SINGLE CHANNEL BLIND SOURCE SEPARATION

Bin Gao

BEng

MSc

A thesis submitted to the university of Newcastle for the degree of

Doctor of Philosophy

School of Electrical, Electronic and Computer Engineering

Faculty of Science, Agriculture and Engineering

April 2011

ABSTRACT

Single channel blind source separation (SCBSS) is an intensively researched field with numerous important applications. This research sets out to investigate the separation of monaural mixed audio recordings without relying on training knowledge. This research proposes a novel method based on variable regularised sparse nonnegative matrix factorization which decomposes an information-bearing matrix into two-dimensional convolution of factor matrices that represent the spectral basis and temporal code of the sources. In this work, a variational Bayesian approach has been developed for computing the sparsity parameters of the matrix factorization. To further improve the previous work, this research proposes a new method based on decomposing the mixture into a series of oscillatory components termed as the intrinsic mode functions (IMF). It is shown that IMFs have several desirable properties unique to SCBSS problem and how these properties can be advantaged to relax the constraints posed by the problem. In addition, this research develops a novel method for feature extraction using psycho-acoustic model. The monaural mixed signal is transformed to a cochleagram using the gammatone filterbank, whose bandwidths increase incrementally as the center frequency increases; thus resulting to non-uniform time-frequency (TF) resolution in the analysis of audio signal. Within this domain, a family of Itakura-Saito (IS) divergence based novel two-dimensional matrix factorization has been developed. The proposed matrix factorizations have the property of scale invariant which enables lower energy components in the cochleagram to be treated with equal importance as the high energy ones. Results show that all the developed algorithms presented in this thesis have outperformed conventional methods.

LIST OF CONTENTS

CHAPTER 1 INTRODUCTION TO THESIS 1

1.1 Background of Source Separation 1
1.1.1 BSS problem formulation 2
1.1.2 Classification of BSS 2
1.1.3 Applications of BSS 3
1.1.4 Single channel source separation (SCSS) 4
1.1.4.1 Time domain SCSS mixing model 4
1.1.4.2 Time-Frequency domain SCSS mixing model 6
1.2 Objectives of Thesis 7
1.3 Thesis Outline 9
1.4 Contribution 10

CHAPTER 2 OVERVIEW OF SINGLE CHANNEL SOURCE SEPARATION 14

2.1 *Supervised* SCSS 17
2.1.1 Frequency model-based SCSS 17
2.1.2 Underdetermined-ICA time model-based SCSS 19
2.2 *Unsupervised* SCSS 21
2.2.1 CASA-based SCBSS 21
2.2.2 Nonnegative matrix factorization based SCBSS 22
2.2.3 Independent subspace analysis based SCBSS 25
2.2.4 Empirical mode decomposition based SCBSS 27
2.3 Summary 28

CHAPTER 3 SINGLE CHANNEL BLIND SOURCE SEPARATION USING VARIABLE REGULARISED SPARSE FEATURES 30

3.1 Background 31
3.1.1 Two-dimensional nonnegative matrix factorization 31
3.1.2 Two-dimensional sparse nonnegative matrix factorization 32
3.2 Proposed Method 33
3.2.1 Formulation of the v-SNMF2D 36
3.2.1.1 Estimation of the spectral basis and temporal code 37
3.2.1.2 Estimation of the variable regularization parameter 38
3.3 Single Channel Blind Source Separation 45

3.3.1 Estimated sources.....45
3.3.2 Experiment set up.....46
3.3.3 Quality evaluation.....46
3.3.4 Impact of sparsity.....48
3.3.4.1 Estimated spectral basis and temporal code.....49
3.3.4.2 Audio source separation results.....51
3.3.4.3 Adaptive behavior of sparsity parameter.....54
3.3.5 Comparison with other sparse NMF-based SCBSS methods.....55
3.4 Summary.....**58**

CHAPTER 4 SINGLE CHANNEL BLIND SOURCE SEPARATION USING EMD-SUBBAND VARIABLE REGULARISED SPARSE FEATURES.....**59**

4.1 Background.....**60**
4.1.1 Empirical mode decomposition.....60
4.2 Proposed Separation Method.....**62**
4.2.1 Matrix representation of IMFs in TF domain.....66
4.2.2 Estimation of sub-sources.....68
4.3 Results and Analysis.....**70**
4.3.1 Effects on audio mixtures separation *with/without* EMD preprocessing.....71
4.3.2 Impacts of sparsity selection.....75
4.3.3 Comparison with other SCSS methods.....81
4.3.3.1 Underdetermined-based ICA SCSS method.....81
4.3.3.2 Hilbert subspace decomposition SCBSS method.....82
4.3.3.3 Comparison results.....82
4.3.3.4 Comparison with NMF-based SCBSS methods.....85
4.4 Summary.....**88**

CHAPTER 5 SINGLE CHANNEL BLIND SOURCE SEPARATION USING GAMMATONE FILTERBANK AND ITAKURA-SAITO MATRIX FACTORIZATION.....**89**

5.1 Time-Frequency Representation.....**91**
5.1.1 Classic spectrogram.....91
5.1.2 Log-frequency spectrogram (constant-Q transform).....92
5.1.3 Gammatone filterbank and Cochleagram.....93

5.1.4 Difference between classic spectrogram, log-frequency spectrogram and cochleagram 95
5.1.5 Separability analysis 97
5.1.5.1 Mixture of two sources 99
5.1.5.2 Mixture of multiple sources 101
5.2 Itakura-Saito based Two-dimensional Nonnegative Matrix Factorization Algorithms **103**
5.2.1 Quasi-EM based two-dimensional nonnegative matrix factorization using the IS divergence 104
5.2.1.1 Estimation of the spectral basis and temporal code using Quasi-EM method 105
5.2.2 Two-dimensional sparse nonnegative matrix factorization using the IS divergence 109
5.2.2.1 Estimation of the spectral basis and temporal code 111
5.2.3 Variable regularised two-dimensional nonnegative matrix factorization using the IS divergence 115
5.2.3.1 Estimation of the spectral basis and temporal code 119
5.2.4 Summary of the proposed algorithms 122
5.2.5 Estimation of sources 124
5.3 Experimental Results and Analysis **124**
5.3.1 Effects on separation based on different TF representation 125
5.3.2 Impacts of convolutive factors and different update methods 133
5.3.3 Impacts of NMF2D using different cost function 136
5.3.4 Impacts of regularizations selection 139
5.3.5 Comparison with other SCBSS methods 143
5.4 Summary **144**

CHAPTER 6 CONCLUSION OF THE THESIS **146**

6.1 Proposed *Unsupervised* Learning SCSS Methods **146**
6.2 Comparison of the Proposed SCBSS Methods **148**
6.3 Future Work **150**
6.3.1 Development of SCBSS method for non-stationary mixing model 150
6.3.2 Development of signal dependent TF representation 151
6.3.3 Development of Quasi-EM IS-vRNMF2D 152

REFERENCE **155**

LIST OF FIGURES

Figure 2.1: A general framework for *supervised* SCSS methods……. 15

Figure 2.2: A general framework for *unsupervised* SCSS methods……. 16

Figure 2.3: Basis functions derived from ICA algorithm……. 20

Figure 3.1: 3D-representation for $\mathbf{D}$, $\mathbf{H}$ and $\mathbf{\Lambda}$ …….. 35

Figure 3.2: Time-domain representation and TF representation of signal……... 49

Figure 3.3: Estimated $\mathbf{D}_i^r$ and $\mathbf{H}_i^\phi$ for case (i)…….. 50

Figure 3.4: Estimated $\mathbf{D}_i^r$ and $\mathbf{H}_i^\phi$ for case (ii)……. 50

Figure 3.5: Estimated $\mathbf{D}_i^r$ and $\mathbf{H}_i^\phi$ for case (iii)…… 50

Figure 3.6: Separated signals in spectrogram……... 51

Figure 3.7: Separated signals in time-domain…….. 52

Figure 3.8: Convergence trajectory of the sparsity……... 55

Figure 3.9: Separated signals in spectrogram……... 56

Figure 3.10: Separated signals in time-domain…… 56

Figure 4.1: EMD of male-female speech mixture showing the first six IMFs..... 62

Figure 4.2: Spectrogram of the IMFs……... 65

Figure 4.3: Core procedure of the proposed method……..... 70

Figure 4.4: Overall separation results of different mixtures *without* EMD preprocess… 72

Figure 4.5: Overall separation results of different mixtures *with* EMD preprocess…..… 72

Figure 4.6: Separation results *without* applying EMD preprocess….…….. 74

Figure 4.7: Estimated sub-sources for male speech……. 75

Figure 4.8: Estimated sub-sources for female speech…….. 75

Figure 4.9: Histogram of regularization parameter for each IMF…… 78

Figure 4.10: The separation results using different methods…… 79

Figure 4.11: Separation results of EMD-SNMF2D by using uniform regularization….. 80

Figure 4.12: Separation results of EMD-based SNMF2D using regularization schemes 80

Figure 4.13: Separation results based on different SCBSS methods…… 83

Figure 4.14: Overall comparison results…….. 84

Figure 4.15: Average ISNR using different number of components…… 87

Figure 5.1: Frequency responses of different filter bank……. 96

Figure 5.2: Overall separability performance for each mixture type……... 102

Figure 5.3: The flow chart of the proposed IS divergence based NMF2D algorithms… 123

Figure 5.4: Separation results in spectrogram…….. 127

Figure 5.5: Time-domain separated results based on spectrogram……... 127

Figure 5.6: Separation results in log-frequency spectrogram……... 128

Figure 5.7: Time-domain separated results based on log-frequency spectrogram……... 129

Figure 5.8: Separation results in cochleagram……. 130

Figure 5.9: Time-domain separated results based on cochleagram…….. 130

Figure 5.10: Estimated $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ based on cochleagram……. 131

Figure 5.11: Estimated $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ based on spectrogram……. 132

Figure 5.12: Estimated $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ based on log-frequency spectrogram…….. 132

Figure 5.13: Separation results by using the family of IS-based matrix factorizations... 136

Figure 5.14: Decomposition results using NMF2D with different cost function……. 138

Figure 5.15: Comparison results between IS-SNMF2D and IS-vRNMF2D…… 141

Figure 6.1: The proposed multi-dimensional signal dependent TF transform……. 152

LIST OF TABLES

Table 2.1: Summarise of learning methods for SCSS…….. 38

Table 3.1: Proposed v-SNMF2D algorithm…….. 45

Table 3.2: Performance comparison between different sparsity methods…… 53

Table 3.3: Performance comparison between different methods……. 57

Table 4.1: Domination proportion of each source signal to each IMF……. 64

Table 4.2: Assignment of regularization parameter…….. 76

Table 5.1: Overall separability performance…… 99

Table 5.2: Separability under different window length…… 100

Table 5.3: Quasi-EM IS-NMF2D algorithm…… 109

Table 5.4: IS-SNMF2D algorithm…… 114

Table 5.5: IS-vRNMF2D algorithm……. 122

Table 5.6: Summary of the proposed IS divergence based NMF2D algorithms…….. 123

Table 5.7: Separation results based on different TF representation…….. 126

Table 5.8: Separation results using different matrix factorization algorithm……... 134

Table 5.9: Separation results using NMF2D with different cost function…… 137

Table 5.10: Separation results using different regularization based matrix factorization 142

Table 5.11: Separation results using different SCBSS methods……... 143

Table 6.1: Summary of the proposed SCBSS methods…… 148

Table 6.2: Separation results using different SCBSS methods……. 149

ACKNOWLEDGEMENT

I wish to express my deepest gratitude to my supervisors Doctor Wai Lok Woo and Professor Satnam Dlay. I am very appreciating with my supervisors who bring me lots of invaluable support, guidance, and encouragement over the last few years. They have helped me understand complex signal processing concepts through our discussions. They have also welcomed my constant barrage of questions both in and out of their classes.

Thank-you also to my research colleagues and friends Jingyi Zhang, Xueyi Zhao, Bing Ji and Peng Liu for their support and encouragement.

Thank-you to my thesis examination committee members, for their time, constructive criticism, and feedback. This thesis is much the better for it.

I wish to thank my father Chun Gao and mother Yongping Jin for their support in all my efforts. I have been truly blessed to have them as my parents. They have provided me with unending love and support over the years and have encouraged me to follow my passion and go after my dreams. My warmest thanks go to my wife Xiangying Lu, who has been supporting me all the time.

ABBREVIATIONS/ACRONYMS

SS:	Source Separation
BSS:	Blind Source Separation
SCSS:	Single Channel Source Separation
MIR	Music Information Retrieval
AR	Auto-Regressive
CASA	Computational Auditory Scene Analysis
SCBSS	Single Channel Blind Source Separation
ICA	Independent Component Analysis
PCA	Principle Component Analysis
SVD	Singular Value Decomposition
ISA	Independent Subspace Analysis
EEG	Electroencephalography
MEG	Magnetoencephalography
ECG	Electrocardiogram
EMD	Empirical Mode Decomposition
HS	Hilbert Spectrum
TF	Time-Frequency
STFT	Short Time Fourier Transform
IS	Itakura-Saito
LS	Least Square

KL	Kullback-Leibler
GMM	Gaussian Mixture Models
NMF	Nonnegative Matrix Factorization
v-SNMF2D	Variable Regularised Two-dimensional Sparse Nonnegative Matrix Factorization
SNMF2D	Two-dimensional Sparse Nonnegative Matrix Factorization
NMF2D	Two-dimensional Nonnegative Matrix Factorization
KL-NMF2D	Kullback-Leibler Two-dimensional Nonnegative Matrix Factorization
LS-NMF2D	Least Square Two-dimensional Nonnegative Matrix Factorization
IS-NMF2D	Itakura-Saito Two-dimensional Nonnegative Matrix Factorization
IS-SNMF2D	Itakura-Saito Two-dimensional Sparse Nonnegative Matrix Factorization
IS-vRNMF2D	Itakura-Saito Variable Regularised Two-dimensional Nonnegative Matrix Factorization
IS-NMF	Itakura-Saito Nonnegative Matrix Factorization
IS-SNMF	Itakura-Saito Sparse Nonnegative Matrix Factorization
MSE:	Mean Square Error
ML:	Maximum Log-likelihood
MAP:	Maximum a Posterior

IBM	Ideal Binary Mask
ERB	Equivalent Rectangular Bandwidth
Fro	Frobenius Norm
WDO	W-Disjoint Orthogonality
ISNR	Improvement of Signal-to-Noise Ratio
PSR	Preserved Signal Ratio
SIR	Signal-to-Interference Ratio
SDR	Signal-to-Distortion Ratio
SAR	Source-to-Artifacts Ratio

LIST OF PUBLICATIONS

- Bin Gao, W.L. Woo, S.S. Dlay, "Single Channel Source Separation Using EMD-Subband Variable Regularised Sparse Features", *IEEE Transactions on Audio, Speech and Language Processing (Regular paper accepted and in press).*

- Bin Gao, W.L. Woo, S.S. Dlay, "Single Channel Audio Source Separation", *WSEAS Transactions on Signal Process*, vol 4, col. 4, Jan, 2008.

- Bin Gao, W.L. Woo, S.S. Dlay, "Performance Analysis of Single Channel Blind Source Separation", *The 2nd WSEAS International Conference on COMPUTER ENGINEERING and APPLICATIONS (CEA'08)*, Cambridge (UK), 20th Feb 2008.

- Bin Gao, W.L. Woo, S.S. Dlay, "Single Channel Blind Source Separation using the Best Characteristic Basis", *3d International Conference on Information & Communication Technologies: from Theory to Applications – (ICTTA'08)* , Syria, 2008, pp:1-5.

- Bin Gao, W.L. Woo, S.S. Dlay, "Single Channel Audio Source Separation by Exploiting Characteristic Filters", *International Symposium on communications systems, networks and digital signal processing (CSNDSP 2008)*, Graz, Jan, 27th, 2008, pp: 572-576.

- A.M. Darsono, Bin Gao, W.L. Woo, S.S. Dlay, "NONLINEAR SINGLE CHANNEL SOURCE SEPARATION", *International Symposium on communications systems, networks and digital signal processing (CSNDSP 2010)*, 2010, pp: 507-511.

- Bin Gao, W.L. Woo, S.S. Dlay, "Variational Bayesian Regularized Two-Dimensional Nonnegative Matrix Factorization", *submitted to IEEE Transactions on Neural Networks.*

CHAPTER 1

INTRODUCTION TO THESIS

1.1 Background of Source Separation

In recent years, source separation (SS) has received considerable attention from both signal processing and neutral network researchers. Source separation means when given a mixture signal, it can be separated with independent components. In this area, the methods to solve SS problem can be categorized either as *supervised* SS methods or *unsupervised* SS methods. The terms "*supervised"* and "*unsupervised"* denote the requirement of training information and without training information, respectively. Blind (or *unsupervised*) source separation (BSS) refers to the powerful technique of separating a mixture of underline sources without training data nor a priori knowledge about the original sources and parameters of the mixing system. During the last decade, tremendous developments have been achieved in the area of BSS [1-11] and BSS has become one of the most promising and exciting topics with solid theoretical foundations and potential applications in the fields of neural computation, advanced statistics, and signal processing. BSS has been successfully applied in various fields such as speech enhancement, recognition, biomedical image processing, image processing, remote sensing, communication systems, exploration seismology, geophysics, econometrics, data mining and neural networks.

1.1.1 BSS problem formulation

A general BSS problem can be mathematically defined as follows: A set of observations $\mathbf{y}(t)=\left[y_1(t)\ y_2(t)\cdots y_{N_o}(t)\right]^{\mathbf{T}}$ which are random processes is generated as a mixture of underlying source signals $\mathbf{x}(t)=\left[x_1(t)\ x_2(t)\cdots x_{N_s}(t)\right]^{\mathbf{T}}$ according to:

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

where $\mathbf{M}$ is the unknown mixing matrix of dimension $N_o \times N_s$ and t is the time or sample index. The technique of BSS aims to estimate both the original sources $\mathbf{x}(t)=\left[x_1(t)\ x_2(t)\cdots x_{N_s}(t)\right]^{\mathbf{T}}$ and the mixing matrix $\mathbf{M}$ using only the observations $\mathbf{y}(t)=\left[y_1(t)\ y_2(t)\cdots y_{N_o}(t)\right]^{\mathbf{T}}$. It is noted that (1.1) represents a simplified model which may not be an accurate representation of the real environment. Issues such as nonlinear distortions, propagation delay of signals and noise should be taken into account and evaluated in order to present a realistic model. Hence the need for further research has led to various branches of research in BSS.

1.1.2 Classification of BSS

A review of current literature shows that there exits three main classification of BSS. These include: Linear and Nonlinear BSS; Instantaneous and Convolutive BSS; Overcomplete and Underdetermined BSS. In the first classification, linear algorithms dominate the BSS research field due to its simplicity in analysis and its explicit separability. Linear BSS assumes that the mixture is represented by a linear combination of sources [1,

6, 12], as defined in (1.1). Extension of BSS for solving nonlinear mixtures has also been introduced [13-19]. This model which takes nonlinear distorted signals into consideration offers a more accurate representation of a realistic environment. In the second classification, when the observed signals consist of combinations of multiple time-delayed versions of the original sources and/or mixed signals themselves, the system is referred as the convolutive mixture. Otherwise, the absence of time delays results in the instantaneous mixture of observed signals. An example of the simplest and conventional form of linear instantaneous BSS model is the linear mixture, which is expressed in (1.1). Finally, when the number of observed signals more than the number of independent sources ($N_o > N_s$), this refers to overcomplete BSS. On the other hand, when the number of observed signals smaller than the number of independent sources ($N_o < N_s$), this becomes to underdetermined BSS.

1.1.3 Applications of BSS

Due to the diverse promising and exciting applications, BSS has attracted a substantial amount of attention in both the academic field as well as the industry area. During the last decade, tremendous developments have been achieved in the application of BSS, particularly in wireless communication, medical signal processing, geophysical exploration and image enhancement/recognition [3, 20-36]. The so-called “cocktail party” problem within the BSS context refers to the phenomenon of extracting original voice signals of the speakers from the mixed signals recorded from several microphones. Similar examples in the field of radio communication involve the observations which correspond to the outputs

of several antenna elements in response to several transmitters which represents the original signals. In the analysis of medical signals, electroencephalography (EEG), magnetoencephalography (MEG) and electrocardiogram (ECG) [3, 11, 20, 22] represents the observations and BSS is used as a signal processing tool to assist noninvasive medical diagnosis. BSS has also been applied to the data analysis in other areas such as finance and seismology. Further evidence of these applications can be found in [23-36]. In addition, BSS [11] has been applied in chemometrics, for example to determine the spectra and concentration profiles of chemical components in an unresolved mixture. Especially, in most audio applications, applying simple processing to a certain source within a polyphonic mixture is virtually impossible to separate signals. This creates a need for source separation methods, which first separate the mixture into sources, and then continue the separated sources individually. These applications include audio coding, analysis, and manipulation of audio signals.

1.1.4 Single channel source separation (SCSS)

1.1.4.1 Time domain SCSS mixing model

In this thesis, the special case of instantaneous underdetermined SS problem termed as single channel source separation is focused. In general case and for many practical applications (e.g. audio processing) only one-channel recording is available and in such cases conventional source separation techniques are not appropriate. However, this is the most interesting case seen from a hearing instrument industry point of view such that the

specific applications [37] are described as following:

1. It is often desirable to process a single instrument in a recording. For example, in a single microphone recording of vocals and acoustic guitar, we might want to adjust the volume of the guitar or shift the pitch of the vocals. Thus, if the individual instruments can be distinguished from a mixture, they can be processed individually.
2. Speech recognition in the presence of noise, particularly heavy non-stationary noise, is a challenging problem. Speech recognition performance could improve if the speech can be distinguished from the noise and perform recognition on the portion of the mixture that corresponds to speech.
3. Musicians often spend large amounts of time trying to listen to a song and learn the part of a specific instrument by ear. This task becomes more difficult when the given piece of music has numerous parts by numerous instruments (which is often the case). If the instrument of interest can be extracted, it could simplify the task of the musician. In practice, this is a common problem for guitar players that try to learn their parts from recordings of bands.
4. Automatic music transcription of polyphonic music is a challenging problem. If each of the instruments in the mixture can be modeled, they can be transcript individually.
5. A number of music information retrieval (MIR) tasks involve extracting information from individual sources. For example, guitar and piano parts could be good indicators of the key of a song. However, the percussion part will rarely have any useful information for this task. Although, the sound mixture can be directly used for many of these tasks, extracting the information from the right source could improve the

performance.

Other field such as neuroscience (spike sorting) [38, 39] seeks to elucidate concerns the mechanisms used by dedicated parts of brains to perform specific tasks. This is also performed by single channel. This leads to the SCSS research area where the problem can be simply treated as one observation instantaneous mixed with several unknown sources:

$$y(t) = \sum_{i=1}^{N_s} x_i(t) \tag{1.2}$$

where $i = 1, \ldots, N_s$ denotes number of sources and the goal is to estimate the sources $x_i(t)$ when only the observation signal $y(t)$ is available. This is an underdetermined system of equation problem. Recently, new advances have been achieved in SCSS and this can be categorized either as *supervised* SCSS methods or *unsupervised* SCSS methods. More details of the above methods will be reviewed in Chapter 2.

1.1.4.2 Time-Frequency domain SCSS mixing model

Audio mixtures of sources in the time domain can be modeled as (1.2). In the TF domain, the mixture (1.2) becomes:

$$Y(f, t_s) = \sum_{i=1}^{N_s} X_i(f, t_s) \tag{1.3}$$

where $Y(f, t_s)$, $X_i(f, t_s)$ denote TF components which can be obtained by applying short time Fourier transform (STFT) or other TF analysis methods. The analysis of different TF transform will be discussed in detail in Chapter 5. Here, the time slots are given by $t_s = 1, 2, \ldots, T_s$ while frequencies are given by $f = 1, 2, \ldots, F$. F and T_s represent the total frequency units and time slots in the TF domain, respectively. Note that in (1.3),

each component is a function of t_s and f. The power spectrogram is defined as the squared magnitude of (1.3):

$$\left|Y(f,t_s)\right|^2 = \sum_{i,j(i\neq j)}^{N_s} \left(\left|X_i(f,t_s)\right|^2 + \left|X_j(f,t_s)\right|^2 + 2\left|X_i(f,t_s)\right|\left|X_j(f,t_s)\right|\cos\left(\theta_{i,j}(f,t_s)\right)\right) \qquad (1.4)$$

where $\theta_{i,j}(f,t_s)$ measures the projection of $\left|X_i(f,t_s)\right|$ onto $\left|X_j(f,t_s)\right|$ [8]. For a large sample size, $X_i(f,t_s)$ and $X_j(f,t_s)$ are assumed orthogonal and hence, the cross term $\theta_{i,j}(f,t_s)=\pi/2$. However, for finite sample size, this assumption on $\theta_{i,j}(f,t_s)=\pi/2$ may not hold and $2\left|X_i(f,t_s)\right|\left|X_j(f,t_s)\right|\cos\left(\theta_{i,j}(f,t_s)\right)$ is treated as the residual noise. In our simulation experiments, all testing recordings are using large sample size and thus a matrix representation for (1.4) is given as follows:

$$\left|\mathbf{Y}\right|^{.2} \approx \sum_{i=1}^{N_s} \left|\mathbf{X}\right|_i^{.2} \qquad (1.5)$$

where $\left|\mathbf{Y}\right|^{.2} = \left[\left|Y(f,t_s)\right|^2\right]_{t_s=1,2,\ldots,T_s}^{f=1,2,\ldots,F}$ and $\left|\mathbf{X}\right|_i^{.2} = \left[\left|X_i(f,t_s)\right|^2\right]_{t_s=1,2,\ldots,T_s}^{f=1,2,\ldots,F}$ are two-dimensional matrices (row and column vectors represent the time slots and frequencies or frequency bins, respectively) which denotes the power TF representation of (1.2). The superscript "." is element-wise operation. Eqn. (1.5) is a synthesis equation since it describes how $\left|\mathbf{Y}\right|^{.2}$ is generated as a mixing of $\left|\mathbf{X}\right|_i^{.2}$.

1.2 Objectives of Thesis

The aims of this thesis are to investigate SCSS methods in terms of its fundamental theory, assumptions, applications and limitations as well as further develop new frameworks of single channel blind source separation (SCBSS) for audio mixture. Three novel methods have been imposed, namely, SCBSS using variable regularised sparse

features; SCBSS using empirical model decomposition (EMD) subband variable regularised sparse features; and SCBSS using cochleagram and Itakura-Saito divergence based matrix factorization. Rigorous mathematical derivations and simulations are carried out to substantiate the efficacy of the proposed algorithms.

The objectives of this thesis are listed as follows:

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

ii). To develop useful audio signal analysis algorithms that have desirable properties unique to SCBSS problem and use these properties can be advantaged to relax the constraints posed by the problem.

iii). To develop a measure for audio signal separability and analysis the source separation in different TF representation.

iv). To develop novel methods for SCBSS which addresses the following:

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

1.3 Thesis Outline

This research is carried out with the focus predominantly on single channel audio mixtures. Three novel generative methods for SCBSS are proposed in this thesis. Real time testing has been conducted and it is shown that the proposed methods gives superior performance over other existing approaches.

In Chapter 2, an overview of recent SCSS methods is given, which reviews a major SCSS methods. The start of this chapter is by introducing *supervised* SCSS methods which includes both time and frequency model based methods. The current *unsupervised* SCSS methods have also been reviewed in this chapter.

In Chapter 3, a new *unsupervised* SCSS method is developed to separate music instantaneous mixture. A novel matrix factorization algorithm is proposed to decompose an information-bearing matrix (TF representation of mixture) into two-dimensional convolution of factor matrices that represent the spectral basis and temporal code of the sources. In addition, a variational Bayesian approach is derived to compute the sparsity parameters for optimizing the matrix factorization. Simulation of single channel music source separation is carried out to effectiveness of the proposed method.

In Chapter 4, a new *unsupervised* SCSS method is developed to separate audio instantaneous mixture (the audio mixture include music&music, speech&music and speech&speech). The idea is based on decomposing the mixture into a series of oscillatory components termed as the intrinsic mode functions. In order to decompose the

sub-mixtures (IMF), the proposed variable regularised sparse two-dimensional matrix factorization (as detailed in Chapter 3) is incorporated. Simulation of single channel audio source separation is carried out to effectiveness of the proposed method.

In Chapter 5, a new *unsupervised* SCSS method is developed to separate music&music and speech&music instantaneous mixture. The idea is based on time-frequency analysis and feature extraction. The monaural mixed signal is transformed to a cochleagram using the gammatone filterbank, whose bandwidths increase incrementally as the center frequency increases; thus resulting to non-uniform TF resolution in the analysis of audio signal. In addition, a family of IS divergence based novel two-dimensional matrix factorization algorithms has been derived to estimate the spectral basis and temporal code of the sources. The proposed method is a more complete and recovers high quality estimates of the individual sources. Several comparisons and simulation are carried out to effectiveness of the proposed method.

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

1.4 Contribution

The contribution of this thesis is to generate novel solutions for SCBSS of audio mixtures. Hence, the proposed methods overcome the limitations associated with the conventional approaches. This thesis presents three novel methods with a significant improvement in performance in terms of both accuracy and versatility. The following

outlines the contributions of this thesis:

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

ii). A novel variable regularised two-dimensional sparse nonnegative matrix factorization (v-SNMF2D) is proposed. The proposed model allows overcomplete representation by allowing many spectral and temporal shifts which are not inherent in the nonnegative matrix factorization (NMF) and sparse NMF (SNMF) models. In addition, imposing sparseness is necessary to give unique and realistic representations of the non-stationary audio signals. Unlike the conventional two-dimensional sparse NMF factorization (SNMF2D), the proposed model imposes sparseness on temporal code $\mathbf{H}$ element-wise so that *each individual code* has its own distribution. Therefore, the sparsity parameter can be individually optimized for each code. This overcomes the problem of under- and over-sparse factorization. In addition, each sparsity parameter in the proposed model is learned and adapted as part of the matrix factorization. This bypasses the need of manual selection as in the case conventional approach.

iii). A new framework for SCBSS based on the EMD and v-SNMF2D is proposed.

- Audio signals are mostly non-stationary and the EMD decomposes the mixed signal into a collection of oscillatory basis components termed as intrinsic mode functions (IMFs) which contain the original source basic properties. In the proposed scheme, instead of processing the mixed signal directly, the IMFs are utilized as the new set of observations. The impetus behind this is that the degree of mixing of the sources in the IMF domain is now less ambiguous and thus, the

dominating source in the mixture is more easily detected. Moreover, the spectral and temporal patterns (i.e. the spectral bases and temporal codes, respectively) associated with each IMF are now simpler and sparser than that of the mixed signal. As such, these patterns can be extracted using a suitably designed sparse algorithm.

- The proposed v-SNMF2D benefits conventional SNMF2D in terms of improved accuracy in resolving spectral basis and temporal code. This benefit has been extended to single channel source separation by merging the proposed v-SNMF2D with EMD.

iv). A novel framework to solve SCBSS based on the cochleagram TF representation and a family of IS divergence based novel two-dimensional nonnegative matrix factorization is proposed.

- Construction of 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 than in the spectrogram and log-frequency spectrogram. A measurement of separability in the TF domain has been derived for SCSS and a quantitative performance measure has been developed to evaluate how separable the sources are given the monaural mixed signal. In addition, the ideal condition has been identified when the sources are perfectly separable.

- A family of IS based novel two dimensional nonnegative matrix factorization algorithms has been developed to extract the spectral basis and temporal code. The proposed factorization is scale invariant whereby the lower energy components in the TF representation can be treated with equal importance as the higher energy components. Within the context of SCBSS, this property enables the spectral-temporal features of the sources that are characterized by a large dynamic range to be estimated with higher accuracy. This is to be contrasted with the matrix factorization based on Least Square (LS) distance [29] and Kullback-Leibler (KL) divergence [30] where both methods favor the high-energy components but neglect the low-energy components.

CHAPTER 2

OVERVIEW OF SINGLE CHANNEL SOURCE SEPARATION

This chapter gives an overview of the existing learning methods for SCSS which have proven to produce applicable separation results in the case of audio signals. These methods can be categorized either as *supervised* SCSS methods or *unsupervised* SCSS methods.

For *supervised* SCSS methods, the probabilistic models of the source are trained as a prior knowledge by using some or the entire source signals. The mixture is first transformed into an appropriate representation, in which the source separation is performed. The source models are either constructed directly based on knowledge of the signal sources, or by learning from training data (e.g. using Gaussian mixture model construct source models either directly based on knowledge of signal sources, or by learning from isolated training data). In the inference stage, the models and data are combined to yield estimates of the sources. This category predominantly includes the frequency model-based SCSS methods [40–44] where the prior bases are modeled in time-frequency domain (e.g. spectrogram or power spectrogram), and the underdetermined-ICA time model-based SCSS method [45–47] which the prior bases are modeled in time domain. Figure 2.1 shows a general framework for *supervised* SCSS methods.

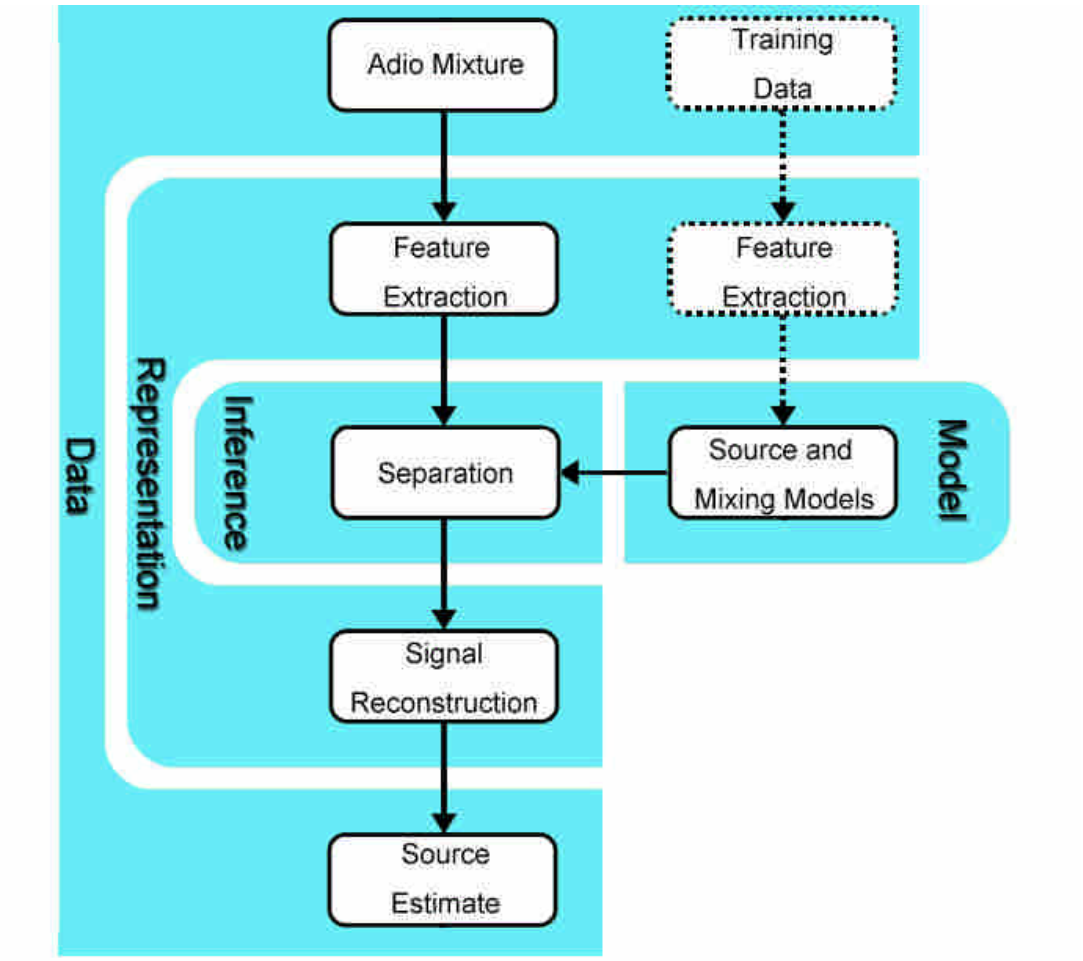


Figure 2.1: A general framework for *supervised* SCSS methods.

In Figure 2.1, the input to the separation system is the audio mixture and the relative training data for the source models. The mixture is transformed into a suitable representation and combined with the source models and mixing model in the inference stage, that either directly or through a signal reconstruction method computes estimates of the separated sources.

For *unsupervised* SCSS methods, this denotes the separation of completely unknown sources without using additional training information. These methods typically rely on the assumption that the sources are non-redundant, and the methods are based on, for example, decorrelation, statistical independence, or the minimum description length principle. Figure

2.2 shows a general framework for *unsupervised* SCSS methods.

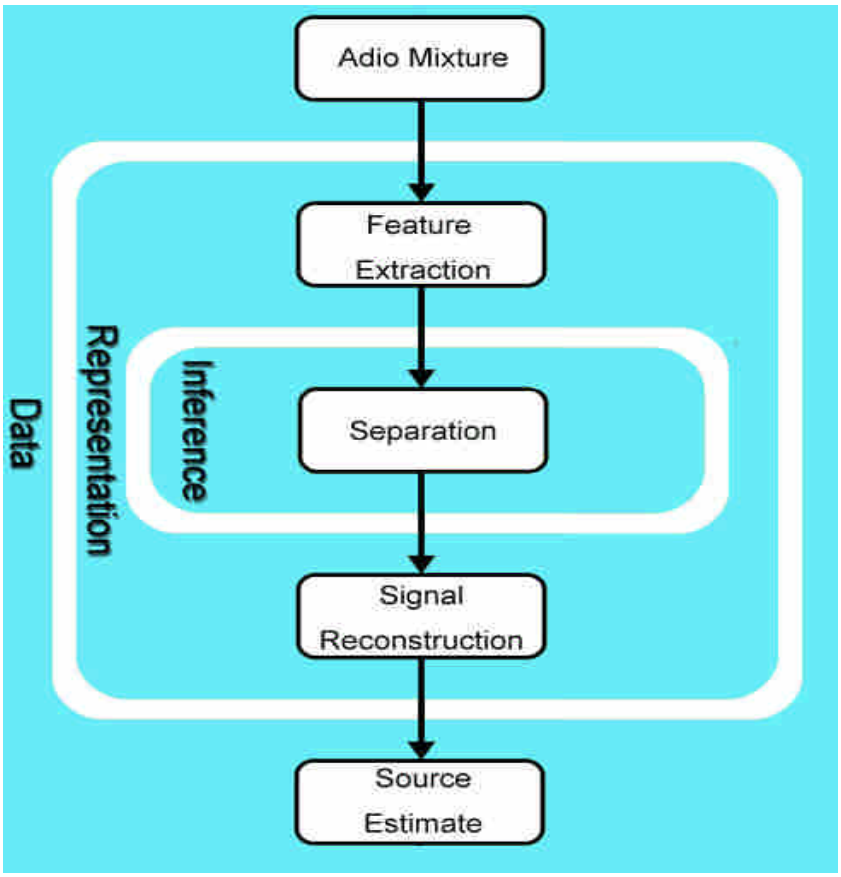


Figure 2.2: A general framework for *unsupervised* SCSS methods.

In Figure 2.2, 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. This category includes several widely used methods: Firstly, the CASA-based SCBSS methods [48–54] whose goal is to replicate the process of human auditory system by exploiting signal processing approaches (e.g. notes in music recordings) and grouping them into auditory streams using psycho-acoustical cues. Secondly, the subspace technique based SCBSS methods using NMF [55] or independent subspace analysis (ISA) [56] which usually factorizes the spectrogram of the input signal into elementary components. Of special interest, EMD [57]

based SCBSS methods which can separate audio mixed signal in time domain and recover sources by combing other data analysis tools, e.g. independent component analysis (ICA) or principle component analysis (PCA).

All the methods mentioned above (*supervised* SCSS and *unsupervised* SCSS methods) are formulated using a linear instantaneous signal model which is explained in Section 1.1.4.

2.1 *Supervised* SCSS

Here, the *supervised* SCSS refers to the single channel source separation applications where prior (or training) information about the sources is available. For example, the source instruments can be defined manually by the user, and in this case it is usually advantageous to optimize the algorithm by using training signals where each instrument (or source) is present in isolation.

2.1.1 Frequency model-based SCSS

The frequency model-based SCSS methods [40] are similar to the model-based audio signal enhancement techniques. These methods exploit the hidden Markov models (HMM) or other algorithms such as e.g. nonnegative matrix factorization, sparse code, etc to generate codebook of audio signals. The HMM based methods are widely used and the heart of these frequency model-based SCSS methods is the approximation of the posterior $p(\underline{\mathbf{x}}_{i,t_s},\ldots,\underline{\mathbf{x}}_{N_s,t_s}\,|\,\underline{\mathbf{y}}_{t_s})$ by Gaussian distribution [58, 59]. The posterior distribution can be

expressed as:

$$p\left(\underline{\mathbf{x}}_{i,t_s},\ldots,\underline{\mathbf{x}}_{N_s,t_s}\,|\,\underline{\mathbf{y}}_{t_s}\right)\propto p\left(\underline{\mathbf{y}}_{t_s}\,|\,\underline{\mathbf{x}}_{i,t_s},\ldots,\underline{\mathbf{x}}_{N_s,t_s}\right)\prod_{i=1}^{N_s}p\left(\underline{\mathbf{x}}_{i,t_s}\right) \tag{2.1}$$

where $\underline{\mathbf{x}}_{i,t_s}=\begin{bmatrix}|X_i(1,t_s)|^2\\ \vdots\\ |X_i(F,t_s)|^2\end{bmatrix}$, $\underline{\mathbf{x}}_{N_s,t_s}=\begin{bmatrix}|X_{N_s}(1,t_s)|^2\\ \vdots\\ |X_{N_s}(F,t_s)|^2\end{bmatrix}$ and $\underline{\mathbf{y}}_{t_s}=\begin{bmatrix}|Y(1,t_s)|^2\\ \vdots\\ |Y(F,t_s)|^2\end{bmatrix}$ are the power spectrum vectors. The priori information for the sources in probability density functions is assumed as Gaussian mixture models (GMM) is defined as:

$$p(\underline{\mathbf{x}}_{i,t_s})=\sum_{k_i}w_{i,k_i}^{GMM}\,N\left(\underline{\mathbf{x}}_{i,t_s};\underline{\mathbf{u}}_{i,k_i},\boldsymbol{\Sigma}_{i,k_i}\right) \tag{2.2}$$

$$\text{where}\quad N\left(\underline{\mathbf{x}}_{i,t_s};\underline{\mathbf{u}}_{i,k_i},\boldsymbol{\Sigma}_{i,k_i}\right)=\frac{\exp\left(-\frac{1}{2}\left(\underline{\mathbf{x}}_{i,t_s}-\underline{\mathbf{u}}_{i,k_i}\right)^{\mathrm{T}}\boldsymbol{\Sigma}_{i,k_i}^{-1}\left(\underline{\mathbf{x}}_{i,t_s}-\underline{\mathbf{u}}_{i,k_i}\right)\right)}{(2\pi)^{F/2}\det\left(\boldsymbol{\Sigma}_{i,k_i}\right)^{1/2}} \tag{2.3}$$

where $\underline{\mathbf{u}}_{i,k_i}$ is the mean vector, $\boldsymbol{\Sigma}_{i,k_i}$ is the covariance matrice, $w_{i,k_i}^{GMM}\geq 0$ is the weight (satisfying $\sum_{k_i}w_{i,k_i}^{GMM}=1$), $\{k_i\}$ denotes the hidden states of i^{th} source, ‘det’ denotes determinant and ‘T’ is matrix transpose. In frequency model-based SCSS methods, the $\underline{\mathbf{u}}_{i,k_i}$ and $\boldsymbol{\Sigma}_{i,k_i}$ of each source are trained before separation process. Within these prior parameters of each source, the factorial hidden Markov model (FHMM) can be employed to separate the mixture. Good separation requires detailed source models that might employ thousands of full spectral states e.g. in [58], GMMs with 8000 states were required to accurately represent one person’s speech for a source separation task. The large state space is required because it attempts to capture every possible instance of the signal. However, these model-based SCSS techniques are computationally intensive not only for training the prior parameters but also for presenting many difficult challenges during both the learning and inference stages.

2.1.2 Underdetermined-ICA time model-based SCSS

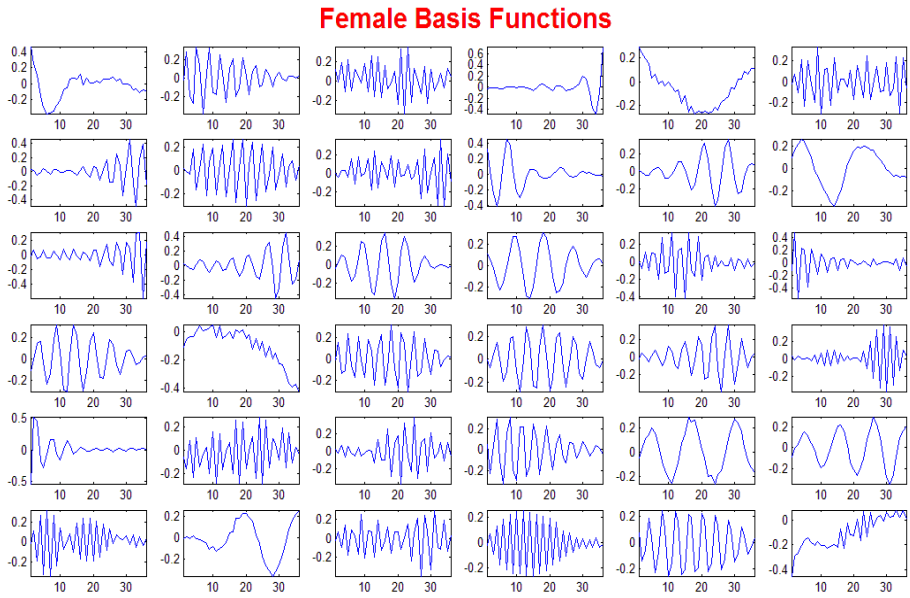
In the underdetermined-ICA time model-based SCSS method [47], the key point is to exploit a priori knowledge of sources such as the basis functions to generate sparse coding. The sources are then projected onto a set of basis functions whose coefficients are as sparse as possible. Thus the separation algorithm use hybrid of maximum likelihood (ML) and maximum a posteriori (MAP) estimators to recover the independent components. In this case, the observation model is expressed as:

$$y(t) = \sum_{i=1}^{N_s} m_i x_i(t) \tag{2.4}$$

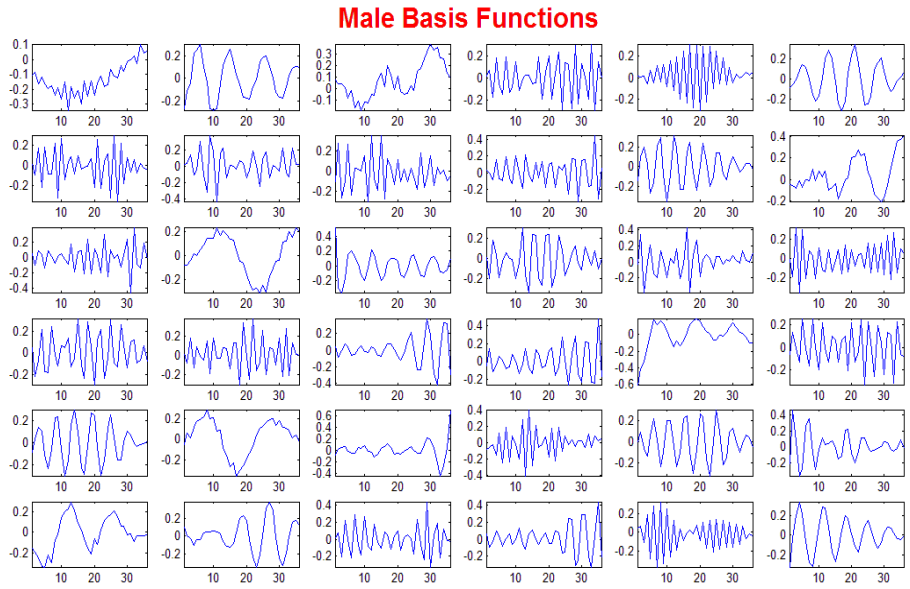
where m_i is the i^{th} source mixing coefficient, while the individual sources are constructed by basis functions and their coefficients. The basis functions and coefficients learned by ICA constitute an efficient representation of the given time-ordered sequences of a sound source by estimating the maximum likelihood densities. Hence the individual sources can be expressed as:

$$\mathbf{x}_i^t = \sum_{n_u=1}^{N_u} \mathbf{a}_{i,n_u}^{ICA} s_{i,n_u}^t = \mathbf{A}_i^{ICA} \mathbf{s}_{i,n_u}^t \quad \text{where} \quad \mathbf{s}_i^t = \mathbf{W}_i^{ICA} \mathbf{x}_i^t \tag{2.5}$$

where N_u is number of basis functions, $\mathbf{a}_{i,n_u}^{ICA}$ is the n_u^{th} basis functions of i^{th} source in the form of O dimensional column vector. Here small length O with $O << T$ from independent source is employed to analysis. The time duration is from t to $t+O-1$, an O dimensional column vector $\mathbf{x}_i^t = \left[x_i(t), x_i(t+1), \ldots, x_i(t+O-1) \right]^{\mathbf{T}}$. $\mathbf{A}_i^{ICA} = \left[\mathbf{a}_{i1}^{ICA}, \mathbf{a}_{i2}^{ICA}, \ldots, \mathbf{a}_{i,N_u}^{ICA} \right]$ is the basis matrix which contains basis functions of i^{th} source signal and $\mathbf{s}_i^t$ is the basis coefficient. An example of basis functions based on ICA algorithm is shown in Figures 2.3 (A) and (B).



(A)



(B)

Figure 2.3: (A) and (B) represent male and female basis functions derived from ICA algorithm.

For simplicity, the authors consider two sources mixed in single channel (i.e. $N_s = 2$). After obtaining prior source basis functions, the estimated sources can be recovered by employing maximum likelihood estimator follows as:

$$\begin{aligned} L^U &= \log\left\{ p\left(x_1(1,\ldots,T)\middle|\mathbf{W}_1^{ICA}\right) p\left(x_2(1,\ldots,T)\middle|\mathbf{W}_2^{ICA}\right)\right\} \\ &\propto \sum_{t=1}^{Tn}\sum_{n_u=1}^{N_u}\left[\log p\left(s_{1,n_u}^t\right)+\log p\left(s_{2,n_u}^t\right)\right] \end{aligned} \tag{2.6}$$

Thus the estimated sources can be obtained by following gradients-based learning rule:

$$\begin{aligned} \frac{\partial L^U}{\partial \hat{z}_1^t} &\propto \sum_{o=1}^{O}\left[m_2\sum_{n_u=1}^{N_u}\left\{\varphi(s_{1,n_u}^{tn})w_{1,n_u,o}^{ICA}\right\} - m_1\sum_{n_u=1}^{N_u}\left\{\varphi(s_{2,n_u}^{tn})w_{2,n_u,o}^{ICA}\right\}\right] \\ \frac{\partial L^U}{\partial \hat{z}_2^t} &\propto \sum_{o=1}^{O}\left[m_1\sum_{n_u=1}^{N_u}\left\{\varphi(s_{1,n_u}^{tn})w_{1,n_u,o}^{ICA}\right\} - m_2\sum_{n_u=1}^{N_u}\left\{\varphi(s_{2,n_u}^{tn})w_{2,n_u,o}^{ICA}\right\}\right] \end{aligned} \tag{2.7}$$

with $tn = t-o+1 \quad \forall o = 1,\ldots,O$, $w_{i,n_u,o}^{ICA} = \mathbf{W}_i^{ICA}(n_u,o)$, $z_i^t = m_i x_i^t$ and $\varphi(s) = \dfrac{\partial \log p(s\,|\,q_m,u,\sigma)}{\partial s}$. The coefficients of Gaussian exponential density model $p(s\,|\,q_m,u,\sigma) \propto \exp\left[-\left|\dfrac{s-u}{\sigma}\right|^{q_m}\right]$ is determined by parameters mean u, exponent q_m and $\sigma = \sqrt{E\left[(s-u)^2\right]}$ where $E[\bullet]$ denotes the expectation.

