectrogram at time frame t

V_{ft}	the magnitude spectrogram of sound source at time t and frequency f
q_t	The hidden state at time t
f_t	the observed output at time t
$P(q_{t+1}\|q_t)$	the state transition probability
$P(f_t\|q_t)$	the distribution of the observation given the state
$P(q_1)$.	an initial state distribution
$P_t(\overline{\boldsymbol{f}})$	the probability of all the observations
$P(f\|z,q)$	a multinomial distribution of frequencies
$P_t(z_t\|q_t)$	a multinomial distribution of the weights
$P(v_t\|q_t)$,	the energy distribution
$\overline{\boldsymbol{f}}$	the observations across all time frames
$\overline{\boldsymbol{v}}$	the number of draws over all time frames
$p(\mathbf{W})$	the multivariate rectified Gaussian
$a_{f,t}^{(k)}$	a magnitude spectrum
$\phi_{f,t}^{(k)}$	phase spectra
θ	the optimal parameters
$\epsilon_{f,t}$	a modeling error for each source
$y_i(t)$	the multichannel audio mixtures
$x_{id}(t)$	unknown sources
$d = 1, \dots, d_{max}$	the source number
a_{id}	a mixing coefficient
$\mathcal{N}_c(\cdot)$	the proper complex Gaussian distribution
$d(x\|y)$	a scalar cost function
$\|\mathbf{Y}\|^{\mathbf{2}}$	the power spectrograms
$\doteq$	equality up to constant
$r = 1, \dots, R$	a certain “resemblance”
$c_{rk}(t)$	the components

$\boldsymbol{\Sigma}_W$	covariance matrix
$N_m(\cdot)$	modified multivariate Gaussian
$vec(\cdot)$	the column vectorization
$\boldsymbol{U}(\boldsymbol{w})$	the multivariate Gaussian cumulative distribution function
$E[\cdot]$	the statistical expectation operator
σ_k^2	the variance of the basis vector
$c_{k,j}$	the cross-covariance between $\boldsymbol{w}_k$ and $\boldsymbol{w}_j$
λ_{kt_s}	the regularization parameter
$\Phi(\bullet)$	the multivariate Gaussian cumulative distribution function
$\boldsymbol{G}_-$	the negative part of the derivative of the criterion
$\boldsymbol{G}_+$	its positive part of the derivative of the criterion
$\bar{\lambda}_k$	threshold of the k^{th} row of $\mathbf{H}$ (equivalently k^{th} column of $\mathbf{W}$)
x_{target}	actual source estimate
e_{interf}	interference signal
e_{artif}	artifacts of the separation algorithm
$\mathbb{C}$	Complex number
$f(\theta)$	Auxiliary function
$Q(\underline{\boldsymbol{h}})$	The Gibbs distribution
N_s	the total number of source signals
β	The weighted parameter
δ	the time-delay
M_j	the maximum AR order
$c_{x_j}(z)$	the zth order AR coefficient of the jth source signal at time t
$e_j(t)$	an independent identically distributed (i.i.d.) random signal with variance σ^2 and zero mean
$a_j(t;\delta,\beta) \Leftrightarrow a_j(t)$	the mixing attenuation of the j^{th} source
$r_j(t;\delta,\beta) \Leftrightarrow r_j(t)$	the residue of the j^{th} source
$\forall(\cdot)$	All elements

δ_{max}	the maximum time delay
f_{max}	the maximum frequency present in the sources
f_s	the sampling frequency
$Y_j(f, t_s)$	the STFT of $y_j(t)$
$R_j(f, t_s)$	the STFT of $r_j(t)$
$X_j(f, t_s)$	the STFT of $x_j(t)$
$\mathcal{L}_i$	the selected minimum function of the i^{th} channel
l	a number of group of classified (f, t_s) units
h_k	a vector of activation coefficient
$\mathbf{V} \Leftrightarrow V_i(f, t_s)$	the $I \times F \times T_s$ tensor with coefficients
$\hat{\mathbf{V}} \Leftrightarrow \hat{V}_i(f, t_s)$	the estimated $I \times F \times T_s$ tensor with coefficients
$\mathbf{L} = \{l_{jk}\}$	the $J \times K$ "labelling matrix" with only one nonzero value per column
c	a constant
$\delta(\boldsymbol{w})$	the delta function
$E[\cdot]$	the statistical expectation operator
$c_{k,n}$	the cross-covariance between $\boldsymbol{w}_k$ and $\boldsymbol{w}_n$
$\hat{c}_{k,n}$	the estimated cross-covariance between $\boldsymbol{w}_k$ and $\boldsymbol{w}_n$
μ_{kn}	the correlation between the basis vectors
$\hat{\mu}_{kn}$	The estimated correlation between the basis vectors
$\boldsymbol{P}_k^{-1}$	a Toeplitz matrix corresponding to the kth diagonal sub-matrix of $\boldsymbol{\Omega}_{diag}^{W}$ where $\boldsymbol{\Omega}_{diag}^{W} = \left[\boldsymbol{\Sigma}_{diag}^{(W)}\right]^{-1}$
$\boldsymbol{P}_k$	exponentially decaying
$\boldsymbol{G}$	the $I \times F \times T_s$ tensor with entries $g_{ift_s} = d'_{IS}(V_i(f, t_s) \vert \hat{V}_i(f, t_s))$
$\bullet$	an element-wise product
$\boldsymbol{\Xi}^{\boldsymbol{T}}$	a $K \times K$ matrix
α	the smoothing parameter
$\epsilon = 10^{-9}$	a small number to prevent division by zero

$\hat{C}_{ik}(f, t_s)$	the reconstructed source of the component k in the channel i
ρ_k	the first-order correlation of $\boldsymbol{w}_k$
$\hat{x}_{ij}(t)$	The estimated sources

ABBREVIATIONS/ACRONYMS

2D	2-Dimensional
AIC	Akaike Information Criterion
APBNTF	Automatic Pruning Bayesian Non-negative Tensor Factorization
AR	Autoregressive
ARD	Automatic Relevance Determination
ASR	Automatic/Automated Speech Recognition
BASS	Blind Audio Source Separation
BIC	Bayesian Information Criterion
BSS	Blind Source Separation
CASA	Computational Auditory Scene Analysis
CMF	Complex Matrix Factorization
cNTF	Cluster Non-negative Tensor Factorization
DUET	Degenerate Estimation Technique
EM	Expectation-Maximization
EEG	Electroencephalogram
HMM	Hidden Markov Model
ICA	Independent Component Analysis
IS	Itakura-Saito

ISM	Imitated Stereo Mixture
ISM-NTF	Imitated Stereo Mixture non-negative Tensor Factorization
ISM-RNTF	Imitated Stereo Mixture Regularized non-negative Tensor Factorization
KL	Kullback-Leibler
LS	Least Squares
MAP	Maximum A Posteriori
MEG	Magenetoencephalogram
ML	Maximum Likelihood
MSE	Mean - Square Error
MU	Multiplicative Update
N-HMM	Non-negative Hidden Markov Model
N-FHMM	Non-negative Factorial Hidden Markov Model
NMF	Non-negative Matrix Factorization
NTF	Non-negative Tensor Factorization
PARAFAC	Parallel Factor Analysis
RWC	Real World Computing
SAR	Signal-to- Artifacts Ratio
SASS	Stereo Audio Source Separation
SC	Sparse Coding
SCBSS	Single Channel Blind Source Separation

SCICA	Single - Channel Independent Component Analysis
SCASS:	Single Channel Audio Source Separation
SDR	Signal-to-Distortion Ratio
SiSEC	Signal Separation Evaluation Campaign
SIR	Signal-to-Interference Ratio
SNMF	Sparse Non-negative Matrix Factorization
STFT	Short Time Fourier Transform
TF	Time Frequency
VB	Variational Bayesian
vL_1-SCMF	Variational L_1-Sparse Complex Matrix Factorization
WDO	Windowed - Disjoint Orthogonality

LIST OF PUBLICATIONS

- P. Parathai, W.L. Woo, and S.S. Dlay, "Monaural Blind Separation By Exploiting Source Temporal Correlation And Nonnegative Tensor Factorization Signal Processing", *accepted for Signal Processing*

- P. Parathai, W.L. Woo, and S.S. Dlay, "Single-Channel Blind Separation using L_1-Sparse Complex Nonnegative Matrix Factorization for Acoustic Signals*" accepted for Journal of the Acoustical Society of America*

- P. Parathai W.L. Woo, and S.S. Dlay, "Single-Channel Source Separation using Temporal Correlation with Regularized Nonnegative Tensor Factorization", *IEEE Transactions on Neural Networks and Learning System (under review)*

- P. Parathai, W.L. Woo, and S.S. Dlay, "MAP-based Regularized Nonnegative Tensor Factorization for Multichannel Source Separation," *IEEE Transactions on Audio, Speech and Language Processing (under review)*

- P. Parathai, W.L. Woo, and S.S. Dlay, "Single Channel Source Separation using Variational L_p-Sparse Complex Matrix Factorization", *IEEE Transactions on Neural Networks and Learning System (under review)*

CHAPTER 1

INTRODUCTION TO THE THESIS

1.1 Background of Blind Source Separation

Humans are very good at focusing their attention on the speech of a single speaker, even in the presence of other speakers and background noise. One classical problem of blind source separation (BSS) [1] is the so-called "cocktail party problem" is a psychoacoustic phenomenon that refers to the remarkable human ability to selectively attend to and recognize one source of auditory input in a noisy environment, where the hearing interference is produced by competing speech sounds or a variety of noises that are often assumed to be independent of each other. Although the human brain and auditory system can handle this everyday problem with ease it is very hard to solve with computational algorithms. There are attempt to imitate the human performance with a machine by simplifying the complex perceptual task as a learning problem for tractable computational solution.

Speaker separation has conventionally been treated as a problem of Blind Source Separation. BSS [2] is an approach to unveil independent source signals from their mixtures without any prior information on the sources or the parameters of the mixed signal. Many methods for BSS have been proposed to reconstruct source signals for example computational auditory scene analysis (CASA) relies on the development of a computational model of the auditory scene to automatically extract and track a sound signal of interest in a cocktail party environment, independent component analysis (ICA). ICA is a data driven method that makes good use of multiple inputs and relaxes the strong characteristic frequency structure assumptions.

The ICA algorithms find the independent components by maximizing the statistical independence of the estimated components. However, ICA algorithms perform best when the number of observed signals is greater than or equal to the number of sources [3]. BSS is broadly applied in different disciplines such as in order to deploy automatic speech recognition (ASR) effectively in real world scenarios it is necessary to handle hostile environments with multiple speech and noise sources. Current state-of-the-art ASR systems are trained on clean single talker speech and therefore inevitably have serious difficulties when confronted with noisy multi-talker environments [4].

Non-negative Matrix Factorization (NMF) has been ubiquitously used in many applications with great success for recovering underlying source signals given by a single sensor. The NMF method was invented by two scientists Lee and Seung [5] for factorizing a matrix into a product of two non-negative matrices. NMF has been applied extensively with considerable success to various problem domains, such as monaural sound source separation [6], polyphonic music transcription [7], face detection [8] and other signal-processing applications. NMF can project all signals that have the homogeneous spectral shape on a single basis, allowing one to represent a variety of phenomena efficiently using a very compact set of spectrum bases and there has been a plenty of work in modeling audio using non-negative matrix factorization and its probabilistic counterparts as they yield rich models that are very useful for source separation and automatic music transcription. Learning time-varying spectra with standard NMF would require using a large number of basis vectors, and some post-processing to group the basis vectors into a single spectral vector. NMF and its extension are the prominent methods for linear combination that algorithms have been applied to solve the practical problems of BSS in many applications. Existing approaches have been successful in different conditions but none of them are yet satisfactory for speech application.

1.1.1 Blind Source Separation Problem using Nonnegative Matrix Factorization

Nonnegative matrix factorization (NMF) is an unsupervised data decomposition technique with considerable research success in the fields of blind source separation (BSS) [6, 9], data classification [10], data mining [11], pattern recognition [12], object detection [13], and dimensionality reduction [14]. Conventional NMF starts with a data matrix $\mathbf{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_{\boldsymbol{T}}] \in \Re_+^{F\times T}$with $\mathbf{Y}_{f,t} > 0$. NMF factorizes this matrix into a product of two nonnegative matrices i.e. $\mathbf{W} \in \Re_+^{F\times K}$ and $\mathbf{H} \in \Re_+^{K\times T}$ such that

$$\mathbf{Y} \approx \mathbf{WH} \tag{1.1}$$

where F and T denote the total number of rows and columns in matrix $\mathbf{Y}$, respectively. Generally, K is arbitrary set to be smaller than F and T. Thus, matrix $\mathbf{Y}$ is the output from the multiplication of the compressed matrix $\mathbf{W}$ weighted by the component of $\mathbf{H}$.

As such, the K of $\mathbf{W}$ is important for approximating the data in $\mathbf{Y}$. Therefore, the matrix $\mathbf{W}$ is considered as a set of basis vectors. NMF was initially developed using the multiplicative update (MU) algorithm to solve its parametric optimization based on the least square (LS) distance and Kullback-Liebler (KL) divergence as a cost function [15-16]. Other families of cost functions have been proposed, such as the Beta divergence [17], Csiszar's divergences [18], and Itakura-Saito (IS) divergence [19]. Additionally, iterative gradient update was presented [20] and a sparseness constraint can be included into the cost function by regularization using the L_1-norm based on minimizing penalized least squares [21] and using different sparsity constraints for $\mathbf{W}$ and $\mathbf{H}$ [22]. However, the sparsity parameter is manually determined the above proposed methods. Approximate sparsity is an important factor which represents significant information in $\mathbf{Y}$. Many sparse solutions have been proposed in the last decade. Nonetheless, the optimal sparse solution remains an open issue.

1.1.2 Applications of BSS

BSS has been a hot topic in signal processing during last few decades. Applications of BSS have been reported in many fields. Such is the case when a sensor array records acoustic or electronic communications signals emanating from a number of different sources. BSS is an important technique used in applications such as a front-end for robust automatic speech recognition (ASR) where many proposed methods are based on independent component analysis (ICA). However, the performances of these methods degrade seriously particularly under extreme reverberant conditions. The experimental results reveal that the separation performance of the ICA method proposed in [23, 24] using subarray processing is improved as the number of microphones is increased. An automatic music transcription task involves extracting information from individual sources. This task becomes more challenging when a given musical recording has numerous parts by numerous instruments. The one solution based on NMF approach [25] is to analyze frequency spectra of music signals and perform instrument separation or note transcription. If each of the instruments in the mixture can be modelled, they can be transcripted individually. This application is useful for a musician attempting to practice a particular instrument of interest directly from a multiinstrument recording.

Another example of the blind source separation application can be seen in medical applications where electrodes on the scalp record a mixture of signals generated by various sources of activity within the brain and are combined with sources of interference, such as signals generated by muscle activity. The BSS problem arises e.g. in analysis of the electric potentials on the scalp surface (electroencephalogram (EEG)), recording the magnetic fields near the surface of the head (magenetoencephalogram (MEG)). The analysis of the data is complex because it is possible that multiple neural generators are simultaneously active, and the potentials and magnetic fields from these sources overlap the footprint of the detectors [26]. In these cases, the BSS solution has been used to un-mix the data into signals representing the behaviors of the original individual generators. Recent research has also shown the feasibility of BSS techniques for various medical applications in [27, 28].

Blind source separation (BSS) has been attempted in robot audition, using a microphone array. In these cases, the use of multiple sensors array improves significantly the performance of the source separation algorithm. Microphones that should be used for robot audition given a specified array geometry i.e. the microphones are located around the robot's head [29].

1.1.3 Blind Audio Source Separation

In this thesis, the special case of audio mixtures problem termed as blind channel source separation is focused. For blind audio source separation (BASS) methods, this denotes the separation of completely unknown sources without using additional training information. Most audio signals are mixtures of several audio sources (speech and music). This method consists in recovering one or several source signals from a given mixture signal. Figure 1.1 shows a general framework for BASS methods.

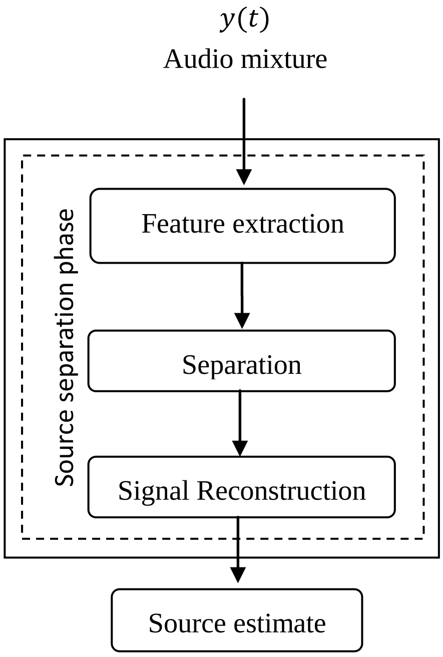


Figure 1.1: Overview of BASS approach.

In Figure 1.1, the input to the separation system is only the audio mixture. The mixture is transformed into a suitable representation that directly through a signal reconstruction method to compute the estimates of the separated sources.

1.2 Objectives of Thesis

This thesis aims to investigate blind source separation methods in fundamental theories, assumptions, applications and limitations terms and further develop new algorithms of BSS for audio mixture. With this goal, the objectives are briefly explained in the following:

i). To present a unified perspective of the widely used state-of-the-art nonnegative matrix factorization (NMF) approaches. The theoretical aspects of BSS are presented to provide sufficient background knowledge relevant to the thesis.

ii). To develop rigorous new theories, mathematical derivations, and algorithms to recover the information about the original sources.

iii). To carry out analysis and comparisons of the proposed algorithms performance with state-of-the-art BSS methods by using objective as well as perceptual evaluation of audio quality using metrics such as the Signal-to-Distortion ratio (SDR).

1.3 Thesis Outline

Three novel methods based on NMF constitute the main contribution of the thesis. The thesis outline is as follows:

Chapter 2 provides an overview of recent source separation methods based on NMF approach is given. This chapter begins with an introduction to NMF methods and includes hidden Markov model (HMM) and complex components. The extention of NMF which is known as nonnegative tensor factorization (NTF) method is also reviewed.

Chapter 3 introduces a novel approach to Bayesian regularized cluster nonnegative tensor factorization under a parallel factor analysis (PARAFAC) structure. The basis for the proposed tensor factorization is developed under the framework of maximum *a posteriori* probability which is further adaptively fine-tuned using the hierarchical Bayesian approach. This chapter will show that this method enables: 1) a generalized criterion for variable sparseness to be imposed onto each element of the temporal code; and 2) modified multivariate rectified Gaussian prior information to be explicitly incorporated into the basis features. This chapter also addresses the important issue of efficiency by using a framework of model selection for pruning unnecessary components and a novel Bayesian regularized cluster nonnegative tensor factorization under a PARAFAC structure with Itakura-Saito divergence. The proposed method is demonstrated further via experiments on underdetermined linear instantaneous stereo mixtures.

Chapter 4 covers a novel single-channel audio source separation (SCASS) which has been developed to extract better quality of audio separated signals. This chapter will introduce this approach that exploits the variational L_1-sparse complex matrix factorization (vL_1-SCMF) to offer the advantages of the CMF and a variational L_1-sparse approach, simultaneously and decomposes an information-bearing matrix into complex factor matrices that represent the spectral dictionary and temporal codes. The derivation of a variational Bayesian approach to compute the sparsity parameters for optimizing the matrix factorization will be presented. The method is then demonstrated on separating audio mixtures recorded from a single channel and its performance is compared with other existing sparse factorization methods. The performance of the developed algorithms will be measured using real-time audio signals in terms of the signal-to-distortion ratio.

Chapter 5 presents, a novel approach to solve the single-channel blind source separation (SCBSS) problem in which a new imitated-stereo mixture is formulated by weighting and time-shifting the original single-channel mixture. This chapter will show how paves the way to employ nonnegative tensor factorization applies to the monaural-channel problem by creating an artificial mixing system whose parameters can be estimated via a proposed nonnegative tensor factorization. The proposed tensor factorization is further developed under the framework of maximum *a posteriori* probability and is adaptively fine-tuned under a PARAFAC structure with Itakura-Saito divergence. In addition, the separability analysis of the proposed imitated-stereo mixture is derived. Experimental testing on real-audio sources has been conducted to verify the capability of the proposed method.

Chapter 6 provides the closing remarks as well as future avenues for research.

1.4 Contribution

This thesis contributes three novel solutions for the BSS problem which can be summarised as follows:

i). A unified approach for the existing BSS methods based on nonnegative matrix factorization.

ii). A novel framework for multichannel blind source separation is proposed:

- Unlike the conventional NTF approach, the proposed framework assigns a probability distribution to each element of $\mathbf{H} = \{h_{ij}\}$ and a sparsity parameter associated with each probability distribution. This sets up a platform to enable the sparsity parameter to be individually optimized for each element code.
- It automatically detects the optimal number of components K of the individual source (i.e. K_d, $d = 1, \dots, d_{max}$ where d_{max} is the maximum number of sources). It designates a prior distribution on $\mathbf{H}$ and determines the desirable K_d in $\mathbf{W}$ by pruning the irrelevant K_d from $\mathbf{W}$. The term $\mathbf{W}$ with the proper K_d is used for estimating the source which renders the better separation performance than $\mathbf{W}$ without the proper K_d.
- It incorporates prior information of the basis vectors using the modified multivariate rectified Gaussian. This benefits the overall algorithm in terms of better estimation accuracy and more meaningful feature extraction that pertain to the data. Since each pattern in the observed mixture has its own features,

designing the appropriate basis to match these features is imperative.

iii). A novel algorithm to solve SCBSS based on CMF is proposed.

- Unlike the CMF, the proposed model is assigned a probability distribution to each element of $\mathbf{H} = \{h_{ij}\}$ and a sparsity parameter associated with each probability distribution. This sets up a platform to enable the sparsity parameter to be individually optimized for each element code.
- The proposed algorithm enables the phase of constituent signals to be estimated more accurately for feature extraction. Since each pattern in $\mathbf{Y}$ has its own features, designing the appropriate phase to match these features is imperative. Incorporating the phase parameter will give the better recovered sources than without using phase information.
- Each sparsity parameter in our model is learned and adapted as part of the matrix factorization.

iv). A novel method for single-channel blind source separation (SCBSS) based NTF is proposed.

- The novel imitated-stereo mixture lights the way to reformulate NTF approaches into the single mixture. This relaxes the under-determined ill-conditions associated with monaural source separation.
- The proposed solution separates sources from a single channel without relying on training information about the original sources.

CHAPTER 2

OVERVIEW OF BLIND SOURCE SEPARATION

The following sections will provide an overview of existing algorithms of the Single-channel Independent Component Analysis (SCICA), nonnegative matrix factorization (NMF) composed of single channel NMF [30-32] and multi-channel NMF which is known as Nonnegative Tensor Factorization (NTF). NTF is a multidimensional model with nonnegativity constraints. Generally, the term 'tensor' denotes multi-way arrays and the order of a tensor is the number of modes, also known as ways or dimensions. The details of these approaches are discussed in Sections 2.1, 2.2 and 2.3.

2.1 Single-channel Independent Component Analysis

The ICA-based methods [33 - 35] show very successful, and perhaps, the most widely used~~,~~ for performing blind source separation in the general case. Single-channel independent component analysis (SCICA) is a BSS technique that extracts statistically independent sources from a single-channel recording. SCICA is an adaptation of the standard ICA algorithm to one observed sensor, which has already been proposed in [24, 36, 37]. The mixtures can then be separated by only employing the standard ICA. The observation model is expressed as:

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{2.1}$$

where the $m \times n$ matrix $\mathbf{A}$ is an unknown constant matrix called the mixing coefficient matrix. The task is to identify the mixing coefficient matrix $\mathbf{A}$, and separate the source signals $\mathbf{x}$ while only knowing a sample of observed vectors $\mathbf{y}$. The term $\mathbf{x}$ represents independent signals. Generally, the original signals can be separated from $\mathbf{y}$ as shown in the following:

$$\mathbf{x} = \mathbf{W}\mathbf{y} \qquad \text{where } \mathbf{W} = \mathbf{A}^{-\mathbf{1}} \tag{2.2}$$

For SCICA, the observed mixture $\mathbf{y}$ is broken up into a sequence of contiguous blocks k with length N. These are treated as a sequence of vector mixtures:

$$\boldsymbol{y}(k) = [y(k\tau), \dots, y(k\tau + N - 1)]^T \tag{2.3}$$

where $k = 1,2, \dots, K$ is the block index τ is a time delay, and $(K\tau + N - 1)$ is the length of the original signal. The matrix $\mathbf{Y}$ is then formed as a set of mixtures $\boldsymbol{y}(k)$ as the following:

$$\mathbf{Y} = [y(1), \dots, y(K)]^T \tag{2.4}$$

The FastICA algorithm [38-40] can then be applied to $\mathbf{Y}$ to compute the mixing and unmixing matrix $\mathbf{A}$ and $\mathbf{W}$. For a perfect reconstruction decomposition, the separation process must be performed in the mixture domain where each signal is discovered via $\mathbf{A}$ and $\mathbf{W}$ as:

$$\boldsymbol{y}_{\boldsymbol{x}}^{(j)} = \mathbf{A}_{(:,j)}\mathbf{W}_{(j,:)}\boldsymbol{y} \tag{2.5}$$

where $\boldsymbol{y}_{\boldsymbol{x}}^{(j)}$ is the originalj^{th} signal in the mixture domain i.e. $\mathbf{y} = \sum_{\boldsymbol{j}} \boldsymbol{y}_{\boldsymbol{x}}^{(j)}$. The j^{th} signal is consecutively estimated and subtracted from $\boldsymbol{y}$ one at a time where the subtracted $\mathbf{y}$ is redefined as a new obtained mixture $\mathbf{y}$. The algorithm is repeated to extract the second signal and so on which is presented in Table 2.1.

Table 2.1: Algorithm of SCICA

1. Break up an observed mixture $\mathbf{y}$ into a sequence of adjacent blocks $\mathbf{Y}$

2. Apply the FastICA algorithm to this matrix~~,~~ to compute the unmixing matrix $\mathbf{W}$

3. Extract the particular signal $\boldsymbol{x}(i)$ of interest by filtering the mixture $\boldsymbol{y}$ with the corresponding row of the matrix $\mathbf{W}$

4. Recover the original signal $\boldsymbol{y}_{\boldsymbol{x}}^{(j)}$ by multiply the extracted signal $\boldsymbol{x}(i)$ with the i^{th} column of the matrix $\mathbf{A}$

5. Subtract the recovered signal $\boldsymbol{y}_{\boldsymbol{x}}^{(j)}$ from the mixture $\mathbf{y}$, redefine the substracted mixture as $\mathbf{y}$, and repeat the steps from 1 – 4 to further extract the remaining signals.

However, SCICA has two major drawbacks: first, the algorithm assumes stationary sources, and second, the sources are assumed to be disjoint in the frequency domain.

2.2 Nonnegative Matrix Factorization Approaches

In recent years, there is growing interest in the field of BSS using factorization-based approaches [41–45]. Non-Negative Matrix factorization (NMF) is a data-adaptive linear representation method for 2-D matrices as presented in Figure 2.1. NMF decomposes a non-negative data matrix $\mathbf{V}$ into the product of two non-negative matrix factors $\mathbf{W}$ and $\mathbf{H}$:

$$\mathbf{V}_{F\times T} \approx \mathbf{W}_{F\times Z}\mathbf{H}_{Z\times T} \tag{2.6}$$

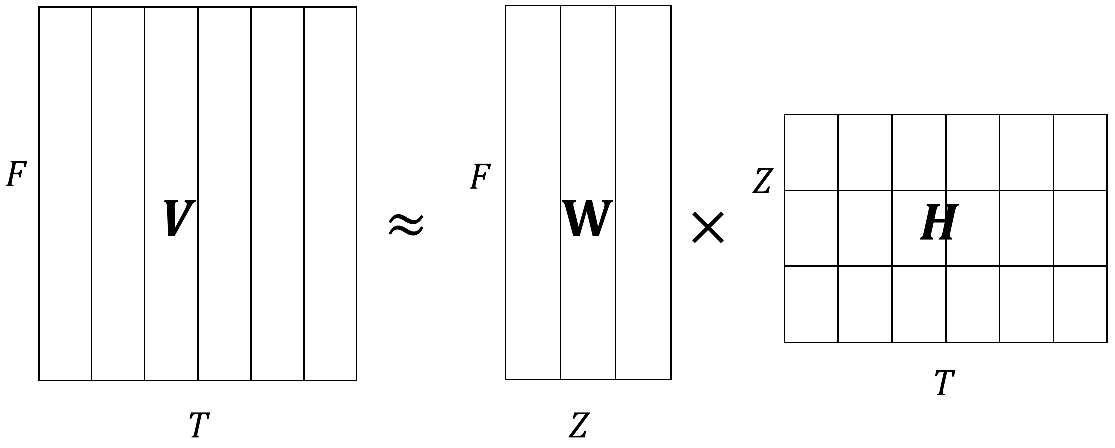


Figure 2.1: Diagrame of NMF

Where $\mathbf{W}$ plays the role of the basic matrix, while $\mathbf{H}$ represents the weight matrix. The Z parameter indicates the number of basis used to represent the original matrix. The basis can be considered as spectral patterns which are frequently observed. NMF is an additive model which does not allow subtraction. To find such a pair of $\mathbf{W}$ and $\mathbf{H}$ which minimizes the error of the approximation in (2.6), two alternative cost functions are defined: Euclidean distance, C, and Divergence, D:

$$C = \|\mathbf{V} - \mathbf{WH}\| = \sum_{f,t}\left(\mathbf{V}_{f,t} - (\mathbf{WH})_{f,t}\right)^2 \tag{2.7}$$

$$D = \sum_{f,t}\left(\mathbf{V}_{f,t}\, log\frac{\mathbf{V}_{f,t}}{(\mathbf{WH})_{f,t}} - \mathbf{V}_{f,t} + (\mathbf{WH})_{f,t}\right) \tag{2.8}$$

NMF aims to calculate the factor of the matrix $\mathbf{V}$ in the form of the product of matrix $\mathbf{W}$. Z is any positive integer which is less than F or T [46] chosen for finding components. For the problem of sound separation at any position, (f, t) of the matrix $\mathbf{V}$ is the amplitude of each frequency f at different time t when $1 \leq f \leq F$ and $1 \leq t \leq T$ presented in spectrogram as shown in (2.6) and can be explained by using linear algebra. Figure 2.1 shows the matrix $\mathbf{V}$ as column vector of length F with T vectors. The T columns of $\mathbf{V}$ consists of F-dimensional data vectors. The Z columns of $\mathbf{W}$ contains basis vectors of dimension F. Each T-dimensional column vector of the approximation $\mathbf{V}$ as (2.6) is a linear combination of all basis vectors, whereby the coefficients are the entries of the corresponding Z-dimensional column vector of $\mathbf{H}$.

After estimating the matrix $\mathbf{W}$ and $\mathbf{H}$ in a source separation, the next step is to select only the basis vector of the sound source from both of the matrices such as column at Z and row at Z of the matrix $\mathbf{W}$ and $\mathbf{H}$ respectively. Next, multiplication of $\mathbf{W}$ and $\mathbf{H}$ yields a new matrix size $F\times T$ to be used in calculating the spectrum of the target sound with various methods to follow.

Thus, NMF algorithms aim to find a local minimum of the divergences. Commonly used cost functions for NMF are the generalized Kullback-Leibler (KL) divergence and Least Square (LS) distance which have been introduced in [5], respectively, as:

$$C_{KL}\left(|\mathbf{Y}|^{\mathbf{2}}\middle\||\widehat{\mathbf{Y}}|^{\mathbf{2}}\right)=\sum_{f,t}\left(\left|\mathbf{Y}_{f,t}\right|^{2}log\frac{\left|\mathbf{Y}_{f,t}\right|^{\mathbf{2}}}{\left|\widehat{\mathbf{Y}}_{f,t}\right|^{\mathbf{2}}}-\left|\mathbf{Y}_{f,t}\right|^{\mathbf{2}}+\left|\widehat{\mathbf{Y}}_{f,t}\right|^{\mathbf{2}}\right)$$

$$C_{LS}\left(|\mathbf{Y}|^{\mathbf{2}}\middle\||\widehat{\mathbf{Y}}|^{\mathbf{2}}\right)=\frac{1}{2}\Sigma_{f,t}\left(\left|\mathbf{Y}_{f,t}\right|^{\mathbf{2}}-\left|\widehat{\mathbf{Y}}_{f,t}\right|^{2}\right)^{2} \qquad (2.9)$$

where $\mathbf{V}=|\mathbf{Y}|^{\mathbf{2}}$ is the power TF representation of mixture $y(t)$ which can be further factorized as the product of two nonnegative matrices $\mathbf{W}$ and $\mathbf{H}$ and $\left|\widehat{\mathbf{Y}}\right|^{\mathbf{2}}=\mathbf{WH}$. From the above equations, C_{KL} is equivalent to an-assumed Poisson noise model for the data and C_{LS} is equivalent to the maximum likelihood estimation of $\mathbf{W}$ and $\mathbf{H}$ in additive independent and identically distributed (i.i.d.) Gaussian noise. The widely used estimation algorithms of Lee and Seung [5] minimize the chosen cost function by initializing the entries of $\mathbf{W}$ and $\mathbf{H}$ with random positive values, and then update those iteratively using multiplicative rules. Each update decreases the value of the cost function until the algorithm converges. The update rule for KL divergence is given by:

$$\mathbf{W} \leftarrow \mathbf{W} \bullet \frac{(|\mathbf{Y}|^2./\mathbf{WH})\mathbf{H}^\mathbf{T}}{\mathbf{1H}^\mathbf{T}} \tag{2.10}$$

and

$$\mathbf{H} \leftarrow \mathbf{H} \bullet \frac{(|\mathbf{Y}|^2./\mathbf{WH})\mathbf{W}^\mathbf{T}}{\mathbf{1W}^\mathbf{T}} \tag{2.11}$$

where '$\bullet$' and '$./$' denote the element-wise product multiplication and division, respectively. '$\mathbf{1}$' is an all-one F by T matrix. The update rule for LS distance is given by:

$$\mathbf{W} \leftarrow \mathbf{W} \bullet \frac{|\mathbf{Y}|^2\mathbf{H}^\mathbf{T}}{\mathbf{WHH}^\mathbf{T}} \tag{2.12}$$

and

$$\mathbf{H} \leftarrow \mathbf{H} \bullet \frac{\mathbf{W}^\mathbf{T}|\mathbf{Y}|^2}{\mathbf{W}^\mathbf{T}\mathbf{WH}} \tag{2.13}$$

On the one hand, the following are advantages of NMF: The mixing model is defined in the magnitude spectrum domain. Because of the phase-invariant nature of magnitude spectra, NMF is able to project all signals that have the same spectral shape onto a single basis. This allows us to represent a variety of acoustic phenomena efficiently using a very compact set of spectrum bases. On the other hand, NMF cannot estimate the phase spectra of underlying constituent signals, which certainly limits its range of applications.

2.2.1 NMF using Hidden Markov Model

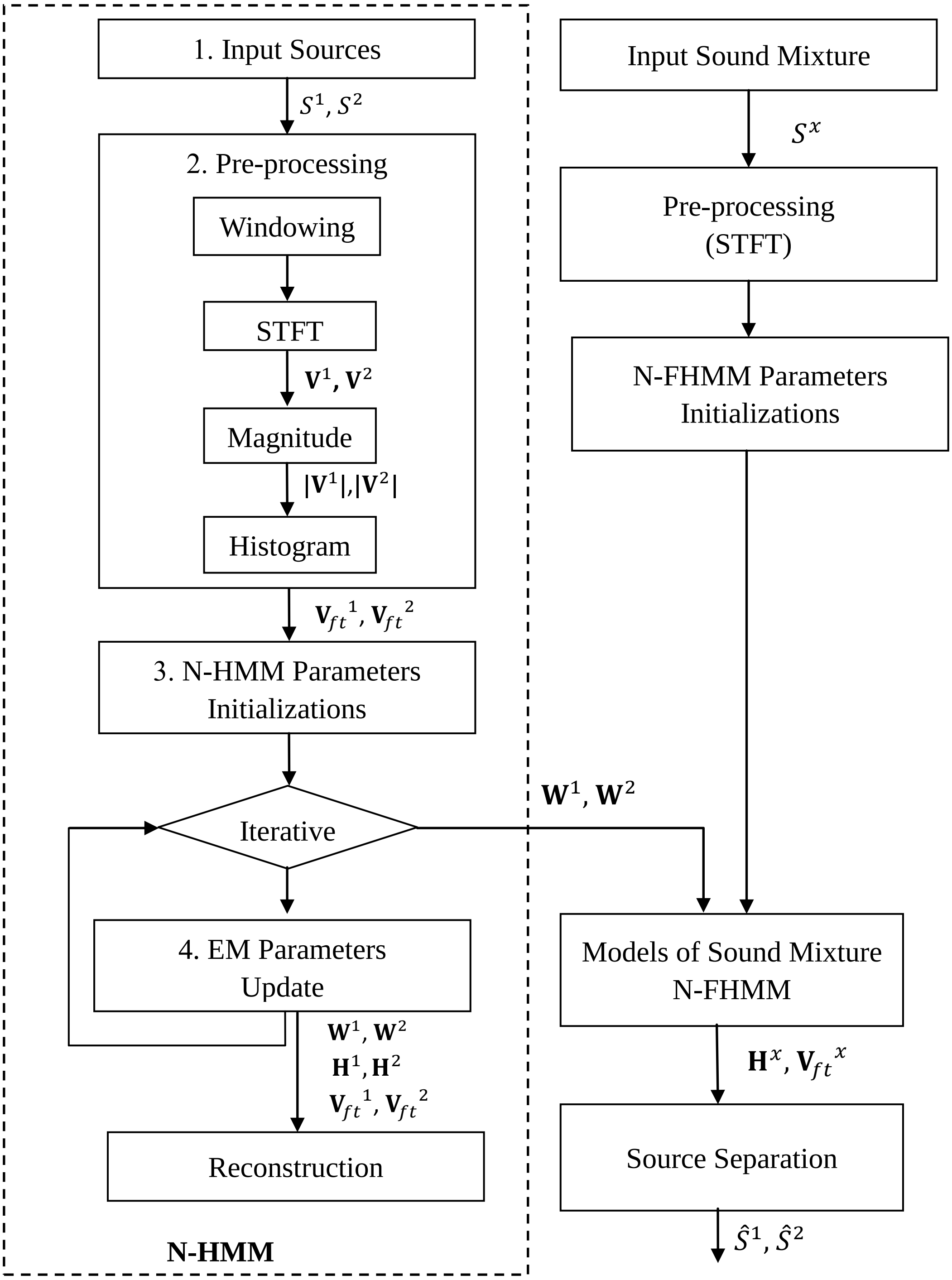

1. Input Sources
S^1, S^2
2. Pre-processing
Windowing
STFT
$\mathbf{V}^1, \mathbf{V}^2$
Magnitude
$|\mathbf{V}^1|, |\mathbf{V}^2|$
Histogram
$\mathbf{V}_{ft}{}^1, \mathbf{V}_{ft}{}^2$
3. N-HMM Parameters Initializations
Iterative
4. EM Parameters Update
$\mathbf{W}^1, \mathbf{W}^2$
$\mathbf{H}^1, \mathbf{H}^2$
$\mathbf{V}_{ft}{}^1, \mathbf{V}_{ft}{}^2$
Reconstruction
N-HMM
Input Sound Mixture
S^x
Pre-processing (STFT)
N-FHMM Parameters Initializations
$\mathbf{W}^1, \mathbf{W}^2$
Models of Sound Mixture N-FHMM
$\mathbf{H}^x, \mathbf{V}_{ft}{}^x$
Source Separation
$\hat{S}^1, \hat{S}^2$

Figure 2.2: Overview of N-HMM method.

This section presents a model of single source sounds, Non-negative hidden Markov model (N-HMM) [47], which combines the rich spectral representative power of NMF and the temporal structure modeling of traditional HMMs [48]. The overview of the N-HMM method is presented in Figure 2.2. N-HMM is consistent with the non-stationary nature of audio as a multiple learning small dictionaries of spectral components to describe different features of the sound source. Furthermore, it can be used to model the temporal dynamics of the sound source between dictionaries by learning a Markov chain.

Table 2.2 Average SDR results for 3 types of mixtures in difference model N-HMM & N-FHMM and NMF.

Mixture	Method	SDR (dB)
Music and Music	N-HMM & N-FHMM	7.42
	NMF	5.17
Music and Speech	N-HMM & N-FHMM	5.73
	NMF	3.55
Male Speech and Female Speech	N-HMM & N-FHMM	2.56
	NMF	2.06

The overall comparison results between the N-HMM & N-FHMM and NMF methods have been summarized in Table 2.2. According to the table, the N-HMM& N-FHMM tends to yield better result than NMF method. The average performance improvement of the N-HMM& N-FHMM method against the NMF method: 1) for the music and music mixtures, the improvement SDR per source is 2.25dB. 2) For the music and speech mixtures, the improvement per source in term of SDR is 2.18dB. 3) For the male speech and female speech mixtures, the improvement per source in term of SDR is 0.5dB.

The results demonstrate that this approach is applicable to model the individual sources by learning several small dictionaries using NMF method and a Markov chain. Sources have been modeled via HMM in time-frequency domain obtain prior information of the original signals. Good separation performance with high SDR has been obtained using the method. However, the mixing model implicitly assumes the added-magnitude spectra which only approximately hold, although attempts are made to mitigate the non-additive problem with respect to NMF. NMF cannot estimate the phase spectra of underlying constituent signals, which certainly limits its range of applications. Moreover, the phase coherence between frequency components can be easily destroyed as a result of many factors. It is difficult to capture high-level structural elements from observations through the use of complex-spectrum bases.

2.2.2 Bayesian Non-negative Matrix Factorization

Bayesian NMF [49] assumes a Gaussian likelihood, independent exponential priors on $\mathbf{W}$ and $\mathbf{H}$ with scales $\alpha_{f,k}$, $\beta_{k,t}$ and derive an efficient Gibbs sampler to approximate the posterior density of the NMF factors. It is assumed that the residuals are i.i.d. zero mean normal with variance σ^2, which gives rise to the likelihood

$$P(\mathbf{Y}|\theta) = \prod_{f,t} \mathcal{N}(\boldsymbol{Y}_{f,t}; (\mathbf{WH})_{f,t}, \sigma^2) \tag{2.14}$$

$$P(\mathbf{W}) = \prod_{f,k} \alpha_{f,k} \exp(-\alpha_{f,k}\mathbf{W}_{f,k}) \tag{2.15}$$

and

$$P(\mathbf{H}) = \prod_{k,t} \beta_{k,t} \exp(-\beta_{k,t}\mathbf{H}_{k,t}) \tag{2.16}$$

where $\theta = \{\mathbf{W}, \mathbf{H}, \sigma^2\}$ denotes all parameters in the model. The prior for the noise variance is chosen as an inverse gamma density with shape k and scale θ,

$$P(\sigma^2) = \prod_{k,t} \frac{\theta^k}{\Gamma(k)} (\sigma^2)^{-\mathrm{k}-1} \exp\left(-\frac{\theta}{\sigma^2}\right) \tag{2.17}$$

According to Bayes' rule, the posterior distribution of all parameters in the model is given by

$$P(\mathbf{W}, \mathbf{H}, \sigma^2 | \mathbf{Y}) \propto P(\mathbf{Y} | \mathbf{W}, \mathbf{H}, \sigma^2)\, P(\mathbf{W}) P(\mathbf{H})\, P(\sigma^2) \tag{2.18}$$

The joint posterior density is approximated in a sampling scheme by iteratively sampling one parameter while keeping all others fixed. Expressions are derived for the conditional posterior densities of the model parameters

$$P\left(\mathbf{W}_{f,k} \middle| \mathbf{Y}, \mathbf{W}_{-f,k}, \mathbf{H}, \sigma^2\right) \tag{2.19}$$

$$P\left(\mathbf{H}_{k,t} \middle| \mathbf{Y}, \mathbf{W}, \mathbf{H}_{-k,t}, \sigma^2\right) \tag{2.20}$$

and

$$P(\sigma^2 | \mathbf{Y}, \mathbf{W}, \mathbf{H},) \tag{2.21}$$

where the index –*f*,*k* depicts all entries of a matrix except entry *f*,*k*.

A Gibbs sampler is used to maximize the posterior density of the parameters by iteratively drawing samples from these conditional posterior distributions which converge towards the joint posterior distribution. Fortunately, there are closed forms of the densities to be drawn from and hence no samples need to be stored and the normalization constant can be computed. The authors demonstrate that the procedure is able to determine the correct number of components in a toy example and in a chemical shift imaging (CSI) dataset.