2.2 *Unsupervised* SCSS

Here, the un*supervised* SCSS (or SCBSS) means--in single channel source separation applications, the training information about the sources is not provided.

2.2.1 CASA-based SCBSS

The Computational Auditory Scene Analysis (CASA)-based SCBSS methods [48] whose goal is to replicate the process of human auditory system by exploiting signal processing approaches (e.g. notes in music recordings) and grouping them into auditory streams using psycho-acoustical cues. The main idea is based on exploiting an appropriate transformation such as STFT or cochleagram TF representation whereby the observation mixture is

segmented into time-frequency cells which are then used to characterize note objects by harmonicity, common onset, correlated modulation and duration of sinusoidal partials, and finally build note streams based on pitch proximity [60-62]. Hence, they segregate the instruments playing different pitch range into different streams. The estimated sources are then reconstructed by using some criteria to group the clusters of similar features in the TF domain. Nevertheless, CASA-based SCBSS techniques cannot efficiently segregate instruments playing in the same pitch range into different streams. They also cannot replicate the entire process performed in the auditory system since the process beyond the auditory nerve is not well studied. In addition, it is difficult to group the sources if one of them is assumed to be fully voiced.

2.2.2 Nonnegative matrix factorization based SCBSS

Recently, solutions to SCBSS using factorization-based approaches have gained popularity [63–71]. They exploit an appropriate TF analysis on the mono input recordings, (1.5) yielding a TF representation which can be decomposed as:

$$\left|\mathbf{Y}\right|^{.2} \approx \mathbf{DH} \tag{2.8}$$

where $\left|\mathbf{Y}\right|^{.2} \in \Re_{+}^{F\times T_s}$ is the power TF representation of mixture $y(t)$ which can be further factorized as the product of two nonnegative matrices, $\mathbf{D} \in \Re_{+}^{F\times I}$ and $\mathbf{H} \in \Re_{+}^{I\times T_s}$. If I is chosen to be $I = T_s$, no benefit is achieved in terms of representation. Thus the idea is to determine $I < T_s$ so the data matrix $\mathbf{D}$ can be compressed and reduced to its integral components such as $\mathbf{D}$ is a matrix containing only a set of spectral basis vectors, and $\mathbf{H}$ is an encoding matrix which describes the amplitude of each basis vector at each time point.

Nonnegative matrix factorization (NMF) [72–74] has been proven to be a very useful tool in a variety of signal processing fields. Recently, NMF methods have successfully been exploited for separating drums from polyphonic music [75] and automatic transcription of polyphonic music [76]. In addition, NMF gives a parts-based decomposition [77]. Commonly used cost functions for NMF are the generalized Kullback-Leibler (KL) divergence and Least Square (LS) distance which have been introduced in [74], respectively, as:

$$\begin{aligned} C_{KL}(\left|\mathbf{Y}\right|^{.2}\left\|\left|\hat{\mathbf{Y}}\right|^{.2}\right.) &= \sum_{f,t_s}\left(\left|\mathbf{Y}_{f,t_s}\right|^2 \log\frac{\left|\mathbf{Y}_{f,t_s}\right|^2}{\left|\hat{\mathbf{Y}}_{f,t_s}\right|^2} - \left|\mathbf{Y}_{f,t_s}\right|^2 + \left|\hat{\mathbf{Y}}_{f,t_s}\right|^2\right) \\ C_{LS}(\left|\mathbf{Y}\right|^{.2}\left\|\left|\hat{\mathbf{Y}}\right|^{.2}\right.) &= \frac{1}{2}\sum_{f,t_s}\left(\left|\mathbf{Y}_{f,t_s}\right|^2 - \left|\hat{\mathbf{Y}}_{f,t_s}\right|^2\right)^2 \end{aligned} \tag{2.9}$$

where $\left|\hat{\mathbf{Y}}\right|^{.2} = \mathbf{DH}$. In above, C_{KL} is equivalent to assuming a Poisson noise model for the data and C_{LS} is equivalent to the maximum likelihood estimation of $\mathbf{D}$ and $\mathbf{H}$ in additive independent and identically distributed (i.i.d.) Gaussian noise. The widely used estimation algorithms of Lee and Seung [74] minimize the chosen cost function by initializing the entries of $\mathbf{D}$ 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:

$$\mathbf{D} \leftarrow \mathbf{D}\bullet\frac{\left(\left|\mathbf{Y}\right|^{.2}./\mathbf{DH}\right)\mathbf{H}^{\mathbf{T}}}{\mathbf{1}\mathbf{H}^{\mathbf{T}}} \quad \text{and} \quad \mathbf{H} \leftarrow \mathbf{H}\bullet\frac{\left(\left|\mathbf{Y}\right|^{.2}./\mathbf{DH}\right)\mathbf{D}^{\mathbf{T}}}{\mathbf{1}\mathbf{D}^{\mathbf{T}}} \tag{2.10}$$

where '$\bullet$' and '$./$' denote the element-wise multiplication and division, respectively. '$\mathbf{1}$' is an all-one F by T_s matrix, and '$\frac{\mathbf{A}_{NMF}}{\mathbf{B}_{NMF}}$' denotes the element-wise division of matrices $\mathbf{A}_{NMF}$ and $\mathbf{B}_{NMF}$. The update rule for LS distance is given by:

$$\mathbf{D} \leftarrow \mathbf{D} \bullet \frac{|\mathbf{Y}|^{.2}\,\mathbf{H}^{\mathbf{T}}}{\mathbf{D}\mathbf{H}\mathbf{H}^{\mathbf{T}}} \quad \text{and} \quad \mathbf{H} \leftarrow \mathbf{H} \bullet \frac{\mathbf{D}^{\mathbf{T}}|\mathbf{Y}|^{.2}}{\mathbf{D}^{\mathbf{T}}\mathbf{D}\mathbf{H}} \tag{2.11}$$

A sparseness constraint can be added to the above cost functions and this termed as sparse NMF (SNMF) where the penalty term is given as:

$$P^{SNMF} = \lambda \sum_{i,t_s} f\left(\mathbf{H}_{i,t_s}\right) \tag{2.12}$$

where λ is a parameter that controls the trade-off between sparseness and reconstruction error and $f(\bullet)$ is a function that measures sparseness. A typical choice [78] is $f\left(\mathbf{H}_{i,t_s}\right) = \left|\mathbf{H}_{i,t_s}\right|$, which is also known as an L_1 norm regularization. This corresponds to the assumption that the elements in $\mathbf{H}$ are i.i.d. one-sided exponential.

Several other types of prior over $\mathbf{D}$ and $\mathbf{H}$ are defined e.g. in [79-82], it is assumed that the prior of $\mathbf{D}$ and $\mathbf{H}$ satisfy the exponential density and the prior for the noise variance is chosen as an inverse gamma density. In [83], Gaussian distributions are chosen for both $\mathbf{D}$ and $\mathbf{H}$. The model parameters and hyperparameters are adapted by using the Markov chain Monte Carlo (MCMC) [81-83]. In all cases, a fully Bayesian treatment is applied to approximate inference for both model parameters and hyperparameters. While these approaches increase the accuracy of matrix factorization, it only works when large sample dataset is available. Moreover, it consumes significantly high computational complexity at each iteration to adapt the parameters and its hyperparameters. Other cost functions for audio spectrograms factorization have also been introduced such as that of Abdallah and Plumbley [84] which assumes multiplicative gamma-distributed noise in power spectrograms, while Parry and Issa [85] attempt to incorporate phase into the factorization by using a probabilistic phase model. Families of parameterized cost

functions, such as the Beta divergence [86] and Csiszar's divergences [87] have been presented for the separation of audio signals. After factorization, the recovered i^{th} source TF representation can be estimated as: $\left|\tilde{\mathbf{X}}\right|_i^{.2} = \mathbf{D}_i \mathbf{H}_i$, where $\mathbf{D}_i$ represents the spectral basis of i^{th} source in TF representation and $\mathbf{H}_i$ represents the code for each spectral basis element. Regardless of the cost function being used, in order to achieve audio source separation, some methods are required to group the basis functions by source or instrument. Different grouping methods have been proposed in [37], but in practice, if the sources overlap in the TF domain, it is difficult to obtain the correct clustering. This issue is discussed in [88]. In addition, most of the above techniques developed so far work only for music separation and thus, they have some important limitations that explicitly use training knowledge about the sources. As a consequence, those methods could deal only with a very specific set of signals and situations.

2.2.3 Independent subspace analysis based SCBSS

An alternative approach to SCBSS is based on independent subspace analysis techniques [89, 90]. The main idea of subspace analysis methods is to decompose the time-frequency space of the mixed signal as the sum of independent source subspaces. Given the power spectrogram of mixture TF representation $\left|\mathbf{Y}\right|^{.2}$, Each time frame of the mixture power spectrogram is expressed as a weighted sum of N_{ISA} independent basis vectors, $\mathbf{z}_\rho^{ISA}$:

$$\underline{\mathbf{y}}_{t_s} = \sum_{\rho=1}^{N_{ISA}} w_{\rho,t_s}^{ISA} \mathbf{z}_\rho^{ISA} \tag{2.13}$$

where each basis vector is weighted by a time-varying scalar w_{ρ,t_s}^{ISA}. Thus each source is

spanned by a subset of such basis vectors that define a subspace which is a matrix with basis vectors in columns $\mathbf{Z}_i^{ISA} = \left[\mathbf{z}_{1,i}^{ISA}, \ldots, \mathbf{z}_{N_{ISA}^i,i}^{ISA} \right]$. In ISA methods, the weight coefficients are obtained by projection of the input $\underline{\mathbf{y}}_{t_s}$ onto each basis component in the subspace. Assume orthonormal components, namely:

$$\mathbf{w}_i^{ISA^{\mathbf{T}}} = \mathbf{Z}_i^{ISA^{\mathbf{T}}} \underline{\mathbf{y}}_{t_s} \tag{2.14}$$

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

$$\underline{\mathbf{y}}_{t_s} = \sum_{i=1}^{I} \mathbf{Z}_i^{ISA} \mathbf{w}_i^{ISA^{\mathbf{T}}} \tag{2.15}$$

To extend the method to all time frames of power spectrogram, the source spectrogram can be estimated as $\left|\tilde{\mathbf{X}}\right|_i^{.2} = \mathbf{Z}_i^{ISA} \mathbf{W}_i^{ISA^{\mathbf{T}}}$ where $\mathbf{W}_i^{ISA} = \left[\mathbf{w}_{i,t_s}^{ISA}, \ldots, \mathbf{w}_{i,T_s}^{ISA} \right]$. Finally, use inverse STFT to reconstruct $\left|\tilde{\mathbf{X}}\right|_i^{.2}$ back to time domain source signal. However, these techniques employ the STFT to construct the TF plane which leads to a remarkable amount of cross-spectral terms due to the harmonic phenomena and the window overlapping between successive time frames. This drawback implies that it is difficult to represent the mixture as the sum of individual source subspaces. The separation efficiency [40] is greatly affected by the cross-spectral energy introduced by STFT. Another limitation of subspace analysis based SCBSS techniques is that this process works well on extracting drums from a mixture because they tend to account for most of the variance in musical signals. However, because of the way in which the model represents the data, it is limited to stationary pitch sounds such as drums.

2.2.4 Empirical mode decomposition based SCBSS

The EMD has recently gained reputation as a method for analysing nonlinear and non-stationary time series data. By combining other data analysis tools, the EMD can be employed to separate the audio sources from a single mixture. Molla and Hirose [40] proposed a subspace decomposition based SCBSS method using EMD and Hilbert spectrum (HS). The EMD decompose the mixture into a sum of band-limited functions, also labeled as IMFs, namely:

$$y(t) = \sum_{n=1}^{N} c_n(t) + r_N^{EMD}(t) \tag{2.16}$$

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

$$\theta_n^H(f) = \frac{\left| \xi_n^H(f) \right|^2}{\varphi_y^H(f)\varphi_{n,c}^H(f)} \quad \text{For} \quad n = 1,\ldots,N \tag{2.17}$$

where $\xi^H(f)$ is the cross spectrum of $y(t)$ and $c_n(t)$, $\varphi_y^H(f) = \sum_{t_s=1}^{T_s} Y_{f,t_s}^H$ and $\varphi_{n,c}^H(f) = \sum_{t_s=1}^{T_s} C_{n,f,t_s}^H$. Thus we can arrange the spectral projection vectors as individual column of a matrix $\mathbf{W}^H = \left[\boldsymbol{\theta}_1^H, \ldots, \boldsymbol{\theta}_N^H \right]$ and then derive spectral independent bases from $\mathbf{W}^H$ by applying PCA and ICA. Once these sets of spectral independent bases are obtained, the KL divergence based *k*-means clustering is used to group the bases into (number of sources) subsets. Finally, synthesis time domain estimated sources $\tilde{x}_i(t)$.

The performance of the above EMD based SCBSS method rely too heavily on the

derived independent basis vectors which are only stationary over time. Therefore, good separation results can be obtained only if basis vectors are statistical independent over time. For some source (e.g. male and female speeches), the features can be very similar and hence, it becomes difficult to obtain the independent basis vectors by PCA or ICA.

2.3 Summary

In this chapter, a wide variety of methods for SCSS have been surveyed. In general sense, all approaches can be considered as forms of constrained optimization where sources are estimated to be consistent with the observed mixture under constraints such as mutual statistical independence. The underlying theme is that as the number of observations decreases and the similarity of the underlying sources increases, the more challenge the separation methods must be. This is a big challenge for the extreme case in separating monaural mixtures of multiple sources. The *supervised* SCSS methods are more accuracy and reliable since they rely on access to source-specific training data to learn tight characteristic in the form of source models. In addition, the *supervised* SCSS methods can be used for all types of mixture if the prior knowledge or training data of the source models are provided. However, in most real applications, only observation signal is available in such case the *supervised* SCSS methods cannot separate it efficiently because of the lack of the prior knowledge of source models. On the contrary, the *unsupervised* SCSS methods can solve this limitation for the specific types of the mixture. In addition, most *unsupervised* SCSS methods are less computation intensive than *supervised* SCSS methods. Thus, for most real applications, this draws to big research interests on the *unsupervised*

SCSS methods. The main research focus can be summarised as the following issues: *i).* How to develop the *unsupervised* SCSS methods for all types of the mixture? *ii).* How to achieve the high accuracy separation performance? *iii).* How to learn source model and automatically detect the number of sources when only mixture signal is available? In this thesis, three novel *unsupervised* SCSS methods have been developed and the design of each method will be described in the next three chapters.

CHAPTER 3

SINGLE CHANNEL BLIND SOURCE SEPARATION USING VARIABLE REGULARISED SPARSE FEATURES

In this chapter, a novel variable regularised two-dimensional sparse nonnegative matrix factorization (v-SNMF2D) is proposed. The proposed model allows overcomplete representation by allowing many spectral and temporal shifts which are not inherent in the NMF and SNMF models. Thus, imposing sparseness is necessary to give unique and realistic representations of the non-stationary audio signals. Unlike the conventional SNMF2D, the proposed model imposes sparseness on temporal code $\mathbf{H}$ element-wise so that *each individual code* has its own distribution. Therefore, the sparsity parameter can be individually optimized for each code. This overcomes the problem of under- and over-sparse factorization. In addition, each sparsity parameter in the proposed model is learned and adapted as part of the matrix factorization. This bypasses the need of manual selection of these parameters as in the case of SNMF2D. The proposed method is tested on the application of SCBSS and the experimental results show that the proposed method can give superior separation performance.

The chapter is organized as follows: Section 3.1 introduces the background of NMF2D and SNMF2D. In Section 3.2, the new v-SNMF2D model is derived. Experimental results coupled with a series of comparison with other NMF techniques are presented in Section

3.3. Finally, Section 3.4 concludes the chapter.

3.1 Background

3.1.1 Two-dimensional nonnegative matrix factorization

The recently developed two-dimensional NMF factorization (NMF2D) model [91] extends the NMF model to be a two-dimensional convolution of $\mathbf{D}$ and $\mathbf{H}$. The factorization is based on a model that represents temporal structure and pitch change which occur when an instrument plays different notes. In audio source separation, the model represents each instrument compactly by a single time-frequency profile convolved in both time and frequency by a time-pitch weight matrix. This model dramatically decreases the number of components needed to model various instruments and effectively solves the SCBSS problem. The two basic cost functions are given in the following:

(Least square) $$C_{LS}^{NMF2D} = \frac{1}{2}\sum_{f,t_s}\left(|\mathbf{Y}|_{f,t_s}^2 - \mathbf{Z}_{f,t_s}\right)^2 \tag{3.1}$$

(Kullback-Leibler) $$C_{KLd}^{NMF2D} = \frac{1}{2}\sum_{f,t_s}|\mathbf{Y}|_{f,t_s}^2 \log\frac{|\mathbf{Y}|_{f,t_s}^2}{\mathbf{Z}_{f,t_s}} - |\mathbf{Y}|_{f,t_s}^2 + \mathbf{Z}_{f,t_s} \tag{3.2}$$

for $\forall f \in F, \forall t_s \in T_s$ where $\mathbf{Z} = \sum_{\tau,\phi} \overset{\downarrow\phi}{\mathbf{D}}^{\tau} \overset{\rightarrow\tau}{\mathbf{H}}^{\phi}$. Here the vertical arrow in $\overset{\downarrow\phi}{\mathbf{D}}^{\tau}$ denotes downward shift which moves each element in the matrix $\mathbf{D}^{\tau}$ down by ϕ rows, and the horizontal arrow in $\overset{\rightarrow\tau}{\mathbf{H}}^{\phi}$ denotes right shift which moves each element in the matrix $\mathbf{H}^{\phi}$ to the right by τ columns. This can be interpreted as the follows, i.e.:

$$\mathbf{A}^{E}=\begin{bmatrix}1&1&1&1\\2&2&2&2\\3&3&3&3\\4&4&4&4\end{bmatrix}\quad \mathbf{A}^{\downarrow 1}_{E}=\begin{bmatrix}0&0&0&0\\1&1&1&1\\2&2&2&2\\3&3&3&3\end{bmatrix}\quad \mathbf{A}^{\rightarrow 1}_{E}=\begin{bmatrix}0&1&1&1\\0&2&2&2\\0&3&3&3\\0&4&4&4\end{bmatrix}$$

In above, $\mathbf{D}^{\tau}=\left\{\mathbf{D}^{\tau}_{f,i}\mid f=1,...,F \text{ and } i=1,...,I\right\}$ denotes the τ^{th} slice of $\mathbf{D}$ and $\mathbf{H}^{\phi}=\left\{\mathbf{H}^{\phi}_{i,t_s}\mid i=1,...,I \text{ and } t_s=1,...,T_s\right\}$ denotes the ϕ^{th} slice of $\mathbf{H}$ which can be derived using the cost functions (3.1) or (3.2).

3.1.2 Two-dimensional sparse nonnegative matrix factorization

The use of sparse representation is strongly related to the principle of parsimony, i.e., among all possible accounts the simplest is considered the best. If no formal prior information is given, parsimony can be considered a reasonable guiding principle to avoid overfitting. Thus, the NMF2D model can be extended to SNMF2D [92] model whereas the two basic cost functions (3.1) and (3.2) with sparse penalty term on $\mathbf{H}$ are given in the following:

(Least square) $$C_{LS}^{SNMF2D}=\frac{1}{2}\sum_{f,t_s}\left(\left|\mathbf{Y}\right|^2_{f,t_s}-\tilde{\mathbf{Z}}_{f,t_s}\right)^2+\lambda f\left(\mathbf{H}\right) \tag{3.3}$$

(Kullback-Leibler) $$C_{KLd}^{SNMF2D}=\frac{1}{2}\sum_{f,t_s}\left|\mathbf{Y}\right|^2_{f,t_s}\log\frac{\left|\mathbf{Y}\right|^2_{f,t_s}}{\tilde{\mathbf{Z}}_{f,t_s}}-\left|\mathbf{Y}\right|^2_{f,t_s}+\tilde{\mathbf{Z}}_{f,t_s}+\lambda f\left(\mathbf{H}\right) \tag{3.4}$$

for $\forall f\in F,\forall t_s\in T_s$ where $\tilde{\mathbf{Z}}=\sum_{\tau,\phi}\tilde{\mathbf{D}}^{\tau}\overset{\downarrow\phi}{\mathbf{H}}\overset{\rightarrow\tau}{\mathbf{H}}^{\phi}$, $\tilde{\mathbf{D}}^{\tau}_{f,i}=\mathbf{D}^{\tau}_{f,i}\Big/\sqrt{\sum_{\tau,f}(\mathbf{D}^{\tau}_{f,i})^2}$ and $f(\mathbf{H})$ can be any function with positive derivative such as $L_{\alpha}-norm(\alpha>0)$ given by $f(\mathbf{H})=\left\|\mathbf{H}\right\|_{\alpha}=\left(\sum_{\phi,i,t_s}\left|\mathbf{H}^{\phi}_{i,t_s}\right|^{\alpha}\right)^{1/\alpha}$. The SNMF2D is more effective than NMF2D in some situations that the structure of a factor in $\mathbf{H}^{\phi}$ can be input into the signature of the same

factor in $\mathbf{D}^{\tau}$ and vice versa. Hence, this leads to ambiguity that can be only resolved by forcing the structure on $\mathbf{D}^{\tau}$ through imposing sparseness on $\mathbf{H}^{\phi}$. However, the drawbacks of SNMF2D originate from its lack of a generalized criterion for controlling the sparsity of $\mathbf{H}$. In practice, the sparsity parameter λ is set manually. When SNMF2D imposes uniform sparsity on all temporal codes, this is equivalent to enforcing each temporal code to be identical to a fixed distribution according to the selected sparsity parameter. In addition, by assigning the fixed distribution onto each individual code, this is equivalent to constraining all codes to be stationary. However, audio signals are non-stationary in the TF domain and have different temporal structure and sparsity. Hence, they cannot be realistically enforced by a fixed distribution. These characteristics are even more pronounced between different types of audio signals. In addition, since the SNMF2D introduces many temporal shifts, this will result in more temporal codes to deviate from the fixed distribution. Therefore, within the context of SCBSS, when SNMF2D imposes uniform sparsity on all the temporal codes, this will inevitably result in under- or over-sparse factorization which will subsequently lead to ambiguity in separating audio mixtures. Thus, the above suggests that the present form of SNMF2D is still technically lacking and is not readily suited for SCBSS especially mixtures involving more types of audio signals.

3.2 Proposed Method

In this section, a new factorization method is derived and it is named as the variable regularised two-dimensional sparse nonnegative matrix factorization. The model is given

as follows, where $\mathbf{D}_i^{\tau}$ is the i^{th} column of $\mathbf{D}^{\tau}$, $\mathbf{H}_i^{\phi}$ is the i^{th} row of $\mathbf{H}^{\phi}$

$$\left|\mathbf{Y}\right|^{.2} = \sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}} \mathbf{D}^{\tau} \overset{\downarrow\phi}{\mathbf{H}^{\phi}} \overset{\rightarrow\tau}{} + \mathbf{V}^{No} = \sum_{i=1}^{I}\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}} \mathbf{D}_i^{\tau} \overset{\downarrow\phi}{\mathbf{H}_i^{\phi}} \overset{\rightarrow\tau}{} + \mathbf{V}^{No} \tag{3.5}$$

$$\text{where } \mathbf{H}^{\phi} \sim p\left(\mathbf{H}^{\phi} \mid \mathbf{\Lambda}^{\phi}\right) = \prod_{i=1}^{I}\prod_{t_s=1}^{T_s} \lambda_{i,t_s}^{\phi} \exp\left(-\lambda_{i,t_s}^{\phi} \mathbf{H}_{i,t_s}^{\phi}\right)$$

In this model (3.5), it is worth pointing out that *each individual element* in $\mathbf{H}^{\phi}$ is constrained to a exponential distribution with independent decay parameter λ_{i,t_s}^{ϕ}. we first define $\mathbf{D} = \left[\mathbf{D}^0 \ \mathbf{D}^1 \ \cdots \ \mathbf{D}^{\tau_{\max}}\right]$, $\mathbf{H} = \left[\mathbf{H}^0 \ \mathbf{H}^1 \ \cdots \ \mathbf{H}^{\phi_{\max}}\right]$ and $\mathbf{\Lambda} = \left[\mathbf{\Lambda}^0 \ \mathbf{\Lambda}^1 \ \cdots \ \mathbf{\Lambda}^{\phi_{\max}}\right]$. $\mathbf{\Lambda}^{\phi} = \left\{\lambda_{i,t_s}^{\phi} \mid i = 1,...,I \text{ and } t_s = 1,...,T_s\right\}$ denotes the ϕ^{th} slice of sparse parameter $\mathbf{\Lambda}$ and $\mathbf{V}^{No}$ is assumed to be independently and identically distributed (i.i.d.) as Gaussian distribution with noise having variance σ^2. The terms $\tau_{\max}$ and $\phi_{\max}$ are the maximum number of τ shifts and ϕ shifts, respectively. This is in contrast with the conventional SNMF2D where λ_{i,t_s}^{ϕ} is simply set to a fixed constant i.e. $\lambda_{i,t_s}^{\phi} = \lambda$ for all i, t_s, ϕ. Such setting imposes uniform constant sparsity on all temporal codes $\mathbf{H}^{\phi}$ which enforces each temporal code to be identical to a fixed distribution according to the selected constant sparsity parameter. The 3D-representation for $\mathbf{D}$, $\mathbf{H}$ and $\mathbf{\Lambda}$ are presented in Figure 3.1.

The consequence of this uniform constant sparsity has already been discussed in Section 3.1. In Section 3.3, the details of the sparsity analysis for source separation and evaluate its performance against with other existing methods will be presented.

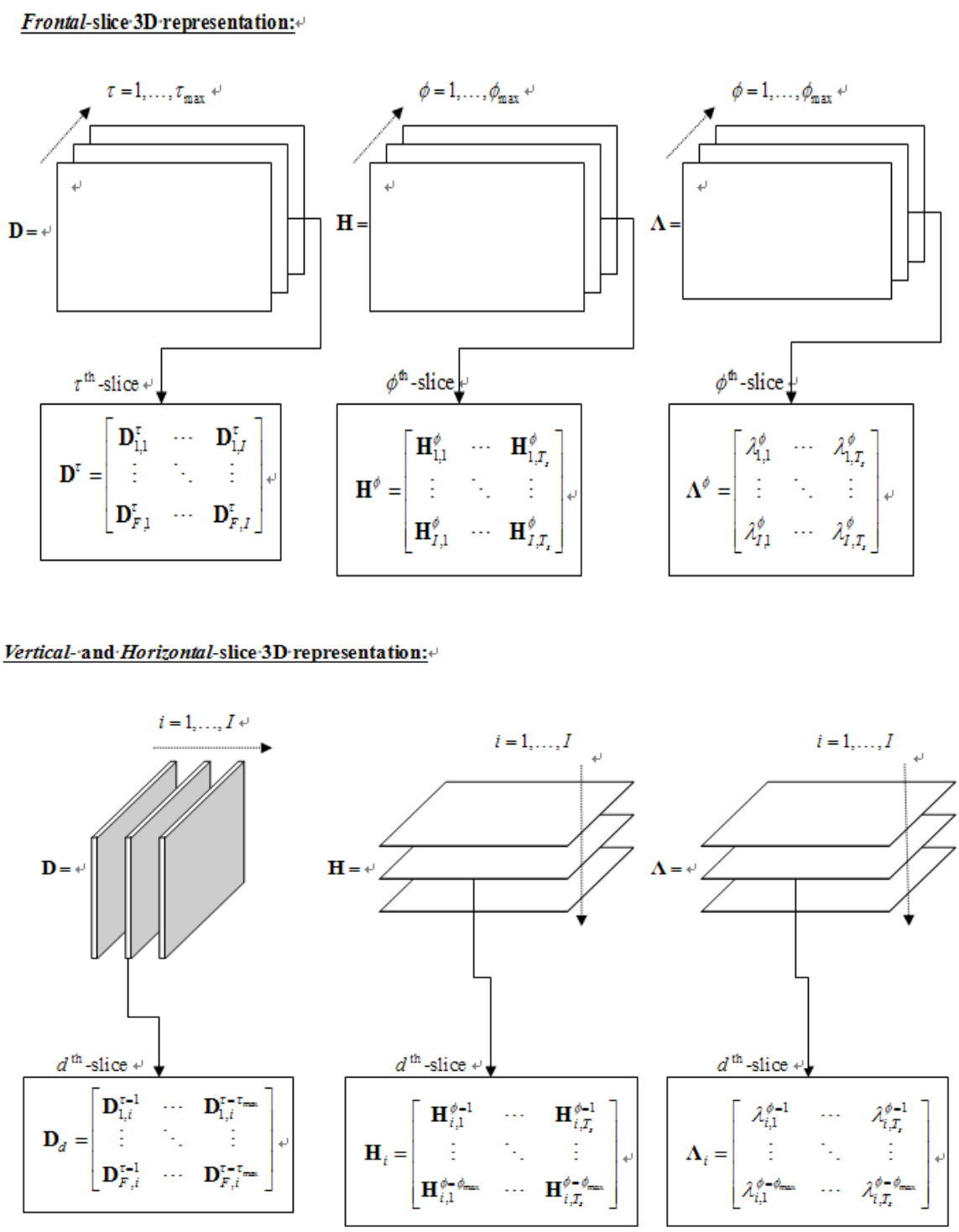


Figure 3.1: 3D-representation for $\mathbf{D}$, $\mathbf{H}$ and $\mathbf{\Lambda}$.

In above, $\mathbf{D}$ has been sliced in two directions, namely, the frontal (i.e. τ^{th}-slice) and vertical (i.e. i^{th}-slice). It is valid that $\mathbf{D}$ can also have horizontal slice representation but this has not been shown as it is not needed in the development of the proposed algorithm.

Similarly, $\mathbf{H}$ and $\mathbf{\Lambda}$ have been sliced in two directions, frontal (i.e. ϕ^{th}-slice) and horizontal (i.e. i^{th}-slice). However, the vertical slice representation has not been shown as it is not required in the development of the algorithm.

3.2.1 Formulation of the v-SNMF2D

To facilitate such spectral basis with variable sparse code, we choose a prior distribution $p(\mathbf{D},\mathbf{H})$ over the factors $\{\mathbf{D},\mathbf{H}\}$ in the analysis equation. The posterior can be found by using Bayes' theorem as follows:

$$p(\mathbf{D},\mathbf{H}\,|\,|\mathbf{Y}|^{.2}) = \frac{p(|\mathbf{Y}|^{.2}\,|\,\mathbf{D},\mathbf{H})\,p(\mathbf{D},\mathbf{H})}{P(|\mathbf{Y}|^{.2})} \tag{3.6}$$

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

$$\log p(\mathbf{D},\mathbf{H}\,|\,|\mathbf{Y}|^{.2}) = \log p(|\mathbf{Y}|^{.2}\,|\,\mathbf{D},\mathbf{H}) + \log p(\mathbf{D},\mathbf{H}) + \text{const} \tag{3.7}$$

where 'const' denotes constant. The noise is assumed to be independently and identically distributed with Gaussian distribution having variance σ^2. Thus, the likelihood of the observations given $\mathbf{D}$ and $\mathbf{H}$ can be written [1] as $p(|\mathbf{Y}|^{.2}\,|\,\mathbf{D},\mathbf{H}) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\left\||\mathbf{Y}|^{.2} - \sum_i\sum_\tau\sum_\phi \mathbf{D}_i^{\tau\downarrow\phi}\mathbf{H}_i^{\phi\rightarrow\tau}\right\|_{Fro}^2 \Big/ 2\sigma^2\right]$ where $\|\cdot\|_{Fro}$ denotes the Frobenius norm. The second term consists of the prior distribution of $\mathbf{H}$ and $\mathbf{D}$ where they are jointly independent. Each element of $\mathbf{H}$ is constrained to be exponential distributed with independent decay parameters, namely, $p(\mathbf{H}\,|\,\mathbf{\Lambda}) = \prod_{i,\phi,t_s}\lambda_{i,t_s}^{\phi}\exp\left(-\lambda_{i,t_s}^{\phi}\mathbf{H}_{i,t_s}^{\phi}\right)$ so that $f(\mathbf{H}) = \sum_{i,\phi,t_s}\lambda_{i,t_s}^{\phi}\mathbf{H}_{i,t_s}^{\phi}$. The prior over $\mathbf{D}$ is flat with each column constrained to have

[1] To avoid cluttering the notation, we shall remove the upper limits from the summation terms. The upper limits can be inferred from (3.3).

unit length. Hence, the negative posterior serves as the Least Square (LS) cost function which is defined as:

$$\begin{aligned} C_{LS}^{vSNMF\,2D} &\propto \frac{1}{2\sigma^2}\left\| |\mathbf{Y}|_{f,t_s}^2 - \sum_i \sum_\tau \sum_\phi \overset{\downarrow\phi}{\mathbf{D}}_i^\tau \overset{\rightarrow\tau}{\mathbf{H}}_i^\phi \right\|_{Fro}^2 + f(\mathbf{H}) \\ &= \frac{1}{2\sigma^2}\left\| |\mathbf{Y}|_{f,t_s}^2 - \sum_i \sum_\tau \sum_\phi \overset{\downarrow\phi}{\mathbf{D}}_i^\tau \overset{\rightarrow\tau}{\mathbf{H}}_i^\phi \right\|_{Fro}^2 + \sum_{\phi,i,t_s} \lambda_{i,t_s}^\phi \mathbf{H}_{i,t_s}^\phi \end{aligned} \tag{3.8}$$

The sparsity term $f(\mathbf{H})$ forms the L_1-norm regularization to resolve the ambiguity by forcing all structure in $\mathbf{H}$ onto $\mathbf{D}$. Therefore, the sparseness of the solution in (3.8) is highly dependent on the regularization parameters λ_{i,t_s}^ϕ.

3.2.1.1 Estimation of the spectral basis and temporal code

In the matrix factorization, each spectral basis is constrained to be of unit length. Hence, this can be represented by $\tilde{\mathbf{Z}} = \sum_i \sum_\tau \sum_\phi \overset{\downarrow\phi}{\tilde{\mathbf{D}}}_i^\tau \overset{\rightarrow\tau}{\mathbf{H}}_i^\phi$ where $\tilde{\mathbf{D}}_{f,i}^\tau = \mathbf{D}_{f,i}^\tau \Big/ \sqrt{\sum_{\tau,f} (\mathbf{D}_{f,i}^\tau)^2}$ is factor-wise normalized to $\mathbf{D}^\tau$. The derivatives of (3.8) corresponding to $\mathbf{D}^\tau$ and $\mathbf{H}^\phi$ of v-SNMF2D are given by:

$$\begin{aligned} \frac{\partial C_{LS}^{vSNMF\,2D}}{\partial \mathbf{D}_{f',i'}^{\tau'}} &= \frac{\partial}{\partial \mathbf{D}_{f',i'}^{\tau'}} \left(\frac{1}{2\sigma^2} \sum_{f,t_s} \left(|\mathbf{Y}|_{f,t_s}^2 - \tilde{\mathbf{Z}}_{f,t_s} \right)^2 + f(\mathbf{H}) \right) \\ &= -\frac{1}{\sigma^2} \sum_{\phi,t_s} \left(|\mathbf{Y}|_{f'+\phi,t_s}^2 - \tilde{\mathbf{Z}}_{f'+\phi,t_s} \right) \mathbf{H}_{i',t_s-\tau'}^\phi \end{aligned} \tag{3.9}$$

$$\begin{aligned} \frac{\partial C_{LS}^{vSNMF\,2D}}{\partial \mathbf{H}_{i',t_s'}^{\phi'}} &= \frac{\partial}{\partial \mathbf{H}_{i',t_s'}^{\phi'}} \left(\frac{1}{2\sigma^2} \sum_{f,t_s} \left(|\mathbf{Y}|_{f,t_s}^2 - \tilde{\mathbf{Z}}_{f,t_s} \right)^2 + f(\mathbf{H}) \right) \\ &= -\frac{1}{\sigma^2} \sum_{\tau,f} \tilde{\mathbf{D}}_{f-\phi',i'}^\tau \left(|\mathbf{Y}|_{f,t_s'+\tau}^2 - \tilde{\mathbf{Z}}_{f,t_s'+\tau} \right) + \frac{\partial f(\mathbf{H})}{\partial \mathbf{H}_{i',t_s'}^{\phi'}} \end{aligned} \tag{3.10}$$

Thus, by applying the standard gradient decent approach:

$$\mathbf{D}^{\tau'}_{f',i'} \leftarrow \tilde{\mathbf{D}}^{\tau'}_{f',i'} - \eta_D \frac{\partial C^{vSNMF\,2D}_{LS})}{\partial \mathbf{D}^{\tau'}_{f',i'}}$$
$$\mathbf{H}^{\phi'}_{i',t'_s} \leftarrow \mathbf{H}^{\phi'}_{i',t'_s} - \eta_H \frac{\partial C^{vSNMF\,2D}_{LS}}{\partial \mathbf{H}^{\phi'}_{i',t'_s}} \tag{3.11}$$

where η_D and η_H are positive learning rates which can be obtained by following the approach of Lee and Seung [74], namely:

$$\eta_D = \frac{\sigma^2 \tilde{\mathbf{D}}^{\tau'}_{f',i'}}{\sum_{\phi,t_s} \tilde{\mathbf{Z}}_{f'+\phi,t_s} \mathbf{H}^{\phi}_{i',t_s-\tau'}} \quad \text{and} \quad \eta_H = \frac{\sigma^2 \mathbf{H}^{\phi'}_{i',t'_s}}{\sum_{\tau,f} \tilde{\mathbf{D}}^{\tau}_{f-\phi',i'} \tilde{\mathbf{Z}}_{f,t'_s+\tau} + \frac{\partial f(\mathbf{H})}{\partial \mathbf{H}^{\phi'}_{i',t'_s}}} \tag{3.12}$$

Thus, the multiplicative learning rules become:

$$\mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi} \bullet \frac{\sum_{\tau} \overset{\downarrow\phi^{\mathbf{T}}}{\tilde{\mathbf{D}}^{\tau}} \overset{\leftarrow\tau}{|\mathbf{Y}|^{.2}}}{\sum_{\tau} \overset{\downarrow\phi^{\mathbf{T}}}{\tilde{\mathbf{D}}^{\tau}} \overset{\leftarrow\tau}{\tilde{\mathbf{Z}}} + \frac{\partial f(\mathbf{H})}{\partial \mathbf{H}^{\phi}}} \quad \text{where} \quad f(\mathbf{H}) = \sum_{i,\phi,t_s} \lambda^{\phi}_{i,t_s} \mathbf{H}^{\phi}_{i,t_s} \tag{3.13}$$

$$\mathbf{D}^{\tau} \leftarrow \tilde{\mathbf{D}}^{\tau} \bullet \frac{\sum_{\phi} \overset{\uparrow\phi}{|\mathbf{Y}|^{.2}} \overset{\rightarrow\tau^{\mathbf{T}}}{\mathbf{H}^{\phi}}}{\sum_{\phi} \overset{\uparrow\phi}{\tilde{\mathbf{Z}}} \overset{\rightarrow\tau^{\mathbf{T}}}{\mathbf{H}^{\phi}}} \quad \text{where} \quad \tilde{\mathbf{D}}^{\tau}_{f,i} = \frac{\mathbf{D}^{\tau}_{f,i}}{\sqrt{\sum_{\tau,f} (\mathbf{D}^{\tau}_{f,i})^2}} \tag{3.14}$$

In (3.13) and (3.14), '$\bullet$' is the element wise product, the column vectors of $\mathbf{D}^{\tau}$ will be factor-wise normalized to unit length.

3.2.1.2 Estimation of the variable regularization parameter

Since $\overset{\rightarrow\tau}{\mathbf{H}^{\phi}}$ are obtained directly from the original sparse code matrix $\overset{\rightarrow 0}{\mathbf{H}^{\phi}}$, it suffices to compute the regularization parameters associated only with $\overset{\rightarrow 0}{\mathbf{H}^{\phi}}$. Therefore, the cost function in (3.8) with $\tau_{\max} = 0$ can be set:

$$\begin{aligned} F(\mathbf{H}) &= \frac{1}{2\sigma^2} \left\| |\mathbf{Y}|^{.2} - \sum_{\phi=0}^{\phi_{\max}} \overset{\downarrow\phi}{\mathbf{D}} \mathbf{H}^{\phi} \right\|^2_{Fro} + \sum_{\phi=0}^{\phi_{\max}} \sum_{i=1}^{N_s} \sum_{t_s=1}^{T_s} \lambda^{\phi}_{i,t_s} \mathbf{H}^{\phi}_{i,t_s} \\ &= \frac{1}{2\sigma^2} \left\| Vec(|\mathbf{Y}|^{.2}) - \sum_{\phi=0}^{\phi_{\max}} (\mathbf{I} \otimes \overset{\downarrow\phi}{\mathbf{D}}) Vec(\mathbf{H}^{\phi}) \right\|^2_{Fro} + \sum_{\phi=0}^{\phi_{\max}} Vec(\mathbf{\Lambda}^{\phi})^{\mathbf{T}} Vec(\mathbf{H}^{\phi}) \end{aligned} \tag{3.15}$$

with $Vec(\cdot)$ representing the column vectorization, '$\otimes$' is the Kronecker product, and $\mathbf{I}$

is the identity matrix. Defining the following terms:

$$\underline{\mathbf{y}} = Vec(|\mathbf{Y}|^{.2}), \quad \overline{\mathbf{D}} = \left[\mathbf{I} \otimes \overset{\downarrow 0}{\mathbf{D}} \vdots \mathbf{I} \otimes \overset{\downarrow 1}{\mathbf{D}} \vdots \cdots \vdots \mathbf{I} \otimes \overset{\downarrow \phi_{\max}}{\mathbf{D}} \vdots \right],$$

$$\underline{\mathbf{h}} = \begin{bmatrix} Vec(\mathbf{H}^0) \\ \cdots\cdots \\ Vec(\mathbf{H}^1) \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ Vec(\mathbf{H}^{\phi_{\max}}) \end{bmatrix}, \quad \underline{\boldsymbol{\lambda}} = \begin{bmatrix} \underline{\boldsymbol{\lambda}}^0 \\ \cdots\cdots \\ \underline{\boldsymbol{\lambda}}^1 \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ \underline{\boldsymbol{\lambda}}^{\phi_{\max}} \end{bmatrix}, \quad \underline{\boldsymbol{\lambda}}^{\phi} = \begin{bmatrix} \lambda_{1,1}^{\phi} \\ \lambda_{2,1}^{\phi} \\ \vdots \\ \lambda_{I,T_s}^{\phi} \end{bmatrix} \tag{3.16}$$

Thus, (3.15) can be rewritten in terms of $\underline{\mathbf{h}}$ as:

$$F(\underline{\mathbf{h}}) = \frac{1}{2\sigma^2} \left\| \underline{\mathbf{y}} - \overline{\mathbf{D}} \underline{\mathbf{h}} \right\|^2 + \underline{\boldsymbol{\lambda}}^{\mathrm{T}} \underline{\mathbf{h}} \tag{3.17}$$

Note that $\underline{\mathbf{h}}$ and $\underline{\boldsymbol{\lambda}}$ are vectors of dimension $R \times 1$ where $R = I \times T_s \times (\phi_{\max} + 1)$. To determine $\underline{\boldsymbol{\lambda}}$, the Expectation-Maximization (EM) algorithm can be used and treat $\underline{\mathbf{h}}$ as the hidden variable where the log-likelihood function can be optimized with respect to $\underline{\boldsymbol{\lambda}}$. To reiterate our aim, we are not developing a full Bayesian inference on the generative model in (3.8). Rather, the proposed Bayesian inference is only focused on the approximation of the posterior distribution of $\mathbf{H}$. Using the Jensen's inequality, it can be shown that for any distribution $Q(\underline{\mathbf{h}})$, the log-likelihood function satisfies the following:

$$\ln p(\underline{\mathbf{y}} \mid \underline{\boldsymbol{\lambda}}, \overline{\mathbf{D}}, \sigma^2) \geq \int Q(\underline{\mathbf{h}}) \ln \left(\frac{p(\underline{\mathbf{y}}, \underline{\mathbf{h}} \mid \underline{\boldsymbol{\lambda}}, \overline{\mathbf{D}}, \sigma^2)}{Q(\underline{\mathbf{h}})} \right) d\underline{\mathbf{h}} \tag{3.18}$$

One can easily check that the distribution that maximizes the right hand side of (3.18) is given by $Q(\underline{\mathbf{h}}) = p(\underline{\mathbf{h}} \mid \underline{\mathbf{y}}, \underline{\boldsymbol{\lambda}}, \overline{\mathbf{D}}, \sigma^2)$ which is the posterior distribution of $\underline{\mathbf{h}}$. In this method, the posterior distribution in the form of Gibbs distribution is expressed as:

$$Q(\underline{\mathbf{h}}) = \frac{1}{Z_h} \exp\left[-F(\underline{\mathbf{h}})\right] \quad \text{where} \quad Z_h = \int \exp\left[-F(\underline{\mathbf{h}})\right] d\underline{\mathbf{h}} \tag{3.19}$$

The functional form of the Gibbs distribution in (3.19) is expressed in terms of $F(\underline{\mathbf{h}})$ and

this is crucial as it will enable us to simplify the variational optimization [93, 94] of $\underline{\boldsymbol{\lambda}}$. The maximum likelihood estimation of $\underline{\boldsymbol{\lambda}}$ can be expressed by:

$$\begin{aligned}\underline{\boldsymbol{\lambda}}^{ML} &= \arg\max_{\underline{\boldsymbol{\lambda}}} \ln p\left(\underline{\mathbf{y}} \mid \underline{\boldsymbol{\lambda}}, \overline{\mathbf{D}}, \sigma^2\right) \\ &= \arg\max_{\underline{\boldsymbol{\lambda}}} \int Q\left(\underline{\mathbf{h}}\right) \ln p\left(\underline{\mathbf{y}}, \underline{\mathbf{h}} \mid \underline{\boldsymbol{\lambda}}, \overline{\mathbf{D}}, \sigma^2\right) d\underline{\mathbf{h}} - \int Q\left(\underline{\mathbf{h}}\right) \ln Q\left(\underline{\mathbf{h}}\right) d\underline{\mathbf{h}} \\ &= \arg\max_{\underline{\boldsymbol{\lambda}}} \int Q\left(\underline{\mathbf{h}}\right) \ln p\left(\underline{\mathbf{y}}, \underline{\mathbf{h}} \mid \underline{\boldsymbol{\lambda}}, \overline{\mathbf{D}}, \sigma^2\right) d\underline{\mathbf{h}} \\ &= \arg\max_{\underline{\boldsymbol{\lambda}}} \int Q\left(\underline{\mathbf{h}}\right) \left(\ln p\left(\underline{\mathbf{y}} \mid \underline{\mathbf{h}}, \sigma^2, \overline{\mathbf{D}}\right) + \ln p\left(\underline{\mathbf{h}} \mid \underline{\boldsymbol{\lambda}}\right)\right) d\underline{\mathbf{h}} \\ &= \arg\max_{\underline{\boldsymbol{\lambda}}} \int Q\left(\underline{\mathbf{h}}\right) \ln p\left(\underline{\mathbf{h}} \mid \underline{\boldsymbol{\lambda}}\right) d\underline{\mathbf{h}}\end{aligned} \tag{3.20}$$

Similarly:

$$\begin{aligned}\sigma^{2(ML)} &= \arg\max_{\sigma^2} \int Q(\underline{\mathbf{h}}) \left(\ln p\left(\underline{\mathbf{y}} \mid \underline{\mathbf{h}}, \sigma^2, \overline{\mathbf{D}}\right) + \ln p\left(\underline{\mathbf{h}} \mid \underline{\boldsymbol{\lambda}}\right)\right) d\underline{\mathbf{h}} \\ &= \arg\max_{\sigma^2} \int Q(\underline{\mathbf{h}}) \ln p\left(\underline{\mathbf{y}} \mid \underline{\mathbf{h}}, \sigma^2, \overline{\mathbf{D}}\right) d\underline{\mathbf{h}}\end{aligned} \tag{3.21}$$

Since each element of $\mathbf{H}$ is constrained to be exponential distributed with independent decay parameters, this gives $p(\underline{\mathbf{h}} \mid \underline{\boldsymbol{\lambda}}) = \prod_p \lambda_p \exp(-\lambda_p h_p)$ and therefore, (3.18) becomes:

$$\underline{\boldsymbol{\lambda}}^{ML} = \arg\max_{\underline{\boldsymbol{\lambda}}} \int Q\left(\underline{\mathbf{h}}\right) \left(\ln \lambda_p - \lambda_p h_p\right) d\underline{\mathbf{h}} \tag{3.22}$$

The Gibbs distribution $Q\left(\underline{\mathbf{h}}\right)$ treats $\underline{\mathbf{h}}$ as the dependent variable while assuming all other parameters to be constant. Solve $\frac{\partial \int Q\left(\underline{\mathbf{h}}\right) \left(\ln \lambda_p - \lambda_p h_p\right) d\underline{\mathbf{h}}}{\partial \lambda_p} = 0$. As such, the functional optimization of $\underline{\boldsymbol{\lambda}}$ in (3.22) is obtained by differentiating the terms within the integral with respect to λ_p and the end result is given by:

$$\lambda_p = \frac{1}{\int h_p Q\left(\underline{\mathbf{h}}\right) d\underline{\mathbf{h}}} \qquad \text{for} \quad p = 1, 2, \ldots, R \tag{3.23}$$

where λ_p is the p^{th} element of $\underline{\boldsymbol{\lambda}}$. Since:

$$p(\underline{\mathbf{y}} \mid \underline{\mathbf{h}}, \sigma^2, \overline{\mathbf{D}}) = \frac{1}{(2\pi\sigma^2)^{N_p/2}} \exp\left(-\frac{1}{2\sigma^2} \left\| \underline{\mathbf{y}} - \overline{\mathbf{D}} \underline{\mathbf{h}} \right\|^2\right) \tag{3.24}$$

where $N_p = F \times T_s$, the iterative update rule for $\sigma^{2(ML)}$ is given by:

$$\begin{aligned}\sigma^{2(ML)} &= \arg\max_{\sigma^2} \int Q\ (\underline{\mathbf{h}}) \left(-\frac{N_p}{2} \ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \left\| \underline{\mathbf{y}} - \overline{\mathbf{D}}\underline{\mathbf{h}} \right\|^2 \right) d\underline{\mathbf{h}} \\ &= \frac{1}{N_p} \int Q\ (\underline{\mathbf{h}}) \left(\left\| \underline{\mathbf{y}} - \overline{\mathbf{D}}\underline{\mathbf{h}} \right\|^2 \right) d\underline{\mathbf{h}}\end{aligned} \tag{3.25}$$

Despite the simple form of (3.23) and (3.25), the integral is difficult to compute analytically and therefore, an approximation to $Q\ (\underline{\mathbf{h}})$ can be found. It is noted that the solution $\underline{\mathbf{h}}$ naturally partition its elements into distinct subsets $\underline{\mathbf{h}}_P$ and $\underline{\mathbf{h}}_M$ 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 $F(\underline{\mathbf{h}})$ can be expressed as following:

$$\begin{aligned}F(\underline{\mathbf{h}}) &= \frac{1}{2\sigma^2} \left\| \underline{\mathbf{y}} - \overline{\mathbf{D}}_P \underline{\mathbf{h}}_P - \overline{\mathbf{D}}_M \underline{\mathbf{h}}_M \right\|^2 + \underline{\boldsymbol{\lambda}}_P^{\mathbf{T}} \underline{\mathbf{h}}_P + \underline{\boldsymbol{\lambda}}_M^{\mathbf{T}} \underline{\mathbf{h}}_M \\ &= \frac{1}{2\sigma^2} \left[\left\| \underline{\mathbf{y}} - \overline{\mathbf{D}}_M \underline{\mathbf{h}}_M \right\|^2 - 2\left(\underline{\mathbf{y}} - \overline{\mathbf{D}}_M \underline{\mathbf{h}}_M \right)^{\mathbf{T}} \left(\overline{\mathbf{D}}_P \underline{\mathbf{h}}_P \right) + \left\| \overline{\mathbf{D}}_P \underline{\mathbf{h}}_P \right\|^2 \right] + \underline{\boldsymbol{\lambda}}_P^{\mathbf{T}} \underline{\mathbf{h}}_P + \underline{\boldsymbol{\lambda}}_M^{\mathbf{T}} \underline{\mathbf{h}}_M \\ &= \frac{1}{2\sigma^2} \left[\left\| \underline{\mathbf{y}} - \overline{\mathbf{D}}_M \underline{\mathbf{h}}_M \right\|^2 + \left\| \underline{\mathbf{y}} - \overline{\mathbf{D}}_P \underline{\mathbf{h}}_P \right\|^2 - \left\| \underline{\mathbf{y}} \right\|^2 - 2\left(\underline{\mathbf{y}} - \overline{\mathbf{D}}_M \underline{\mathbf{h}}_M \right)^{\mathbf{T}} \left(\overline{\mathbf{D}}_P \underline{\mathbf{h}}_P \right) + 2\underline{\mathbf{y}}^{\mathbf{T}} \left(\overline{\mathbf{D}}_P \underline{\mathbf{h}}_P \right) \right] + \underline{\boldsymbol{\lambda}}_P^{\mathbf{T}} \underline{\mathbf{h}}_P + \underline{\boldsymbol{\lambda}}_M^{\mathbf{T}} \underline{\mathbf{h}}_M \\ &= \underbrace{\frac{1}{2\sigma^2} \left\| \underline{\mathbf{y}} - \overline{\mathbf{D}}_M \underline{\mathbf{h}}_M \right\|^2 + \underline{\boldsymbol{\lambda}}_M^{\mathbf{T}} \underline{\mathbf{h}}_M}_{F(\underline{\mathbf{h}}_M)} + \underbrace{\frac{1}{2\sigma^2} \left\| \underline{\mathbf{y}} - \overline{\mathbf{D}}_P \underline{\mathbf{h}}_P \right\|^2 + \underline{\boldsymbol{\lambda}}_P^{\mathbf{T}} \underline{\mathbf{h}}_P}_{F(\underline{\mathbf{h}}_P)} + \underbrace{\frac{1}{2\sigma^2} \left[2\left(\overline{\mathbf{D}}_M \underline{\mathbf{h}}_M \right)^{\mathbf{T}} \left(\overline{\mathbf{D}}_P \underline{\mathbf{h}}_P \right) - \left\| \underline{\mathbf{y}} \right\|^2 \right]}_{\psi} \\ &= F(\underline{\mathbf{h}}_M) + F(\underline{\mathbf{h}}_P) + \psi\end{aligned} \tag{3.26}$$

In (3.26), the term $\left\| \underline{\mathbf{y}} \right\|^2$ in ψ is simply a constant which does not affect the optimization while $\left(\overline{\mathbf{D}}_M \underline{\mathbf{h}}_M \right)^{\mathbf{T}} \left(\overline{\mathbf{D}}_P \underline{\mathbf{h}}_P \right)$ measures the orthogonality between $\overline{\mathbf{D}}_M \underline{\mathbf{h}}_M$ and $\overline{\mathbf{D}}_P \underline{\mathbf{h}}_P$ which is assumed to be uncorrelated. Therefore, (3.26) can be simplified to $F(\underline{\mathbf{h}}) \approx F(\underline{\mathbf{h}}_M) + F(\underline{\mathbf{h}}_P)$. Hence, $Q\ (\underline{\mathbf{h}})$ can be decomposed as:

$$\begin{aligned}Q\ (\underline{\mathbf{h}}) &= \frac{1}{Z_h} \exp\left[-F(\underline{\mathbf{h}}) \right] \\ &\approx \frac{1}{Z_h} \exp\left[-\left(F(\underline{\mathbf{h}}_P) + F(\underline{\mathbf{h}}_M) \right) \right] \\ &= \frac{1}{Z_h} \exp\left[-F(\underline{\mathbf{h}}_P) \right] \exp\left[-F(\underline{\mathbf{h}}_M) \right] \\ &= \frac{1}{Z_P} \exp\left[-F(\underline{\mathbf{h}}_P) \right] \frac{1}{Z_M} \exp\left[-F(\underline{\mathbf{h}}_M) \right] \\ &= Q_P(\underline{\mathbf{h}}_P) Q_M(\underline{\mathbf{h}}_M)\end{aligned} \tag{3.27}$$

where $Z_P = \int \exp\left[-F\left(\underline{\mathbf{h}}_P\right)\right] d\underline{\mathbf{h}}_P$ and $Z_M = \int \exp\left[-F\left(\underline{\mathbf{h}}_M\right)\right] d\underline{\mathbf{h}}_M$. In order to characterize $Q_P(\underline{\mathbf{h}}_P)$ we need to allow some positive deviation to $\underline{\mathbf{h}}_P$ (any negative values of $\underline{\mathbf{h}}_P$ will be rejected since NMF only allow nonnegative values). Hence, $\underline{\mathbf{h}}_P$ must take on zero and positive values in $Q_P(\underline{\mathbf{h}}_P)$. The distribution $Q_P(\underline{\mathbf{h}}_P)$ can be approximated by using the Taylor expansion about the MAP estimate, $\underline{\mathbf{h}}^{MAP}$ is given by (3.13):

$$\begin{aligned}
&Q_P(\underline{\mathbf{h}}_P \geq 0) = Q(\underline{\mathbf{h}})\Big|_{\underline{\mathbf{h}}_M = \underline{\mathbf{h}}_M^{MAP}} \\
&\propto \frac{1}{Z_P} \exp\left[-\left\{ F\left(\underline{\mathbf{h}}^{MAP}\right) + \left(\underline{\mathbf{h}} - \underline{\mathbf{h}}^{MAP}\right)^{\mathbf{T}} \frac{\partial F\left(\underline{\mathbf{h}}\right)}{\partial \underline{\mathbf{h}}}\Big|_{\underline{\mathbf{h}}=\underline{\mathbf{h}}^{MAP}} + \frac{1}{2}\left(\underline{\mathbf{h}} - \underline{\mathbf{h}}^{MAP}\right)^{\mathbf{T}} \frac{\partial^2 F\left(\underline{\mathbf{h}}\right)}{\partial \underline{\mathbf{h}} \partial \underline{\mathbf{h}}^{\mathbf{T}}}\Big|_{\underline{\mathbf{h}}=\underline{\mathbf{h}}^{MAP}} \left(\underline{\mathbf{h}} - \underline{\mathbf{h}}^{MAP}\right) \right\}_P \right] \\
&\propto \exp\left\{ -\left[\left(\frac{\partial F}{\partial \underline{\mathbf{h}}} \right)\Big|_{\underline{\mathbf{h}}^{MAP}} \right]_P^{\mathbf{T}} \underline{\mathbf{h}}_P - \frac{1}{2} \underline{\mathbf{h}}_P^{\mathbf{T}} \overline{\mathbf{\Theta}}_P \underline{\mathbf{h}}_P \right\} \\
&\propto \exp\left[-\left(\overline{\mathbf{\Theta}} \underline{\mathbf{h}}^{MAP} - \frac{1}{\sigma^2} \overline{\mathbf{D}}^{\mathbf{T}} \underline{\mathbf{y}} + \underline{\boldsymbol{\lambda}} \right)_P^{\mathbf{T}} \underline{\mathbf{h}}_P - \frac{1}{2} \underline{\mathbf{h}}_P^{\mathbf{T}} \overline{\mathbf{\Theta}}_P \underline{\mathbf{h}}_P \right]
\end{aligned} \tag{3.28}$$

where $\overline{\mathbf{\Theta}} = \frac{1}{\sigma^2} \overline{\mathbf{D}}^{\mathbf{T}} \overline{\mathbf{D}}$, $\overline{\mathbf{\Theta}}_{n,P}$ is the sub-matrix of $\overline{\mathbf{\Theta}}$ corresponds to $\underline{\mathbf{h}}_P$. Although $Q_P(\underline{\mathbf{h}}_P)$ is obtained in the form of (3.28), its integral is difficult to evaluate and does not yield closed analytical form of the moments which subsequently prohibits inference of the sparsity parameters. To overcome this problem, we propose to variationally approximate $Q_P(\underline{\mathbf{h}}_P)$ using the mean-field approximation with a factorized exponential distribution as:

$$\hat{Q}_P(\underline{\mathbf{h}}_P \geq 0) = \prod_{p \in P} \frac{1}{u_p} \exp(-h_p / u_p) \tag{3.29}$$

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

$$\begin{aligned}
\underline{\mathbf{u}} &= \min_{\underline{\mathbf{u}}} \int \hat{Q}_P(\underline{\mathbf{h}}_P) \ln \frac{\hat{Q}_P(\underline{\mathbf{h}}_P)}{Q_P(\underline{\mathbf{h}}_P)} d\underline{\mathbf{h}}_{n,P} \\
&= \arg\min_{\underline{\mathbf{u}}} \int \hat{Q}_P(\underline{\mathbf{h}}_P) \left[\ln \hat{Q}_P(\underline{\mathbf{h}}_P) - \ln Q_P(\underline{\mathbf{h}}_P) \right] d\underline{\mathbf{h}}_P
\end{aligned} \tag{3.30}$$

where

$$\begin{aligned}
&\int \hat{Q}_P(\underline{\mathbf{h}}_P)\ln\left[\hat{Q}_P(\underline{\mathbf{h}}_P)\right]d\underline{\mathbf{h}}_P = \sum_{p\in P}\int \hat{Q}_P(h_p)\ln\left[\hat{Q}_P(h_p)\right]dh_p \\
&= \sum_{p\in P}\int_0^\infty \frac{1}{u_p}\exp\left(-h_p/u_p\right)\left(-\ln u_p - h_p/u_p\right)dh_p \\
&= -\sum_{p\in P}\ln u_p \int_0^\infty \exp\left(-h_p/u_p\right)d\left(\frac{h_p}{u_p}\right) - \sum_{p\in P}\int_0^\infty \frac{h_p}{u_p}\exp\left(-h_p/u_p\right)d\left(\frac{h_p}{u_p}\right) \\
&= -\sum_{p\in P}\ln u_p + 1
\end{aligned} \tag{3.31}$$

and

$$\begin{aligned}
\int \hat{Q}_P(\underline{\mathbf{h}}_P)\ln\left[Q_P(\underline{\mathbf{h}}_P)\right]d\underline{\mathbf{h}}_P &= -\int\left[\left(\overline{\boldsymbol{\Theta}}\underline{\mathbf{h}}^{MAP} - \frac{1}{\sigma^2}\overline{\mathbf{D}}^{\mathbf{T}}\underline{\mathbf{y}} + \underline{\boldsymbol{\lambda}}\right)_P^{\mathbf{T}}\underline{\mathbf{h}}_P + \frac{1}{2}\underline{\mathbf{h}}_P^{\mathbf{T}}\overline{\boldsymbol{\Theta}}_P\underline{\mathbf{h}}_P\right]\hat{Q}_P(\underline{\mathbf{h}}_P)d\underline{\mathbf{h}}_P \\
&= -\sum_{p\in P,m\in M}\frac{1}{2}\left(\overline{\boldsymbol{\Theta}}\right)_{pm}\left\langle h_p h_m\right\rangle - \sum_{p\in P}\left(\overline{\boldsymbol{\Theta}}\underline{\mathbf{h}}^{MAP} - \frac{1}{\sigma^2}\overline{\mathbf{D}}^{\mathbf{T}}\underline{\mathbf{y}} + \underline{\boldsymbol{\lambda}}\right)_p\left\langle h_p\right\rangle
\end{aligned} \tag{3.32}$$

with $\langle\bullet\rangle$ denotes the expectation under the distribution of $\hat{Q}_P(\underline{\mathbf{h}}_P)$ such that $\langle h_p h_m\rangle = u_p u_m$ and $\langle h_p\rangle = u_p$ which leads to:

$$\min_{u_p}\hat{\underline{\mathbf{b}}}_P^{\mathbf{T}}\underline{\mathbf{u}} + \frac{1}{2}\underline{\mathbf{u}}^{\mathbf{T}}\hat{\boldsymbol{\Theta}}\underline{\mathbf{u}} - \sum_{p\in P}\ln u_p \tag{3.33}$$

where $\hat{\underline{\mathbf{b}}}_P = \left(\overline{\boldsymbol{\Theta}}\underline{\mathbf{h}}^{MAP} - \frac{1}{\sigma^2}\overline{\mathbf{D}}^{\mathbf{T}}\underline{\mathbf{y}} + \underline{\boldsymbol{\lambda}}\right)_P$, $\hat{\boldsymbol{\Theta}} = \overline{\boldsymbol{\Theta}}_p + diag\left(\overline{\boldsymbol{\Theta}}_p\right)$ and '$diag(\bullet)$' denotes a matrix with the argument on the diagonal The optimization of (3.33) can be accomplished using nonnegative quadratic programming method [94] as following:

$$G\left(\underline{\mathbf{u}},\tilde{\underline{\mathbf{u}}}\right) = \hat{\underline{\mathbf{b}}}_P^{\mathbf{T}}\underline{\mathbf{u}} + \frac{1}{2}\sum_{p\in P}\frac{\left(\hat{\boldsymbol{\Theta}}\tilde{\underline{\mathbf{u}}}\right)_p}{\tilde{u}_p}u_p^2 - \sum_{p\in P}\ln u_p \tag{3.34}$$

Taking the derivative of $G\left(\underline{\mathbf{u}},\tilde{\underline{\mathbf{u}}}\right)$ in (3.34) with respect to $\underline{\mathbf{u}}$ and setting it to be zero, this gives:

$$\frac{\left(\hat{\boldsymbol{\Theta}}\tilde{\underline{\mathbf{u}}}\right)_p}{\tilde{u}_p}u_p + \hat{b}_p - \frac{1}{u_p} = 0 \tag{3.35}$$

The above equation is equivalent to the following quadratic equations:

$$\frac{\left(\hat{\boldsymbol{\Theta}}\tilde{\underline{\mathbf{u}}}\right)_p}{\tilde{u}_p}u_p^2+\hat{b}_p u_p-1=0 \tag{3.36}$$

Solving (3.36) for u_p leads to the following update equation:

$$u_p \leftarrow u_p \frac{-\hat{b}_p+\sqrt{\hat{b}_p^2+4\dfrac{\left(\hat{\boldsymbol{\Theta}}\underline{\mathbf{u}}\right)_p}{\tilde{u}_p}}}{2\left(\hat{\boldsymbol{\Theta}}\underline{\mathbf{u}}\right)_p} \tag{3.37}$$

As for components $\underline{\mathbf{h}}_M$, $Q_M(\underline{\mathbf{h}}_M)$ has the functional form equivalent to a multivariate Gaussian distribution. Therefore, we propose to approximate $Q_M(\underline{\mathbf{h}}_M)$ as as joint Gaussian with mean $\underline{\mathbf{h}}_M^{MAP}$. Thus using the factorized approximation $Q(\underline{\mathbf{h}})=\hat{Q}_P(\underline{\mathbf{h}}_P)Q_M(\underline{\mathbf{h}}_M)$ in (3.32), the following is obtained:

$$\lambda_p=\begin{cases}\dfrac{1}{h_p^{MAP}} & if\ \ p\in M\\ \dfrac{1}{u_p} & if\ \ p\in P\end{cases} \tag{3.38}$$

for $p=1,2,\ldots,R$ and h_p^{MAP} is the p^{th} element of sparse code $\underline{\mathbf{h}}_P$ computed from (3.13). and its covariance $\mathbf{C}^{ov}$:

$$C_{p,m}^{ov}=\begin{cases}\left(\overline{\boldsymbol{\Theta}}_P^{-1}\right)_{p,m} & if\ \ p,m\in M\\ u_p^2 & \text{Otherwise}\end{cases} \tag{3.39}$$

Thus, the update rule for σ^2 can be obtained as:

$$\sigma^2=\frac{1}{N_0}\left[\left(\underline{\mathbf{y}}-\overline{\mathbf{D}}\underline{\hat{\mathbf{h}}}\right)^{\mathbf{T}}\left(\underline{\mathbf{y}}-\overline{\mathbf{D}}\underline{\hat{\mathbf{h}}}\right)+\mathrm{Tr}\left(\overline{\mathbf{D}}^{\mathbf{T}}\overline{\mathbf{D}}\mathbf{C}^{ov}\right)\right] \ \text{ where } \ \hat{h}_p=\begin{cases}h_p^{MAP} & if\ \ p\in M\\ u_p & if\ \ p\in P\end{cases} \tag{3.40}$$

where '$\mathrm{Tr}(\bullet)$' denotes *trace* function. Table 3.1 shows the specific steps of the proposed v-SNMF2D method. In the table, $\mathbf{G}=|\mathbf{Y}|^{.2}$ with $\mathbf{H}_p^{\phi}$ and $\mathbf{U}_p^{\phi}$ corresponding to the matrix representation of (3.38).

Table 3.1: Proposed v-SNMF2D algorithm

1. Initialize $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ with nonnegative random values.
2. Define $\tilde{\mathbf{D}}^{\tau}_{f,i} = \mathbf{D}^{\tau}_{f,i} \Big/ \sqrt{\sum_{\tau,f} (\mathbf{D}^{\tau}_{f,i})^2}$.
3. Compute $\tilde{\mathbf{Z}} = \sum_i \sum_{\tau} \sum_{\phi} \overset{\downarrow\phi}{\tilde{\mathbf{D}}}{}^{\tau}_{i} \overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi}_{i}$.
4. Compute $\overline{\mathbf{\Theta}}_P = \frac{1}{\sigma^2} \overline{\mathbf{D}}_P^{\mathbf{T}} \overline{\mathbf{D}}_P$ Minimize $\min_{u_p} \hat{\underline{\mathbf{b}}}_P^{\mathbf{T}} \underline{\mathbf{u}} + \frac{1}{2} \underline{\mathbf{u}}^{\mathbf{T}} \hat{\mathbf{\Theta}} \underline{\mathbf{u}} - \sum_{p \in P} \ln u_p$ respect to u_p.
5. Assign $\boldsymbol{\lambda}^{\phi}_{p} = \begin{cases} \dfrac{1}{\mathbf{H}^{\phi}_{p}} & if \ p \in M \\ \dfrac{1}{\mathbf{U}^{\phi}_{p}} & if \ p \in P \end{cases}$.
6. Assign $\sigma^2 = \frac{1}{N_0} \left[\left(\underline{\mathbf{y}} - \overline{\mathbf{D}} \hat{\underline{\mathbf{h}}} \right)^{\mathbf{T}} \left(\underline{\mathbf{y}} - \overline{\mathbf{D}} \hat{\underline{\mathbf{h}}} \right) + \mathrm{Tr}\left(\overline{\mathbf{D}}^{\mathbf{T}} \overline{\mathbf{D}} \mathbf{C}^{ov} \right) \right]$.
7. Update $\mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi} \bullet \frac{\sum_{\tau} \overset{\downarrow\phi^{\mathbf{T}} \leftarrow\tau}{\mathbf{D}^{\tau} \ \mathbf{G}}}{\sum_{\tau} \overset{\downarrow\phi^{\mathbf{T}} \leftarrow\tau}{\mathbf{D}^{\tau} \ \tilde{\mathbf{Z}}} + \boldsymbol{\lambda}^{\phi}_{p}}$.
8. Compute $\tilde{\mathbf{Z}} = \sum_i \sum_{\tau} \sum_{\phi} \overset{\downarrow\phi}{\tilde{\mathbf{D}}}{}^{\tau}_{i} \overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi}_{i}$.
9. Update $\mathbf{D}^{\tau} \leftarrow \tilde{\mathbf{D}}^{\tau} \bullet \frac{\sum_{\phi} \overset{\uparrow\phi \ \rightarrow\tau^{\mathbf{T}}}{\mathbf{G} \ \mathbf{H}^{\phi}}}{\sum_{\phi} \overset{\uparrow\phi \ \rightarrow\tau^{\mathbf{T}}}{\tilde{\mathbf{Z}} \ \mathbf{H}^{\phi}}}$.
10. Repeat steps 2 to 9 until convergence.

3.3 Single Channel Blind Source Separation

3.3.1 Estimated sources

The matrices to determine are $\left\{ |\mathbf{X}_i|^{.2} \right\}_{i=1}^{I}$ and this will be obtained by using the proposed matrix factorization as $\left| \tilde{\mathbf{X}}_i \right|^{.2} = \sum_{\tau} \sum_{\phi} \overset{\downarrow\phi}{\mathbf{D}}{}^{\tau}_{i} \overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi}_{i}$ with $\mathbf{D}^{\tau}_{i}$ and $\mathbf{H}^{\phi}_{i}$ estimated using (3.13) and (3.14). Once these matrices are estimated, the i^{th} binary mask according to $Mask_i(f, t_s) = 1$ is formed if $\left| \tilde{X}_i(f, t_s) \right|^{.2} > \left| \tilde{X}_j(f, t_s) \right|^{.2}$ and zero otherwise. Finally, the estimated time-domain sources are obtained as $\tilde{\mathbf{x}}_i = \mathrm{Resynthesize}\left(\mathbf{Mask}_i \bullet \mathbf{Y} \right)$ where 'Resynthesize'

[92] denotes the inverse mapping of the log-frequency axis to the original frequency axis and followed by the inverse STFT back to the time domain [35] and $\tilde{\mathbf{x}}_i = \left[\tilde{x}_i(1),\ldots,\tilde{x}_i(T)\right]^{\mathbf{T}}$ denotes the i^{th} estimated audio sources in time-domain.

3.3.2 Experiment set up

The proposed method is tested by separating audio sources. Several experimental simulations under different conditions have been designed to investigate the efficacy of the proposed method. All simulations and analyses are performed using a PC with Intel Core 2 CPU 6600 @ 2.4GHz and 2GB RAM. MATLAB is used as the programming platform. To generate mixed signal, a 4 second polyphonic music containing trumpet and piano is analysed. The mixed signal is sampled at 16 kHz sampling rate. The TF representation is computed by normalizing the time-domain signal to unit power and computing the STFT using Hamming window of length 1024 point with 50% overlap between two frames. The frequency axis of the obtained spectrogram is then logarithmically scaled and grouped into 175 frequency bins in the range of 50Hz to 8kHz with 24 bins per octave. This corresponds to twice the resolution of the equal tempered musical scale. For the v-SNMF2D parameters, the convolutive components in time and frequency are selected to be $\tau = \{0,1,2,3\}$ and $\phi = \{0,\ldots,31\}$, respectively. The corresponding sparse factor was determined by (3.38).

3.3.3 Quality Evaluation

The separation performance in terms of the signal-to-distortion ratio (SDR) is used for

evaluation. This is a global measure that unifies signal-to-interference ratio (SIR), and signal-to-artifacts ratio (SAR) [95, 96]. Specifically, the above three metrics are described as follow:

1. Signal-to-distortion ratio (SDR) — this is an overall measure of performance as it accounts for both of the SIR and SAR criteria.
2. Signal-to-interference ratio (SIR) — this is a measure of the suppression of the unwanted source.
3. Signal-to-artifacts ratio (SAR) — this is a measure of the artifacts (such as musical noise) that have been introduced by the separation process.

The goal is to maximize SIR (as this is the measure of the actual separation) while trying to keep SAR as high as possible (in order to prevent the introduction of artifacts). In order to compute these metrics, a given estimated time domain signal $\tilde{x}_i(t)$ is decomposed as a sum of the following parts:

1. x_{target} : actual source estimate.
2. e_{interf} : interference signal (i.e. the unwanted source).
3. e_{artif} : artifacts of the separation algorithm.

The decomposition is done up to a constant scaling factor. Using these terms, the metrics are computed as follows:

$$\text{SIR}=\frac{\left\|x_{\text{target}}\right\|^2}{\left\|e_{\text{interf}}\right\|^2},\quad \text{SAR}=\frac{\left\|x_{\text{target}}+e_{\text{interf}}\right\|^2}{\left\|e_{\text{artif}}\right\|^2}\quad \text{and}\quad \text{SDR}=\frac{\left\|x_{\text{target}}\right\|^2}{\left\|e_{\text{interf}}+e_{\text{artif}}\right\|^2} \tag{3.41}$$

3.3.4 Impact of sparsity

In this implementation, several experiments have been conducted to compare the performance of the proposed method with SNMF2D under different sparsity regularization. To investigate the impact of sparsity regularization on source separation performance, three cases[2] are conducted:

Case (i): Uniform constant sparsity with low sparseness, $\lambda^{\phi}_{i,t_s} = \lambda = 0.01$ for all i, t_s, ϕ.

Case (ii): Uniform constant sparsity with high sparseness, $\lambda^{\phi}_{i,t_s} = \lambda = 100$ for all i, t_s, ϕ.

Case (iii): Proposed adaptive sparsity according to (3.38).

The time and TF domain of the original trumpet, piano music and its mixture are shown in Figure 3.2. The trumpet and the piano play a different short melodic passage each consisting of three distinct notes. However, both trumpet and piano overlap in time, and the piano notes are interspersed in frequency with the trumpet notes. Hence, this is a challenging task for single channel source separation which will test the impact of sparsity for matrix factorization.

[2] Cases (i) and (ii) correspond to the two-dimensional sparse nonnegative matrix deconvolution (SNMF2D) [92]. This section therefore presents the comparison of our proposed method with the SNMF2D with uniform constant sparsity.

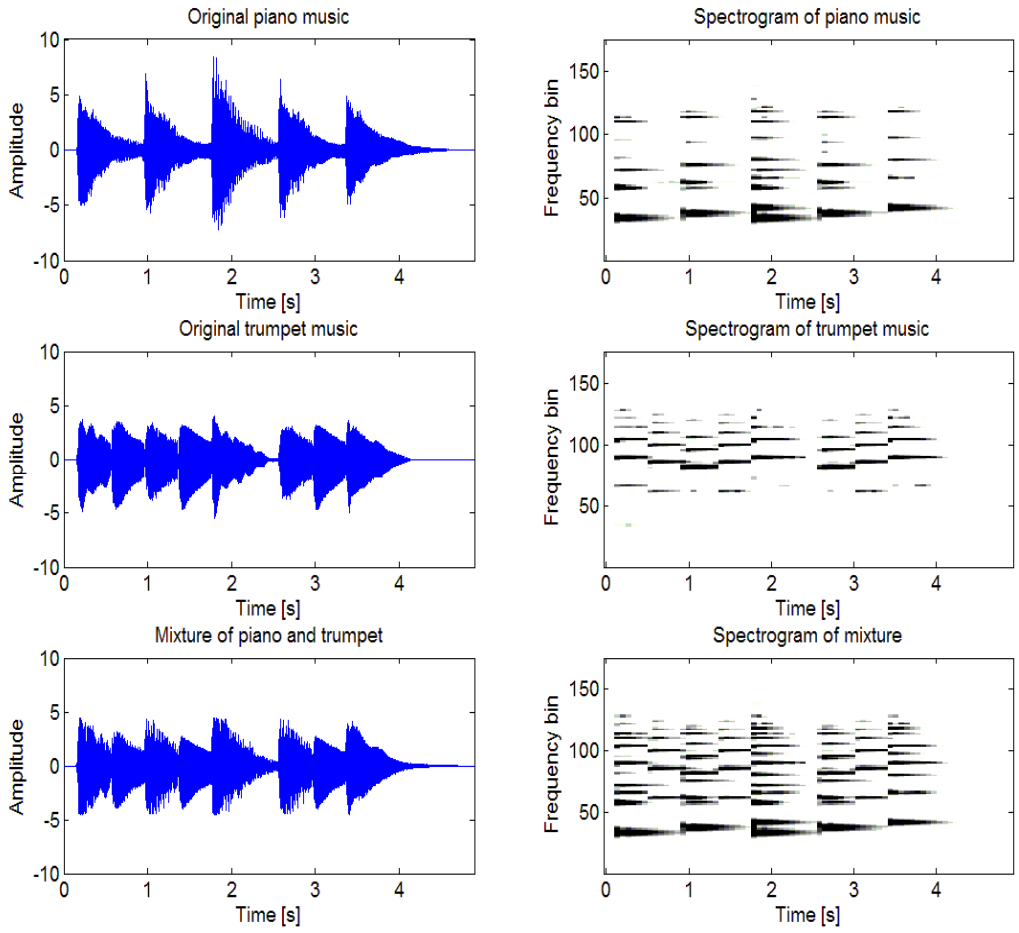


Figure 3.2: Time-domain representation and spectrogram of the piano music (top panels), trumpet music (middle panels) and mixed signal (bottom panels).

3.3.4.1 Estimated spectral basis and temporal code

Figures 3.3 to 3.5 show the results of the matrix factorization in terms of spectral basis $\mathbf{D}_i^{\tau}$ and temporal code $\mathbf{H}_i^{\phi}$ for cases (i) to (iii), respectively. Figure 3.3 shows the case of 'under-sparse' factorization which is clearly evident by the spreading of the estimated temporal codes. Figure 3.4 shows the case of 'over-sparse' factorization where some of the temporal codes have been discarded. On the other hand, Figure 3.5 shows the case of 'optimally-sparse' factorization based on the proposed adaptive tuning of the sparsity parameter.

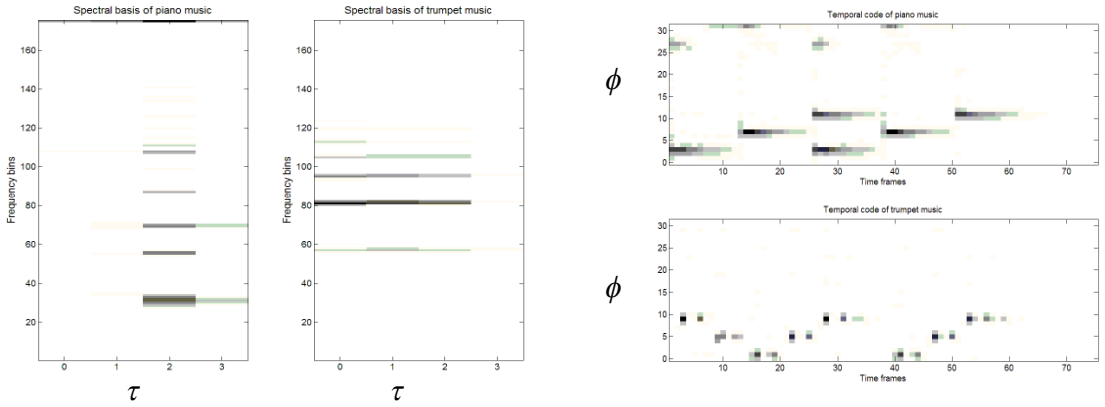


Figure 3.3: Estimated $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ for Case (i).

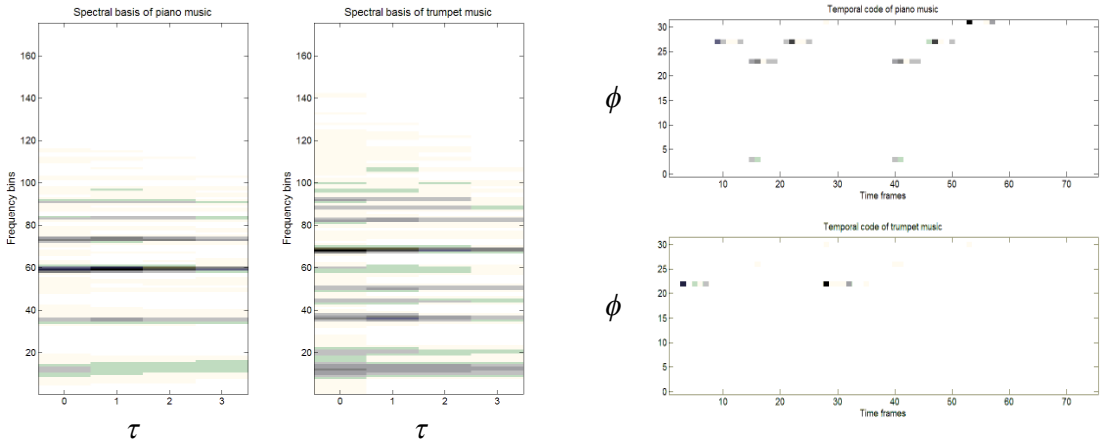


Figure 3.4: Estimated $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ for Case (ii).

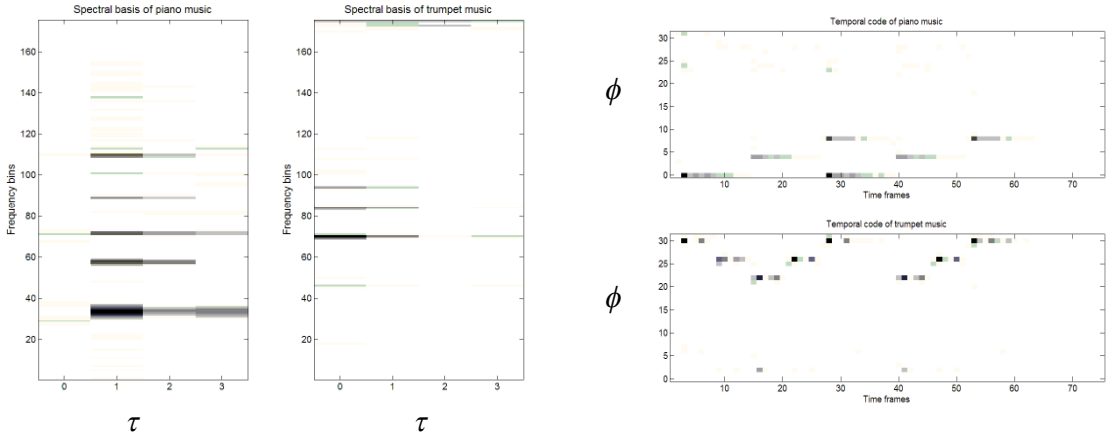


Figure 3.5: Estimated $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ for Case (iii).

3.3.4.2 Audio source separation results

In above, the analysis of the sparsity factorization was presented in terms of $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$. In the following, the audio source separation results for each case are shown. Figures 3.6 and 3.7 show the separated sources in terms of spectrogram and time-domain representation, respectively.

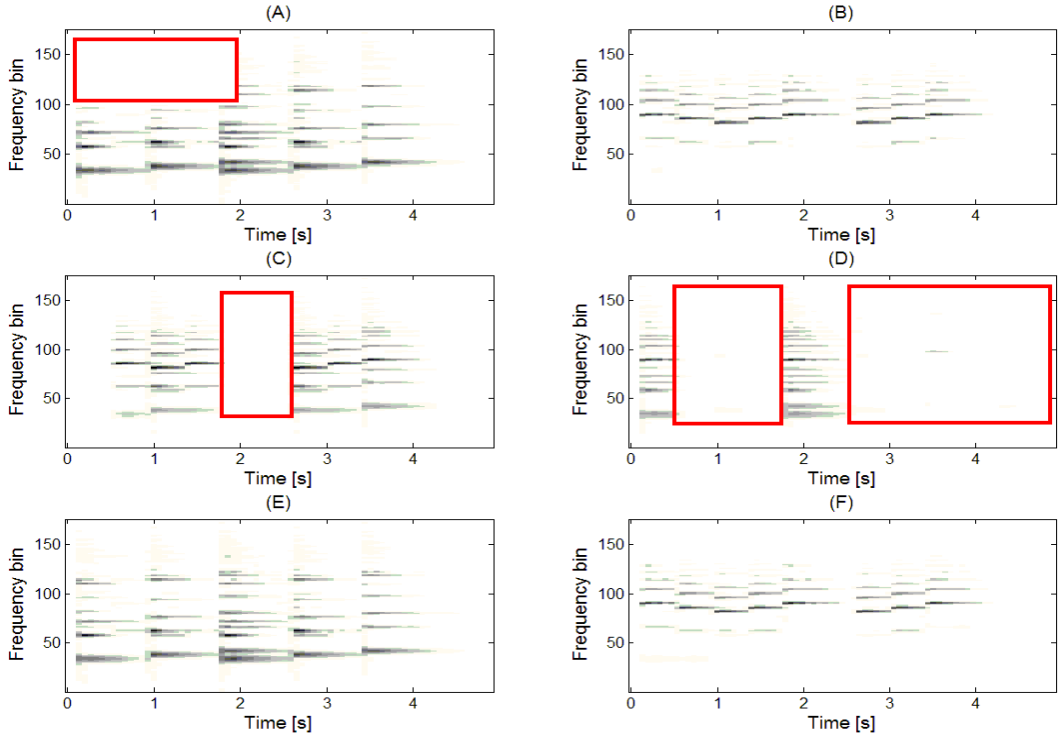


Figure 3.6: Separated signals in spectrogram. (A)-(B): piano and trumpet music for case (i). (C)-(D): piano and trumpet music for case (ii). (E)-(F): piano and trumpet music for case (iii).

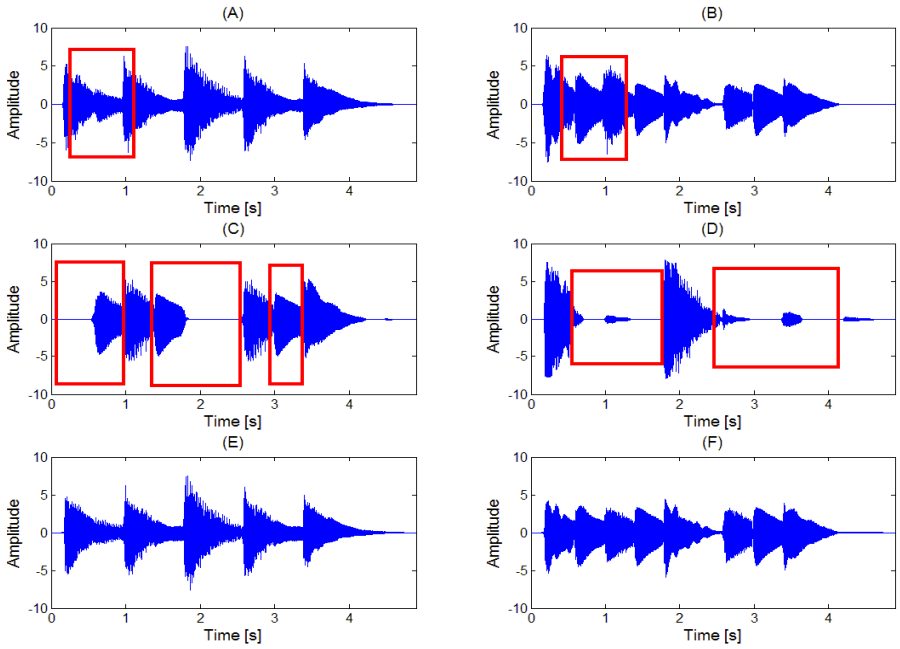


Figure 3.7: Separated signals in time-domain. (A)-(B): piano and trumpet music for case(i). (C)-(D): piano and trumpet music for case(ii). (E)-(F): piano and trumpet music for case(iii).

Panels (A)-(D) in both Figures 3.6 and 3.7 clearly show that better source separation results require careful selection of the sparsity regularization. In the case of 'under-sparse' factorization (e.g. (A)-(B)), the factorization still contains the mixed components (as indicated by the red box marked area) in each separated source. In the case of over-sparse factorization (e.g. (C)-(D)), the spectral basis of the source occurs too rarely in the spectrogram and this results in lesser information which do not fully recover the original source as noted in the middle panels (indicated by the red box marked area). In the case of the proposed method (e.g. (E)-(F)), it assigns a regularization parameter to each temporal code which is individually and adaptively tuned to yield the optimal number of times the spectral basis of a source recurs in the spectrogram. The sparsity on $\mathbf{H}_i^{\phi}$ is imposed *element-wise* in the proposed model so that each individual code in $\mathbf{H}_i^{\phi}$ is optimally sparse in the L_1-norm. In the conventional SNMF2D method, the sparsity is not fully

controlled but is imposed uniformly on all the codes. The ensuing consequence is that the temporal codes are no longer optimal and this leads to 'under-sparse' or 'over-sparse' factorization which eventually results in inferior separation performance.

The analysis for cases (i) and (ii) in Figures 3.6 and 3.7 is based on a single fixed uniform sparsity parameter i.e. $\lambda_{i,t_s}^{\phi} = \lambda$ for all i, t_s, ϕ where λ is set to be either very high or very low. It might be argued that such settings of uniform sparsity parameter are unrealistic for source separation. Therefore, in this sub-section, the performance comparison will be investigated when the uniform constant sparsity parameter is progressively varied from 0 to 10 with every increment of 0.1 (i.e. $\lambda = 0, 0.1, 0.2, \ldots, 10$) and the best result is retained and tabulated in Table 3.2.

Table 3.2: Performance comparison between different sparsity methods

Estimated sources	**Methods**	**SDR**	**SAR**	**SIR**
Recovered trumpet music	Proposed sparsity	10.1	12.3	12.6
	(Best) Uniform sparsity	8.2	10.4	10.1
Recovered piano music	Proposed sparsity	11.2	13.4	13.8
	(Best) Uniform sparsity	8.6	10.1	10.5

From Table 3.2, the performance improvement of the proposed method against the uniform constant sparsity method can be summarised as follows: (i) For the recovered trumpet music, the improvement per source in terms of the SDR is 1.9dB. (ii) For the recovered piano music, the improvement per source in terms of SDR is 2.6dB. Analysing the separation results, there is clear indication that when the sparse parameter is uncontrolled, this will result in poorer separation results than that based on adaptive

sparsity. Compared with the uniform constant sparsity, the proposed method renders a more accurate part based regularised factorization as indicated in Table 3.2.

3.3.4.3 Adaptive behavior of sparsity parameter

In this sub-section, we will show the obtained results of the sparsity parameters adapted by using the proposed method. Several sparsity parameters have been selected to illustrate its adaptive behavior. Figure 3.8 shows the convergence trajectory of four adaptive sparsity parameters $\lambda_{1,1}^{\phi=0}$, $\lambda_{1,5}^{\phi=0}$, $\lambda_{1,10}^{\phi=0}$ and $\lambda_{1,15}^{\phi=0}$ corresponding to their respected temporal codes. All sparsity parameters are initialized as $\lambda_{i,t_s}^{\phi}=10$ for all i,t_s,ϕ and are subsequently adapted according (3.38). After 150 iterations, the above sparsity parameters converge to their steady-states. It is noted that these values are significantly different for each sparsity parameter e.g. $\lambda_{1,1}^{\phi=0}=24.4$, $\lambda_{1,5}^{\phi=0}=1.98$, $\lambda_{1,10}^{\phi=0}=5.87$ and $\lambda_{1,15}^{\phi=0}=17.46$. In addition, it is worth pointing out that the SDR result scales up to 10.6dB when λ_{i,t_s}^{ϕ} is adaptive. This represents a 2dB improvement over the case of uniform constant sparsity (which is only 8.4dB in Table 3.2). In summary, the above results are clear to indicate that the performance of source separation have been undermined when uniform constant sparsity is imposed on all temporal codes. On the other hand, significant improved performances can be obtained by allowing the sparsity parameters to be individually adapted for each temporal code.

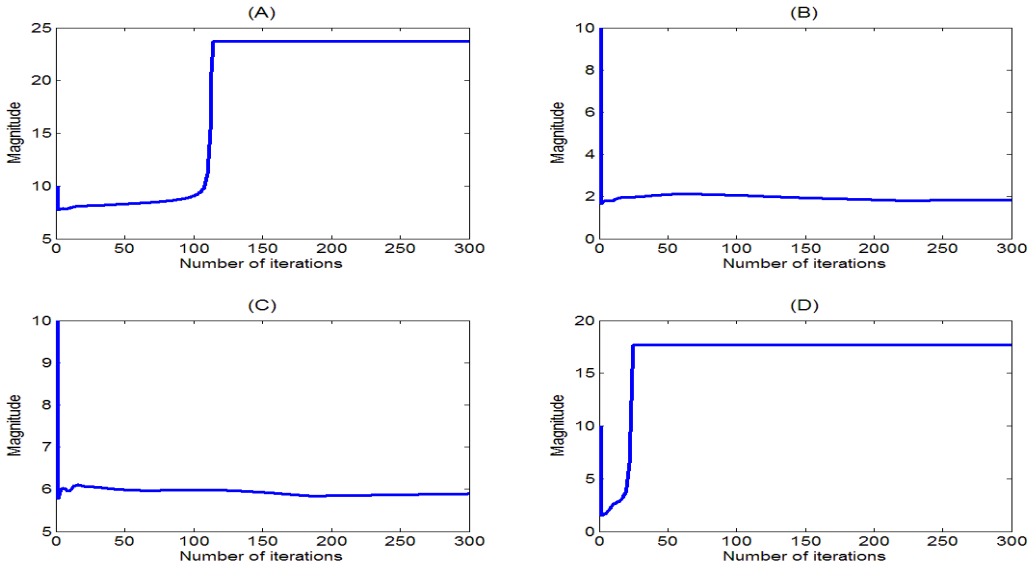


Figure 3.8: Convergence trajectory of the sparsity: (A) $\lambda_{1,1}^{\phi=0}$, (B) $\lambda_{1,5}^{\phi=0}$, (C) $\lambda_{1,10}^{\phi=0}$, (D) $\lambda_{1,15}^{\phi=0}$.

3.3.5 Comparison with other sparse NMF-based SCBSS methods

In section 3.3.4, analysis has been carried out to investigate effects between adaptive sparsity and uniform constant sparsity on source separation. In this evaluation, the proposed method will be compared with other sparse NMF-based SCBSS methods. These consist of the following:

- NMF with Temporal Continuity and Sparseness Criteria [37] (NMF-TCS) is based on factorizing the magnitude spectrogram of the mixed signal into a sum of components, which include the temporal continuity and sparseness criteria into the separation framework.
- SNMF (a multiplicative update algorithm by Lee and Seung [74]).
- Automatic Relevance Determination NMF (NMF-ARD) [97] exploits a hierarchical Bayesian framework SNMF that amounts to imposing an exponential prior for pruning and thereby enables estimation of the NMF model order.

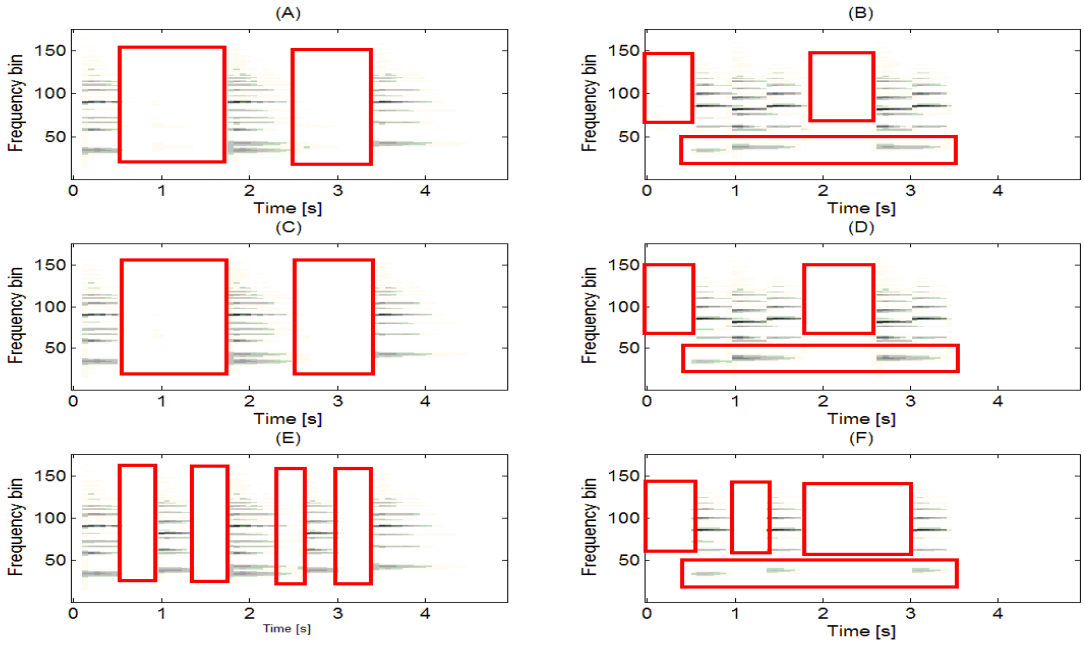


Figure 3.9: Separated signals in spectrogram. (A)-(B): piano and trumpet music using SNMF. (C)-(D): piano and trumpet music using NMF-ARD. (E)-(F): piano and trumpet music using NMF-TCS.

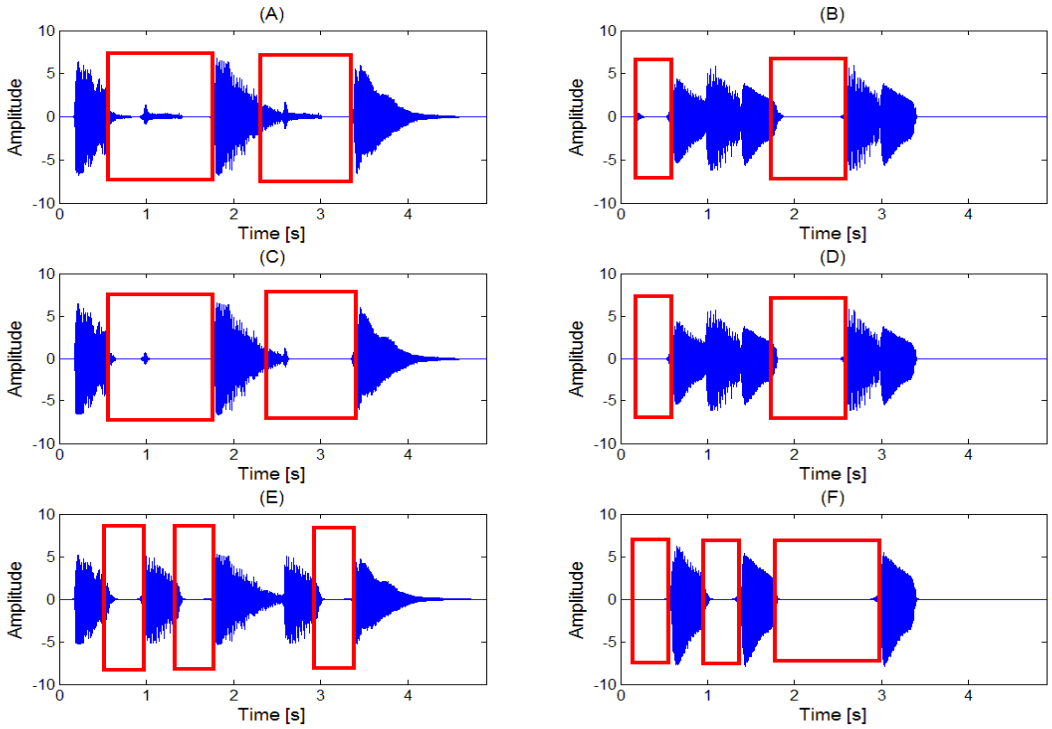


Figure 3.10: Separated signals in time-domain. (A)-(B): piano and trumpet music using SNMF. (C)-(D): piano and trumpet music using NMF-ARD. (E)-(F): piano and trumpet music using NMF-TCS.

In Figure 3.9 and 3.10, panels (A)-(F) show that the NMF and SNMF are weak models since it does not take into account the relative position of each spectrum thereby discarding the temporal information. Better separation results require the model that can represent both temporal structure and the pitch change which occurs when an instrument plays different notes simultaneously. If the temporal structure and the pitch change are not considered in the model, the mixing ambiguity will still contain (marked red box) in each separated source. Table 3.3 further gives the SDR, SAR and SIR comparison results between our proposed method and the above three sparse NMF methods.

Table 3.3: Performance comparison between different methods

Mixtures	**Methods**	**SDR**	**SAR**	**SIR**
Recovered trumpet music	Proposed method	10.1	12.3	12.6
	NMF-TCS	4.6	7.6	7.8
	SNMF	3.7	5.4	6.5
	NMF-ARD	3.3	6.2	6.9
Recovered piano music	Proposed method	11.2	13.4	13.8
	NMF-TCS	4.3	7.2	5.2
	SNMF	2.8	5.1	3.4
	NMF-ARD	3.1	6.5	4.1

The improvement of the proposed method compared with NMF-TCS, SNMF and NMF-ARD can be summarised as follows: (i) for the recovered trumpet music, the average improvement in terms of SDR is 5.5dB (ii) for the recovered piano music, the average improvement in terms of SDR is 7dB. Analysing the separation results, the proposed method leads to the best separation performance for both recovered sources. The SNMF method performs with poorer results whereas the separation performance by the NMF-TCS method is slightly better than the NMF-ARD and SNMF methods. The proposed method

gives significantly better performance than the NMF-TCS, SNMF and NMF-ARD methods. The reasons are: Firstly, the SNMF and NMF-ARD do not have convolutive factors $\tau, \phi = \{0\}$. As such, SNMF and NMF-ARD are weak models since they do not take into account the relative position of each spectrum thereby discarding the temporal information. The spectral basis obtained via NMF-TCS, SNMF and NMF-ARD methods are not adequate to capture the temporal dependency of the frequency patterns within the audio signal. Secondly, the NMF-TCS, SNMF and NMF-ARD do not model 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.