2.2.3 NMF with Automatic Relevance Determinant

In [62] presented automatic relevance determination for KL-NMF for model order selection which does not need to evaluate the evidence by formulating a MAP criterion:

$$C_{MAP}(\mathbf{W},\mathbf{H},\beta) = -\ln P\,(\mathbf{W},\mathbf{H},\beta|\mathbf{Y})$$
$$= -\ln P\,(\mathbf{Y}|\mathbf{W},\mathbf{H}) - \ln P\,(\mathbf{W}|\beta) - \ln P\,(\mathbf{H}|\beta)\ \ - \ln P\,(\beta) \qquad (2.22)$$

using KL-divergence log likelihood and independent half-normal priors on each column of $\mathbf{W}$ and row of $\mathbf{H}$ with precision parameter β_k

$$P(\mathbf{W}_{f,k}) = \sqrt{\frac{2}{\beta_k\pi}}\exp\left(-\frac{1}{2}\beta_k\mathbf{W}_{f,k}^{\mathbf{2}}\right), \qquad \mathbf{W}_{f,k} \geq 0 \qquad (2.23)$$

$$P(\mathbf{H}_{k,t}) = \sqrt{\frac{2}{\beta_k\pi}}\exp\left(-\frac{1}{2}\beta_k\mathbf{H}_{k,t}^{\mathbf{2}}\right), \qquad \mathbf{H}_{k,t} \geq 0 \qquad (2.24)$$

The precision parameters β_k are provided with a Gamma prior

$$P(\beta_k|a_k,b_k) = \frac{b_k^{a_k}}{\Gamma(a_k)}\,\beta_k^{a_k-1}\exp(-\beta_k b_k), \qquad \beta_k \geq 0 \qquad (2.25)$$

with fixed hyperparameters a and b.

A multiplicative algorithm optimizes C_{MAP} in (2.22) by iteratively updating **W**, **H** and β. The data automatically determines the optimal values of the hyperparameters β. The algorithm is initialized with a relatively large value K of components and successively drives unnecessary components to extinction. This property results from Bayesian inference: a subset of the precision parameters will be driven to an upper bound which corresponds to a sharp peak at zero for the priors on $\mathbf{W}_{*,k}$ and row $\mathbf{H}_{k,*}$ and leads to an effective extinction of column $\mathbf{W}_{*,k}$ and row $\mathbf{H}_{k,*}$. The effective number of components is determined by the number parameters β_k which are not driven to an upper bound during the iterations.

2.3 Complex Non-negative Matrix Factorization

This section presents a mixture model defined in the complex time-frequency domain. Complex Non-Negative Matrix Factorization model (CMF) [50-52, 80] is a sparse representation for acoustic signals which offers the advantages of the sparse coding (SC) [81-83] and non-negative matrix factorization (NMF) [5] concurrently. It can extract the recurrent patterns of magnitude spectra and the phase estimates of constituent signals, and can be performed with an efficient iterative algorithm. CMF shares with NMF the ability to generate non-negative matrices $\mathbf{W}$ and $\mathbf{H}$, while the input matrix $\mathbf{Y}$ is assumed to be a complex matrix and the algorithm also generates a third-rank complex-valued tensor as the following

$$\boldsymbol{Z}_{f,t}^{k} = e^{j\phi_{f,t}^{(k)}} \tag{2.26}$$

It is assumed that the short-time Fourier transform (STFT) of an audio signal, $\boldsymbol{X}_{f,t} \in \mathbb{C}$ in frequency bin f and time frame t, is composed of K complex-valued elements

$$\boldsymbol{X}_{f,t} \cong \sum_{k=1}^{K} \left|a_{f,t}^{k}\right| e^{j\phi_{f,t}^{k}} \tag{2.27}$$

Each $a_{f,t}^{k}$ is assumed to have a magnitude spectrum which is constant up to the gain over time:

$$\left|a_{f,t}^{k}\right| = \boldsymbol{W}_{f}^{k}\boldsymbol{H}_{t}^{k}(\forall_{f,k}\,\boldsymbol{W}_{f}^{k} \geq 0, \forall_{k,t}\,\boldsymbol{H}_{t}^{k} \geq 0\,) \tag{2.28}$$

and a time-varying phase spectrum

$$arg\left(a_{f,t}^{k}\right) = \phi_{f,t}^{k} \tag{2.29}$$

The CMF model can be expressed as

$$\boldsymbol{X}_{f,t} = \sum_k \boldsymbol{W}_f^k \boldsymbol{H}_t^k \,.\, e^{j\phi_{f,t}^{(k)}} \quad (\forall_{f,k}\, \boldsymbol{W}_f^k \ \geq 0, \forall_{k,t}\, \boldsymbol{H}_t^k \geq 0\ ) \tag{2.30}$$

where $\boldsymbol{W}_f^k$ corresponds to recurring magnitude spectral pattern, $\boldsymbol{H}_t^k$ to time-varying activation coefficients and $\phi_{f,t}^k$ to time-varying phase spectra and assume

$$\sum_f \boldsymbol{W}_f^k = 1 \quad (k = 1, \ldots, K) \tag{2.31}$$

in order to eliminate an indeterminacy in the scaling between $\boldsymbol{W}_f^k$ and $\boldsymbol{H}_t^k$.

The CMF method in [50] can be summarized as follows:

(1) Transform the single channel mixture of two sources: $x_1(t) = s_1(t) + s_2(t)$ form the time domain to the time-frequency domain using STFT.

(2) Initialize $\mathbf{W}, \mathbf{H}$ and $\boldsymbol{\phi}$.

(3) Update $\boldsymbol{\beta}_{f,t}^k = \frac{\boldsymbol{W}_f^k \boldsymbol{H}_t^k}{\sum_n \boldsymbol{W}_f^n \boldsymbol{H}_t^n}$.

(4) Stabilize the algorithm by running NMF at the beginning of the iteration, which can be performed simply by fixing the value of ϕ at $e^{j\phi_{f,t}^k} \leftarrow \frac{Y_{f,t}}{|Y_{f,t}|}$.

(5) The iterative algorithm is summarized as follows:

1. Update $\bar{\boldsymbol{\theta}} = \{\bar{\mathbf{Y}}, \bar{\mathbf{H}}\}$ by computing $\bar{\boldsymbol{Y}}_{f,t}^k = \boldsymbol{W}_f^k \boldsymbol{H}_t^k . e^{j\phi_{f,t}^k} + \boldsymbol{\beta}_{f,t}^k (\boldsymbol{Y}_{f,t} - \boldsymbol{X}_{f,t})$, and $\bar{\boldsymbol{H}}_t^k = \boldsymbol{H}_t^k$.

2. Update $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{H}, \boldsymbol{\phi}\}$ by computing

$$\boldsymbol{H}_t^k \leftarrow \frac{\sum_t \frac{\boldsymbol{H}_t^k}{\boldsymbol{\beta}_{f,t}^k} Re\left[\bar{Y}_{f,t}^{k\;*} . e^{j\phi_{f,t}^k}\right]}{\sum_t \frac{\boldsymbol{H}_t^{k\;2}}{\boldsymbol{\beta}_{f,t}^k}}, \ \boldsymbol{W}_f^k \leftarrow \frac{\sum_t \frac{\boldsymbol{W}_f^k}{\boldsymbol{\beta}_{f,t}^k} Re\left[\bar{Y}_{f,t}^{k\;*} . e^{j\phi_{f,t}^k}\right]}{\sum_t \frac{\boldsymbol{W}_f^{k\;2}}{\boldsymbol{\beta}_{f,t}^k} + \lambda p \left|\boldsymbol{H}_t^k\right|^{p-2}}, \text{ and } e^{j\phi_{f,t}^k} \leftarrow \frac{\bar{Y}_{f,t}^k}{\left|\bar{Y}_{f,t}^k\right|}.$$

3. Update $\boldsymbol{\beta}_{f,t}^k$ according to the equation $\boldsymbol{\beta}_{f,t}^k = \frac{\boldsymbol{W}_f^k \boldsymbol{H}_t^k}{\sum_n \boldsymbol{W}_f^n \boldsymbol{H}_t^n}$ and return to Step 1.

(6) Obtain the estimation of each source $\hat{S}_1$ and $\hat{S}_2$ by applying two different methods

(6.1) Atom selection method

The magnitude of atomic spectrum closest to the true spectrum was selected for each frame, and the framewise signals, each of which constructed using the selected atom and the corresponding activation coefficient and phase spectrum, were concatenated to synthesize the whole signal stream [50].

(6.2) Reconstruction

The reconstruction is calculated by multiplying the row of the spectral components $W_f^{(k)}$ with the corresponding column of the mixture weights $H_t^{(k)}$ and time-varying phrase spectrum $e^{j\phi_{f,t}^{(k)}}$. Then, convert the time-frequency represented sources back into time domain.

An experimental results of the CMF method showed that reasonably good separation performance on the single-channel audio source separation can be obtained.

2.4 Nonnegative Tensor Factorization Approach

In this case, the extension of NMF for solving multichannel mixtures has been regarded by stacking up the spectrograms of each channel into a single matrix [53]. This approach is considered as nonnegative tensor factorization (NTF), also called nonnegative parallel factor analysis (PARAFAC), where the channel spectrograms are jointly modeled by a 3-valence tensor [54]. NTF was introduced by Shashua and Hazan in [55] and has become a popular technique for data analysis and dimensionality reduction, parts-based representation of nonnegative data. Algorithms for NTF such as PARAFAC have been used for audio source separation in [54, 56]. Regardless of the cost function used, in order to achieve audio source separation, some methods require grouping of the basis functions according to the sources or instruments. Different grouping methods have been proposed by Casey [57] and Virtanen [6], but in practice, if the sources overlap in the time-frequency (TF) domain, it is difficult to obtain the correct clustering. This issue is discussed in [58]. Clustering of the spatial cues to group the NTF components (cNTF) [56] was developed for multichannel audio source separation. In most applications, it is crucial that the “right” model order K is selected. If K is too small, the data does not fit the model well. Conversely, if K is too large, then overfitting occurs. It is aimed to find an elegant solution for this dichotomy between data fidelity and overfitting. Choosing the right model is in particular challenging in the PARAFAC model as the number of components is specified for each modality separately. This delivers heuristics such as the Bayesian information criterion (BIC) [59] and Akaike information criterion (AIC) [60]. Both techniques cannot account for additional constraints such as non-negativity. Furthermore, a Bayesian approach of automatic relevance determination (ARD) was introduced by Mackay [61] to determine the relevant number of explanatory variables in the context of regression. This technique was used in [62 - 64] based on NMF model and multi-way models as in [65]. The spectral dictionary obtained via NMF-ARD [66] methods is not adequate to capture the temporal dependency of the frequency patterns within the audio signal. In addition, the NMF-ARD does not model musical notes but rather unique events only. Thus, if two notes are always played simultaneously they will be modeled as one component. Also, some components might not correspond to notes but rather to the model, e.g., background noise.

Nonnegative tensor factorization (NTF) has been proven to be a very useful tool in a variety of signal processing fields. Recently, NTF methods have successfully been exploited for data mining, dimensionality reduction, pattern recognition, object detection, gene clustering, sparse nonnegative representation and coding, and blind source separation (BSS) [67–75].

Given a data tensor $\mathbf{Y} \in \Re_{+}^{I\times F\times T}$ and the positive index K, the goal is to find three-component matrices, also called loading matrices, $\mathbf{Q} = [q_1, \dots, q_K] \in \Re_{+}^{I\times K}$, $\mathbf{W} = [w_1, \dots, w_K] \in \Re_{+}^{F\times K}$ and $\mathbf{H} = [h_1, \dots, h_K] \in \Re_{+}^{T\times K}$ which performs the following approximate factorization. $\mathbf{V}$ is the $I \times F \times T_s$ tensor with coefficient $V_i(f, t_s) = |Y_i(f, t_s)|^2$ and $\widehat{\mathbf{V}}$ is estimated $I \times F \times T_s$ tensor with coefficient $\hat{V}_i(f, t_s) = \sum_{k=1}^{K} q_{ik} w_{fk} h_{kt_s}$. The NTF under PARAFAC structure can be formulated in the element-wise form as follows

$$y_{ift_s} = \sum_{\boldsymbol{k=1}}^{K} q_{ik} w_{fk} h_{t_s k} + e_{ift_s} \tag{2.32}$$

A PARAFAC model is given by the matrices of $\mathbf{Q}$, $\mathbf{W}$, and $\mathbf{H}$ with elements q_{ik}, w_{fk} and $h_{t_s k}$, respectively. The trilinear model is found to minimize the sum of squares of the residuals, e_{ift_s} in the model. Figure 2.3 illustrates the principle of the PARAFAC model.

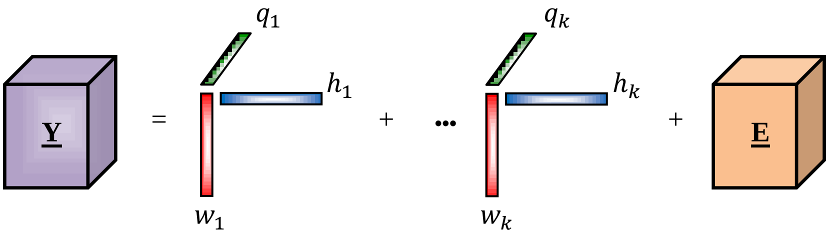


Figure 2.3: A graphical representation the principle of the decomposition of a 3-way data cube according to the PARAFAC model.

2.5 Summary

The state-of the art blind source separation method have been explained in this chapter. Generally speaking, more sparseness of the constructing components yields the better approximation. The NMF method is a flexible approach which can be developed as a new cost function, sparsity updating, and a new factorization for quantity analyzing of data to render better separation performance. Solving the BSS problem by using NMF approach has drawn huge interest from researchers in last two decades. However, the qualities of the reconstructed sources are not enough to launch the NMF solution in a real application.

CHAPTER 3

MAP-BASED REGULARIZED NONNEGATIVE TENSOR FACTORIZATION FOR MULTICHANNEL SOURCE SEPARATION

In this chapter, a novel approach to Bayesian regularized cluster nonnegative tensor factorization under a PARAFAC structure is presented. The proposed tensor factorization is developed under the framework of maximum *a posteriori* probability and is adaptively fine-tuned using the hierarchical Bayesian approach. The method enables: 1) a generalized criterion for variable sparseness to be imposed onto each element of the temporal code; and 2) modified multivariate rectified Gaussian prior information to be explicitly incorporated into the basis features. Underlying all factorization algorithms is the principal difficulty in estimating the adequate number of latent components for each source. This method takes the advantage of the combination of the automatic detection of the optimal K_d through both the pruning technique and the prior information on $\mathbf{W}$ to estimate the signature parameter of the original sources. This chapter addresses this important issue by using a framework of model selection for pruning unnecessary components and a novel Bayesian regularized cluster nonnegative tensor factorization under a PARAFAC structure with Itakura-Saito divergence. The experiments were designed to demonstrate on underdetermined linear instantaneous stereo mixtures of musical sources. The proposed method gives an average performance improvement of at least twice better than the state-of-art Itakura-Saito Nonnegative Tensor Factorization (IS-NTF) and Itakura-Saito with Cluster Nonnegative Tensor Factorization (IS-cNTF) methods, respectively.

The chapter is organized as follows: Section 3.1 introduces the background of NTF with IS Divergence. Section 3.2 describes the generative model of the proposed method, and the formulation of the NTF algorithm. Experimental results with a series of performance comparison with other NTF techniques are presented in Section 3.3. Finally, Section 3.4 concludes the chapter.

3.1 Background

3.1.1 NTF with Itakura-Saito Divergence

A statistical IS-NTF model of the observation $y_i(t)$ can be expressed as

$$y_i(t) = \sum_{k=1}^{K} m_{ik} c_{rk}(t) \tag{3.1}$$

where m_{ik} is defined as $m_{ik} = a_{id}$ if and only if $k \in K_d$. The a_{id} corresponds to mixing coefficient, $\{ a_{id}\}$ in a $I \times d_{max}$ mixing matrix. The components $c_{rk}(t)$ will be characterized by a spectral shape w_k and a vector of activation coefficient h_k through a statistical model and

$$C_{rk}(f, t_s) \sim \mathcal{N}_c(0|w_{fk}h_{kt_s}) \tag{3.2}$$

where $\mathcal{N}_c(\cdot)$ denotes the proper complex Gaussian distribution and $w_{fk}h_{kt_s}$ is the variance.

The factorization [15] is usually achieved through the minimization problem

$$minimize_{\mathbf{Q},\mathbf{W},\mathbf{H}}\, D(\mathbf{V}|\mathbf{QWH}) = D\big(\mathbf{V}\big|\widehat{\mathbf{V}}\big) = \sum_{ift} d\big(v_{ft_s}\big|\hat{v}_{ft_s}\big) \;\; subject\ to\ \ \mathbf{Q},\mathbf{W},\mathbf{H} \geq 0 \tag{3.3}$$

where $d(x|y)$ is a scalar cost function.

In this section, the term $d_{IS}(x|y)$ is exploited to be the IS divergence which is defined as [19]

$$d_{IS}(x|y) = \frac{x}{y} - log\frac{x}{y} - 1 \tag{3.4}$$

Thus, log-likelihood of the factor $\mathbf{W},\ \mathbf{H}$ and $\mathbf{Q}$ can be written as

$$\begin{aligned} -log\, p(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{Q}) &\doteq D_{IS}\big(|\mathbf{Y}|^{\mathbf{2}}\big|\mathbf{QWH}\big) \\ &= \sum_{ift_s} d_{IS}\Big(V_i(f,t_s)\Big|\hat{V}_i(f,t_s)\Big) \\ &= \sum_{ift_s}\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)} - log\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)} - 1 \end{aligned} \tag{3.5}$$

where $|\mathbf{Y}|^{\mathbf{2}}$ is the matrix with entries $\left|y_{ift}\right|^2$and “≐” denotes equality up to constant.

3.2 Proposed APBNTF Method

3.2.1 Generative Model

Under the linear instantaneous mixing and the point-sources assumption, the multichannel audio mixtures $y_i(t)$ can be generated by several unknown sources $x_{rd}(t)$ such that

$$y_i(t) = \sum_{d=1}^{d_{max}} a_{id} x_{rd}(t) \tag{3.6}$$

where $i = 1, \dots, I$ denotes the channel number, $d = 1, \dots, d_{max}$ denotes the source number, $t = 1, 2, \dots, T$ denotes time index, and $r = 1, \dots, R$ are assumed to share a certain "resemblance", as modelled by being two different realizations of the same random process, characterizing their time-frequency behavior, as opposed to be the same realization. In this work, the source can be further modeled as a sum of elementary components themselves, so that

$$x_{rd}(t) = \sum_{k \in K_d} c_{rk}(t) \tag{3.7}$$

where K_d represents the number of latent components associated with the d^{th} source, $d = 1, \dots, d_{max}$ where d_{max} is the maximum number of sources and $[K_1, \dots, K_{d_{max}}]$ denotes a nontrivial partition of $[1, \dots, K]$. Thus, the observation $y_i(t)$ can be expressed as

$$y_i(t) = \sum_{k=1}^{K} m_{ik} c_{rk}(t) \tag{3.8}$$

where m_{ik} is defined as $m_{ik} = a_{id}$ if and only if $k \in K_d$. The TF representation of the mixture in (3.8) is given by

$$Y_i(f, t_s) = \sum_{k=1}^{K} m_{ik} C_{rk}(f, t_s) \tag{3.9}$$

where $Y_i(f, t_s)$ and $C_{rk}(f, t_s)$ denote the TF components which are obtained by applying the linearity of short time Fourier transform (STFT) to the mixture. The time slots are given by $t_s = 1, 2, \dots, T_s$ while frequencies by $f = 1, 2, \dots, F$. Since each component is a function of t_s and f, the 3-valence tensor of mixture STFT $\mathbf{Y}_{\boldsymbol{i}} = [Y_i(f, t_s)]_{t_s=1,2,\dots,T_s}^{f=1,2,\dots,F}$, of size $I \times F \times T_s$, is modeled as a sum of K_d complex-valued latent tensor components $\boldsymbol{C}_{rk} = [C_{rk}(f, t_s)]_{t_s=1,2,\dots,T_s}^{f=1,2,\dots,F}$. In this case, the power spectrograms $|\mathbf{Y}_{\boldsymbol{i}}|^{\mathbf{2}}$ are approximated by a linear combination of nonnegative spectrograms $|C_{rk}(f, t_s)|^2 \simeq w_{fk} h_{kt_s}$ such that

$$\begin{aligned} |Y_i(f, t_s)|^2 &= \sum_{k=1}^{K} |m_{ik}|^2 |C_{rk}(f, t_s)|^2 \\ &= \sum_{k=1}^{K} q_{ik} |C_{rk}(f, t_s)|^2 \\ &\simeq \sum_{k=1}^{K} q_{ik} w_{fk} h_{kt_s} \end{aligned} \tag{3.10}$$

where $q_{ik} = |m_{ik}|^2$. Denoting the non-negative matrices $\mathbf{W} = \{w_{fk}\}$, $\mathbf{H} = \{h_{kt_s}\}$ and $\mathbf{Q} = \{q_{ik}\}$, the problem to solve is to separate the sources $x_{rd}(t)$ given by $|Y_i(f, t_s)|^2$ in (3.10).

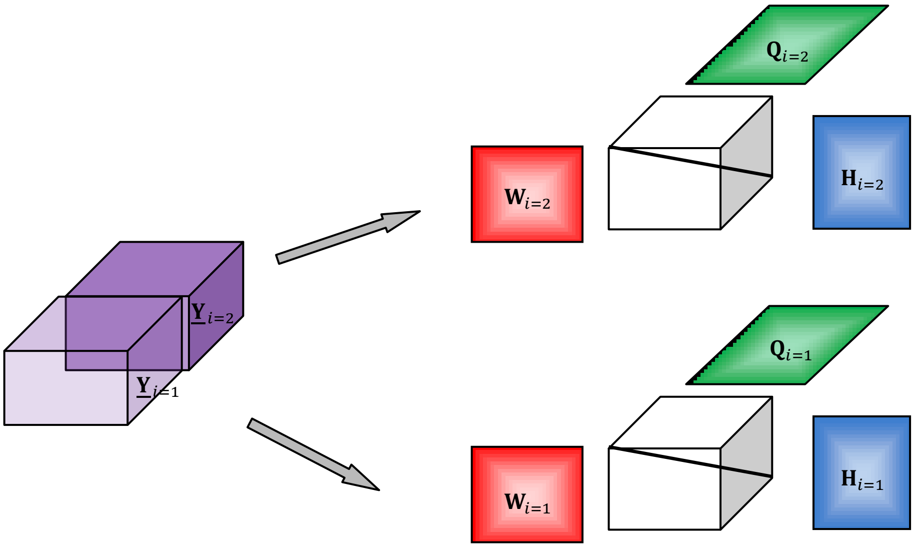


Figure 3.1: Illustration of the proposed method by using PARAFAC model for two channels ($i = 2$) source separation.

Figure 3.1 show the proposed method based on PARAFAC model for two channels source separation. The proposed method focuses on the estimate the three parameters $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ from two sub-tensors $\underline{\mathbf{Y}}_1$ and $\underline{\mathbf{Y}}_2$.

The proposed method focuses on the estimation of unknown parameters $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ of each sources. The estimates of $\mathbf{Q}$, $\mathbf{W}$ and $\mathbf{H}$ are used to reconstruct the original sources which are presented in Section 3.2.2.

3.2.2 Formulation of the Proposed Algorithm

In order to formulate the proposed algorithm, the parameters are firstly defined: $\mathbf{V}$ is the $I \times F \times T_s$ tensor with coefficients$V_i(f,t_s) = |Y_i(f,t_s)|^2$, $\hat{\mathbf{V}}$ is estimated the $I \times F \times T_s$ tensor with coefficients $\hat{V}_i(f,t_s) = \sum_{k=1}^{K} q_{ik} w_{fk} h_{kt_s}$, $\mathbf{P} = \{|a_{id}|^2\}$ is the $I \times d_{max}$ mixing matrix, $\mathbf{L} = \{l_{dk}\}$ is the $d_{max} \times K$ “labelling matrix” with only one nonzero value per column, i.e., such that

$$l_{dk} = \begin{cases} 1, if\ k \in K_d \\ 0, otherwise \end{cases} \tag{3.11}$$

and nonnegative vector $\boldsymbol{\lambda} = \{\lambda_{kt_s}\}$. The term $\mathbf{Q}$ can be expressed as follows:

$$\mathbf{Q} = \mathbf{PL} = \{|a_{id}|^2 l_{dk}\} \tag{3.12}$$

Thus, a prior distribution $p(\mathbf{W},\mathbf{H})$ is chosen over the factors $\{\mathbf{W},\mathbf{H}\}$. It shows a that the following optimization problem needs to be solved

$$min_{\mathbf{Q},\mathbf{W},\mathbf{H}}\, C_{MAP}\,(\mathbf{W},\mathbf{H}) \doteq -log\, p(\mathbf{W},\mathbf{H}|\mathbf{Y},\boldsymbol{\lambda},\mathbf{Q}) \tag{3.13}$$

The posterior can be found by using Bayes' theorem as

$$p(\mathbf{W},\mathbf{H}|\mathbf{Y},\boldsymbol{\lambda},\mathbf{Q}) = \frac{p(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{Q})\mathrm{p}(\mathbf{W},\mathbf{H}|\boldsymbol{\lambda})}{P(\mathbf{Y})} \tag{3.14}$$

where the denominator is a constant and therefore, the log-posterior can be expressed as

$$log\, p(\mathbf{W},\mathbf{H}|\mathbf{Y},\boldsymbol{\lambda},\mathbf{Q}) = log\, p(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{Q}) + log\, p(\mathbf{W},\mathbf{H}|\boldsymbol{\lambda}) + const \tag{3.15}$$

Thus, log-likelihood of the factor $\mathbf{W},\mathbf{H}$ and $\mathbf{Q}$ can be written as

$$-log\, p(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{Q}) \doteq \sum_{ift_s} d_{IS}(V_i(f,t_s)|\hat{V}_i(f,t_s))$$

$$= \sum_{ift_s} \frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)} - log\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)} - 1 \tag{3.16}$$

The second term on the right hand side of (3.15) consists of the prior distribution of $\mathbf{W}$ and $\mathbf{H}$ where it is assumed that they are jointly independent. In our proposed model, the prior over $\mathbf{W}$ is assumed to be distributed as $N_m(\mathbf{W}|0,\boldsymbol{\Sigma}_W)$ i.e. zero-mean modified multivariate Gaussian with covariance matrix $\boldsymbol{\Sigma}_W$ which will be developed. Since $\mathbf{W}$ is nonnegative, it is natural to assume that it satisfies the multivariate rectified Gaussian unlike other research which use the exponential distribution or the normal Gaussian distribution. However, the exponential distribution gives poorer sparseness than the Gaussian distribution. For a likelihood method based on Gaussian distribution, this is a simple Bayesian criterion for NMF. The Gaussian distribution causes the NMF many locally optimal solutions. Furthermore, it does not fit with the multiplicative NMF algorithm. On the other hand, the rectified Gaussian, where priors are conjugated to the

Gaussian likelihood, provides more flexible shapes of prior distribution. This benefits the distribution model to better suit the signals. Therefore, we propose the multivariate rectified Gaussian defined as

$$p(\mathbf{W}) = \Phi(-diag^{-1}(\boldsymbol{\Sigma}_W)\boldsymbol{u}_W)\delta(\boldsymbol{w}) + \left(\sqrt{2\pi}|\boldsymbol{\Sigma}_W|^2\right)^{(-1/2FK)} exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{u}_W)^{\boldsymbol{T}}\boldsymbol{\Sigma}_W^{-1}(\boldsymbol{w}-\boldsymbol{u}_W)\right)\boldsymbol{U}(\boldsymbol{w}) \tag{3.17}$$

where $\boldsymbol{w} = vec(\mathbf{W}) = [\boldsymbol{w}_1^{\boldsymbol{T}} \vdots\ \boldsymbol{w}_2^{\boldsymbol{T}} \vdots \cdots \vdots\ \boldsymbol{w}_K^{\boldsymbol{T}}]^{\boldsymbol{T}}$, $\delta(\boldsymbol{w}) = \begin{cases} +\infty, & \boldsymbol{w} = 0 \\ 0, & \boldsymbol{w} \neq 0 \end{cases}$, '$\boldsymbol{T}$' denotes matrix transpose, $vec(\cdot)$ represents the column vectorization, $\boldsymbol{U}(\boldsymbol{w}) = \begin{cases} 0, \boldsymbol{w} \leq 0 \\ 1, \boldsymbol{w} > 0 \end{cases}$, and $\Phi(\bullet)$ denotes the multivariate Gaussian cumulative distribution function. Considering the zero mean of the rectified Gaussian distribution (i.e. set $\boldsymbol{u}_W = 0$) on the latent variable would better suit most of the real world data and can enable the induction of sparse positive factors, the expression (3.17) results in

$$p(\boldsymbol{W}) = 0.5\delta(\boldsymbol{w}) + \left(\sqrt{2\pi}|\boldsymbol{\Sigma}_W|^2\right)^{(-1/2FK)} exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{u}_W)^{\boldsymbol{T}}\,\boldsymbol{\Sigma}_W^{-1}(\boldsymbol{w}-\boldsymbol{u}_W)\right) \tag{3.18}$$

In applications, $\mathbf{W}$ represents the basis vectors that span the domain of the input matrix $\mathbf{V}_i$. Although the exact values of $\mathbf{W}$ are case specific, one is almost warranted that in most cases the probability of having zero-valued basis vectors i.e. $p(\mathbf{W} = \mathbf{0})$ is very rare. Thus, the above takes the form of

$$p(\mathbf{W}) \propto \begin{cases} exp\left(-\frac{1}{2}\boldsymbol{w}^T\boldsymbol{\Sigma}_W^{-1}\boldsymbol{w}\right), & \boldsymbol{w} \geq 0 \\ 0\ , & \boldsymbol{w} < 0 \end{cases} \tag{3.19}$$

where $\boldsymbol{\Sigma}_W = \begin{bmatrix} \boldsymbol{\Sigma}_{1,1} & \cdots & \boldsymbol{\Sigma}_{1,K} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{I,1} & \cdots & \boldsymbol{\Sigma}_{K,K} \end{bmatrix}$ is the covariance matrix of $\boldsymbol{w} = vec(\mathbf{W})$ and $\boldsymbol{\Sigma}_{k,j} = E\left[\boldsymbol{w_k}\boldsymbol{w}_j^{\boldsymbol{T}}\right]$ is the cross-correlation matrix between the basis vectors $\boldsymbol{w_k}$ and $\boldsymbol{w_j}$, "$E[\cdot]$" denotes the statistical expectation operator. The covariance matrix $\boldsymbol{\Sigma}_W$ can be partitioned as

$$\boldsymbol{\Sigma}_W = \underbrace{\begin{bmatrix} \boldsymbol{\Sigma}_{1,1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{2,2} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \ddots & \vdots \\ \mathbf{0} & \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{\Sigma}_{K,K} \end{bmatrix}}_{\boldsymbol{\Sigma}_{diag}^{\boldsymbol{W}}} + \underbrace{\begin{bmatrix} \mathbf{0} & \boldsymbol{\Sigma}_{1,2} & \cdots & \cdots & \boldsymbol{\Sigma}_{1,K} \\ \boldsymbol{\Sigma}_{2,1} & \mathbf{0} & \boldsymbol{\Sigma}_{2,3} & \cdots & \vdots \\ \vdots & \boldsymbol{\Sigma}_{3,2} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \boldsymbol{\Sigma}_{K-1,K} \\ \boldsymbol{\Sigma}_{K,1} & \cdots & \cdots & \boldsymbol{\Sigma}_{K,K-1} & \mathbf{0} \end{bmatrix}}_{\boldsymbol{\Sigma}_{off}^{\boldsymbol{W}}} \tag{3.20}$$

In the above, $\mathbf{0}$ is a $K \times K$ matrix with zero elements. The inverse covariance matrix can be approximated as

$$\begin{aligned} \boldsymbol{\Sigma}_W^{-1} &= (\boldsymbol{\Sigma}_{diag}^W + \boldsymbol{\Sigma}_{off}^W)^{-1} \\ &\approx \boldsymbol{\Sigma}_{diag}^{-1(W)} - \boldsymbol{\Sigma}_{diag}^{-1(W)}\boldsymbol{\Sigma}_{off}^W\boldsymbol{\Sigma}_{diag}^{-1(W)} \\ &= \boldsymbol{\Omega}_{diag}^W - \boldsymbol{\Omega}_{off}^W \end{aligned} \tag{3.21}$$

where $\boldsymbol{\Sigma}_{diag}^{-1(W)}$ is the inverse covariance matrix of $\boldsymbol{\Sigma}_{diag}^W$, $\boldsymbol{\Omega}_{diag}^W = \boldsymbol{\Sigma}_{diag}^{-1(W)}$ and $\boldsymbol{\Omega}_{off}^W = \boldsymbol{\Sigma}_{diag}^{-1(W)}\boldsymbol{\Sigma}_{off}^W\boldsymbol{\Sigma}_{diag}^{-1(W)}$. The $(k,j)^{th}$ sub-matrix of $\boldsymbol{\Omega}_{off}^W$ is given by

$$\boldsymbol{\Omega}_{off,k,j}^W = \boldsymbol{\Sigma}_{k,k}^{-1(W)}\boldsymbol{\Sigma}_{k,j}^W\boldsymbol{\Sigma}_{j,j}^{-1(W)} \tag{3.22}$$

It can be shown that when the elements within the same basis vector are uncorrelated, the above matrices simplify to $\boldsymbol{\Sigma}_{k,k}^W = \sigma_k^2\boldsymbol{I}$, $\boldsymbol{\Sigma}_{j,j}^W = \sigma_j^2\boldsymbol{I}$ and $\boldsymbol{\Sigma}_{k,j}^W = c_{k,j}\boldsymbol{I}$ where σ_k^2 is the variance of the basis vector $\boldsymbol{w}_k$ and $c_{k,j}$ is the cross-covariance between $\boldsymbol{w}_k$ and $\boldsymbol{w}_j$. Thus, $\boldsymbol{\Omega}^W_{off,k,j}$ can be simplified to

$$\boldsymbol{\Omega}^W_{off,k,j} = \mu_{kj}\boldsymbol{I} \tag{3.23}$$

where

$$\mu_{kj} = \frac{\gamma_{kj}}{\sigma_k\sigma_j} \; and \quad \gamma_{kj} = \frac{c_{k,j}}{\sigma_k\sigma_j} \tag{3.24}$$

Using the above, (3.19) can be cast into two terms:

$$\begin{aligned} -\,log\, p(\mathbf{W}) &\doteq \frac{1}{2}\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\Omega}^W_{diag}\boldsymbol{w} - \frac{1}{2}\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\Omega}^W_{off}\boldsymbol{w} \\ &= \gamma - \tfrac{1}{2}\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\Omega}^W_{off}\boldsymbol{w} \end{aligned} \tag{3.25}$$

The first term $\boldsymbol{\gamma} = \frac{1}{2}\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\Omega}^W_{diag}\boldsymbol{w}$ relates only to the power of $\boldsymbol{w}$ while the second term $\frac{1}{2}\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\Omega}^W_{off}\boldsymbol{w} = \frac{1}{2}\sum_{k,j,(k\neq j)}\mu_{kj}\boldsymbol{w}_k^{\boldsymbol{T}}\boldsymbol{w}_{\boldsymbol{j}}$ measures the sum of weighted correlation between $\boldsymbol{w}_{\boldsymbol{k}}$ and $\boldsymbol{w}_{\boldsymbol{j}}$ for al$l\; k, j, (k \neq j)$. Hence, the interesting information is actually contained in the second term which represents the prior information of the basis vectors. By including this term, the underlying correlation between the different basis vectors can be incorporated into the matrix factorization to yield results that reflect on this prior information. Therefore, with the factorial model in (3.25) the desired constraint assumes the following form:

$$f(\mathbf{W}) = -\,log\, p(\mathbf{W}) \doteq -\tfrac{1}{2}\sum_{k,j,(k\neq j)}\mu_{kj}\boldsymbol{w}_k^{\boldsymbol{T}}\boldsymbol{w}_{\boldsymbol{j}} \tag{3.26}$$

In this section, the probabilistic framework is used for the purpose of developing a platform to incorporate the statistical correlation between $\boldsymbol{w_k}$ and $\boldsymbol{w_j}$ into the matrix factorization as part of the regularization. In feature extraction, such constraint is required in order to fully extract the basis especially in the situation where the patterns contain overlapping features. Despite of the proposed prior model for **W** stems from the modified Gaussian distribution, it is a combination of constrained and unconstrained parameterization of the inverse covariance matrix.

In order to turn off excess components thereby optimizing for K, the component-wise exponential distribution prior is imposed on **H**, namely,

$$p(\mathbf{H}|\boldsymbol{\lambda}) = \prod_k \prod_{t_s} \lambda_{kt_s} exp\,(-\lambda_{kt_s} h_{kt_s}) \tag{3.27}$$

Following (3.27), the negative log prior on **H** is defined as

$$\begin{aligned} f(\mathbf{H}) &= -\,log\, p(\mathbf{H}|\boldsymbol{\lambda}) = -\sum_k \sum_{t_s} \left\{ log\, \lambda_{kt_s} - \lambda_{kt_s} h_{kt_s} \right\} \\ &= -\sum_k \sum_{t_s} log\, \lambda_{kt_s} + \sum_k \sum_{t_s} \lambda_{kt_s} h_{kt_s} \end{aligned} \tag{3.28}$$

By substituting (3.16), (3.19) and (3.27) into (3.14), the negative log posterior of **W** and $\boldsymbol{H}$ is given by the following:

$$-\,log\, p(\mathbf{W}, \mathbf{H}|\mathbf{Y}, \boldsymbol{\lambda}, \mathbf{Q}) \doteq -\,log\, p(\mathbf{Y}|\mathbf{W}, \mathbf{H}, \mathbf{Q}) - log\, p(\mathbf{W}) - log\, p(\mathbf{H}|\boldsymbol{\lambda}) \tag{3.29}$$

From (3.16), (3.26) and (3.28), the above can be written as

$$L \doteq \sum_{ift_s} d_{IS}(V_i(f,t_s)|\widehat{V}_i(f,t_s)) + f(\boldsymbol{W}) + f(\boldsymbol{H})$$

$$= \sum_{ift_s} \frac{V_i(f,t_s)}{\widehat{V}_i(f,t_s)} - log\frac{V_i(f,t_s)}{\widehat{V}_i(f,t_s)} - 1 - \frac{1}{2}\sum_{k,j,(k\neq j)} \mu_{kj}\boldsymbol{w}_k^{\boldsymbol{T}}\boldsymbol{w_j} - \sum_k \sum_{t_s} log\,\lambda_{kt_s} \;+\; \sum_k \sum_{t_s} \lambda_{kt_s} h_{kt_s} \tag{3.30}$$

The sparsity term $\sum_k \sum_{t_s} \lambda_{kt_s} h_{kt_s}$ forms the L_1-norm regularization to resolve the permutation ambiguity by forcing all structure in **H** onto **W** Therefore, the sparseness of the solution in (3.30) is highly dependent on the regularization parameter λ_{kt_s}.