3.4 Summary

This chapter has presented a new variable regularised two-dimensional sparse nonnegative matrix factorization. The impetus behind this is that the sparsity achieved by conventional NMF, SNMF, NMF2D and SNMF2D methods is not enough; in such situations it is useful to control the degree of sparseness explicitly. In the proposed method, the regularization term is adaptively tuned using a variational Bayesian approach to yield desired sparse decomposition, thus enabling the spectral basis and temporal codes of non-stationary audio signals to be estimated more efficiently. This has been verified based on the simulations. In addition, the proposed method has yielded significant improvements in single channel audio source separation when compared with other sparse NMF-based source separation methods.

CHAPTER 4

SINGLE CHANNEL BLIND SOURCE SEPARATION USING EMD-SUBBAND VARIABLE REGULARISED SPARSE FEATURES

In the previous chapter, the novel v-SNMF2D based SCBSS method has been proposed to separate music mixtures only. In this chapter, a new framework for SCBSS to separate all types of audio mixtures based on the EMD and v-SNMF2D is proposed. The proposed solution separates audio sources from single channel without relying on training information about the original sources. Audio signals are mostly non-stationary and the EMD decomposes the mixed signal into a collection of oscillatory basis components termed as intrinsic mode functions (IMFs) which contain the basic properties of the original source (e.g. amplitude and frequency). In the proposed scheme, instead of processing the mixed signal directly, the IMFs are utilized as the new set of observations. The impetus behind this is that the degree of mixing of the sources in the IMF domain is now less ambiguous and thus, the dominating source in the mixture is more easily detected. Moreover, the spectral and temporal patterns (i.e. the spectral bases and temporal codes, respectively) associated with each IMF are now simpler and sparser than that of the mixed signal. As such, these patterns can be extracted using a suitably designed sparse algorithm. To this end, the proposed v-SNMF2D is used to complete the separation process. The proposed variable regularization benefits conventional SNMF2D in terms of improved

accuracy in resolving spectral bases and temporal codes which were previously not possible by using SNMF2D alone. This benefit has been extended to SCSS by merging the proposed v-SNMF2D *with* EMD.

The chapter is organized as follows: Section 4.1 introduces the background of EMD. In Section 4.2, the proposed source separation framework is fully developed. Experimental results coupled with a series of performance comparison with other SCBSS techniques are presented in Section 4.3. Finally, Section 4.4 concludes this chapter.

4.1 Background

4.1.1 Empirical mode decomposition

EMD is a signal processing tool for decomposing any non-stationary signal into oscillating components by empirically identifying the physical time scales intrinsic to the data. These oscillating components are termed as the intrinsic mode functions (IMF). For in-depth information on EMD, interested readers are referred to [98-104]. In principle, the IMFs satisfy two fundamental conditions: Firstly, in the whole dataset, the number of extrema (minima and maxima) and the number of zero crossing must be same or differ at most by one. Secondly, the mean value of envelop defined by the local minima is always zero. The first condition is obvious; it is similar to the traditional narrow band requirements for a stationary Gaussian process. The second condition is a relatively new idea for non-stationary data; it modifies the classical global requirement to a local one. The specific steps to decompose arbitrary data series into IMF components [40] can be summarised as:

i). Determine all the maxima and minima of the series $y(t)$.

ii). Generate the lower $Low(t)$ and upper $High(t)$ envelops for connecting the maxima and minima with cubic *spline* function.

iii). Point by point averaging the two envelops to calculate the local mean series as $\varphi^{EMD}(t) = \frac{\left(Low(t) + High(t)\right)}{2}$.

iv). A new data series $h_1^{EMD}(t)$ can be obtained by subtracting the local mean series from $h_1^{EMD}(t) = y(t) - \varphi^{EMD}(t)$. Check the properties of: if not an IMF, replace $y(t)$ with $h_1^{EMD}(t)$ and repeat χ times when the procedures from step one until the local mean envelop is approximate to zero. The first IMF component, $c_1(t)$ then can be extracted from data $c_1(t) = h_\chi^{EMD}(t)$ and its residue $r_1^{EMD}(t)$ are evaluated as: $r_1^{EMD}(t) = y(t) - c_1(t)$.

v). Once the first IMF is obtained which represents the highest frequency component of the original series. The residual signal still contains information of $y(t)$. The procedure is repeated for all subsequent residues until the range below a predetermined level or the residue has a monotonic trend.

The final results is: $r_2^{EMD}(t) = r_1^{EMD}(t) - c_2(t), \cdots, r_n^{EMD}(t) = r_{n-1}^{EMD}(t) - c_n(t)$. At the end of decomposition, the mixed signal can be represented as:

$$y(t) = \sum_{n=1}^{N} c_n(t) + r_N^{EMD}(t) \tag{4.1}$$

where $c_n(t)$ is the n^{th} IMF, N is the total number of IMFs, and $r_N^{EMD}(t)$ is the final residue. Figure 4.1 shows the EMD of a signal mixture (panel (A)) containing a male and a female speech. The IMFs (panels (B)-(G)) are similar to the bandlimited functions for representing the time series data. Therefore, EMD is suitable for analysing non-stationary data and can be considered as a dyadic filterbank with each narrow band contains most energy of one

dominating source. Also, the frequency of IMFs decreases as the order increases e.g. the 6th IMF contains lower frequency components of the mixture than that of the 5th IMFs.

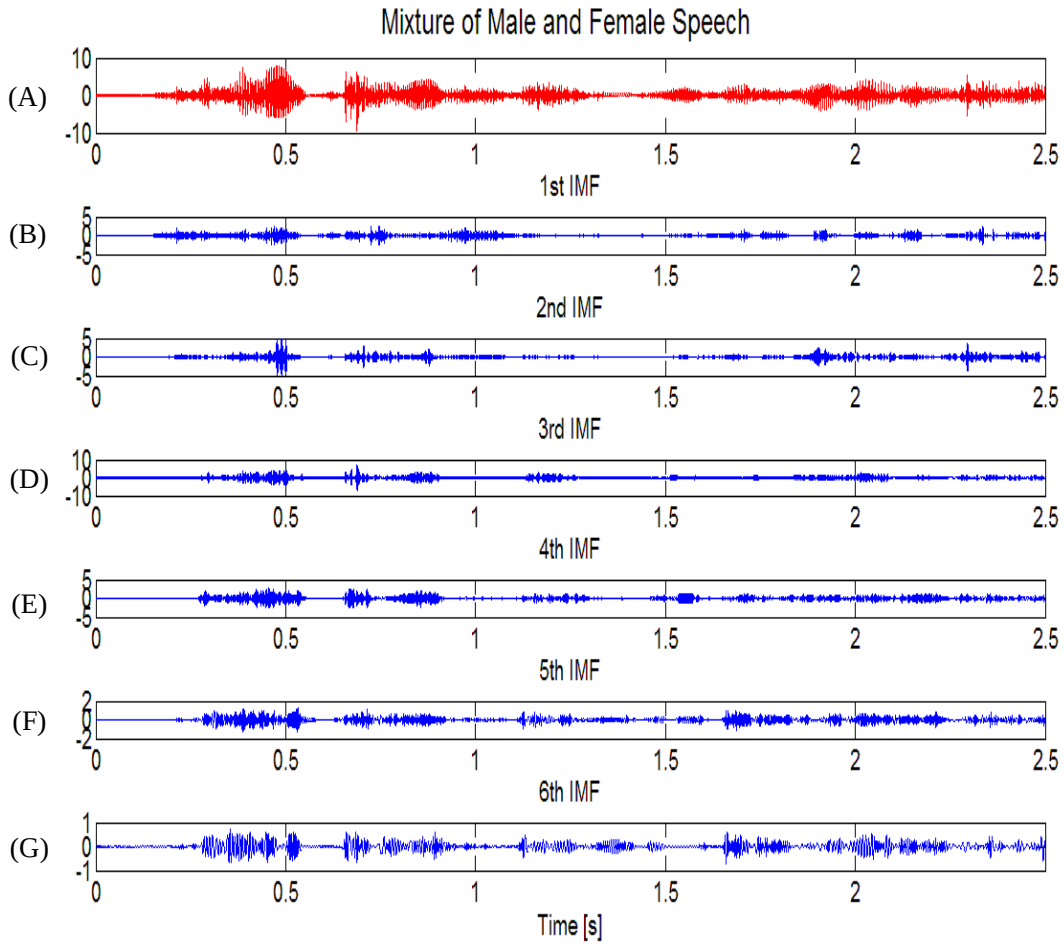


Figure 4.1: EMD of male-female speech mixture showing the first six (out of 10) IMFs.

4.2 Proposed Separation Method

In this section, the foundation of how EMD and matrix factorization can be unified within the context of SCBSS. Three benefits will be obtained from this merger. The EMD decomposes the audio mixture signal as a collection of IMFs as follows:

$$y(t) = \sum_{n=1}^{N} c_n^y(t) + r_N^{EMD}(t) \tag{4.2}$$

These IMFs which are derived from the data can serve as the basis of expansion, which can be linear or nonlinear as dictated by the data. In addition, it is complete and almost orthogonal. Thus, the extracted IMFs are real-valued signals [98] that contain the basic properties of the original source. From the filtering point of view, the EMD process can be considered as a dynamic filterbank where the bandwidths are ranged automatically and dependent on the input signal. This is unlike the conventional filterbank which has fixed bandwidths that are independent of the input signal. Given the nature of this dynamic filterbank, the first benefit EMD brings to SCBSS is as follows: For each IMF of the mixed signal, the degree of mixing from the original sources is considerably reduced in that particular sub-band of frequencies. To validate this finding, the $\vartheta_{n,i}$ is defined to measure the dominating factor of the i^{th} original source on the n^{th} IMF as follows:

$$\vartheta_{n,i} = 1 - \frac{\sum_{t} \left| x_i(t) - c_n^y(t) \right|^2}{\sum_{i=1}^{2} \sum_{t} \left| x_i(t) - c_n^y(t) \right|^2} \tag{4.3}$$

In this analysis, a mixture of male ($x_1(t)$) and female ($x_2(t)$) speeches is used. The dominating factor of each source to each IMF is tabulated in Table 4.1. The higher value of $\vartheta_{n,i}$, the more contribution from the i^{th} source is to the n^{th} IMF. From Table 4.1, it is observed that the value is high on either $\vartheta_{n,1}$ or $\vartheta_{n,2}$ which indicates that the mixing at the IMF levels is dominated either by source 1 or source 2, respectively. In this example, it is clear that source 1 dominates in the 1st and 5th – 7th IMFs while source 2 dominates in the 2nd-4th IMFs.

Table 4.1: Domination proportion of each source signal to each IMF

n^{th} IMF	$\vartheta_{n,1}$ (%)	$\vartheta_{n,2}$ (%)
1st IMF	64.38%	35.62%
2nd IMF	42.53%	57.47%
3rd IMF	32.64%	67.36%
4th IMF	36.61%	63.39%
5th IMF	66.82%	33.18%
6th IMF	66.01%	33.99%
7th IMF	66.03%	33.97%

The second benefit EMD brings to SCBSS is that since each IMF corresponds to a filtered signal bounded within a particular range of sub-band frequencies, the complexity of the spectral basis and temporal code associated with each IMF will be simpler and sparser than that of the mixed signal. The degree of sparsity depends on the sources and the order of the IMF. Not only that, it is also found that the sparsity varies across all the IMF order. This is shown in Figure 4.2. This effectively means that in the TF domain of each IMF there is a relatively clear distinction of the spectral basis and temporal code between the dominating source and the less dominating one. As a result, lesser number of components is used in the NMF and yet able to maintain a robust source separation performance. This will be elaborated in Section 4.3. In addition, the sparseness of the IMF suits the proposed v-SNMF2D method since it enables the user to correctly select the model order for the convolutive factors ($\tau_{\max}$ and $\phi_{\max}$ in (3.5)). Finally, the third benefit is since all IMFs are almost orthogonal; the statistical contents in each IMF are relatively decoupled from each other. Therefore, each IMF can be treated independently; when any error is resulted from the processing, this will be confined to that particular IMF only. At the source reconstruction stage, this error will be averaged over all the IMFs; thus the

contribution of this error to the reconstructed source will be minimized.

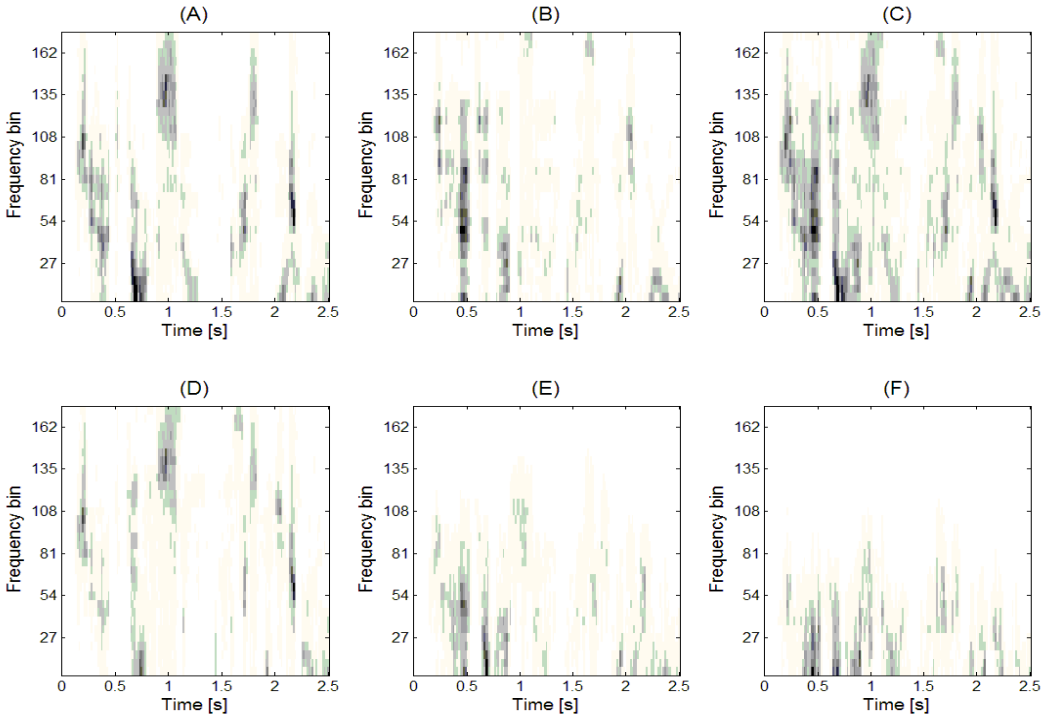


Figure 4.2: (A)-(B) denote the spectrogram of male and female speeches, respectively. (C) denotes the spectrogram of mixed speech (male + female). (D)-(F) denote the spectrogram of the first three IMFs decomposed by EMD. The spectral and temporal patterns' complexity associated with each IMF (D)-(F) is simpler and sparser than the mixed speech (C).

During the decomposition, the maximum IMF order is determined by assessing whether the n^{th} IMF is of acceptable quality as judged by its power ($10\log_{10}\left(\sum_{t=1}^{T}\left|c_n^y(t)\right|^2\right)$) relative to the mixture's power $10\log_{10}\left(\sum_{t=1}^{T}\left|y(t)\right|^2\right)$. In this thesis, a threshold has been set at 5% of the mixture's power. For example, if the n^{th} IMF power is less than a pre-specified threshold of mixture signal, this particular IMF will be rejected. By using this threshold approach, it is possible to consistently select the most significant IMFs. For simplicity, N is assumed as the maximum order and therefore, the mixture signal can be modeled as:

$$\hat{y}(t) = \sum_{n=1}^{N} c_n^y(t) \tag{4.4}$$

In vector form, (4.4) can be written as:

$$\hat{\mathbf{y}} = \mathbf{C}_{imf}^{y}\mathbf{1}_{N} \tag{4.5}$$

where $\mathbf{C}_{imf}^{y} = \left[\mathbf{c}_1^{y}, \mathbf{c}_2^{y}, \cdots, \mathbf{c}_N^{y}\right]$ with $\mathbf{c}_n^{y} = \left[c_n^{y}(1), \ldots, c_n^{y}(T)\right]^{\mathbf{T}}$, $\hat{\mathbf{y}} = \left[\hat{y}(1), \hat{y}(2), \ldots \hat{y}(T)\right]^{\mathbf{T}}$ and $\mathbf{1}_N$ is a vector $\mathbf{1}_N = [1, \ldots, 1]^{\mathbf{T}}$ consist of N components of unit scalar. Similarly, the original sources can be decomposed using the EMD as:

$$\mathbf{x}_1 = \mathbf{C}_{imf}^{x_1}\mathbf{1}_{N_1} \qquad \text{and} \qquad \mathbf{x}_2 = \mathbf{C}_{imf}^{x_2}\mathbf{1}_{N_2} \tag{4.6}$$

where $\mathbf{C}_{imf}^{x_1} = \left[\mathbf{c}_1^{x_1}, \mathbf{c}_2^{x_1}, \ldots, \mathbf{c}_{N_1}^{x_1}\right]$ and $\mathbf{C}_{imf}^{x_2} = \left[\mathbf{c}_1^{x_2}, \mathbf{c}_2^{x_2}, \ldots, \mathbf{c}_{N_2}^{x_2}\right]$ which contains N_1 and N_2 number of IMFs, respectively. The $\left\{\mathbf{c}_n^{x_1}\right\}$ and $\left\{\mathbf{c}_n^{x_2}\right\}$ are defined as the *sub-sources* of $x_1(t)$ and $x_2(t)$, respectively. The aim is to estimate these sub-sources given only $\left\{\mathbf{c}_n^{y}\right\}$, assign each of them to the correct source class and finally reconstruct the estimated sources in the time domain.

4.2.1 Matrix representation of IMFs in TF domain

To estimate the sub-sources, $\mathbf{c}_n^{y}$ from (4.6) is projected into the TF domain, in which the mixed signal becomes:

$$C_n^{y}(f, t_s) = C_n^{x_1}(f, t_s) + C_n^{x_2}(f, t_s) \qquad \text{for } n = 1,2,\ldots,N \tag{4.7}$$

where $C_n^{y}(f, t_s)$, $C_n^{x_1}(f, t_s)$ and $C_n^{x_2}(f, t_s)$ denote the TF components obtained by applying the STFT e.g. $C_n^{z}(f, t_s) = STFT\left(c_n^{z}(t)\right)$ for $z = y$, x_1 and x_2. In practice, the frequency axis of the spectrogram for the audio signals is logarithmically scaled and this convention has been adopted in this chapter. The power spectrogram is defined as the squared magnitude of (4.7):

$$\left|C_n^{y}(f, t_s)\right|^2 = \left|C_n^{x_1}(f, t_s)\right|^2 + \left|C_n^{x_2}(f, t_s)\right|^2 + 2\left|C_n^{x_1}(f, t_s)\right|\left|C_n^{x_2}(f, t_s)\right|\cos\left(\theta_n(f, t_s)\right) \tag{4.8}$$

where $\theta_n(f,t_s)$ measures the projection of $C_n^{x_1}(f,t_s)$ onto $C_n^{x_2}(f,t_s)$. For large sample size, the $C_n^{x_1}(f,t_s)$ and $C_n^{x_2}(f,t_s)$ are assumed as orthogonal and hence, $\theta_n(f,t_s)=\pi/2$. However, for finite sample size, $\theta_n(f,t_s)=\pi/2$ may not hold and the $2\left|C_n^{x_1}(f,t_s)\right|\left|C_n^{x_2}(f,t_s)\right|\cos\left(\theta_n(f,t_s)\right)$ is treated as the residual noise. Note that in (4.8) each component is a function of f and t_s variables. As such, a matrix representation for each component can be represented as $\mathbf{C}^z_{n(f,t_s)}=\left[C_n^z(f,t_s)\right]_{t_s=1,2,\ldots T_s}^{f=1,2,\ldots F}$ where row and column vector represents the time slots and frequency bins respectively. Hence, (4.8) becomes:

$$\text{(Synthesis)} \qquad \left|\mathbf{C}^y_{n(f,t_s)}\right|^2=\left|\mathbf{C}^{x_1}_{n(f,t_s)}\right|^2+\left|\mathbf{C}^{x_2}_{n(f,t_s)}\right|^2+\mathbf{V}_n^{No} \tag{4.9}$$

where $\mathbf{V}_n^{No}$ is the residual noise. Eqn. (4.9) is a synthesis equation since it describes how $\left|\mathbf{C}^y_{n(f,t_s)}\right|^2$ is generated as a mixing of $\left|\mathbf{C}^{x_1}_{n(f,t_s)}\right|^2$, $\left|\mathbf{C}^{x_2}_{n(f,t_s)}\right|^2$ and $\mathbf{V}_n^{No}$. Note that all elements in $\left|\mathbf{C}^{x_1}_{n(f,t_s)}\right|^2$ and $\left|\mathbf{C}^{x_2}_{n(f,t_s)}\right|^2$ are nonnegative whereas the elements in $\mathbf{V}_n^{No}$ could be both positive and negative. However, the overall sum in (4.9) is always nonnegative and therefore, an analysis equation in a form of matrix factorization can be constructed. The model of the proposed factorization algorithm termed as the v-SNMF2D is given as follows:

$$\text{(Analysis)} \qquad \left|\mathbf{C}^y_{n(f,t_s)}\right|^2=\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}}\mathbf{D}_n^{\tau\downarrow\phi}\,\mathbf{H}_n^{\phi\rightarrow\tau}+\mathbf{V}_n^{No}=\sum_{i=1}^{I}\sum_{\tau=0}^{\tau_{\max}}\sum_{\phi=0}^{\phi_{\max}}\mathbf{D}_{n,i}^{\tau\downarrow\phi}\,\mathbf{H}_{n,i}^{\phi\rightarrow\tau}+\mathbf{V}_n^{No} \tag{4.10}$$

$$\text{where}\quad \mathbf{H}_n^{\phi}\sim p(\mathbf{H}_n^{\phi}\mid\boldsymbol{\Lambda}_n^{\phi})=\prod_{i=1}^{I}\prod_{t_s=1}^{T_s}\lambda_{n,i,t_s}^{\phi}\exp\left(-\lambda_{n,i,t_s}^{\phi}\mathbf{H}_{n,i,t_s}^{\phi}\right)$$

The advantages of using v-SNMF2D have already been described in Chapter 3. It is worth pointing out that *each individual element* in $\mathbf{H}_n^{\phi}$ is constrained to a exponential distribution with independent decay parameter λ_{n,i,t_s}^{ϕ}. In (4.10), $\mathbf{D}_{n,i}^{\tau}$ is the i^{th} column of $\mathbf{D}_n^{\tau}$, $\mathbf{H}_{n,i}^{\phi}$ is the i^{th} row of $\mathbf{H}_n^{\phi}$. In terms of interpretation, $\mathbf{D}_{n,i}^{\tau}$ represents the spectral

basis of the n^{th} IMF of the i^{th} source in the spectrogram domain and $\mathbf{H}_{n,i}^{\phi}$ represents the temporal sparse code for each spectral basis element. In the proposed algorithm, the two matrices to separate are $\left|\mathbf{C}_{n(f,t_s)}^{x_1}\right|^2$ and $\left|\mathbf{C}_{n(f,t_s)}^{x_2}\right|^2$ in the synthesis equation. This estimation corresponds to the case of $i=\{1,2\}$ in the analysis equation.

4.2.2 Estimation of sub-sources

The n^{th} order sub-sources $\left|\mathbf{C}_{n(f,t_s)}^{x_1}\right|^2$ and $\left|\mathbf{C}_{n(f,t_s)}^{x_2}\right|^2$ are estimated as $\left|\widetilde{\mathbf{C}}_{n(f,t_s)}^{x_1}\right|^2=\sum_{\tau}\sum_{\phi}\mathbf{D}_{n,1}^{\tau}\overset{\downarrow\phi}{\mathbf{H}}_{n,1}^{\phi}\overset{\rightarrow\tau}{}$ and $\left|\widetilde{\mathbf{C}}_{n(f,t_s)}^{x_2}\right|^2=\sum_{\tau}\sum_{\phi}\mathbf{D}_{n,2}^{\tau}\overset{\downarrow\phi}{\mathbf{H}}_{n,2}^{\phi}\overset{\rightarrow\tau}{}$. In the default setting, $\mathbf{D}_{n,i}^{\tau}$ is the i^{th} column of $\mathbf{D}_n^{\tau}$ that corresponds to the i^{th} row of $\mathbf{H}_{n,i}^{\phi}$ where $i=\{1,2\}$ for the case of two sources. If more components are considered in the v-SNMF2D e.g. $\mathbf{D}_n^{\tau}=\left[\mathbf{d}_{n,1}^{\tau},\ldots\mathbf{d}_{n,I_s}^{\tau}\right]\ \forall I_s>2$, this necessitates an efficient clustering method to group the column vectors $\mathbf{d}_{n,i_s}^{x}$ to their respective sources. The details of the clustering methods will be presented in Section 4.3. Once $\left|\mathbf{C}_{n(f,t_s)}^{x_1}\right|^2$ and $\left|\mathbf{C}_{n(f,t_s)}^{x_2}\right|^2$ are estimated, the time-domain sub-sources $\tilde{\mathbf{c}}_n^{x_i}$ can be reconstructed as follows:

$$\begin{aligned}\tilde{\mathbf{c}}_n^{x_1}&=\text{Resynthesize}(\mathbf{Mask}_n^{x_1}\bullet\mathbf{C}_{n(t,f)}^{y})\\ \tilde{\mathbf{c}}_n^{x_2}&=\text{Resynthesize}(\mathbf{Mask}_n^{x_2}\bullet\mathbf{C}_{n(t,f)}^{y})\end{aligned}\tag{4.11}$$

where ‘Resynthesize’ denotes the inverse mapping of the log-frequency axis to the original frequency axis and followed by the inverse STFT back to the time domain [92]. The mask signals are determined element wise by:

$$\mathbf{Mask}_{n,f,t_s}^{x_i}=\begin{cases}1\ ,\ \text{if}\ \left|\left[\tilde{\mathbf{C}}_{n(f,t_s)}^{x_i}\right]_{f,t_s}\right|^2>\left|\left[\tilde{\mathbf{C}}_{n(f,t_s)}^{x_j}\right]_{f,t_s}\right|^2\\ 0\ ,\ \text{otherwise.}\end{cases}\tag{4.12}$$

The estimated sub-sources in (4.11) are subsequently clustered into groups according to the number of sources. The Kullback-Leibler divergence (KLd) based *k*-means clustering algorithm [40] is used for grouping the subsets of the sub-sources. The sub-sources are firstly represented as vectors which are then normalized to unit length and transformed into their corresponding probability mass function. They are then grouped into κ clusters according to the entropy contained by individual vectors. In this paper, the symmetric KLd is used to measure the relative entropy between two probability mass function $p_1(\varepsilon)$ and $p_2(\varepsilon)$ over a random variable Ω:

$$\mathrm{KLd}\left(p_1, p_2\right) = \frac{1}{2}\left[\sum_{\varepsilon\in\Omega} p_1(\varepsilon)\log\frac{p_1(\varepsilon)}{p_2(\varepsilon)} + \sum_{\varepsilon\in\Omega} p_2(\varepsilon)\log\frac{p_2(\varepsilon)}{p_1(\varepsilon)}\right] \tag{4.13}$$

After convergence, all sub-sources will be grouped into their respective clusters which are given as $\hat{\mathbf{C}}_{imf}^{x_1} = \left\{\tilde{\mathbf{c}}_1^{x_1}, \tilde{\mathbf{c}}_2^{x_1}, \ldots, \tilde{\mathbf{c}}_{N_1}^{x_1}\right\}$ and $\hat{\mathbf{C}}_{imf}^{x_2} = \left\{\tilde{\mathbf{c}}_1^{x_2}, \tilde{\mathbf{c}}_2^{x_2}, \ldots, \tilde{\mathbf{c}}_{N_2}^{x_2}\right\}$. The estimated time-domain signal of the i^{th} source is then obtained by summing up the sub-sources from each cluster as:

$$\hat{\mathbf{x}}_1 = \hat{\mathbf{C}}_{imf}^{x_1}\mathbf{1}_{N1} \qquad \text{and} \qquad \hat{\mathbf{x}}_2 = \hat{\mathbf{C}}_{imf}^{x_2}\mathbf{1}_{N2} \tag{4.14}$$

The core procedure of the proposed method is summarised in Figure 4.3.

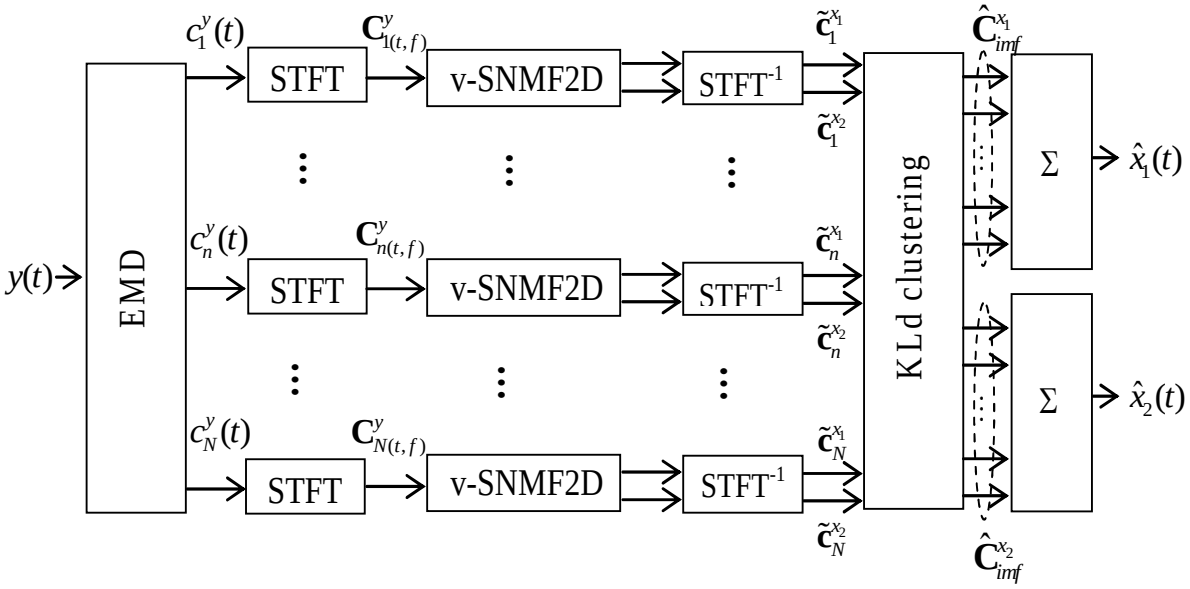


Figure 4.3: Core procedure of the proposed method.

4.3 Results and Analysis

The proposed monaural source separation method is tested by separating audio sources. Several experimental simulations under different conditions have been designed to investigate the efficacy of the proposed method. To generate mixtures, 40 sentences of the target speakers (20 male and 20 female sentences from 8 male and 8 female subjects) are selected from the TIMIT speech database and 20 music signals including 10 Jazz and 10 piano signals are selected from the RWC [100] database. Three types of mixture have been generated: (i) Jazz mixed with piano, (ii) speech mixed with music and (iii) speech mixed with speech. 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. All mixed signals are sampled at 16 kHz sampling rate and the audio mixture is divided into blocks of length 0.65s. Smaller-size

blocks perform better when the signal spectra are frequently changing. The TF representation is computed by normalizing the time-domain signal to unit power and computing the STFT using 1024 point Hamming window FFT with 50% overlap. The frequency axis of the obtained spectrogram is then logarithmically scaled and grouped into 175 frequency bins in the range of 50Hz to 8kHz with 24 bins per octave. This corresponds to twice the resolution of the equal tempered musical scale. For the v-SNMF2D parameters, the convolutive components in time and frequency are selected to be $\tau = \{0,\ldots,4\}$ and $\phi = \{0,\ldots,4\}$, respectively. The distortion measure between the original and estimated source is computed by using the improvement of signal-to-noise ratio (ISNR) [57] which is defined as:

$$ISNR_i = 10\log_{10}\frac{\sum_t |x_i(t)|^2}{\sum_t |x_i(t)-\hat{x}_i(t)|^2} - 10\log_{10}\frac{\sum_t |x_i(t)|^2}{\sum_t |y(t)-x_i(t)|^2} \quad (4.15)$$

where $\hat{x}_i(t)$ denotes the estimated i^{th} sources. The ISNR is used as the quantitative performance measure for separation, and the average ISNR will be tabulated in the evaluation graphs. The ISNR represents the degree of suppression of the interfering signals to improve the quality of the target one. The higher value of ISNR indicates better separation performance.

4.3.1 Effects on audio mixtures separation *with/without* EMD preprocess

In this section, we first investigate the performance of the proposed method *without* using the EMD preprocessing for separating audio mixtures. This is motivated by the fact that in the IMF subband domain, the spectral and temporal patterns of each IMF are

simpler and sparser than that of the mixed signal. Therefore, the spectral and temporal patterns of the dominating source and the less dominating one can be separated by using the matrix factorization methods (i.e. SNMF2D or v-SNMF2D). In addition, any error resulted in the IMF subband during the source separation can be alleviated at the source reconstruction stage. Thus, it is hypothesized that *with* the EMD preprocessing, the audio source separation will be significantly enhanced. Figure 4.4 and 4.5 shows the performance of our proposed method *without* and *with* the EMD preprocessing, respectively, under various audio mixtures.

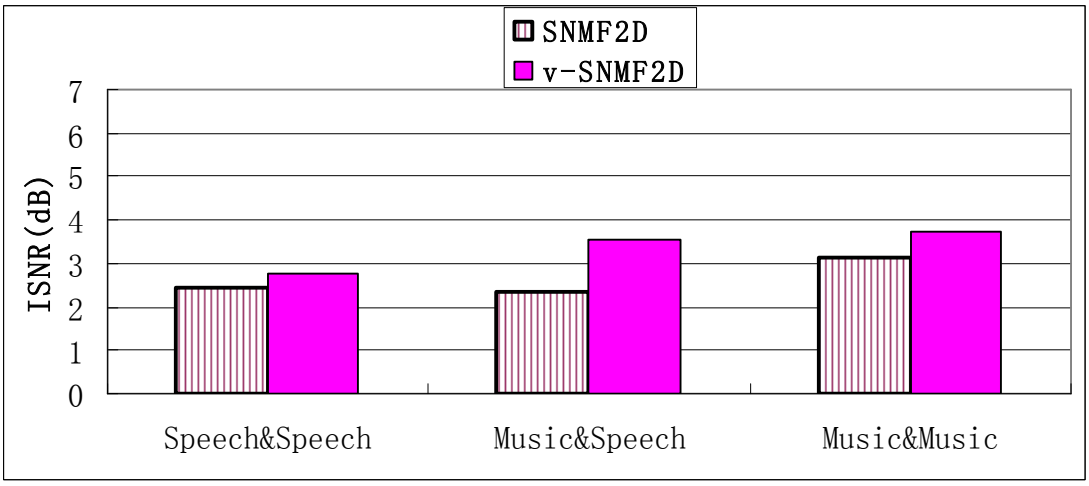


Figure 4.4: Overall separation results of different mixtures *without* EMD preprocess.

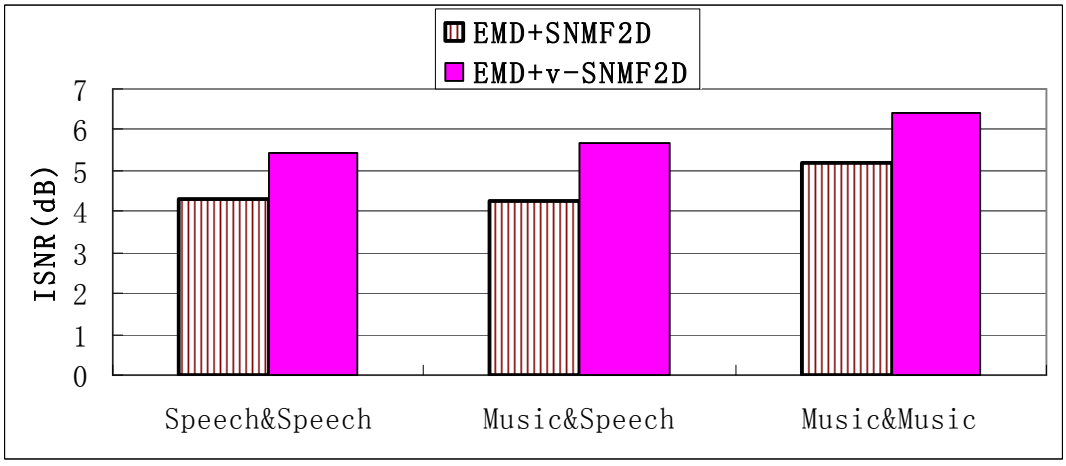


Figure 4.5: Overall separation results of different types of mixtures *with* EMD preprocess.

Figure 4.4 shows that *without* the EMD preprocessing, the ISNR is degraded substantially since the mixing ambiguity has been highly affected by the level of spectral overlap between $|\mathbf{X}_1|^2$ and $|\mathbf{X}_2|^2$. This is evidenced in Figure 4.6 which illustrates the mixture of original male and female speeches (top panels), the single channel mixed signal (middle panel), and the separated speeches (bottom panels) using the v-SNMF2D *without* the EMD preprocessing. The ISNR for the separated speeches, on average, is calculated to be 2.7dB per source. The ambiguity between the two speeches is highlighted in the red box marked area. Figure 4.10 (D)-(E) further illustrate this observation on the TF plane by means of another mixture of male speech and Jazz music. By visual inspection, a considerable level of spectral overlap has not been correctly separated. On the other hand, Figure 4.5 shows a large improvement gain in ISNR by incorporating the EMD preprocessing. An average improvement of 2.5dB per source has been obtained across all the different type of mixtures by using the v-SNMF2D *with* EMD preprocessing as compared to using the v-SNMF2D alone. Similarly, an average improvement of 2dB per source is obtained for the SNMF2D *with* EMD preprocessing as compared to using the SNMF2D alone.

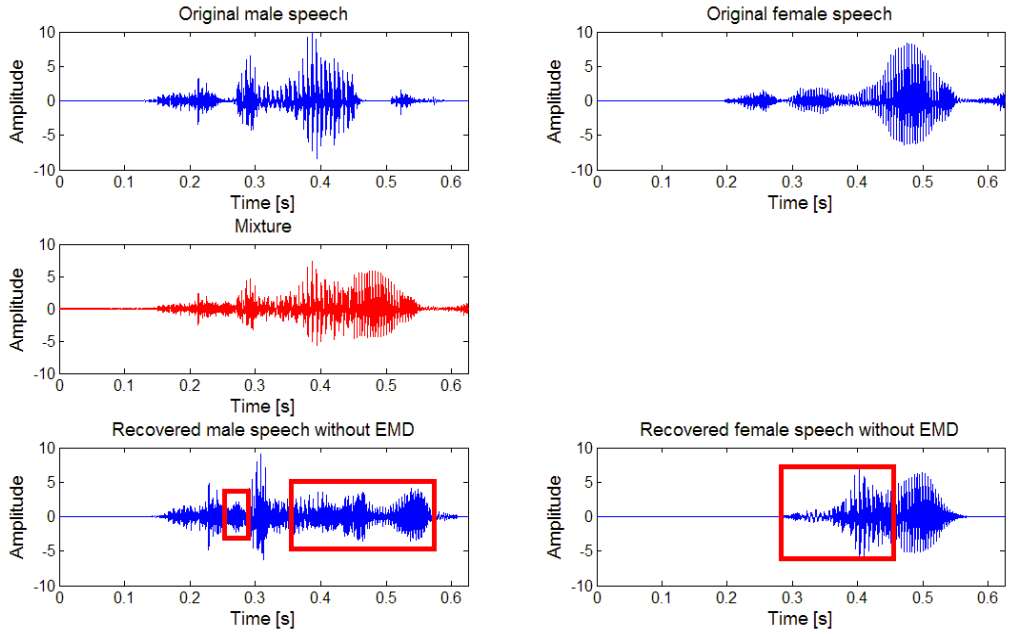
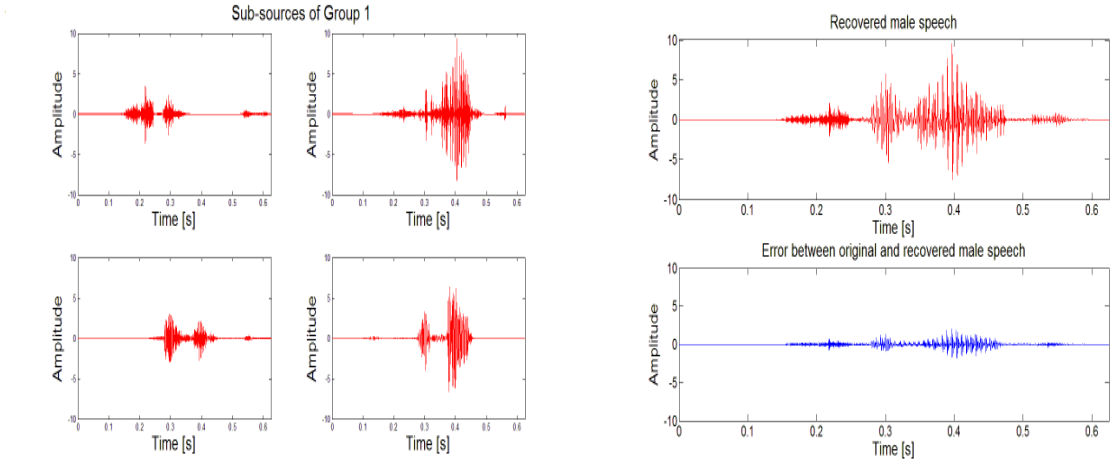


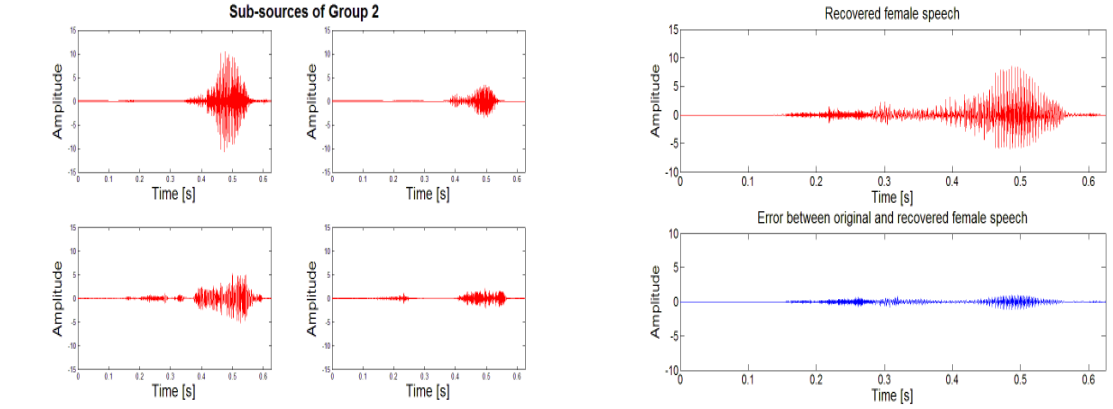
Figure 4.6: Separation results *without* applying EMD preprocess.

In the following, the results of the v-SNMF2D *with* EMD preprocessing are shown. Figures 4.7 and 4.8 show the time-domain separation results. In both figures, subplots(a) show the estimated sub-sources by exploiting the hybrid EMD and v-SNMF2D while subplots(b) show the reconstructed speech signals and the error between the original and the reconstructed signals based on the four estimated sub-sources (e.g. $\tilde{\mathbf{c}}_n^{x_i}$). The mean square error (MSE) between the original and the reconstructed speech is 0.34 and 0.32 for male speech and female speech, respectively. It is also found that as the number of estimated sub-sources increases (e.g. 6), the error becomes progressively smaller (MSE = 0.31 and 0.28 for male and female speeches, respectively).



(A) (B)

Figure 4.7: (A) Estimated sub-sources for male speech. (B) Reconstructed male speech and error.



(A) (B)

Figure 4.8: (A) Estimated sub-sources for female speech. (B) Reconstructed female speech and error.

4.3.2 Impacts of sparsity selection

In this section, the impact of sparsity selection is investigated. Choosing $f(\mathbf{H}_n)$ as well as each of the scalar regularization parameter $\underline{\boldsymbol{\lambda}}_n = \left\{\lambda_{n,i,t_s}^{\phi}\right\}$ will have significant impact on the matrix factorization and the final separation results. The proposed algorithm resolves this difficulty by using the EMD to reduce the mixing ambiguity in each sub-band. In addition, since the sparsity of each IMF on the TF plane varies across different IMF order, the sparseness constraint of $\mathbf{H}_n$ that impacts each IMF ought to be optimally controlled.

Table 4.2 shows the value of the sparse regularization parameter that corresponds to each IMFs of different mixtures. In Table 4.2, $\{J,P,M,F\}$ represent Jazz, piano music, male and female speech.

Table 4.2: Assignment of regularization parameter

Regularization parameter in vector form for each IMF	J & P	$(J\ or\ P)$ & $(M\ or\ F)$	M & $(M\ or\ F)$
$\underline{\boldsymbol{\lambda}}_1$	0.1	5	5
$\underline{\boldsymbol{\lambda}}_2$	0.05	5	5
$\underline{\boldsymbol{\lambda}}_3$	0	1	5
$\underline{\boldsymbol{\lambda}}_4$	0	1	5
$\underline{\boldsymbol{\lambda}}_5$	0	1	1
$\underline{\boldsymbol{\lambda}}_6$	0	0	1
$\underline{\boldsymbol{\lambda}}_7$	0	0	0

For mixture of piano and speech, the regularization parameters can be set similarly to the ones used for jazz and speech mixture. Table 4.2 shows that as the IMF order increases, lower values can be assigned to $\underline{\boldsymbol{\lambda}}_n$ for each type of mixture. This is evidenced from the fact that the EMD can automatically range the bandwidths so that in each sub-band only one source with the most energy is retained. This allows the selection of the sparseness in each $\mathbf{H}_n$. It is also found that different types of audio mixtures require different selection of the sparseness regularization. Using the mixture of music and speech as an example, it is well documented that music pitches jumped discretely while speech pitches do not so that $\underline{\boldsymbol{\lambda}}_n$ can be set to zero from the 6th IMF onwards since these correspond to the lower

frequency bands and are dominated with most energy from the speech components. In the lower frequency bands, very little mixing exists between the music and speech signal so that imposing sparseness will lead to over-sparse code and eventually render less efficiency in estimating the speech signal components. On the contrary, it is difficult to set $\underline{\boldsymbol{\lambda}}_n$ equal to zero for mixture of male and female speeches since the fundamental pitches of both signals are too similar for the SNMF2D to separate. It should be noted that the above regularization parameters are set empirically and by no means, are the optimal values. The selection of $\lambda^{\phi}_{n,i,t_s} = \lambda$ for all i, t_s, ϕ of n^{th} individual IMF is based on Monte-Carlo simulation over many different realizations of audio mixture. The selection proceeds as follows: Firstly, a threshold is set for a target ISNR e.g. ISNR = 4dB. Secondly, the value of λ for each IMF that renders signal separation with ISNR above this target threshold will be accepted while the ones that do not will be discarded. Thirdly, this process is repeated for different sources of the same type of mixture. Finally, the λ for each IMF is selected by averaging over all realizations. In the following figure, the results are obtained using this Monte-Carlo simulation.

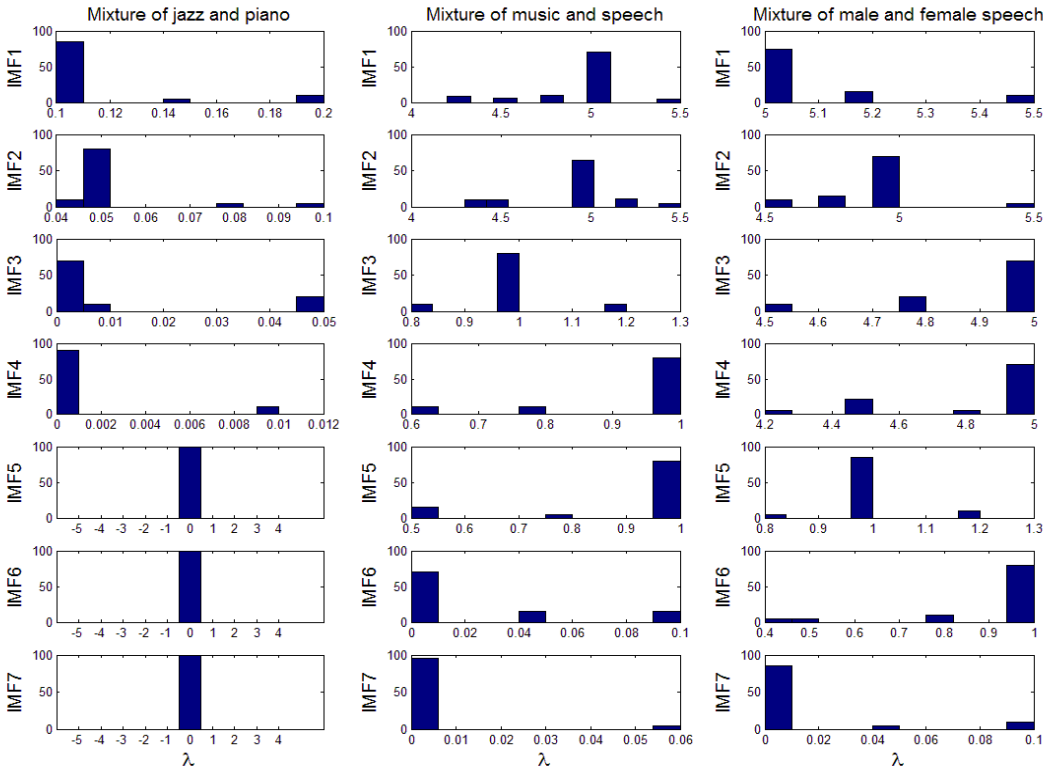


Figure 4.9: Histogram of regularization parameter in SNMF2D for each IMF.

Figure 4.9 shows the histogram of the regularization parameter for each IMF using the Monte-Carlo simulation. Each column in the above figure represents the histogram of selective λ over all realizations for IMF order from 1^{st} to 7^{th}. Based on the above histogram, the selective λ assigned to each IMF is thus obtained in Table 4.2. However, the Monte-Carlo approach to obtain these regularization parameters is not as optimal as our proposed method in terms of signal separation.

The proposed method resolves this issue by adaptively updating these sparse regularization parameters while the spectral bases and the temporal codes are still being learned. To study the effects of sparsity regularization on the separation results, Figure 4.10 shows the spectrograms computed using the EMD SNMF2D and EMD v-SNMF2D.

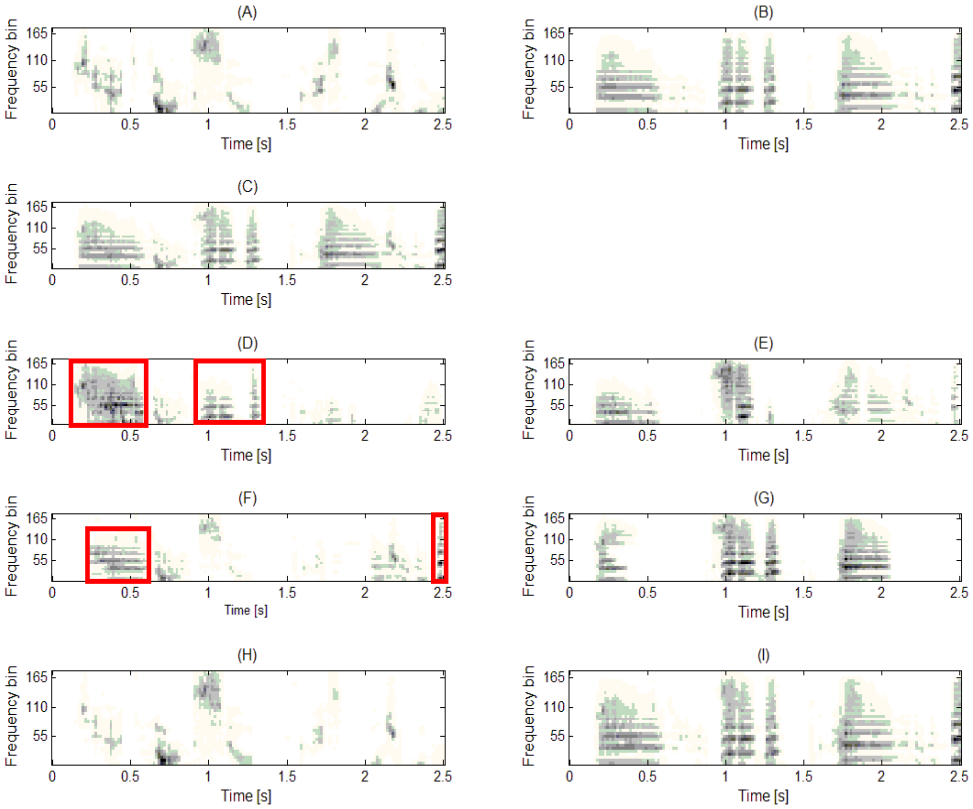


Figure 4.10: (A)-(B) denote the original spectrogram of male speech and Jazz music respectively. (C) denotes the spectrogram of the mixture. (D)-(E), (F)-(G), and (H)-(I) denote the reconstructed spectrogram of male speech and Jazz music by directly using the SNMF2D method (*without* EMD), EMD SNMF2D method, and EMD v-SNMF2D method, respectively.

In Figure 4.10, it is noted that errors still present in the estimated male speech spectrogram by using the SNMF2D and the EMD SNMF2D methods. The components in the red box marked region in (D) and (F) definitely belong to the Jazz music but have been attributed to the male speech instead. As a result, the estimated male speech contains interference from the Jazz music whereas the estimated Jazz music loses some of its information. Because of the 'under- or over-sparse' resolution, the estimates are only coarse by using the EMD *with* SNMF2D. Consequently, this leads to ambiguity in the TF region which reduces the separation efficiency. On the other hand, the performance has been significantly improved when the decomposition of spectral bases and temporal codes

are performed using the variable sparse regularization. It is noted that the level of mixing ambiguity has been progressively reduced from using the SNMF2D *without* EMD preprocessing to the proposed v-SNMF2D *with* EMD preprocessing.

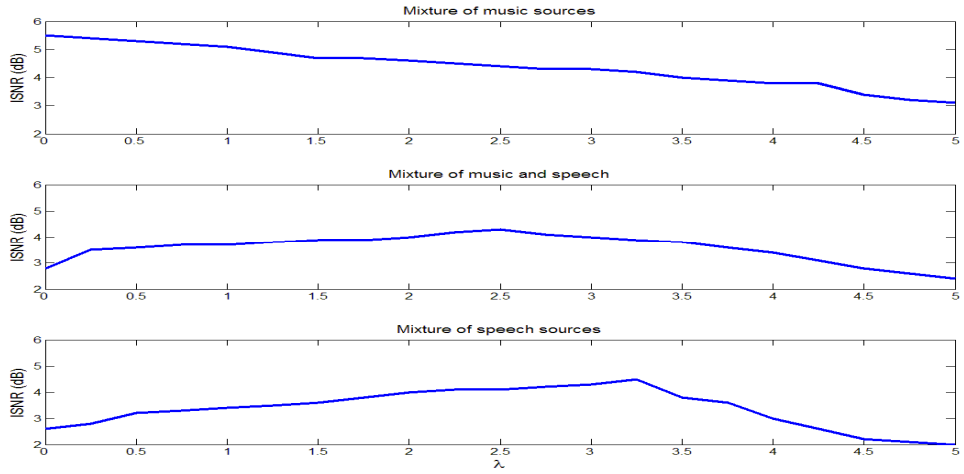


Figure 4.11: Separation results of EMD-SNMF2D by using different uniform regularization.

Figure 4.11 shows the impact of sparsity regularization on the separation results in terms of the ISNR under different uniform regularization. In this implementation, the uniform regularization for all IMF is chosen as i.e. $\underline{\boldsymbol{\lambda}}_1 = \underline{\boldsymbol{\lambda}}_2 = \cdots = \underline{\boldsymbol{\lambda}}_7 = c$, $c = 0,0.5,\ldots,5$. Figure 4.12 summarises the average separation results of the EMD-NMF2D, EMD-SNMF2D, selective uniform regularization EMD-SNMF2D based on Table 4.2 and EMD v-SNMF2D methods.

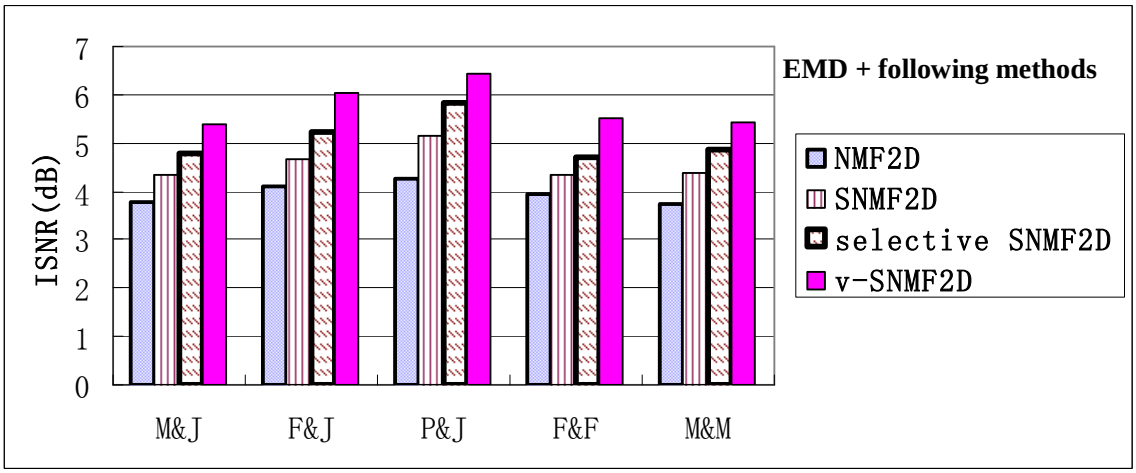


Figure 4.12: Separation results of EMD-based SNMF2D using regularization schemes.

For comparison purpose, the average performance improvement of the proposed method has been summarised based on Figure 4.12 as follows: (i) for mixture of music signals, the average improvement is 1.4dB per source, (ii) for mixture of speech and music signal, the average improvement is 1.6dB per source, and (iii) for mixture of speech signals, the average improvement is 1.7dB per source. The above results clearly indicate that the best performance is achieved by the EMD preprocessing *with* v-SNMF2D.

4.3.3 Comparison with other SCSS methods

4.3.3.1 Underdetermined-based ICA SCSS method

In the underdetermined-ICA time model-based SCSS method [47], the key point is to exploit the prior knowledge of the sources such as the basis functions to generate the sparse codes. In this work, these basis functions are obtained in two stages: (i) *Training stage*: the basis functions are obtained by performing ICA on each concatenated sources. In our experiments, a set of 64 basis functions is derived for each type of source[3]. For example, to generate the ICA speech basis functions, 10 male and 10 female speeches from TIMIT speech database are used. Similarly, to generate the ICA music basis functions, 5 Jazz and 5 piano signals from RWC database are used. These training data exclude the target sources which have been exclusively used to generate the mixture signals. (ii) *Adaptation stage*: the obtained ICA basis functions from the training stage are further adapted based on the current estimated sources during the separation process. At this stage, both the estimated sources and the ICA basis functions are jointly optimized by

[3] Here the types of source signals are the male speech, female speech, jazz and piano music.

maximizing the log-likelihood of the current mixture signal until it converges to the steady-state solution.

4.3.3.2 Hilbert subspace decomposition SCBSS method

The method of [57] performs source separation without training information by decomposing the Hilbert spectrum of the mixed signal into independent source subspaces. Once a set of independent basis vectors is obtained by means of PCA and ICA, the KLd based *k*-means clustering algorithm is utilized for grouping purpose and the Hilbert spectrum of individual source is constructed by each group subset. The time-domain estimated sources are calculated from the Hilbert spectrum of each of the extracted signals.

4.3.3.3 Comparison Results

Figure 4.13 shows the separated male and female speeches based on the above two SCSS methods. Figure 4.14 shows the comparison results between the proposed method and the above two SCSS methods in terms of the ISNR. In the case of the underdetermined-ICA time model-based SCSS method, it is noted that the recovered sources have not been clearly separated and the mixing ambiguity region is still large when compared with the original speeches in Figure 3.13 (top panels). The proposed method has yielded considerable improvement over the underdetermined-ICA time model-based SCSS method and this is summarised as follows: (i) for mixture of music signals, the proposed method results in an average improvement of 2.3dB per source, (ii) for mixture of speech and music signal, an average improvement of 2.9dB per source, and (iii) for mixture of speech signals, an average improvement of 4.1dB per source. The performance of the

underdetermined-ICA time model-based SCSS method relies on the ICA-derived time domain basis functions. Figure 4.14 indicates that high level performance is achieved only when the basis functions of each source are sufficiently distinct. The result becomes considerably less robust in separating mixture where the original sources are of the same type e.g. mixture of speeches [101]. Speech basis functions learned from the ICA exhibit waveforms that resemble Gabor wavelets; however, the set of basis functions from the male speech has high degree of correlation with that of the female speech. Therefore, these two sets of basis functions overlap significantly with each other. Hence, the underdetermined-ICA time model-based SCSS method is less efficient in resolving the mixing ambiguity in portions of the speech mixture where the basis functions for the male and female are very similar.

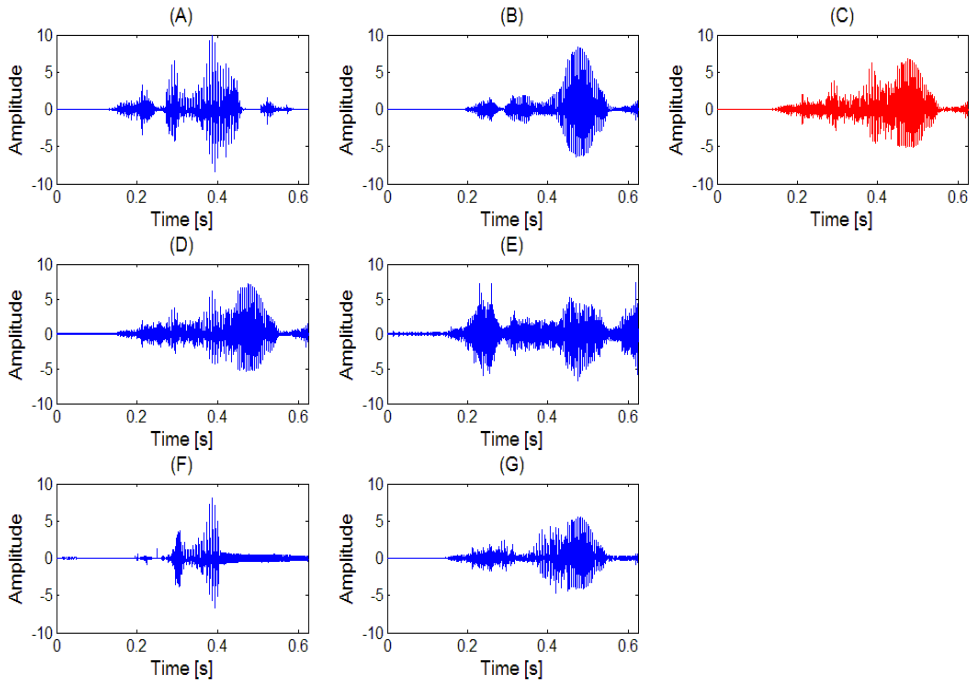


Figure 4.13: (A)-(C) denote the original male, female speeches and mixture, respectively. (D)-(E) denote the recovered male and female speeches by using the underdetermined-ICA SCBSS method. (F)-(G) denote the recovered male and female speeches by using the Hilbert SCBSS method.

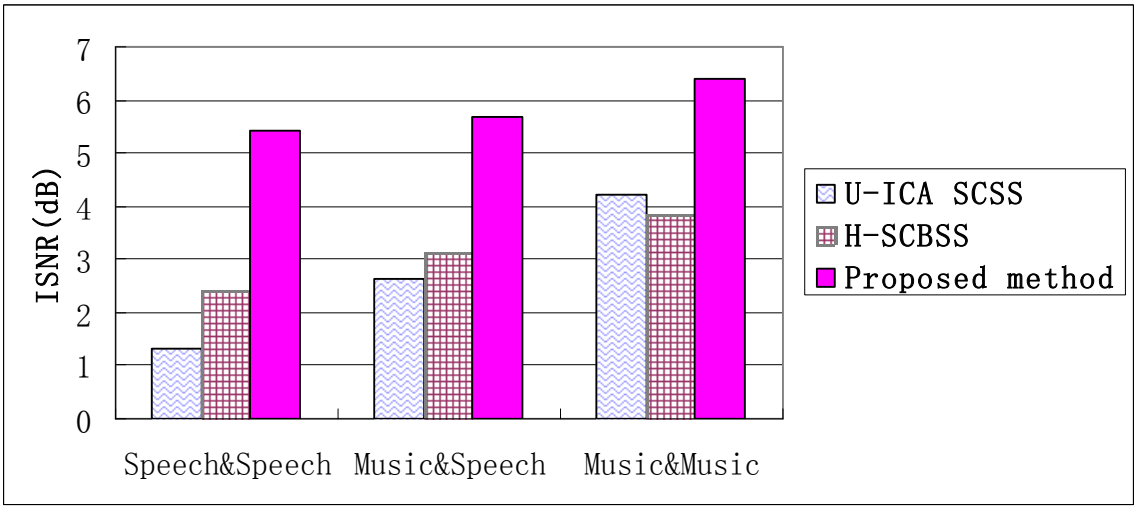


Figure 4.14: Overall results between the proposed method, underdetermined-ICA time model-based SCSS and Hilbert SCBSS methods.

In Figure 4.14, 'U-ICA SCSS' and 'H-SCBSS' denote the Underdetermined-ICA time model-based SCSS and Hilbert SCBSS methods, respectively. The decomposition obtained by the Hilbert SCBSS method shows that this technique leads to better separation results than the underdetermined-ICA time model-based SCSS method. However, it is noted that the separated speeches still contain high level of mixing ambiguity and therefore, it degrades the separation performance. This is evidenced in Figure 4.14 which shows the comparison of the proposed method with the Hilbert SCSS method: (i) for mixture of music signals, the average improvement is 2.4dB per source, (ii) for mixture of speech and music signal, the average improvement is 2.5dB per source, and (iii) for mixture of speech signals, the average improvement is 3.2dB per source. The performance of the Hilbert SCBSS method relies too heavily on the derived frequency independent basis vectors which are stationary over time. Therefore, good separation results can be obtained only if the basis vectors are statistical independent within the processing window. The distinctiveness of the corresponding amplitude weighting vectors is also highly dependent

on the independence of the basis vectors. Thus, if the frequency features are too similar, it becomes difficult to obtain the independent basis vectors by using the ICA. This explains the reason Figure 4.14 shows a relatively poorer performance when separating mixture that contains speech sources. Comparing with the Hilbert SCBSS method, the proposed v-SNMF2D yields an optimally sparse part-based decomposition that is unique under certain conditions e.g. sparse and nonnegative component, making it unnecessary to impose constraints in the form of statistical independence between the sources. Furthermore, the spectral bases $\mathbf{D}_n^{\tau}$ and sparse code $\mathbf{H}_n^{\phi}$ in the proposed method are derived separately at each individual IMF. Thus, these spectral bases and temporal codes are non-stationary over time leading to more robust separation results compared with the stationary basis vectors obtained from the Hilbert SCBSS method.

4.3.3.4 Comparison with NMF-based SCBSS methods

In this evaluation, the following NMF-based SCBSS methods are used for comparison:

- NMF with Temporal Continuity and Sparseness Criteria [37] (NMF-TCS) based SCBSS method as described in Chapter 3.
- Automatic Relevance Determination NMF (NMF-ARD) [97] based SCBSS method as described in Chapter 3.

Currently, there are no reliable NMF methods for automatic estimation of the number of components (e.g. the basis vectors in $\mathbf{D}$) and normally, this has to be set manually. As discussed in Section 4.2, each IMF is separated into a number of components that corresponds exactly to the number of sources. However, in this implementation, more

components than the number of sources are used for evaluating the efficiency of the proposed method. In order to obtain the baseline comparison of each method, all NMF algorithms are tested by factorizing the mixture signal into $I_s = 2,4,\ldots,10$ components. In the case of NMF-ARD, the threshold has been modified such that it accepts all the initialized components. Since more than two components are used and the tested methods are blind, there is no information to tell which component belongs to which source. Thus, the clustering method proposed in [57] is utilized where the original sources are used as reference to create component clusters for each source. However, a large number of components i.e. $I_s > 10$ may not necessarily produce better results since more sub-sources need to be classified. If the recovered sub-sources are incorrectly clustered, then these sub-sources will become interference to the supposedly correct estimated source. We have carried out additional analysis to compare the KLd-based *k*-means clustering method [57] with the supervised clustering method in [37]. The finding shows that if the sub-sources are too sparse, both methods will introduce errors during the clustering process. For example, beyond the 7^{th} stage decomposition by the EMD, the TF sub-sources are too sparse to assign them to the correct sources. If wrongly clustered, this particular sub-source will become interference to the intended source. To mitigate this situation, a power threshold is set as described in Section 4.2 to judge whether the IMF is of acceptable quality. The findings have shown that the results based on KLd *k*-means clustering method are identical to the supervised clustering method in [37] except in special circumstances where the sub-sources are overly too sparse in the TF domain. Figure 4.15 shows the ISNR performance between the proposed method and the NMF-TCS, NMF-ARD methods under

different mixture types, and the increasing number of components from $I_s = 2,4,6,8,10$.

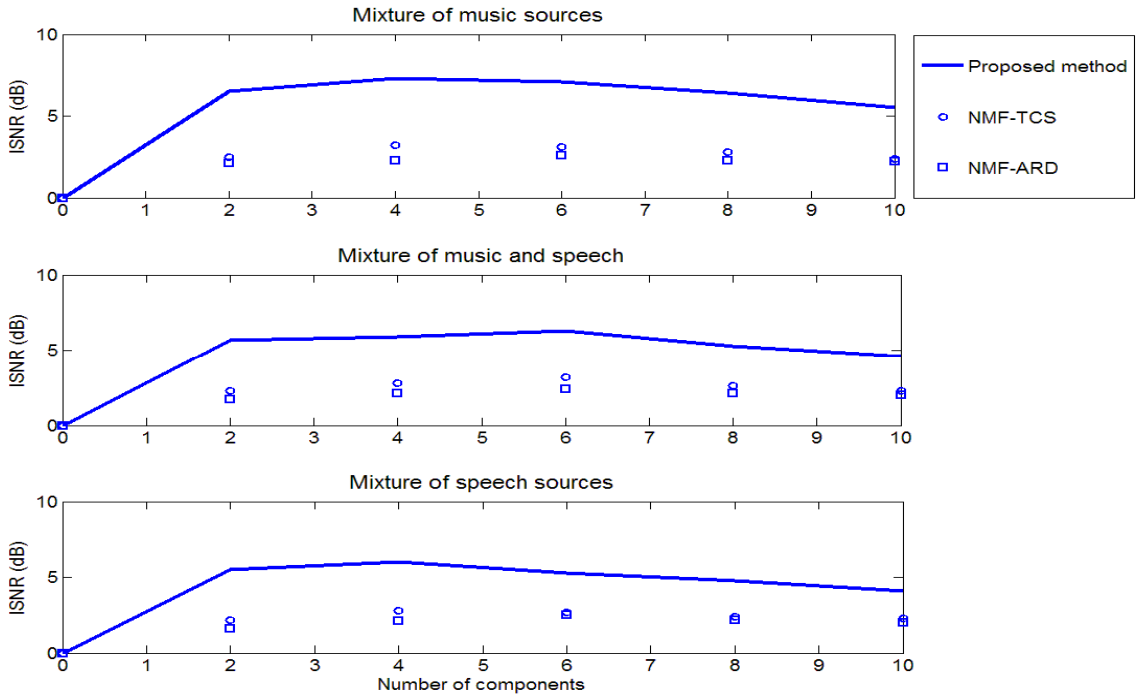


Figure 4.15: Average ISNR using different number of components.

In Figure 4.15, the ISNR improvement of the proposed method compared with NMF-TCS and NMF-ARD can be summarised as follows: (i) for mixture of music signals, the average improvement is 4.3dB per source, (ii) for mixture of speech and music signal, the average improvement is 3.1dB per source, and (iii) for mixture of speech signals, the average improvement is 3.3dB per source. Analysing the separation results, NMF-ARD performs with poorer results whereas the separation performance by NMF-TCS is slightly better than NMF-ARD. The common feature among these two methods is that they do not incorporate the preprocessing step that benefits the nonnegative matrix factorization. This renders the performance less efficient especially in terms of separating mixture that contains speech sources. The result indicates that *without* the EMD preprocessing, it

becomes difficult to obtain the unique spectral basis **D** especially when the spectral overlapping between the sources in TF domain is large since each column in **D** may contain the combination spectral information of both sources. In this case, by directly using NMF methods, the separation of sources is no longer efficient.

4.4 Summary

This chapter has presented a novel framework of amalgamating EMD *with* v-SNMF2D for single channel source separation. In this chapter, it is shown that the IMFs have several desirable properties unique to single channel source separation problem: (*i*) the degree of mixing in each IMF is less ambiguous than the mixed signal, (*ii*) the IMFs has simpler and sparser spectral and temporal patterns which allows the proposed v-SNMF2D algorithm to efficiently track them, and (*iii*) the IMFs serve as the orthogonal temporal bases for signal separation; hence errors resulted from any IMF will be averaged over all the IMFs leading to smaller errors at the signal reconstruction stage. In the proposed v-SNMF2D algorithm, the sparsity parameters are individually optimized and adaptively tuned using the variational Bayesian approach to yield the optimal sparse codes. The proposed framework enjoys at least two significant advantages: Firstly, it avoids the strong constraints of separating blind source among all types of audio mixture without training knowledge. Secondly, the v-SNMF2D algorithm gives a robust sparse decomposition and under non-negativity condition, the decomposition is unique making it unnecessary to impose constraints in the form of statistical independence of the sources.

CHAPTER 5

SINGLE CHANNEL BLIND SOURCE SEPARATION USING GAMMATONE FILTERBANK AND ITAKURA-SAITO MATRIX FACTORIZATION

In this chapter, a novel framework to solving SCBSS based on the cochleagram TF representation and a family of IS divergence based novel two-dimensional nonnegative matrix factorization algorithms are proposed. The proposed solution separates audio sources from a single channel without relying on training information about the original sources. The uniqueness of the proposed work can be summarised as follows:

(i) Using the gammatone filterbank to construct audio signal TF representation. 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.

(ii) The separability theory has been derived in the TF domain and a quantitative performance measure has been developed to evaluate how separable the sources in the monaural mixed signal. In particular, the ideal condition has been identified when the sources are perfectly separable. We also proposed a separation framework using the gammatone filterbank. The latter produces a non-uniform TF domain termed as the cochleagram whereby each TF unit has different resolution unlike the classical spectrogram which deals only with uniform resolution. Towards this end, it is shown

that the mixed signal is significantly more separable in the cochleagram than the classic spectrogram and the log-frequency spectrogram (constant-Q transform).

(iii) A family of IS divergence based novel two-dimensional nonnegative matrix factorization algorithms has been developed to extract the spectral and temporal features of the sources. The proposed factorizations are scale invariant whereby the lower energy components in the cochleagram can be treated with equal importance as the higher energy components. Within the context of SCBSS, this property is highly desirable as it enables the spectral-temporal features of the sources that are usually characterized by large dynamic range of energy to be estimated with significantly higher accuracy. This is to be contrasted with the matrix factorization based on LS distance and KL divergence where both methods favor the high-energy components but neglect the low-energy components.

This chapter is organized as follows: Section 5.1 introduces the different TF matrix representations and the separability theory is developed. In Section 5.2, the family of IS divergence based NMF2D and regularised NMF2D algorithms are derived. The proposed source separation framework is fully developed. Experimental results and a series of performance comparison with other matrix factorization methods are presented in Section 5.3. Finally, Section 5.4 concludes the work of this chapter.

5.1 Time-Frequency Representation

The section sets out to investigate effective TF representations to enhance the separability of SCSS. It is generally accepted that TF analysis is the core technique for characterizing and manipulating audio signals. In the task of audio source separation, one critical decision is to choose a suitable TF domain to represent the time-varying contents of the signals. In this section, we concentrate on the analysis of three widely used TF representations classic spectrogram, log-frequency spectrogram and cochleagram. In order to analyse the impacts of these TF representations, the separability analysis of source separation in the TF domain has been developed.

5.1.1 Classic spectrogram

The signal $y(t)$ is first multiplied by a finite length window function, and the Fourier Transform is taken as the window is slid along the time axis, resulting in a two-dimensional power representation of the signal, namely:

$$\left|Y\left(f,\tau_w\right)\right|^2 = \left|\int_{-\infty}^{\infty} y(t)\, win(t-\tau_w)\, e^{-j2\pi ft}\, dt\right|^2 \tag{5.1}$$

where $win(t)$ is the window function and τ is the time-shift. The classic spectrogram as computed by the STFT is equivalent to a bank of K_{STFT} filters equally spaced at the frequencies:

$$f_{k_{\text{stft}}}^{STFT} = k_{\text{stft}} \frac{f_s}{K_{\text{STFT}}} \quad \text{for} \quad k_{\text{stft}} = 1,\ldots,K_{\text{STFT}} \tag{5.2}$$

with constant bandwidth:

$$v^{\text{STFT}} = B_w \frac{f_s}{K_{\text{STFT}}} \tag{5.3}$$

where B_w is the main-lobe width in bins, a parameter given for each type of impulse response. (e.g. For Hanning windows, the main-lobe width is $B_w = 4$ bins) and f_s denotes the sampling frequency. The details on the classic spectrogram analogy can be found in [106].

5.1.2 Log-frequency spectrogram (constant-Q transform)

The classic spectrogram decomposes signals to components of linearly spaced frequencies. However, in western music the typically used frequencies are geometrically spaced. Thus, getting an acceptable low-frequency resolution is absolutely necessary, while a resolution that is geometrically related to the frequency is desirable, although not critical. The constant Q transform as introduced in [105], tries to solve both issues. If f_{fund} is the fundamental frequency of one note, then the center frequencies are geometrically spaced as:

$$f_{k_Q}^{Q} = f_{\text{fund}} \cdot 2^{k_Q / K_Q} \tag{5.4}$$

where K_Q denotes the maximum number of filters per octave. In addition, the bandwidth of the k_Q th filter is:

$$v_{k_Q}^{Q} = f_{k_Q}^{Q} \left(2^{1/K_Q} - 1 \right) \tag{5.5}$$

Thus, the filters cover the whole frequency range without overlapping. This yields a constant Q ratio of frequency to resolution, which is expressed as:

$$Q^{\text{const}} = \frac{f_{k_Q}^{Q}}{v_{k_Q}^{Q}} = \left(2^{1/K_Q} - 1 \right)^{-1} \tag{5.6}$$

In general, the twelve-tone equal tempered scale which forms the basis of modern western music divides each octave into twelve half notes where the frequency ratio between each successive half note is equal. The fundamental frequency of the note which is k_Q halfnotes above can be expressed as $f_{k_Q}^{Q} = f_{\text{fund}} \cdot 2^{k_Q/24}$. Taking the logarithmic, it gives $\log f_{k_Q}^{Q} = \log f_{\text{fund}} + \frac{k_Q}{24}\log 2$. Thus, in a log-frequency representation the notes are linearly spaced. In the method, the frequency axis of the obtained spectrogram is logarithmically scaled and grouped into 175 frequency bins in the range of 50Hz to 8kHz (given $f_s = 16\text{kHz}$) with 24 bins per octave and the bandwidth follows the constant-Q rule [105].

5.1.3 Gammatone filterbank and Cochleagram

Gammatone filterbank was previously proposed in [107, 108] as a model to cochlear filtering which decomposes the time-domain input into the frequency domain. The impulse response of a gammatone filter centered at frequency f is given by:

$$g(f,t) = \begin{cases} t^{\ell-1} e^{-2\pi v t} \cos(2\pi f t) \ , \ t \geq 0 \\ 0 \ , \ else \end{cases} \tag{5.7}$$

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

$$x(c,t) = x(t) * g(f_c, t) \tag{5.8}$$

where '$*$' represents convolution. The response is shifted backwards by $(\ell-1)/(2\pi v)$ to compensate for the filter delay. The output of each filter channel is divided into time frame

with 50% overlap between consecutive frames [49]. The resulting outputs form the time-frequency spectra which are then constructed to form the cochleagram. This is supported by the physiological studies [110, 111] of auditory nerve tuning curves [112] and psychophysical studies of critical bandwidth [113]. Both studies have indicated that auditory filters are distributed in frequency according to their bandwidths, which increase quasi logarithmically with increasing center frequency. Thus, the bandwidth of each filter is set according to its equivalent rectangular bandwidth (ERB) which is a psychophysical measurement of the critical bandwidth in human subjects (as described in Glasberg and Moore [113]) as:

$$ERB(f) = 24.7(4.37f/1000 + 1) \tag{5.9}$$

More specifically, they define $b_c = 1.019ERB(f_c)$ where b_c determines the rate of decay of the impulse response, which is related to bandwidth. Additionally, the gains of the filters are adjusted according to the ISO standard for equal loudness contours [114] in order to model the pressure gains of the outer and middle ears. Thus, the use of the gammatone filter is consistent according to the neurobiological modeling perspective. Equation (5.7) provides a close approximation to the experimentally derived auditory nerve fiber impulse responses, as measured by [115] using a reverse-correlation technique. Furthermore, the fourth-order gammatone filter provides a good match to psychophysically derived "rounded-exponential" models of human auditory filter shape [116].

5.1.4 Difference between classic spectrogram, log-frequency spectrogram and cochleagram

The classical spectrogram as computed by the STFT has an equal-spaced bandwidth across all frequency channels. Since speech signals are characterized as highly non-stationary and non-periodic whereas music changes continuously; therefore, application of the Fourier transform will produce errors especially when complicated transient phenomena such as the mixing of speech and music occur in the analysed signal. Unlike the spectrogram, the log-frequency spectrogram possesses non-uniform TF resolution. However, it does not exactly match to the nonlinear resolution of the cochlear since their centre frequencies are distributed logarithmically along the frequency axis and all filters have constant-Q factor [105]. On the other hand, the gammatone filters used in the cochlear model (3) are approximately *logarithmically* spaced with constant-Q for frequencies from $f_s/10$ to $f_s/2$ and approximately *linearly* spaced for frequencies below $f_s/10$. Hence, this characteristic results in selective *non-uniform* resolution in the TF representation of the analysed audio signal. Figure 5.1 shows an example of frequency response for different types transform.

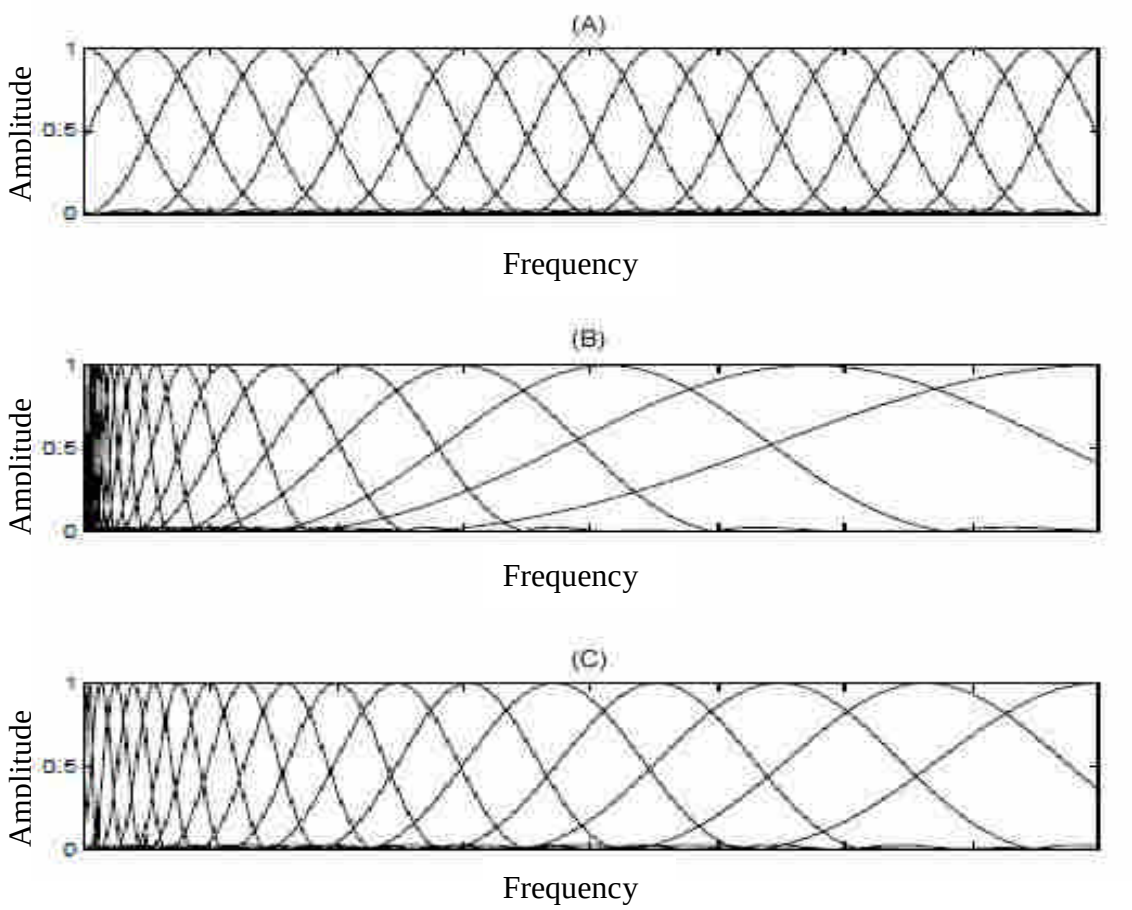


Figure 5.1: (A) Normalized frequency responses of 17-channel STFT filter bank. (B) Normalized frequency responses of 17-channel constant-Q filter bank. (C) Normalized frequency responses of 17-channel gammatone filter bank.

From Figure 5.1, it can be seen that the classic spectrogram which is based on the STFT yields a time-frequency representation with only uniform frequency and time resolutions. On the other hand, the log-frequency spectrogram based on constant-Q transform has non-uniform time-frequency resolution and the time-resolution trade-off is strongly biased towards improving frequency resolution in the low-frequency region. The cochleagram based on gammatone filter bank also has non-uniform time-frequency resolution while it is more balanced between the high and low frequency areas when compared to the constant-Q representation. In the next sub section, the comparison of separability based on above three TF representations is carried out and the cochleagram is found as the most

suitable TF tool when using NMF2D model for audio source separation.

5.1.5 Separability analysis

For separation, one generates a TF mask corresponding to each source and applies the generated mask to the mixture to obtain the estimated source TF representation. In particular, when the sources do not overlap in the TF domain, an optimum mask $Mask_i^{opt}(f,t_s)$ exists which allows one to extract the i^{th} original source from the mixture as:

$$X_i(f,t_s) = Mask_i^{opt}(f,t_s)Y(f,t_s) \tag{5.10}$$

where '*opt*' denotes optimum. Given any TF mask $Mask_i(f,t_s)$ such that $0 \le Mask_i(f,t_s) \le 1$ for all (f,t_s), we define the separability in the TF domain for target source $x_i(t)$ in the presence of the interfering sources $p_i(t) = \sum_{j=1, j\neq i}^{N} x_j(t)$:

$$S_{Mask_i}^{Y\to X_i,P_i} \triangleq \frac{\left\|Mask_i(f,t_s)X_i(f,t_s)\right\|_{Fro}^2}{\left\|X_i(f,t_s)\right\|_{Fro}^2} - \frac{\left\|Mask_i(f,t_s)P_i(f,t_s)\right\|_{Fro}^2}{\left\|X_i(f,t_s)\right\|_{Fro}^2} \tag{5.11}$$

where $X_i(f,t_s)$ and $P_i(f,t_s)$ is the TF representation of $x_i(t)$ and $p_i(t)$,respectively. In addition, the separability of the mixture with respect to all the N_s sources is defined as:

$$S_{Mask_1,\ldots,Mask_{N_s}}^{Y\to X_1,\ldots,X_{N_s}} = \frac{1}{N_s}\sum_{i=1}^{N_s} S_{Mask_i}^{Y\to X_i,P_i} \tag{5.12}$$

Eqn. (5.11) is equivalent to measuring the ability of extracting the i^{th} source $X_i(f,t_s)$ from the mixture $Y(f,t_s)$ given the TF mask $Mask_i(f,t_s)$. Eqn. (5.12) measures the ability of extracting all the N_s sources simultaneously from the mixture. To further study the separability, the following two criteria [51] are used: (i) Preserved signal ratio (PSR)

which determines how well the mask preserves the source of interest and (ii) Signal-to-interference ratio (SIR) which indicates how well the mask suppresses the interfering sources:

$$PSR_{M_i}^{X_i} \triangleq \frac{\left\| Mask_i(f,t_s)X_i(f,t_s) \right\|_{Fro}^2}{\left\| X_i(f,t_s) \right\|_{Fro}^2} \quad \text{and} \quad SIR_{M_i}^{X_i} \triangleq \frac{\left\| Mask_i(f,t_s)X_i(f,t_s) \right\|_{Fro}^2}{\left\| Mask_i(f,t_s)P_i(f,t_s) \right\|_{Fro}^2} \tag{5.13}$$

Using (5.13), (5.11) can be expressed as:

$$S_{Mask_i}^{Y \to X_i, P_i} = PSR_{Mask_i}^{X_i} - PSR_{Mask_i}^{X_i} \Big/ SIR_{Mask_i}^{X_i} \tag{5.14}$$

Analysing the terms in (5.14), namely:

$$\begin{aligned} PSR_{Mask_i}^{X_i} &:= \begin{cases} 1 & , \ if \ \operatorname{supp} Mask_i^{opt} = \operatorname{supp} Mask_i \\ <1 & , \ if \ \operatorname{supp} Mask_i^{opt} \subset \operatorname{supp} Mask_i \end{cases} \\ SIR_{Mask_i}^{X_i} &:= \begin{cases} \infty & , \ if \ \operatorname{supp}\left[Mask_i X_i \right] \cap \operatorname{supp} P_i = \varnothing \\ finite & , \ if \ \operatorname{supp}\left[Mask_i X_i \right] \cap \operatorname{supp} P_i \neq \varnothing \end{cases} \end{aligned} \tag{5.15}$$

where 'supp' denotes the support. When $S_{Mask_i}^{Y \to X_i, P_i} = 1$ (i.e. $PSR_{Mask_i}^{X_i} = 1$ and $SIR_{Mask_i}^{X_i} = \infty$), this indicates that the mixture $y(t)$ is separable with respect to the i^{th} source $x_i(t)$. In other words, $X_i(f,t_s)$ does not overlap with $P_i(f,t_s)$ and the TF mask $Mask_i(f,t_s)$ has perfectly separated the i^{th} source $X_i(f,t_s)$ from the mixture $Y(f,t_s)$. This corresponds to $Mask_i(f,t_s) = Mask_i^{opt}(f,t_s)$ in (5.10). Hence, this is the maximum attainable $S_{Mask_i}^{Y \to X_i, P_i}$ value. For other cases of $PSR_{Mask_i}^{X_i}$ and $SIR_{Mask_i}^{X_i}$, $S_{Mask_i}^{Y \to X_i, P_i} < 1$. Using the above concept, we can extend the analysis for the case of separating N_s sources. A mixture $y(t)$ is said to be *fully* separable to all the N_s sources if and only if $S_{Mask_1,\ldots,Mask_{N_s}}^{Y \to X_1,\ldots,X_{N_s}} = 1$ in (5.12). For the case $S_{Mask_1,\ldots,Mask_{N_s}}^{Y \to X_1,\ldots,X_{N_s}} < 1$, this implies that some of the sources overlap with each other in the TF domain and therefore, they cannot be fully separated. Thus, $S_{Mask_1,\ldots,Mask_{N_s}}^{Y \to X_1,\ldots,X_{N_s}}$ provides the quantitative performance measure to evaluate how separable the mixture is in the TF domain. In the following, we show the analysis of how cochleagram, log-frequency

spectrogram and classic spectrogram affect the separability of the mixture. To make the comparison fair, the ideal binary mask (IBM) [109] from the original sources is generated for comparing all TF representations.

5.1.5.1 Mixture of two sources

In this experiment, three types of mixture are generated: (i) music mixed with music, (ii) speech mixed with music and (iii) speech mixed with speech. All source signals are sampled at 16kHz. The speech signals are selected from TIMIT database and normalized to unit energy. The music sources are selected from the RWC [100] database and similarly normalized to unit energy as well. Two sources are randomly chosen from the databases and the mixed signal is generated by live mixing the two sources. All mixed signals are sampled at 16 kHz sampling rate. The separability results for a mixture of two sources are tabulated in Table 5.1.

Table 5.1: Overall separability performance for mixture of two sources

Types of TF domain	**Mixtures**	**PSR**	**SIR**	$S_{M_1,M_2}^{Y\to X_1,X_2}$
Cochleagram	music and music	0.996	275.8	0.993
	music and speech	0.995	186.8	0.989
	speech and speech	0.984	184.2	0.979
Log-frequency spectrogram	music and music	0.958	165.5	0.953
	music and speech	0.942	118.5	0.947
	speech and speech	0.943	20.2	0.934
Spectrogram	music and music	0.885	55.8	0.869
	music and speech	0.882	53.6	0.865
	speech and speech	0.871	50.83	0.854

TF representation using different window length has also been investigated and the

evaluation results are tabulated in Table 5.2.

Table 5.2: Separability under different window length

Types of TF domain	**Window Length**	$S_{M_1,M_2}^{Y\to X_1,X_2}$
Cochleagram	20ms (320)	**0.985**
	32ms (512)	0.972
	64ms (1024)	0.965
	128ms (2048)	0.892
Log-frequency spectrogram	20ms (320)	0.813
	32ms (512)	0.874
	64ms (1024)	**0.948**
	128ms (2048)	0.912
Spectrogram	20ms (320)	0.801
	32ms (512)	0.834
	64ms (1024)	**0.864**
	128ms (2048)	0.842

Table 5.2 shows the average sparability results for all types of the mixture when using different window length. The bracketed number shows the number of data points corresponding to the particular window length. It is quite clear that, for both spectrogram and log-frequency spectrogram settings, the STFT with 1024-point window length is the best setting to analyse the separability performance. The results of PSR, SIR and separability for each TF domain are obtained by averaging over 300 realizations. From the listening performance test in [51], it was concluded that $S_{Mask_i}^{Y\to X_i,P_i} > 0.8$ implies acceptable separation performance. From the results in Table 5.1, it is noted that that all TF representations satisfy this condition. Analysing the separability results, it is seen that the spectrogram performs with the relatively poorer results with $S_{Mask_1,Mask_2}^{Y\to X_1,X_2} \approx 0.86$ while the log-frequency spectrogram shows better results $S_{Mask_1,Mask_2}^{Y\to X_1,X_2} \approx 0.94$ than the spectrogram. However, cochleagram exhibits the best separability among the three TF representations

with $S^{Y\to X_1,X_2}_{Mask_1,Mask_2} \approx 0.98$. In addition, it should be noted from Table 5.1 that the average SIR of cochleagram exhibits much higher value than those of spectrogram and log-frequency spectrogram. This implies that the amount of interference between any two sources is lesser in the cochleagram.

5.1.5.2 Mixture of multiple sources

The analysis conducted in 5.1.4.1 is based on a mixture of two sources. In below, we extend the separability analysis over a number of sources from 2 to 8. For mixture of music and speech sources, the number of music sources is selected to balance with number of speech sources (e.g. for mixture of 8 sources, 4 are drawn from music and another 4 from speech; for mixture of 7 sources, either 3 (or 4) are drawn from music and another 4 (or 3) from speech). This is shown in Figure 5.2. Similar to the first experiment, the separability performance for each TF representation is obtained by averaging over 300 realizations.

In Figure 5.2, it is observed that for all number of sources, the cochleagram can be singled out to show the best separability performance across all different types of mixture. It is worth pointing out that the cochleagram always retain a high level of separability even when the number of sources increases. Also, the curve of separability decreases steadily as the number of sources increases. On the contrary, other TF representations fail to separate the mixture when large number of sources is present, e.g. for mixture of music and speech (8 sources mixed), $S^{Y\to X_1,\ldots,X_{N_s}}_{Mask_1,\ldots,Mask_{N_s}} \approx 0.65$ for spectrogram and $S^{Y\to X_1,\ldots,X_{N_s}}_{Mask_1,\ldots,Mask_{N_s}} \approx 0.3$ for log-frequency spectrogram. They are all below the acceptable level of separability (which is $S^{Y\to X_1,\ldots,X_{N_s}}_{Mask_1,\ldots,Mask_{N_s}} \geq 0.8$). On the other hand, cochleagram maintains at $S^{Y\to X_1,\ldots,X_{N_s}}_{Mask_1,\ldots,Mask_{N_s}} \approx 0.9$

which is manifold well above the rest. It is noted that the curve of separability for the log-frequency spectrogram decreases very sharply as number of sources increases. In Table 5.1, it is seen that the log-frequency spectrogram leads to better separability than the classic spectrogram. However, this is not always the case especially when the number of sources in the mixture is increased from four onwards. The curve in Figure 5.2 is clear to indicate that the separability of the spectrogram degrades more gracefully as compared with the log-frequency spectrogram. Finally, of all the three mixture types and over the range of number of sources, only the cochleagram preserves the separability larger than 0.8. Therefore, based on this study, it can be concluded that the cochleagram is the most separable TF transform for SCBSS among the above three TF representations.

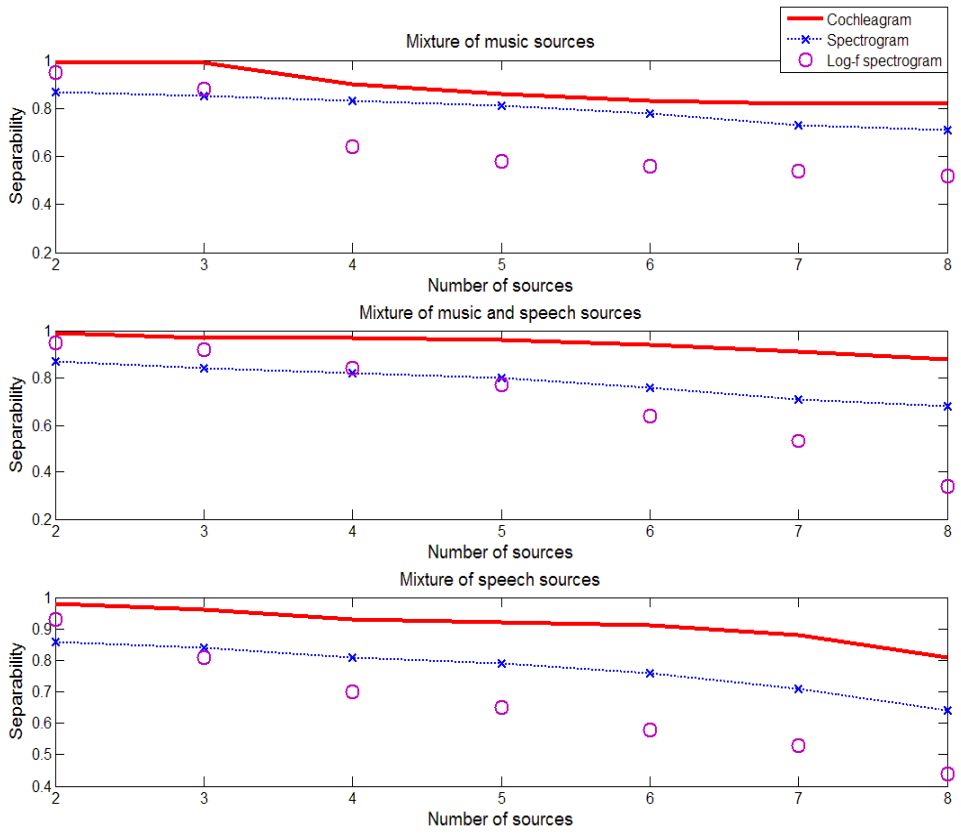


Figure 5.2: Overall separability performance for each mixture type.

5.2 Itakura-Saito based Two-dimensional Nonnegative Matrix Factorization Algorithms

In this section, a family of IS divergence based novel two-dimensional nonnegative matrix factorization will be proposed. These algorithms consist of Quasi-EM two-dimensional nonnegative matrix factorization using the IS divergence (Quasi-EM IS-NMF2D); multiplicative update rule based regularised two-dimensional sparse nonnegative matrix factorization using the IS divergence (IS-SNMF2D) and multiplicative update rule based variable regularised two-dimensional nonnegative matrix factorization using the IS divergence (IS-vRNMF2D). The IS divergence [117] is obtained from the maximum likelihood (ML) estimation of a short-time speech spectra under the autoregressive modeling. It was originally presented as a measure of the goodness of fit between two spectra and has been proven to be efficient especially in terms obtaining the good perceptual properties of the reconstructed audio signals. The IS divergence also leads to desirable statistical interpretations of the NMF [118]. Most significantly, the NMF with IS divergence is scale invariant which enables low energy components of $|\mathbf{Y}|^{\cdot 2}$ bear the same relative importance as high energy ones. This is relevant to situations where the coefficients of $|\mathbf{Y}|^{\cdot 2}$ have a large dynamic range such as in audio short-term spectra. The IS divergence is formally defined as:

$$d_{IS}(a|b) = \frac{a}{b} - \log\frac{a}{b} - 1 \tag{5.16}$$

The IS divergence is a limit case of the β -divergence [86] which is defined as:

$$d_\beta(a\,|\,b)=\begin{cases}\dfrac{1}{\beta(\beta-1)}\left(a^\beta+(\beta-1)b^\beta-\beta ab^{\beta-1}\right) & ,\quad \beta\in\Re\setminus\{0,1\}\\ a\left(\log a-\log b\right)+\left(b-a\right) & ,\quad \beta=1\\ \dfrac{a}{b}-\log\dfrac{a}{b}-1 & ,\quad \beta=0\end{cases} \tag{5.17}$$

It is interesting to note that for $\beta=2$, the Euclidean distance is obtained as expressed by the Frobenius norm and for $\beta=1$ the generalized Kullback-Leibler divergence is defined. Therefore, the β-divergence can be simply represented as $d_\beta(\gamma a\,|\,\gamma b)=\gamma^\beta d_\beta(a\,|\,b)$. For $\beta=0$, this results to the IS divergence which is unique to the β-divergence as it holds the property of scale invariant, namely:

$$d_{IS}(\gamma a\,|\,\gamma b)=d_{IS}(a\,|\,b) \tag{5.18}$$

Eqn. (5.18) shows that a good fit of the factorization for a lower energy component a will cost as much as higher energy component b. On the other hand, factorizations by exploiting LS distance or KL divergence are highly dependent on the high-energy components but less emphasis the low-power components. This inadvertently leads to less precision in the overall estimation of the TF patterns in $|\mathbf{Y}|^{\cdot 2}$.