3.2.2.1 Estimation of the mixing, basis and code

In this section, the estimations of **W**, **H** and $\mathbf{P} = \{|a_{id}|^2\}$ are presented. The derivative of (3.30) with respect to **W** of the proposed model is given by:

$$\begin{aligned}
\frac{\partial L}{\partial w_{fk}} &= \frac{\partial}{\partial w_{fk}}\begin{pmatrix} \sum_{ift_s} \frac{V_i(f,t_s)}{\widehat{V_\iota}(f,t_s)} - log\frac{V_i(f,t_s)}{\widehat{V_\iota}(f,t_s)} - 1 - \frac{1}{2}\sum_{k,j,(k\neq j)} \mu_{kj}\boldsymbol{w}_k^{\boldsymbol{T}}\boldsymbol{w_j} \\ -\sum_k \sum_{t_s} log\,\lambda_{kt_s} + \sum_k \sum_{t_s} \lambda_{kt_s} h_{kt_s} \end{pmatrix} \\
&= \sum_{ift_s} \frac{\partial}{\partial w_{fk}}\left(\frac{V_i(f,t_s)}{\widehat{V_\iota}(f,t_s)}\right) - \frac{\partial}{\partial w_{fk}} log\left(\frac{V_i(f,t_s)}{\widehat{V_\iota}(f,t_s)}\right) - \frac{1}{2}\sum_{k,j,(k\neq j)} \mu_{kj}\boldsymbol{w_j}\frac{\partial}{\partial w_{fk}}\boldsymbol{w}_k^{\boldsymbol{T}} \\
&= \sum_{ift_s} V_i(f,t_s)\frac{\partial\,(\widehat{V_\iota}(f,t_s))^{-1}}{\partial w_{fk}} - \frac{\partial\,(logV_i(f,t_s) - log\,\widehat{V_\iota}(f,t_s))}{\partial w_{fk}} - \frac{1}{2}\sum_{j\neq k} \mu_{kj} w_{fj} \\
&= -\sum_{ift_s} \frac{V_i(f,t_s)}{\widehat{V_\iota}(f,t_s)^2}\frac{\partial\,\widehat{V_\iota}(f,t_s)}{\partial w_{fk}} + \frac{\partial\,log\,\widehat{V_\iota}(f,t_s)}{\partial w_{fk}} - \frac{1}{2}\sum_{j\neq k} \mu_{kj} w_{fj} \\
&= -\sum_{ift_s} \frac{V_i(f,t_s)}{\widehat{V_\iota}(f,t_s)^2}\frac{\partial\,\sum_k q_{ik}h_{kt_s}w_{fk}}{\partial w_{fk}} + \frac{1}{\widehat{V_\iota}(f,t_s)}\frac{\partial\,\widehat{V_\iota}(f,t_s)}{\partial w_{fk}} - \frac{1}{2}\sum_{j\neq k} \mu_{kj} w_{fj} \\
&= -\sum_{it_s} q_{ik}h_{kt_s}\frac{V_i(f,t_s)}{\widehat{V_\iota}(f,t_s)^2} + \frac{1}{\widehat{V_\iota}(f,t_s)}\frac{\partial\,\sum_k q_{ik}h_{kt_s}w_{fk}}{\partial w_{fk}} - \frac{1}{2}\sum_{j\neq k} \mu_{kj} w_{fj} \\
&= -\sum_{it_s} q_{ik}h_{kt_s}\frac{V_i(f,t_s)}{\widehat{V_\iota}(f,t_s)^2} + \sum_k q_{ik}h_{kt_s}\frac{1}{\widehat{V_\iota}(f,t_s)} - \frac{1}{2}\sum_{j\neq k} \mu_{kj} w_{fj} \\
&= \sum_{it_s} q_{ik}h_{kt_s}\left[\frac{1}{\widehat{V_\iota}(f,t_s)^2} - \frac{V_i(f,t_s)}{\widehat{V_\iota}(f,t_s)^2}\right] - \frac{1}{2}\sum_{j\neq k} \mu_{kj} w_{fj} \\
&= \sum_{it_s} q_{ik}h_{kt_s}d_{IS}'\big(V_i(f,t_s)|\widehat{V_\iota}(f,t_s)\big) - \frac{1}{2}\sum_{j\neq k} \mu_{kj} w_{fj}
\end{aligned} \tag{3.31}$$

Similarly, the derivative of (3.30) with respect to **H** is given by

$$
\begin{aligned}
\frac{\partial L}{\partial h_{kt_s}} &= \frac{\partial}{\partial h_{kt_s}}\left(\begin{array}{c}\sum_{ift_s}\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)} - log\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)} - 1 - \frac{1}{2}\sum_{k,j,(k\neq j)}\mu_{kj}\boldsymbol{w}_k^{\boldsymbol{T}}\boldsymbol{w}_{\boldsymbol{j}} \\ -\sum_k\sum_{t_s} log\,\lambda_{kt_s} + \sum_k\sum_{t_s}\lambda_{kt_s}h_{kt_s}\end{array}\right) \\
&= \sum_{ift_s}\frac{\partial}{\partial h_{kt_s}}\left(\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)}\right) - \frac{\partial}{\partial h_{kt_s}} log\left(\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)}\right) + \frac{\partial}{\partial h_{kt_s}}\sum_k\sum_{t_s}\lambda_{kt_s}h_{kt_s} \\
&= \sum_{ift_s}V_i(f,t_s)\frac{\partial\,(\hat{V}_i(f,t_s))^{-1}}{\partial h_{kt_s}} - \frac{\partial\, log(V_i(f,t_s) - log\,\hat{V}_i(f,t_s))}{\partial h_{kt_s}} + \lambda_{kt_s} \\
&= -\sum_{ift_s}\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2}\frac{\partial\ \hat{V}_i(f,t_s)}{\partial h_{kt_s}} + \frac{\partial\ log\,\hat{V}_i(f,t_s)}{\partial h_{kt_s}} + \lambda_{kt_s} \\
&= -\sum_{ift_s}\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2}\frac{\partial\sum_k q_{ik}h_{kt_s}w_{fk}}{\partial h_{kt_s}} + \frac{1}{\hat{V}_i(f,t_s)}\frac{\partial\hat{V}_i(f,t_s)}{\partial h_{kt_s}} + \lambda_{kt_s} \\
&= -\sum_{t_sf} q_{ik}w_{fk}\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2} + \frac{1}{\hat{V}_i(f,t_s)}\frac{\partial\sum_k q_{ik}h_{kt_s}w_{fk}}{\partial h_{kt_s}} + \lambda_{kt_s} \\
&= -\sum_{t_sf} q_{ik}w_{fk}\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2} + \sum_k q_{ik}w_{fk}\frac{1}{\hat{V}_i(f,t_s)} + \lambda_{kt_s} \\
&= \sum_{t_sf} q_{ik}w_{fk}\left[\frac{1}{\hat{V}_i(f,t_s)^2} - \frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2}\right] + \lambda_{kt_s} \\
&= \sum_{it_s} q_{ik}w_{fk}d'_{IS}\big(V_i(f,t_s)|\hat{V}_i(f,t_s)\big) + \lambda_{kt_s} \qquad (3.32)
\end{aligned}
$$

The derivative of (3.30) with respect to $\mathbf{P} = \{|a_{id}|^2\}$ is given by

$$
\begin{aligned}
\frac{\partial L}{\partial p_{id}} &= \frac{\partial}{\partial p_{id}}\left(\begin{array}{c}\sum_{ift_s}\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)} - log\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)} - 1 - \frac{1}{2}\sum_{k,j,(k\neq j)}\mu_{kj}\boldsymbol{w}_k^{\boldsymbol{T}}\boldsymbol{w}_{\boldsymbol{j}} \\ -\sum_k\sum_{t_s} log\,\lambda_{kt_s} + \sum_k\sum_{t_s}\lambda_{kt_s}h_{kt_s}\end{array}\right) \\
&= \sum_{ift_s}\frac{\partial}{\partial p_{id}}\left(\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)}\right) - \frac{\partial}{\partial p_{id}} log\left(\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)}\right) \\
&= \sum_{ift_s}V_i(f,t_s)\frac{\partial\,(\hat{V}_i(f,t_s))^{-1}}{\partial p_{id}} - \frac{\partial\,(logV_i(f,t_s) - log\,\hat{V}_i(f,t_s))}{\partial p_{id}} \\
&= -\sum_{ift_s}\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2}\frac{\partial\ \hat{V}_i(f,t_s)}{\partial p_{id}} + \frac{\partial\ log\,\hat{V}_i(f,t_s)}{\partial p_{id}} \\
&= -\sum_{ift_s}\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2}\frac{\partial\sum_k q_{ik}h_{kt_s}w_{fk}}{\partial p_{id}} + \frac{1}{\hat{V}_i(f,t_s)}\frac{\partial\hat{V}_i(f,t_s)}{\partial p_{id}} \\
&= -\sum_{ift_s}\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2}\frac{\partial\sum_k p_{id}l_{dk}h_{kt_s}w_{fk}}{\partial p_{id}} + \frac{1}{\hat{V}_i(f,t_s)}\frac{\partial\sum_k q_{ik}h_{kt_s}w_{fk}}{\partial p_{id}} \\
&= -\sum_{ift_s}\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2}\frac{\sum_k l_{dk}h_{kt_s}w_{fk}\partial\, p_{id}}{\partial p_{id}} + \frac{1}{\hat{V}_i(f,t_s)}\frac{\partial\sum_k p_{id}l_{dk}h_{kt_s}w_{fk}}{\partial p_{id}} \\
&= -\sum_k l_{dk}\sum_{f,t_s} w_{fk}h_{kt_s}\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2} + \sum_k l_{dk}\sum_{f,t_s} w_{fk}h_{kt_s}\frac{1}{\hat{V}_i(f,t_s)} \\
&= \sum_k l_{dk}\sum_{f,t_s} w_{fk}h_{kt_s}\left[\frac{1}{\hat{V}_i(f,t_s)^2} - \frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2}\right] \\
&= \sum_k l_{dk}\sum_{f,t_s} w_{fk}h_{kt_s}d'_{IS}\big(V_i(f,t_s)|\hat{V}_i(f,t_s)\big) \qquad (3.33)
\end{aligned}
$$

From (3.31), (3.32) and (3.33), these then obtain

$$\begin{aligned} \nabla_{w_{fk}} &= \sum_{i,t_s} q_{ik} h_{kt_s} d_{IS}^{'}(V_i(f,t_s)|\hat{V}_i(f,t_s)) - \frac{1}{2}\sum_{j\neq k} \mu_{kj} w_{fj} \\ \nabla_{h_{kt_s}} &= \sum_{i,f} q_{ik} w_{fk} d_{IS}^{'}(V_i(f,t_s)|\hat{V}_i(f,t_s)) + \lambda_{kt_s} \\ \nabla_{P_{id}} &= \sum_k l_{dk} \sum_{f,t_s} w_{fk} h_{kt_s} d_{IS}^{'}(V_i(f,t_s)|\hat{V}_i(f,t_s)) \end{aligned} \tag{3.34}$$

The term $\mathbf{G}$ is defined as $I \times F \times T_s$ tensor with entries $g_{ift_s} = d'_{IS}(V_i(f,t_s)|\hat{V}_i(f,t_s))$, namely

$$d'_{IS}\left(V_i(f,t_s)\middle|\hat{V}_i(f,t_s)\right) = \frac{1}{\hat{V}_i(f,t_s)} - \frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2} \tag{3.35}$$

Generally, the multiplicative algorithms is importantly for updating θ. Each parameter θ is estimated by multiplying its value at previous iteration by the ratio of the negative and positive part of the derivative criterion with respect to this parameter as,

$$\theta \longleftarrow \theta \frac{[\nabla_\theta D(\mathbf{V}|\hat{\mathbf{V}})]_-}{[\nabla_\theta D(\mathbf{V}|\hat{\mathbf{V}})]_+} \tag{3.36}$$

where

$$\nabla_\theta D(\mathbf{V}|\hat{\mathbf{V}}) = [\nabla_\theta D(\mathbf{V}|\hat{\mathbf{V}})]_+ - [\nabla_\theta D(\mathbf{V}|\hat{\mathbf{V}})]_- \tag{3.37}$$

and the terms in the right hand side of (3.37) are both positive [59]. Here, the term $\mathbf{G}_-$ follows the multiplicative update (MU) rule that denotes the negative part of the derivative of the criterion e.g. $\mathbf{G}_- = \left[d'_{IS}\left(V_i(f,t_s)\middle|\hat{V}_i(f,t_s)\right)\right]_- = \frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2}$ and $\mathbf{G}_+$ denotes its positive part of the derivative of the criterion, $\mathbf{G}_+ = \left[d'_{IS}\left(V_i(f,t_s)\middle|\hat{V}_i(f,t_s)\right)\right]_+ = \frac{1}{\hat{V}_i(f,t_s)}$. The term $\mathbf{Q} \circ \mathbf{H}$ denotes

$I \times T_s \times K$ tensor with elements $q_{ik}h^{\boldsymbol{T}}_{kt_s}$. Similarly, $\mathbf{Q} \circ \mathbf{W}$ denotes $I \times F \times K$ tensor with elements $q_{ik}w_{fk}$ and $\mathbf{W} \circ \mathbf{H}$ denotes $F \times T_s \times K$ tensor with elements $w_{fk}h^{\boldsymbol{T}}_{kt_s}$.

Notice that $\langle \bar{\mathbf{A}}, \bar{\mathbf{B}} \rangle_{k_{\bar{A}}, k_{\bar{B}}}$ denotes the contracted product between tensors $\bar{\mathbf{A}}$ with size $I_1 \times \ldots \times I_M \times J_1 \times \ldots \times J_N$ and $\bar{\mathbf{B}}$ with size $I_1 \times \ldots \times I_M \times K_1 \times \ldots \times K_P$ and $k_{\bar{A}}$ and $k_{\bar{B}}$ are the sets of mode indices over which the summation take place. The contracted product $\langle \bar{\mathbf{A}}, \bar{\mathbf{B}} \rangle_{\{1,\ldots,M\},\{1,\ldots,M\}}$ is a tensor of size $J_1 \times \ldots \times J_N \times K_1 \times \ldots \times K_P$ given by

$$\langle \bar{\mathbf{A}}, \bar{\mathbf{B}} \rangle_{\{1,\ldots,M\},\{1,\ldots,M\}} = \sum_{i_1=1}^{I_1} \cdots \sum_{i_M}^{I_M} \bar{a}_{i_1,\ldots,i_M,j_1,\ldots,j_N} \bar{b}_{i_1,\ldots,i_M,k_1,\ldots,k_P} \tag{3.38}$$

The contracted tensor product is a form of a generalized dot product of two tensors along common modes of same dimensions. For example, in this chapter the contracted tensor product along the mode $\{1,3\}$ of a tensor $\underline{\mathbf{G}} \in \mathbb{R}^{I \times F \times T_s}$ and the mode $\{1,2\}$ of a tensor $\mathbf{Q} \circ \mathbf{H} \in \mathbb{R}^{I \times T_s \times K}$ returns a tensor $\langle \underline{\mathbf{G}}, \mathbf{Q} \circ \mathbf{H} \rangle_{\{1,3\},\{1,2\}} \in \mathbb{R}^{F \times K}$ as in Figure 3.2.

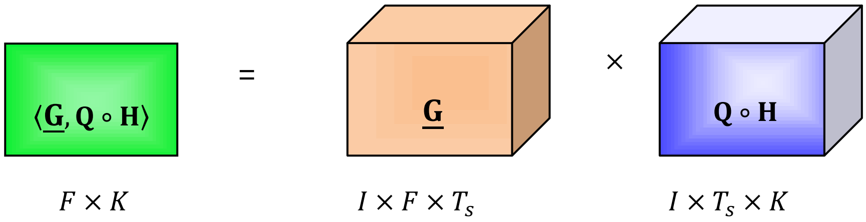


Figure 3.2: Illustration of mode 2 multiplications for the case of 3rd order tensor $\underline{\mathbf{G}}$ and $\mathbf{Q} \circ \mathbf{H}$ results in a 2-way tensor (a matrix) $\langle \underline{\mathbf{G}}, \mathbf{Q} \circ \mathbf{H} \rangle$.

Using (3.30), (3.32), (3.33) and (3.38) the MU rule for $\mathbf{P}$ is obtained as

$$\mathbf{P}_{id} \leftarrow \mathbf{P} \bullet \frac{\langle \boldsymbol{G}_{-}, \mathbf{W} \circ \mathbf{H} \rangle_{\{2,3\},\{1,2\}} \boldsymbol{L}^{\boldsymbol{T}}}{\langle \boldsymbol{G}_{+}, \mathbf{W} \circ \mathbf{H} \rangle_{\{2,3\},\{1,2\}} \boldsymbol{L}^{\boldsymbol{T}}}$$

$$= \mathbf{P}_{id} \bullet \frac{\sum_{t_s k} \boldsymbol{G}_{-ift_s} (\mathbf{W} \circ \mathbf{H})_{ft_s k} \boldsymbol{L}^{\boldsymbol{T}}_{id}}{\sum_{t_s k} \boldsymbol{G}_{+ift_s} (\mathbf{W} \circ \mathbf{H})_{ft_s k} \boldsymbol{L}^{\boldsymbol{T}}_{id}} \tag{3.39}$$

Similarly, the MU rule for $\mathbf{H}$ is

$$\mathbf{H}_{t_s k} \leftarrow \mathbf{H} \bullet \frac{\langle \boldsymbol{G}_{-}, \mathbf{Q} \circ \mathbf{W} \rangle_{\{1,2\},\{1,2\}}}{\langle \boldsymbol{G}_{+}, \mathbf{Q} \circ \mathbf{W} \rangle_{\{1,2\},\{1,2\}} + \boldsymbol{\lambda} \mathbf{1}^{\boldsymbol{T}}}$$

$$= \mathbf{H}_{t_s k} \bullet \frac{\sum_{fk} \boldsymbol{G}_{-i f t_s} (\boldsymbol{Q} \circ \boldsymbol{W})_{ifk}}{\sum_{fk} \boldsymbol{G}_{+i f t_s} (\boldsymbol{Q} \circ \boldsymbol{W})_{ifk} + \boldsymbol{\lambda}_{t_s k} \mathbf{1}_{kk}^{\boldsymbol{T}}} \tag{3.40}$$

and $\mathbf{W}$ is updated by using

$$\mathbf{W}_{fk} \leftarrow \mathbf{W} \bullet \frac{\langle \boldsymbol{G}_{-}, \mathbf{Q} \circ \mathbf{H} \rangle_{\{1,3\},\{1,2\}}}{\langle \boldsymbol{G}_{+}, \mathbf{Q} \circ \mathbf{H} \rangle_{\{1,3\},\{1,2\}} + \mathbf{W} \boldsymbol{\Xi}^{\boldsymbol{T}}}$$

$$= \mathbf{W}_{fk} \bullet \frac{\sum_{ik} \boldsymbol{G}_{-i f t_s} (\mathbf{Q} \circ \mathbf{H})_{i t_s k}}{\sum_{ik} \boldsymbol{G}_{+i f t_s} (\mathbf{Q} \circ \mathbf{H})_{i t_s k} + \mathbf{W}_{fk} \boldsymbol{\Xi}_{kk}^{\boldsymbol{T}}} \tag{3.41}$$

where '•' is element-wise product and $\boldsymbol{\Xi}^{\boldsymbol{T}}$ is a $K \times K$ matrix whose $(k,j)^{th}$ element is given by μ_{kj} except its diagonal elements being zeros.

3.2.2.2 Estimation of the adaptive sparsity parameter

The update of λ_k follows by solving $\frac{\partial L}{\partial \lambda_{k t_s}} = 0$, this gives

$$\frac{\partial L}{\partial \lambda_{k t_s}} = \frac{1}{\lambda_k} - h_{k t_s}$$

$$\lambda_{k t_s} = \frac{1}{h_{k t_s}} \tag{3.42}$$

Note that the sparsity term $\sum_k \sum_{t_s} \lambda_{k t_s} h_{k t_s}$ forms the sparse NTF objectives while the normalization term $\sum_k \sum_{t_s} log\, \lambda_{k t_s}$ is given to learn the degree of regularization from data, i.e. tune the pruning parameter, λ_{kt_s}. The concept introduced in [63] based on the assumption that the factorization in (3.10) has been used for an approximation error of $\sum_{i=1}^{I}\sum_{f=1}^{F}\sum_{t_s=1}^{T_s}|Y_i(f,t_s)|^2/IFT_s$. The pruning will turn off excess components thereby optimizing K_d by following component-wise exponential prior on **H** and the model parameters based on the MAP estimation. As a result of inference in (3.30), a subset of the λ_{kt_s} will be driven to a large upper bound, with the corresponding columns of **W** and rows of **H** driven to small values. The effective dimensionality can be deduced from the distribution of the λ_{kt_s} such that, it has been found in practice, two clusters clearly emerge: A group of values in same order of magnitude corresponding to relevant components on columns of **W** and rows of **H**, and a group of similar values of much higher magnitude corresponding to irrelevant components. Furthermore, for components which had become zero or close to zero, the term λ_{kt_s} is set equal to $\frac{1}{\epsilon}$ where ϵ =10^{-9}. Thus the pruning parameter can be determined by the following:

$$\bar{\lambda}_k \triangleq \frac{1}{T_s}\sum_{t_s=1}^{T_s}\lambda_{kt_s} > 10^9 \cdot \sqrt{\frac{\sum_{i=1}^{I}\sum_{f=1}^{F}\sum_{t_s=1}^{T_s}|Y_i(f,t_s)|^2}{IFT_s}} \tag{3.43}$$

Eq. (3.43) is a threshold defining which k^{th} row of **H** (equivalently k^{th} column of **W**) is to be removed. This allows us to estimate the effective number of component. If the prior assumptions are slightly violated or even if the likelihood function differs from the model assumption, the correct factorization rank can be determined by either evaluating the above bound by the pruning parameter.

3.2.2.3 Estimation of sound sources

For every algorithm of the proposed method, given the estimates of $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{P}$ that yield the smallest cost value, the reconstruction is executed in the time domain component by using the Wiener filtering [56]. The Short Time Fourier Transform (STFT) estimate of a $\hat{C}_{rk}(f, t_s)$ of the component k in channel i is reconstructed through

$$\begin{aligned}\hat{C}_{rk}(f, t_s) &\overset{\text{def}}{=} E\{C_{rk}(f, t_s)|\mathbf{P}, \mathbf{W}, \mathbf{H}, \mathbf{Y}\} \\ &= \frac{q_{ik} w_{fk} h_{kt_s}}{\sum_{k=1}^{K} q_{ik} w_{fk} h_{kt_s}} Y_i(f, t_s) \\ &= \frac{q_{ik} w_{fk} h_{kt_s}}{\hat{V}_i(f, t_s)} Y_i(f, t_s) \qquad (3.44)\end{aligned}$$

where $\hat{V}_i(f, t_s) = \sum_{k=1}^{K} q_{ik} w_{fk} h_{kt_s}$. The decomposition is conservative in the sense that it satisfies

$$y_i(t_s) = \sum_{k=1}^{K} \hat{c}_{rk}(t_s) \qquad (3.45)$$

The estimated sources are reconstructed by using inverse-STFT of $\hat{C}_{rk}(f, t_s)$ for all r and k leads to a set of time-domain components $\{\hat{c}_k(t), \ldots, \hat{c}_K(t)\}$, with

$$\hat{c}_k(t) = \begin{bmatrix} \hat{c}_{rk}(t) \\ \vdots \\ \hat{c}_{Rk}(t) \end{bmatrix} \qquad (3.46)$$

and sources estimates can be obtained as

$$\hat{x}_{rd}(t) = \sum_{k \in K_d} \hat{c}_{rk}(t) \qquad (3.47)$$

The proposed algorithm is summarized in Table 3.1.

Table 3.1: Overview proposed algorithm of APBNTF.

1) Initialize $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{P}$ with nonnegative random values.

2) Define $\mathbf{Q} = \mathbf{PL}$.

3) Compute $Y_i(f, t_s)$ using STFT on the audio mixture, from power spectrogram $V_i(f, t_s) = |Y_i(f, t_s)|^2$, and compute $\hat{V}_i(f, t_s) = \sum_{k=1}^{K} q_{ik} w_{fk} h_{kt_s}$.

4) Compute $\boldsymbol{G}_+$ and $\boldsymbol{G}_-$ according to (3.35). Update model parameters $\mathbf{P}, \mathbf{W}$ and $\mathbf{H}$ as follows:

$$\mathbf{P}_{id} \leftarrow \mathbf{P}_{id} \bullet \frac{\sum_{t_s k} \boldsymbol{G}_{-ift_s} (\mathbf{W} \circ \mathbf{H})_{ft_s k} \boldsymbol{L}_{id}^{\boldsymbol{T}}}{\sum_{t_s k} \boldsymbol{G}_{+ift_s} (\mathbf{W} \circ \mathbf{H})_{ft_s k} \boldsymbol{L}_{id}^{\boldsymbol{T}}}$$

$$\mathbf{W}_{fk} \leftarrow \mathbf{W}_{fk} \bullet \frac{\sum_{ik} \boldsymbol{G}_{-ift_s} (\mathbf{Q} \circ \mathbf{H})_{it_s k}}{\sum_{ik} \boldsymbol{G}_{+ift_s} (\mathbf{Q} \circ \mathbf{H})_{it_s k} + \boldsymbol{W}_{fk} \boldsymbol{\Xi}_{kk}^{\boldsymbol{T}}}$$

$$\mathbf{H}_{t_s k} \leftarrow \mathbf{H}_{t_s k} \bullet \frac{\sum_{fk} \boldsymbol{G}_{-ift_s} (\mathbf{Q} \circ \mathbf{W})_{ifk}}{\sum_{fk} \boldsymbol{G}_{+ift_s} (\mathbf{Q} \circ \mathbf{W})_{ifk} + \boldsymbol{\lambda}_{t_s k} \mathbf{1}_{kk}^{\boldsymbol{T}}}$$

5) Update $\boldsymbol{\lambda} = \frac{\mathbf{1}}{\mathbf{H}}$, $\bar{\lambda}_k = \frac{1}{T_s} \boldsymbol{\delta}_k^{\boldsymbol{T}} \boldsymbol{\lambda}\, \mathbf{1}$

6) Prune the irrelevant components of $\mathbf{W}$ and $\mathbf{H}$ using the criteria

$$\bar{\lambda}_k \triangleq \frac{1}{T_s} \sum_{t_s=1}^{T_s} \lambda_{kt_s} > 10^9 \cdot \sqrt{\frac{\sum_{i=1}^{I} \sum_{f=1}^{F} \sum_{t_s=1}^{T_s} |Y_i(f, t_s)|^2}{IFT_s}}.$$

7) Repeat steps 3 to 6 until it converges or reaches the predefine number of iteration.

8) Reconstruct components and sources using

$$\hat{x}_{rd}(t) = \sum_{k \in K_d} \hat{c}_{rk}(t)$$

where $\hat{c}_{rk}(t) = STFT^{-1}\{\hat{C}_{rk}(f, t_s)\}$ and $\hat{C}_{rk}(f, t_s) = \frac{q_{ik} w_{fk} h_{kt_s}}{\hat{V}_i(f, t_s)} Y_i(f, t_s)$.

3.3 Results and Analysis

3.3.1 Toy Examples

In this section, the proposed APBNTF method will be tested in their ability to extract the basis and code from a simulated mixed data. The simulated data is generated to have a high degree of pattern overlap. To investigate the effects of the pruning technique and prior on $\mathbf{W}, \mu_{kj}$ on the performance of feature extraction, the following three experiments have been developed.

1) Without pruning and without prior on $\mathbf{W}$, i.e., μ_{kj} = 0.

2) With pruning and without prior on $\mathbf{W}$.

3) With pruning and with prior on $\mathbf{W}$, $\mu_{kj} = 0.50$.

Figure 3.3 shows the real basis (i.e., vertical panels) and code (i.e., horizontal panels) of the simulated mixed pattern. The basis **W** consists of one circle and one cross features. These features are convolved with the code **H** given at the top panels to yield the data matrix **Y** which is a mixture of both patterns.

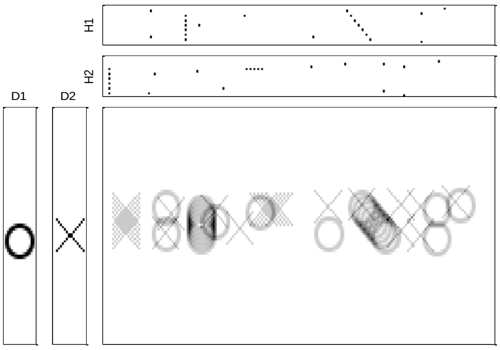


Figure 3.3: Real basis and code of the simulated mixed data.

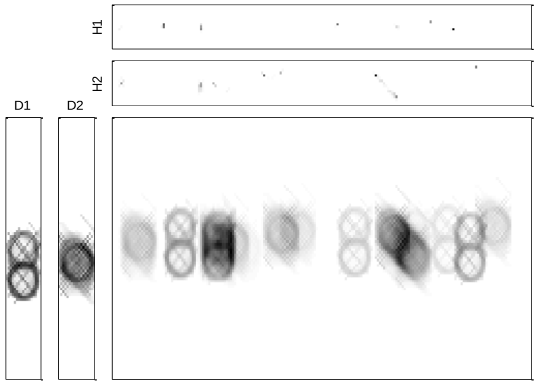


Figure 3.4: Estimated results based on the proposed method without pruning and without prior on $\mathbf{W}$.

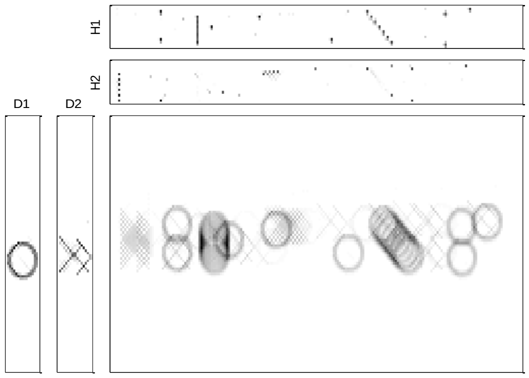


Figure 3.5: Estimated results based on the proposed method with pruning and without prior on $\mathbf{W}$.

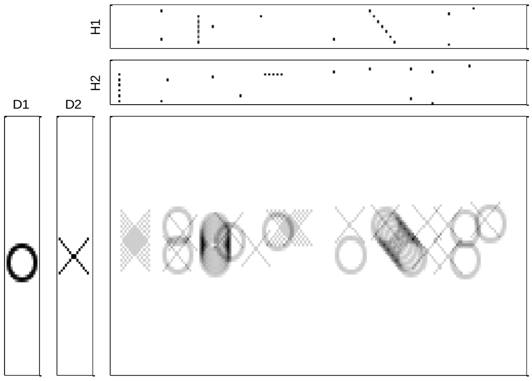


Figure 3.6: Estimated results based on adaptive sparsity factorization and with prior on $\mathbf{W}$, i.e., μ_{kj}= 0.50.

Figures 3.3 – 3.6 show the matrix factorization results corresponding to each of the above experiments. It is seen that the proposed method without pruning and without prior on $\mathbf{W}$ has failed to identify the correct basis and code. The major reason stems from the high degree of pattern overlap between the circle and the cross features in the mixed dataset. Since the sparseness is uncontrolled, the larger parts of the pattern overlap will cause more errors in estimating the basis while the code tends to be more ambiguous. This decreases the possibility of correct assignment of the basis to each feature, and subsequently results in poorer extraction and reconstruction performance as shown in Figures 3.4 and 3.5. For example, one could see the extracted codes (i.e., upper panels of Figure 3.4) are almost identical and thereby cause parts of the circle and the cross features missing from the figure. On the other hand, Figure 3.5 shows a better extraction result by using only the pruning (without prior on $\mathbf{W}$), while Figure 3.6 shows the best result when both regularizations (i.e., pruning and modified Gaussian prior) are used. However, if only the pruning is adopted, this may yield a sub-optimal performance, which is evident from Figure 3.5, where the cross feature has not been fully extracted. Nonetheless, the performance of feature extraction also depends on the correlation between the real bases of the mixed pattern. Visual inspection of Figure 3.3 shows that the real basis shares some degree of commonality and therefore induces correlation. Thanks to the modified Gaussian prior, this correlation is explicitly modeled by μ_{kj} in the proposed method. This enables the estimated basis vectors $\mathbf{W}_1$ and $\mathbf{W}_2$ to take advantage of the correlation in learning the real basis directly from the mixed pattern. This explains the reason as to why Figure 3.6 shows better performance than Figure 3.5. Therefore, the analysis results have unanimously indicated the importance of selecting the correct sparseness λ_{kt_s} for each element code and of incorporating the correlation μ_{kj} between the different basis vectors in order to arrive at the optimal performance of feature extraction. In the next section, the proposed method will be further tested on real application of multi-channel BSS. A series of performance comparison with other matrix factorization methods will also be presented.

3.3.2 Real Application

The performance of proposed method is demonstrated by separating music sources. Several experimental simulations under different conditions have been designed to investigate the efficacy of the proposed method. All experiments are conducted using a PC with Intel® Core™ i5 CPU 650 at 3.2 GHz and 4 GB RAM. MATLAB is used as the programming platform. The TF representation is computed by using the STFT of 1024-point Hanning window with 50% overlap. the proposed method has been evaluated and compared with IS-cNTF and IS-NTF method [56] where 3 linear instantaneous stereo mixtures of 3 sources taken from the Signal Separation Evaluation Campaign (SiSEC 2010) “Underdetermined speech and music mixtures” task development dataset [76]. Three types of mixture have been considered and are described as: 1) wdrums, a linear instantaneous stereo mixture (with positive mixing coefficients) of 2 drum sources and 1 bass line. 2) nodrums, a linear instantaneous stereo mixture (with positive mixing coefficients) of 1 rhythmic acoustic guitar, 1 electric lead guitar and 1 bass line. Both datasets correspond to the test data for the 2007 Stereo Audio Source Separation Evaluation Campaign

(SASSEC'07) [77]. It also coincides with development dataset dev2 of SiSEC'08 "underdetermined speech and music mixtures" task. All mixtures are 10 seconds-long and sampled at 16 kHz. The instantaneous mixing is characterized by static positive gains. The STFT has been applied with sine bell of length 64 ms (1024 samples) leading to $F = 513.$ 3) Shannonsongs Sunrise, a linear instantaneous stereo mixture of $d_{max} = 3$ musical sources (drums, lead vocals and piano) created using 17 seconds-excerpts of original separated tracks from the song "Sunrise" by S. Hurley, available under a Creative Commons License at [78] and downsampled to 16 kHz. The mixing parameters (instantaneous mixing matrix) were taken from the 2008 Signal Separation Evaluation Campaign (SiSEC'08) "underdetermined speech and music mixtures" task development datasets [77].

All experiments have used the linear instantaneous stereo mixture of wdrums, nodrums and Shannonsongs Sunrise datasets and set the number of components per each source to $K_d = 20$ with d_{max} = 3 sources. An initial NTF decomposition is computed from the power spectrograms using the Kullback-Leibler (KL) divergence [19]. The number of iterations was set to 50 for initialization parameters, and 400 for updating the parameters. The separation performance has been evaluated in terms of the Signal-to-Distortion Ratio (SDR) which is a global measure unifying the Source-to-Interference Ratio (SIR) and the Sources-to-Artifacts Ratio (SAR) criteria expressed in decibels (dB), defined as

$$\text{SDR} = \left(\frac{\|s_{target}\|^2}{\|e_{interf}+e_{artif}\|^2}\right), \text{SIR} = \left(\frac{\|s_{target}\|^2}{\|e_{interf}\|^2}\right) \text{ and } \text{SAR} = \left(\frac{\|s_{target}+e_{interf}\|^2}{\|e_{artif}\|^2}\right) \tag{3.48}$$

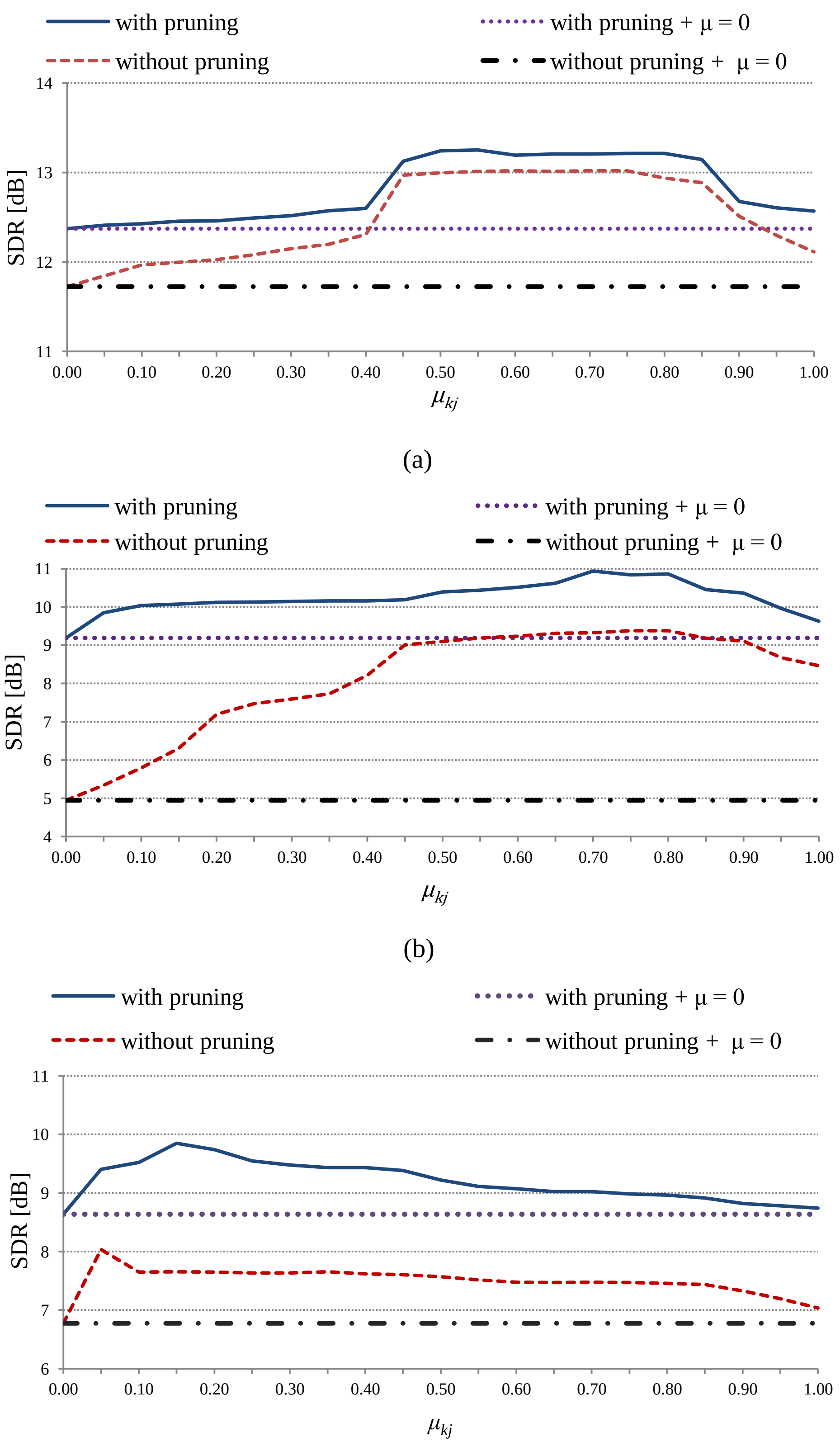
where s_{target} is the actual source estimate, e_{interf} represent the interference from other sources and e_{artif} is the artifacts of the separation algorithm. MATLAB routines for computing these criteria are obtained from the SiSEC'08 webpage [77, 79].

3.3.3 Effects on audio mixtures separation with/without pruning

In this section, the performance of the proposed method has been investigated with and without the pruning technique and with and without the prior information on the basis $\mathbf{W}$ for separating audio mixtures. It is hypothesized that with the pruning, the audio source separation will be significantly enhanced. Figures 3.7 (a)-(c) show the performance of the proposed algorithm with 1) pruning, 2) without pruning and 3) without pruning + without prior on $\mathbf{W}$ **(**e.g., $\mu_{kj} = 0$) for wdrums, nodrums and Shannonsongs Sunrise datasets respectively, under various μ_{kj} parameters.

Using $\mu_{kj} = \frac{\gamma_{kj}}{\sigma_k \sigma_j}$ leads to very close between the optimal performances and that estimated using above. This is verified by Figure 3.7. As the basis $\mathbf{W}$ also depended on whether pruning is activated or not, the estimate of the optimal prior on $\mathbf{W}$ (which inherently depends on the basis) will likewise change depending on whether pruning is activated or not.

The wdrums, nodrums and Shannonsongs Sunrise datasets have been used for the above cases.



(a)

(b)

(c)

Figure 3.7: Separation results of the proposed method with pruning, without pruning and without pruning + without prior on $\mathbf{W}$ e.g., $\mu_{kj} = 0$: (a) wdrums mixture, (b) nodrums mixtures, (c) Shannonsongs Sunrise mixtures.

Figures 3.7(a)-(c) illustrates that with the pruning yields better separation performance than without the pruning for all mixtures across all μ_{kj}. The reason is that in the case of without the pruning, the term K_d is constant at 20 which may be either too small or too large for each source. The constant K_d leads to the accumulation of unsuitable dictionary of the source which resulted in poor re-constructed source. In the case of with the pruning, the number of K_d is approximately determined for each source. The proper K_d corresponding to each source will avoid under- or over-fit. If setting K_d too low, K_d is selected to be under-fit. This means that the model is too simplistic and the data does not fit the model well. In this case, it can result in a factorization where multiple basis from different sources are approximated by a single factor, which in turn leads to incorrect separation. Conversely, if K_d too large, over-fit occurs when the model has too many parameters relative to the number of the mixtures which will causes excessive computational complexity. In this case, if any of the input points are varied slightly, it could result in an extremely different model and this can cause problems at the grouping stage. Hence, the over-fit model will generally have poorer predictive performance.

The separation results in terms of the SDR are given in Figures 3.7(a)-(c). According to the figures, the proposed method with pruning tends to yield the best overall performance, where the average improvement over the without pruning in three cases can be summarized as follows: 1) for wdrums mixture, the average SDR improvement is 0.89 dB per source; 2) for nodrums mixture, the average SDR improvement is 1.49dB per source; and 3) for mixtures of music and vocal (Shannonsongs Sunrise), the improvement is 1.28dB per source. The results have also clearly indicated that there are certain values of μ_{kj} where the algorithm performs the best. In the case of wdrums mixture, the best performance is obtained when μ_{kj} ranges from 0.41 to 0.93 (within 5% from highest SDR) with the highest SDR of 13.25dB. As for nodrums mixture, the best performance is obtained when μ_{kj} ranges from 0.52 to 0.87 with the highest SDR of 10.94dB and in the case of music and vocal mixture, the best performance is obtained when μ_{kj} ranges from 0.04 to 0.42 with the highest SDR of 9.85dB. Of the above findings, we can conclude that for music mixtures, the best performance is obtained when μ_{kj} ranges from 0.41 to 0.93 and in the case of music and vocal mixture, the best performance is obtained when μ_{kj} ranges from 0.04 to 0.42. On the contrary, it is noted that when μ_{kj} is set either too low or high, the separation performance tends to degrade. It is also worth pointing out that the separation results are rather coarse when the factorization is nonregularized (i.e., without prior on $\mathbf{W}$) and without pruning. Here, the average SDR of the proposed method without prior on $\mathbf{W}$ and without pruning is the lowest among the three methods across $\mu_{kj} > 0$. As evidence, Figure 3.7 shows the SDR of the without prior on $\mathbf{W}$ and without pruning method as for wdrums mixture: 11.73dB per source, for nodrums: 4.95dB per source, and for Shannonsongs Sunrise: 6.78dB per source. It can be summarized that the average improvement of the proposed method (with prior on $\mathbf{W}$ and with pruning) against the case of without pruning and without prior on $\mathbf{W}$: 1) For wdrums mixture, the improvement per source in terms of the SDR is 1.88dB. 2) For nodrums mixture, the improvement per source in terms of the SDR is 5.27dB. 3) For Shannonsongs Sunrise mixture, the improvement per source in terms of the SDR is 2.72dB.

Figure 3.8 shows the original hi-hat, drum and bass music and its separation results.

The mean square errors (MSE) between the original and estimated sub-sources are 0.04, 0.81 and 0.01 for hi-hat, drum and bass music, respectively. On dataset wdrums, the highest separation result in terms of SDR is obtained with the proposed method. The average SDR are 12.79dB, 10.04dB and 9.92dB for wdrums, nodrums and Shannonsongs Sunrise, respectively. Additionally, it is found that hi-hat and bass give high separation performance with SDR of 14.33dB and 20.01dB, respectively.

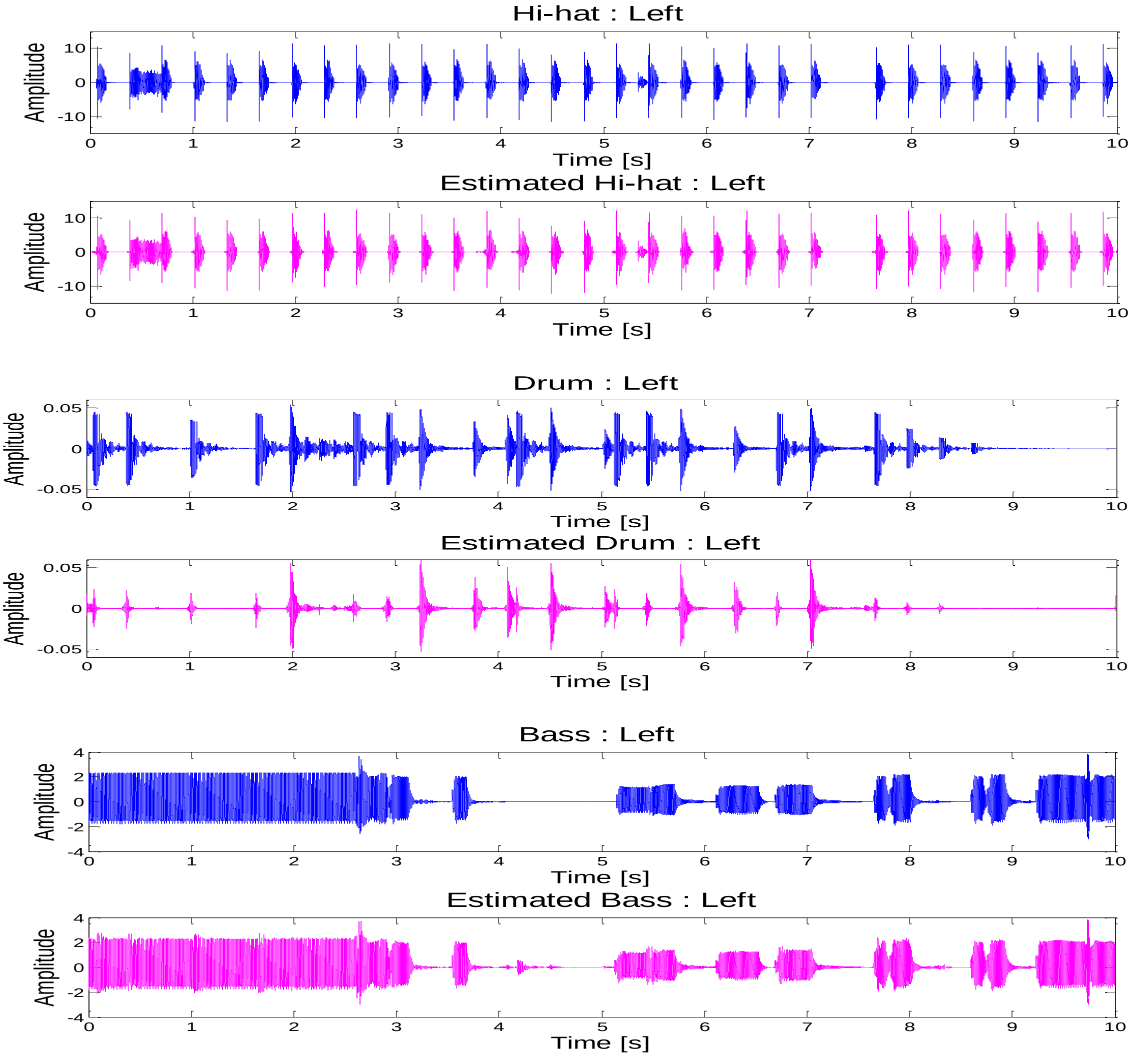
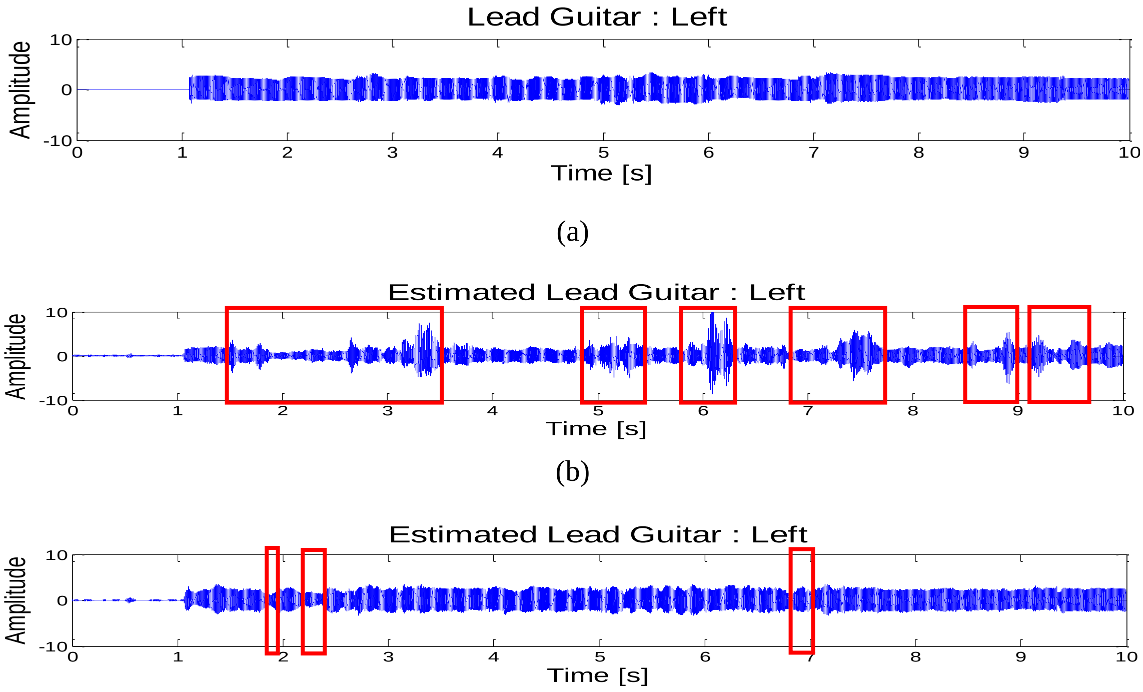


Figure 3.8: Original sources and the estimated sources from left microphone using the proposed method with $\mu_{kj} = 0.50$.

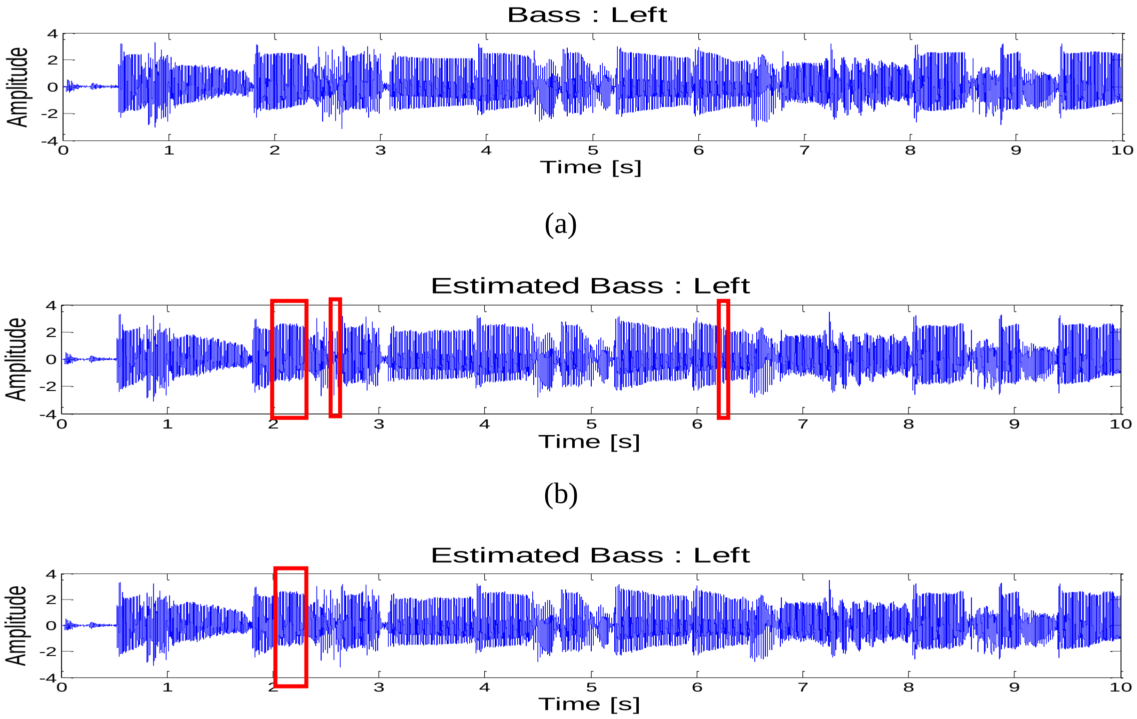


(a)

(b)

(c)

Figure 3.9: Time-domain representation of (a) the original source (Lead Guitar.) of nodrums mixture, (b) and the estimated Lead Guitar from left microphone using the proposed method without pruning and (c) with pruning.



(a)

(b)

(c)

Figure 3.10: Time-domain representation of (a) the original sources (Bass) of nodrums mixture ,(b) the estimated Bass signal from left microphone using the proposed method without pruning and (c) with pruning.

Figures 3.9 and 3.10 show the separated results of the proposed algorithm with and without pruning for lead guitar and bass, respectively. In both figures, panel (a) shows the original sub-sources, panels (b) shows the estimated sub-sources by using the proposed method without pruning while panel (c) shows the estimated sub-sources by exploiting the hybrid pruning and the proposed method. The mean square error (MSE) between the original and estimated source of Lead Guitar from nodrums separation is 0.18 and 0.72 for the proposed method with pruning and the proposed method without pruning, respectively. Also the MSE between the original and estimated source of bass from nodrums separation is 0.01 and 0.04 for the proposed method with pruning and the proposed method without pruning, respectively.

3.3.4 Comparison with Other NTF-Based Multichannel Source Separation Methods

The separation performance of the proposed method has been evaluated by comparing with the NTF-based unsupervised multichannel audio source separation methods i.e. IS-cNTF [9] and IS-NTF [56] method. The numbers of latent components for the three methods have been set for three mixtures as the following: 1) K_d = 20 and K_d = 3 for the proposed method, 2) K_d = 20 and K_d = 3 for IS-cNTF, and 3) K = 60 and K = 9 for IS-NTF .

Each of the six algorithms was run 20 times from 50 random initializations for 400 iterations. The separation performance of the proposed method has also been presented in this section. The separation performance is calculated from the average of 10 experiments under the same mixture. The IS-cNTF and IS-NTF parameters are set as follows [9, 56]:

numbers of components per each source are 3 and 20, respectively for all datasets. The TF domain used in IS-cNTF and IS-NTF are based on the log-frequency spectrogram. Cost function of IS-cNTF and IS-NTF are based on the Itakura-Saito divergence.

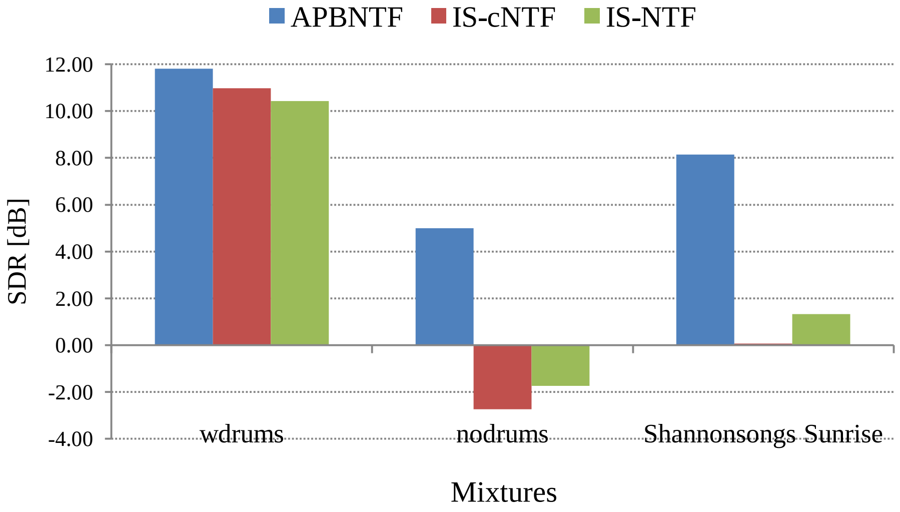


Figure 3.11: Comparison of average SDR performance on wdrums, nodrums, and Shannonsongs Sunrise of three audio sources between IS-NTF, IS-cNTF, and the proposed method (APBNTF) with $K_d = 3$ per source and $\mu_{kj} = 0.50\ (wdrums), 0.75\ (nodrums),\ 0.15\ (Shannonsongs\ Sunrise)$.

In Figure 3.11, K_d was set to 3 according to [56], shows negative results for nodrums and low separation performance for Shannonsongs Sunrise when using IS-cNTF and IS-NTF. The proposed method, on the other hand, has led to better separation performance as it takes the advantages of the more meaningful feature extraction that pertain to the data through prior information on basis vectors $\mathbf{W}$ and the unique sparsity that has been individually optimized for each element code with automatic pruning. However, it should be noted that setting $K_d = 3$ does not necessarily guarantee that this is the optimal number of components associated with each source. To further investigate this, the initial K_d has been increased to 20 and allow the pruning technique to determine the appropriate number of components. The final results in term of SDR are shown in

Figure 3.12.

From Figure 3.12, when the initial K_d is set to 20, it is seen that the proposed method yields superior separation performance among all three methods. K_d is determined based on pruning which is optimally selected for each source i.e. for wdrums: $\{K_{Hi-hat} = 13$, $K_{drums} = 11,\ K_{bass} = 12\}$, for nodrums: $\{K_{bass} = 12, K_{lead\ G.} = 16,\ K_{rhmthmix\ G.} = 19\}$, and for Shannonsongs Sunrise: $\{K_{drum} = 17,\ K_{vocal} = 18,\ K_{pano} = 15\}$. The average SDR improvement of the proposed method with K_d=20 over IS-cNTF and IS-NTF method are 11.8dB, 10.1dB, and 7.8dB per source for wdrums, nodrums, and Shannonsongs Sunrise, respectively. Hence, setting $K_d = 3$ is not adequate for modeling the components of the sources. To sum up, combining prior information on **W** and the pruning technique benefit the proposed method with better separation performance than using either only one of them.

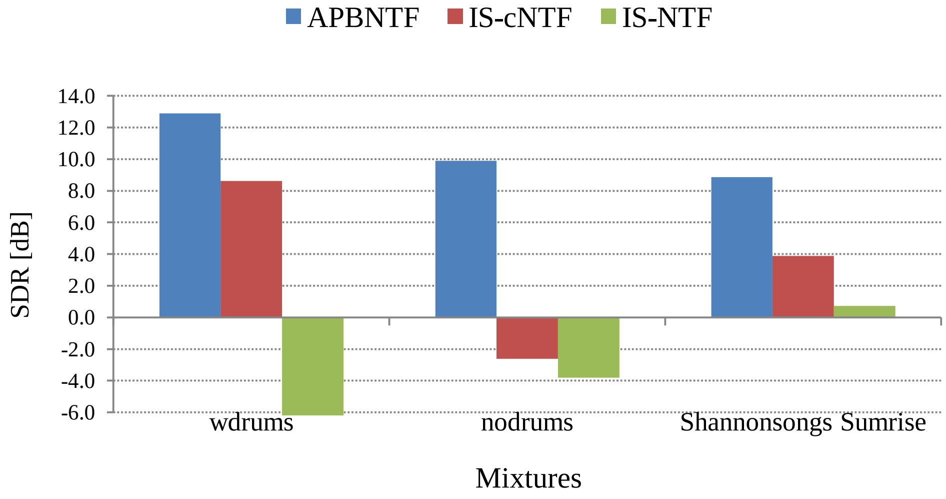


Figure 3.12: Comparison of average SDR performance on wdrums, nodrums, and Shannonsongs Sunrise of three audio sources between IS-NTF, IS-cNTF, and the proposed method (APBNTF) with $K_d = 20$ per source and $\mu_{kj} = 0.50(wdrums), 0.75(nodrums), 0.15(Shannonsongs\ Sunrise)$.

Figure 3.13 shows the Comparison of average SDR performance of estimated Hihat, Drum and Bass from the wdrums dataset. It is clearly shown that, the proposed method can separate the sources from the wdrums mixture more efficiently than the two methods. The average SDR improvements of the proposed method over the IS-NTF and IS-cNTF methods are 2.5dB per source and 1.9dB per source, respectively.

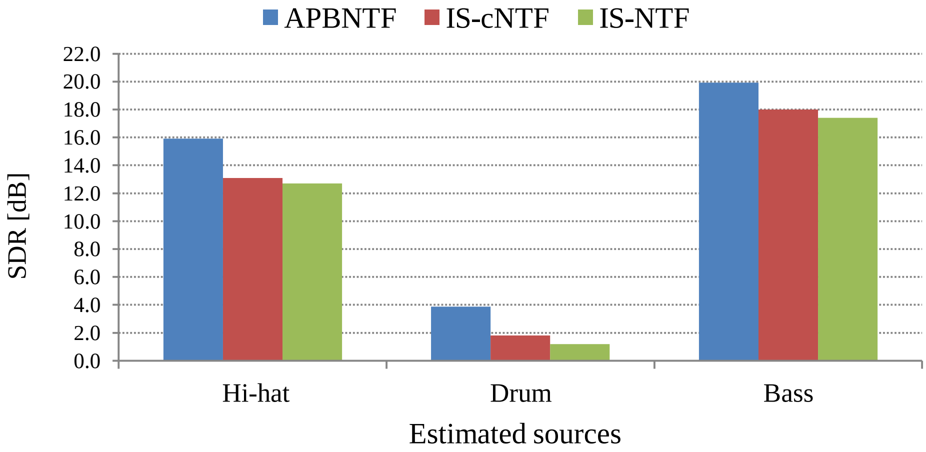
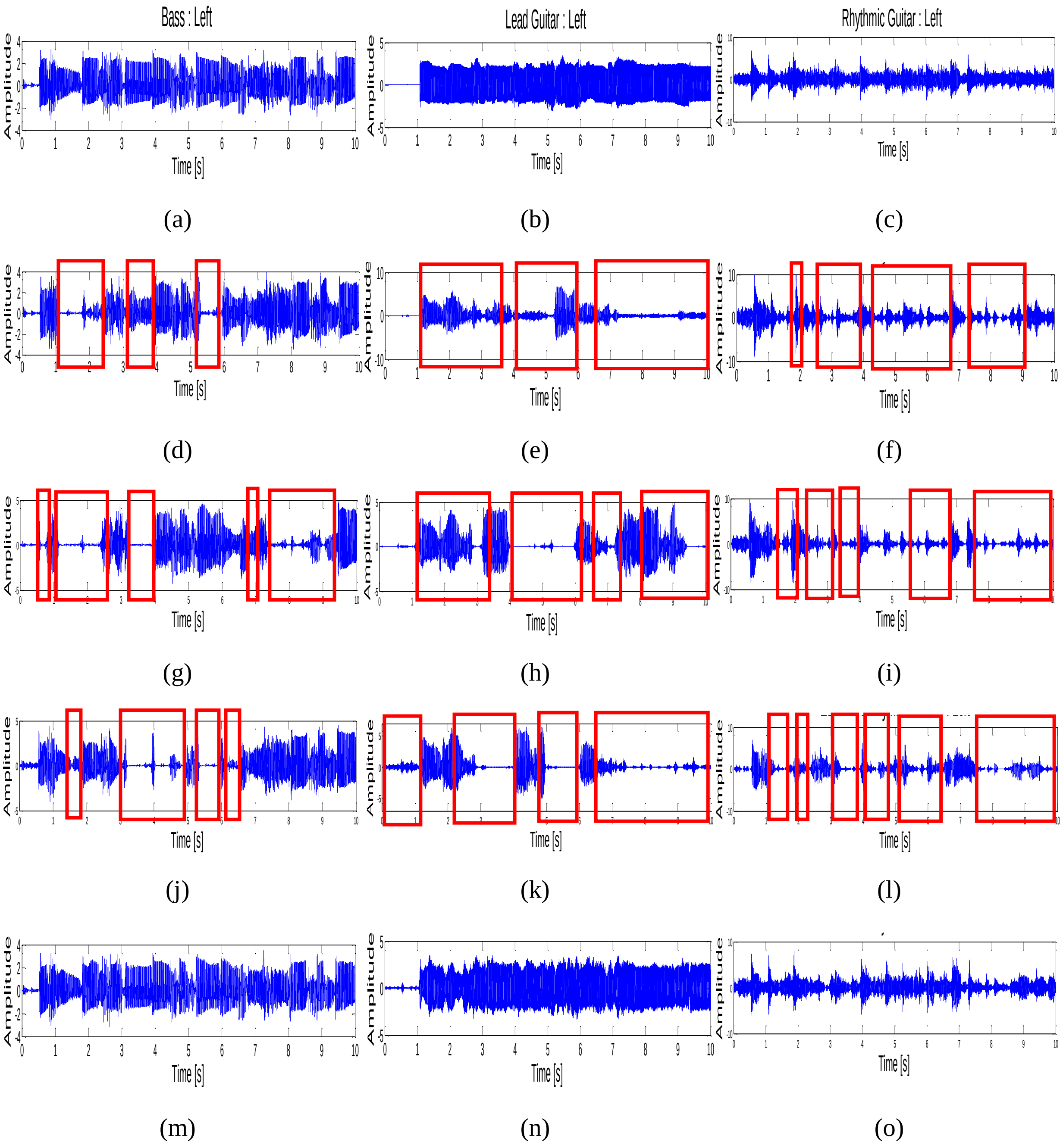


Figure 3.13: Comparison of average SDR performance of estimated Hi-hat, Drum and Bass from wdrums dataset between IS-NTF, IS-cNTF, and proposed method (APBNTF).

In Figure 3.14, panels (a)-(c) show the original sources of the nodrums mixture which are the bass, lead guitar and rhythmic guitar, respectively. Panels (d)-(o) shows the estimated sources using the IS-NTF, IS-cNTF and proposed method. Analyzing panels (d)-(l), it visibly clear that that source separation using latent components methods require a joint optimization of the sparsity regularization for each element code h_{kt_s}, number of latent components per source, and data-adapted correlated basis vectors in $\mathbf{W}$. Panels (m)-(o) show the case of the separations results of the proposed method when joint process is optimized while panels (d)-(l) show that the other NTF-based methods did not fully separate the music mixture. Many spectral and temporal components are missing from the recovered sources and these have been highlighted (marked red box) in all panels. The other NTF-based methods fail to take into account of the data-correlated basis vectors and specific sparsity associated with each code, and this has resulted in ambiguous estimation of each source spectrum and thereby discarding the temporal information. When the temporal structure and the pitch change are not properly estimated in the model, the mixing ambiguity is still contained in each separated source.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

(m) (n) (o)

Figure 3.14 Separated signals of nodrums in time-domain. (a)-(c): original bass, Lead Guitar and Rhythmic Guitar music. (d)- (e): estimated sources using the proposed method (initial K_d=3 and $\mu_{kj} = 0.50$ for all sources). (g)- (i): estimated sources using IS-cNTF. (j)- (l): estimated sources using IS-NTF. (m)- (o): estimated sources using the proposed method (initial K_d=20 and $\mu_{kj} = 0.54, 0.75, 0.72$ for bass, lead guitar and rhythmic guitar, respectively).

Table 3.2: Performance comparison between other NTF based multichannel audio source separation methods and the proposed method.

Mixtures	Methods	SDR (dB)	SIR (dB)	SAR (dB)
wdrums (Hi-hat/drums/bass)	Proposed method (APBNTF)	**12.8**	**38.1**	**12.8**
	Proposed method (APBNTF) without pruning	11.8	37.2	11.8
	IS-cNTF	10.9	18.1	11.3
	IS-NTF	10.4	17.8	10.1
nodrums (bass/lead G /rhythmic G)	Proposed method (APBNTF)	**10.0**	**34.4**	**10.1**
	Proposed method (APBNTF) without pruning	4.9	31.5	4.9
	IS-cNTF	3.7	27.9	3.9
	IS-NTF	-1.7	-0.3	3.5
Shannonsongs Sunrise (drum/vocal/piano)	Proposed method (APBNTF)	**10.1**	**31.8**	**10.1**
	Proposed method without pruning	8.2	29.9	8.2
	IS-cNTF	0.1	19.4	0.2
	IS-NTF	1.3	20.7	1.4

Table 3.2 further gives the SDR, SIR, and SAR comparison results between our proposed method and the other NTF methods. The improvement of our method compared with the proposed method without pruning, IS-cNTF and IS-NTF can be summarized as follows: 1) for the mixture of wdrums, the average improvement per source in terms of the SDR is 1.0dB, 1.9dB, and 2.4dB per source, respectively; 2) for the mixture of

nodrums, the average improvement per source in terms of SDR is 5.1dB, 6.3dB, and 11.7dB per source, respectively; 3) for the Shannonsongs Sunrise mixture, the average improvement per source in terms of SDR is 1.9dB, 10.0dB, and 8.8dB per source, respectively. In a nutshell, the proposed method gives an average performance improvement of at least twice better than the state-of-art IS-NTF and IS-cNTF methods, respectively. Analyzing the separation results, the proposed method leads to the best separation performance for all recovered sources. The IS-cNTF method performs with poorer results whereas the separation performance by the proposed method without pruning is slightly better than the IS-NTF and IS-cNTF methods. Our proposed method gives significantly better performance than the proposed method without pruning, IS-cNTF and IS-NTF methods. The spectral dictionary obtained via the proposed method without pruning, IS-cNTF and IS-NTF methods are not adequate to capture the temporal dependency of the frequency patterns within the audio signal.

3.3.5 Determination of Optimal K_d and μ_{kj} of the Proposed Method

In this section, the proposed method which described in the previous section has been applied to determine the optimal K_d and μ_{kj} of each source for separating mixtures of wdrums, nodrums and Shannonsongs Sunrise. The proposed method has been tested by using three mixtures with various μ_{kj} values from 0 to 1 with every increment of 0.05 and retained the value of K_d and μ_{kj} associated with the best SDR for each estimated source. The results are tabulated in Table 3.3.

Table 3.3 shows that the pruned numbers of K_d and μ_{kj} values for each source are different. Analyzing the results, the proposed method with the pruned numbers of K_d and μ_{kj} are compared to the method with fixed $\mu_{kj} = 0.1$. The average SDR improvement of the proposed method with pruned K_d and μ_{kj} over the fixed μ_{kj} method are as follow: 1) for wdrums mixtures, the average SDR improvement is 0.57dB per source; and 2) for nodrums mixtures, the average SDR improvement is 1.39dB per source; and 3) for Shannonsongs Sunrise mixtures, the average SDR improvement is 0.97dB per source. The results have also clearly indicated that the proposed method yields superb separation performance when the value of μ_{kj} for each source is properly selected. Moreover, it can be seen that the optimal values of μ_{kj} for each source consistent with the inference in Section 3.3.2 i.e. in the case of music mixture, the best performance is obtained when μ_{kj} ranges from 0.41 – 0.93 and for music and vocal mixtures, the best performance is obtained when μ_{kj} ranges from 0.04 – 0.42. This should be made feasible using automatic selection for the appropriate value of μ_{kj} for each source in the future work.

Table 3.3: Optimal number of K_d and μ_{kj} of wdrum, nodrum and Shannonsongs Sunrise.

Mixtures	SDR (dB)				No. Comp. (K_d)			μ_{kj}		
	x_1	x_2	x_3	Avg	x_1	x_2	x_3	x_1	x_2	x_3
wdrums (Hi-hat/ drum/ bass)	15.9	4.0	20.1	**13.4**	13	11	12	0.8	0.4	0.6
nodrums (bass/ lead G./ rhythmic G.)	19.0	9.9	5.4	**11.4**	12	16	19	0.5	0.8	0.7
Shannonsongs Sunrise (drum/ vocal/ piano)	15.2	5.8	10.4	**10.4**	17	18	15	0.4	0.2	0.7

3.4 Summary

This chapter has presented a new framework of amalgamating pruning and Bayesian regularized cluster NTF under a PARAFAC structure with Itakura-Saito divergence for multichannel audio source separation. The impetus behind the proposed work is that sparseness achieved by the conventional NTF is not efficient enough; in source separation, it is very necessary to yield control over the degree of sparseness explicitly for each element code $\{h_{ij}\}$. In addition, it does not incorporate correlation information between different basis vectors into the factorization process. Underlying all factorization algorithms is the principal difficulty in estimating the adequate number of latent components for each source. The proposed method addresses this issue by using the principle of pruning. The proposed method offers at least four advantages: First, the sparse regularization term is adaptively tuned using a hierarcs hical Bayesian approach. This yields the desired sparse decomposition, thus the proposed method is enable more efficient estimation of the spectral dictionary and temporal codes of nonstationary audio signals. Second, the modified Gaussian prior is formulated to express the correlation between different basis vectors. Third, the proposed algorithm can automatically detect the optimal number of latent components of the individual source. Finally, it avoids the strong constraints of separating blind source without training knowledge. Hence, the work is a step forward to realizing optimal BASS. This has been verified concretely based on experiments which is very promising results. In addition, the separation performance of the proposed method yields significant improvement of SDR on multichannel audio separation compared with other NTF-based source separation methods.

CHAPTER 4

SINGLE-CHANNEL AUDIO SEPARATION USING VARIATIONAL $\boldsymbol{L_1}$ −SPARSE COMPLEX MATRIX FACTORIZATION

In Chapter 4, an extreme case of blind source separation was regarded when a sole recording is available. A novel single-channel blind source separation (SCBSS) has been developed to extract better quality of audio separated signals. This approach will exploit the variational L_1-sparse complex matrix factorization (vL_1-SCMF) which offers the advantages of the CMF and a variational L_1-sparse approach simultaneously. CMF is based on a mixing model defined in the complex-spectrum domain and estimates recurring patterns in the observed magnitude spectra, their activations and their phases. The proposed factorization decomposes an information-bearing matrix into complex factor matrices that represent the spectral dictionary and temporal codes. A variational Bayesian approach was derived for computing the sparsity parameters for optimizing the matrix factorization. The method is demonstrated on separating audio mixtures recorded from a single channel that it yields superior performance compared with other existing sparse factorization methods. The performance of the developed algorithms will be measured using real-time audio signals in terms of the signal-to-distortion ratio.

This chapter is organized as follows: The formulation of the proposed Variational L_1-sparse CMF are articulated in Section 4.1. Experimental source separation results and a series of performance comparison with other existing BSS methods are presented in Section 4.2. Finally, Section 4.3 concludes the work of this chapter.

4.1 The Proposed Method

4.1.1 Generative Model

In this section, the problem is now, given an observed complex spectrum, $\boldsymbol{Y}_{f,t} \in \mathbb{C}$, to estimate the optimal parameters $\theta = \{\mathbf{W}, \mathbf{H}, \boldsymbol{\phi}\}$ of the model. We derive a new factorization method termed as the variational L_1-sparse complex non-negative matrix factorization (vL_1-SCMF). The generative model is given by

$$\boldsymbol{Y}_{f,t} = \sum_{k=1}^{K} \boldsymbol{W}_f^k \boldsymbol{H}_t^k \boldsymbol{Z}_{f,t}^k + \boldsymbol{\epsilon}_{f,t} \tag{4.1}$$

where $\boldsymbol{Z}_{f,t}^k = e^{j\phi_{f,t}^k}$ and the reconstruction error $\boldsymbol{\epsilon}_{f,t} \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$, is assumed to be independently and identically distributed (i.i.d.) as complex Gaussian distribution with white noise having mean 0 and variance σ^2 which used to denote a modeling error for each source.

The likelihood of $\theta = \{\mathbf{W}, \mathbf{H}, \boldsymbol{\phi}\}$ is thus written as

$$P(\mathbf{Y}|\theta) = \prod_{f,t} \frac{1}{\pi\sigma^2} exp(-\frac{|\boldsymbol{Y}_{f,t} - \boldsymbol{X}_{f,t}|^2}{\sigma^2}) \tag{4.2}$$

We assume that the prior distributions for $\mathbf{W}, \mathbf{H}$ and $\boldsymbol{\phi}$ are statistically independent, which yields

$$P(\theta|\boldsymbol{\lambda}) = P(\mathbf{W})P(\mathbf{H}|\boldsymbol{\lambda})P(\boldsymbol{\phi}) \tag{4.3}$$

$P(\mathbf{H}|\boldsymbol{\lambda})$ corresponds to the sparsity cost, for which a natural choice is a generalized Gaussian prior:

$$P(\mathbf{H}|\boldsymbol{\lambda}) = \prod_{k,t} \frac{p\,\lambda_t^k}{2\Gamma(1/p)} \exp\left(-\left(\lambda_t^k\right)^p |\mathbf{H}_t^k|^p\right) \tag{4.4}$$

where λ_t^k and p are the shape parameters of the distribution.

In this section, we assume that $P(\mathbf{H}|\boldsymbol{\lambda})$ promotes sparsity when $\mathrm{p} = 1$ and the norm of $\mathbf{H}$ is bounded. The posterior density defined as

$$P(\theta|\mathbf{Y}) \propto P(\mathbf{Y}|\theta)P(\mathbf{H}|\boldsymbol{\lambda}) \tag{4.5}$$

Maximum *a posteriori* (MAP) estimation problem leads to solve the following optimization problem:

$$minimize\ f(\theta) = \sum_{f,t}\left|\boldsymbol{Y}_{f,t} - \boldsymbol{X}_{f,t}\right|^2 + 2\lambda \sum_{k,t}\left|\boldsymbol{H}_t^k\right|^p \tag{4.6}$$

subject to $\sum_f \boldsymbol{W}_f^k = 1 \quad (k = 1, \ldots, K)$.

The CMF model parameters have been optimized by using an efficient iterative algorithm relied on an auxiliary function which its detail presents in [50]. Auxiliary function for $f(\theta)$ is given as follow: for any auxiliary variables with $\sum_k \overline{\boldsymbol{Y}}_{f,t}^k = \boldsymbol{Y}_{f,t}$, for any $\beta_{f,t}^k > 0$, $\sum_k \beta_{f,t}^k = 1$, for any $\boldsymbol{H}_t^k \in \Re$, $\overline{\boldsymbol{H}}_t^k \in \Re$ and $p = 1$. We can illustrate that $f(\theta) \leq f^+(\theta, \bar{\theta})$ with an auxiliary function defined as

$$f^{+}(\theta,\bar{\theta}) \equiv \sum_{f,k,t} \frac{\left|\bar{\boldsymbol{Y}}_{f,t}^{k} - \boldsymbol{W}_{f}^{k}\boldsymbol{H}_{t}^{k}.e^{j\phi_{f,t}^{k}}\right|^{2}}{\beta_{f,t}^{k}} + \lambda \sum_{k,t}\left(p\left|\bar{\boldsymbol{H}}_{t}^{k}\right|^{p-2}\boldsymbol{H}_{t}^{k^{2}} + (2-p)\left|\bar{\boldsymbol{H}}_{t}^{k}\right|^{p}\right) \quad (4.7)$$

and $\bar{\theta} = \left\{\bar{\boldsymbol{Y}}_{f,t}^{k}, \bar{\boldsymbol{H}}_{t}^{k} \mid 1 \le f \le F, 1 \le t \le T, 1 \le k \le K\right\}$, $f^{+}(\theta,\bar{\theta})$ is minimized w.r.t. $\bar{\theta}$ when

$$\bar{\boldsymbol{Y}}_{f,t}^{k} = \boldsymbol{W}_{f}^{k}\boldsymbol{H}_{t}^{k}.e^{j\phi_{f,t}^{k}} + \beta_{f,t}^{k}(\boldsymbol{Y}_{f,t} - \boldsymbol{X}_{f,t}) \quad (4.8)$$

$$\bar{\boldsymbol{H}}_{t}^{k} = \boldsymbol{H}_{t}^{k} \quad (4.9)$$

4.1.2 Formulation of the Proposed Variational L_1-sparse CMF

To facilitate such spectral dictionaries with adaptive sparse coding, we first define $\mathbf{W} = [\boldsymbol{W_1}, \boldsymbol{W_2} \ldots, \boldsymbol{W_F}]$, $\mathbf{H} = [\boldsymbol{H_1}, \boldsymbol{H_2} \ldots, \boldsymbol{H_T}]$, and $\boldsymbol{\lambda} = [\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2} \ldots, \boldsymbol{\lambda_T}]$. Hence, the negative log likelihood serves as the cost function in (4.6) defined as

$$L \propto \frac{1}{2\sigma^2}\left\|\boldsymbol{Y} - \sum_{k,f,t}\boldsymbol{W}_{f}^{k}\boldsymbol{H}_{t}^{k}e^{j\phi_{f,t}^{k}}\right\|_{F}^{2} + \lambda \sum_{k,t}\left|\boldsymbol{H}_{t}^{k}\right|^{p}$$

$$= \frac{1}{2\sigma^2}\left\|\boldsymbol{Y} - \sum_{k,f,t}\boldsymbol{W}_{f}^{k}\boldsymbol{H}_{t}^{k}e^{j\phi_{f,t}^{k}}\right\|_{F}^{2} - \sum_{k,t} log\, \lambda_{t}^{k} + \sum_{k,t}\lambda_{t}^{k}\boldsymbol{H}_{t}^{k} \quad (4.10)$$

when $p = 1$, we assume that $P(\mathbf{H})$ promotes sparsity so that

$$f(\mathbf{H}) = -\sum_{k,t} log\, \lambda_{t}^{k} + \sum_{k,t}\lambda_{t}^{k}\boldsymbol{H}_{t}^{k} \quad (4.11)$$

The sparsity term $f(\mathbf{H})$ forms the L_1-norm regularization which is used to resolve the ambiguity by forcing all structure in $\mathbf{H}$ onto $\mathbf{W}$. Therefore, the sparseness of the solution in (4.11) is highly dependent on the regularization parameter λ_t^k.

4.1.2.1 Estimation of the Dictionary and Temporal Code

In [50], the update rule for θ is derived by differentiating $f^+(\theta,\bar{\theta})$ partially w.r.t. $\boldsymbol{W}_f^k$ and $\boldsymbol{H}_t^k$, and setting them at zero then the updates for $\mathbf{W}$ and $\mathbf{H}$,respectively, become:

$$f^+(\theta,\bar{\theta}) \equiv \sum\nolimits_{f,k,t} \frac{\left|\bar{\boldsymbol{Y}}_{f,t}^k - \boldsymbol{W}_f^k \boldsymbol{H}_t^k . e^{j\phi_{f,t}^k}\right|^2}{\beta_{f,t}^k} + \lambda \sum\nolimits_{k,t} \left(p \left|\bar{\boldsymbol{H}}_t^k\right|^{p-2} {\boldsymbol{H}_t^k}^2 + (2-p)\left|\bar{\boldsymbol{H}}_t^k\right|^p \right)$$

$$\frac{\partial f^+(\theta,\bar{\theta})}{\partial \boldsymbol{W}_f^k} = \frac{\partial}{\partial \boldsymbol{W}_f^k} \sum\nolimits_{k,f,t} \frac{\left|\bar{Y}_{f,t}^k\right|^2 - 2\boldsymbol{W}_f^k \boldsymbol{H}_t^k Re\left[{\bar{\boldsymbol{Y}}_{f,t}^k}^* . e^{j\phi_{f,t}^k}\right] + {\boldsymbol{W}_f^k}^2 {\boldsymbol{H}_t^k}^2}{\beta_{f,t}^k}$$

$$= \sum\nolimits_t \frac{-2\boldsymbol{H}_t^k Re\left[{\bar{\boldsymbol{Y}}_{f,t}^k}^* . e^{j\phi_{f,t}^k}\right] + 2\boldsymbol{W}_f^k {\boldsymbol{H}_t^k}^2}{\beta_{f,t}^k}$$

$$= -2\sum\nolimits_t \frac{\boldsymbol{H}_t^k Re\left[{\bar{\boldsymbol{Y}}_{f,t}^k}^* . e^{j\phi_{f,t}^k}\right]}{\beta_{f,t}^k} + 2\boldsymbol{W}_f^k \sum\nolimits_t \frac{\boldsymbol{H}_t^k}{\beta_{f,t}^k}$$

Setting the above to zero leads to

$$\boldsymbol{W}_f^k \sum\nolimits_t \frac{\boldsymbol{H}_t^k}{\beta_{f,t}^k} = \sum\nolimits_t \frac{\boldsymbol{H}_t^k Re\left[{\bar{\boldsymbol{Y}}_{f,t}^k}^* . e^{j\phi_{f,t}^k}\right]}{\beta_{f,t}^k}$$

$$\boldsymbol{W}_f^k \;=\frac{\sum_t \frac{\boldsymbol{H}_t^k}{\beta_{f,t}^k} Re\left[\overline{\boldsymbol{Y}}_{f,t}^{k\;*} . e^{j\phi_{f,t}^k}\right]}{\sum_t \frac{\boldsymbol{H}_t^{k^2}}{\beta_{f,t}^k}} \tag{4.12}$$

Similarly,

$$\frac{\partial\, f^+(\theta,\,\overline{\theta})}{\partial\, \boldsymbol{H}_t^k} = \frac{\partial}{\partial\, \boldsymbol{H}_t^k}\left(\begin{array}{c}\displaystyle\sum_{k,f,t} \frac{\left|\overline{Y}_{f,t}^k\right|^2 - \,2\boldsymbol{W}_f^k \boldsymbol{H}_t^k Re\left[\overline{\boldsymbol{Y}}_{f,t}^{k\;*} . e^{j\phi_{f,t}^k}\right] + \boldsymbol{W}_f^{k^2} \boldsymbol{H}_t^{k^2}}{\beta_{f,t}^k} \\ \displaystyle + \lambda \sum_{k,t} \left(p\left|\overline{\boldsymbol{H}}_t^k\right|^{p-2} \boldsymbol{H}_t^{k^2} + (2-p)\left|\overline{\boldsymbol{H}}_t^k\right|^p\right)\end{array}\right)$$

$$= \sum_f \frac{-2\boldsymbol{W}_f^k Re\left[\overline{Y}_{f,t}^{k\;*} . e^{j\phi_{f,t}^k}\right] + 2\boldsymbol{W}_f^{k^2} \boldsymbol{H}_t^k}{\beta_{f,t}^k} + \;2\lambda\, p\left|\overline{\boldsymbol{H}}_t^k\right|^{p-2} \boldsymbol{H}_t^k$$

$$= -2\sum_f \frac{\boldsymbol{W}_f^k Re\left[\overline{Y}_{f,t}^{k\;*} . e^{j\phi_{f,t}^k}\right]}{\beta_{f,t}^k} + 2\boldsymbol{H}_t^k \sum_f \frac{\boldsymbol{W}_f^{k^2}}{\beta_{f,t}^k} + 2\lambda\, p\left|\overline{\boldsymbol{H}}_t^k\right|^{p-2} \boldsymbol{H}_t^k$$

Setting the above to zero yields

$$\boldsymbol{H}_f^k \left(\sum_f \frac{\boldsymbol{W}_f^{k^2}}{\beta_{f,t}^k} + \lambda\, p\left|\overline{\boldsymbol{H}}_t^k\right|^{p-2}\right) \;= \sum_f \frac{\boldsymbol{W}_f^k Re\left[\overline{Y}_{f,t}^{k\;*} . e^{j\phi_{f,t}^k}\right]}{\beta_{f,t}^k}$$

$$\boldsymbol{H}_f^k \;= \frac{\sum_f \frac{\boldsymbol{W}_f^k}{\beta_{f,t}^k} Re\left[\overline{\boldsymbol{Y}}_{f,t}^{k\;*} . e^{j\phi_{f,t}^k}\right]}{\sum_f \frac{\boldsymbol{W}_f^{k^2}}{\beta_{f,t}^k} + \lambda\, p\left|\overline{\boldsymbol{H}}_t^k\right|^{p-2}} \tag{4.13}$$

The update rule for the phase, $\phi_{f,t}^k$, are derived by using (4.7) that can be simply written as follows

$$f^{+}(\theta,\bar{\theta}) = \sum\nolimits_{k,f,t} \frac{\left|\bar{Y}_{f,t}^{k}\right|^{2} - 2\boldsymbol{W}_{f}^{k}\boldsymbol{H}_{t}^{k} Re\left[\bar{\boldsymbol{Y}}_{f,t}^{k} .e^{-j\phi_{f,t}^{k}}\right] + {\boldsymbol{W}_{f}^{k}}^{2} {\boldsymbol{H}_{t}^{k}}^{2}}{\beta_{f,t}^{k}}$$

$$+ \lambda \sum\nolimits_{k,t} \left(p\left|\bar{\boldsymbol{H}}_{t}^{k}\right|^{p-2} {\boldsymbol{H}_{t}^{k}}^{2} + (2-p)\left|\bar{\boldsymbol{H}}_{t}^{k}\right|^{p}\right)$$

$$= A - 2\ \sum\nolimits_{k,f,t} \frac{\boldsymbol{W}_{f}^{k}\boldsymbol{H}_{t}^{k} Re\left[\bar{\boldsymbol{Y}}_{f,t}^{k}.e^{-j\phi_{f,t}^{k}}\right]}{\beta_{f,t}^{k}} \left(\frac{\left|\bar{\boldsymbol{Y}}_{f,t}^{k}\right|}{\left|\bar{\boldsymbol{Y}}_{f,t}^{k}\right|}\right)$$

$$= A - 2\ \sum\nolimits_{k,f,t} \frac{\boldsymbol{W}_{f}^{k}\boldsymbol{H}_{t}^{k}\left|\bar{\boldsymbol{Y}}_{f,t}^{k}\right|}{\beta_{f,t}^{k}} \left(\frac{Re\left[\bar{\boldsymbol{Y}}_{f,t}^{k}.e^{-j\phi_{f,t}^{k}}\right]}{\left|\bar{\boldsymbol{Y}}_{f,t}^{k}\right|}\right)$$

$$= A - 2\sum\nolimits_{k,f,t}\left|\boldsymbol{B}_{f,t}^{k}\right| \frac{Re\left[\left({\bar{\boldsymbol{Y}}_{f,t}^{k}}^{(r)} + j{\bar{\boldsymbol{Y}}_{f,t}^{k}}^{(i)}\right)\left(\cos\phi_{f,t}^{k} - j\sin\phi_{f,t}^{k}\right)\right]}{\left|\bar{\boldsymbol{Y}}_{f,t}^{k}\right|}$$

$$= A - 2\sum\nolimits_{k,f,t}\left|\boldsymbol{B}_{f,t}^{k}\right| \frac{Re\left[{\bar{\boldsymbol{Y}}_{f,t}^{k}}^{(r)} \cos\phi_{f,t}^{k} - j{\bar{\boldsymbol{Y}}_{f,t}^{k}}^{(r)} \sin\phi_{f,t}^{k} + j{\bar{\boldsymbol{Y}}_{f,t}^{k}}^{(i)} \cos\phi_{f,t}^{k} + {\bar{\boldsymbol{Y}}_{f,t}^{k}}^{(i)} \sin\phi_{f,t}^{k}\right]}{\left|\bar{\boldsymbol{Y}}_{f,t}^{k}\right|}$$

$$= A - 2\sum\nolimits_{k,f,t}\left|\boldsymbol{B}_{f,t}^{k}\right| \left(\frac{Re\left[\bar{\boldsymbol{Y}}_{f,t}^{k}\right]\cos\phi_{f,t}^{k}}{\left|\bar{\boldsymbol{Y}}_{f,t}^{k}\right|} + \frac{Im\left[\bar{\boldsymbol{Y}}_{f,t}^{k}\right]\sin\phi_{f,t}^{k}}{\left|\bar{\boldsymbol{Y}}_{f,t}^{k}\right|}\right)$$

$$= A - 2\sum\nolimits_{k,f,t}\left|\boldsymbol{B}_{f,t}^{k}\right| \cos\phi_{f,t}^{k}\cos\Omega_{f,t}^{k} + \sin\phi_{f,t}^{k}\sin\Omega_{f,t}^{k}$$

$$= A - 2\sum\nolimits_{k,f,t}\left|\boldsymbol{B}_{f,t}^{k}\right| \cos\left(\phi_{f,t}^{k} - \Omega_{f,t}^{k}\right) \tag{4.14}$$

where A denotes the terms that irrelevant with $\boldsymbol{B}_{f,t}^{k} = \frac{\boldsymbol{W}_{f}^{k}\boldsymbol{H}_{t}^{k}\bar{\boldsymbol{Y}}_{f,t}^{k}}{\beta_{f,t}^{k}}$, $\cos\Omega_{f,t}^{k} = \frac{Re\left[\bar{\boldsymbol{Y}}_{f,t}^{k}\right]}{\left|\bar{\boldsymbol{Y}}_{f,t}^{k}\right|}$, $\sin\Omega_{f,t}^{k} = \frac{Im\left[\bar{\boldsymbol{Y}}_{f,t}^{k}\right]}{\left|\bar{\boldsymbol{Y}}_{f,t}^{k}\right|}$ and $\phi_{f,t}^{k}$. The auxiliary function, $f^{+}(\theta,\bar{\theta})$ in (4.7) is minimized when $\cos\left(\phi_{f,t}^{k} - \Omega_{f,t}^{k}\right) = \cos\phi_{f,t}^{k}\cos\Omega_{f,t}^{k} + \sin\phi_{f,t}^{k}\sin\Omega_{f,t}^{k} = 1$, namely,