5.2.1 Quasi-EM based two-dimensional nonnegative matrix factorization using the IS divergence

To facilitate the factorization, the following generative model [118] is defined by:

$$\mathbf{y}_{t_s}=\sum_{k=1}^{K}\boldsymbol{\upsilon}_{k,t_s}\quad\text{where}\quad \boldsymbol{\upsilon}_{k,t_s}\sim N_c\left(0\,,\sum_{\tau,\phi}h^{\phi}_{k,t_s-\tau}\;diag\left(\begin{matrix}\downarrow\phi\\ \mathbf{d}^{\tau}_{k}\end{matrix}\right)\right)\qquad \forall t_s=1,\ldots,T_s \tag{5.19}$$

where $\mathbf{d}^{\tau}_{k}$ is the k^{th} column of $\mathbf{D}^{\tau}$ and $\mathbf{h}^{\phi}_{k}$ is the k^{th} row of $\mathbf{H}^{\phi}$. where $\mathbf{y}_{t_s}\in\mathbb{C}^{F\times 1}$, $\boldsymbol{\upsilon}_{k,t_s}\in\mathbb{C}^{F\times 1}$ and $N_c(u,\Sigma)$ denotes the proper multivariate complex Gaussian distribution

and the components $\boldsymbol{\upsilon}_{1,t_s},\ldots,\boldsymbol{\upsilon}_{K,t_s}$ are both mutually and individually independent. The Expectation-Maximization (EM) framework will be developed for the ML estimation of $\boldsymbol{\theta}=\left\{\mathbf{D}^{\tau},\mathbf{H}^{\phi}\right\}$. Due to the additive structure of the generative model (5.19), the parameters describing each component $\boldsymbol{\upsilon}_k=[\boldsymbol{\upsilon}_{k,1},\ldots,\boldsymbol{\upsilon}_{k,T_s}]$ can be updated separately. To perform the latter, the SAGE algorithm [119] is used. We now consider a partition of the parameter space $\boldsymbol{\theta}=\bigcup_{k=1}^{K}\boldsymbol{\theta}_k$ as: $\boldsymbol{\theta}_k=\left\{\mathbf{d}_k^{\tau},\mathbf{h}_k^{\phi}\right\}$ The SAGE algorithm works by formulating the conditional expectation of the minus log likelihood of $\boldsymbol{\upsilon}_k$ as follows:

$$Q_k^{ML}\left(\boldsymbol{\theta}_k\,|\,\boldsymbol{\theta}'\right)=-\int_{\boldsymbol{\upsilon}_k}p\left(\boldsymbol{\upsilon}_k\,|\,\mathbf{Y},\boldsymbol{\theta}'\right)\log p\left(\boldsymbol{\upsilon}_k\,|\,\boldsymbol{\theta}_k\right)d\boldsymbol{\upsilon}_k \tag{5.20}$$

where $\boldsymbol{\theta}'$ always contains the most up-to-date parameter values $\left\{\mathbf{d}^{\tau},\mathbf{h}^{\phi}\right\}$.

5.2.1.1 Estimation of the spectral basis and temporal code using Quasi-EM method

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

$$\begin{aligned}-\log p\left(\boldsymbol{\upsilon}_k\,|\,\boldsymbol{\theta}_k\right)&=-\sum_{t_s=1}^{T_s}\sum_{f=1}^{F}\log N_c\left(\upsilon_{k,f,t_s}\,\middle|\,0,\sum_{\tau,\phi}d_{f-\phi,k}^{\tau}\,h_{k,t_s-\tau}^{\phi}\right)\\&\doteq\sum_{t_s=1}^{T_s}\sum_{f=1}^{F}\log\left(\sum_{\tau,\phi}d_{f-\phi,k}^{\tau}\,h_{k,t_s-\tau}^{\phi}\right)+\frac{\left|\upsilon_{k,f,t_s}\right|^2}{\sum_{\tau,\phi}d_{f-\phi,k}^{\tau}\,h_{k,t_s-\tau}^{\phi}}\end{aligned} \tag{5.21}$$

where $\doteq$ in the second line denotes equality up to constant terms. Then, the hidden-data posterior is obtained through Wiener filtering:

$$p\left(\boldsymbol{\upsilon}_k\,|\,\mathbf{Y},\boldsymbol{\theta}\right)=\prod_{t_s=1}^{T_s}\prod_{f=1}^{F}N_c\left(\upsilon_{k,f,t_s}\,\middle|\,u_{k,f,t_s}^{post}\,\sigma_{k,f,t_s}^{post}\right) \tag{5.22}$$

for a fixed *k*. Thus, the E-step merely includes computing the posterior power $\mathbf{V}_k^{post}$ of component $\boldsymbol{\upsilon}_k$, defined as $[\mathbf{V}_k^{post}]_{f,t_s}=v_{k,f,t_s}=\left|u_{k,f,t_s}^{post}\right|^2+\sigma_{k,f,t_s}^{post}$ where u_{k,f,t_s}^{post} and σ_{k,f,t_s}^{post} are the posterior mean and variance of υ_{k,f,t_s}, given by:

$$u_{k,f,t_s}^{post} = \frac{\sum_{\tau,\phi} d_{f-\phi,k}^{\tau} h_{k,t_s-\tau}^{\phi}}{\sum_{\tau,\phi,l} d_{f-\phi,l}^{\tau} h_{l,t_s-\tau}^{\phi}} y_{f,t_s} \tag{5.23}$$

$$\sigma_{k,f,t_s}^{post} = \frac{\sum_{\tau,\phi} d_{f-\phi,k}^{\tau} h_{k,t_s-\tau}^{\phi}}{\sum_{\tau,\phi,l} d_{f-\phi,l}^{\tau} h_{l,t_s-\tau}^{\phi}} \sum_{\tau,\phi,l\neq k} d_{f-\phi,l}^{\tau} h_{l,t_s-\tau}^{\phi} \tag{5.24}$$

The M-step can be conducted to treat as the following one-component NMF problem:

$$\begin{aligned} Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}') &\doteq \sum_{t_s=1}^{T_s}\sum_{f=1}^{F} \log\left(\sum_{\tau,\phi} d_{f-\phi,k}^{\tau} h_{k,t_s-\tau}^{\phi}\right) + \frac{\left|u_{k,f,t_s}^{post'}\right|^2 + \sigma_{k,f,t_s}^{post'}}{\sum_{\tau,\phi} d_{f-\phi,k}^{\tau} h_{k,t_s-\tau}^{\phi}} \\ &\doteq \sum_{t_s=1}^{T_s}\sum_{f=1}^{F} D_{IS}\left(\left|u_{k,f,t_s}^{post'}\right|^2 + \sigma_{k,f,t_s}^{post'} \,|\, \sum_{\tau,\phi} d_{f-\phi,k}^{\tau} h_{k,t_s-\tau}^{\phi}\right) \end{aligned} \tag{5.25}$$

Given (5.25), the E-step merely includes computing the posterior power $\mathbf{V}_k^{post}$ of component $\boldsymbol{\upsilon}_k$, defined as $[\mathbf{V}_k^{post}]_{f,t_s} = v_{k,f,t_s} = \left|u_{k,f,t_s}^{post}\right|^2 + \sigma_{k,f,t_s}^{post}$ where u_{k,f,t_s}^{post} and σ_{k,f,t_s}^{post} are the posterior mean and variance of υ_{k,f,t_s} defined in (5.21) and (5.22). Table 5.3 shows the pseudo MATLAB code of the proposed Quasi-EM IS-NMF2D algorithm. The M-step thus amounts to minimising $D_{IS}\left(\mathbf{V}_k^{post\prime} \,|\, \sum_{\tau,\phi} \mathbf{d}_k^{\tau\downarrow\phi\rightarrow} \mathbf{h}_k^{\phi}\right)$ in (5.25) where $\mathbf{V}_k^{post\prime}$ denotes $\mathbf{V}_k^{post}$ as computed from $\boldsymbol{\theta}'$. The derivative of a given element of $g_{k,f,t_s} = \sum_{\tau,\phi} d_{f-\phi,k}^{\tau} h_{k,t_s-\tau}^{\phi}$ with respect to $d_{f,k}^{\tau}$ and h_{k,t_s}^{ϕ} is defined as:

$$\frac{\partial g_{k,f,t_s}}{\partial d_{f',k'}^{\tau'}} = \frac{\partial \sum_{\tau,\phi} d_{f-\phi,k}^{\tau} h_{k,t_s-\tau}^{\phi}}{\partial d_{f',k'}^{\tau'}} = h_{k',t_s-\tau'}^{f-f'} \tag{5.26}$$

$$\frac{\partial g_{k,f,t_s}}{\partial h_{k',t_s'}^{\phi'}} = \frac{\partial \sum_{\tau,\phi} d_{f-\phi,k}^{\tau} h_{k,t_s-\tau}^{\phi}}{\partial h_{k',t_s'}^{\phi'}} = d_{f-\phi',k'}^{t_s-t_s'} \tag{5.27}$$

The derivatives corresponding to $d_{f,k}^{\tau}$ and h_{k,t_s}^{ϕ} can be obtained as:

$$\begin{aligned}\frac{\partial Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}')}{\partial d_{f',k'}^{\tau'}} &= \frac{\partial}{\partial d_{f',k'}^{\tau'}}\left(\sum_{f,t_s}\log\left(g_{k,f,t_s}\right)+\frac{v'_{k,f,t_s}}{g_{k,f,t_s}}\right)\\ &= \sum_{f,t_s}\frac{h_{k',t_s-\tau'}^{f-f'}}{g_{k,f,t_s}}-\frac{v'_{k,f,t_s}}{g_{k,f,t_s}^2}h_{k',t_s-\tau'}^{f-f'}\\ &= \sum_{f,t_s}\left(\frac{g_{k,f,t_s}-v'_{k,f,t_s}}{g_{k,f,t_s}^2}\right)h_{k',t_s-\tau'}^{f-f'}\\ &= \sum_{\phi,t_s}\left(\frac{g_{k,f'+\phi,t_s}-v'_{k,f'+\phi,t_s}}{g_{k,f'+\phi,t_s}^2}\right)h_{k',t_s-\tau'}^{\phi}\end{aligned} \tag{5.28}$$

and

$$\begin{aligned}\frac{\partial Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}')}{\partial h_{k',t_s'}^{\phi'}} &= \frac{\partial}{\partial h_{k',t_s'}^{\phi'}}\left(\sum_{f,t_s}\log\left(g_{k,f,t_s}\right)+\frac{v'_{k,f,t_s}}{g_{k,f,t_s}}\right)\\ &= \sum_{f,t_s}\left(\frac{g_{k,f,t_s}-v'_{k,f,t_s}}{g_{k,f,t_s}^2}\right)d_{f-\phi',k'}^{t_s-t_s'}\\ &= \sum_{\tau,f}\left(\frac{g_{k,f,t_s'+\tau}-v'_{k,f,t_s'+\tau}}{g_{k,f,t_s'+\tau}^2}\right)d_{f-\phi',k'}^{\tau}\end{aligned} \tag{5.29}$$

Unlike the conventional EM algorithm [118], it is not possible in (5.28) and (5.29) to directly set $\partial Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}')/d_{f',k'}^{\tau'}=0$ and $\partial Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}')/h_{k',t_s'}^{\phi'}=0$. Therefore, closed form expressions for estimating $d_{f,k}^{\tau}$ and h_{k,t_s}^{ϕ} cannot be accomplished. To overcome this problem, we develop the following update rules and unify it as part of the M-step. The rules are derived as follows:

$$d_{f',k'}^{\tau'} \leftarrow d_{f',k'}^{\tau'} - \eta_D\frac{\partial Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}')}{\partial d_{f',k'}^{\tau'}} \quad \text{and} \quad h_{k',t_s'}^{\phi'} \leftarrow h_{k',t_s'}^{\phi'} - \eta_H\frac{\partial Q_k^{ML}(\boldsymbol{\theta}_k|\boldsymbol{\theta}')}{\partial h_{k',t_s'}^{\phi'}} \tag{5.30}$$

where η_D and η_H are positive learning rates which can be obtained by following the work in [74]:

$$\eta_D = \frac{d_{f',k'}^{\tau'}}{\sum_{\phi,t_s}(g_{k,f'+\phi,t_s})^{-1}h_{k',t_s-\tau'}^{\phi}} \quad \text{and} \quad \eta_H = \frac{h_{k',t_s'}^{\phi'}}{\sum_{\tau,f}d_{f-\phi',k'}^{\tau}(g_{k,f,t_s'+\tau})^{-1}} \tag{5.31}$$

Thus, the update rule is obtained as:

$$
\begin{aligned}
d_{f',k'}^{\tau'} &\leftarrow d_{f',k'}^{\tau'} - \frac{-d_{f',k'}^{\tau'} \sum_{\phi,t_s} \left(\left(g_{k,f'+\phi,t_s} \right)^{-2} \left(v'_{k,f'+\phi,t_s} - g_{k,f'+\phi,t_s} \right) \right) h_{k',t_s-\tau'}^{\phi}}{\sum_{\phi,t_s} \left(g_{k,f'+\phi,t_s} \right)^{-1} h_{k',t_s-\tau'}^{\phi}} \\
&= d_{f',k'}^{\tau'} \left(\frac{\sum_{\phi,t_s} \left(g_{k,f'+\phi,t_s} \right)^{-1} h_{k',t_s-\tau'}^{\phi} + \sum_{\phi,t_s} \left(\left(g_{k,f'+\phi,t_s} \right)^{-2} \left(v'_{k,f'+\phi,t_s} - g_{k,f'+\phi,t_s} \right) \right) h_{k',t_s-\tau'}^{\phi}}{\sum_{\phi,t_s} \left(g_{k,f'+\phi,t_s} \right)^{-1} h_{k',t_s-\tau'}^{\phi}} \right) \\
&= d_{f',k'}^{\tau'} \frac{\sum_{\phi,t_s} \left(g_{k,f'+\phi,t_s} \right)^{-2} v'_{k,f'+\phi,t_s} h_{k',t_s-\tau'}^{\phi}}{\sum_{\phi,t_s} \left(g_{k,f'+\phi,t_s} \right)^{-1} h_{k',t_s-\tau'}^{\phi}}
\end{aligned}
\tag{5.32}
$$

Similar, the update rules in $h_{k',t'_s}^{\phi'}$ writes:

$$
\begin{aligned}
h_{k',t'_s}^{\phi'} &\leftarrow h_{k',t'_s}^{\phi'} - \frac{-h_{k',t'_s}^{\phi'} \sum_{\tau,f} d_{f-\phi',k'}^{\tau} \left(\left(g_{k,f,t'_s+\tau} \right)^{-2} \left(v'_{k,f,t'_s+\tau} - g_{k,f,t'_s+\tau} \right) \right)}{\sum_{\tau,f} d_{f-\phi',k'}^{\tau} \left(g_{k,f,t'_s+\tau} \right)^{-1}} \\
&= h_{k',t'_s}^{\phi'} \left(\frac{\sum_{\tau,f} d_{f-\phi',k'}^{\tau} \left(g_{k,f,t'_s+\tau} \right)^{-1} + \sum_{\tau,f} d_{f-\phi',k'}^{\tau} \left(\left(g_{k,f,t'_s+\tau} \right)^{-2} \left(v'_{k,f,t'_s+\tau} - g_{k,f,t'_s+\tau} \right) \right)}{\sum_{\tau,f} d_{f-\phi',k'}^{\tau} \left(g_{k,f,t'_s+\tau} \right)^{-1}} \right) \\
&= h_{k',t'_s}^{\phi'} \frac{\sum_{\tau,f} d_{f-\phi',k'}^{\tau} \left(g_{k,f,t'_s+\tau} \right)^{-2} v'_{k,f,t'_s+\tau}}{\sum_{\tau,f} d_{f-\phi',k'}^{\tau} \left(g_{k,f,t'_s+\tau} \right)^{-1}}
\end{aligned}
\tag{5.33}
$$

Comparity with the standard gradient descent approach, the above update rules have an advantage of ensuring the nonnegativity constraints of $d_{f,k}^{\tau}$ and h_{k,t_s}^{ϕ} are always maintained during every iteration. In matrix notation, the above can be written as follows:

$$
\mathbf{d}_k^{\tau} \leftarrow \mathbf{d}_k^{\tau} \bullet \frac{\sum_{\phi} \left(\left(\overset{\uparrow\phi}{\mathbf{G}}_k \right)^{\cdot -2} \bullet \overset{\uparrow\phi}{\mathbf{V}}_k \right) \overset{\rightarrow\tau^{\mathbf{T}}}{\mathbf{h}_k^{\phi}}}{\sum_{\phi} \left(\overset{\uparrow\phi}{\mathbf{G}}_k \right)^{\cdot -1} \overset{\rightarrow\tau^{\mathbf{T}}}{\mathbf{h}_k^{\phi}}} \quad \text{and} \quad \mathbf{h}_k^{\phi} \leftarrow \mathbf{h}_k^{\phi} \bullet \frac{\sum_{\tau} \overset{\downarrow\phi^{\mathbf{T}}}{\mathbf{d}_k^{\tau}} \left(\left(\overset{\leftarrow\tau}{\mathbf{G}}_k \right)^{\cdot -2} \bullet \overset{\leftarrow\tau}{\mathbf{V}}_k \right)}{\sum_{\tau} \overset{\downarrow\phi^{\mathbf{T}}}{\mathbf{d}_k^{\tau}} \left(\overset{\leftarrow\tau}{\mathbf{G}}_k \right)^{\cdot -1}}
\tag{5.34}
$$

The specific steps of update rule for Quasi-EM IS-NMF2D algorithm are summarised as follow:

Table 5.3: Quasi-EM IS-NMF2D algorithm

Initialize $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ with nonnegative values

for *iter=1:iteration*

for *k=1:K*

Compute $\Upsilon_k = \sum_{\tau}\sum_{\phi} \mathbf{d}_k^{\tau} \mathbf{h}_k^{\phi} \Big/ \sum_{\tau}\sum_{\phi} \mathbf{D}^{\tau} \mathbf{H}^{\phi}$ (E-step)

Compute $\mathbf{V}_k = \Upsilon_k^{.2} \bullet |\mathbf{Y}|^{.2} + (1-\Upsilon_k) \bullet \left(\sum_{\tau}\sum_{\phi} \mathbf{d}_k^{\tau} \mathbf{h}_k^{\phi} \right)$ (E-step)

Run M-step until the convergence is achieved

$\mathbf{h}_k^{\phi} \leftarrow \mathbf{h}_k^{\phi} \bullet \left(\sum_{\tau} \mathbf{d}_k^{\tau} \left(\left(\mathbf{G}_k \right)^{.-2} \bullet \mathbf{V}_k \right) \Big/ \sum_{\tau} \mathbf{d}_k^{\tau} \left(\mathbf{G}_k \right)^{.-1} \right)$ for all τ, ϕ (M-step)

Normalize $\mathbf{h}_k^{\phi}$

$\mathbf{d}_k^{\tau} \leftarrow \mathbf{d}_k^{\tau} \bullet \left(\sum_{\phi} \left(\left(\mathbf{G}_k \right)^{.-2} \bullet \mathbf{V}_k \right) \mathbf{h}_k^{\phi} \Big/ \sum_{\phi} \left(\mathbf{G}_k \right)^{.-1} \mathbf{h}_k^{\phi} \right)$ for all τ, ϕ (M-step)

Normalize $\mathbf{d}_k^{\tau}$

end

end

In Table 5.3, the term $\sum_{\tau}\sum_{\phi} \mathbf{D}^{\tau} \mathbf{H}^{\phi}$ needs to be computed only once at initialization, and subsequently be updated as $\sum_{\tau}\sum_{\phi} \mathbf{D}^{\tau} \dot{\mathbf{H}}^{\phi} - \sum_{\tau}\sum_{\phi} \dot{\mathbf{d}}_k^{\tau} \dot{\mathbf{h}}_k^{\phi} + \sum_{\tau}\sum_{\phi} \ddot{\mathbf{d}}_k^{\tau} \ddot{\mathbf{h}}_k^{\phi}$ where $\dot{a}$ represents the old updates and $\ddot{a}$ denotes the new updates.

5.2.2 Two-dimensional sparse nonnegative matrix factorization using the IS divergence

In this sub-section, we consider to directly use multiplicative update (MU) approach for IS divergence based two-dimensional nonnegative matrix factorization. In addition, the sparse parameter will be analysed and the family of (MU) IS-based two-dimensional nonnegative matrix factorization algorithms will be developed. To facilitate the decomposition in (5.12), the following generative model [118] is considered:

$$\left|\mathbf{Y}\right|^{\cdot 2} = \left(\sum_{\tau} \sum_{\phi} \mathbf{D}^{\tau}{}^{\downarrow\phi} \, \mathbf{H}^{\phi}{}^{\rightarrow\tau} \right) \bullet \mathbf{E} \tag{5.35}$$

where "$\bullet$" is element-wise product and $\mathbf{E}$ is a matrix of multiplicative independent and identically-distributed (i.i.d.) Gamma noise with mean unity i.e. $p(\mathbf{E}_{f,t_s}) = \xi^G(\mathbf{E}_{f,t_s} \mid \alpha, \varsigma)$ where $\xi^G(\mathbf{E}_{f,t_s} \mid \alpha, \varsigma)$ denotes the gamma probability density function (pdf) defined as:

$$\xi^G(\mathbf{E}_{f,t_s} \mid \alpha, \varsigma) = \frac{\varsigma^{\alpha}}{\Gamma(\alpha)} \left(\mathbf{E}_{f,t_s}\right)^{\alpha-1} \exp\left(-\varsigma \mathbf{E}_{f,t_s}\right), \mathbf{E}_{f,t_s} \geq 0 \tag{5.36}$$

Next, choose a prior distribution $p_{\mathbf{D},\mathbf{H}}(\mathbf{D},\mathbf{H})$ over the factors $\{\mathbf{D},\mathbf{H}\}$ in the model. The posterior is found using Bayes rule, namely:

$$p(\mathbf{D},\mathbf{H} \mid \left|\mathbf{Y}\right|^{\cdot 2}) = \frac{p(\left|\mathbf{Y}\right|^{\cdot 2} \mid \mathbf{D},\mathbf{H}) \, p_{\mathbf{D},\mathbf{H}}(\mathbf{D},\mathbf{H})}{P(\left|\mathbf{Y}\right|^{\cdot 2})} \tag{5.37}$$

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

$$\log p(\mathbf{D},\mathbf{H} \mid \left|\mathbf{Y}\right|^{\cdot 2}) = \log p(\left|\mathbf{Y}\right|^{\cdot 2} \mid \mathbf{D},\mathbf{H}) + \log p_{\mathbf{D},\mathbf{H}}(\mathbf{D},\mathbf{H}) + \text{const} \tag{5.38}$$

Under the independent and identically distributed (i.i.d.) noise assumption, the first term of the right hand side of (5.38) can be expanded as:

$$\begin{aligned}
-\log p\left(\mathbf{Y} \mid \mathbf{D},\mathbf{H}\right) &= -\sum_{t_s=1}^{Ts} \sum_{f=1}^{F} \log \xi^G \left(\frac{|\mathbf{Y}|^{\cdot 2}_{f,t_s}}{\sum_i \sum_{\tau} \sum_{\phi} \mathbf{D}^{\tau}_{f,i}{}^{\downarrow\phi} \, \mathbf{H}^{\phi}_{i,t_s}{}^{\rightarrow\tau}} \middle| \alpha, \varsigma \right) \Bigg/ \sum_i \sum_{\tau} \sum_{\phi} \mathbf{D}^{\tau}_{f,i}{}^{\downarrow\phi} \, \mathbf{H}^{\phi}_{i,t_s}{}^{\rightarrow\tau} \\
&\doteq \varsigma \sum_{t_s=1}^{Ts} \sum_{f=1}^{F} \frac{|\mathbf{Y}|^{\cdot 2}_{f,t_s}}{\sum_i \sum_{\tau} \sum_{\phi} \mathbf{D}^{\tau}_{f,i}{}^{\downarrow\phi} \, \mathbf{H}^{\phi}_{i,t_s}{}^{\rightarrow\tau}} - \frac{\alpha}{\varsigma} \log \frac{|\mathbf{Y}|^{\cdot 2}_{f,t_s}}{\sum_i \sum_{\tau} \sum_{\phi} \mathbf{D}^{\tau}_{f,i}{}^{\downarrow\phi} \, \mathbf{H}^{\phi}_{i,t_s}{}^{\rightarrow\tau}} - 1 \\
&= D_{IS}\left(|\mathbf{Y}|^{\cdot 2} \middle| \sum_i \sum_{\tau} \sum_{\phi} \mathbf{D}^{\tau}_{f,i}{}^{\downarrow\phi} \, \mathbf{H}^{\phi}_{i,t_s}{}^{\rightarrow\tau} \right)
\end{aligned} \tag{5.39}$$

where $\doteq$ in the third line denotes equality up to a positive scale and a constant. The ratio α/ς is simply the mean of the Gamma distribution which by definition is equal to unity. Thus, the last line of (5.39) is obtained by setting $\alpha/\varsigma = 1$. The second term of (5.38)

consists of the prior distribution of $\mathbf{H}$ and $\mathbf{D}$ where they are jointly independent. The prior over $\mathbf{H}$ which is assumed to be one-sided exponential i.e. $p_{\mathbf{H}}(\mathbf{H}\,|\,\lambda)=\prod_{\phi}\prod_{i}\prod_{t_s}\lambda\exp\left(-\lambda\mathbf{H}^{\phi}_{i,t_s}\right)$ with scale parameter λ which weights the importance of the sparsity term to the reconstruction and the prior over $\mathbf{D}$ is assumed to be flat with each column constrained to have unit norm. The IS-divergence cost function for SNMF2D is defined as the negative log likelihood of $p(\mathbf{Y}\,|\,\mathbf{D},\mathbf{H})$ with prior over $\mathbf{H}$:

$$\begin{aligned} C_{IS}^{SNMF\,2D}(\mathbf{D},\mathbf{H}) &= -\log p(\mathbf{Y}\,|\,\mathbf{D},\mathbf{H}) - \log p_{\mathbf{H}}(\mathbf{H}) \\ &\doteq D_{IS}\left(|\mathbf{Y}|^{\cdot 2}\,\middle|\,\sum_{i}\sum_{\tau}\sum_{\phi}\overset{\downarrow\phi}{\mathbf{D}}{}_{i}^{\tau}\overset{\rightarrow\tau}{\mathbf{H}}{}_{i}^{\phi}\right)+\lambda f(\mathbf{H}) \qquad (5.40) \\ &= \sum_{t_s=1}^{Ts}\sum_{f=1}^{F}\frac{|\mathbf{Y}|^{2}_{f,t_s}}{\sum_{i}\sum_{\tau}\sum_{\phi}\overset{\downarrow\phi}{\mathbf{D}}{}_{i}^{\tau}\overset{\rightarrow\tau}{\mathbf{H}}{}_{i}^{\phi}}-\log\frac{|\mathbf{Y}|^{2}_{f,t_s}}{\sum_{i}\sum_{\tau}\sum_{\phi}\overset{\downarrow\phi}{\mathbf{D}}{}_{i}^{\tau}\overset{\rightarrow\tau}{\mathbf{H}}{}_{i}^{\phi}}-1+\lambda f(\mathbf{H}) \end{aligned}$$

where $f(\mathbf{H})=\|\mathbf{H}\|_1=\sum_{\phi,i,t_s}\left|\mathbf{H}^{\phi}_{i,t_s}\right|$ is L_1-norm regularization which can resolve the ambiguity by forcing all structures in $\mathbf{H}$ onto $\mathbf{D}$ giving the correct components. Finally, $C_{IS}^{SNMF\,2D}(\mathbf{D},\mathbf{H})$ is equivalent to $D_{IS}\left(|\mathbf{Y}|^{\cdot 2}\,\middle|\,\sum_{i}\sum_{\tau}\sum_{\phi}\overset{\downarrow\phi}{\mathbf{D}}{}_{i}^{\tau}\overset{\rightarrow\tau}{\mathbf{H}}{}_{i}^{\phi}\right)+\lambda f(\mathbf{H})$ up to a positive factor and a constant. Hence the scale invariance of the IS-divergence can be interpreted by the multiplicative noise equivalence. In fact, it is that the noise acts as a scale factor on $\left|\hat{\mathbf{Y}}\right|^{\cdot 2}_{f,t_s}$, here $\left|\hat{\mathbf{Y}}\right|^{\cdot 2}_{f,t_s}=\sum_{\tau}\sum_{\phi}\overset{\downarrow\phi}{\mathbf{D}}{}^{\tau}\overset{\rightarrow\tau}{\mathbf{H}}{}^{\phi}$.

5.2.2.1 Estimation of the spectral basis and temporal code (IS-SNMF2D)

In the matrix factorization, each spectral basis is constrained to be of unit length. Hence, this can be represented by $\tilde{\mathbf{Z}}=\sum_{i}\sum_{\tau}\sum_{\phi}\overset{\downarrow\phi}{\tilde{\mathbf{D}}}{}_{i}^{\tau}\overset{\rightarrow\tau}{\mathbf{H}}{}_{i}^{\phi}$ where $\tilde{\mathbf{D}}^{\tau}_{f,i}=\mathbf{D}^{\tau}_{f,i}\Big/\sqrt{\sum_{\tau,f}(\mathbf{D}^{\tau}_{f,i})^{2}}$ is

factor-wise normalized to $\mathbf{D}^{\tau}$. In view of this constraint, the (5.40) can be re-defined:

$$L_{IS}^{SNMF\,2D}(\mathbf{D},\mathbf{H}) = \sum_{f,t_s}\left(\frac{|\mathbf{Y}|_{f,t_s}^2}{\tilde{\mathbf{Z}}_{f,t_s}} - \log\frac{|\mathbf{Y}|_{f,t_s}^2}{\tilde{\mathbf{Z}}_{f,t_s}} - 1\right) + \lambda f(\mathbf{H}) \tag{5.41}$$

Using the above, the derivatives of (5.41) corresponding to $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ are given by:

$$\begin{aligned}\frac{\partial L_{IS}^{SNMF\,2D}}{\partial \mathbf{D}_{f',i'}^{\tau'}} &= \frac{\partial}{\partial \mathbf{D}_{f',i'}^{\tau'}}\left(\sum_{f,t_s}\left(\frac{|\mathbf{Y}|_{f,t_s}^2}{\tilde{\mathbf{Z}}_{f,t_s}} - \log\frac{|\mathbf{Y}|_{f,t_s}^2}{\tilde{\mathbf{Z}}_{f,t_s}} - 1\right) + \lambda f(\mathbf{H})\right) \\ &= \sum_{f,t_s}\left(\left(\tilde{\mathbf{Z}}_{f,t_s}\right)^{-2}\left(\tilde{\mathbf{Z}}_{f,t_s} - |\mathbf{Y}|_{f,t_s}^2\right)\right)\mathbf{H}_{i',t_s-\tau'}^{f-f'} \\ &= -\sum_{\phi,t_s}\left(\left(\tilde{\mathbf{Z}}_{f'+\phi,t_s}\right)^{-2}\left(|\mathbf{Y}|_{f'+\phi,t_s}^2 - \tilde{\mathbf{Z}}_{f'+\phi,t_s}\right)\right)\mathbf{H}_{i',t_s-\tau'}^{\phi}\end{aligned} \tag{5.42}$$

Similarly:

$$\begin{aligned}\frac{\partial L_{IS}^{SNMF\,2D}}{\partial \mathbf{H}_{i',t'}^{\phi'}} &= \sum_{f,t_s}\tilde{\mathbf{D}}_{f-\phi',i'}^{t_s-t_s'}\left(\left(\tilde{\mathbf{Z}}_{f,t_s}\right)^{-2}\left(\tilde{\mathbf{Z}}_{f,t_s} - |\mathbf{Y}|_{f,t_s}^2\right)\right) + \lambda \\ &= -\sum_{\tau,f}\tilde{\mathbf{D}}_{f-\phi',i'}^{\tau}\left(\left(\tilde{\mathbf{Z}}_{f,t_s'+\tau}\right)^{-2}\left(|\mathbf{Y}|_{f,t_s'+\tau}^2 - \tilde{\mathbf{Z}}_{f,t_s'+\tau}\right)\right) + \lambda\end{aligned} \tag{5.43}$$

Consequently, by applying the standard gradient decent approach, namely:

$$\mathbf{D}_{f',i'}^{\tau'} \leftarrow \tilde{\mathbf{D}}_{f',i'}^{\tau'} - \eta_D\frac{\partial L_{IS}^{SNMF\,2D}}{\partial \mathbf{D}_{f',i'}^{\tau'}} \quad \text{and} \quad \mathbf{H}_{i',t_s'}^{\phi'} \leftarrow \mathbf{H}_{i',t_s'}^{\phi'} - \eta_H\frac{\partial L_{IS}^{SNMF\,2D}}{\partial \mathbf{H}_{i',t_s'}^{\phi'}} \tag{5.44}$$

where η_D and η_H are positive learning rates which can be obtained by following the approach of Lee and Seung [74] , namely:

$$\eta_D = \frac{\tilde{\mathbf{D}}_{f',i'}^{\tau'}}{\sum_{\phi,t_s}\left(\tilde{\mathbf{Z}}_{f'+\phi,t_s}\right)^{-1}\mathbf{H}_{i',t_s-\tau'}^{\phi}} \quad \text{and} \quad \eta_H = \frac{\mathbf{H}_{i',t_s'}^{\phi'}}{\sum_{\tau,f}\tilde{\mathbf{D}}_{f-\phi',i'}^{\tau}\left(\tilde{\mathbf{Z}}_{f,t_s'+\tau}\right)^{-1} + \lambda} \tag{5.45}$$

Inserting (5.45) into (5.44) leads to the multiplicative update rules:

$$\mathbf{D}^{\tau'}_{f',i'} \leftarrow \tilde{\mathbf{D}}^{\tau'}_{f',i'} - \frac{-\tilde{\mathbf{D}}^{\tau'}_{f',i'} \sum_{\phi,t_s} \left(\left(\tilde{\mathbf{Z}}_{f'+\phi,t_s} \right)^{-2} \left(|\mathbf{Y}|^{.2}_{f'+\phi,t_s} - \tilde{\mathbf{Z}}_{f'+\phi,t_s} \right) \right) \mathbf{H}^{\phi}_{i',t_s-\tau'}}{\sum_{\phi,t_s} \left(\tilde{\mathbf{Z}}_{f'+\phi,t_s} \right)^{-1} \mathbf{H}^{\phi}_{i',t_s-\tau'}}$$

$$= \tilde{\mathbf{D}}^{\tau'}_{f',i'} \left(\frac{\sum_{\phi,t_s} \left(\tilde{\mathbf{Z}}_{f'+\phi,t_s} \right)^{-1} \mathbf{H}^{\phi}_{i',t_s-\tau'} + \sum_{\phi,t_s} \left(\left(\tilde{\mathbf{Z}}_{f'+\phi,t_s} \right)^{-2} \left(|\mathbf{Y}|^{.2}_{f'+\phi,t_s} - \tilde{\mathbf{Z}}_{f'+\phi,t_s} \right) \right) \mathbf{H}^{\phi}_{i',t_s-\tau'}}{\sum_{\phi,t_s} \left(\tilde{\mathbf{Z}}_{f'+\phi,t_s} \right)^{-1} \mathbf{H}^{\phi}_{i',t_s-\tau'}} \right) \tag{5.46}$$

$$= \tilde{\mathbf{D}}^{\tau'}_{f',i'} \frac{\sum_{\phi,t_s} \left(\tilde{\mathbf{Z}}_{f'+\phi,t_s} \right)^{-2} |\mathbf{Y}|^{.2}_{f'+\phi,t_s} \mathbf{H}^{\phi}_{i',t_s-\tau'}}{\sum_{\phi,t_s} \left(\tilde{\mathbf{Z}}_{f'+\phi,t_s} \right)^{-1} \mathbf{H}^{\phi}_{i',t_s-\tau'}}$$

Similarly, the update rules for $\mathbf{H}^{\phi'}_{i',t'_s}$ gives:

$$\mathbf{H}^{\phi'}_{i',t'_s} \leftarrow \mathbf{H}^{\phi'}_{i',t'_s} - \frac{-\mathbf{H}^{\phi'}_{i',t'_s} \left(\sum_{\tau,f} \tilde{\mathbf{D}}^{\tau}_{f-\phi',i'} \left(\left(\tilde{\mathbf{Z}}_{f,t'_s+\tau} \right)^{-2} \left(|\mathbf{Y}|^{.2}_{f,t'_s+\tau} - \tilde{\mathbf{Z}}_{f,t'_s+\tau} \right) \right) + \lambda \right)}{\sum_{\tau,f} \tilde{\mathbf{D}}^{\tau}_{f-\phi',i'} \left(\tilde{\mathbf{Z}}_{f,t'_s+\tau} \right)^{-1} + \lambda}$$

$$= \mathbf{H}^{\phi'}_{i',t'_s} \left(\frac{\sum_{\tau,f} \tilde{\mathbf{D}}^{\tau}_{f-\phi',i'} \left(\tilde{\mathbf{Z}}_{f,t'_s+\tau} \right)^{-1} + \lambda + \sum_{\tau,f} \tilde{\mathbf{D}}^{\tau}_{f-\phi',i'} \left(\left(\tilde{\mathbf{Z}}_{f,t'_s+\tau} \right)^{-2} \left(|\mathbf{Y}|^{.2}_{f,t'_s+\tau} - \tilde{\mathbf{Z}}_{f,t'_s+\tau} \right) \right) - \lambda}{\sum_{\tau,f} \tilde{\mathbf{D}}^{\tau}_{f-\phi',i'} \left(\tilde{\mathbf{Z}}_{f,t'_s+\tau} \right)^{-1} + \lambda} \right) \tag{5.47}$$

$$= \mathbf{H}^{\phi'}_{i',t'_s} \frac{\sum_{\tau,f} \tilde{\mathbf{D}}^{\tau}_{f-\phi',i'} \left(\left(\tilde{\mathbf{Z}}_{f,t'_s+\tau} \right)^{-2} \left(|\mathbf{Y}|^{.2}_{f,t'_s+\tau} \right) \right)}{\sum_{\tau,f} \tilde{\mathbf{D}}^{\tau}_{f-\phi',i'} \left(\tilde{\mathbf{Z}}_{f,t'_s+\tau} \right)^{-1} + \lambda}$$

In terms of matrix notation, the multiplicative learning rules in (5.46) and (5.47) can be written as:

$$\mathbf{D}^{\tau} \leftarrow \tilde{\mathbf{D}}^{\tau} \bullet \frac{\sum_{\phi} \left(\left(\overset{\uparrow\phi}{\tilde{\mathbf{Z}}} \right)^{.-2} \bullet |\mathbf{Y}|^{\uparrow\phi}_{.2} \right)^{\rightarrow\tau^{\mathrm{T}}} \mathbf{H}^{\phi}}{\sum_{\phi} \left(\overset{\uparrow\phi}{\tilde{\mathbf{Z}}} \right)^{.-1} \mathbf{H}^{\phi}} \quad \text{and} \quad \mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi} \bullet \frac{\sum_{\tau} \overset{\downarrow\phi^{\mathrm{T}}}{\tilde{\mathbf{D}}^{\tau}} \left(\left(\overset{\leftarrow\tau}{\tilde{\mathbf{Z}}} \right)^{.-2} \bullet |\mathbf{Y}|^{\leftarrow\tau}_{.2} \right)}{\sum_{\tau} \overset{\downarrow\phi^{\mathrm{T}}}{\tilde{\mathbf{D}}^{\tau}} \left(\overset{\leftarrow\tau}{\tilde{\mathbf{Z}}} \right)^{.-1} + \lambda} \tag{5.48}$$

Table 5.4 summarises the basic steps of the proposed IS-SNMF2D algorithm.

Table 5.4: IS-SNMF2D algorithm

1. Initialize $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ with nonnegative random values
2. Initialize λ with positive values.
3. $\tilde{\mathbf{D}}^{\tau}_{f,i} = \mathbf{D}^{\tau}_{f,i} \Big/ \sqrt{\sum_{\tau,f} (\mathbf{D}^{\tau}_{f,i})^2}$
4. Calculate $\tilde{\mathbf{Z}} = \sum_i \sum_{\tau} \sum_{\phi} \tilde{\mathbf{D}}_i^{\tau} \overset{\downarrow\phi}{\mathbf{H}}_i^{\phi} \overset{\rightarrow\tau}{}$
3. $\mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi} \bullet \dfrac{\sum_{\tau} \overset{\downarrow\phi\mathbf{T}}{\tilde{\mathbf{D}}^{\tau}} \left(\left(\overset{\leftarrow\tau}{\tilde{\mathbf{Z}}} \right)^{\cdot -2} \bullet \left| \mathbf{Y} \right|^{\cdot 2} \right)}{\sum_{\tau} \overset{\downarrow\phi\mathbf{T}}{\tilde{\mathbf{D}}^{\tau}} \left(\overset{\leftarrow\tau}{\tilde{\mathbf{Z}}} \right)^{\cdot -1} + \lambda}$
5. Calculate $\tilde{\mathbf{Z}} = \sum_i \sum_{\tau} \sum_{\phi} \tilde{\mathbf{D}}_i^{\tau} \overset{\downarrow\phi}{\mathbf{H}}_i^{\phi} \overset{\rightarrow\tau}{}$
6. $\mathbf{D}^{\tau} \leftarrow \tilde{\mathbf{D}}^{\tau} \bullet \dfrac{\sum_{\phi} \left(\left(\overset{\uparrow\phi}{\tilde{\mathbf{Z}}} \right)^{\cdot -2} \bullet \left| \mathbf{Y} \right|^{\cdot 2} \right) \overset{\rightarrow\tau\mathbf{T}}{\mathbf{H}^{\phi}}}{\sum_{\phi} \left(\overset{\uparrow\phi}{\tilde{\mathbf{Z}}} \right)^{\cdot -1} \overset{\rightarrow\tau\mathbf{T}}{\mathbf{H}^{\phi}}}$
7. Repeat from step 3 to 6 until convergence.

Using the IS-SNMF2D as the starting point, a family of IS divergence based nonnegative matrix factorization algorithms can be obtained. Firstly, by constraining the convolutive factors $\tau, \phi = \{0\}$ and sparse regularization $\lambda = 0$ in (5.48), this yields the IS divergence based nonnegative matrix factorization (IS-NMF):

$$\text{(IS-NMF)} \quad \mathbf{D} \leftarrow \mathbf{D} \bullet \frac{\left((\mathbf{DH})^{\cdot -2} \bullet |\mathbf{Y}|^{\cdot 2} \right) \mathbf{H}^{\mathbf{T}}}{(\mathbf{DH})^{\cdot -1} \mathbf{H}^{\mathbf{T}}} \quad \text{and} \quad \mathbf{H} \leftarrow \mathbf{H} \bullet \frac{\mathbf{D}^{\mathbf{T}} \left((\mathbf{DH})^{\cdot -2} \bullet |\mathbf{Y}|^{\cdot 2} \right)}{\mathbf{D}^{\mathbf{T}} (\mathbf{DH})^{\cdot -1}} \tag{5.49}$$

Secondly, by constraining the convolutive factors $\tau, \phi = \{0\}$ and enabling the sparsity term on $f(\mathbf{H})$ with L_1-norm, the IS divergence based sparse nonnegative matrix factorization (IS-SNMF) is obtained as follow:

$$\text{(IS-SNMF)}\quad \mathbf{D} \leftarrow \tilde{\mathbf{D}} \bullet \frac{\left(\left(\tilde{\mathbf{D}}\mathbf{H}\right)^{\cdot -2} \bullet \left|\mathbf{Y}\right|^{\cdot 2}\right)\mathbf{H}^{\mathbf{T}}}{\left(\tilde{\mathbf{D}}\mathbf{H}\right)^{\cdot -1}\mathbf{H}^{\mathbf{T}}} \quad \text{and} \quad \mathbf{H} \leftarrow \mathbf{H} \bullet \frac{\tilde{\mathbf{D}}^{\mathbf{T}}\left(\left(\tilde{\mathbf{D}}\mathbf{H}\right)^{\cdot -2} \bullet \left|\mathbf{Y}\right|^{\cdot 2}\right)}{\tilde{\mathbf{D}}^{\mathbf{T}}\left(\tilde{\mathbf{D}}\mathbf{H}\right)^{\cdot -1} + \lambda} \tag{5.50}$$

where $\tilde{\mathbf{D}}_{f,i} = \mathbf{D}_{f,i} \Big/ \sqrt{\sum_{\tau,f} (\mathbf{D}_{f,i})^2}$. Finally, by setting the sparse regularization $\lambda = 0$ in (5.48), this gives the IS divergence based two-dimensional nonnegative matrix factorization (MU IS-NMF2D), namely:

$$\text{(MU IS-NMF2D)}\ \mathbf{D}^{\tau} \leftarrow \mathbf{D}^{\tau} \bullet \frac{\sum_{\phi}\left(\left(\overset{\uparrow\phi}{\mathbf{Z}}\right)^{\cdot -2} \bullet \left|\overset{\uparrow\phi}{\mathbf{Y}}\right|^{\cdot 2}\right)\overset{\rightarrow\tau^{\mathbf{T}}}{\mathbf{H}^{\phi}}}{\sum_{\phi}\left(\overset{\uparrow\phi}{\mathbf{Z}}\right)^{\cdot -1}\overset{\rightarrow\tau^{\mathbf{T}}}{\mathbf{H}^{\phi}}} \quad \mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi} \bullet \frac{\sum_{\tau}\overset{\downarrow\phi^{\mathbf{T}}}{\mathbf{D}^{\tau}}\left(\left(\overset{\leftarrow\tau}{\mathbf{Z}}\right)^{\cdot -2} \bullet \left|\overset{\leftarrow\tau}{\mathbf{Y}}\right|^{\cdot 2}\right)}{\sum_{\tau}\overset{\downarrow\phi^{\mathbf{T}}}{\mathbf{D}^{\tau}}\left(\overset{\leftarrow\tau}{\mathbf{Z}}\right)^{\cdot -1}} \tag{5.51}$$

where $\mathbf{Z} = \sum_{i}\sum_{\tau}\sum_{\phi}\overset{\downarrow\phi}{\mathbf{D}_i^{\tau}}\overset{\rightarrow\tau}{\mathbf{H}_i^{\phi}}$. In the result section, we will conduct an experimental study on the efficacy of all the above algorithms for source separation and subsequently analyse their performance in terms of the sparsity λ and the convolutive factors $\{\tau, \phi\}$.

5.2.3 Variable regularised two-dimensional nonnegative matrix factorization using the IS divergence

In the Section 5.2.2, the IS-SNMF2D algorithm has been developed. However, the drawbacks of IS-SNMF2D originate from its lack of a generalized criterion for controlling the sparsity of $\mathbf{H}$. In practice, the sparsity parameter is set manually. In this section, we proposed our model imposes sparseness on $\mathbf{H}$ element-wise so that *each individual code* has its own distribution. Therefore, the sparsity parameter can be individually optimized for each code. This overcomes the problem of under- and over-sparse factorization. In

addition, each sparsity parameter in the proposed model is learned and adapted as part of the matrix factorization. This bypasses the need of manual selection as in the case of IS-SNMF2D. Secondly, as each audio signal has its own temporal dependency of the frequency patterns, the basis vectors in $\mathbf{D}$ have to be designed to match the characteristics of these patterns efficiently. Hence, we incorporate a suitably designed Gaussian prior on $\mathbf{D}$ to allow those frequency patterns to be expressed for each audio source. To facilitate the decomposition in (5.12), given nonnegative two-dimensional observation matrix $|\mathbf{Y}|^{.2}$, a prior distribution $p(\mathbf{D},\mathbf{H})$ is chosen over the factors $\{\mathbf{D},\mathbf{H}\}$ in the model. The posterior is found using Bayes rule, namely:

$$p\left(\mathbf{D},\mathbf{H}\middle\|\mathbf{Y}|^{.2},\boldsymbol{\Lambda}\right)=\frac{p\left(|\mathbf{Y}|^{.2}\,|\,\mathbf{D},\mathbf{H}\right)p(\mathbf{D})\,p\left(\mathbf{H}\,|\,\boldsymbol{\Lambda}\right)}{P\left(|\mathbf{Y}|^{.2}\right)} \tag{5.52}$$

where the denominator is a constant and $\boldsymbol{\Lambda}=\left[\boldsymbol{\Lambda}^1\,\boldsymbol{\Lambda}^2\cdots\boldsymbol{\Lambda}^{\phi_{\max}}\right]$ with $\boldsymbol{\Lambda}^{\phi}=\left\{\lambda_{i,t_s}^{\phi}\,|\,i=1,...,I \text{ and } t_s=1,...,T_s\right\}$. The $\mathbf{D}$ and $\mathbf{H}$ are assumed jointly independent, so that the log-posterior can be expressed as:

$$\log p\left(\mathbf{D},\mathbf{H}\middle\|\mathbf{Y}|^{.2},\boldsymbol{\Lambda}\right)=\log p\left(|\mathbf{Y}|^{.2}\,|\,\mathbf{D},\mathbf{H}\right)+\log p(\mathbf{D})+\log p\left(\mathbf{H}\,|\,\boldsymbol{\Lambda}\right)+\text{const} \tag{5.53}$$

Under the independent and identically distributed (i.i.d.) noise assumption, the minus log likelihood $-\log p\left(|\mathbf{Y}|^{.2}\,|\,\mathbf{D},\mathbf{H}\right)$ is defined as (5.39). In the proposed model, the prior over $\mathbf{D}$ is a factorial model where each τ^{th} slice of $\mathbf{D}$ is assumed to be zero-mean multivariate rectified Gaussian with covariance matrix $\boldsymbol{\Sigma}_{\tau}$ which can be expressed as:

$$\begin{aligned} p_D(\mathbf{D}) &= \prod_{\tau=0}^{\tau_{\max}} p_{D^{\tau}}(\mathbf{d}^{\tau}) \\ p_{D^{\tau}}(\mathbf{d}^{\tau}) &= \begin{cases} \dfrac{2}{\left(2\pi|\boldsymbol{\Sigma}_{\tau}|^2\right)^{(1/2FI)}}\exp\left(-\dfrac{1}{2}\mathbf{d}^{\tau\mathbf{T}}\boldsymbol{\Sigma}_{\tau}^{-1}\mathbf{d}^{\tau}\right), & \mathbf{d}^{\tau}\geq 0 \\ 0, & \mathbf{d}^{\tau}<0 \end{cases} \end{aligned} \tag{5.54}$$

where $\mathbf{d}^{\tau} = Vec\left(\mathbf{D}^{\tau}\right) = \left[\mathbf{D}_1^{\tau\mathbf{T}} \vdots \mathbf{D}_2^{\tau\mathbf{T}} \vdots \cdots \vdots \mathbf{D}_I^{\tau\mathbf{T}} \vdots\right]^{\mathbf{T}}$ and $\mathbf{\Sigma}_{\tau} = \begin{bmatrix} \mathbf{\Sigma}_{1,1,\tau} & \cdots & \mathbf{\Sigma}_{1,I,\tau} \\ \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_{I,1,\tau} & \cdots & \mathbf{\Sigma}_{I,I,\tau} \end{bmatrix}$ is the covariance matrix of $Vec\left(\mathbf{D}^{\tau}\right)$. Here $\mathbf{\Sigma}_{i,j,\tau} = E\left[\mathbf{D}_i^{\tau}\mathbf{D}_j^{\tau\mathbf{T}}\right]$ $\{i,j\} \in I$, '$E[\bullet]$' denotes the expectation. In the case of source separation, we can assume that $\mathbf{\Sigma}_{i,i,\tau}$ is large whereas $\mathbf{\Sigma}_{i,j,\tau}$ $i \neq j$ is small. Therefore, the inverse covariance matrix can be approximated as:

$$\begin{aligned} \mathbf{\Sigma}_{\tau}^{-1} &= \left(\mathbf{\Sigma}_{diag,\tau} + \mathbf{\Sigma}_{off,\tau}\right)^{-1} \\ &\approx \mathbf{\Sigma}_{diag,\tau}^{-1} - \mathbf{\Sigma}_{diag,\tau}^{-1}\mathbf{\Sigma}_{off,\tau}\mathbf{\Sigma}_{diag,\tau}^{-1} \\ &= \Omega_{diag,\tau} + \Omega_{off,\tau} \end{aligned} \tag{5.55}$$

where $\Omega_{diag,\tau} = \mathbf{\Sigma}_{diag,\tau}^{-1}$, $\Omega_{off,\tau} = -\mathbf{\Sigma}_{diag,\tau}^{-1}\mathbf{\Sigma}_{off,\tau}\mathbf{\Sigma}_{diag,\tau}^{-1}$ and

$$\mathbf{\Sigma}_{diag,\tau} = \begin{bmatrix} \mathbf{\Sigma}_{1,1,\tau} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_{2,2,\tau} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \ddots & \vdots \\ \mathbf{0} & \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{\Sigma}_{I,I,\tau} \end{bmatrix}, \quad \mathbf{\Sigma}_{off,\tau} = \begin{bmatrix} \mathbf{0} & \mathbf{\Sigma}_{1,2,\tau} & \cdots & \cdots & \mathbf{\Sigma}_{1,I,\tau} \\ \mathbf{\Sigma}_{2,1,\tau} & \mathbf{0} & \mathbf{\Sigma}_{2,3,\tau} & \cdots & \vdots \\ \vdots & \mathbf{\Sigma}_{3,2,\tau} & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \mathbf{\Sigma}_{I-1,I,\tau} \\ \mathbf{\Sigma}_{I,1,\tau} & \cdots & \cdots & \mathbf{\Sigma}_{I,I-1,\tau} & \mathbf{0} \end{bmatrix}$$

In above, $\mathbf{0} = \begin{bmatrix} 0_{11} & \cdots & 0_{1F} \\ \vdots & \ddots & \vdots \\ 0_{F1} & \cdots & 0_{FF} \end{bmatrix}$ is a $F \times F$ matrix with zero elements and $\mathbf{\Sigma}_{diag,\tau}^{-1}$ is the inverse covariance matrix of $\mathbf{\Sigma}_{diag,\tau}$. Thus, the $(i,j)^{\text{th}}$ sub matrix of $\Omega_{off,\tau}$ is given by $\Omega_{off,i,j,\tau}$ which measure the correlation between the different basis vectors. Here, we propose that each sub matrix $\Omega_{off,i,j,\tau}$ is constrained to $\mu_{ij\tau}\mathbf{I}$, where $\mathbf{I}$ is identity matrix and $\mu_{ij\tau}$ is a scalar. The goal is to simplify the process for the end user to exercise control over the correlation between the different basis vectors by using $u_{i,j,\tau}$. Thus, we may cast the (5.55) into two parts as:

$$\begin{aligned} -\log p_{D^{\tau}}(\mathbf{d}^{\tau}) &\approx \frac{1}{2}Vec\left(\mathbf{D}^{\tau}\right)^{\mathbf{T}}\Omega_{diag,\tau}Vec\left(\mathbf{D}^{\tau}\right) + \frac{1}{2}Vec\left(\mathbf{D}^{\tau}\right)^{\mathbf{T}}\Omega_{off,\tau}Vec\left(\mathbf{D}^{\tau}\right) \\ &= \gamma + \frac{1}{2}Vec\left(\mathbf{D}^{\tau}\right)^{\mathbf{T}}\Omega_{off,\tau}Vec\left(\mathbf{D}^{\tau}\right) \end{aligned} \tag{5.56}$$

The term $\gamma = \frac{1}{2} Vec\left(\mathbf{D}^{\tau}\right)^{\mathbf{T}} \Omega_{diag,\tau} Vec\left(\mathbf{D}^{\tau}\right)$ relates only to the power of $\mathbf{D}^{\tau}$. On the other hand, the desired constraint lies in the second term of (5.56) and since $\Omega_{off,\tau}$ is an off-diagonal matrix, then $Vec\left(\mathbf{D}^{\tau}\right)^{\mathbf{T}} \Omega_{off,\tau} Vec\left(\mathbf{D}^{\tau}\right)$ simply reduces to $\sum_{i,j,(i\neq j)} \mu_{ij\tau} \mathbf{D}_i^{\tau^{\mathrm{T}}} \mathbf{D}_j^{\tau}$. Thus, with the factorial model in (5.56), the desired constraint may assume the following form:

$$f(\mathbf{D}) = -\sum_{\tau=0}^{\tau_{\max}} \log p_{D^{\tau}}(\mathbf{d}^{\tau}) \approx \sum_{\substack{i,j \\ (i\neq j)}} \sum_{\tau} \mu_{ij\tau} \mathbf{D}_i^{\tau^{\mathrm{T}}} \mathbf{D}_j^{\tau} \tag{5.57}$$

where $\mu_{ij\tau}$ is a scalar that determines the importance of the prior over $\mathbf{D}$.