$cos\,\phi_{f,t}^{k} = cos\,\Omega_{f,t}^{k}$ and $sin\,\phi_{f,t}^{k} = sin\,\Omega_{f,t}^{k}$. The update formula for $e^{j\phi_{f,t}^{k}}$ leads eventually to

$$e^{j\phi_{f,t}^{k}} = cos\,\phi_{f,t}^{k} + j\,sin\,\phi_{f,t}^{k}$$

$$= cos\,\Omega_{f,t}^{k} + j\,sin\,\Omega_{f,t}^{k}$$

$$= \frac{Re\left[\bar{\boldsymbol{Y}}_{f,t}^{k}\right] + Im\left[\bar{\boldsymbol{Y}}_{f,t}^{k}\right]}{\left|\bar{\boldsymbol{Y}}_{f,t}^{k}\right|}$$

$$= \frac{\bar{\boldsymbol{Y}}_{f,t}^{k}}{\left|\bar{\boldsymbol{Y}}_{f,t}^{k}\right|} \tag{4.15}$$

As in [50], the update formula for $\beta_{f,t}^{k}$ and $\boldsymbol{H}_{t}^{k}$, for projection onto the constraint space, is set to

$$\beta_{f,t}^{k} = \frac{\boldsymbol{W}_{f}^{k}\boldsymbol{H}_{t}^{k}}{\sum_{n}\boldsymbol{W}_{f}^{k}\boldsymbol{H}_{t}^{k}} \tag{4.16}$$

$$\boldsymbol{H}_{t}^{k} \leftarrow \frac{\boldsymbol{H}_{t}^{k}}{\sum_{n}\boldsymbol{H}_{t}^{k}} \tag{4.17}$$

4.1.2.2 Estimation of the Sparsity Parameter

It suffices to compute λ_{t}^{k} just for the regularization parameters associated with $\boldsymbol{H}_{t}^{k}$. Therefore, we can set the cost function in (4.10) as

$$F(\mathbf{H}, \boldsymbol{\lambda}) = \frac{1}{2\sigma^{2}}\left\| vec(\boldsymbol{Y}) - \sum_{k}\left((\boldsymbol{I}\otimes\mathbf{W})^{\circ}e^{j\phi}\right) vec(\mathbf{H})\right\|_{F}^{2}$$

$$- \sum_{k}(log\,\lambda)^{\boldsymbol{T}} + \sum_{k}(\lambda)^{\boldsymbol{T}}\,vec(\mathbf{H}) \tag{4.18}$$

with $vec(\cdot)$ represents the column vectorization, "⊗" is the Kronecker product, "$\circ$" is the Hadamard product, and $\boldsymbol{I}$ is the identity matrix. Defining the following terms:

$$\overline{\boldsymbol{W}} = \left[\boldsymbol{I}\otimes\boldsymbol{W}_f^1 \vdots \boldsymbol{I}\otimes\boldsymbol{W}_f^2 \vdots \cdots \vdots \boldsymbol{I}\otimes\boldsymbol{W}_f^K\right],\ e^{j\overline{\underline{\boldsymbol{\Phi}}}_t} = \left[e^{j\underline{\boldsymbol{\phi}}^1_{:,t}} \vdots e^{j\underline{\boldsymbol{\phi}}^2_{:,t}} \vdots \cdots \vdots e^{j\underline{\boldsymbol{\phi}}^K_{:,t}}\right]$$

$$\underline{\boldsymbol{y}} = vec(\boldsymbol{Y}) = \begin{bmatrix} \underline{\boldsymbol{Y}}_{:,1} \\ \cdots \\ \underline{\boldsymbol{Y}}_{:,2} \\ \cdots \\ \vdots \\ \cdots \\ \underline{\boldsymbol{Y}}_{:,T} \end{bmatrix},\ \underline{\boldsymbol{h}} = \begin{bmatrix} \boldsymbol{H}_t^1 \\ \cdots \\ \boldsymbol{H}_t^2 \\ \cdots \\ \vdots \\ \cdots \\ \boldsymbol{H}_t^K \end{bmatrix},\ \underline{\boldsymbol{\lambda}} = \begin{bmatrix} \boldsymbol{\lambda}_t^1 \\ \cdots \\ \boldsymbol{\lambda}_t^2 \\ \cdots \\ \vdots \\ \cdots \\ \boldsymbol{\lambda}_t^K \end{bmatrix},\ \underline{\boldsymbol{\phi}} = \begin{bmatrix} \boldsymbol{\phi}_{:,t}^1 \\ \cdots \\ \boldsymbol{\phi}_{:,t}^2 \\ \cdots \\ \vdots \\ \cdots \\ \boldsymbol{\phi}_{:,t}^K \end{bmatrix}$$

$$\overline{\boldsymbol{A}} = \begin{bmatrix} \overline{\boldsymbol{W}}\circ e^{j\overline{\underline{\boldsymbol{\Phi}}}_t} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \overline{\boldsymbol{W}}\circ e^{j\overline{\underline{\boldsymbol{\Phi}}}_t} & \boldsymbol{0} & \vdots \\ \vdots & \boldsymbol{0} & \overline{\boldsymbol{W}}\circ e^{j\overline{\underline{\boldsymbol{\Phi}}}_t} & \boldsymbol{0} \\ \boldsymbol{0} & \cdots & \boldsymbol{0} & \overline{\boldsymbol{W}}\circ e^{j\overline{\underline{\boldsymbol{\Phi}}}_t} \end{bmatrix} \tag{4.19}$$

Thus, (4.18) can be rewritten in terms of $\underline{\boldsymbol{h}}$ as

$$F(\underline{\boldsymbol{h}},\underline{\boldsymbol{\lambda}}) = \frac{1}{2\sigma^2}\left\|\underline{\boldsymbol{y}} - \overline{\boldsymbol{A}}\underline{\boldsymbol{h}}\right\|_F^2 + \underline{\boldsymbol{\lambda}}^{\boldsymbol{T}}\underline{\boldsymbol{h}} - \left(log\,\underline{\boldsymbol{\lambda}}\right)^{\boldsymbol{T}}\underline{\boldsymbol{1}} \tag{4.20}$$

Note that $\underline{\boldsymbol{h}}$ and $\underline{\boldsymbol{\lambda}}$ are vectors of dimension $R \times 1$ where $R = F \times T \times K$. To determine $\underline{\boldsymbol{\lambda}}$, we use the Expectation-Maximization (EM) algorithm and treat $\underline{\boldsymbol{h}}$ as the hidden variable where the log-likelihood function can be optimized with respect to $\underline{\boldsymbol{\lambda}}$. Using the Jensen's inequality, it can be shown that for any distribution $Q(\underline{\boldsymbol{h}})$, the log-likelihood function satisfies the following:

$$ln\, p\left(\underline{\boldsymbol{y}}\,\middle|\underline{\boldsymbol{\lambda}}, \overline{\boldsymbol{A}}, \sigma^2\right) \geq \int Q(\underline{\boldsymbol{h}})\, ln\left(\frac{p\left(\underline{\boldsymbol{y}}, \underline{\boldsymbol{h}}\middle|\underline{\boldsymbol{\lambda}}, \overline{\boldsymbol{A}}, \sigma^2\right)}{Q(\underline{\boldsymbol{h}})}\right) d\underline{\boldsymbol{h}} \tag{4.21}$$

One can easily check that the distribution that maximizes the right-hand side of (4.21) is given by $Q(\underline{\boldsymbol{h}}) = p\left(\underline{\boldsymbol{h}}\middle|\underline{\boldsymbol{y}}\,,\underline{\boldsymbol{\lambda}},\overline{\boldsymbol{A}},\sigma^2\right)$ which is the posterior distribution of $\underline{\boldsymbol{h}}$. In this section, we represent the posterior distribution in the form of Gibbs distribution as follows:

$$Q(\underline{\boldsymbol{h}}) = \frac{1}{Z_h} exp\left[-F(\underline{\boldsymbol{h}})\right] \text{ where } Z_h = \int exp\left[-F(\underline{\boldsymbol{h}})\right]\; d\underline{\boldsymbol{h}} \qquad (4.22)$$

The functional form of the Gibbs distribution in (4.22) is expressed in terms of $F(\underline{\boldsymbol{h}})$ and this is crucial as it will enable us to simplify the variational optimization of $\underline{\boldsymbol{\lambda}}$. The maximum-likelihood estimation of λ_t^k can be expressed by

$$\underline{\boldsymbol{\lambda}}^{ML} = arg\,\underset{\underline{\boldsymbol{\lambda}}}{max}\, ln\, p\left(\underline{\boldsymbol{y}}\,\middle|\underline{\boldsymbol{\lambda}},\overline{\boldsymbol{A}},\sigma^2\right)$$

$$= arg\,max_{\underline{\boldsymbol{\lambda}}} \int Q(\underline{\boldsymbol{h}})\, ln\, p\left(\underline{\boldsymbol{y}}\,,\underline{\boldsymbol{h}}\middle|\underline{\boldsymbol{\lambda}},\overline{\boldsymbol{A}},\sigma^2\right) d\underline{\boldsymbol{h}}$$

$$= arg\,max_{\underline{\boldsymbol{\lambda}}} \int Q(\underline{\boldsymbol{h}})\left(ln\, p\left(\underline{\boldsymbol{y}}\,\middle|\underline{\boldsymbol{h}},\overline{\boldsymbol{A}},\sigma^2\right) + ln\, p(\underline{\boldsymbol{h}}|\underline{\boldsymbol{\lambda}})\right) d\underline{\boldsymbol{h}}$$

$$= arg\,max_{\underline{\boldsymbol{\lambda}}} \int Q(\underline{\boldsymbol{h}})\, ln\, p(\underline{\boldsymbol{h}}|\underline{\boldsymbol{\lambda}})\, d\underline{\boldsymbol{h}} \qquad (4.23)$$

Similarly,

$${\sigma^2}_{ML} = arg\,\underset{\sigma^2}{max}\, ln\, p\left(\underline{\boldsymbol{y}}\,\middle|\underline{\boldsymbol{\lambda}},\overline{\boldsymbol{A}},\sigma^2\right)$$

$$= arg\,\underset{\sigma^2}{max} \int Q(\underline{\boldsymbol{h}})\, ln\, p\left(\underline{\boldsymbol{y}}\,,\underline{\boldsymbol{h}}\middle|\underline{\boldsymbol{\lambda}},\overline{\boldsymbol{A}},\sigma^2\right) d\underline{\boldsymbol{h}}$$

$$= arg\,\underset{\sigma^2}{max} \int Q(\underline{\boldsymbol{h}})\left(ln\, p\left(\underline{\boldsymbol{y}}\,\middle|\underline{\boldsymbol{h}},\overline{\boldsymbol{A}},\sigma^2\right) + ln\, p(\underline{\boldsymbol{h}}|\underline{\boldsymbol{\lambda}})\right) d\underline{\boldsymbol{h}}$$

$$= arg\,max_{\sigma^2} \int Q(\underline{\boldsymbol{h}})\, ln\, p\left(\underline{\boldsymbol{y}}\,\middle|\underline{\boldsymbol{h}},\overline{\boldsymbol{A}},\sigma^2\right) d\underline{\boldsymbol{h}} \qquad (4.24)$$

Since each element of $\mathbf{H}$ is constrained to be exponential distributed with independent decay parameters, this gives $p(\underline{\boldsymbol{h}}|\underline{\boldsymbol{\lambda}}) = \prod_g \lambda_g exp\,(-\lambda_g h_g)$ and therefore, (4.23) becomes

$$\underline{\boldsymbol{\lambda}}^{ML} = arg\, max_{\underline{\boldsymbol{\lambda}}} \int Q(\underline{\boldsymbol{h}})\left(ln\,\lambda_g - \lambda_g h_g\right) d\underline{\boldsymbol{h}} \qquad (4.25)$$

The Gibbs distribution $Q(\underline{\boldsymbol{h}})$ treats $\underline{\boldsymbol{h}}$ as the dependent variable while assuming all other parameters to be constant. As such, the functional optimization of $\underline{\boldsymbol{\lambda}}$ in (4.25) is obtained by differentiating the terms within the integral with respect to λ_g and the end result is given by

$$\lambda_g = \frac{1}{\int h_g Q(\underline{\boldsymbol{h}})\, d\underline{\boldsymbol{h}}} \quad \text{for} \quad g = 1,2,\ldots,R \qquad (4.26)$$

where $\lambda_g$is the g^{th}element of $\underline{\boldsymbol{\lambda}}$.

Since $p\left(\underline{\boldsymbol{y}}\,\middle|\underline{\boldsymbol{h}},\overline{\boldsymbol{A}},\sigma^2\right) = (\pi\sigma^2)^{-\boldsymbol{N_0/2}} exp\left(-(1/2\sigma^2)\left\|\underline{\boldsymbol{y}} - \overline{\boldsymbol{A}}\underline{\boldsymbol{h}}\right\|^2\right)$where$N_0 = K \times T$, the iterative update rule for $\sigma^2{}_{ML}$ is given by

$$\sigma^2{}_{ML} = arg\, max_{\sigma^2} \int Q(\underline{\boldsymbol{h}})\left(\frac{-N_0}{2} ln(\pi\sigma^2) - \frac{1}{2\sigma^2}\left\|\underline{\boldsymbol{y}} - \overline{\boldsymbol{A}}\underline{\boldsymbol{h}}\right\|^2\right) d\underline{\boldsymbol{h}}$$

$$= \frac{1}{N_0}\int Q(\underline{\boldsymbol{h}})\left(\left\|\underline{\boldsymbol{y}} - \overline{\boldsymbol{A}}\underline{\boldsymbol{h}}\right\|^2\right) d\underline{\boldsymbol{h}} \qquad (4.27)$$

Despite the simple form of (4.26) and (4.27), the integral is difficult to compute analytically and therefore, we seek an approximation to $Q(\underline{\boldsymbol{h}})$. We note that the solution

$\underline{\boldsymbol{h}}$ naturally partition its elements into distinct subsets $\underline{\boldsymbol{h}}_P$ and $\underline{\boldsymbol{h}}_M$ consisting of components $\forall_p \in P$ such that $h_P = 0$, and components $\forall_m \in M$ such that $h_m > 0$. Thus, the $F(\underline{\boldsymbol{h}})$ can be expressed as follows:

$$F(\underline{\boldsymbol{h}},\underline{\boldsymbol{\lambda}}) = \underbrace{\frac{1}{2\sigma^2}\left\|\underline{\boldsymbol{y}} - \bar{\boldsymbol{A}}_M\underline{\boldsymbol{h}}_M\right\|^2 + \underline{\boldsymbol{\lambda}}_M^{\boldsymbol{T}}\underline{\boldsymbol{h}}_M - (\log\underline{\boldsymbol{\lambda}})_M^{\boldsymbol{T}}\underline{\boldsymbol{1}}M}_{F(\underline{\boldsymbol{h}}_M,\underline{\boldsymbol{\lambda}}_M)}$$

$$+ \underbrace{\frac{1}{2\sigma^2}\left\|\underline{\boldsymbol{y}} - \bar{\boldsymbol{A}}_P\underline{\boldsymbol{h}}_P\right\|^2 + \underline{\boldsymbol{\lambda}}_P^{\boldsymbol{T}}\underline{\boldsymbol{h}}_P - (\log\underline{\boldsymbol{\lambda}})_P^{\boldsymbol{T}}\underline{\boldsymbol{1}}P}_{F(\underline{\boldsymbol{h}}_P,\underline{\boldsymbol{\lambda}}_P)} + \underbrace{\frac{1}{2\sigma^2}\left[2(\bar{\boldsymbol{A}}_M\underline{\boldsymbol{h}}_M)^{\boldsymbol{T}}(\bar{\boldsymbol{A}}_P\underline{\boldsymbol{h}}_P) - \left\|\underline{\boldsymbol{y}}\right\|^2\right]}_{G}$$

$$= F(\underline{\boldsymbol{h}}_M,\underline{\boldsymbol{\lambda}}_M) + F(\underline{\boldsymbol{h}}_P,\underline{\boldsymbol{\lambda}}_P) + G \qquad (4.28)$$

In (4.28), the term $\left\|\underline{\boldsymbol{y}}\right\|^2$ in G is a constant and the cross-term $(\bar{\boldsymbol{A}}_M\underline{\boldsymbol{h}}_M)^{\boldsymbol{T}}(\bar{\boldsymbol{A}}_P\underline{\boldsymbol{h}}_P)$ measures the orthogonality between $\bar{\boldsymbol{A}}_M\underline{\boldsymbol{h}}_M$ and $\bar{\boldsymbol{A}}_P\underline{\boldsymbol{h}}_P$, where $\bar{\boldsymbol{A}}_P$ is the sub-matrix of $\bar{\boldsymbol{A}}$ that corresponds to $\underline{\boldsymbol{h}}_P$, $\bar{\boldsymbol{A}}_M$ is the sub-matrix of $\bar{\boldsymbol{A}}$ that corresponds to $\underline{\boldsymbol{h}}_M$. In this section, we intend to simplify the expression in (4.28) by discounting the contribution from these terms and let $F(\underline{\boldsymbol{h}})$ be approximated as $F(\underline{\boldsymbol{h}},\underline{\boldsymbol{\lambda}}) \approx F(\underline{\boldsymbol{h}}_M,\underline{\boldsymbol{\lambda}}_M) + F(\underline{\boldsymbol{h}}_P,\underline{\boldsymbol{\lambda}}_P)$. Given this approximation, $Q(\underline{\boldsymbol{h}})$ can be decomposed as

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

$$\approx \frac{1}{Z_h}exp\left[-\left(F(\underline{\boldsymbol{h}}_M,\underline{\boldsymbol{\lambda}}_M) + F(\underline{\boldsymbol{h}}_P,\underline{\boldsymbol{\lambda}}_P)\right)\right]$$

$$= \frac{1}{Z_h}exp[-F(\underline{\boldsymbol{h}}_M,\underline{\boldsymbol{\lambda}}_M)]\ exp[-F(\underline{\boldsymbol{h}}_P,\underline{\boldsymbol{\lambda}}_P)]$$

$$= \frac{1}{Z_M}exp[-F(\underline{\boldsymbol{h}}_M,\underline{\boldsymbol{\lambda}}_M)]\ \frac{1}{Z_P}exp[-F(\underline{\boldsymbol{h}}_P,\underline{\boldsymbol{\lambda}}_P)]$$

$$= Q_M(\underline{\boldsymbol{h}}_M)Q_P(\underline{\boldsymbol{h}}_P) \qquad (4.29)$$

where $Z_P = \int exp\left[-F\left(\underline{\boldsymbol{h}}_P, \underline{\boldsymbol{\lambda}}_P\right)\right] d\underline{\boldsymbol{h}}_P$, and $Z_M = \int exp\left[-F\left(\underline{\boldsymbol{h}}_M, \underline{\boldsymbol{\lambda}}_M\right)\right] d\underline{\boldsymbol{h}}_M$. In order to characterize $Q_P\left(\underline{\boldsymbol{h}}_P\right)$, we need to allow some positive deviation to $\underline{\boldsymbol{h}}_P$ (any negative values of $\underline{\boldsymbol{h}}_P$ will be rejected since NMF only allow nonnegative values). Hence, $\underline{\boldsymbol{h}}_P$ must take on zero and positive values in $Q_P\left(\underline{\boldsymbol{h}}_P\right)$. The distribution $Q_P\left(\underline{\boldsymbol{h}}_P\right)$can be approximated by using the Taylor expansion about the maximum a posterior (MAP) estimate, $\underline{\boldsymbol{h}}^{MAP}$ given in (4.13)

$$Q_P\left(\underline{\boldsymbol{h}}_P \geq 0\right) \propto exp\left\{-\left[\left(\frac{\partial F}{\partial \underline{\boldsymbol{h}}}\right)\Bigg|_{\underline{\boldsymbol{h}}^{MAP}}\right]_P^{\boldsymbol{T}} \underline{\boldsymbol{h}}_P - \frac{1}{2}\underline{\boldsymbol{h}}_P^{\boldsymbol{T}}\overline{\boldsymbol{C}}_P\underline{\boldsymbol{h}}_P\right\}$$

$$= exp\left[-\left(\overline{\boldsymbol{C}}\underline{\boldsymbol{h}}^{MAP} - \frac{1}{\sigma^2}\overline{\boldsymbol{A}}^{\boldsymbol{T}}\underline{\boldsymbol{y}} + \underline{\boldsymbol{\lambda}}\right)_P^{\boldsymbol{T}} \underline{\boldsymbol{h}}_P - \frac{1}{2}\underline{\boldsymbol{h}}_P^{\boldsymbol{T}}\overline{\boldsymbol{C}}_P\underline{\boldsymbol{h}}_P\right] \tag{4.30}$$

where $\overline{\boldsymbol{C}}_P = \frac{1}{\sigma^2}\overline{\boldsymbol{A}}_P^{\boldsymbol{T}}\overline{\boldsymbol{A}}_P$ and $\overline{\boldsymbol{C}} = \frac{1}{\sigma^2}\overline{\boldsymbol{A}}^{\boldsymbol{T}}\overline{\boldsymbol{A}}$. Although $Q_P\left(\underline{\boldsymbol{h}}_P\right)$ is obtained in the form of (4.30), its integral is difficult to evaluate and does not yield closed form analytical expression of the moments, which subsequently prohibits inference of the sparsity parameters. Alternatively, we may variationally approximate $Q_P\left(\underline{\boldsymbol{h}}_P\right)$ by using a fixed form distribution that can yield a closed analytical expression of the moments. Since $\underline{\boldsymbol{h}}_P$ takes on zero and positive values only, a suitable fixed form distribution is to use the factorized exponential distribution given by

$$\hat{Q}_P\left(\underline{\boldsymbol{h}}_P \geq 0\right) = \prod_{p\in P}\frac{1}{u_p} exp\left(\frac{-h_p}{u_p}\right) \tag{4.31}$$

The variational parameters $\underline{\boldsymbol{u}} = \{u_p\}$ for $\forall p \in P$ are obtained by minimizing the Kullback–Leibler divergence between Q_P and $\hat{Q}_P$

$$\underline{\boldsymbol{u}} = arg \min_{\underline{\underline{\boldsymbol{u}}}} \hat{Q}_P\left(\underline{\boldsymbol{h}}_P\right) ln \frac{\hat{Q}_P\left(\underline{\boldsymbol{h}}_P\right)}{Q_P\left(\underline{\boldsymbol{h}}_P\right)} d\underline{\boldsymbol{h}}_P$$

$$= arg \min_{\underline{\underline{\boldsymbol{u}}}} \hat{Q}_P\left(\underline{\boldsymbol{h}}_P\right) \left[ln\, \hat{Q}_P\left(\underline{\boldsymbol{h}}_P\right) - ln\, Q_P\left(\underline{\boldsymbol{h}}_P\right)\right] d\underline{\boldsymbol{h}}_P \qquad (4.32)$$

Solving (4.32) for u_p leads to the following update [84]:

$$u_p \leftarrow u_p \frac{-\hat{b}_p + \sqrt{\hat{b}_p^2 + 4\frac{\left(\widehat{\boldsymbol{C}}\underline{\boldsymbol{u}}\right)_p}{\tilde{u}_p}}}{2\left(\widehat{\boldsymbol{C}}\underline{\boldsymbol{u}}\right)_p} \qquad (4.33)$$

The approximate distribution for components $\underline{\boldsymbol{h}}_M$ can be obtained substituting $F\left(\underline{\boldsymbol{h}}_M, \underline{\boldsymbol{\lambda}}_M\right)$ into $Q_M\left(\underline{\boldsymbol{h}}_M\right)$ as follows:

$$Q_M\left(\underline{\boldsymbol{h}}_M\right) = \frac{1}{Z_M} exp\left[-F\left(\underline{\boldsymbol{h}}_M, \underline{\boldsymbol{\lambda}}_M\right)\right]$$

$$\propto\ exp\left[-\left(\frac{1}{2}\underline{\boldsymbol{h}}_M^T \overline{\boldsymbol{C}}_M \underline{\boldsymbol{h}}_M - \frac{1}{\sigma^2}\underline{\boldsymbol{y}}^T \overline{\boldsymbol{A}}_M \underline{\boldsymbol{h}}_M + \underline{\boldsymbol{\lambda}}_M \underline{\boldsymbol{h}}_M\right)\right] \qquad (4.34)$$

In (4.34), $Q_M\left(\underline{\boldsymbol{h}}_M\right)$ has the functional form equivalent to a multivariate Gaussian distribution. Therefore, $Q_M\left(\underline{\boldsymbol{h}}_M\right)$ can be represented as the unconstrained Gaussian with mean $\underline{\boldsymbol{h}}_M^{MAP}$ and covariance $\overline{\boldsymbol{C}}_M^{-1}$, where $\overline{\boldsymbol{C}}_M$ is the sub-matrix of $\overline{\boldsymbol{C}}$.

Substituting (4.29), (4.31), (4.34) into (4.26), the sparsity parameter can be inferred as

$$\lambda_g = \begin{cases} \frac{1}{\int h_g Q_M\left(\underline{\boldsymbol{h}}_M\right) d\underline{\boldsymbol{h}}_M} = \frac{1}{h_g^{MAP}}\ if\ g\ \in\ M \\ \frac{1}{\int h_g \hat{Q}_P\left(\underline{\boldsymbol{h}}_P\right) d\underline{\boldsymbol{h}}_P} = \frac{1}{u_g}\ \ if\ g\ \in\ P \end{cases} \qquad (4.35)$$

and its covariance **X** is given by

$$X_{ab} = \begin{cases} \left(\overline{\boldsymbol{C}}_P^{-1}\right)_{ab},\ if\ a, b\ \in\ M \\ u_p^2 \delta_{ab}\quad,\ Otherwise. \end{cases} \qquad (4.36)$$

Similarly, the inference for σ^2 can be computed from (4.27) as

$$\sigma^2 = \frac{1}{N_0}\int Q(\underline{\boldsymbol{h}})\left(\left\|\underline{\boldsymbol{y}} - \overline{\boldsymbol{A}}\underline{\hat{\boldsymbol{h}}}\right\|^2\right) d\underline{\boldsymbol{h}} \tag{4.37}$$

where $\hat{h}_g = \begin{cases} h_g^{MAP} & if\ g\ \in\ M \\ u_g & if\ g\ \in\ P \end{cases}.$

Table 4.1 presents the main steps of the proposed method. We term the above algorithm as the variational L_1-sparse CMF (vL_1-SCMF). The proposed algorithm for single-channel blind separation is summarized in Table 4.1.

Table 4.1: Overview the proposed vL_1-SCMF algorithm.

1. Compute $\boldsymbol{Y}_{f,t} = STFT(y(t))$.

2. Initialize $\boldsymbol{W}_f^k$, $\boldsymbol{H}_t^k$ and $\boldsymbol{\phi}_{f,t}^k$ with nonnegative random values.

3. Update $\beta_{f,t}^k$ according to (4.16) and fixing the value of ϕ at $e^{j\phi_{f,t}^k} = \frac{\boldsymbol{Y}_{f,t}}{|\boldsymbol{Y}_{f,t}|}$ and update u_p using (4.33).

4. Calculate $\boldsymbol{\lambda}_g$ and σ^2 using (4.35) and (4.37).

5. Update $\bar{\theta} = \{\overline{\mathbf{Y}}, \overline{\mathbf{H}}\}$ according to (4.8), (4.9), update $\theta = \{\boldsymbol{W}_f^k, \boldsymbol{H}_t^k, \boldsymbol{\phi}_{f,t}^k\}$ according to (4.12), (4.13), (4.15) and Update $\beta_{f,t}^k$ and $\boldsymbol{H}_t^k$ according to (4.16) and (4.17).

6. Repeat steps 3 to 5 until convergence is reached.

7. Obtain an estimation of each source by multiplying the respective rows of the spectral components $\boldsymbol{W}_f^k$ with the corresponding columns of the mixture weights $\boldsymbol{H}_t^k$ and time-varying phase spectrum $e^{j\phi_{f,t}^k}$. Convert the time-frequency represented sources into time domain to obtain the separated sources.

4.2 Experimental Results and Analysis

4.2.1 Experimental Environments

In this section, the proposed single channel sources separation method vL_1-SCMF is tested with the real audio sources gen. We erated mixed signal from 30 music signals including 10 drum, 10 jazz, 10 and 10 piano signals are selected from the RWC [85] database and 20 sentences of the target speakers (10 male and 10 female sentences from 8 male and 8 female subjects) are selected from the TIMIT speech database. The sources are randomly chosen from the database and the mixed signal is generated by adding the chosen sources. In all cases, the sources are mixed with equal average power over the duration of the signals. As an example, the mixtures are generated i.e. piano + jazz, piano + drum, jazz + drum, piano + male speech, jazz + male speech, drum + male speech and male speech + female speech. Three types of mixture can be summarized as follows: 1) Music mixed with Music, 2) speech mixed with music, and 3) speech mixed with speech. The TF representation is computed by normalizing the time-domain signal to unit power and computing the STFT using 1024 point Hanning window with 50% overlap. The parameter corresponding to the number of components K is set as 4. All experiments are conducted using a PC with Intel® Core™ i5 CPU 650 at 3.2 GHz and 4 GB RAM.

4.2.2 Source Separation Results of the proposed method

In this section, we have generated the mixtures of two sources which select from piano, drum, jazz, male speech and female speech. Both sources are mixed with equal power to generate the mixture. This is shown in the first three panels of Figure 4.1.

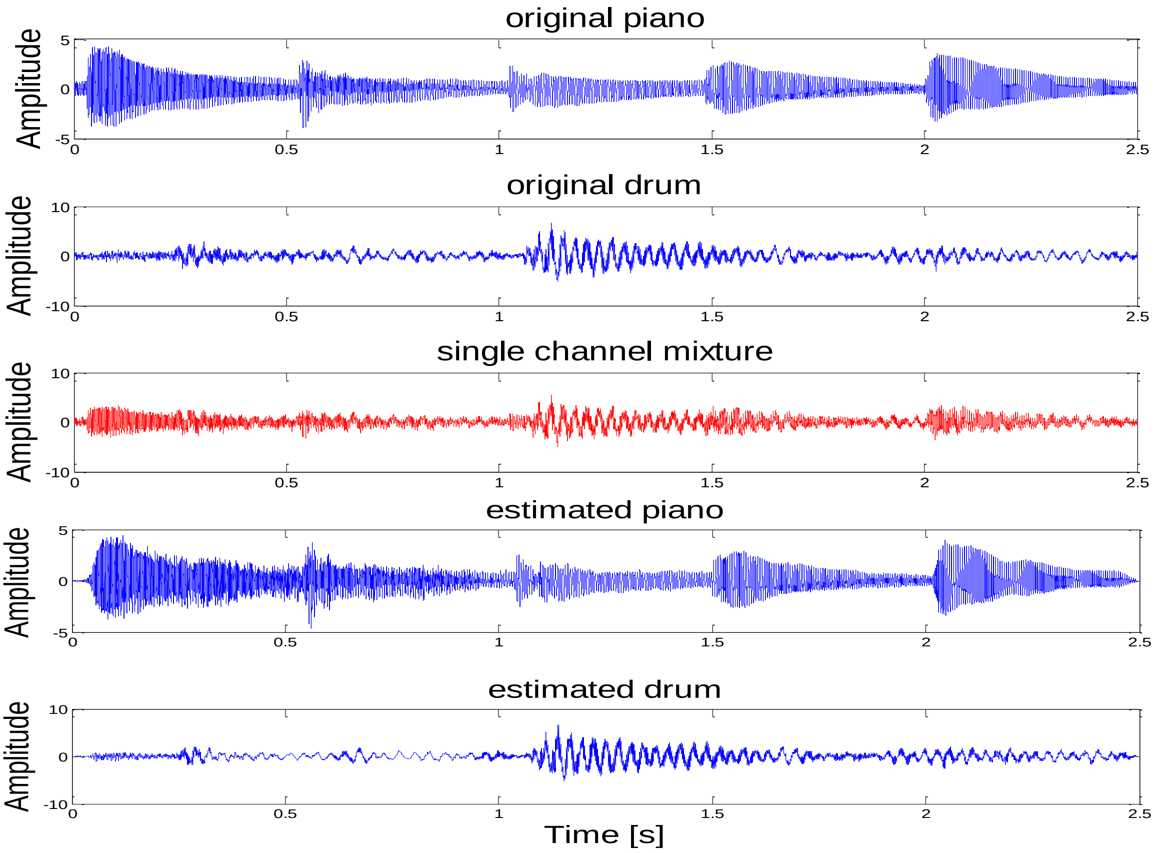


Figure 4.1: Time-domain representation of the original sources, single channel mixture, and estimated sources of music mixture between piano and drum using the proposed method.

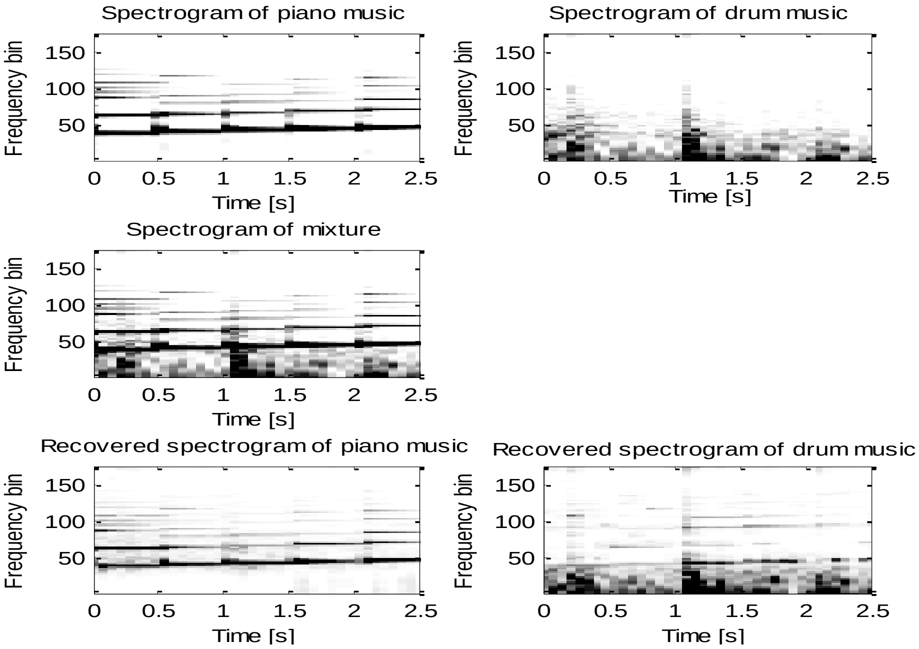
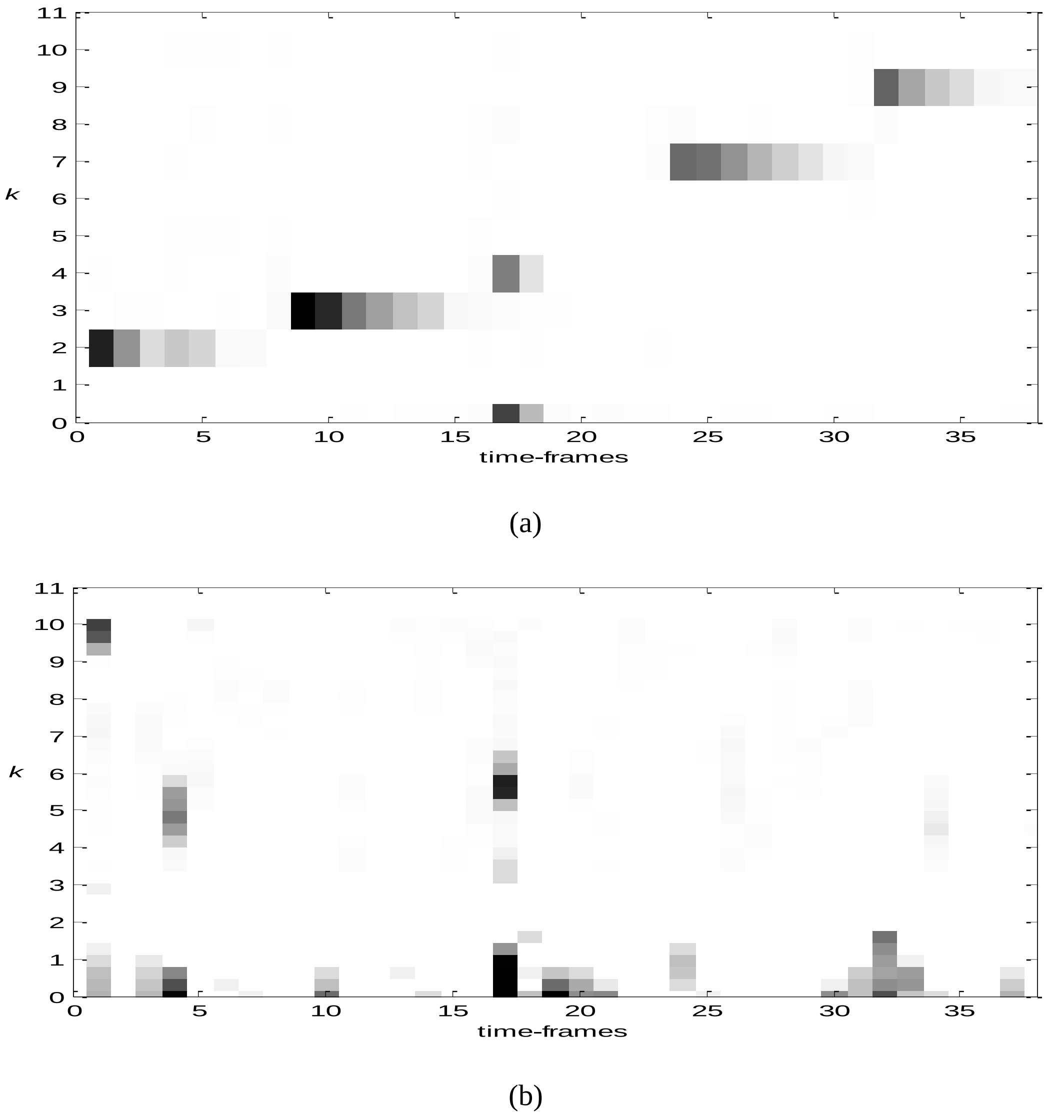


Figure 4.2: TF domain representation of the original piano and drum music (top panels), mixed signal (middle panels), and separated signal piano and drum (bottom panels) using the proposed method.

Figures 4.1 and 4.2 shows of the original piano, drum music, the single channel mixture and the separated sources using the proposed method in terms of spectrogram and time-domain representation, respectively. The estimated sources are plotted in the last two panels of Figure 4.1. From the plots in both Figures 4.1 and 4.2, they are visually evident that the estimated sources resemble closely to the original sources. The mean square error (MSE) between the original and the estimated music is 0.11dB and 0.07dB for piano and drum, respectively.



(a)

(b)

Figure 4.3: Estimated **H.** (a) piano (b) drum.

In this section, seven types of mixture have been generated: 1) piano mixed jazz; 2) piano mixed drum; 3) jazz mixed drum; 4) piano mixed male speech; 5) jazz mixed male speech; 6) drum mixed male speech and 7) male mixed female speech. All separation results have been summarized in Figure 4.3. The separation of piano + drum music mixture is much better than those of other types of mixtures where the average SDR has approached to 14.6dB, 13.5dB for recovered piano music and 15.8dB for recovered drum music. Figure 4.3 shows the matrix factorization results in term of the temporal codes $\mathbf{H}$ in the case of “optimally-sparse” based on the proposed method.

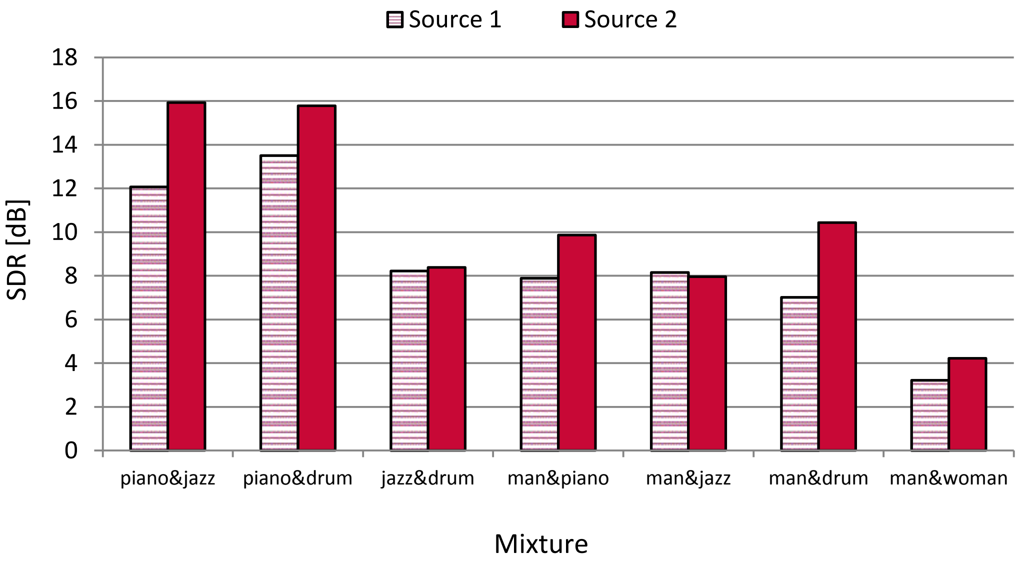
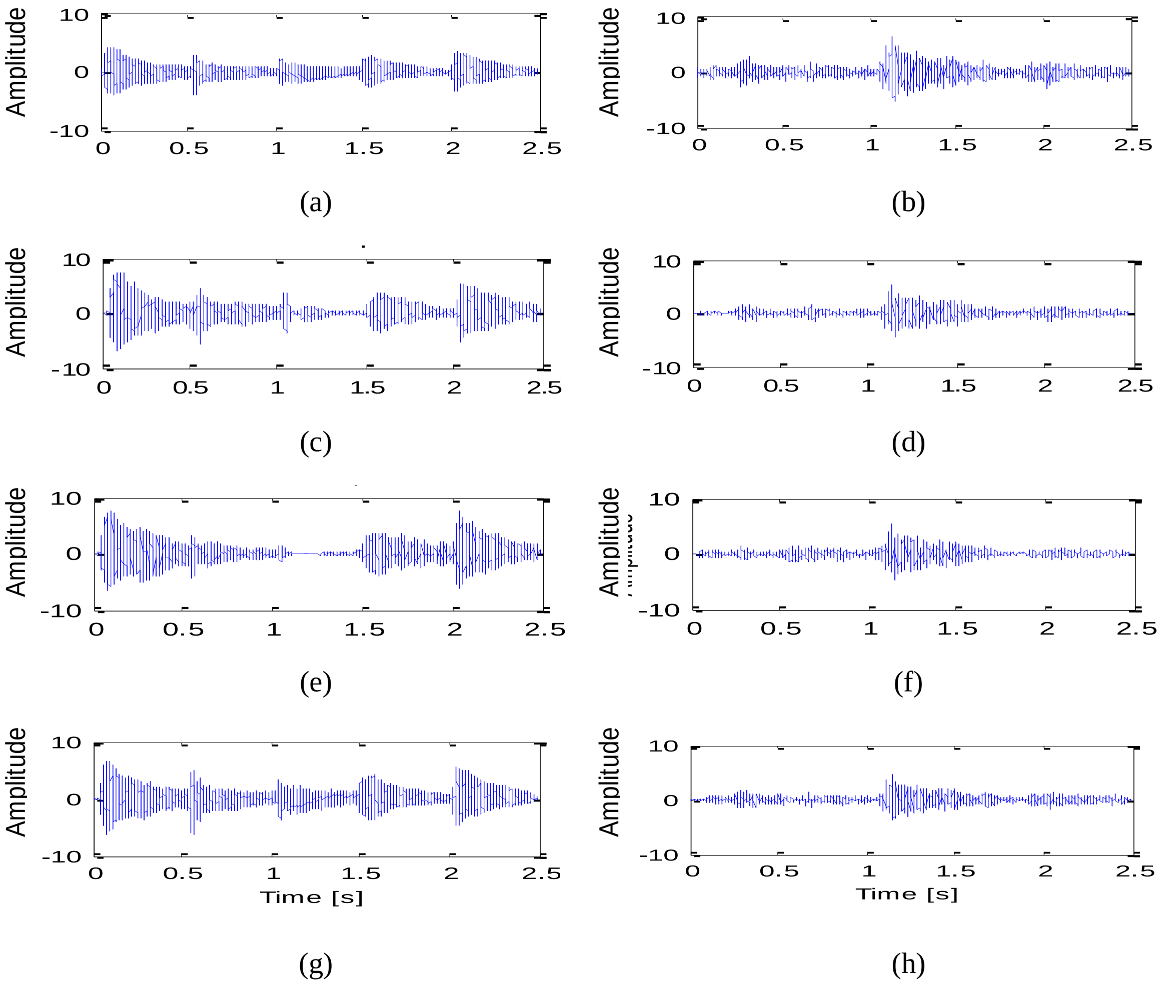


Figure 4.4: Overall separation results of different types of mixtures using the proposed method.