In the proposed prior model, no explicit constraint is imposed on the correlation between any two elements in the same basis vector so that the spectral basis can learn directly from the data. Since each element of $\mathbf{D}_i^{\tau}$ represent part of a feature, it is not necessary to add any constraints to $\mathbf{\Sigma}_{i,i,\tau}$. On the other hand, we may not be able to extract the underlying features correctly from the data if $\mathbf{\Sigma}_{i,i,\tau}$ is constrained to have a certain structure. This is because $\mathbf{\Sigma}_{i,i,\tau}$ represents the covariance matrix of $\mathbf{D}_i^{\tau}$. Hence, when the covariance matrix is constrained, $\mathbf{D}_i^{\tau}$ will be biased accordingly and therefore, part of the feature will not be efficiently extracted. In this paper, the probabilistic framework is used for the purpose of developing a platform to incorporate the statistical correlation between $\mathbf{D}_i^{\tau}$ and $\mathbf{D}_j^{\tau}$ into the matrix factorization as part of the regularization. In source separation, such constraint is required for the basis to be fully expressible (i.e. fully recovered) especially in situation where the patterns overlap each other. Despite the proposed prior model for $\mathbf{D}$ stems from the rectified Gaussian distribution, it is actually a combination of constrained and unconstrained parameterization of the inverse covariance matrix as noted by $\Omega_{diag,\tau}$ and $\Omega_{off,\tau}$. In Sections 5.3.4, we will verify and demonstrate that this prior model works

efficiently for source separation. In the third term of (5.51), each element of $\mathbf{H}$ is constrained to be exponential distributed with independent decay parameters λ_{i,t_s}^{ϕ}, namely:

$$p(\mathbf{H} \mid \mathbf{\Lambda}) = \prod_{\phi} \prod_{i} \prod_{t_s} \lambda_{i,t_s}^{\phi} \exp(-\lambda_{i,t_s}^{\phi} \mathbf{H}_{i,t_s}^{\phi}) \tag{5.58}$$

with $-\log p(\mathbf{H} \mid \mathbf{\Lambda}) = \sum_{\phi,i,t_s} \lambda_{i,t_s}^{\phi} \mathbf{H}_{i,t_s}^{\phi} - \sum_{i,t_s,\phi} \log \lambda_{i,t_s}^{\phi}$ so that $f(\mathbf{H}) = \sum_{\phi,i,t_s} \lambda_{i,t_s}^{\phi} \mathbf{H}_{i,t_s}^{\phi}$. Thus, the goal is to find spectral basis $\mathbf{D}^{\tau}$ and temporal sparse code $\mathbf{H}^{\phi}$. By substituting (5.57) and (5.58) into (5.53), the following cost function L_{IS}^{v} can be expressed as:

$$\begin{aligned} L_{IS}^{v} &\propto \sum_{f,t_s} \left[\frac{|\mathbf{Y}|_{f,t_s}^2}{\mathbf{Z}_{f,t_s}} - \log\left(\frac{|\mathbf{Y}|_{f,t_s}^2}{\mathbf{Z}_{f,t_s}} \right) - 1 \right] + f(\mathbf{H}) - \sum_{i,t_s,\phi} \log \lambda_{i,t_s}^{\phi} + f(\mathbf{D}) \\ &\propto \sum_{f,t_s} \left[\frac{|\mathbf{Y}|_{f,t_s}^2}{\mathbf{Z}_{f,t_s}} - \log\left(\frac{|\mathbf{Y}|_{f,t_s}^2}{\mathbf{Z}_{f,t_s}} \right) - 1 \right] + \sum_{\substack{i,j \\ (i \neq j)}} \sum_{\tau} \mu_{ij\tau} \mathbf{D}_i^{\tau \mathrm{T}} \mathbf{D}_j^{\tau} + f(\mathbf{H}) - \sum_{i,t_s,\phi} \log \lambda_{i,t_s}^{\phi} \end{aligned} \tag{5.59}$$

where $\mathbf{Z} = \sum_{i} \sum_{\tau} \sum_{\phi} \mathbf{D}_i^{\tau \downarrow \phi} \mathbf{H}_i^{\phi \rightarrow \tau}$ and $f(\mathbf{H}) = \sum_{\phi,i,t_s} \lambda_{i,t_s}^{\phi} \mathbf{H}_{i,t_s}^{\phi}$. The sparsity term $f(\mathbf{H})$ forms the L_1-norm regularization to resolve the ambiguity by forcing all structure in $\mathbf{H}$ onto $\mathbf{D}$.Therefore, the sparseness of the solution is highly dependent on the regularization parameters λ_{i,t_s}^{ϕ}.

5.2.3.1 Estimation of the spectral basis and temporal code (IS-vRNMF2D)

Using above, the derivatives of (5.59) corresponding to $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ are given by:

$$\begin{aligned} \frac{\partial L_{IS}^{v}}{\partial \mathbf{D}_{f',i'}^{\tau'}} &= \sum_{f,t_s} \left(\left(\mathbf{Z}_{f,t_s} \right)^{-2} \left(\mathbf{Z}_{f,t_s} - |\mathbf{Y}|_{f,t_s}^2 \right) \right) \mathbf{H}_{i',t_s-\tau'}^{f-f'} + \sum_{j \neq i'} \mu_{i'j\tau'} \mathbf{D}_{f',j}^{\tau'} \\ &= -\sum_{\phi,t_s} \left(\left(\mathbf{Z}_{f'+\phi,t_s} \right)^{-2} \left(|\mathbf{Y}|_{f'+\phi,t_s}^2 - \mathbf{Z}_{f'+\phi,t_s} \right) \right) \mathbf{H}_{i',t_s-\tau'}^{\phi} + \sum_{j \neq i'} \mu_{i'j\tau'} \mathbf{D}_{f',j}^{\tau'} \end{aligned} \tag{5.60}$$

Similarly:

$$\begin{aligned}\frac{\partial L_{IS}^{v}}{\partial \mathbf{H}_{i',t'}^{\phi'}} &= \sum_{f,t_s} \mathbf{D}_{f-\phi',i'}^{t_s-t_s'} \left(\left(\mathbf{Z}_{f,t_s} \right)^{-2} \left(\mathbf{Z}_{f,t_s} - \left| \mathbf{Y} \right|_{f,t_s}^{2} \right) \right) + \lambda_{i',t_s'}^{\phi'} \\ &= -\sum_{\tau,f} \mathbf{D}_{f-\phi',i'}^{\tau} \left(\left(\mathbf{Z}_{f,t_s'+\tau} \right)^{-2} \left(\left| \mathbf{Y} \right|_{f,t_s'+\tau}^{2} - \mathbf{Z}_{f,t_s'+\tau} \right) \right) + \lambda_{i',t_s'}^{\phi'}\end{aligned} \tag{5.61}$$

Consequently, by applying the standard gradient decent approach, namely:

$$\mathbf{D}_{f',i'}^{\tau'} \leftarrow \mathbf{D}_{f',i'}^{\tau'} - \eta_D \frac{\partial L_{IS}^{v}}{\partial \mathbf{D}_{f',i'}^{\tau'}} \quad \text{and} \quad \mathbf{H}_{i',t_s'}^{\phi'} \leftarrow \mathbf{H}_{i',t_s'}^{\phi'} - \eta_H \frac{\partial L_{IS}^{v}}{\partial \mathbf{H}_{i',t_s'}^{\phi'}} \tag{5.62}$$

The term η_D and η_H are positive learning rates which can be derived using the approach of Lee and Seung [74] as:

$$\eta_D = \frac{\mathbf{D}_{f',i'}^{\tau'}}{\sum_{\phi,t_s} \left(\mathbf{Z}_{f'+\phi,t_s} \right)^{-1} \mathbf{H}_{i',t_s-\tau'}^{\phi} + \sum_{j \neq i'} \mu_{i'j\tau'} \mathbf{D}_{f',j}^{\tau'}} \quad \text{and} \quad \eta_H = \frac{\mathbf{H}_{i',t_s'}^{\phi'}}{\sum_{\tau,f} \mathbf{D}_{f-\phi',i'}^{\tau} \left(\mathbf{Z}_{f,t_s'+\tau} \right)^{-1} + \lambda_{i',t_s'}^{\phi'}} \tag{5.63}$$

Inserting (5.63) into (5.62) leads to the multiplicative update rules:

$$\begin{aligned}\mathbf{D}_{f',i'}^{\tau'} &\leftarrow \mathbf{D}_{f',i'}^{\tau'} - \frac{-\mathbf{D}_{f',i'}^{\tau'} \left(\sum_{\phi,t_s} \left(\left(\mathbf{Z}_{f'+\phi,t_s} \right)^{-2} \left(\left| \mathbf{Y} \right|_{f'+\phi,t_s}^{2} - \mathbf{Z}_{f'+\phi,t_s} \right) \right) \mathbf{H}_{i',t_s-\tau'}^{\phi} + \sum_{j \neq i'} \mu_{i'j\tau'} \mathbf{D}_{f',j}^{\tau'} \right)}{\sum_{\phi,t_s} \left(\mathbf{Z}_{f'+\phi,t_s} \right)^{-1} \mathbf{H}_{i',t_s-\tau'}^{\phi} + \sum_{j \neq i'} \mu_{i'j\tau'} \mathbf{D}_{f',j}^{\tau'}} \\ &= \mathbf{D}_{f',i'}^{\tau'} \frac{\sum_{\phi,t_s} \left(\mathbf{Z}_{f'+\phi,t_s} \right)^{-2} \left| \mathbf{Y} \right|_{f'+\phi,t_s}^{2} \mathbf{H}_{i',t_s-\tau'}^{\phi}}{\sum_{\phi,t_s} \left(\mathbf{Z}_{f'+\phi,t_s} \right)^{-1} \mathbf{H}_{i',t_s-\tau'}^{\phi} + \sum_{j \neq i'} \mu_{i'j\tau'} \mathbf{D}_{f',j}^{\tau'}}\end{aligned} \tag{5.64}$$

Similarly, the update rules for $\mathbf{H}_{i',t_s'}^{\phi'}$ gives:

$$\begin{aligned}\mathbf{H}_{i',t_s'}^{\phi'} &\leftarrow \mathbf{H}_{i',t_s'}^{\phi'} - \frac{-\mathbf{H}_{i',t_s'}^{\phi'} \left(\sum_{\tau,f} \mathbf{D}_{f-\phi',i'}^{\tau} \left(\left(\mathbf{Z}_{f,t_s'+\tau} \right)^{-2} \left(\left| \mathbf{Y} \right|_{f,t_s'+\tau}^{2} - \mathbf{Z}_{f,t_s'+\tau} \right) \right) + \lambda_{i',t_s'}^{\phi'} \right)}{\sum_{\tau,f} \mathbf{D}_{f-\phi',i'}^{\tau} \left(\mathbf{Z}_{f,t_s'+\tau} \right)^{-1} + \lambda_{i',t_s'}^{\phi'}} \\ &= \mathbf{H}_{i',t_s'}^{\phi'} \frac{\sum_{\tau,f} \mathbf{D}_{f-\phi',i'}^{\tau} \left(\left(\mathbf{Z}_{f,t_s'+\tau} \right)^{-2} \left(\left| \mathbf{Y} \right|_{f,t_s'+\tau}^{2} \right) \right)}{\sum_{\tau,f} \mathbf{D}_{f-\phi',i'}^{\tau} \left(\mathbf{Z}_{f,t_s'+\tau} \right)^{-1} + \lambda_{i',t_s'}^{\phi'}}\end{aligned} \tag{5.65}$$

The update of $\boldsymbol{\Lambda}$ follows by solving $\frac{\partial L_{IS}^{v}}{\partial \lambda_{i',t_s'}^{\phi'}} = 0$.

$$\frac{\partial L_{IS}^{v}}{\partial \lambda_{i',t_s'}^{\phi'}} = \mathbf{H}_{i',t_s'}^{\phi'} - \frac{1}{\lambda_{i',t_s'}^{\phi'}} \tag{5.66}$$

$$\lambda_{i',t_s'}^{\phi'} = \frac{1}{\mathbf{H}_{i',t_s'}^{\phi'}} \quad \text{where '} \frac{a}{b} \text{' is element wise divide} \tag{5.67}$$

In terms of matrix notation, the multiplicative learning rules in (5.64), (5.65) and (5.67) can be written as:

$$\mathbf{D}^{\tau} \leftarrow \mathbf{D}^{\tau} \bullet \frac{\sum_{\phi} \left(\left(\overset{\uparrow\phi}{\mathbf{Z}} \right)^{\cdot -2} \bullet \left| \overset{\uparrow\phi}{\mathbf{Y}} \right|^{\cdot 2} \right) \overset{\rightarrow\tau^{\mathbf{T}}}{\mathbf{H}^{\phi}}}{\sum_{\phi} \left(\overset{\uparrow\phi}{\mathbf{Z}} \right)^{\cdot -1} \overset{\rightarrow\tau^{\mathbf{T}}}{\mathbf{H}^{\phi}} + \mathbf{D}^{\tau} \boldsymbol{\Xi}^{\tau \mathbf{T}}} \quad \text{and} \quad \mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi} \bullet \frac{\sum_{\tau} \overset{\downarrow\phi^{\mathbf{T}}}{\mathbf{D}^{\tau}} \left(\left(\overset{\leftarrow\tau}{\mathbf{Z}} \right)^{\cdot -2} \bullet \left| \overset{\leftarrow\tau}{\mathbf{Y}} \right|^{\cdot 2} \right)}{\sum_{\tau} \overset{\downarrow\phi^{\mathbf{T}}}{\mathbf{D}^{\tau}} \left(\overset{\leftarrow\tau}{\mathbf{Z}} \right)^{\cdot -1} + \boldsymbol{\Lambda}^{\phi}} \tag{5.68}$$

where $\boldsymbol{\Xi}^{\tau}$ is a $I \times I$ matrix whose $(i,j)^{\text{th}}$ element is given by $\mu_{ij\tau}$ except the diagonal elements being zeros. In (5.68), $\boldsymbol{\Lambda}^{\phi}$ is the matrix representation of λ_{i,t_s}^{ϕ} which is adaptive according to (5.67) and the parameter $\mu_{ij\tau}$ in $\boldsymbol{\Xi}^{\tau}$ is non-adaptive which can be selected manually depending on applications. The above algorithm is termed as the variable regularised two-dimensional nonnegative matrix factorization with IS divergence (IS-vRNMF2D). Table 5.5 summarises the basic steps of the proposed IS-vRNMF2D algorithm.

Table 5.5: IS-vRNMF2D algorithm

1. Initialize $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ with nonnegative random values
2. Initialize $\mu_{ij\tau}$ with positive values.
3. Caculate $\mathbf{Z} = \sum_i \sum_{\tau} \sum_{\phi} \mathbf{D}_i^{\tau\downarrow\phi} \mathbf{H}_i^{\phi\rightarrow\tau}$
4. $\mathbf{H}^{\phi} \leftarrow \mathbf{H}^{\phi} \bullet \dfrac{\sum_{\tau} \mathbf{D}^{\tau\downarrow\phi\mathbf{T}} \left(\left(\overset{\leftarrow\tau}{\mathbf{Z}} \right)^{\cdot -2} \bullet \left| \overset{\leftarrow\tau}{\mathbf{Y}} \right|^{\cdot 2} \right)}{\sum_{\tau} \mathbf{D}^{\tau\downarrow\phi\mathbf{T}} \left(\overset{\leftarrow\tau}{\mathbf{Z}} \right)^{\cdot -1} + \boldsymbol{\Lambda}^{\phi}}$
5. $\boldsymbol{\Lambda}^{\phi} = \dfrac{1}{\mathbf{H}^{\phi}}$.
6. Caculate $\mathbf{Z} = \sum_i \sum_{\tau} \sum_{\phi} \mathbf{D}_i^{\tau\downarrow\phi} \mathbf{H}_i^{\phi\rightarrow\tau}$
7. $\mathbf{D}^{\tau} \leftarrow \mathbf{D}^{\tau} \bullet \dfrac{\sum_{\phi} \left(\left(\overset{\uparrow\phi}{\mathbf{Z}} \right)^{\cdot -2} \bullet \left| \overset{\uparrow\phi}{\mathbf{Y}} \right|^{\cdot 2} \right) \mathbf{H}^{\phi\rightarrow\tau\mathbf{T}}}{\sum_{\phi} \left(\overset{\uparrow\phi}{\mathbf{Z}} \right)^{\cdot -1} \mathbf{H}^{\phi\rightarrow\tau\mathbf{T}} + \mathbf{D}^{\tau} \boldsymbol{\Xi}^{\tau\mathbf{T}}}$
8. Repeat from step 3 until convergence.

5.2.4 Summary of the proposed algorithms

In Section 5.2, a novel family of IS divergence based two-dimensional nonnegative matrix factorization methods to solve SCBSS has been proposed. These include (i). Quasi-EM based NMF2D with IS divergence (Quasi-EM IS-NMF2D). (ii). Multiplicative update based NMF2D with IS divergence (MU IS-NMF2D). (iii). Multiplicative update based SNMF2D with IS divergence (IS-SNMF2D). (iv). Multiplicative update based variable regularised NMF2D with IS divergence (IS-vRNMF2D). The Table 5.6 and Figure 5.3 summarise the proposed IS divergence based NMF2D algorithms.

Table 5.6: Summary of the proposed IS divergence based NMF2D algorithms

Methods	**Cost function**	**Regularization**		**Update method**
		D	**H**	
Quasi-EM IS-NMF2D	ISD	-	-	Quasi-EM
MU IS-NMF2D				MU
IS-SNMF2D	ISD	-	Uniform constant sparsity	MU
IS-vRNMF2D	ISD	Correlation of the basis	Adaptive sparsity	MU

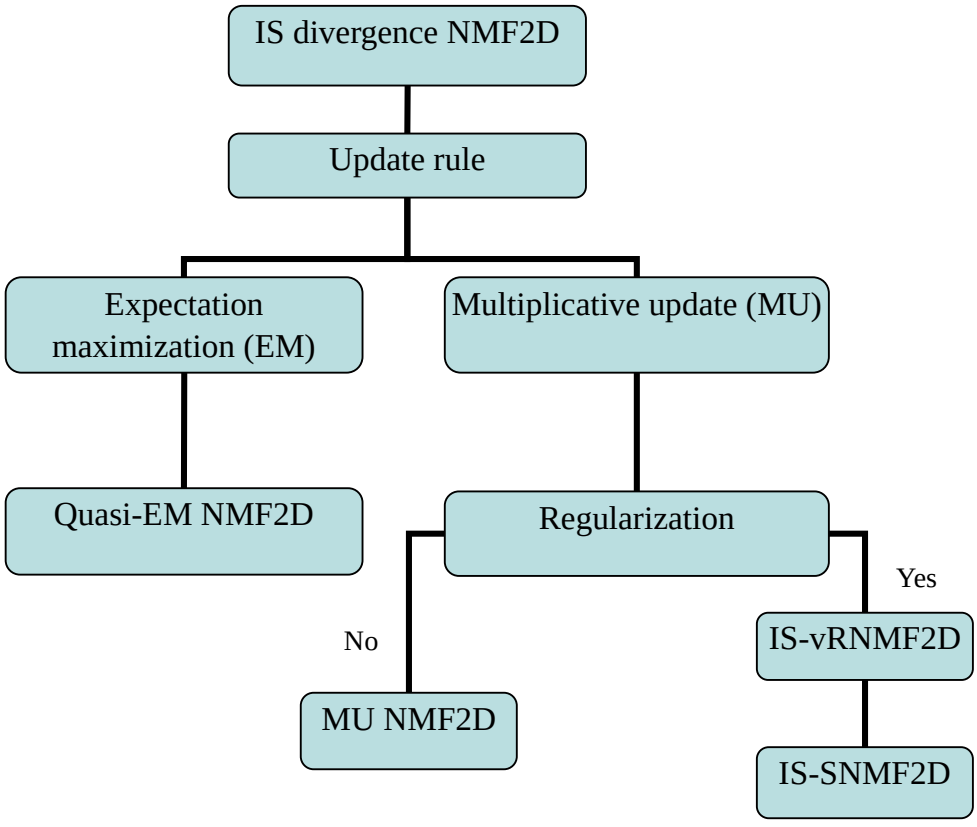


Figure 5.3: The flow chart of the proposed IS divergence based NMF2D algorithms.

5.2.5 Estimation of sources

The binary mask $\mathbf{Mask}_i$ is generated same as Section 3.3.1 and finally, the estimated time-domain signals are obtained as:

$$\tilde{\mathbf{x}}_i = \text{Resynthesize}(\mathbf{Mask}_i \bullet \mathbf{Y}) \tag{5.69}$$

for $i = 1,2$ where $\tilde{\mathbf{x}}_i = [\tilde{x}_i(1),\ldots,\tilde{x}_i(T)]^{\mathbf{T}}$ denotes the i^{th} estimated source. The time-domain estimated sources are re-synthesized using the approach in [120] from the mixture by weighting the mixture cochleagram by the mask and correcting phase shifts introduced during the gammatone filtering.

5.3 Experimental Results and Analysis

The proposed monaural source separation method is tested on recorded audio signals. Several experimental studies have been designed to investigate the efficacy of the proposed approach. For mixture generation, two sentences of the target speakers (male and female) 'fcjf0' and 'mcpm0', were selected from TIMIT speech database and the others including jazz and piano music. All mixtures are sampled at 16 kHz sampling rate. In all cases, the sources are mixed with equal average power over the duration of the signals. In this section, two types of mixtures are used: mixture of music and speech; mixture of different kinds of music. As for the proposed family IS divergence based two-dimensional matrix factorization algorithms, the convolutive components are selected as follows: (i) For jazz and speech mixture, $\tau = \{0,\ldots,4\}$ and $\phi = \{0,\ldots,4\}$. (ii) For jazz and piano mixture, $\tau = \{0,\ldots,6\}$ and $\phi = \{0,\ldots,9\}$. (iii) For piano and speech mixture, $\tau = \{0,\ldots,6\}$ and

$\phi = \{0,\ldots,9\}$. These parameters are selected after conducting the Monte-Carlo simulation over many different realizations of audio mixture. The measure of distortion between the original source and the estimated one is computed by using the SDR, SAR and SIR.

5.3.1 Effects on separation based on different TF representation

In this section, the performance of our proposed Quasi-EM IS-NMF2D algorithm is evaluated by using three types of TF representation: (i) spectrogram (STFT with a 1024-point Hamming windowed FFT and 50% overlap), (ii) log-frequency spectrogram (as described in section 5.1 with 1024-point Hamming window). and (iii) cochleagram based on Gammatone filterbank of 128 channels, filter order of 4 (i.e. $h = 4$ in Eqn.(5.2)), and the output is divided into 20-ms time frame with 50% overlap between consecutive frames. Speech signals and music are used to generate the monoaural mixture. In the following, we show that the separation results based on the cochleagram is significantly more effective than other TF domain. Table 5.7 shows the comparison of the proposed algorithm (quasi-EM IS-NMF2D) based on the spectrogram, log-frequency spectrogram and cochleagram under various audio mixtures.

Table 5.7: Separation results based on different TF representation

Mixtures	TF methods	SDR	SAR	SIR
jazz music and male speech	spectrogram	3.47	6.57	4.86
	log-frequency spectrogram	6.54	9.51	10.53
	cochleagram	**8.87**	**10.31**	**12.62**
jazz music and female speech	spectrogram	-1.41	5.87	0.14
	log-frequency spectrogram	3.97	9.48	6.17
	cochleagram	**9.34**	**9.77**	**14.37**
piano music and male speech	spectrogram	2.10	4.34	6.23
	log-frequency spectrogram	2.31	5.42	6.64
	cochleagram	**7.16**	**8.56**	**12.08**
piano music and female speech	spectrogram	-1.01	5.15	1.13
	log-frequency spectrogram	0.27	8.01	2.25
	cochleagram	**7.44**	**9.18**	**11.38**
jazz music and piano music	spectrogram	-0.59	6.34	0.97
	log-frequency spectrogram	1.21	6.42	4.89
	cochleagram	**7.21**	**13.07**	**8.68**

The separation results for all mixture types based on the spectrogram gives an average SDR of 0.51dB while the log-frequency spectrogram gives an average SDR of 2.8dB. However, a significantly higher performance is attained by the cochleagram with an average SDR of 8dB which leads to a substantial gain improvement of 7.5dB and 5.2dB, respectively. The major reason for the large discrepancy between them is in the mixing ambiguity between $|\mathbf{X}_1|^{.2}$ and $|\mathbf{X}_2|^{.2}$ in the TF domain. The larger the mixing ambiguity between $|\mathbf{X}_1|^{.2}$ and $|\mathbf{X}_2|^{.2}$, the more numerous TF units will be ambiguous which subsequently decreases the possibility of correct assignment of each unit to the sources. This inadvertently results in poorer performance of source separation. Figure 5.4 shows the spectrogram of the original sources, the mixed signal, and the estimated sources using the proposed Quasi-EM IS-NMF2D algorithm. The spectral overlapping between the two sources has resulted in mixing ambiguity in the time-domain as highlighted with red box

marked area in the last two panels of Figure 5.5.

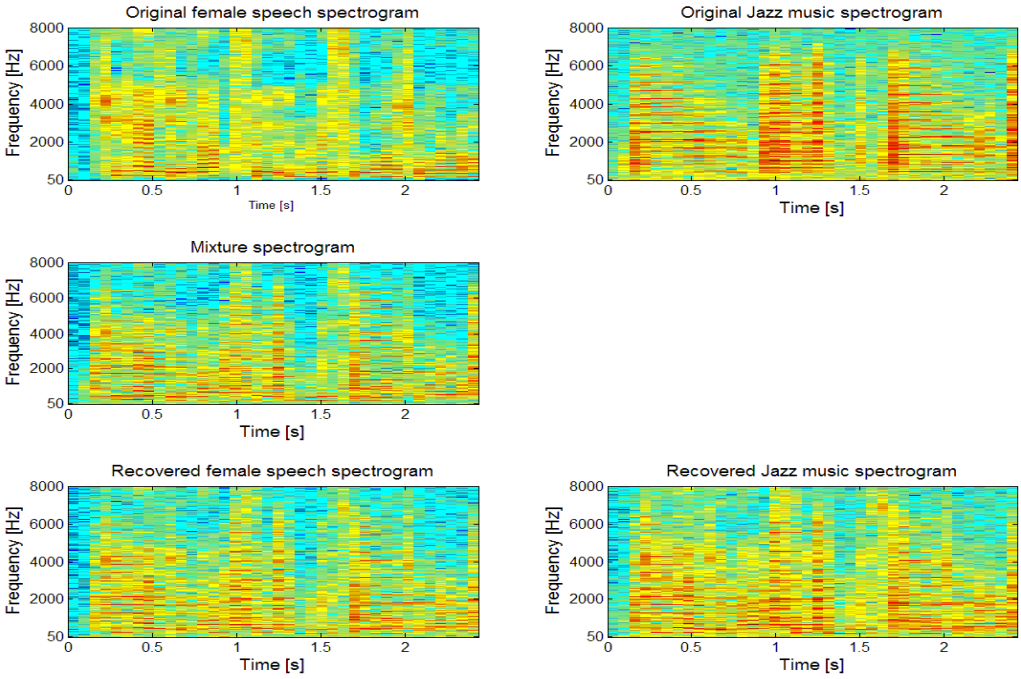


Figure 5.4: Separation results in spectrogram. Top panel: Spectrogram of the original sources. Middle panel: Spectrogram of the mixture. Bottom panel: Spectrogram of the estimated sources.

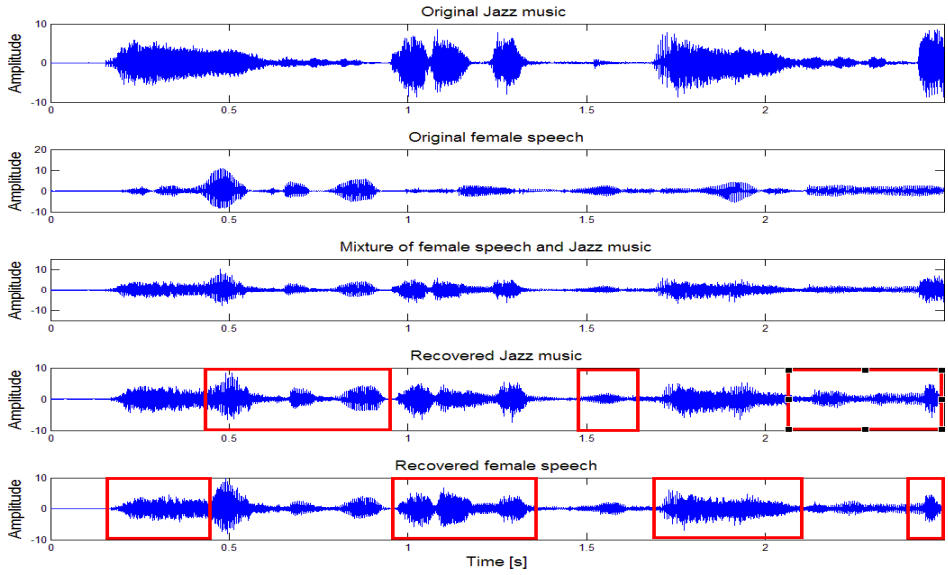


Figure 5.5: Time-domain separated results based on spectrogram.

Figures 5.4 and 5.5 substantiate the fact that STFT lacks provision for further low-level

information about a particular TF unit and therefore, the resulting spectrogram fails to infer the dominating source. This leads to high degree of ambiguity in TF domain and causes lack of uniqueness in extracting the spectral-temporal features of the sources. Figures 5.6 and 5.7 show the separation results based on log-frequency spectrogram. Comparing with spectrogram, the separation performance is better since log-frequency spectrogram has the prosperity of non-uniform time frequency resolution. However, according to the analysis of separability in Section 5.1.4, the transform used by the log-frequency spectrogram is still not be an optimal option for audio source separation. The spectral overlapping based on log-frequency spectrogram between the two sources has resulted in mixing ambiguity in the time-domain as highlighted with red box marked area in the last two panels of Figure 5.8.

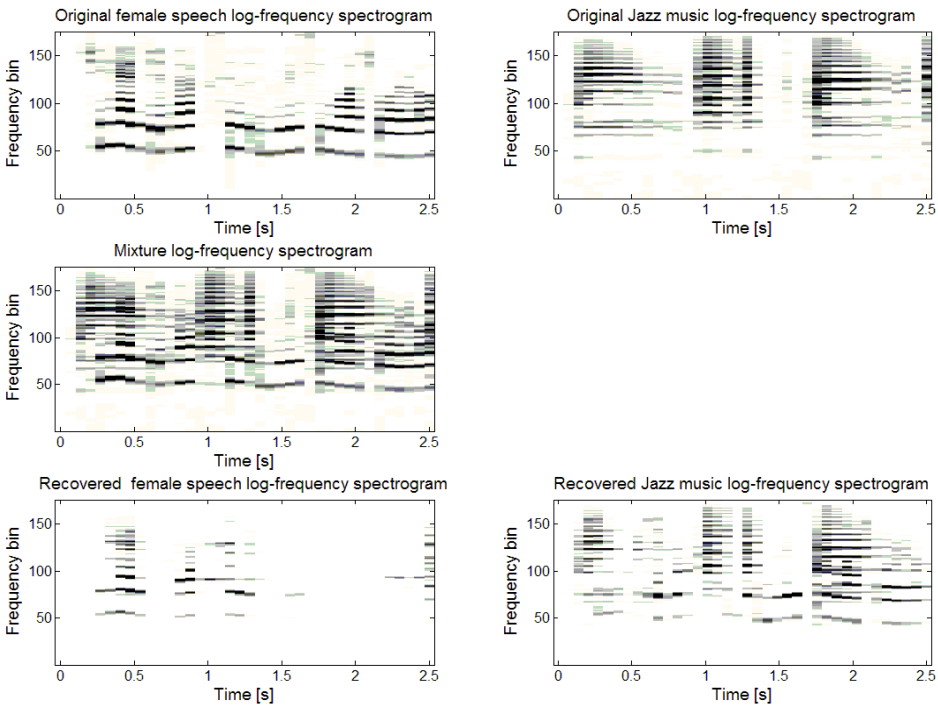


Figure 5.6: Separation results in log- frequency spectrogram. Top panel: Log-frequency spectrogram of the original sources. Middle panel: Log-frequency spectrogram of the mixture. Bottom panel: Log-frequency spectrogram of the estimated sources.

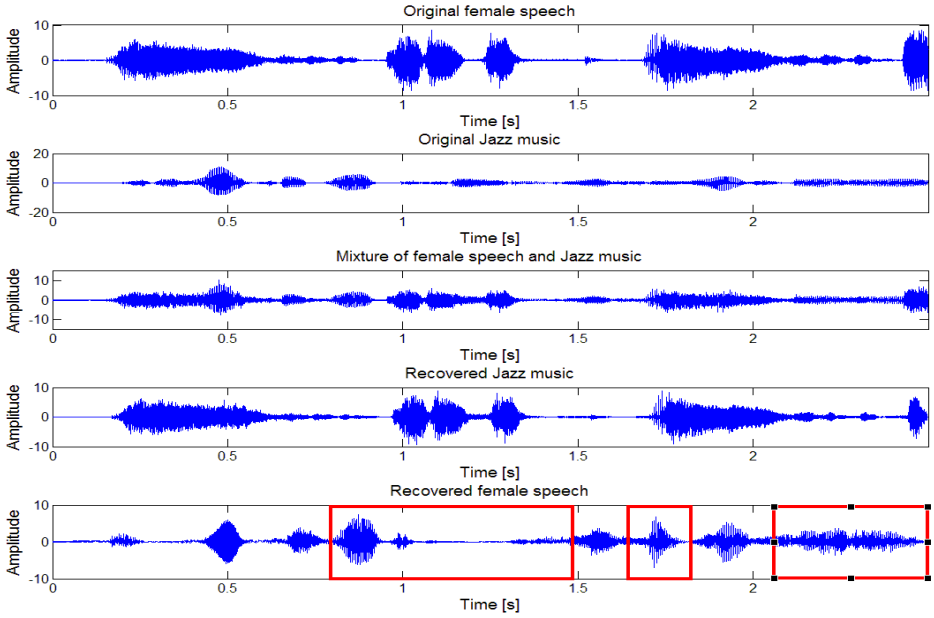


Figure 5.7: Time-domain separated results based on log-frequency spectrogram.

On the other hand, the results of separation in the cochleagram have led to significant SDR improvement. The cochleagram enables the mixed signal to be more separable and thereby reduces the mixing ambiguity between $|\mathbf{X}_1|^{.2}$ and $|\mathbf{X}_2|^{.2}$. This explains the average performance of separating mixture jazz music and female utterance is highest among all the mixtures because both sources have very distinguishable TF patterns in the cochleagram. Figure 5.8 further shows the separation results in the cochleagram. The plot clearly shows the spectral energy of the two audio sources is clustered at different frequencies in the cochleagram due to their different fundamental frequencies. These prominent features have been separated using the proposed Quasi-EM IS-NMF2D algorithm. Figure 5.9 shows the final recovered time-domain sources.

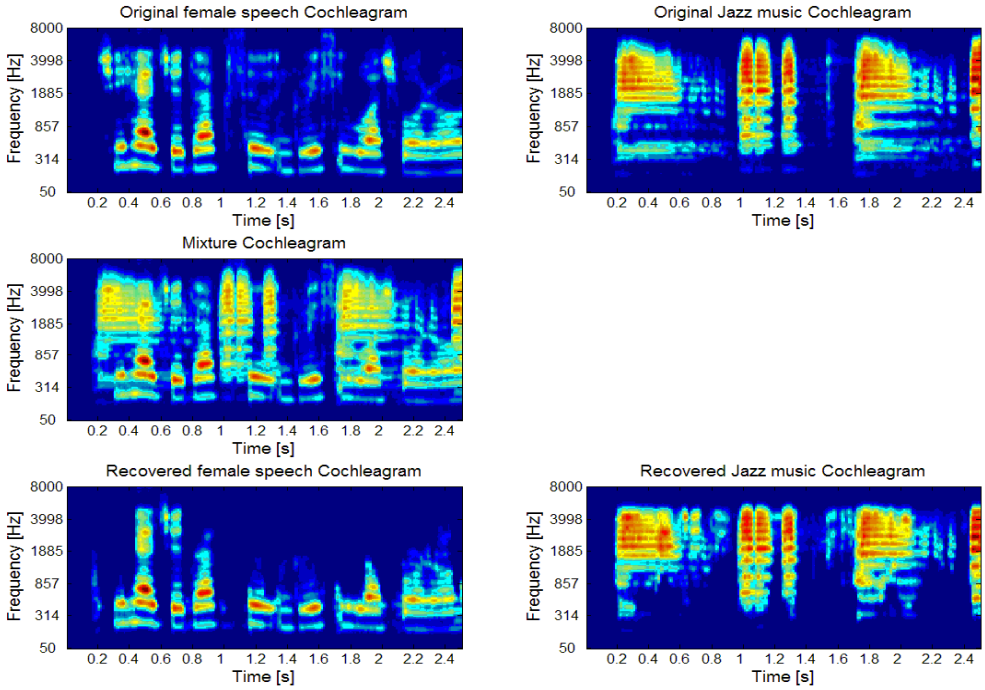


Figure 5.8: Separation results in cochleagram. Top panel: Cochleagram of the original sources. Middle panel: Cochleagram of the mixture. Bottom panel: Cochleagram of the estimated sources.

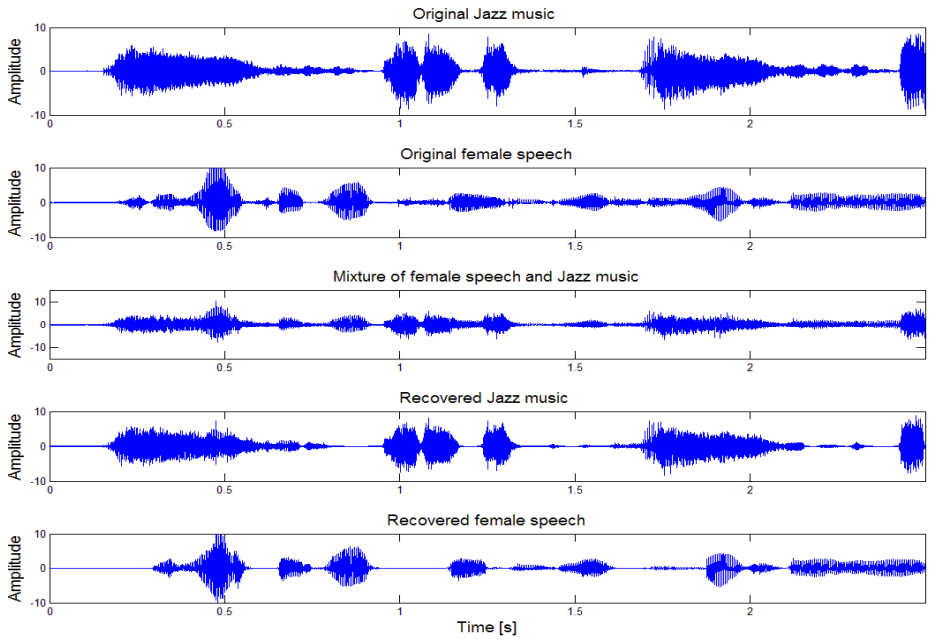


Figure 5.9: Time-domain separated results using the proposed algorithm.

In the proposed method, the performance of source separation depends to an extent on how distinguishable the two spectral bases $\mathbf{D}_1^{\tau}$ and $\mathbf{D}_2^{\tau}$ are from each other. When $\mathbf{D}_1^{\tau}$

and $\mathbf{D}_2^{\tau}$ are distinguishable from each other and since $\left\{\mathbf{H}_i^{\phi}\right\}_{i=1}^{2}$ are sparse, it follows that the mixing ambiguity between $\left|\mathbf{X}_1\right|^{.2}$ and $\left|\mathbf{X}_2\right|^{.2}$ which constitutes the magnitude of interference in the TF domain will be small. Thus, by exploiting the sparse property of $\left\{\mathbf{H}_i^{\phi}\right\}_{i=1}^{2}$, it is possible to determine $\left|\mathbf{X}_1\right|^{.2}$ and $\left|\mathbf{X}_2\right|^{.2}$ from $\left|\mathbf{Y}\right|^{.2}$ provided that $\mathbf{D}_1^{\tau}$ and $\mathbf{D}_2^{\tau}$ are sufficiently distinguishable. Figure 5.10 shows the results of $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ for the above mixture (mixing between female utterance and jazz music) when the factorization is obtained in the cochleagram. In Figure 5.10, panels (A)-(B) refer to $\mathbf{D}_1^{\tau}$ and $\mathbf{D}_2^{\tau}$ which are the estimated spectral bases of jazz music and female utterance, respectively. Panels (C)-(D) refer to $\mathbf{H}_1^{\phi}$ and $\mathbf{H}_2^{\phi}$ which correspond to the estimated temporal code (i.e. time pitch signature) of jazz music and female utterance, respectively. In comparison, the results of $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ have also been included when factorizing the same mixture in the spectrogram and log-frequency spectrogram. These are shown in Figure 5.11 and 5.12, respectively.

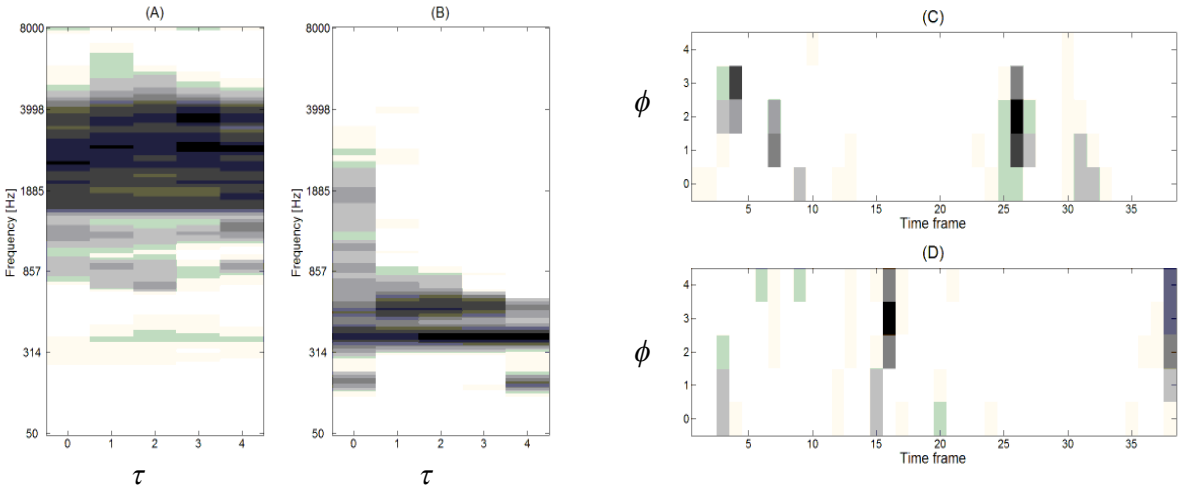


Figure 5.10: Estimated $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ using the proposed algorithm based on cochleagram.

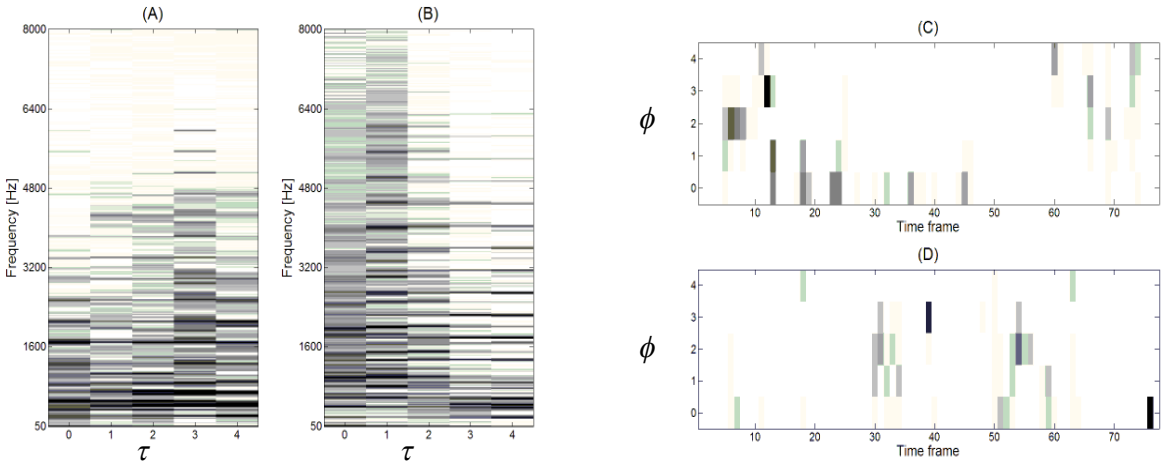


Figure 5.11: Estimated $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ using the proposed algorithm based on spectrogram.

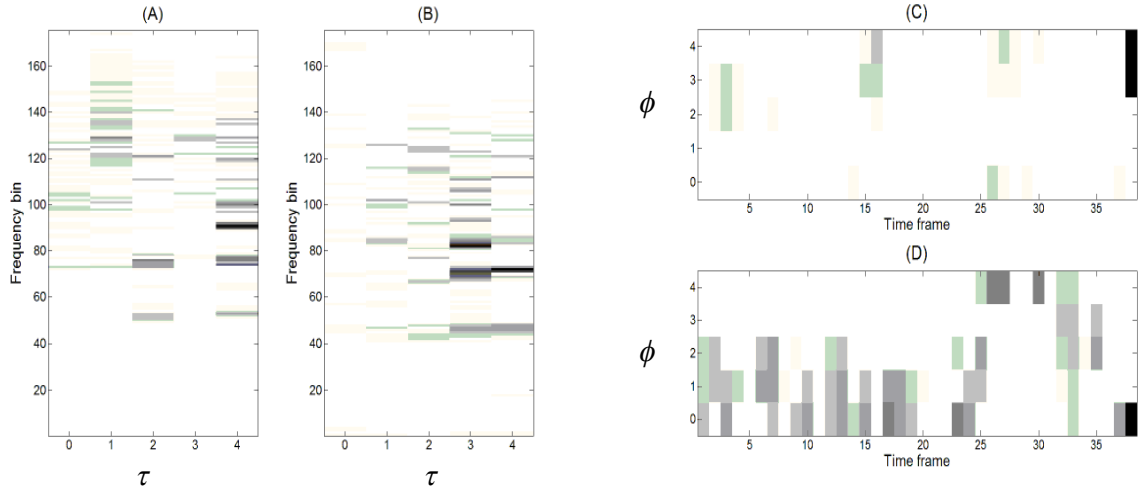


Figure 5.12: Estimated $\mathbf{D}_i^{\tau}$ and $\mathbf{H}_i^{\phi}$ using the proposed algorithm based on log-frequency spectrogram.

In sharp contrast with Figure 5.10, Figures 5.11 amd 5.12 show overlap in the spectral bases between $\mathbf{D}_1^{\tau}$ and $\mathbf{D}_2^{\tau}$. The cochleagram based spectral bases estimation shows less overlap among the all. Hence, the recovered sources are much better as noted by the very high values of SDR in Table 5.7.

5.3.2 Impacts of convolutive factors and different update methods

In the proposed family IS based nonnegative matrix factorization algorithms, the selection of the convolutive factors τ and ϕ has significant impact on the final separation results. This lies in the fact that the NMF is a weak model since it does not take into account the relative position of each spectrum thereby discarding the temporal information. In addition, the NMF does not model notes but rather unique events. 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. The proposed algorithm resolves these problems by extending the NMF model to be a two-dimensional convolution of $\mathbf{D}$ and $\mathbf{H}$ with the IS divergence. It factorizes the cochleagram using a model that represents both temporal structure and the pitch change which occurs when an instrument plays different notes simultaneously. To verify the above, an experimental study has been conducted to evaluate the performance of the different matrix factorization methods: IS-NMF, MU IS-NMF2D and Quasi-EM IS-NMF2D. Table 5.8 shows the performance with different algorithms under various audio mixtures.