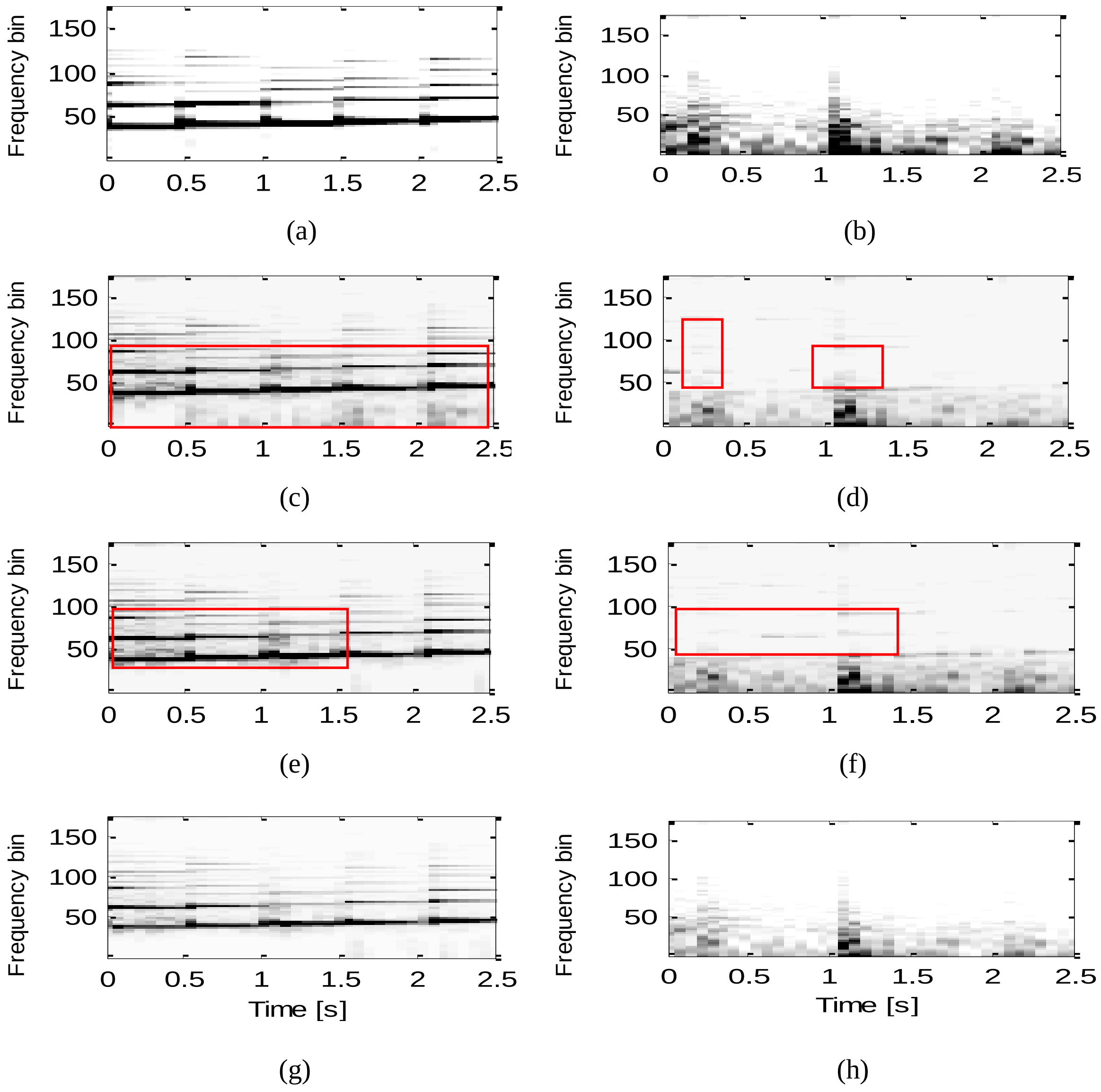
Figure 4.4 summarizes the separation results of the proposed method. It is worth pointing out that because the piano music and drum music have different basis components. Hence, it is easier to separate these signals by using the vL_1-SCMF. Thus, Figure 4.4 shows the better separation results over all the mixtures when audio mixture contains piano and drum music. On the other hand, the frequency range of male speech is very similar to female speech sources and this particular mixture is very difficult to separate which explains the reason why the SDR is relatively low. For separating the male speech and female speech mixture, the vL_1-SCMF yields an average SDR of 3.72dB. However, this performance is still substantially better than using the CMF alone.

4.2.3 Effect on Source Separation with Variational L_1-Sparse and Fixed Sparsity



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 4.5: Time-domain representation of (a)–(b): the original piano and drum music. (c)–(d) and (e)–(f) denote the recovered piano and drum music using uniform sparsity factorization with $\lambda = 0.01$ and $\lambda = 100$, respectively. (g)–(h) denote the recovered piano and drum music using the proposed method.



(a)

(b)

(c)

(d)

(e)

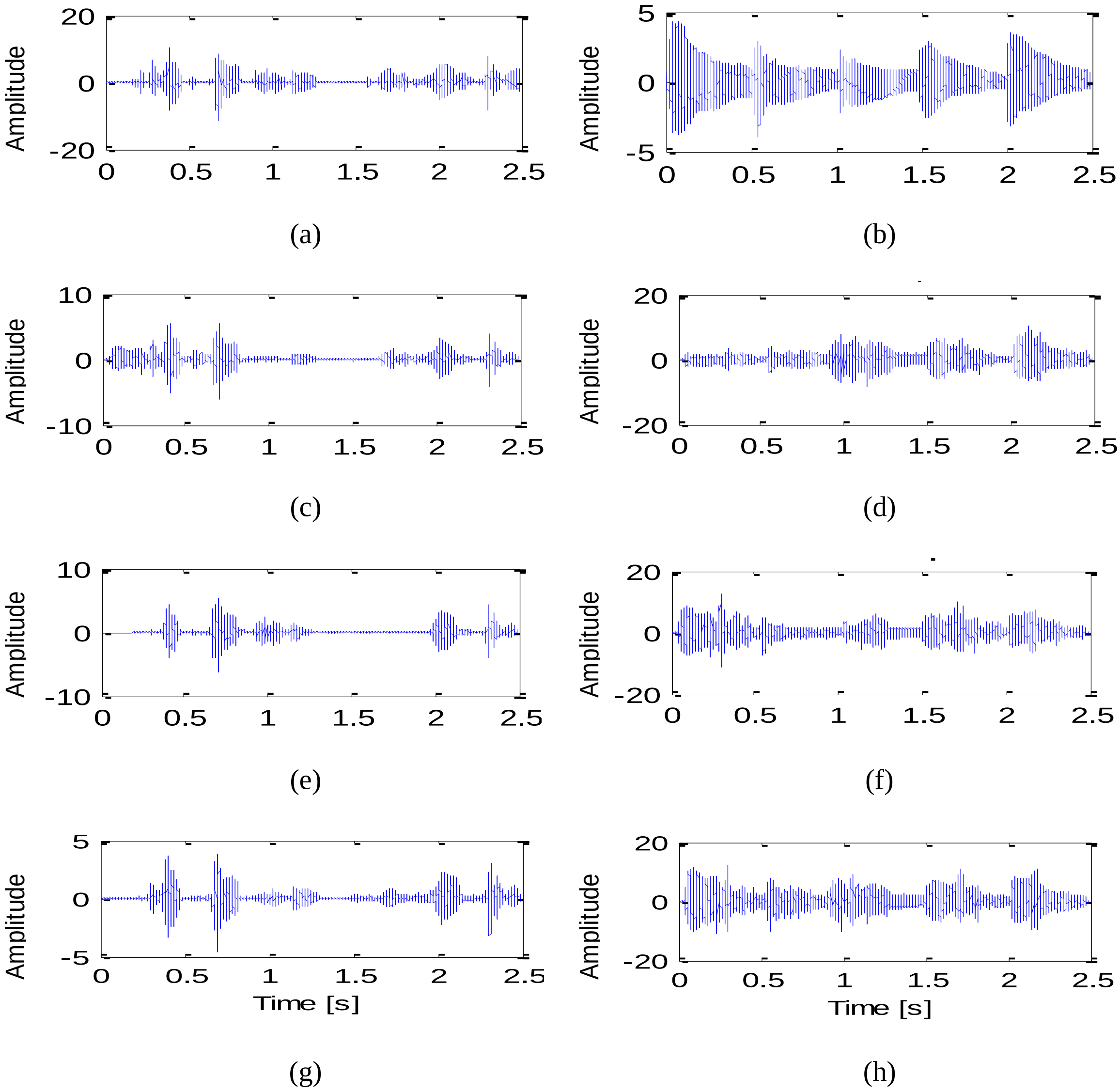
(f)

(g)

(h)

Figure 4.6: Spectrogram of (a)–(b): the original piano and drum music. (c)–(d) and (e)–(f) denote the recovered piano and drum music using uniform sparsity factorization with $\lambda = 0.01$ and $\lambda = 100$, respectively. (g)–(h) denote the recovered piano and drum music using the proposed method.

In this implementation, we have conducted several experiments to compare the performance of the proposed method with unsupervised CMF under different sparsity regularization. To investigate the effect of sparsity regularization on source separation performance, we evaluated and compared the proposed variational L_1-sparse with the case of 1) Uniform constant sparsity with low sparseness e.g., $\lambda_t^k = 0.01$ and 2) Uniform constant sparsity with high sparseness e.g., $\lambda_t^k = 100$. We use an example mixture of piano music and drum music as shown in Figures 4.5 and 4.6. And set the number of components per each source to $K_j = 4$ with $j = 2$ sources. An initial CMF decomposition is random. The number of iterations was set to 200 for updating the parameters. The hypothesized is set that the proposed variational L_1-sparse will significantly yield improvement of the audio source separation compare with fixed sparsity.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 4.7: Time-domain representation of (a)–(b): the original male speech and piano music. (c)–(d) and (e)–(f) denote the recovered male speech and piano music using uniform sparsity factorization with $\lambda = 0.01$ and $\lambda = 100$, respectively. (g)–(h) denote the recovered male speech and piano music using the proposed method.

Figsures 4.7 and 4.8 show the separated source in term of time-domain representation and spectrogram, respectively. Panels (a)-(b) show the original sources of the speech and music mixture which are the male speech and piano, respectively. Panels (c)-(h) shows the estimated sources using uniform constant sparsity factorization with low sparseness ($\lambda_t^k = 0.01$), high sparseness ($\lambda_t^k = 100$) and the proposed variational L_1-sparse parameters.

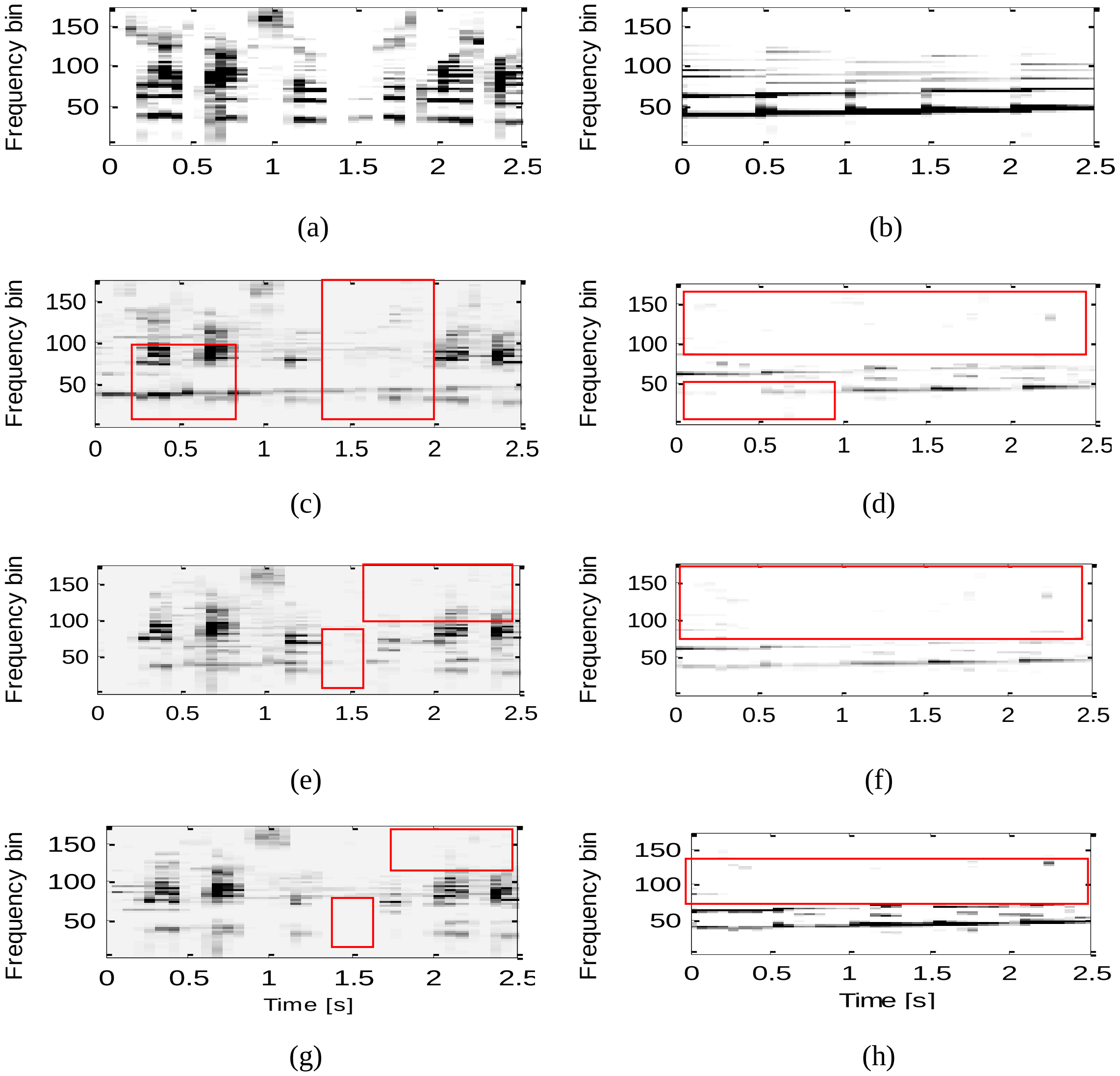


Figure 4.8: Spectrogram of (a)–(b): the original male speech and piano music. (c)–(d) and (e)–(f) denote the recovered male speech and piano music using uniform sparsity factorization with $\lambda = 0.01$ and $\lambda = 100$, respectively. (g)–(h) denote the recovered male speech and piano music using the proposed method.

Table 4.2: Comparison of average SDR and SIR performance on three types of mixtures between uniform regularization methods and the proposed method (vL_1-SCMF).

Mixtures	Methods	SDR (dB)	SIR (dB)
Music and Music	Proposed method (vL_1-SCMF)	12.32	14.87
	(Best)Uniform regularization sparsity	10.89	14.68
Music and Speech	Proposed method (vL_1-SCMF)	8.55	9.54
	(Best)Uniform regularization sparsity	7.15	7.36
Male Speech and Female Speech	Proposed method (vL_1-SCMF)	3.72	5.58
	(Best)Uniform regularization sparsity	2.81	4.53

The overall comparison results between the proposed variational L_1-sparse and uniform sparsity methods have been summarized in Table 4.2. According to the table, CMF with variational L_1-sparse tends to yield better result than the uniform sparsity-based methods. We may summarize the average performance improvement of our method against the uniform constant sparsity method: 1) for the music and music mixtures, the improvements per source in terms of the SDR are 1.4dB and SIR 0.2dB. 2) For the music and speech mixtures, the improvements per source in terms of SDR are 1.4dB and SIR 2.2dB. 3) For the male speech and female speech mixtures, the improvements per source in terms of SDR are 0.9dB and SIR 1.1dB. On a point of interest, the analyses for the case of 1) Uniform constant sparsity with low sparseness e.g., $\lambda_t^k = 0.01$ and 2) Uniform constant sparsity with high sparseness e.g., $\lambda_t^k = 100$ in Figures. 4.5 and 4.6 are based on the single fixed uniform sparsity parameter where is set to be too high and too low, respectively. From these results, it could be argued that such settings of uniform sparsity parameter are unrealistic for source separation. To investigate this further, the impact of sparsity regularization on the separation results in terms of the SDR under different uniform regularization has been undertaken and the results are plotted in Figure 4.9. In this implementation, the uniform regularization for all sparsity parameters e.g., $\lambda_t^k = 0, 0.5, \ldots, 10$. The best result is retained and tabulated in Table 4.2.

In Figure 4.9, the results have clearly indicated that there are certain values of where the unsupervised CMF performs with exceptionally good results. In the case of music and music mixtures, the best performance is obtained when ranges from 0.5 to 2 where the highest SDR is 10.9dB. As for music and speech mixtures, the best performance is obtained when ranges from 1.0 to 3.5 where the highest SDR is 7.2dB and for male speech and female speech mixtures, the best performance is obtained when ranges from 1 to 4 where the highest SDR is 2.8dB. On the contrary, when is set too high, the separation performance tends to degrade. It is also worth pointing out that the separation results are coarse when the factorization is non-regularized. Here, we see that 1) for music and music mixtures, the SDR is only 7.8dB, 2) for music and speech mixtures, the SDR is only 5.7dB, and 3) for male speech and female speech mixtures, the SDR is only 1dB. From above, it is evident that uniform sparsity scheme gives varying performance depending on the value of which in turn depends on the type of mixture. Hence, this poses a practical difficulty in selecting the appropriate level sparseness necessary for matrix factorization to resolve the ambiguity between the sources in the TF domain.

For comparison purposes, we have summarized the average performance improvement of our proposed method against the case of the uniform constant sparsity ($\lambda_t^k = 0, 0.5, \ldots, 10$) and the case of non-regularized ($\lambda_t^k = 0$) based on Figure 4.9 as follows: 1) for mixture of music signals, the average improvements are 4.4dB and 4.5dB per source, respectively 2) for mixture of music and speech signal, the average improvements are 2.7dB and 2.9dB per source, respectively, and 3) for mixture of speech signals, the average improvements are 1.7dB and 2.7dB per source, respectively. The above results clearly indicate that the performances of source separation have been undermined when the uniform constant sparsity scheme is used. On the contrary, improved performances can be obtained by allowing the sparsity parameters to be individually adapted for each element code. This is evident based on source separation performance as indicated in Table 4.2.

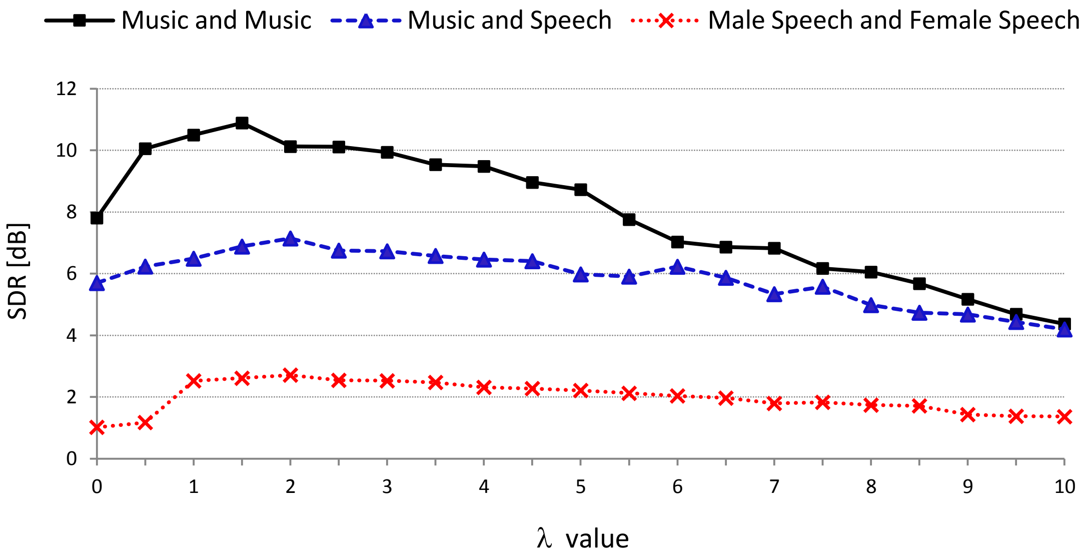


Figure 4.9: Separation results of vL_1-SCMF by using different uniform regularization.

4.2.4 Comparison With Other SCBSS Methods

In this evaluation, we compare the proposed method with similar class of matrix factorization methods, e.g., original CMF [50] and SNMF with second-order optimization

[11], NMF with Itakura-Saito divergence (NMF-ISD) [19], and single-channel ICA (SCICA) [24].

Table 4.3: Comparison of average SDR and SIR performance on three types of mixtures between SCICA, NMF-ISD, SNMF, CMF and the proposed method (vL_1-SCMF).

Mixtures	**Methods**	**SDR (dB)**	**SIR (dB)**
Music and Music	Proposed method (vL_1-SCMF)	12.72	14.87
	CMF [50]	6.11	7.53
	SNMF [11]	5.23	7.14
	NMF-ISD [19]	5.17	6.31
	SCICA [24]	3.85	4.86
Music and Speech	Proposed method (vL_1-SCMF)	8.92	9.84
	CMF [50]	6.06	6.64
	SNMF [11]	4.52	6.11
	NMF-ISD [19]	3.55	6.62
	SCICA [24]	1.43	3.12
Male Speech and Female Speech	Proposed method (vL_1-SCMF)	4.53	5.78
	CMF [50]	3.89	5.65
	SNMF [11]	1.62	3.23
	NMF-ISD [19]	2.06	4.27
	SCICA [24]	-0.56	1.25

Table 4.3 further gives the SDR and SIR comparison results between our proposed method and the above four methods. The improvement of our method compared with

CMF, SNMF, NMF-ISD and SCICA can be summarized as follows: 1) for the music and music, the average improvement per source in SDR is 7.6dB and in SIR 8.4dB; 2) for music and speech, the average improvement per source in SDR is 5.0dB and in SIR 4.2dB; 3) for male speech and female speech, the average improvement per source in SDR is 2.8dB and in SIR 3.5dB. Analyzing the separation results and SDR performance, the proposed method leads to the best separation performance for both recovered sources. In the case of the SCICA method performs with poorer results, we note that the recovered sources have not been clearly separated and the mixing ambiguity region is still large when compared with the original speeches. The SCICA models the sources as sparse combination of a set of time-domain basis functions which are initially derived using the ICA methods. The sources are subsequently estimated by maximizing the log-likelihood with the ICA-derived basis functions. This method renders optimal separation when the ICA basis functions corresponding to each source have minimal time-domain overlap. In the case where the basis functions have significant overlap with each other e.g. mixture of two speech sources where the basis functions for two sources are very similar, this method performs very poorly. Of note is that the CMF method exploits the phase information of the sources which is inherently ignored by SNMF and NMF-ISD and this has led to improved performance about 2dB in SDR. In addition, by carefully adapting the sparsity parameter for each temporal code using the proposed variational L_1-norm method, a considerable interference rejection level has been achieved. This is ostensibly apparent in the SIR criterion. On the other hand, the parts decomposed by the SNMF and NMF-ISD method are not adequate to capture the phase spectra and the temporal dependency of the frequency patterns within the audio signal. Additionally, the CMF and NMF-ISD are not unique if the data does not span the positive octant adequately since a rotation of $\mathbf{W}$ and opposite $\mathbf{H}$ can give the same results. In CMF, the sparsity parameter is set manually and therefore it is difficult to avoid under or over sparse resolution of the factorization.

4.3 Summary

This chapter has presented a novel framework of amalgamating variational L_1-sparse with complex matrix factorization for single channel source separation. The impetus behind the proposed work is that NMF cannot estimate the phase spectra of underlying constituent signals, and sparseness achieved by the conventional CMF is not efficient enough. The proposed method addresses the above and enjoys at least two significant advantages: first, the sparse regularization term is adaptively tuned to obtain the desired sparse decomposition, and second, the proposed method can extract recurrent patterns of magnitude spectra that underlie observed complex spectra and the phase estimates of constituent signals, thus enabling the features of the components to be extracted more efficiently. In addition, we derived analytical update equations through an auxiliary function approach and an experimental evaluation showed that reasonably good separation was obtained with the present method.

CHAPTER 5

SINGLE CHANNEL BLIND SOURCE SEPARATION USING IMITATED STEREO AUDIO MIXTURE WITH REGULARIZED NONNEGATIVE TENSOR FACTORIZATION

In this chapter, nonnegative matrix factorization given by a single observed mixture is extended to multiple-array mixtures. A novel approach for solving the SCBSS problem is developed. The proposed mixing mixture is an analogy of a stereo signal concept given by two microphones, one being the real and another is virtual. An “imitated-stereo” mixture model is developed by weighting and time-shifting the original single-channel mixture. This leads to an artificial mixing system of dual channels which gives rise to a new form of temporal correlation diversity of the sources. The imitated-stereo mixture has further culminated to a new development of the parallel factor analysis (PARAFAC) model. The PARAFAC model yields a time-frequency representation of an artificial dual channels information despite the mixture is a single-channel recording. Underlying all factorization algorithms in SCBSS problems are the principal difficulties in estimating the adequate number of latent components for each source and in preventing the same source from being extracted more than once at the output. This chapter addresses these issues by developing a framework for pruning unnecessary components and incorporating a modified multivariate rectified Gaussian prior information into the spectral basis features.

The parameters of the imitated stereo model are estimated via the proposed PARAFAC regularized nonnegative tensor factorization with Itakura-Saito divergence. In addition, we have derived the separability conditions of the proposed mixture model and demonstrated that the proposed method can indeed separate mixtures of real-audio sources. Experimental testing on real-audio sources has been conducted to verify the capability of the proposed method.

The chapter is organized as follows: Section 5.1 summarizes the imitated-stereo mixing model. The proposed algorithm is fully developed in Section 5.2. Experimental results coupled with a series of performance comparison with other SCBSS method are presented in Section 5.3. Finally, Section 5.4 concludes this chapter.

5.1 Single Channel Mixing Model

5.1.1 Imitated – Stereo Mixture Model

The single-channel blind source separation problem can be expressed as

$$y_1(t) = x_1(t) + x_2(t) + \cdots + x_{N_s}(t) \tag{5.1}$$

where $y_1(t)$ denotes the single observed mixture, $x_j(t)$ denotes the jth source signal, N_s, is the total number of source signals and $t = 1,2,\ldots,T$ denotes the time index. To recover the original signals $x_j(t)$ given only by the sole observed mixture $y_1(t)$, we compose another mixture based on the autoregressive (AR) process of the sources. This idea has been motivated by the most of audio signals that can be modeled by the AR process. This enables us to propose the imitated mixture by time-shifting and weighting the observed mixture as

$$y_2(t) = \frac{1}{1+|\beta|}\left(y_1(t) + \beta y_1(t-\delta)\right) \tag{5.2}$$

where $\beta \in \Re$ is the weight parameter, and δ is the time-delay. The AR process of the signal can be expressed [86] as

$$x_j(t) = -\sum_{z=1}^{M_j} c_{x_j}(z;t)x_j(t-z) + e_j(t) \tag{5.3}$$

where M_j is the maximum AR order, z is the number of AR order, $c_{x_j}(z;t)$ denotes the zth order AR coefficient of the jth source signal at time t and $e_j(t)$ is an independent identically distributed (i.i.d.) random signal with variance σ^2 and zero mean. We establish a ‘imitated-stereo’ term in the mixing model in (5.1) and (5.2) since the mixing model resembles a stereo signal where the attenuation of the sources differs but the sources have identical time delay because given by one location. By using the AR process in (5.3), the imitated mixture can be rewritten in terms of the sources, its coefficients and time-delay as

$$y_2(t) = \sum_{j=1}^{N_s} \frac{\left(-c_{x_j}(\delta)+\beta\right)x_j(t-\delta)}{1+|\beta|} + \frac{-\sum_{\substack{z=1\\ z\neq\delta}}^{M_j} c_{x_j}(z;t)x_j(t-z)+e_j(t)}{1+|\beta|} \tag{5.4}$$

The proposed mixing model in terms of the sources can now be expressed concisely as a function of time

$$\begin{aligned} y_1(t) &= \sum_{j=1}^{N_s} x_j(t) \\ y_2(t) &= \sum_{j=1}^{N_s} a_j x_j(t-\delta) + r_j(t) \end{aligned} \tag{5.5}$$

where $a_j(t;\delta,\beta)$ and $r_j(t;\delta,\beta)$ represent the mixing attenuation and residue of the j^{th} source, respectively.

$$a_j(t;\delta,\beta) = \frac{-c_{x_j}(\delta;t)+\beta}{1+|\beta|} \tag{5.6}$$

$$r_j(t) = \frac{-\sum_{\substack{z=1 \\ z\neq\delta}}^{M_j} c_{x_j}(z;t)x_j(t-z)+e_j(t)}{1+|\beta|} \tag{5.7}$$

Note that the parameterization of $a_j(t)$ and $r_j(t)$ depends on δ and β although this is not shown explicitly. By comparing with the single channel mixture, the imitated stereo mixture $y_2(t)$ contains extra information i.e. $a_j(t),\delta,r_j(t)$ which are used for estimating the sources. For time-frequency (TF) representation, the mixing model can be expressed for $\forall(f,t_s)$ as

$$Y_1(f,t_s) = \sum_{j=1}^{N_s} X_j(f,t_s)$$

$$Y_2(f,t_s) = \sum_{j=1}^{N_s}\left(a_j(t_s)e^{-i2\pi f\delta}X_j(f,t_s-\delta) - \sum_{\substack{z=1 \\ z\neq\delta}}^{M_j}\frac{c_{x_j}(z;t_s)}{1+|\beta|}e^{-i2\pi fz}X_j(f,t_s-z)\right) \tag{5.8}$$

where $Y_1(f,t_s)$ and $Y_2(f,t_s)$ are obtained using the STFT of $y_1(t)$ and $y_2(t)$, respectively. The term $X_j(f,t_s)$ is the j$^{\text{th}}$ source in time-frequence domain. In (5.8), we use the fact that $e_j(t) \ll x_j(t)$, and hence the TF of $r_j(t)$ in (5.7) becomes

$$R_j(f,t_s) = -\sum_{\substack{z=1 \\ z\neq\delta}}^{M_j}\frac{c_{x_j}(z;\tau)e^{-i2\pi fz}}{1+|\beta|}X_j(f,t_s-z) \tag{5.9}$$

From (5.8), it can be seen that the imitated-stereo mixture comprises of $a_j(t_s)e^{-i2\pi f\delta}$ and $X_j(f,t_s-\delta)$. A careful analysis of (5.8) will reveal that even if $X_j(f,t_s)$ is unknown, the signature of each source can be extracted directly from $Y_1(f,t_s)$ using only information of $a_j(t_s)e^{-i2\pi f\delta}$. Care must be exercised in selecting the time-delay δ in the imitated-stereo (5.2). The factor $e^{-i2\pi f\delta}$ is only uniquely specified if $|2\pi f\delta|<\pi$, otherwise this would cause phase-wrap. Selecting improper time-delay δ will lead to phase-wrap if the maximum frequency of the source is exceeded. In order to avoid phase ambiguity, we must satisfy

$$|2\pi f_{max}\delta_{max}/f_s|<\pi \tag{5.10}$$

where δ_{max} is the maximum time delay, f_{max} is the maximum frequency present in the sources and f_s is the sampling frequency. Hence, δ_{max} can be determined from (5.10) according to

$$\delta_{max}<\frac{f_s}{2f_{max}} \tag{5.11}$$

As long as the delay parameter is less than δ_{max}, there will not be any phase ambiguity. For example, for a maximum frequency $f_{max}=3.5\,kHz$, and a sampling frequency $f_s=16\,kHz$, one obtains $\delta_{max}<2.28$ using (5.11). Therefore, phase ambiguity can be avoided provided δ is selected to be either 1 or 2. Additionally, for a maximum frequency $f_{max}=8\,kHz$ the maximum delay δ_{max} is limited to 1 only. This condition will be used to determine the range of δ in formulating the pseudo-stereo mixture.

In the proposed framework, the excitation signal for each source is filtered by a different AR filter. By comparing with the observed mixture $y_1(t)$, the imitated-stereo mixture $y_2(t)$ has extra information of the sources i.e. $a_j(t)$, δ, and $r_j(t)$. This results in a form of temporal correlation diversity of the sources in terms of the AR coefficients. It is noted in (5.7) and (5.8) that the second channel ($y_2(t)$ or equivalently $Y_2(f, t_s)$) is a mixture of the original sources and weighted by the source's temporal correlation. Thus our method in constructing the model enables this diversity to be manifested in the pair of imitated-stereo mixture as noted in $y_1(t)$ and $y_2(t)$. In addition, the residue $r_j(t)$ can be minimized by selecting the appropriate β and δ. This is the time temporal correlation diversity is proposed for solving the SCBSS problem. Our novelty of the imitated-stereo mixture has been the emergence of a new diversity in the form of sources temporal correlation within the context of SCBSS. Furthermore, the concept of temporal correlation admits a tensor representation which is then evolved into a statistical estimation problem. This enables us to treat the single-channel recording as multiple channels and subsequently allow us to develop a NTF approach for estimating the sources.

5.1.2 Method Assumptions

The proposed method focuses on separating sources from one mixture using the Wiener filtering [56]. To achieve this, the following assumptions will be used:

Assumption 1: The sources satisfy the W-disjoint orthogonality (WDO) [87] condition:

$$X_i(f,t_s)X_j(f,t_s) \approx 0, \qquad \forall i \neq j\ ,\ \forall f, t_s \tag{5.12}$$

where $X_j(f,t_s)$ is the Short-Time Fourier Transform (STFT) of $x_j(t)$ defined as

$$X_j(f,t_s) = F^W\left[x_j(t)\right](f,t_s)$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} W(t-t_s)\, x_j(t) e^{-i2\pi f t} dt \tag{5.13}$$

and $W(t)$ is the window function. The STFT is performed on the signal frame-by-frame and thus, t_s represents the window shift.

Assumption 2: The sources satisfy the local stationarity of the time-frequency representation. This refers to the approximation of $X_j(f,t_s-\phi) \approx X_j(f,t_s)$ where ϕ is the maximum time-delay (shift) associated with $F^W(\cdot)$ with an appropriate window function $W(\cdot)$. Ifϕ is small compared with the length of $W(\cdot)$ then $W(\cdot\ -\phi) \approx W(\cdot)$ [88]. Hence, the Fourier transform of a windowed function with shift ϕ yields approximately the same Fourier transform without ϕ. For the proposed method, the imitated-stereo mixture is shifted by δ and by invoking the local stationarity this leads to

$$x_j(t-\delta) \xrightarrow{STFT} e^{-i2\pi f\delta} X_j(f,t_s-\delta)$$

$$\approx e^{-i2\pi f\delta} X_j(f,t_s), \qquad \forall\delta, |\delta| \leq \phi \tag{5.14}$$

Thus, the STFT of $x_j(t-\delta)$ where $|\delta| \leq \phi$ is approximately $e^{-i2\pi f\delta}X_j(f,t_s)$ according to the local stationarity property.

5.1.3 Separability of the Imitated-Stereo Mixture Model

In this section, the imitated-stereo mixture will be examined the separability of the proposed method by considering $a_j(t)$ and $r_j(t)$. To achieve this, we assumed that the sources satisfy the WDO [87] condition:

$$X_i(f,t_s)X_j(f,t_s) \approx 0, \qquad \forall i \neq j\ ,\ \forall f,t_s \tag{5.15}$$

The imitated-stereo mixtures of different cases based on $a_j(t)$ and $r_j(t)$ are evaluated by the selected minimum function $\mathcal{L}_i$. Motivated by the separation step of the proposed algorithm, the minimum-selecting function is derived from the estimated signals in TF domain. This can be expressed by assuming that the j^{th} source dominates at a particular TF unit as

$$\begin{aligned}
\hat{X}_{ij}(f,t_s) &= \frac{\sum_{k\in K_l} q_{ik} d_{fk} h_{kt_s}}{\sum_l q_{ik} \sum_{k\in K_l} d_{fk} h_{kt_s}} Y_i(f,t_s) \\
&= \frac{E[Y_i(f,t_s)X_{il}^*(f,t_s)}{E[|Y_i(f,t_s)|^2]} \sum_j m_{ij} X_{ij}(f,t_s) \\
&= \frac{E[\sum_l m_{il} X_{il}(f,t_s) X_{il}^*(f,t_s)}{E[\sum_l m_{il} X_{il}(f,t_s) \sum_l m_{il}^* X_{il}^*(f,t_s)]} \sum_j m_{ij} X_{ij}(f,t_s) \\
&= \frac{\sum_l m_{il} E[X_{il}(f,t_s) X_{il}^*(f,t_s)]}{E[\sum_l m_{il} X_{il}(f,t_s) \sum_l m_{il}^* X_{il}^*(f,t_s)]} \sum_j m_{ij} X_{ij}(f,t_s) \\
&= \frac{\sum_l m_{il} \sum_j m_{ij} E[|X_{il}(f,t_s)|^2] X_{ij}(f,t_s)}{\sum_l |m_{il}|^2 E[|X_{il}(f,t_s)|^2]}
\end{aligned} \tag{5.16}$$

If $j = l$, we then obtain $\hat{X}_{ij}(f,t_s) = \mathrm{X}_{ij}(f,t_s)$.

In this light, we formulate the proposed minimum-selecting function which can be expressed as:

$$\mathcal{L}_i = min_l \left| X_{ij}(f,t_s) - \frac{\sum_{k\in K_l} q_{1k} d_{fk} h_{kt_s}}{\sum_{l=1}^{2}\sum_{k\in K_l} q_{1k} d_{fk} h_{kt_s}} Y_i(f,t_s) \right|^2$$

$$= min_l \left| X_{ij}(f,t_s) - \hat{X}_{il}(f,t_s) \right|^2 \qquad (5.17)$$

By evaluating the minimum-selecting function, each TF unit is mark to the l^{th} argument that yields the minimum value. Hence, the TF units of the mixture are classified into l groups of (f,t_s) units. The minimum-selected function is further analyzed in the cases of the i^{th} mixture. In the first case where $i = 1$ i.e. $Y_1(f,t_s) = \sum_{j=1}^{2} X_j(f,t_s) = X_1(f,t_s) + X_2(f,t_s)$, the function $\mathcal{L}_1$ can be expressed as

$$\mathcal{L}_1 = \min_l \left| X_{1j}(f,t_s) - \frac{\sum_{k\in K_l} q_{1k} d_{fk} h_{kt_s}}{\sum_{l=1}^{2}\sum_{k\in K_l} q_{1k} d_{fk} h_{kt_s}} Y_1(f,t_s) \right|^2$$

$$= min_l \left| X_{1j}(f,t_s) - \frac{\sum_{k\in K_l} q_{1k} d_{fk} h_{kt_s}}{\sum_{l=1}^{2}\sum_{k\in K_l} q_{1k} d_{fk} h_{kt_s}} \sum_{l=1}^{2} X_l\,(f,t_s) \right|^2 \qquad (5.18)$$

Secondly, when $i = 2$ i.e. $Y_2(f,t_s) = \sum_{j=1}^{2} \overline{a}_j X_j(f,t_s) = \overline{a}_1 X_1(f,t_s) + \overline{a}_2 X_2(f,t_s)$, the function $\mathcal{L}_2$ can be expressed as

$$\mathcal{L}_2 = \min_d \left| \overline{a}_j X_{2j}(f,t_s) - \frac{\sum_{k\in K_l} q_{2k} \mathrm{d}_{fk} h_{kt_s}}{\sum_{l=1}^{2}\sum_{k\in K_l} q_{2k} d_{fk} h_{kt_s}} Y_2(f,t_s) \right|^2$$

$$= min_d \left| \overline{a}_j X_{2j}(f,t_s) - \frac{\sum_{k\in K_l} q_{2k} d_{fk} h_{kt_s}}{\sum_{l=1}^{2}\sum_{k\in K_l} q_{2k} d_{fk} h_{kt_s}} \sum_{l=1}^{2} \overline{a}_l X_l(f,t_s) \right|^2 \qquad (5.19)$$

The functions $\mathcal{L}_1$ and $\mathcal{L}_2$ will then be used for evaluating the separability of the proposed imitated-stereo mixture by considering $a_j(t)$ and $r_j(t)$ in the following three scenarios.

I: If $\forall j \;\; a_j(t) = a(t)$ and $r_j(t) = r(t)$, then $x_2(t) = \left(\frac{a(t)+\beta}{1+|\beta|}\right) x_1(t-\delta) + 2r(t)$.

The first scenario presents a situation where two identical sources are mixed in the single channel. By a weighted and time-shifting of the observed mixture, the imitated mixture is only gained the time-delayed and scalar of the first mixture. This achieves no advantage of the imitated mixture at all. The separability of this case is presented by substituting the imitated-stereo mixture of Scenario I into the functions L_1 and L_2. Since both sources are identical, the minimum-selecting function of each mixture can be evaluated as follow: For $i = 1,\; X_j(f, t_s) = X(f, t_s)\; \forall j$, the L_1 function then becomes

$$L_1 = min_l \left| X(f, t_s) - \frac{\sum_{\forall k} q_{1k} w_{fk} h_{kt_s}}{2\sum_{\forall k} q_{1k} w_{fk} h_{kt_s}} 2X(f, t_s) \right|^2$$

$$= min_l |X(f, t_s) - X(f, t_s)|^2$$

$$= 0 \;\text{ for } \forall l \tag{5.20}$$

For $i = 2$, $a_j(t)$ and $r_j(t)$ are related to the source via $\overline{a}_j$, thus $\overline{a}_j X_j(f, t_s) = \overline{a}\, X(f, t_s)\; \forall j$. Thus the L_2 function becomes:

$$L_2 = min_l \left| \overline{a}\, X(f, t_s) - \frac{\sum_{\forall k} q_{1k} w_{fk} h_{kt_s}}{2\sum_{\forall k} q_{1k} w_{fk} h_{kt_s}} 2\overline{a}\, X(f, t_s) \right|^2$$

$$= min_l |\overline{a}\, X(f, t_s) - \overline{a}\, X(f, t_s)|^2$$

$$= 0 \quad \text{for } \forall l \tag{5.21}$$

As a result, the both minimum-selecting function are zero for all l^{th} arguments i.e. $L_1 = L_2 = 0$. In this case, the function cannot discriminate the lth arguments, the mixture is not separable.

II: If $\forall j$: $a_j(t) = a(t)$ and $r_j(t) \neq r_k(t)$ for $j \neq k$ then $x_2(t) = \left(\frac{a(t)+\beta}{1+|\beta|}\right) x_1(t-\delta) + r_1(t) + r_2(t)$.

Scenario II represents different sources but setting β and δ for the imitated-stereo mixture such that $a_1(t) = \cdots = a_{N_s}(t)$. By following the steps in Case 1, the separability of this mixture can be analyzed using the functions L_1 and L_2 as

$$L_1 = min_l \left| X_{1j}(f,t_s) - \frac{\sum_{k\in K_l} q_{1k} w_{fk} h_{kt_s}}{\sum_{l=1}^{N_s} \sum_{k\in K_l} q_{1k} w_{fk} h_{kt_s}} \sum_{l=1}^{N_s} X_l\,(f,t_s) \right|^2$$

$$= min_l \left| X_{1j}(f,t_s) - X_l(f,t_s) \right|^2 \tag{5.22}$$

Since $r_j(t) \neq r_k(t)$ thus $\overline{a}_j X_j(f,t_s) \neq \overline{a}_k X_k(f,t_s)$ for $j \neq k$, we then obtain

$$L_2 = \min_l \left| \overline{a}_j X_{2j}(f,t_s) - \frac{\sum_{k\in K_l} q_{2k} w_{fk} h_{kt_s}}{\sum_{l=1}^{N_s} \sum_{k\in K_l} q_{2k} w_{fk} h_{kt_s}} \sum_{l=1}^{N_s} \overline{a}_l X_l(f,t_s) \right|^2$$

$$= min_l \left| \overline{a}_j X_{2j}(f,t_s) - \overline{a}_l X_l(f,t_s) \right|^2 \tag{5.23}$$

As a result of $j = l$, the both L_1 and L_2 functions yields a zero value. The minimum-selecting functions are capable to separate the lth arguments although the sources have the same mixing attenuation; $a_1(t) = \cdots = a_{N_s}(t) = a(t)$. Therefore, the mixture of Scenario II is separable.