Table 5.8: Separation results using different matrix factorization algorithm

Mixtures	Algorithms	SDR	SAR	SIR
jazz music and male speech	IS-NMF	4.14	7.54	8.72
	MU IS-NMF2D	7.45	9.23	11.96
	Quasi-EM IS-NMF2D	8.87	10.31	14.62
jazz music and female speech	IS-NMF	4.51	7.13	8.53
	MU IS-NMF2D	7.67	9.82	12.21
	Quasi-EM IS-NMF2D	9.34	9.77	14.37
piano music and male speech	IS-NMF	-0.70	7.57	0.68
	MU IS-NMF2D	5.84	8.21	10.07
	Quasi-EM IS-NMF2D	7.16	8.56	12.08
piano music and female speech	IS-NMF	2.59	6.32	5.12
	MU IS-NMF2D	6.36	8.55	10.42
	Quasi-EM IS-NMF2D	7.44	9.18	11.38
jazz music and piano music	IS-NMF	3.37	7.80	7.29
	MU IS-NMF2D	6.18	10.60	8.29
	Quasi-EM IS-NMF2D	7.21	13.07	8.68

Referring to Table 5.8, it is noted that the SDR performance vary significantly depending on the matrix factorization algorithms used for separation. For all type of mixtures, the IS-NMF algorithm delivers an average SDR of 2.78dB; the MU IS-NMF2D algorithm with an average of SDR 6.7dB and finally, the Quasi-EM IS-NMF2D algorithm with an average SDR of 8dB. The results obtained by using the NMF with convolutive factors outperform the method without the convolutive factors. It is also noted that both MU IS-NMF2D and Quasi-EM IS-NMF2D algorithms exhibit a good reconstruction in terms of SDR. However, the resulting factorizations are not equivalent. This is because the Quasi-EM IS-NMF2D algorithm prohibits zeros in the factors i.e. $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ cannot take entries equal to zero. In particular, in order to minimize $D_{IS}\left(\mathbf{V}'_k \mid \sum_{\tau,\phi} \mathbf{d}_k^{\tau} \overset{\downarrow\phi \ \rightarrow\tau}{\mathbf{h}_k^{\phi}}\right)$, if either $d_{f,k}^{\tau}$ or h_{k,t_s}^{ϕ} is zero then the resulting cost $D_{IS}\left(v'_{k,f,t_s} \mid \sum_{\tau,\phi} d_{f-\phi,k}^{\tau} h_{k,t_s-\tau}^{\phi}\right)$ becomes infinite. On the contrary, this is not

a feature shared by the MU IS-NMF2D algorithm, which does not *a priori* exclude zero coefficients in $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ (excepts for $\mathbf{Z}_{f,t_s}=0$, which would lead to a division by zero). Since zero coefficients are invariant under multiplicative updates, if the MU IS-NMF2D algorithm attains a fixed point solution with zero entries, then it cannot be determined if the limit point is a stationary point. On the other hand, if the limit point does not take zero entries (i.e. belongs to the interior of the parameter space) then it is a stationary point, which may or may not be a local minimum [118]. Consequently, the Quasi-EM IS-NMF2D algorithm can be considered more reliable for updating $\mathbf{D}^{\tau}$ as well as $\mathbf{H}^{\phi}$. Additionally, the Quasi-EM IS-NMF2D algorithm has outperformed all the above algorithms at every type of audio mixture. More precisely, the Quasi-EM IS-NMF2D algorithm leads to an average SDR improvement close to 1.3dB per source across all the different type of mixtures as compared to the MU IS-NMF2D algorithm. To further analyse the performance of all the above matrix factorization methods in separating the mixed signal and capturing the TF patterns of the sources, the cochleagram of the each recovered source has been plotted in Figure 5.13. In Figure 5.13, panels (A)-(B), (C)-(D) and (E)-(F) denote the recovered cochleagram of the female speech and jazz music by using the IS-NMF, MU IS-NMF2D and Quasi-EM IS-NMF2D algorithms, respectively. In particular, panels (A)-(D) imply that IS-NMF algorithm cannot obtain better reconstruction of the sources. On the other hand, it is noted that both MU IS-NMF2D and Quasi-EM IS-NMF2D algorithms exhibit good reconstruction of the female speech as well as the jazz music. However, the MU IS-NMF2D algorithm fails to identify several missing components as indicated in the red box marked area of panel (C)-(D). Hence, less accuracy is obtained in

the estimation of the jazz music as compared with the Quasi-EM IS-NMF2D algorithm which has successfully estimated both sources with high accuracy. In summary, all the results in Table 5.8 and Figures 5.13 unanimously show the importance of using the two-dimensional convolutive model of matrix factorization in order to correctly estimate the spectral and temporal features of each source.

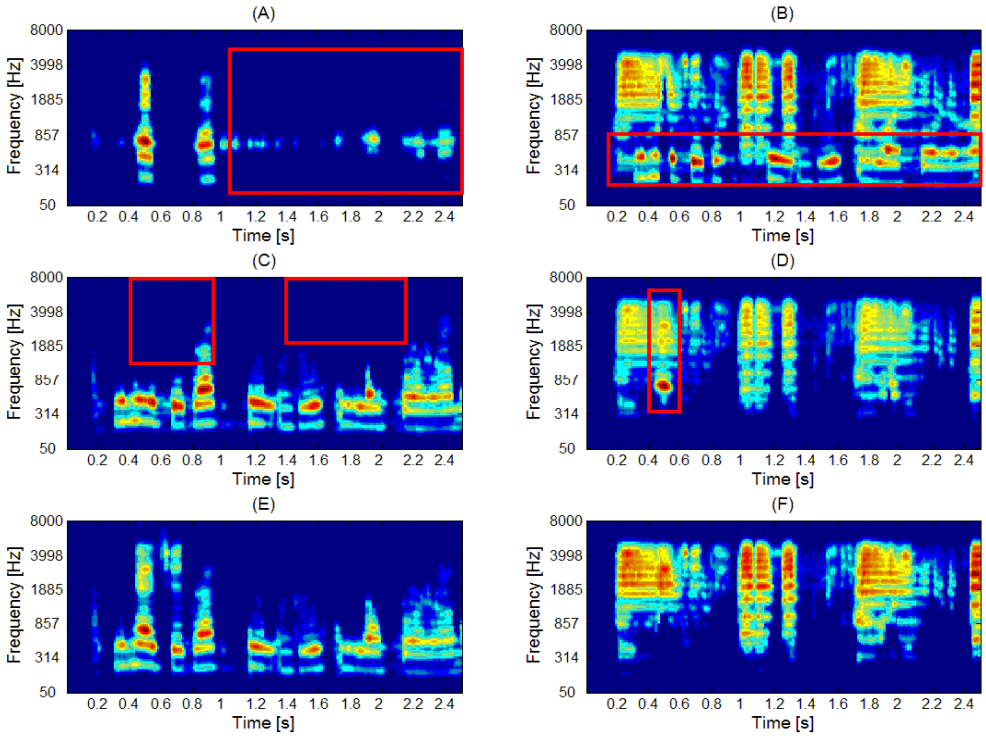


Figure 5.13: Decomposition results by using the family of IS-based nonnegative matrix factorizations (A)-(B) IS-NMF. (C)-(D) MU IS-NMF2D. (E)-(F) Quasi-EM IS-NMF2D.

5.3.3 Impacts of NMF2D using different cost function

Experiments have also been conducted to evaluate the NMF2D under different cost functions. Here, the Least Square (LS) distance and Kullback-Leibler (KL) divergence will be used for evaluation. Figure 5.14 shows the separation results of using NMF2D based on the LS, KL and IS cost functions. The algorithms LS-NMF2D was developed in [96], and

KL-NMF2D in [97].

Table 5.9: Separation results using NMF2D with different cost function

Mixtures	Algorithms	SDR	SAR	SIR
jazz music and male speech	LS-NMF2D	6.15	8.64	10.32
	KL-NMF2D	7.24	10.63	11.45
	Quasi-EM IS-NMF2D	8.87	10.31	14.62
jazz music and female speech	LS-NMF2D	4.69	8.63	10.11
	KL-NMF2D	7.35	11.23	13.27
	Quasi-EM IS-NMF2D	9.34	9.77	14.37
piano music and male speech	LS-NMF2D	5.11	7.28	10.46
	KL-NMF2D	5.42	8.61	9.65
	Quasi-EM IS-NMF2D	7.16	8.56	12.08
piano music and female speech	LS-NMF2D	4.21	7.90	6.22
	KL-NMF2D	5.38	8.32	8.32
	Quasi-EM IS-NMF2D	7.44	9.18	11.38
jazz music and piano music	LS-NMF2D	4.61	7.73	8.21
	KL-NMF2D	5.86	10.01	7.89
	Quasi-EM IS-NMF2D	7.21	13.07	8.68

Table 5.9 shows the overall comparison results among the three algorithms. It is noted that the results obtained by the Quasi-EM IS-NMF2D algorithm outperform those of LS distance and KL divergence on an average SDR of 3.1dB, and 1.8dB, respectively. This is evidenced by the fact that the IS divergence holds a desirable property of scale invariant so that low energy components can be precisely estimated and they bear the same relative importance as the high energy ones. On the contrary, factorizations obtained with LS distance or KL divergence highly dependent on the high energy components but abandon the low energy ones. In the cochleagram, the dynamic range can be large such that the dominating signal at a particular TF unit may manifest as low or high energy components. In addition, these components tend to exist as clusters. As such, when either LS distance- or KL divergence-based NMF2D is used, these clusters with low energy tend to be ignored

in favor of the high energy ones. This leads to mixing ambiguities in the cochleagram especially for low energy ones which subsumed together leads to significant lost of spectral-temporal information of the sources. Figure 5.14 shows the case of how different cost functions have impacted the separation.

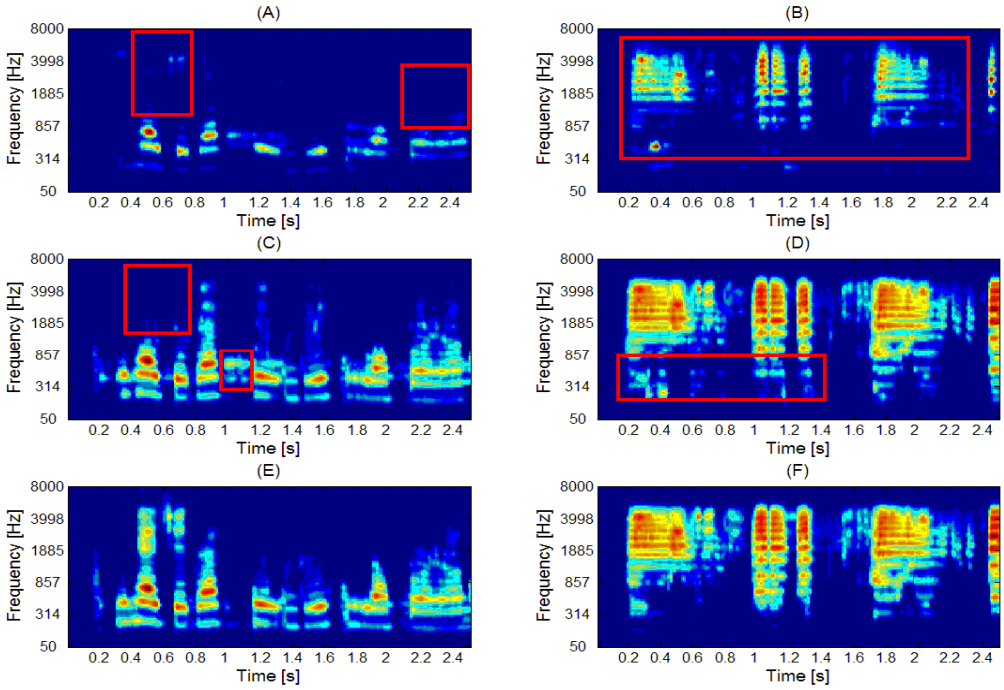


Figure 5.14: Separation results: (A)-(B), (C)-(D) and (E)-(F) denote the recovered female speech and jazz music in the cochleagram by using the LS-NMF2D, KL-NMF2D and Quasi-EM IS-NMF2D algorithms, respectively.

From Figure 5.14, it can be clearly seen that by using the LS-NMF2D algorithm, it fails in determining the correct TF components of each source. Figure 5.14 (A)-(B) also shows a considerable level of mixing ambiguities (red box marked area) which have not been accurately resolved by the LS-NMF2D algorithm. The KL-NMF2D exhibits better performance but ignores some low energy TF components in the red box marked area of (C)-(D). On the other hand, the proposed Quasi-EM IS-NMF2D algorithm has successfully

extracted the low energy components for both female speech and jazz music with high accuracy. This result shows the significance of using the IS divergence as the cost function for NMF2D in the estimation of spectral bases and temporal codes.

5.3.4 Impacts of regularizations selection

In this section, the impacts of regularizations on the factorization performance will be analysed. As mentioned in Section 5.2.3, the SNMF2D imposes uniform sparsity on all temporal codes and this is equivalent to enforcing each temporal code to be identical to a fixed distribution according to the selected sparsity parameter. The drawbacks of using uniform sparsity on all temporal codes are summarised in Chapter 3. Therefore, the above suggests that the current form of SNMF2D is still technically lacking and is not readily suited for SCBSS especially mixtures involving different types of audio signals. The proposed IS-vRNMF2D algorithm overcomes all the limitations associated with the IS-SNMF2D as previously discussed above. As each audio signal has its own temporal dependency of the frequency patterns, the basis vectors in $\mathbf{D}$ have to be designed to match the characteristics of these patterns efficiently. Hence, a suitably designed Gaussian prior on $\mathbf{D}$ is incorporated to allow those frequency patterns to be expressed for each audio signal.

We will show that when the sparse constraints are not controlled, the matrix factorization will be under- or over-sparse, and this will result in ambiguity in the estimation of recovered sources. Figures 5.15 shows the factorization results based on the IS-SNMF2D

and the proposed method. The top and middle panels clearly reveal that good separation performance require suitably controlled sparse regularization. In the case of uncontrolled sparse factorization, the estimated sources still retain redundant information where the two sources are not fully separated. In the case of the IS-vRNMF2D, it assigns a regularization parameter to each temporal code which is individually and adaptively tuned to yield the optimal number of times the spectral basis of a source recurs in the cochleagram. This is noted in the bottom panels which clearly show the optimal separation result.

In Figure 5.15, panels (A)-(D) imply that better separation results require the optimal sparse regularization when using IS-SNMF2D. If it is uncontrolled, the IS-SNMF2D will lead to either ‘under-sparse’ (e.g. (C)-(D)) or ‘over-sparse’ (e.g. (A)-(B)) factorization that still contain the mixed components in each separated sources. Panels (E)-(F) exhibits the recovered sources by using IS-vRNMF2D where it assigns a regularization parameter to each temporal coefficient (code), which is individually optimized and adaptively tuned to yield the optimal sparse and efficient matrix factorization.

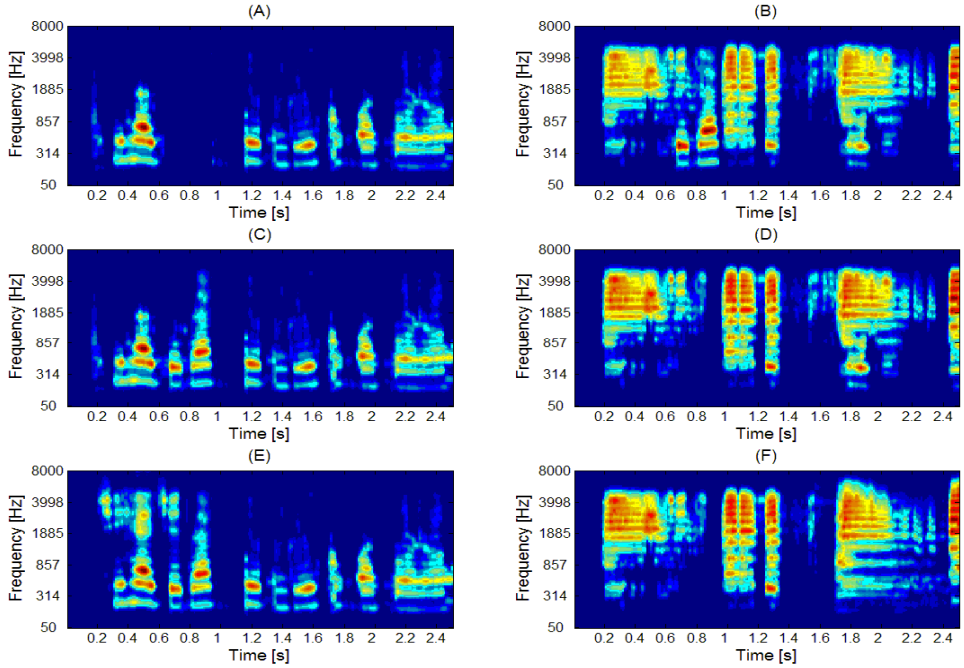


Figure 5.15: Separation results. Panels (A)-(B) denote the recovered cochleagram of the jazz music and female speech by using IS-SNMF2D whereas $\lambda^{\phi}_{i,t_s} = \lambda = 10$, $\mu_{ij\tau} = 0$. (C)-(D) denote the recovered cochleagram of the jazz music and female speech by using IS-SNMF2D whereas $\lambda^{\phi}_{i,t_s} = \lambda = 0.1$, $\mu_{ij\tau} = 0$. (E)-(F) denote the recovered cochleagram of the jazz music and female speech by using IS-vRNMF2D.

To investigate the effects of $\mu_{ij\tau}$ and λ^{ϕ}_{i,t_s} on the separation performance, three cases are conducted:

Case (i): No sparseness $\lambda^{\phi}_{i,t_s} = 0$ and $\mu_{ij\tau}$ is varied as $\mu_{ij\tau} = 0, 0.5, 1.0, \ldots, 5$.

Case (ii): Uniform sparseness $\lambda^{\phi}_{i,t_s} = c$ and $\mu_{ij\tau}$ is varied as $\mu_{ij\tau} = 0, 0.5, 1.0, \ldots, 5$.

Case (iii): Adaptive sparseness and $\mu_{ij\tau}$ is varied as $\mu_{ij\tau} = 0, 0.5, 1.0, \ldots, 5$.

For each case, the optimal results are based on Monte-Carlo simulation over 100 realizations.

Table 5.10: Separation results using different regularization based matrix factorization algorithms

Mixtures	TF methods	SDR	SAR	SIR
jazz music and male speech	Case (i)	7.75	9.14	11.22
	Case (ii)	8.1	9.7	10.1
	Case (iii)	9.4	10.3	10.8
jazz music and female speech	Case (i)	7.92	9.63	10.57
	Case (ii)	8.7	10.2	11.9
	Case (iii)	9.5	10.1	12.2
piano music and male speech	Case (i)	5.91	8.21	10.07
	Case (ii)	6.3	9.2	10.3
	Case (iii)	7.5	9.5	11.1
piano music and female speech	Case (i)	6.65	8.55	10.42
	Case (ii)	7.6	9.6	9.1
	Case (iii)	8.5	10.5	9.9
jazz music and piano music	Case (i)	6.5	10.6	7.29
	Case (ii)	7.7	10.3	7.4
	Case (iii)	8.4	11.1	8.5

From Table 5.10, in comparison, the average performance improvement of IS-vRNMF2D against the IS-SNMF2D and MU IS-NMF2D methods can be concluded as follows: (i) For music mixture, the average improvement per source delivers an average of 1.3dB SDR and (ii) For mixture of speech and music signal, the improvement per source is an average of SDR 1.4dB. Analysing the separation results in term of SDR, it is found that if sparse regularization uncontrolled in IS-SNMF2D that will incurs either ‘under-sparse’ or ‘over-sparse’, it is less robust than IS-vRNMF2D. Compared with the IS-SNMF2D and MU IS-NMF2D, the IS-vRNMF2D method renders a more optimal part based regularised decomposition whereas not only the learning algorithm is motivated by expressing patterns more effectively but also leads to faster convergence and least state value of IS cost function.

5.3.5 Comparison with other SCBSS methods

Table 5.11: Separation results using different SCBSS methods

Mixtures	TF methods	SDR	SAR	SIR
jazz music and male speech	IS-vRNMF2D	9.4	10.3	10.8
	NMF-TCS	3.2	6.7	8.1
	NMF-ARD	2.6	5.9	7.3
jazz music and female speech	IS-vRNMF2D	9.5	10.1	12.2
	NMF-TCS	3.8	6.2	8.9
	NMF-ARD	2.3	6.4	7.8
piano music and male speech	IS-vRNMF2D	7.5	9.5	11.1
	NMF-TCS	2.9	4.2	7.3
	NMF-ARD	1.4	4.3	7.5
piano music and female speech	IS-vRNMF2D	8.5	10.5	9.9
	NMF-TCS	2.5	4.5	7.1
	NMF-ARD	1.3	4.3	7.6
jazz music and piano music	IS-vRNMF2D	8.4	11.1	8.5
	NMF-TCS	2.8	7.3	7.4
	NMF-ARD	1.5	7.7	8.1

In comparison, the average performance improvement of the proposed method over the NMF-TCS and NMF-ARD method can be summarised as follows: (i) for music mixture, the average improvement per source is 5.6dB SDR and 6.9dB SDR, respectively. (ii) for mixture of speech and music signal, the improvement per source is 5.8dB SDR and 6.9dB SDR, respectively. The reasons why NMF-TCS [37] and NMF-ARD [97] obtain the worst separation performance are: Firstly, the NMF-ARD do not have convolutive factors $\tau, \phi = \{0\}$. As such, NMF-ARD are weak models since they do not take into account the relative position of each spectrum thereby discarding the temporal information. The spectral basis obtained via NMF-TCS and NMF-ARD methods are not adequate to capture the temporal dependency of the frequency patterns within the audio signal. Secondly, the NMF-TCS and NMF-ARD do not model 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.

5.4 Summary

In this chapter, a novel family of IS divergence based two-dimensional nonnegative matrix factorization methods to solve SCBSS has been proposed. The chapter presents a Quasi-EM based NMF2D with IS divergence (Quasi-EM IS-NMF2D), Multiplicative update based (non-regularised and regularised) NMF2D with IS divergence (These consist of IS-SNMF2D, MU IS-NMF2D and IS-vRNMF2D). The separation system of cochleagram and the family of IS divergence based factorization algorithms have been developed in a principled manner coupled with the theoretical support of audio signal separability. The proposed method enjoys at least three significant advantages: Firstly, it avoids strong constraints of separating sources without training knowledge where only single channel recording is provided. Secondly, the cochleagram rendered by the gammatone filterbank has non-uniform time-frequency resolution which enables the mixed signal to be more separable and improves the efficiency in source tracking. Finally, the IS divergency holds a desirable property of scale invariant that enables low energy components in the cochleagram bear the same relative importance as the high energy ones. In the comparison of IS based non-regularised and regularised NMF2D algorithms, the proposed IS-vRNMF2D obtains the best separation performance. The impetus behind this work is that, firstly, sparseness achieved by the conventional NMF, SNMF, NMF2D and

SNMF2D is not efficient enough; in source separation it is very necessary to yield control over the degree of sparseness explicitly for each temporal code. Secondly, the modified Gaussian prior is formulated to express the basis vectors more effectively; thus enabling the spectral and temporal features of the sources to be extracted 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 SCSS of linear instantaneous mixtures was presented. Both *supervised* SCSS methods and *unsupervised* SCSS methods that aim to increase the accuracy of the separated sources 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 separation of single channel mixtures to improve the performance at both theoretical and practical levels. This therefore provides the motivation for one of the aims of this thesis, which is to develop new strategies for retrieving single channel mixed sources.

6.1 Proposed *Unsupervised* Learning SCSS Methods

In Chapter 3, a new v-SNMF2D is presented for solving *unsupervised* SCSS problem. The impetus behind this is that the sparsity achieved by NMF is not enough; in such situations it might be useful to control the degree of sparseness explicitly. In the proposed method, the regularization term is adaptively tuned using a variational Bayesian approach

to yield desired sparse factorization, thus enabling the spectral basis and temporal codes of non-stationary audio signals to be estimated more efficiently. This has been verified based on our experiments. In addition, the proposed method has yielded significant improvements in separating single channel music mixture when compared with other sparse NMF-based *unsupervised* SCSS methods.

In Chapter 4, a novel framework of amalgamating EMD *with* v-SNMF2D is presented for solving *unsupervised* SCSS problem. In this chapter, it is shown that the IMFs have several desirable properties unique to single channel source separation problem: (*i*) the degree of mixing in each IMF is less ambiguous than the mixed signal, (*ii*) the IMFs has simpler and sparser spectral and temporal patterns which allows the proposed v-SNMF2D algorithm to efficiently track them, and (*iii*) the IMFs serve as the orthogonal temporal bases for signal separation; hence errors resulted from any IMF will be averaged over all the IMFs leading to smaller errors at the signal reconstruction stage. To this end, we have shown that the proposed method can deliver an acceptable separation performance for all types of single channel audio mixture.

In Chapter 5, a new family of IS divergence based factorization methods to solve *unsupervised* SCSS problem has been proposed. The chapter presents a Quasi-EM based NMF2D with Itakura-Saito divergence (Quasi-EM IS-NMF2D), Multiplicative update based (non-regularised and regularised) NMF2D with Itakura-Saito divergence (These consist of IS-SNMF2D, IS-NMF2D and IS-vRNMF2D). The cochleagram rendered by the gammatone filterbank has non-uniform time-frequency resolution which enables the mixed

signal to be more separable and improves the efficiency in source separation. In addition, the proposed IS-vRNMF2D obtains the best separation performance. These methods are tested on two types of mixture (mixture of music sources and mixture of music and speech). Table 6.1 summarise the proposed methods in this thesis

Table 6.1: Summary of the proposed SCBSS methods

Methods	**TF representation**	**Cost function**	**Regularization**		**Update method**
			D	**H**	
vSNMF2D	log-frequency	LS	-	Adaptive sparsity (VB)	MU
EMD-vSNMF2D	EMD + log-frequency	LS	-	Adaptive sparsity (VB)	MU
Quasi-EM IS-NMF2D	cochleagram	ISD	-	-	Quasi-EM
MU IS-NMF2D					MU
IS-SNMF2D	cochleagram	ISD	-	Uniform constant sparsity	MU
IS-vRNMF2D	cochleagram	ISD	Correlation of the basis	Adaptive sparsity (MAP)	MU

6.2 Comparison of the Proposed SCBSS Methods

In this section, the proposed three SCBSS methods will be tested across all types of mixture and compared in terms of SDR, SAR and SIR. In the proposed third method, the IS-vRNMF2D will be chosen for comparison as it has been proven to be the best method among all types of IS divergence based nonnegative matrix factorization algorithms. The following table summarises the comparison results.

Table 6.2: Separation results using different SCBSS methods

Mixtures	TF methods	SDR	SAR	SIR
jazz music and male speech	v-SNMF2D	6.6	7.2	8.3
	EMD-vSNMF2D	8.8	7.7	9.2
	IS-vRNMF2D	**9.4**	10.3	10.8
jazz music and female speech	v-SNMF2D	6.4	7.7	8.1
	EMD-vSNMF2D	8.7	9.5	10.4
	IS-vRNMF2D	**9.5**	10.1	12.2
piano music and male speech	v-SNMF2D	5.2	6.2	7.1
	EMD-vSNMF2D	6.3	6.6	8.5
	IS-vRNMF2D	**7.5**	9.5	11.1
piano music and female speech	v-SNMF2D	5.4	6.5	7.3
	EMD-vSNMF2D	6.7	7.2	8.3
	IS-vRNMF2D	**8.5**	10.5	9.9
jazz music and piano music	v-SNMF2D	7.1	8.5	9.6
	EMD-vSNMF2D	7.4	10.5	9.1
	IS-vRNMF2D	**8.4**	11.1	8.5
male speech and female speech	v-SNMF2D	2.4	5.3	5.8
	EMD-vSNMF2D	**5.7**	7.1	8.2
	IS-vRNMF2D	3.5	6.2	7.1

In comparison, the IS-vRNMF2D with cochleagram leads to the best separation performance for most types of the mixture except the mixture of male speech and female speech. The EMD-vSNMF2D also performs the relative good results as compared with IS-vRNMF2D. The reasons of using EMD as a preprocessing tool for SCBSS have been described in Chapter 3. However, it is interesting to point that the big advantage of using IS-vRNMF2D with cochleagram is that this method is less complexity intensive than EMD-vSNMF2D method and simultaneously retain a high level of the separation performance. The reasons of relative poorer separation results obtained by v-SNMF2D can be summarised as three points. Firstly, the v-SNMF2D model is based on least square cost function, the drawbacks of using this cost function has already been discussed in Section 5.3.3. Secondly, the v-SNMF2D does not have prior on $\mathbf{D}$ such that the frequency

patterns of each source may not be estimated as well as IS-vRNMF2D. Finally, the v-SNMF2D is performed by using log-frequency spectrogram in such case the separability of this TF representation is worse than cochleagram. However, the adaptive sparsity by using variational Bayesian is more reliable than MAP approachs since the former prohibit zero elements in each adaptive step for sparsity parameter.

6.3 Future Work

6.3.1 Development of SCBSS method for non-stationary mixing model

In the future work, the SCBSS method to separate non-stationary (here non-stationary refers to, the sources not located in the fixed place, e.g. the speaker is talking while he is walking) and reverberant mixing model will be developed. The non-stationary and reverberant mixing model has not been solved by using current SCSS methods. For instantaneous non-stationary mixing model, it gives as follows:

$$y(t) = \sum_{i=1}^{N_s} m_i(t) x_i(t) + n(t) \tag{6.1}$$

where $m_i(t)$ denotes the i^{th} source mixing parameters at t time, and $n(t)$ is additive noise. For non-stationary reverberant mixing model, it gives as follows:

$$y(t) = \sum_{i=1}^{N_s} \sum_{\tau_r=0}^{L_r-1} m_i(\tau_r, t) x_i(t-\tau_r) + n(t) \tag{6.2}$$

where $m_i(\tau_r, t)$ is the finite impulse response of causal filter at t time and τ_r is the time delay. Thus, the power TF representation of matrix representation is given by $|\mathbf{Y}|^2 = \sum_{i=1}^{N_s} |\mathbf{M}_i|^2 \bullet |\mathbf{X}_i|^2 + \mathbf{V}^{No}$. The matrix $\mathbf{M}_i$ is a mixing parameter in TF domain (it is

assumed that the mixing parameter is stationary within a short period such that (i) for instantaneous mixing $\mathbf{M}_i = \left[\mathbf{m}_{i,1},\ldots,\mathbf{m}_{i,T_s}\right]$ where $\mathbf{m}_{i,t_s} = \left[m_{i,1,t_s},\ldots,m_{i,F,t_s}\right]$ and $m_{i,1,t_s} = m_{i,2,t_s},\ldots,= m_{i,F,t_s}$; (ii) for convolutive mixing $\mathbf{M}_i = \left[m_{i,f,t_s}\right]_{t_s=1,2,\ldots,T_s}^{f=1,2,\ldots,F}$, m_{i,f,t_s} is the $(i,f,t_s)^{\text{th}}$ element of $\mathbf{M}_i$ and $\mathbf{V}^{No}$ is the noise. The aim of the developed SCBSS method is to estimate nonstationary mixing model $\mathbf{M}_i$ and the sources $\left|\mathbf{X}_i\right|^{.2}$.

6.3.2 Development of signal dependent TF representation

In this novel idea, the time domain mixed signal will be projected onto signal-dependent multidimensional transform domain where the specific features of each source sparsely and smoothly clustered with maximally distinction. Figure 6.1 shows an example of the proposed structure of the new signal dependent TF transform. It can be divided into two stages. In the first stage, different types of audio signals will be trained to find the most suitable signal dependent TF transform. Once this has been done, when different type of audio sources mixed in single channel, the proposed signal dependent TF transform will learn training information and automatically adjust itself to suit the mixture signal.

In Figure 6.1, '$SDTF_X$' denotes signal dependent TF transform of the source and '$SDTF_Y$' denotes signal dependent TF transform of the mixture. '*F(.)*' denotes the function to analysis the features of each source when using different types of multidimensional representation. For generating signal dependent TF transform, we can consider to use the idea of hybrid compressive sensing method and dynamic filterbank technique to construct the proposed signal dependent TF representation for audio source. This transform will

bring at least two benefits to SCBSS problem: Firstly, the mixture in this domain will be more separable than other conventional TF transform.

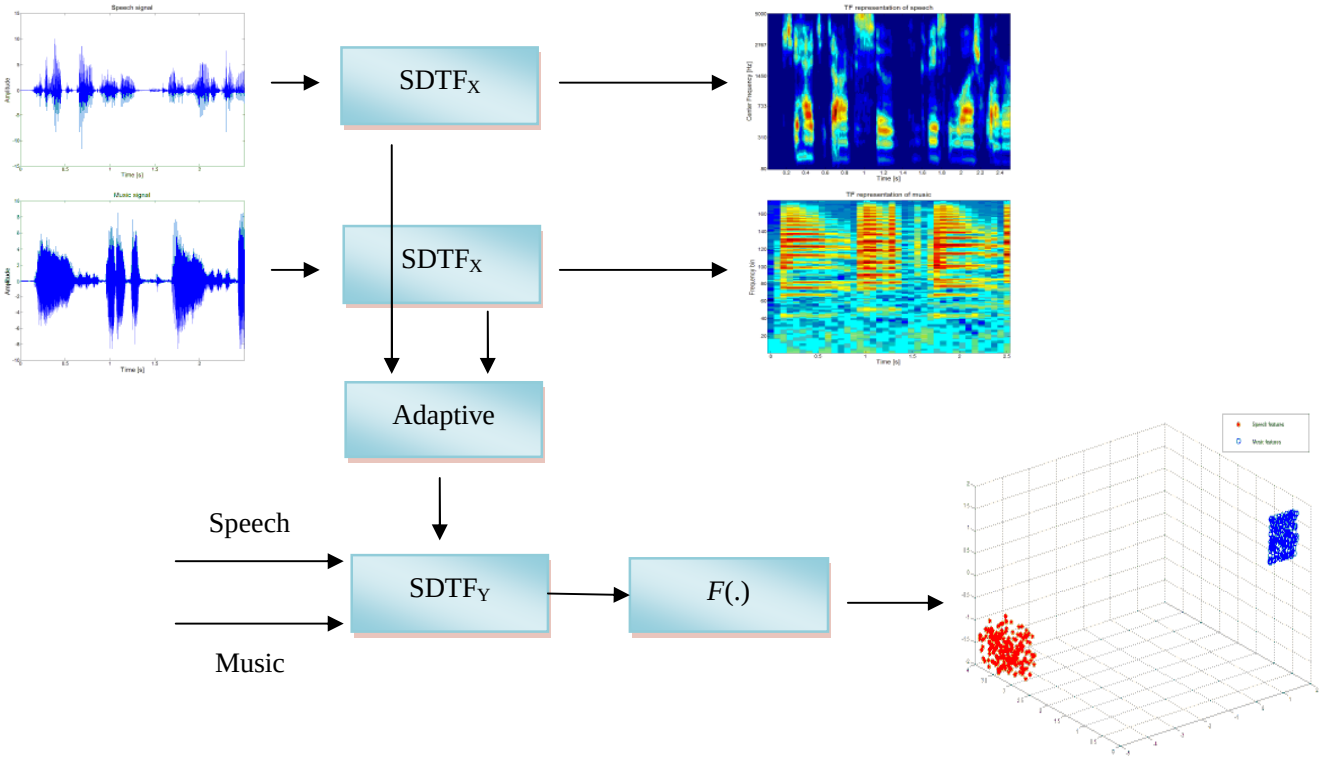


Figure 6.1: The proposed multi-dimensional signal dependent TF transform.

To enhance the analysis of TF transformation, the separability theory of SCSS as described in Chapter 5 can be used. This could be seen as a function of '*F(.)*' that analyses the features of different sources and these features can be clustered with different level of distinction by using different types of TF representation. Secondly, the *analysis of source tracking* (for identifying the TF patterns that belong to a particular original source) should be substantially more effective in signal dependent TF domain than in other TF domains.

6.3.3 Development of Quasi-EM IS-vRNMF2D

As described in Chapter 5, the key difference between Quasi-EM IS-NMF2D and MU

IS-NMF2D is that the former algorithm prohibits zeros in the factors i.e. $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ cannot take entries equal to zero. In particular, in order to minimize $d_{IS}\left(\mathbf{V}'_k \middle| \sum_{\tau,\phi} \mathbf{d}_k^{\tau} \overset{\downarrow\phi}{\mathbf{h}}_k^{\phi} \overset{\rightarrow\tau}{}\right)$, if either $d_{f,k}^{\tau}$ or h_{k,t_s}^{ϕ} is zero then the resulting cost function becomes infinite. On the contrary, this is not a feature shared by the MU IS-NMF2D algorithm, which does not *a priori* exclude zero coefficients in $\mathbf{D}^{\tau}$ and $\mathbf{H}^{\phi}$ (except for $\mathbf{Z}_{f,t_s}=0$, which would lead to a division by zero). Since zero coefficients are invariant under multiplicative updates, if the MU IS-NMF2D algorithm attains a fixed point solution with zero entries, then it cannot be determined if the limit point is a stationary point. On the other hand, if the limit point does not take zero entries (i.e. belongs to the interior of the parameter space) then it is a stationary point, which may or may not be a local minimum. Consequently, the Quasi-EM IS-NMF2D is better than MU IS-NMF2D. In addition, it is desirable to have regularization for imposing the sparseness and constrain the correlation between different spectral bases in the process of matrix factorization. This has been verified in Section 5.2.3. Thus, the development of regularised Quasi-EM IS-NMF2D is necessary to improve the accuracy of separation performance. Consider the generative model in (5.17), the EM algorithm works by formulating the conditional expectation of the negative log likelihood of $\boldsymbol{\upsilon}_k$ as:

$$p\left(\mathbf{d}_k^{\tau},\mathbf{h}_k^{\phi}\middle|\boldsymbol{\upsilon}_k\right)=\frac{p\left(\boldsymbol{\upsilon}_k\middle|\mathbf{d}_k^{\tau},\mathbf{h}_k^{\phi}\right)p\left(\mathbf{d}_k^{\tau}\right)p\left(\mathbf{h}_k^{\phi}\right)}{P\left(\boldsymbol{\upsilon}_k\right)} \tag{6.3}$$

where the denominator is a constant and it is assumed $\mathbf{d}_k^{\tau}$ and $\mathbf{h}_k^{\phi}$ are jointly independent so that EM algorithm (5.18) can be presented as:

$$
\begin{aligned}
Q_k^{MAP}\left(\boldsymbol{\theta}_k \mid \boldsymbol{\theta}'\right) &\doteq -\int_{\mathbf{C}_k} p\left(\boldsymbol{\upsilon}_k \mid \mathbf{Y}, \boldsymbol{\theta}'\right) \log p\left(\boldsymbol{\theta}_k \mid \boldsymbol{\upsilon}_k\right) d\boldsymbol{\upsilon}_k \\
&= -\int_{\mathbf{C}_k} p\left(\boldsymbol{\upsilon}_k \mid \mathbf{Y}, \boldsymbol{\theta}'\right) \left[\log p\left(\boldsymbol{\upsilon}_k \mid \boldsymbol{\theta}_k\right) + \log p\left(\mathbf{d}_k^{\tau}\right) + \log p\left(\mathbf{h}_k^{\phi}\right)\right] d\boldsymbol{\upsilon}_k \\
&= -\int_{\mathbf{C}_k} p\left(\boldsymbol{\upsilon}_k \mid \mathbf{Y}, \boldsymbol{\theta}'\right) \left[\log p\left(\boldsymbol{\upsilon}_k \mid \boldsymbol{\theta}_k\right)\right] d\boldsymbol{\upsilon}_k - \log p\left(\mathbf{d}_k^{\tau}\right) - \log p\left(\mathbf{h}_k^{\phi}\right)
\end{aligned}
\tag{6.4}
$$

In (6.4), the prior distribution over $\mathbf{d}_k^{\tau}$ can be assumed to be zero-mean multivariate rectified Gaussian with covariance matrix $\boldsymbol{\Sigma}_{\tau}^{k}$ and the prior distribution over $\mathbf{h}_k^{\phi}$ can be assumed to be exponential distributed with independent decay parameters λ_{k,t_s}^{ϕ}. With these assumptions, $\mathbf{d}_k^{\tau}$ and $\mathbf{h}_k^{\phi}$ can be optimized by following the approach presented in Section 5.2.3.

REFERENCE

[1] Cichocki and S. I. Amari, *Adaptive Blind Signal and Image Processing – Learning Algorithms and Applications*, J. Wiley & Sons Ltd., 2003.

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

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

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

[5] J. F. Cardoso, "Source Separation using Higher Order Moments", in *Proceedings ICASSP*, Glasgow, 1989, pp. 2109-2112.

[6] J. F. Cardoso, "Blind Signal Separation: Statistical Principles", in *Proceedings of the IEEE*, vol. 86, 1998, pp. 2009-2025.

[7] E. Oja, J. Karhunen, L. Wang and R. Vigario, "Principal and Independent Components in Neural Networks", in *Proc. VII Italian Workshop on Neural Nets WIRN*, Italy, 1995.

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

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

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

[11] A. Hyvarinen, J. Karhunen, and E. Oja, Independent component analysis and blind source separation, John Wiley & Sons pp.20–60, 2001.

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

[13] 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.

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

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

[16] C. Jutten, M. Babaie-Zadeh, and S. Hosseini, "Three Easy Ways for Separating Nonlinear Mixtures? ", *Signal Processing*, vol. 84, no. 2, pp. 217-229, 2004.

[17] M. Solazzi, R. Parisi, and A. Uncini, "Blind Source Separation in Nonlinear Mixtures by Adaptive Spline Neural Networks", in *Proc. Int. Conf. on Independent Component Analysis and Signal Separation (ICA2001)*, San Diego, USA, 2001, pp. 254-259.

[18] A. Hyvarinen, and P. Pajunen, "Nonlinear independent component analysis: Existence and uniqueness results", *Neural Networks*, vol. 12, no. 3, pp. 429-439, Apr. 1999.

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

[20] V. D. Calhoun, T. Adali, L. K. Hansen, J. Larsen, and J. J. Pekar, "ICA of Functional MRI Data: An Overview", in *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, April 2003.

[21] F. Acernese, A. Ciaramella, S. De Martino, R. De Rosa, M. Falanga, and R. Tagliaferri, "Neural Networks for Blind Source Separation of Stromboli Explosion Quakes", *IEEE Transactions on Neural Networks*, vol. 14, pp. 167-175, 2003.

[22] 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, Aug. 2004, pp. 311-312.

[23] 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*, in press.

[24] 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.

[25] 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, April. 2004, pp. 149-152.

[26] 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.

[27] A. D. Back and A. S. Weigend, "A First Application of Independent Component Analysis to Extracting Structure from Stock Returns", *International Journal of Neural Systems*, vol. 8, Issue 4, pp. 474-484, 1997.

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

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

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

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

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

[33] W. L. Woo and S. S. Dlay, "Neural network Approach to Blind Separation Mono-nonlinearly Mixed Sources", *IEEE Trans. On Circuits and System-1*, vol. 52, no. 6, pp. 1236-1247, 2005.

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

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

[36] W. L. Woo, and S. Sali, "General Multilayer Perceptron Demixer Scheme for Nonlinear Blind Signal Separation", in *IEE Proc. On Vision, Image and Signal Processing*, vol. 149, Oct. 2002, pp. 253-262.

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

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

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

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

[41] D. Ellis, "Model-based scene analysis", in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G. Brown, Eds. New York: Wiley/IEEE Press, 2006.

[42] M.J. Reyes-Gomez, D. Ellis, and N. Jojic, "Multiband audio modeling for single channel acoustic source separation", in *Proc. of Intl. Conf. Acoustic, Speech, and Signal Processing (ICASSP'04)*, Montreal, Canada, vol. 5, May 2004, pp. 641–644.

[43] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction", in *Proc. of Intl. Conf. Acoustic, Speech, and Signal Processing (ICASSP'04)*, Montreal, Canada, vol.2, May 2004, pp. 817–820.

[44] M. Mandel, R. Weiss, and D. Ellis, "Model-Based Expectation-Maximization Source Separation and Localization", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 18 no. 2, pp. 382-394, February 2010.

[45] Y. Li, S. Amari, A. Cichocki, D. W.C. Ho, and X. Shengli, "Underdetermined blind source separation based on sparse representation", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 54, no. 2, pp. 423–437, 2006.

[46] C. Fevotte and S. J. Godsill, "A Bayesian approach for blind separation of sparse sources", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2174–2188, Nov. 2006.

[47] 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.

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

[49] G. Hu and D.L. Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation”, *IEEE Trans. Neural Networks*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[50] M.S. Pedersen, D.L. Wang, J. Larsen and U. Kjems, “Two-Microphone Separation of Speech Mixtures”, *IEEE Trans. on Neural Networks*, vol. 19, no. 3, pp. 475–492, March. 2008.

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

[52] D. Ellis, “Prediction-driven computational auditory scene analysis”, Ph.D. dissertation, MIT, 1996.

[53] Woodruff J. and Wang D.L, “Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization”, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1856-1866, 2010.

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

[55] 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.

[56] M. A. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis", in *Proc. Int. Comput. Music Conf*, 2000, pp. 154–161.

[57] Md. K. I. Molla and K. Hirose, "Single-Mixture Audio Source Separation by Subspace Decomposition of Hilbert Spectrum", *IEEE Trans. on Audio, Speech and* Language Processing, vol. 15, no. 3, pp. 893–900, March 2007.

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

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

[60] D. Ellis, "Prediction-driven computational auditory scene analysis", Ph.D. dissertation, MIT, 1996.

[61] P. Bofill and M. Zibulevsky, "Underdetermined blind source separation using sparse representations", *Signal Process*, vol. 81, pp. 2353–2362, 2001.

[62] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization", *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, Oct. 2003.

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

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

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

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

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

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

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

[70] 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.

[71] O. Okun, and H. Priisalu, "Unsupervised data reduction", *Signal Processing*, vol. 87, no. 9, pp. 2260-2267, 2007.

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

[73] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription", in *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '03)*, New Paltz, NY, USA, October 2003, pp. 177–180.

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

[75] 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.

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

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

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

[79] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 91–98, 2006.

[80] A. T. Cemgil, "Bayesian inference for nonnegative matrix factorisation models" *Computational Intelligence and Neuroscience*. doi: 10.1155/2009/785152. 2009.

[81] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret, "Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling" *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4133–4145, Nov 2006.

[82] M. N. Schmidt, O. Winther, and L.K. Hansen "Bayesian non-negative matrix factorization," *Independent Component Analysis and Signal Separation, International Conference on*, 2009, pp. 540-547.

[83] R. Salakhutdinov and A. Mnih "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo" in *Proc of the 25th international conference on Machine learning*, 2008. pp. 880-887.

[84] S. A. Abdallah and M. D. Plumbley, "Polyphonic transcription by non-negative sparse coding of power spectra", in *Proc. 5th Conf. on Music Information Retrieval (ISMIR '04)*, Barcelona, Spain, October 2004, pp. 318–325.

[85] R. M. Parry and I. Essa, "Incorporating phase information for source separation via spectrogram factorization", in *Proc. Intl. Conf. Acoustics, Speech and Signal Processing (ICASSP '07)*, vol. 2, Hawaii, USA, April 2007, pp. 661–664.

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

[87] A. Cichocki, R. Zdunek, and S.-I. Amari, "Csisz´ar's divergences for non-negative matrix factorization: family of new algorithms", in *Proc. 6th Intl. Conf. on Independent Component Analysis and Signal Separation (ICA '06)*, Charleston, USA, March 2006, pp. 32–39.

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

[89] A. Hyv¨arinen and P. Hoyer, "Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces", *Neural Computation*, vol.12, col. 7, pp.1705–1720, 2000.

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

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

[92] M. Morup and M. N. Schmidt, "Sparse non-negative matrix factor 2-D deconvolution", *Tech. Rep Technical University of Denmark*, Copenhagen, Denmark, 2006.

[93] Yuanqing Lin, "l1-norm sparse Bayesian learning: theory and applications", Ph.D. Thesis, University of Pennsylvania, 2008.

[94] Yuanqing Lin, Daniel D. Lee, "Bayesian Regularization And Nonnegative Deconvolution (BRAND) for room impulse response estimation", *IEEE Trans. Signal Processing*, vol. 54, col. 3, pp. 839-847, 2006.

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

[96] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation", *IEEE Trans. on Audio, Speech, and Language Processing*. vol. 14, no. 4, pp. 1462–1469, Jul. 2005.

[97] M. Mørup and K.L. Hansen "Tuning Pruning in Sparse Non-negative Matrix Factorization", in *Proc. of 17th European Signal Processing Conference (EUSIPCO'09)*, Glasgow, Scotland, 2009.

[98] N.E. Huang, Z. Shen, S.R. Long, M.L. Wu, H.H. Shih, Q. Zheng, N.C. Yen, C.C. Tung and H.H. Liu, "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis", in *Proc. Royal Soc. London A*, vol.454, pp.903–995, 2002.

[99] B.Z. Wu, and N.E. Huang, "A study of the characteristics of white noise using the empirical mode decomposition method" in *Proc of. Royal Soc. London A,* no. 460, pp. 1597–1611, 2004.

[100] M.Goto, H.Hashiguchi, T.Nishimura, and R.Oka. "RWC music database: Music genre database and musical instrument sound database", in *Proc. of Intl. Symp. on Music Information Retrieval (ISMIR),* Baltimore, Maryland, USA, October 2003, pp. 229–230.

[101] Bin Gao, W.L. Woo, S.S. Dlay "Single channel blind source separation using the best characteristic basis", in *Proc of 3rd Int. Conf. on Information and Communication Technologies: From Theory to Applications (ICTTA'08),* Syria, 2008, pp. 1–5.

[102] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions", in *Proc. of Automatic Speech Recognition: Challenges for the new Millennium (ISCA ITRW ASR2000),* Paris, France, 2000, pp. 29-32.

[103] K. Khaldi, A. Boudraa, A. Bouchikhi, and M. T. Alouane, "Speech enhancement via EMD", *EURASIP Journal on Advances in Signal Processing,* vol. 2008, Article ID 873204, 2008.

[104] E.Deger, Md. K. I. Molla, K.Hirose and N.Minematsu, "Speech enhancement using soft-thresholding with DCT-EMD based hybrid algorithm", in *Proc. of 15th European Signal Processing Conference (EUSIPCO 2007),* 2007.

[105] Judith C. Brown, "Calculation of a constant Q spectral transform", *J. Acoust. Soc. Am.,* vol. 89, no 1, pp. 425–434, 1991.

[106] K. Gröchenig, *Foundations of Time-Frequency Analysis,* Birkhäuser, Boston, 2001.

[107] Z. Jin and D.L Wang, "A supervised learning approach to monaural segregation of reverberant speech", *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, pp. 625-638. 2009.

[108] G. Hu and D. L. Wang, "Auditory segmentation based on onset and offset analysis", *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 2, pp. 396–405, Feb. 2007.

[109] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis", in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, pp. 181–197, 2005.

[110] Roads, Curtis et al, *The computer music tutorial*, The MIT Press, 1996.

[111] S. Schulz and T. Herfet, "Binaural source separation in non-ideal reverberant environments", in *Proc. of 10th Intl. Conf. Digital Audio Effects (DAFx-07)*, Bordeaux, France, September, 2007, pp. 10-15.

[112] A. R. Palmer, "Physiology of the cochlear nerve and cochlear nucleus", in *Hearing, M. P. Haggard and E. F. Evans*, Eds., Edinburgh: Churchill Livingstone, 1987.

[113] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data", *Hearing Research*, vol. 47, pp. 103–138, 1990.

[114] ISO, Normal Equal-Loudness Level Contours for Pure Tones Under Free-Field Listening Conditions (ISO 226), International Standards Organization.

[115] E. de Boer and H. D. de Jongh, "On cochlear encoding: Potentialities and limitations of the reverse correlation technique", *Journal of Acoust. Soc. Am.*, vol. 63, pp. 115–135, 1978.

[116] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, APU Report 2341: An Efficient Auditory Filterbank Based on the Gammatone Function, Cambridge: Applied Psychology Unit, 1988.

[117] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method", in *Proc 6th Intl. Congress on Acoustics*, Tokyo, Japan, Aug. 1968. pp. C–17-20.

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

[119] J.A. Fessler and A.O. Hero, "Space-alternating generalized expectation-maximization algorithm", *IEEE Trans. on Signal Processing*, vol.42, no. 10, pp.:2664–2677, Oct. 1994.

[120] Y. Li and D. L. Wang. "Musical sound separation based on binary time-frequency masking", *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, Article ID 130567, 2009.