III: If $a_j(t) \neq a_k(t)$ and $r_j(t) \neq r_k(t)$ for $j \neq k$ then

$$x_2(t) = \sum_{j=1}^{N_s} \left(\frac{a_j(\delta)+\beta}{1+|\beta|}\right) x_j(t-\delta) + r_j(t)$$

This scenario corresponds to the most general case where the sources are distinct, and β and δ are determined arbitrarily such that the mixing attenuations and residues are also different. The L_1 function is firstly treated where the original signals differ i.e. $X_j(f, t_s) \neq X_k(f, t_s)$. Hence, the L_1 function of Scenario III obtains the same as Scenario II in (5.22) i.e. $L_1 = min_l \left|X_{1j}(f, t_s) - X_l(f, t_s)\right|^2$.

Since the mixing attenuations $a_j(t_s)$ and $a_k(t_s)$ correspond respectively to $x_j(t)$ and $x_k(t)$, thus $\overline{a}_j X_j(f, t_s) \neq \overline{a}_k X_k(f, t_s)$ and $r_j(t) \neq r_k(t)$. By following similar line of the L_2 function in Scenario II, we then have

$$L_2 = min_l \left|\overline{a}_j X_{2j}(f, t_s) - \overline{a}_l X_l(f, t_s)\right|^2 \qquad (5.24)$$

For $j \neq l$, the L_1 and L_2 functions in Scenario III render a non-zero value. Hence, this mixture can be separated by the minimum-selecting function.

5.2 Proposed Separation Method

5.2.1 Separation Model

The proposed method aim to estimate the original signals $\left[x_1(t) x_2(t) \cdots x_{N_s}(t)\right]^T$ by formulating an imitated stereo mixture and using the proposed method given only one observed mixture, $y_1(t)$. The process of the proposed method is illustrated in Figure 5.1.

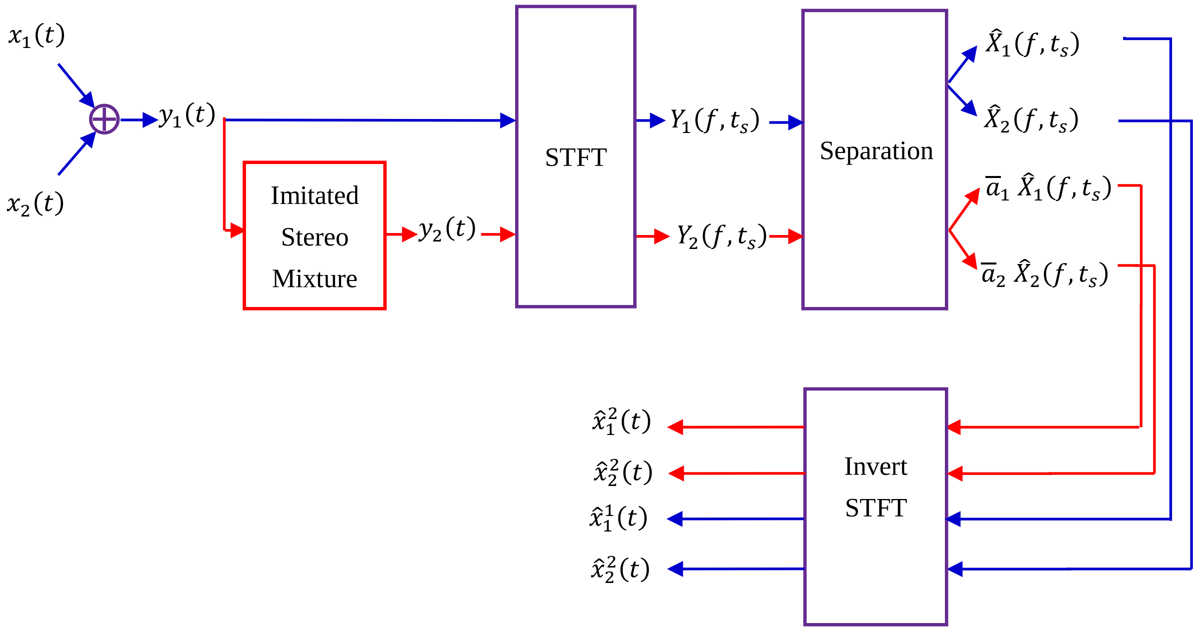


Figure 5.1: The process of the proposed method for $N_s = 2$.

5.2.2 Formulation of the Proposed Algorithm

In order to formulate the proposed algorithm, we choose a prior distribution $p(\mathbf{W},\mathbf{H})$ over the factors $\{\mathbf{W},\mathbf{H}\}$. It can be shown that the following optimization problem needs to be solved

$$min_{\mathbf{Q},\mathbf{W},\mathbf{H}}\, C_{MAP}\,(\mathbf{W},\mathbf{H}) \doteq -log\, p(\mathbf{W},\mathbf{H}|\mathbf{Y},\boldsymbol{\lambda},\mathbf{Q}) \tag{5.25}$$

where

$$log\, p(\mathbf{W},\mathbf{H}|\mathbf{Y},\boldsymbol{\lambda},\mathbf{Q}) = log\, p(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{Q}) + log\, p(\mathbf{W},\mathbf{H}|\boldsymbol{\lambda}) + c$$

and

$$-log\, p(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{Q}) = \sum_{ift_s} \frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)} - log\,\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)} - 1 \tag{5.26}$$

Similarly (3.1) to (3.24), the negative log prior on $\mathbf{W}$ is defined as

$$-\,log\, p(\mathbf{W}) \doteq -\,log\,\delta(w) + \tfrac{1}{2}\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\Omega}_{diag}^{W}\boldsymbol{w} - \tfrac{1}{2}\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{\Omega}_{off}^{W}\boldsymbol{w} \tag{5.27}$$

Analyzing the above, the second term $\boldsymbol{w^T\Omega}^W_{diag}\boldsymbol{w} = \sum_k \boldsymbol{w}_k^{\boldsymbol{T}}\mathbf{P}_k^{-1}\boldsymbol{w}_k$ where $\mathbf{P}_k^{-1}$ is a Toeplitz matrix corresponding to the kth diagonal sub-matrix of $\boldsymbol{\Omega}^W_{diag}$. Since the source signals are modelled as AR processes, it is natural that $\mathbf{P}_k$ assumes the AR autocorrelation matrix of the following form[1]:

$$\mathbf{P}_k = \sigma_k^2 \begin{bmatrix} 1 & \rho_k & \cdots & \rho_k^{F-1} \\ \rho_k & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_k \\ \rho_k^{F-1} & \cdots & \rho_k & 1 \end{bmatrix} \tag{5.28}$$

where ρ_k is the first-order correlation of $\boldsymbol{w}_k$. For the third term, we note that $\boldsymbol{\Omega}^W_{off,k,n} \overset{\text{def}}{=} \boldsymbol{\Sigma}_{k,k}^{-1(W)}\boldsymbol{\Sigma}^W_{k,n}\boldsymbol{\Sigma}_{n,n}^{-1(W)} = \mathbf{P}_k^{-1}\boldsymbol{\Sigma}^W_{k,n}\mathbf{P}_n^{-1}$. Since the elements in $\mathbf{P}_k$ are exponentially decaying, we can make a crude approximation that $\boldsymbol{P}_k^{-1}\boldsymbol{\Sigma}^W_{k,n}\boldsymbol{P}_n^{-1} \cong \mu_{kn}\boldsymbol{I}$ where $\mu_{kn} = \sigma_k^{-2}\sigma_n^{-2}c_{k,n}$ and $c_{k,n}$ is the correlation between the k^{th} and n^{th} basis vectors. Thus the term $\boldsymbol{w^T\Omega}^W_{off}\boldsymbol{w} = \sum_{k,n,(k\neq n)} \mu_{kn}\boldsymbol{w}_k^{\boldsymbol{T}}\boldsymbol{w}_n$ measures the sum of weighted correlation between $\boldsymbol{w}_k$ and $\boldsymbol{w}_n$ for all $k, n, (k \neq n)$. Hence, by including both of these terms, the underlying statistical correlation within and between the basis vectors can be incorporated into the matrix factorization to yield results that reflect on prior information of the AR sources. Therefore, with the factorial model in (5.27) the desired constraint assumes the following form:

$$f(\mathbf{W}) = -\,log\, p(\mathbf{W}) \doteq -\sum_k log\, \delta(\boldsymbol{w}_k) + \frac{1}{2}\sum_k \boldsymbol{w}_k^{\boldsymbol{T}}\boldsymbol{P}_k^{-1}\boldsymbol{w}_k - \frac{1}{2}\sum_{k,n,(k\neq n)} \mu_{kn}\boldsymbol{w}_k^{\boldsymbol{T}}\boldsymbol{w}_n \tag{5.29}$$

The use of multivariate rectified Gaussian prior $p(\mathbf{W})$ enables the matrix factorization

[1] In practice, $|\rho_k| \ll 1$ and the terms associated with ρ_k^n $(n > 1)$ in $\boldsymbol{P}_k$ decay exponentially. Thus, in implementation $\boldsymbol{P}_k$ can assume a symmetric tri-diagonal matrix and the explicit inverse of such matrix is well documented in the literature.

to leverage on the statistical first order AR correlation within and between the basis vectors. Once the basis $\boldsymbol{w}_k$ has successfully extracted a particular spectral basis associated with a source signal, subsequent basis vectors $\{\boldsymbol{w}_{j\neq k}\}$ will leverage on $\boldsymbol{w}_k$ to extract other spectral components of the same source. However, care must be exercised in order that the basis vectors do not extract the same spectral component. Thus this necessitates us to monitor the correlation between the basis vectors i.e. μ_{kn}, and as this value gets larger, the more imperative it is to introduce pruning to prevent the basis vectors from extracting the same spectral component. This will be elaborated in Section 5.2.2. 2). In order to turn off excess components thereby optimizing K, we choose a component-wise exponential distribution prior is imposed on $\mathbf{H}$, namely,

$$p(\mathbf{H}|\boldsymbol{\lambda}) = \prod_k \prod_{t_s} \lambda_{kt_s} exp\,(-\lambda_{kt_s} h_{kt_s}) \tag{5.30}$$

Following (5.30), the negative log prior on $\mathbf{H}$ is defined as

$$f(\mathbf{H}) = -\,log\,p(\mathbf{H}|\boldsymbol{\lambda}) = -\sum_k \sum_{t_s} \left\{ log\,\lambda_{kt_s} - \lambda_{kt_s} h_{kt_s} \right\}$$

$$= -\sum_k \sum_{t_s} log\,\lambda_{kt_s} + \sum_k \sum_{t_s} \lambda_{kt_s} h_{kt_s} \tag{5.31}$$

By substituting (5.26) and (5.30) into (5.25), the negative log posterior of $\mathbf{W}$ and $\mathbf{H}$ is given by the following:

$$-\,log\,p(\mathbf{W},\mathbf{H}|\mathbf{Y},\boldsymbol{\lambda},\mathbf{Q}) \doteq -\,log\,p(\mathbf{Y}|\mathbf{W},\mathbf{H},\mathbf{Q}) - log\,p(\mathbf{W}) - log\,p(\mathbf{H}|\boldsymbol{\lambda}) \tag{5.32}$$

From (5.29) and (5.31), the above can be written as

$$L \doteq \sum_{ift_s} d_{IS}\left(V_i(f,t_s)\middle|\hat{V}_i(f,t_s)\right) + f(\mathbf{W}) + f(\mathbf{H})$$

$$= \sum_{ift_s} \frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)} - log\frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)} - 1 - \sum_k log\,\delta(\boldsymbol{w}_k) + \frac{1}{2}\sum_k \boldsymbol{w}_k^{\boldsymbol{T}}\boldsymbol{P}_k^{-1}\boldsymbol{w}_k$$

$$-\frac{1}{2}\sum_{k,j,(k\neq n)} \mu_{kn}\boldsymbol{w}_k^{\boldsymbol{T}}\boldsymbol{w}_n - \sum_k\sum_{t_s} log\,\lambda_{kt_s} + \sum_k\sum_{t_s}\lambda_{kt_s}h_{kt_s} \tag{5.33}$$

The sparsity term $\sum_k\sum_{t_s}\lambda_{kt_s}h_{kt_s}$ forms the L_1-norm regularization to resolve the permutation ambiguity by forcing all structure in **H** onto **W**. Therefore, the sparseness of the solution in (5.33) is highly dependent on the regularization parameter λ_{kt_s}.

5.2.2.1 Estimation of the mixing coefficient, basis and code

In this section, we will derive the estimation of **W**, **H** and $\mathbf{P} = \left\{\left|a_{ij}\right|^2\right\}$. The derivative of (5.33) with respect to **W** of the proposed model is given by:

$$\frac{\partial L}{\partial w_{fk}} = \sum_{it_s} q_{ik}h_{kt_s}d'_{IS}\left(V_i(f,t_s)|\hat{V}_i(f,t_s)\right) + \sum_n p_{k,fn}w_{fn} - \sum_{n\neq k}\mu_{kn}w_{fn}$$
$$-\delta'\left(w_{fk}\right)/\delta\left(w_{fk}\right) \tag{5.34}$$

where $p_{k,fn}$ is the $(f,n)^{th}$ component of the $\mathbf{P}_k^{-1}$ matrix.

Similarly, the derivative of (5.33) with respect to **H** is given by

$$\frac{\partial L}{\partial h_{kt_s}} = \sum_{it_s} q_{ik}w_{fk}d'_{IS}\left(V_i(f,t_s)|\hat{V}_i(f,t_s)\right) + \lambda_{kt_s} \tag{5.35}$$

The derivative of (5.33) with respect to $\mathbf{P} = \left\{ \left| a_{ij} \right|^2 \right\}$ is given by

$$\frac{\partial L}{\partial p_{ij}} = \sum_k l_{jk} \sum_{f,t_s} w_{fk} h_{kt_s} d'_{IS}\left(V_i(f,t_s) | \widehat{V}_i(f,t_s) \right) \tag{5.36}$$

We define the term $\boldsymbol{G}$ is $I \times F \times T_s$ tensor with entries $g_{ift_s} = d'_{IS}(V_i(f,t_s)|\widehat{V}_i(f,t_s))$, namely

$$d'_{IS}\left(V_i(f,t_s) \middle| \widehat{V}_i(f,t_s) \right) = \frac{1}{\widehat{V}_i(f,t_s)} - \frac{V_i(f,t_s)}{\widehat{V}_i(f,t_s)^2} \tag{5.37}$$

We note $\langle \overline{\boldsymbol{A}}, \overline{\boldsymbol{B}} \rangle_{k_{\overline{\boldsymbol{A}}}, k_{\overline{\boldsymbol{B}}}}$ the contracted product between tensors $\overline{\boldsymbol{A}}$ with size $I_1 \times \ldots \times I_M \times J_1 \times \ldots \times J_N$ and $\overline{\boldsymbol{B}}$ with size $I_1 \times \ldots \times$ $I_M \times K_1 \times \ldots \times K_P$ and $k_{\overline{\boldsymbol{A}}}$ and $k_{\overline{\boldsymbol{B}}}$ are the sets of mode indices over which the summation take place. The contracted product $\langle \overline{\boldsymbol{A}}, \overline{\boldsymbol{B}} \rangle_{\{1,\ldots,M\},\{1,\ldots,M\}}$ is a tensor of size $J_1 \times \ldots \times J_N \times K_1 \times \ldots \times K_P$ given by

$$\langle \overline{\boldsymbol{A}}, \overline{\boldsymbol{B}} \rangle_{\{1,\ldots,M\},\{1,\ldots,M\}} = \sum_{i_1=1}^{I_1} \cdots \sum_{i_M}^{I_M} \bar{a}_{i_1,\ldots,i_M,j_1,\ldots,j_N} \bar{b}_{i_1,\ldots,i_M,k_1,\ldots,k_P} \tag{5.38}$$

The contracted tensor product is a form a generalized dot product of two tensors along common modes of same dimensions. Using (5.38), the multiplicative (**MU**) learning rules in matrix notation for **W**, **H**, and **P** become

$$\mathbf{W} \leftarrow \mathbf{W} \bullet \frac{\langle \boldsymbol{G}_{-}, \mathbf{Q} \circ \mathbf{H} \rangle_{\{1,3\},\{1,2\}}}{\langle \boldsymbol{G}_{+}, \mathbf{Q} \circ \mathbf{H} \rangle_{\{1,3\},\{1,2\}} + \left[\delta'(\boldsymbol{W}) ./ \delta(\boldsymbol{W}) \right] + \boldsymbol{W} \boldsymbol{\Xi}^{\boldsymbol{T}}}$$

$$\mathbf{H} \leftarrow \mathbf{H} \bullet \frac{\langle \boldsymbol{G}_{-}, \mathbf{Q} \circ \mathbf{W} \rangle_{\{1,2\},\{1,2\}}}{\langle \boldsymbol{G}_{+}, \mathbf{Q} \circ \mathbf{W} \rangle_{\{1,2\},\{1,2\}} + \boldsymbol{\lambda} \mathbf{1}^{\boldsymbol{T}}}$$

$$\mathbf{P} \leftarrow \mathbf{P} \bullet \frac{\langle \boldsymbol{G}_{-}, \mathbf{W} \circ \mathbf{H} \rangle_{\{2,3\},\{1,2\}} \, \boldsymbol{L}^{\boldsymbol{T}}}{\langle \boldsymbol{G}_{+}, \mathbf{W} \circ \mathbf{H} \rangle_{\{2,3\},\{1,2\}} \, \boldsymbol{L}^{\boldsymbol{T}}} \tag{5.39}$$

which has a strikingly similar form with the conventional NMF update rules. In (5.39), '•' is element-wise product and $\boldsymbol{\Xi^T}$ is a $K \times K$ matrix whose $(k, n)^{th}$ element is given by $P_{k,fn}$ and μ_{kn}. Here $\mathbf{G}_{-}$ follows the MU rule that denotes the negative part of the derivative of the criterion e.g. $\mathbf{G}_{-} = \left[d'_{IS}\left(V_i(f,t_s)\middle|\hat{V}_i(f,t_s)\right)\right]_{-} = \frac{V_i(f,t_s)}{\hat{V}_i(f,t_s)^2}$ and $\boldsymbol{G}_{+}$ denotes its positive part. The term $\mathbf{Q} \circ \mathbf{H}$ denotes $I \times K \times T_s$ tensor with elements $q_{ik}h_{kt_s}$. Similarly, $\mathbf{Q} \circ \mathbf{W}$ denotes $F \times I \times K$ tensor with elements $q_{ik}w_{fk}$ and $\mathbf{W} \circ \mathbf{H}$ denotes $F \times K \times T_s$ tensor with elements $w_{fk}h_{kt_s}$.

5.2.2.2 Estimation of the Adaptive Sparsity Parameter

The update of λ_k follows by solving $\frac{\partial L}{\partial \lambda_{kt_s}} = 0$, this gives

$$\frac{\partial L}{\partial \lambda_{kt_s}} = \frac{1}{\lambda_k} - h_{kt_s}$$

$$\lambda_{kt_s} = \frac{1}{h_{kt_s}} \tag{5.40}$$

However, this may cause abrupt changes in the level of sparsity. An adaptive first-order implementation that smooth over time can be obtained as follows:

$$\lambda_{kt_s}(t) = \alpha\lambda_{kt_s}(t-1) + (1-\alpha)\frac{1}{h_{kt_s}+\epsilon} \tag{5.41}$$

where α is the smoothing parameter and is normally set to 0.95, and $\epsilon = 10^{-9}$ is a small number to prevent division by zero. As mentioned in Section 5.2.2., pruning is exercised to prevent the basis vectors from extracting the same spectral component. First, note that the sparsity term $\sum_k \sum_{t_s} \lambda_{kt_s} h_{kt_s}$ forms the sparse NTF objectives while the normalization term $\sum_k \sum_{t_s} log\, \lambda_{kt_s}$ are given to learn the degree of regularization from data, i.e. tune the pruning parameter, λ_{kt_s}. Second, let us assume that the factorization in (5.29) has an approximation error of $\sum_{i=1}^{I} \sum_{f=1}^{F} \sum_{t_s=1}^{T_s} |Y_i(f,t_s)|^2 / IFT_s$. As a result of inference in (5.33), a subset of the λ_{kt_s} will be driven to a large upper bound, with the corresponding columns of $\mathbf{W}$ and rows of $\mathbf{H}$ driven to small values. The effective dimensionality can be deduced from the distribution of the λ_{kt_s}. We have found in practice, two clusters clearly emerge: A group of values in same order of magnitude corresponding to relevant components on columns of $\mathbf{W}$ and rows of $\mathbf{H}$, and a group of similar values of much higher magnitude corresponding to irrelevant components. Furthermore, for components which had become to zero or close to zero we set $\lambda_{kt_s} = \frac{1}{\epsilon}$. Thus, based on the above empirical observation, we propose the following pruning threshold: Let $\bar{\lambda}_k \triangleq \frac{1}{T_s} \sum_{t_s=1}^{T_s} \lambda_{kt_s}$ be the average sparseness value associated with the k^{th} row of $\mathbf{H}$. If

$$\bar{\lambda}_k > \epsilon^{-1} \cdot \sqrt{\frac{\sum_{i=1}^{I} \sum_{f=1}^{F} \sum_{t_s=1}^{T_s} |Y_i(f,t_s)|^2}{IFT_s}} \qquad (5.42)$$

then the kth row of $\mathbf{H}$ (equivalently kth column of $\mathbf{W}$) is to be removed. This method allows us to estimate the effective number of component. If the prior assumptions are slightly violated or even if the likelihood function differs from the model assumption, the correct factorization rank can be determined by evaluating the above bound by the pruning threshold.

5.2.2.3 Estimation of Source Signals

For the proposed method, we obtain the estimates of **W**, **H** and **P** that yield the smallest cost value. To reconstruct the source signals, the term $\hat{C}_{ik}(f,t_s)$ of the component k in channel i is reformulated by using the Wiener filtering [56] as

$$\begin{aligned}\hat{C}_{ik}(f,t_s) &\stackrel{\text{def}}{=} E\{C_{ik}(f,t_s)|\mathbf{P},\mathbf{W},\mathbf{H},\mathbf{Y}\}\\ &= \frac{q_{ik}w_{fk}h_{kt_s}}{\sum_{k=1}^{K} q_{ik}w_{fk}h_{kt_s}} Y_i(f,t_s)\\ &= \frac{q_{ik}w_{fk}h_{kt_s}}{\hat{V}_i(f,t_s)} Y_i(f,t_s)\end{aligned} \tag{5.43}$$

where $\hat{V}_i(f,t_s) = \sum_{k=1}^{K} q_{ik}w_{fk}h_{kt_s}$. The decomposition is conservative in the sense that it satisfies

$$y_i(t_s) = \sum_{k=1}^{K} \hat{c}_{ik}(t_s) \tag{5.44}$$

The estimated sources are reconstructed by using inverse-STFT of $\hat{C}_{ik}(f,t_s)$ for all i and k leads to a set of time-domain components $\{\hat{c}_1(t),\dots,\hat{c}_K(t)\}$, with

$$\hat{c}_k(t) = \begin{bmatrix} \hat{c}_{ik}(t) \\ \vdots \\ \hat{c}_{Ik}(t) \end{bmatrix} \tag{5.45}$$

and sources estimates can be obtained as

$$\hat{x}_{ij}(t) = \sum_{k\in K_j} \hat{c}_{ik}(t) \tag{5.46}$$

The proposed algorithm is summarized in Table 5.1.

Table 5.1: Overview proposed algorithm of ISM-RNTF.

1. Generate the mixture $y_2(t)$ from (5.2) and compute the STFT of $Y_1(f,t_s)$ and $Y_2(f,t_s)$.
2. Apply spectral subtraction on $|Y_i(f,t_s)|^2$, $i = 2,\dots,N_s$.
3. Initialize $\mathbf{W}$, $\mathbf{H}$ and $\mathbf{P}$ with nonnegative random values and define $\mathbf{L} = \{l_{jk}\}$, $l_{jk} = \begin{cases} 1, if\ k \in K_j \\ 0, otherwise \end{cases}$ and $\mathbf{Q} = \mathbf{PL}$.
4. Compute the followings:
   - power spectrogram $V_i(f,t_s) = |Y_i(f,t_s)|^2$,
   - $\hat{V}_i(f,t_s) = \sum_{k=1}^{K} q_{ik} w_{fk} h_{kt_s}$
   - $\boldsymbol{G}_+$ and $\boldsymbol{G}_-$ according to (5.46).
   - ρ_k and μ_{kn}
5. Update model parameters as follows:

$$\mathbf{W} \leftarrow \mathbf{W} \bullet \frac{\langle \boldsymbol{G}_-, \mathbf{Q} \circ \mathbf{H} \rangle_{\{1,3\},\{1,2\}}}{\langle \boldsymbol{G}_+, \mathbf{Q} \circ \mathbf{H} \rangle_{\{1,3\},\{1,2\}} + [\delta'(\mathbf{W})./\delta(\mathbf{W})] + \mathbf{W}\boldsymbol{\Xi}^{\boldsymbol{T}}}$$

$$\mathbf{H} \leftarrow \mathbf{H} \bullet \frac{\langle \boldsymbol{G}_-, \mathbf{Q} \circ \mathbf{W} \rangle_{\{1,2\},\{1,2\}}}{\langle \boldsymbol{G}_+, \mathbf{Q} \circ \mathbf{W} \rangle_{\{1,2\},\{1,2\}} + \boldsymbol{\lambda} \mathbf{1}^{\boldsymbol{T}}}$$

$$\mathbf{P} \leftarrow \mathbf{P} \bullet \frac{\langle \boldsymbol{G}_-, \mathbf{W} \circ \mathbf{H} \rangle_{\{2,3\},\{1,2\}} \boldsymbol{L}^{\boldsymbol{T}}}{\langle \boldsymbol{G}_+, \mathbf{W} \circ \mathbf{H} \rangle_{\{2,3\},\{1,2\}} \boldsymbol{L}^{\boldsymbol{T}}}$$

$$\boldsymbol{\lambda} = \alpha \boldsymbol{\lambda} + (\mathbf{1} - \alpha) \frac{\mathbf{1}}{\mathbf{H} + \epsilon} \quad , \quad \bar{\lambda}_k = \frac{1}{T_s} \boldsymbol{\delta}_k^{\boldsymbol{T}} \boldsymbol{\lambda}\, \mathbf{1}$$

6. Prune the irrelevant components of $\mathbf{W}$ and $\mathbf{H}$ using the criteria (5.42). Normalize $\mathbf{W}$ and $\mathbf{H}$.
7. Repeat steps 5 to 7 until it converges or reaches the predefine number of iteration.
8. Formulate $\hat{C}_{ik}(f,t_s) = \frac{q_{ik} w_{fk} h_{kt_s}}{\hat{V}_i(f,t_s)} Y_i(f,t_s)$.
9. Convert $\hat{C}_{ik}(f,t_s)$ from TF domain into time domain $\hat{c}_{ik}(t)$ and reconstruct

components and sources using $\hat{x}_{ij}(t) = \sum_{k\in K_j} \hat{c}_{ik}(t)$.

5.3 Results and Analysis

5.3.1 Experiment Setup

The proposed ISM-RNTF method is demonstrated by separating real-audio sources. The real-audio sources which are inherently non-stationary include vocal and music signals. All experiments are conducted using a PC with Intel® Core™ i5 CPU 650 at 3.2 GHz and 4 GB RAM. MATLAB is used as the programming platform. The TF representation is computed by using the STFT of 1024-point Hanning window with 50% overlap. The experiments consist of 7 type of mixtures are generated i.e. male speech + female speech, male speech + jazz, male speech + drum, male speech + piano, jazz + drum, and drum + piano. The male speech, female speech and music sources are selected from the RWC [85] database and 3 linear instantaneous stereo mixtures of 3 sources taken from the Signal Separation Evaluation Campaign (SiSEC 2010) "Underdetermined speech and music mixtures" task development dataset [76]. Three audio datasets have been considered and are described as: 1) wdrums, a linear instantaneous stereo mixture (with positive mixing coefficients) of 2 drum sources and 1 bass line. 2) nodrums, a linear instantaneous stereo mixture (with positive mixing coefficients) of 1 rhythmic acoustic guitar, 1 electric lead guitar and 1 bass line. It also coincides with development dataset dev2 of SiSEC'08 "underdetermined speech and music mixtures" task. Both mixtures are 10 seconds-long and sampled at 16 kHz. The instantaneous mixing is characterized by static positive gains. We applied a STFT with sine bell of length 64 ms (1024 samples) leading to $F = 513$. 3)

Shannonsongs Sunrise, a linear instantaneous stereo mixture of $d_{max} = 3$ musical sources (drums, lead vocals and piano) created using 3.12 seconds-excerpts of original separated tracks from the song "Sunrise" by S. Hurley and downsampled to 16 kHz. We have evaluated our separation performance by measuring the distortion between original source and the estimated one according to the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR) and source-to-artifacts ratio (SAR). MATLAB routines for computing these criteria are obtained from the SiSEC'08 webpage [77 79]. The proposed method will be compared with 1) the other SCBSS method as the sparse nonnegative matrix 2-dimensional factorization (SNMF2D) [91] and the single-channel independent component analysis (SCICA) [24]. The SNMF2D parameters are set as follows [92]: number of factors is 2, sparsity weight of 1.1, number of phase shift and time shift is 31 and 7, respectively for music. As for speech, both shifts are set to 4. The TF domain used in SNMF2D is based on the log-frequency spectrogram. Cost function of SNMF2D is based on the Kullback-Leibler divergence. As for the SCICA, the number of block is 10 with time delay set to unity, 2) the other NTF-based multichannel method and NTF-based SCBSS.

5.3.2 Single Channel Mixture

5.3.2.1 Single Channel Sources

In this section, Audio sources can be characterized as non-stationary AR processes since their AR coefficients vary with time [93]. We have generated the mixtures of two sources which select from male speech, female speech, jazz, piano and drum. Both sources are mixed with equal power to generate the mixture. This is shown in the first three panels of Figure 5.2. Figure 5.2 shows the time-domain separation results of jazz and drum mixtures. We have investigated the separation performance by using $\beta = 1$ and $\delta = 1$. The estimated sources are plotted in the last two panels of Figure 5.2. Visually, the estimated sources resemble closely to the original sources. The mean square error (MSE) between the original and the estimated music is 0.08dB and 0.16dB for jazz and drum, respectively.

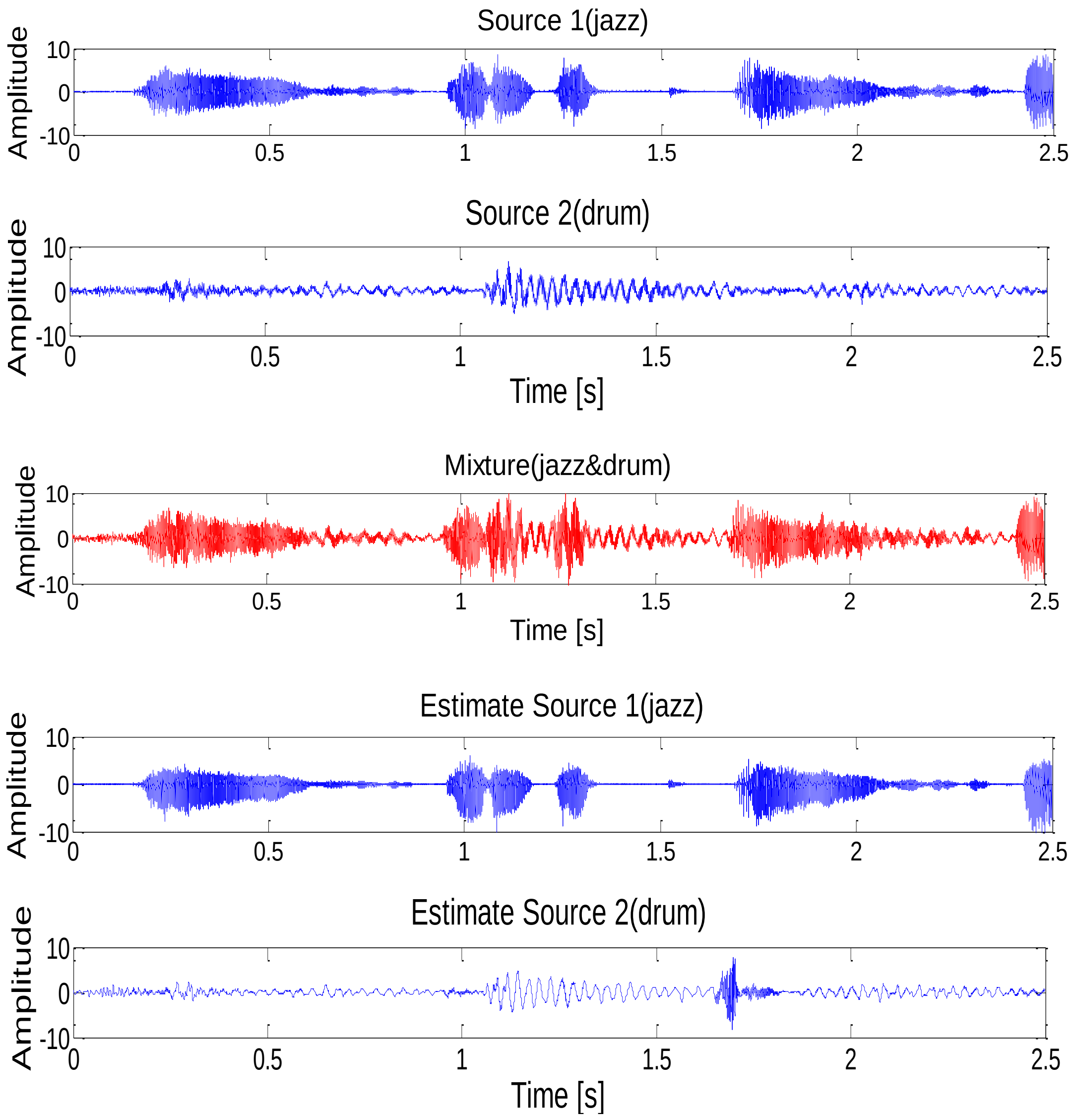


Figure 5.2: Original sources, single channel mixture, and estimated sources of music mixture between jazz and drum using the proposed method with $\beta = 1$ and $\delta = 1$.

Figure 5.3 illustrates the results between the SCICA and the SNMF2D method and the proposed method. In comparison, the average performance improvement of the proposed music mixture, the average SDR improvement is 2.6dB and 9.2dB per source, respectively; 2) for mixture of speech and music signal, the SDR improvement 4.6dB and 7.9dB per source, respectively; and 3) for speech mixture, the average SDR improvement is 2.1dB and 3.4dB per source, respectively. Figure 5.3 shows separation performance of different types of mixtures. Compared with the SNMF2D and the SCICA method, the proposed method renders a more optimal part-based factorization. The factorization is unique under certain conditions (e.g., adaptive sparse and nonnegative component), making it unnecessary to impose constrains in the form of statistical independence between original sources. Furthermore, the proposed method can automatically detect the optimal number of components of the individual source, thus leading to more robust separation results compared to the SNMF2D and the SCICA method.

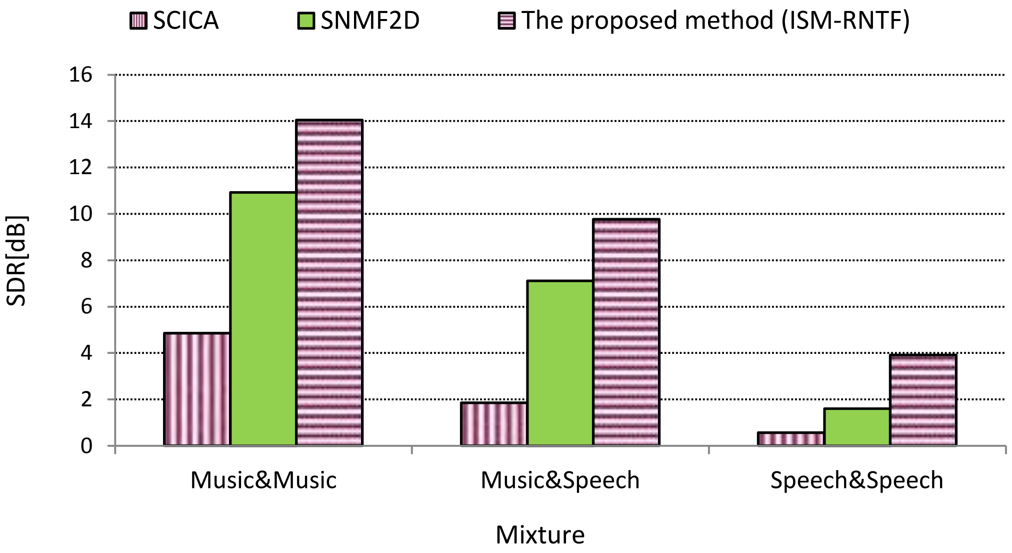


Figure 5.3: Comparison of average SDR performance on mixture of two audio sources with SNMF2D, SCICA, and the proposed method (ISM-RNTF) with $\beta = 1$ and $\delta = 1$.

5.3.2.2 Real Stereo Signal (left channel only)

In this evaluation, three stereo signals wdrums, nodrums and Shannonsongs Sunrise are used to demonstrate the effectiveness of the proposed method in dealing with having one signal from left channel of stereo signals. $x_1(t)$ is a "left channel mixture" of stereo signal. $x_2(t)$ is a imitated-stereo mixture which was generated from (5.3) for a "right channel mixture".

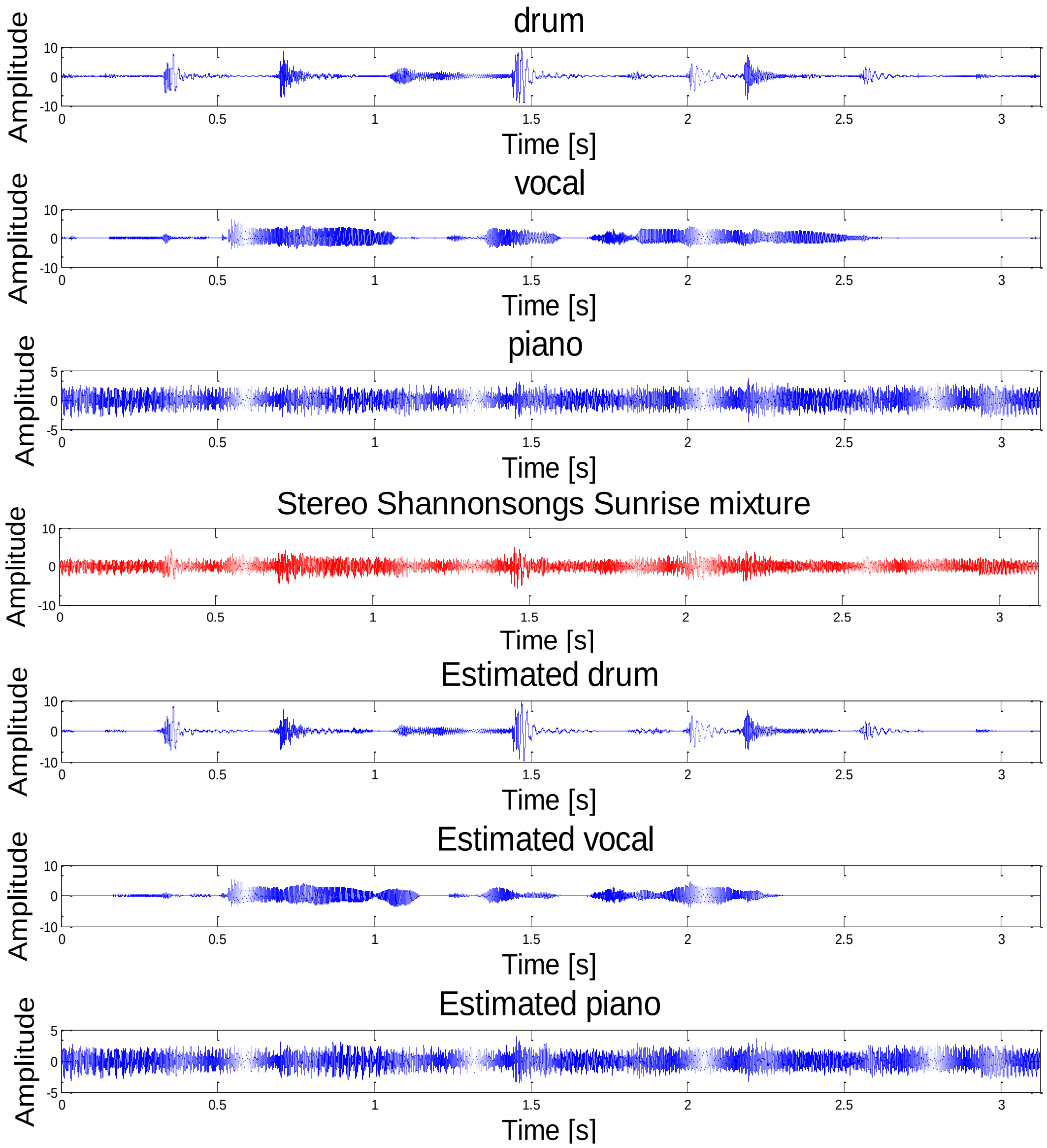


Figure 5.4: Original sources, single channel mixture, and estimated sources of Shannonsongs Sunrise mixture using the proposed method with $\beta = 1$ and $\delta = 1$.

Figure 5.4 shows the three original sources, the single channel mixture and the separated sources using the proposed method with $\beta = 1$ and $\delta = 1$. From the plots, it is visually evident that the mixture has been clearly separated comparing with the original sources.

5.3.3 Impact of weight ($\boldsymbol{\beta}$) and time-delay ($\boldsymbol{\delta}$) parameters on Matrix Factorization and Source Separation

The imitated stereo mixture is formulated via determining the weight β and the time-delay δ parameters. The weight β parameter acts as a controlling factor to maintain the difference of the sources' AR coefficients and to control the amount of the residues $r_j(t; \delta, \beta)$.

The impact of determination of values for β and δ parameters will be investigated in this section. In this implementation, the selection of values for β and δ parameters will depend on the type of sources and require manual setting. A set of experiments has been conducted to determine the β and δ pairs by using wdrums, nodrums and Shannonsongs Sunrise mixtures. A finite range of β and δ is selected to be [-4, 4] (excluding $\beta = 0$) and $[1, 2]$, respectively. The reason is in the extreme case of $\beta = 0$, this leads to $y_1(t) = y_2(t)$ where the imitated stereo mixture cannot be formulated. In practice, the AR coefficients of sources are generally unknown. However, if one knows the source category then β and δ can be chosen from τ. Moreover, if specific information of the sources such as piano or drum is known in advance then the AR coefficients can be determined by randomly sample the signals that belong to those groups. Hence, this enables the algorithm to estimate β and δ for the specific type of sources.

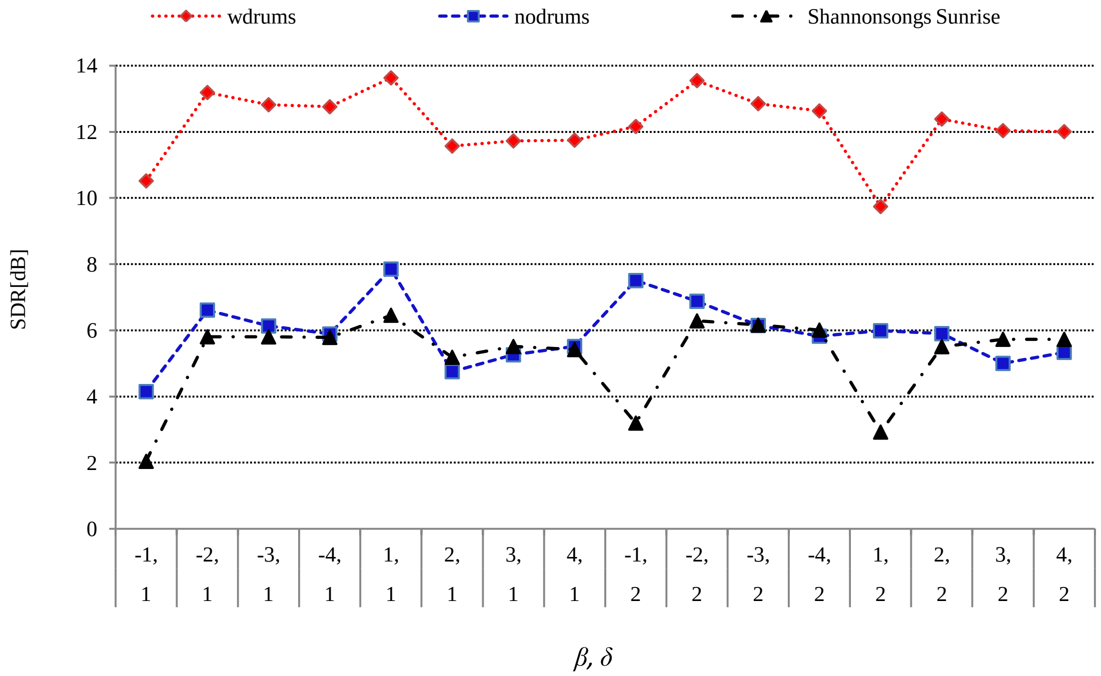
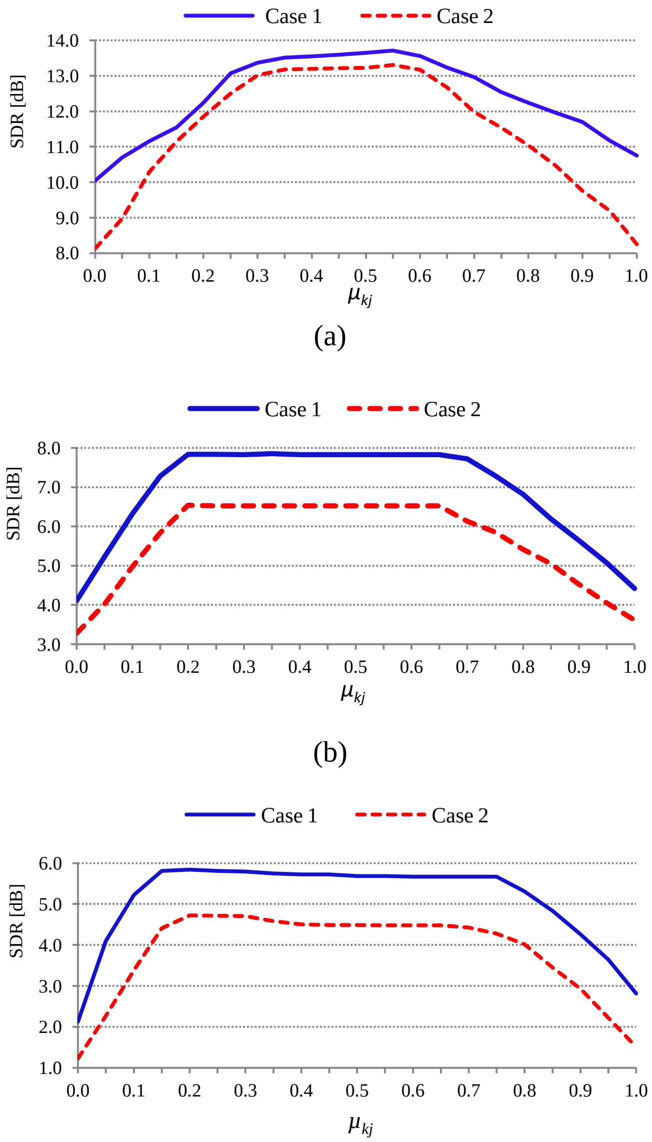


Figure 5.5: Separation results of the proposed method by using different weight (β) and time-delay (δ) parameters.

Figure 5.5 shows the separation results in terms of the SDR for the mixtures of wdrums, nodrums and Shannonsongs Sunrise with 16 pairs of β and δ defined as $\tau = \begin{Bmatrix} (-1,1),(-2,1),(-3,1),(-4,1),(1,1),(2,1),(3,1),(4,1), \\ (-1,2),(-2,2),(-3,2),(-4,2),(1,2),(2,2),(3,2),(4,2) \end{Bmatrix}$. As a result in Figure 5.5, it can be seen that there are certain values of β and δ where the algorithm performs the best. In the case of wdrums mixture, the best performance is obtained when the pairs of $\beta\ and\ \delta$ ranges are (1,1), (1,-2) and (2,-2) (within 5% from highest SDR) with the highest average SDR is 13.63dB which $\beta\ and\ \delta$ is (1,1). As for nodrums mixture, the best performance is obtained when the pairs of $\beta\ and\ \delta$ ranges are (1,1) and (2,-1) (within 5% from highest SDR) with the highest average SDR is 7.85dB which $\beta\ and\ \delta$ is (1,1) and in the case of Shannonsongs Sunrise mixture, the best performance is obtained when the pairs of $\beta\ and\ \delta$ ranges are (1,1), (2,-2) and (2,-3) (within 5% from highest SDR) with the highest average SDR is 6.46dB which $\beta\ and\ \delta$ is (1,1). In the case where the type of sources is unknown, then choosing $(\beta, \delta) = (1,1)$ will yield the best possible SDR since this particular pair overlaps with all the three categories. The results indicate that only the low order AR coefficients i.e. $\delta = 1$ are beneficial for separation. This is not surprising since speech and music are mainly characterized by the initial few AR coefficients and these coefficients tend to vary for different sources.

5.3.4 Impact of μ_{kj} on Separation Performance



(c)

Figure 5.6: SDR results as a function of μ_{kj}.(a) wdrums. (b) nodrums. (c) Shannonsongs Sunrise.

In this section, the impact of μ_{kj} will be investigated. In practice, the actual statistics for computing the prior on $\mathbf{W}$ (μ_{kj}) given in (5.34) is unknown. In this case, the selection of μ_{kj}will depend on the type of sources and require estimation. Hence, we investigate the effects of μ_{kj} in conjunction with the pruning method on the separation performance. Firstly, we estimate $\hat{\mu}_{kn} = \hat{\sigma}_k^{-2}\hat{\sigma}_n^{-2}\hat{c}_{k,n}$ using $\hat{c}_{k,n} = \boldsymbol{w}_k^{\boldsymbol{T}}\boldsymbol{w}_n$ and $\hat{\sigma}_j^{-2} = 1/\|\boldsymbol{w}_j\|^{\mathbf{2}}$ for $j = k, n$. We then compare the estimated $\hat{\mu}_{kn}$ with manual setting. The following two cases are considered: Case 1) with pruning and μ_{kj} is varied from 0, 0.05, 0.1,…, 1.0 Case 2) without pruning and μ_{kj} is varied from 0, 0.05, 0.1,…, 1.0. The wdrums, nodrums and Shannonsongs Sunrise datasets have been used for the above cases.

The separation results in terms of the SDR are given in Figures 5.6(a)-5.6(c). According to Figures 5.6(a)-5.6(c), Case 1 yields the best overall performance, where the average improvement over Case 2 can be summarized as follows: 1) for wdrums mixture, the average SDR improvement is 0.96dB per source; 2) for nodrums mixture, the average SDR improvement is 1.26dB per source; and 3) for mixtures of music and vocal (Shannonsongs Sunrise), the improvement is 1.29dB per source. The results have also clearly indicated that there are certain values of μ_{kj} where the algorithm performs the best. In the case of wdrums mixture, the best performance is obtained when μ_{kj} ranges from 0.33 to 0.64 (within 2% from highest SDR) with the highest average SDR is 13.71dB. The $\hat{\mu}_{kn}$ rendered from data estimation is 0.47 which very closely approaches the optimum SDR, which is at $\mu_{kj} =$ 0.52. As for nodrums mixture, the best performance is obtained when μ_{kj} ranges from 0.17 to 0.72 with the highest average SDR is 7.85dB. The $\hat{\mu}_{kn}$ rendered from data estimation is 0.39 which very closely approaches the optimum SDR, which is at $\mu_{kj} = 0.34$. In the case of Shannonsongs Sunrise which is a music and vocal mixture, the best performance is obtained when μ_{kj} ranges from 0.13 to 0.46 with the highest average SDR is 5.84dB. The $\hat{\mu}_{kn}$ rendered from data estimation is 0.23 which very closely approaches the optimum SDR, which is at $\mu_{kj} = 0.21$.

From the above findings, we can conclude that for music mixtures, the best performance is obtained when μ_{kj} ranges from 0.17 to 0.72 and in the case of music and vocal mixture, the best performance is obtained when μ_{kj} ranges from 0.13 to 0.46. On the contrary, it is noted that when μ_{kj} is set either too low or high, the separation performance tends to degrade. It is also worth pointing out that the separation results are rather coarse when the factorization is non-regularized (i.e., without prior on $\mathbf{W}$) and without pruning. Here, we can see that the average SDR of without prior on $\mathbf{W}$ and without pruning is the lowest among the three methods across $\mu_{kj} > 0$. As evidence, Figure 5.6 shows the SDR of the without prior on $\mathbf{W}$ and without pruning method as for wdrums mixture: 8.13dB per source, for nodrums: 3.28dB per source, and for Shannonsongs Sunrise: 1.22dB per source. We may summarize the average improvement of the proposed method (with prior on $\mathbf{W}$ and with pruning) against the case of without pruning and without prior on $\mathbf{W}$: 1) For wdrums mixture, the improvement per source in terms of the SDR is 4.26dB. 2) For nodrums mixture, the improvement per source in terms of the SDR is 3.60dB. 3) For Shannonsongs Sunrise mixture, the improvement per source in terms of the SDR is 3.38dB.

By incorporating regularization (i.e., using $\mu_{kj} > 0$ and pruning), the performance increases significantly for all types of mixture. This is clearly evident in Figures 5.6 (a)-5.6(c) where the average SDR result for separation three mixtures scales up to 9.1dB while for the case of without regularization the average SDR result is only 7.6dB. This amounts to a significant 1.5dB performance improvement using the proposed regularization than that without regularization. Thanks to the modified Gaussian prior, this correlation is explicitly modeled by μ_{kj} in the proposed method. This enables the estimated basis vectors $\mathbf{w}_1$ and $\mathbf{w}_2$to take advantage of the correlation in learning the real basis directly from the mixed pattern. This explains the reason as to why that the proposed method with pruning and with prior on $\mathbf{W}$ shows better performance than the proposed method with pruning and without prior on $\mathbf{W}$. Therefore, the analysis have unanimously indicated the importance of selecting the correct number of components and of incorporating the correlation μ_{kj} between the different basis vectors in order to arrive at the optimal performance of feature extraction.

5.3.5 Comparison With Other NTF-Based BSS Methods

In this section, comparison of the proposed method with other NTF-based source separation methods will be undertaken. These consist of the following methods. 1) The proposed method based on multichannel source separation using stereo source mixture. 2) The proposed method without pruning. 3) ISM with NTF (ISM-NTF) is based on factorizing the power spectrogram of the mixed signal into a sum of components given by ISM.

Table 5.2: Comparison of average SDR, SIR and SAR performance on three mixtures between the proposed method (ISM-RNTF), the proposed method (ISM-RNTF) without pruning, and the ISM-NTF.

Mixtures	Methods	SDR (dB)	SIR (dB)	SAR (dB)
wdrums (Hi-hat/drums/bass)	Proposed method (ISM-RNTF)	13.63	37.95	13.65
	Proposed method (ISM-RNTF) without pruning	13.13	37.44	13.15
	ISM-NTF	12.40	36.63	12.41
nodrums (bass/lead G. /rhythmic G.)	Proposed method (ISM-RNTF)	7.85	30.85	7.84
	Proposed method (ISM-RNTF) without pruning	6.58	30.74	6.60
	ISM-NTF	6.28	28.40	6.27
Shannonsongs Sunrise (drum/vocal/piano)	Proposed method (ISM-RNTF)	5.79	25.83	5.85
	Proposed method (ISM-RNTF) without pruning	4.77	25.12	4.83
	ISM-NTF	4.32	24.36	4.39

The comparison results are tabulated in Table 5.2. In general, the above methods deliver good competitive results, especially in terms of the SDR. However, the best overall separation performance is still wielded by the proposed method. By analyzing the results, we may summarize the average improvement of the proposed method over the proposed method without pruning and ISM-NTF method as follows: 1) for wdrums mixture, the average improvements are 0.5 and 1.2 dB per source, respectively; 2) for nodrums mixture, the improvements are 1.3 and 1.6 dB per source, respectively; and 3) for mixture of Shannonsongs Sunrise, the improvements are 1.0 and 1.5 dB per source, respectively. This clearly shows that the factorization coupled with the adaptive assignment of the sparsity parameters and the inclusion of the correlation among the different basis vectors as in the proposed method which enforces the uniqueness of the factorization leading to higher accuracy in estimating the temporal information and frequency patterns of the audio signals.

5.3.6 Comparison With Other SCBSS Methods

Table 5.3 further gives the SDR, SIR, and SAR comparison results between our proposed method and the other SCBSS methods. In comparison, the average performance improvement of the proposed method over the the sparse nonnegative matrix 2-dimensional factorization (SNMF2D) and the single-channel independent component analysis (SCICA) methods can be summarized as follows: 1) for the mixture of wdrums, the average improvement per source in terms of the SDR is 9.0dB, and 10.2dB per source, respectively; 2) for the mixture of nodrums, the average improvement per source in terms of SDR is 3.4dB, and 6.4dB per source, respectively; 3) for the Shannonsongs Sunrise mixture, the average improvement per source in terms of SDR is 2.9dB, and 9.3dB per source, respectively.

Table 5.3: Comparison of average SDR, SIR and SAR performance on three mixtures of three audio sources between SNMF2D, SCICA and the proposed method (ISM-RNTF).

Mixtures	**Methods**	**SDR (dB)**	**SIR (dB)**	**SAR (dB)**
wdrums (Hi-hat/drums/bass)	Proposed method (ISM-RNTF)	13.63	37.95	13.65
	SNMF2D	4.62	11.90	6.45
	SCICA	3.47	12.28	4.03
nodrums (bass/lead G. /rhythmic G.)	Proposed method (ISM-RNTF)	7.85	30.85	7.84
	SNMF2D	4.45	12.15	6.13
	SCICA	1.43	13.50	2.57
Shannonsongs Sunrise (drum/vocal/piano)	Proposed method (ISM-RNTF)	5.79	25.83	5.85
	SNMF2D	2.86	7.24	6.14
	SCICA	-3.50	8.31	-0.62

Analyzing the separation performance and SDR results, the proposed method yields the best separation performance for all recovered sources. The SCICA method performs with poorer results whereas the separation performance by the SNMF2D method is slightly better than the SCICA method. Our proposed method gives significantly better performance than the SNMF2D and SCICA methods. The reasons are the spectral dictionary obtained via SNMF2D methods are not pruning for estimating adequate number of latent components for each the audio source. In addition, the SNMF2D and SCICA methods have been separated individual sources from only single channel mixture. On the other hand, our proposed method can be created the two channels mixture from a single mixture by using ISM technique which provides the separation process with more information than the SNMF2D and SCICA methods. This lead to robust separation method from just a single channel mixture signal is possible.

5.4 Summary

A novel solution for the single channel blind source separation problem has been presented. An imitated stereo mixture is proposed by weighting and delaying the observed mixture where the source signals can be modelled by the AR processes. Experiments have been conducted successfully to separate real-audio mixtures. In this work, the separability analysis of the imitated stereo mixture has been derived based on Wiener masking. The proposed method has demonstrated high level separation performance for real-audio sources. The proposed method enjoys at least three advantages: Firstly, it can be employed to separate individual sources from the imitated stereo mixture by using a single channel mixture. Secondly, the modified Gaussian prior is formulated to express the correlation between different basis vectors. Finally, our proposed algorithm can automatically detect the optimal number of latent components of the individual source, thus enabling the spectral dictionary and temporal codes of the individual source to be estimated more efficiently.

CHAPTER 6

CONCLUSION OF THE THESIS

The work in this thesis has fulfilled all the aims and objectives set out in Chapter 1. In Chapter 2, an overview of the BSS of linear instantaneous mixtures was presented. Both SCBSS and multi-CBSS methods that aim to increase the accuracy of the separated signals through various techniques were summarised and organised into a unifying framework. However, the practicality of these approaches still has several unresolved challenges which therefore limit the applications in reality. These problems have been summarised in Chapter 2. Hence, this requires the development of reliable solutions for the BSS of single channel and multi-channel mixtures to improve the performance at both theoretical and practical issues. This fact is the impetus behind the three proposed method of this thesis, which is to develop novel algorithms for retrieving single channel and muli-channel mixed sources.

6.1 Proposed BSS Methods

In Chapter 3, the novel framework of amalgamating pruning and Bayesian regularized cluster nonnegative tensor factorization under a PARAFAC structure with Itakura-Saito divergence has been proposed for multichannel audio source separation. The impetus behind the proposed work is that sparseness achieved by the conventional NTF is not efficient enough in source separation, it is necessary to yield control over the degree of sparseness explicitly for each element code $\{h_{ij}\}$. In addition, it does not incorporate correlation information between different basis vectors into the factorization process. Underlying all factorization algorithms is the principal difficulty in estimating the adequate number of latent components for each source. The proposed method addresses this issue by using the principle of pruning. The proposed method offers at least four advantages: first, the sparse regularization term is adaptively tuned using a hierarchical Bayesian approach to yield the desired sparse decomposition, thus enabling the spectral dictionary and temporal codes of nonstationary audio signals to be more efficient estimate, second, the modified Gaussian prior is formulated to express the correlation between different basis vectors , third, our proposed algorithm can automatically detect the optimal number of latent components of the individual source, and finally, it avoids the strong constraints of separating blind source without training knowledge. Hence, the work is a step forward to realizing optimal BASS. This has been verified concretely based on experiments which produced very promising results. In addition, the separation performance of the proposed method yields significant improvement of SDR on multichannel audio separation compared with other NTF-based source separation methods.

Chapter 4 introduced a novel framework of amalgamating variational L_1-sparse with complex matrix factorization for single channel source separation. The impetus behind the proposed work is that NMF cannot estimate the phase spectra of underlying constituent signals, and sparseness achieved by the conventional CMF is not efficient enough. The proposed method addresses the above and enjoys at least two significant advantages: first, the sparse regularization term is adaptively tuned to obtain the desired sparse decomposition, and second, the proposed method can extract recurrent patterns of magnitude spectra that underlie observed complex spectra and the phase estimates of constituent signals, thus enabling the features of the components to be extracted more efficiently. In addition, the analytical update equations were derived through an auxiliary function approach and an experimental evaluation showed that reasonably good separation was obtained with the present method.

In Chapter 5, a novel solution for the single channel blind source separation problem has been presented. An imitated stereo mixture is proposed by weighting and delaying the observed mixture where the source signals can be modelled by the AR processes. Experiments have been conducted successfully to separate real-audio mixtures. In this work, the separability analysis of the imitated stereo mixture has been derived based on Wiener masking. The proposed method has demonstrated high separation performance for real-audio sources. The proposed method enjoys at least three advantages: Firstly, it can be employed to separate individual sources from the imitated stereo mixture by using a single channel mixture. Secondly, the modified Gaussian prior is formulated to express the correlation between different basis vectors. Finally, our proposed algorithm can automatically detect the optimal number of latent components of the individual source, thus enabling the spectral dictionary and temporal codes of the individual source to be estimated more efficiently.

In conclusion, the three proposed methods are summarized in Table 6.1.

Table 6.1: Summary of the proposed BSS methods.

Methods	**Type of BSS**	**TF representation**	**Cost function**	**Regularization**		**Update method**
				W	**H**	
APBNTF	Multi-Channel BSS	Spectrogram	ISD	Correlation of the basis	Adaptive sparsity (MAP)	MU
vL_1-SCMF	SCBSS	Spectrogram	LS	-	Adaptive sparsity (VB)	MU
ISM-RNTF	SCBSS	Spectrogram	ISD	Correlation of the basis	Adaptive sparsity (MAP)	MU

6.2 Comparison of the Proposed SCBSS Methods

In this section, the proposed BSS methods will be tested across all types of mixture and compared in terms of SDR, SAR and SIR. The following table summarises the comparison results. In comparison, the ISM-RNTF leads to the best separation performance for the music & music mixtures and music & speech mixtures. The reasons of using ISM model for SCBSS have been described in Chapter 5. However, it is interesting to point out that the big advantage of using ISM-RNTF with temporal correlation is that this method is more meaningful features extraction that pertains to the data than vL_1-SCMF method and analogous to the stereo signal concept given by one microphone, thus simultaneously retain a high level of the separation performance. On the other hand, the vL_1-SCMF does not have prior on $\mathbf{W}$ such that the frequency patterns of each source may not be estimated as well as ISM-RNTF. Additionally, the vL_1-SCMF is separated individual sources by using the single channel. However, the vL_1-SCMF performs good results for male speech and female speech as compared with ISM-RNTF owing to incorporating the phase parameter will give the better recovered sources than without using phase information.

Table 6.2: Separation results using different SCBSS methods.

Mixtures	**TF methods**	**SDR (dB)**	**SAR (dB)**	**SIR (dB)**
music and music	vL_1-SCMF	12.7	12.8	14.9
	ISM-RNTF	14.4	14.3	16.8
music and speech	vL_1-SCMF	8.9	8.5	9.8
	ISM-RNTF	9.8	10.1	12.2
male speech and female speech	vL_1-SCMF	4.5	6.6	7.8
	ISM-RNTF	2.4	2.5	13.3

The overall comparison results between the APBNTF and ISM-RNTF methods where proposed in Chapter 3 and Chapter 5, respectively, have been summarized in Table 6.3.

Table 6.3: Separation results using different Proposed ISM-RNTF and APBNTF methods based on NTF.

Mixtures	**TF methods**	**SDR (dB)**	**SAR (dB)**	**SIR (dB)**
wdrum	APBNTF	12.8	12.8	38.1
	ISM-RNTF	13.6	13.7	37.9
nodrum	APBNTF	10.0	10.1	34.4
	ISM-RNTF	8.9	8.8	31.9
Shannonsongs Sunrise	APBNTF	10.1	10.1	31.8
	ISM-RNTF	3.8	3.9	12.8

According to the table, the overall results indicate that the APBNTF method gives the outstanding separation performance over the ISM-RNTF method at the average SDR 6.6dB per source. The APBNTF achieves the good results for nodrum and Shannonsongs Sunrise as compared with ISM-RNTF. While ISM-RNTF yield better results than the APBNTF methods for the wdrum mixtures. This may because: Firstly the selecting the β and δ parameters for the imitated stereo mixture causes the coefficients of the j$^{\text{th}}$ sources a_j of ISM-RNTF differ from one another than the natural coefficients given by the real stereo mixture. Secondly, the wdrums signal consists of Hi-hat, drum and bass which their coefficients are more distinguish than the musical instruments in the nodrum (i.e. bass, lead guitar, and rhmthmix guitar) and Shannonsongs Sunrise (i.e. drum, vocal and piano). Thus, this allows some space that the estimated coefficients of the j^{th} sources can be diveraged from its actual value. On the other hands, the nodrum and Shannonsongs Sunrise requires the accurate coefficients of the j^{th} sources that available in the APBNTF method.

6.3 Future Work

6.3.1 Development of BSS Method for Noisy Mixture Enhancement

As described in Chapter 5, the proposed BSS algorithms are derived for noise-free condition. Hence, The method may not able to solve the BSS problem in noisy environments since the presence of noise seriously degrades the performance. In a realistic scenario of audio applications, desired signals will be corrupted by additive background noise. In the future work, a novel framework to solving BSS based on ISM-RNTF in noisy environments [94] will be developed. In an instantaneous linear problem of source separation, the unknown source signals and the observed data are related to the single observed mixing model in terms of the sources and a noise in time domain is considered as

$$y_1(t) = \sum_{j=1}^{N_s} x_j(t) \ + n_1(t) \tag{6.1}$$

For the virtual mixture by weighting and time-shifting the single channel mixture $x_1(t)$, it gives as follows:

$$y_2(t) = \textstyle\sum_{j=1}^{N_s} a_j x_j(t-\delta) + r_j(t) + n_2(t) \tag{6.2}$$

where $x_j(t)$ denotes the jth source signal, $a_j(t;\delta,\beta)$ and $r_j(t;\delta,\beta)$ represent the mixing attenuation and residue of the j^{th} source, respectively and $n_j(t)$ is additive noise of the j^{th} source. This method will be considered a wide class of the cost functions and efficient NTF algorithms with only single parameter to tune. The optimal choice of the parameter in the cost function depends and on a statistical distribution of data and additive noise, thus an updating rules algorithm should be applied for estimating the basis matrix and the source matrices, depending on *a priori* knowledge about the statistics of noise.

The aim of the developed SCBSS based on ISM-RNTF method is to estimate the original signals from the noisy mixture by including a preprocess to eliminate noise components from the observed signal and then performing the separation process.

6.3.2 Development of a BSS method for non-stationary mixing model

Most of the algorithms for the BSS approach are based on a model of stationary sources. Non-stationary Blind Separation (NBS) represents the separation of independent source signals with non-stationary mixing from the single sensor or multi-sensor and time-varying temporally correlated sources. The speakers or equivalently, the microphones can move. The problem is cast in terms of the mixing proportions of sources which can be tracked by using particle filter [95, 96].

Future works will consider of the generalization of the proposed estimator for the non-stationary blind source separation problem. It addresses the problem of separating the sources when speakers or microphones are moving, and developing a generative model for analysis of non-stationary multivariate time series. The objective of this future work is to perform BSS in time-varying mixing process of linear instantaneous mixture of independent temporally correlated, non-stationary sources.

The BSS method to separate non-stationary mixing model based on ISM-RNTF will be developed. The non-stationary mixing model has not been solved by using current BSS methods. For instantaneous non-stationary single observed mixing model, it gives as follows:

$$y_1(t) = \sum_{j=1}^{N_s} m_j(t)x_j(t) \ + n_1(t) \tag{6.3}$$

where $m_j(t)$ denotes the j$^{\text{th}}$ source mixing parameters at t time, and $n_1(t)$ is additive noise. Thus, the power TF representation of matrix representation is given by $|\boldsymbol{Y}|^2 = \sum_{j=1}^{N_s} \left|\mathbf{M}_j\right|^2 \bullet \left|\boldsymbol{X}_j\right|^2 + \mathbf{N}$. The matrix $\mathbf{M}_j$ is a mixing parameter in TF domain (it is assumed that the mixing parameter is stationary within a short period. The aim of the developed BSS method is to estimate nonstationary mixing model $\mathbf{M}_j$ and the sources $\left|\mathbf{X}_j\right|^2$.

6.3.3 Development of Informed Speech Separation based on a Cochleagram TF Representation

The cochleargram modelled by using the gammatone filterbank was proposed in [92, 97-99] to decompose the time-domain input into the frequency domain. It produces a non-uniform time-frequency resolution while it is more balanced between the high and low frequency areas when compared to the classic spectrogram and log-frequency spectrogram (constant-Q transform).The impulse response of a gammatone filter centered at frequency f is given by:

$$g(f,t) = \begin{cases} t^{l-1}e^{-2\pi vt}\cos 2\pi ft\,, & t \geq 0 \\ 0, & else \end{cases} \tag{6.4}$$

where l represents the order of filter, v denotes the rectangular bandwidth which increases with the center frequency f. The filter output response $x(c,t)$ can be expressed with regards to a particular filter channel c as:

$$x(c,t) = x(t) * g(f_c,t) \tag{6.5}$$

where f_c denotes the center frequency, and '*' indicates a convolution operator.

The development will construct the audio signal TF representation using the gammatone filterbank. It produces a non-uniform TF domain termed as the cochleagram whereby each TF unit has different resolution unlike the classic spectrogram which deals only with uniform resolution. The mixed audio signal is more separable in the cochleagram. This property befits to an NTF method which requires sparsity. Moreover, in the separating process, an exemplar masking will be provided to turning the masking of the original sources. Therefore, the development of imitated-stereo mixing informed speech separation using Cochleagram will improve the accuracy of the speech separation performance.

REFERENCE

[1] E. C. Cherry, “Some experiments on the recoginition of speech, with one and two ears,” *Journal of the Acoustic Society of America*, vol. 25, no. 5, pp. 975 – 979, 1953.

[2] J. F. Cardoso, “Blind Signal Separation: Statistical Principles,” *in Proceedings of the IEEE*, vol. 86, no. 10, pp. 2009-2025, 1998.

[3] Gil-Jin Jang and Te-Won Lee, “A Maximum Likelihood Approach to Single-channel Source Separation,” *Journal of Machine Learning Research*, pp.1365-1392, 2003.

[4] M. Kühne, R. Togneri, and S. Nordholm, “Time-Frequency Masking: Linking Blind Source Separation and Robust Speech Recognition,” *Speech Recognition, Technologies and Applications,* pp. 550, Austria, Nov. 2008.

[5] D.D. Lee and H.S. Seung, “Learning the parts of objects with nonnegative matrix factorization,” *Nature*, vol. 401, pp. 788–791, 1999.

[6] T. Virtanen, “Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Trans. Audio, Speech, and Lang. Process.,* vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[7] P. Smaragdis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” *in IEEE Workshop on Appl. of Signal Processing to Audio and Acoustics*, pp. 177–180, 2003.

[8] I. Kotsia, S. Zafeiriou, and I. Pitas, “A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems,” *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 588–595, 2007.

[9] A. Ozerov and C. Févotte, “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation,” *IEEE Trans. Audio, Speech, and Lang. Process*., vol. 18, no. 3, pp. 550–563, Mar. 2010.

[10] Y.C. Cho and S. Choi, “Nonnegative features of spectro-temporal sounds for classification,” *Pattern Recognit. Lett*., vol. 26, pp. 1327–1336, 2005.

[11] R. Zdunek and A. Cichocki, “Nonnegative matrix factorization with constrained second-order optimization,” *Signal Process*., vol. 87, no.8, pp. 1904–1916, Aug. 2007.

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

[13] F.J. Theis and G.A. García, “On the use of sparse signal decomposition in the analysis of multi-channel surface electromyograms,” *Signal Process*., vol. 86, no. 3, pp. 603–623, Mar. 2006.

[14] O. Okun and H. Priisalu, “Unsupervised data reduction,” *Signal Process*., vol. 87, no. 9, pp. 2260–2267, Sep. 2007.

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

[16] D.D. Lee and H.S. Seung, “Algorithms for non-negative matrix factorization,” *in Proc. NIPS*, pp. 556–562, 2000.

[17] R. Kompass, “A generalized divergence measure for nonnegative matrix factorization,” *Neural Comput.*, vol. 19, no. 3, pp. 780–791, 2007.

[18] A. Cichocki, R. Zdunek, and S.I. Amari, “Csisz´ar’s divergences for non-negative matrix factorization: Family of new algorithms,” *in Proc. Int. Conf. Ind. Compon. Anal. Blind Signal Separat. (ICABSS’06)*, Charleston, SC, vol. 3889, pp. 32–39, Mar. 2006.

[19] C. Févotte, N. Bertin, and J.-L. Durrieu, “Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis,” *Neural Computation*, vol. 21, pp. 793–830, Mar. 2009.

[20] C.-J. Lin, “Projected gradient methods for nonnegative matrix factorization,” *Int. J. Neural Comput*., vol. 19, no. 10, pp. 2756–2779, Oct. 2007.

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

[22] K. Stadlthanner, F. J. Theis, C. G. Puntonet, and E. W. Lang, “Extended sparse nonnegative matrix factorization,” *in Proc. Artif. Neural Netw*., vol. 3512, pp. 249–256, 2005,

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

[24] M. E. Davies and C. J. James, “Source separation using single channel ICA,” *Signal Process*., vol. 87, no. 8, pp. 1819–1832, 2007.

[25] M. Helén and T. Virtanen, “Separation of drums from polyphonic music using nonnegative matrix factorization and support vector machine,” *in Proc of 13th European Signal Processing*, 2005.

[26] M. Burghoff and P. Van Leeuwen, “Separation of Fetal and Maternal Magnetocardiographic Signals in Twin Pregnancy using Independent Component Analysis (ICA),” *in Biomag 2004, Boston, USA*, pp. 311-312, Aug. 2004.

[27] N. Correa, T. Adali, and V. D. Calhoun, “Performance of Blind Source Separation Algorithms for fMRI Analysis using a Group ICA Method,” *Magnetic Resonance Imaging*, vol. 25, no. 5, pp. 684-694, 2007.

[28] J. V. Stone, J. Porrill, N. R. Porter and I. D. Wilkinson, “Spatiotemporal Independent Component Analysis of Event-Related fMRI Data using Skewed Probability Density Functions,” *Neuroimage*, vol. 15, no. 2, pp. 407-421, Feb. 2002.

[29] M. Maazaoui, K. Abed-Meraim and Y. Grenier, “Blind source separation for robot audition using fixed HRTF beamforming,” *EURASIP J. Adv. Sig. Proc*., 2012.

[30] B. Gao, W. L. Woo, and S. S. Dlay, “Variational regularized 2-D nonnegative matrix factorization,” *IEEE Trans. Neural Netw. and Learning Sys*., vol. 23, no. 5, pp. 703–716, May 2012.

[31] B. Gao, W. L. Woo, and S. S. Dlay, “Single-channel source separation using EMD-subband variable regularized sparse features,” *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 4, pp. 961-976, May 2011.

[32] B. Gao, W. L. Woo, and S. S. Dlay, “Adaptive sparsity nonnegative matrix factorization for single channel source separation,” *IEEE Journal of Selected Topics in Signal Process*., vol. 5, no. 5, pp. 989-1001, 2011.

[33] J. Koikkalainen and J. Lotjonen, “Image Segmentation with the Combination of the PCA- and ICA-Based Modes of Shape Variation”, in *IEEE International Symposium on Biomedical Imaging: Nano to Macro*, vol. 1, pp. 149-152, April 2004.

[34] C. Beckmann and S. Smith, “Probability Independent Component Analysis for Functional Magnetic Resonance Imaging”, *IEEE Transactions on Medical Imaging*, vol. 23, pp. 137-152, 2004.

[35] C. Liu and h. Wechsler, “Independent Component Analysis of Gabor features for face recognition”, *IEEE Trans. On Neural Netw.*, vol. 14, pp. 919-928, 2003.

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

[37] B. Mijovic, M. D. Vos, I. Gligorijevic, J. Taelman, and S. V. Haffel, “Source separation from single-channel recordings by combining empirical-mode decomposition and

independent component analysis," *IEEE Trans. Biomedical Eng*., vol. 57, no. 9, Sep. 2010

[38] K.-K. Shyu, M.-H. Lee, Y.-T. Wu, and P.-L. Lee, "Implementation of pipelined FastICA on FPGA for Real-time Blind source separation," *IEEE Trans. Neural Netw.,*vol.19, no.6, Jun. 2008.

[39] S. C. Douglas, "A statistical convergence analysis of the fastica algorithm for two-source mixtures," *IEEE Trans. Neural Netw*., vol. 14, pp. 943-949, Jul. 2003.

[40] Z. Koldovský, P. Tichavský, and E. Oja, "Efficient variant of algorithm fastICA for independent component analysis attaining the Cramer-Rao lower bound," *IEEE Trans. Neural Netw.*, vol. 17, pp. 1265- 1277, 2006.

[41] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 3, pp. 538-5493, 2010.

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

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

[44] W. Liu, D. P. Mandic, and A. Cichocki, "Blind Second-order Source Extraction of Instantaneous Noisy Mixtures", *IEEE Trans. on Circuits and Systems II*, vol. 53, no. 9, pp. 931-935, 2006.

[45] S. Rickard and A. Cichocki, “When is non-negative matrix decomposition unique?”, *Information Sciences and Systems, CISS 2008*, 42nd Annual Conference, pp. 1091 – 1092, Mar. 2008.

[46] N. Bertin, R. Badeau, and G. Richard, “Blind Signal Decompositions for Automatic Transcription of Polyphonic Music: NMF and K-SVD on the Benchmark,” *in Proc. ICASSP*, pp.65-68, 2007.

[47] G. J. Mysore, P. Smaragdis, B. Raj, “Non-negative Hidden Markov Modeling of Audio with Application to Source Separation,” *In Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA / ICA)*, pp.140-148, 2010.

[48] L.R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,” *Proceeding of the IEEE*, Vol. 88, No. 2, Feb. 1989.

[49] M. N. Schmidt, O. Winther, and L. K. Hansen, “Bayesian non-negative matrix factorization,” *In ICA '09: Proc. of the 8th International Conference on Independent Component Analysis and Signal Separation*, pp. 540-547, 2009.

[50] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, “Complex NMF: A new sparse representation for acoustic signals,” *in Proc. 2009 IEEE Intl. Conf. Acoust., Signal, Speech Process. (ICASSP’09)*, pp. 3437-3440, 2009.

[51] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, “Formulations and algorithms for multichannel complex NMF,” *in Proc. 2011 IEEE Intl. Conf. Acoust. Speech, Signal Process.*, 2011.

[52] 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," *IEEE (TENCON'11)*, pp. 806-809, 2011.

[53] R. M. Parry, and I. A. Essa, "Estimating the spatial position of spectral components in audio," *in Proc. 6th International Conference on Independent Component Analysis and Blind Signal Separation (ICA'06)*, Charleston SC, USA, pp. 666–673, Mar. 2006.

[54] D. FitzGerald, M. Cranitch, and E. Coyle, "Non-negative Tensor Factorization for Sound Source Separation," *in Proc. of the Irish Signals and Systems Conference*, Dublin, Ireland, 2005.

[55] A. Shashua and T. Hazan, "Non-Negative Tensor Factorization with Applications to Statistics and Computer Vision," *International Conference of Machine Learning (ICML)*, 2005.

[56] C. Févotte and A. Ozerov, "Notes on nonnegative tensor factorization of the spectrogram for audio source separation: statistical insights and towards self-clustering of the spatial cues," *in 7th International Symposium on Computer Music Modeling and Retrieval (CMMR 2010)*, 2010.

[57] M. A. Casey, and A. Westner, "Separation of Mixed Audio Sources by Independent Subspace Analysis," *in Proc. Of ICMC 2000,* pp. 154-161, Berlin, Germany, 20006.

[58] D. FitzGerald, "Automatic drum transcription and source separation," *Ph.D. dissertation, Dublin Inst. of Technol.*, Dublin, Ireland, 2004.

[59] G. Schwarz, “Estimating the dimension of a model,” *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[60] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Automat. Contr.*,vol AC-19 , pp. 716–723, Dec.1974.

[61] D. J. C. Mackay, “Probable networks and plausible predictions – a review of practical Bayesian models for supervised neural networks,” *Network: Computation in Neural Systems*, vol. 6, no. 3, pp. 469–505, 1995.

[62] V. Y. F. Tan and C. Févotte, “Automatic relevance determination in nonnegative matrix factorization with the β-Divergence,” *in Proc. NIPS,* Nov. 2011.

[63] M. Mørup and L. K. Hansen, “Automatic relevance determination for multiway models,” *Journal of Chemometrics*, vol. 23, no.7-8, pp. 352–363, 2009.

[64] C. Févotte, “Itakura-Saito nonnegative factorizations of the power spectrogram for music signal decomposition,” *In Wang, W., ed.: Machine Audition: Principles, Algorithms and Systems*, pp.266-296, IGI Global Press, 2010.

[65] M. Mørup and L. K. Hansen, “Sparse Coding and Automatic Relevance Determination for Multi-way models,” *Spars'09,* 2009.

[66] M. Mørup, and L. K. Hansen, “Tuning pruning in sparse non-negative matrix factorization,” *in Proc. 17th Eur. Signal Process. Conf.*, pp. 1–5, Glasgow, Scotland, 2009.

[67] A. Smilde, R. Bro, and P. Geladi, “Multi-way Analysis: Applications in the Chemical Sciences*,” John Wiley and Sons*, New York, 2004.

[68] F. Miwakeichi, E. Martnez-Montes, P. Valds-Sosa, N. Nishiyama, H. Mizuhara, and Y. Yamaguchi, “Decomposing EEG data into space−time−frequency components using parallel factor analysis,” *NeuroImage*, vol.22, no.3, pp.1035–1045, 2004.

[69] T. Hazan, S. Polak, and A. Shashua, “Sparse image coding using a 3D non-negative tensor factorization,” *International Conference of Computer Vision (ICCV)*, pp. 50–57, 2005.

[70] A. Shashua, R. Zass, and T. Hazan, “Multi-way clustering using super-symmetric non-negative tensor factorization,” *European Conference on Computer Vision (ECCV)*, Graz, Austria, May 2006.

[71] J. Sun, D. Tao, and C. Faloutsos, “Beyond streams and graphs: dynamic tensor analysis,” *Proc.of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.374–383, 2006.

[72] M. Heiler and C. Schnoerr, “Controlling sparseness in non-negative tensor factorization,” *Springer LNCS*, vol.3951, pp.56–67, 2006.

[73] M. Mørup, L.K. Hansen, C.S. Herrmann, J. Parnas, and S.M. Arnfred, “Parallel factor analysis as an exploratory tool for wavelet transformed event-related EEG,” *NeuroImage*, vol.29, no.3, pp.938–947, 2006.

[74] A. Cichocki, R. Zdunek, S. Choi, R. Plemmons, and S. Amari, “Nonnegative tensor factorization using Alpha and Beta divergencies,” *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP07)*, Honolulu, Hawaii, USA, pp.1393–1396, April 15–20 2007.

[75] A. Cichocki, R. Zdunek, A.H. Phan, and S. Amari, “Nonnegative Matrix and Tensor Facorizarions and Beyond,” *Wiely*, Chichester, 2009.

[76] Signal Separation Evaluation Campaign (SiSEC 2010). (2010) [Online]. Available: http://sisec.wiki.irisa.fr

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

[78] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, “First stereo audio source separation evaluation campaign: Data, algorithms and results.” *in Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA’07). Springer*, pp. 552–559, 2007.

[79] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2005.

[80] B. King and L. Atlas, “Single-channel source separation using complex matrix factorization,” *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 19, no. 8, pp. 2591 - 2597, Nov. 2011.

[81] B.A. Olshausen and D.J. Field, “Emergence of simple-cell receptive field properties by learning a sparse code for natural images,” *Nature*, 381, 607–609, 1996.

[82] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online Learning for Matrix Factorization and Sparse Coding,” *Journal of Machine Learning Research 11 (2010)*, pp. 19-60, 2010.

[83] H. Lee, A. Battle, R. Raina, and A.Y. Ng, “Efficient sparse coding algorithms,” *In Advances in Neural Information Processing Systems 19 (NIPS),* 2007.

[84] D. P. Bertsekas, “Nonlinear Programming,” *2nd ed. Belmont, MA: Athena Scientific,* 1999.

[85] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “RWC music database: Music genre database and musical instrument sound database,” *in Proc. Int. Symp. Music Inf. Retrieval (ISMIR),* Baltimore, MD, pp. 229–230, Oct. 2003.

[86] Y. Xiang, S. K. Ng, and V. K. Nguyen, “Blind Separation of Mutually Correlated Sources Using Precoders,” *IEEE Trans. Neural Netw.,* vol. 21, no. 1, pp. 82–90, Jan. 2010.

[87] R. de Frein and S. Rickard, “The synchronized short-time-Fourier-transform: properties and definitions for multichannel source separation,” *IEEE Trans. Signal Process*., vol. 59, no. 1, pp. 91-103, Jan. 2011.

[88] R. Balan, J. Rosca, S. Rickard, and J. O’Ruanaidh, “The influence of windowing on time delay estimates,” *in Proc. Conf. Inform. Sci. Syst*., vol. 1, pp. WP1-15 –17, Princeton, NJ, Mar. 2000.

[89] K. Hu and D. L. Wang, “Unvoiced speech separation from nonspeech interference via CASA and spectral subtraction,” *IEEE Trans. Audio, Speech, and Lang. Process*., vol. 19, no. 6, pp. 1600-1609, Aug. 2011.

[90] M. S. Khan, S. M. Naqvi, and J. A. Chambers, “A new cascaded spectral subtraction approach for binaural speech dereverberation and its application in source separation,” *in Proc. IEEE ICASSP*, pp. 6566-6570, May 2013.

[91] M. N. Schmidt and M. Morup, “Nonnegative matrix factor 2-D deconvolution for blind single channel source separation,” *in Proc. ICABSS 2006*, vol. 3889, pp. 700–707, Mar. 2006.

[92] 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 Trans. Circuits and Sys. I*, vol. 60, no. 3, pp. 662-675, 2013.

[93] N. Tengtrairat, B. Gao, W. L. Woo, and S. S. Dlay, “ Single-channel blind separation using complex 2-D histogram,” *IEEE Trans. Neural Netw*., vol. 24, no. 11, pp. 1722-1735, Nov. 2013.

[94] J. Woodruff and D.L. Wang, “Binaural localization of multiple sources in reverberant and noisy environments,” *IEEE Trans. Audio, Speech, and Lang. Process*., vol. 20, no. 5, pp. 1503–1512, Jul. 2012.

[95] R. Everson, and S.J. Roberts, “Particle Filters for Non-Stationary ICA,” *Advances in Independent Component Analysis. Springer*, pp. 25–41, 2000.

[96] A. Ahmed, C. Andrieu, A. Doucet and P.J.W. Rayner, “On-line non-stationary ICA using mixture models,” *IEEE International Conference on Acoustics, Speech, and Signal Process.*, pp.3148-3151, 2000.

[97] B. Gao, “Single channel blind source separation,” *Ph.D. Thesis*, Newcastle University, 2011.

[98] Z. Jin and D.L Wang, “A supervised learning approach to monaural segregation of reverberant speech”, *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 17, pp. 625-638. 2009.

[99] G. Hu and D. L. Wang, “Auditory segmentation based on onset and offset analysis”, *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 2, pp. 396–405, Feb. 2007.