ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

PERCEPTUAL AUDIO SOURCE SEPARATION BY SUBSPACE LEARNING

Ph.D. THESIS

Serap KIRBIZ

Department of Electronics and Communications

Telecommunication Engineering Programme

Thesis Advisor: Prof. Dr. Bilge GÜNSEL

JANUARY 2013

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF SCIENCE ENGINEERING AND TECHNOLOGY

PERCEPTUAL AUDIO SOURCE SEPARATION BY SUBSPACE LEARNING

Ph.D. THESIS

Serap KIRBIZ
(504052304)

Department of Electronics and Communications

Telecommunication Engineering Programme

Thesis Advisor: Prof. Dr. Bilge GÜNSEL

JANUARY 2013

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

ALTUZAY ÖĞRENME İLE ALGISAL SES KAYNAK AYRIŞTIRMA

DOKTORA TEZİ

Serap KIRBIZ
(504052304)

Elektronik ve haberleşme Mühendisliği Anabilim Dalı

Telekominikasyon Mühendisliği Programı

Tez Danışmanı: Prof. Dr. Bilge GÜNSEL

OCAK 2013

Serap KIRBIZ, **a Ph.D.** student of ITU **Graduate School of Science Engineering and Technology** 504052304, successfully defended the **dissertation** entitled "**PERCEPTUAL AUDIO SOURCE SEPARATION BY SUBSPACE LEARNING**", which she prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor :	**Prof. Dr. Bilge GÜNSEL** İstanbul Technical University	.....
Jury Members :	**Prof. Dr. Ahmet Hamdi KAYRAN** İstanbul Technical University	.....
	Prof. Dr. Bülent SANKUR Boğaziçi University	.....
	Prof. Dr. Muhittin Gökmen İstanbul Technical University	.....
	Prof. Dr. Ayşın Baytan ERTÜZÜN Boğaziçi University	.....

Date of Submission : 28 September 2012
Date of Defense : 8 January 2013

FOREWORD

First of all, I wish to express my deep gratitude to my supervisor Prof. Dr. Bilge GÜNSEL. I appreciate the effort that she put on this study, her inspiring advices, valuable comments and continuous guidance to improve the quality of my thesis.

I would also like to express my deep appreciation to my thesis progress committee members Prof. Dr. Ahmet Hamdi KAYRAN and Prof. Dr. Bülent SANKUR for guiding me on my way and being with me during hard times and also for their direction, and invaluable advices during this thesis. I would like to thank the rest of my thesis committee members Prof. Dr. Muhittin GÖKMEN and Prof. Dr. Ayşın Baytan ERTÜZÜN for reviewing this work and providing helpful feedback.

I am grateful to Prof. Dr. Barbara Shinn-CUNNINGHAM for accepting me to her research laboratory at Boston University.

I am also grateful to Assist. Prof. Paris SMARAGDIS for his motivating support, time, and effort during my studies at Boston University. He always provided valuable comments and outstanding directions for the efficiency and the quality of my work. I am honored to be one of his students.

I would also like to thank to Assoc. Prof. A. Taylan CEMGİL for introducing me to new research areas and for his guidance.

I also want to extend my thanks to the friends at the Department of Electronics and Communications, Istanbul Technical University (present and past) for their support and kind presence whenever needed: Ebru Sinem ÇETİN, Suat AKSU, Süleyman BAYKUT, Özgün ÇIRAKMAN, Sezer KUTLUK, Ömer Deniz AKYILDIZ, Ozan GÜRSOY, Yener ÜLKER, Koray KAYABOL, Özgür ORUÇ and Cercis Özgür SOLMAZ.

I would never forget the help and support I got from Müge BALKAYA, Gökce ÇÖL, Derya GÜNEY and Aytül ERDALOĞLU for helping me get through the difficult times and for all the emotional support, entertainment and caring they provided.

Sincere gratitude and special thanks are for my family who always support and encourage me to step forward and share my success and happiness.

January 2013

Serap KIRBIZ
(M.Sc.)

TABLE OF CONTENTS

Page

ABBREVIATIONS

ADF : Adaptive Decorrelation Filtering
APS : Artifacts-related Perceptual Score
BNMFD : Bayesian Non-negative Matrix Factor Deconvolution
BS : Broadcasting Service
BSSEVAL : Blind Source Separation EVALuation
CASA : Computational Auditory Scene Analysis
CD : Compact Disk
CNMF : Clustered Non-Negative Matrix Factorization
CP : CANDECOMP/PARAFAC
CQT : Constant-Q-Transform
dB : Desibel
DFT : Discrete Fourier Transform
EM : Expectation Maximization
ERB : Equivalent Rectangular Bandwidth
FDICA : Frequency-Domain Independent Component Analysis
FIR : Finite Impulse Response
GHz : Giga-Hertz
HAS : Human Auditory System
Hz : Hertz
ICA : Independent Component Analysis
IID : Interchannel Intensity Difference
IPS : Interference-related Perceptual Score
IS : Itakura-Saito
ISA : Independent Subspace Analysis
ISTFT : Inverse Short-Time Fourier Transform
ITD : Interchannel Time Difference
ITU : International Telecommunication Union
ITU-R : International Telecommunication Union Recommendation
kHz : kilo-Hertz
KL : Kullback-Leibler
MAP : Maximum-a-Posteriori
ML : Maximum Likelihood
MRFD : MultiResolution Frequency Domain
MR-NTF : MultiResolution Non-negative Tensor Factorization
NMF : Non-negative Matrix Factorization
NMF2D : Non-negative Matrix Factor 2-D Deconvolution

NMFD : Non-negative Matrix Factor Deconvolution
NMR : Noise-to-Mask-Ratio
NTF : Non-negative Tensor Factorization
OLA : OverLap-Add
OPS : Overall Perceptual Score
PARAFAC : Parallel Factor Analysis
PC : Personal Computer
NTPCA : Non-negative Tensor Pricipal Component Analysis
PEAQ : Perceptual Evaluation of Audio Quality
PEASS : Perceptual Audio Source Separation
PEMO-Q : PErception MOdel based Quality
PSM : Perceptual Similarity Measure
PW-CNMF : Perceptually Weighted Clustered Non-Negative Matrix Factorization
PW-NMF2D : Perceptually Weighted Non-negative Matrix Factor 2-D Deconvolution
PW-MRNTF : Perceptually Weighted MultiResolution Non-negative Tensor Factorization
SAR : Signal-to-Artifacts-Ratio
SDR : Signal-to-Distortion-Ratio
SIR : Signal-to-Interference-Ratio
SISEC : SIgnal Separation Evaluation Campaign
SNMF : Shifted Non-negative Matrix Factorization
SNTF : Shifted Non-negative Tensor Factorization
SPL : Sound Pressure Level
STFT : Short-Time Fourier Transform
VB : Variational Bayes
WP : Wavelet Packet

LIST OF SYMBOLS

a	:	Sparsity parameter
b	:	Mean of the Gamma distribution
b_p	:	Bias of sigmoid p
c	:	Constant coefficient
$\mathbf{e}_j^{artif}$	:	Artifacts for the $j-$th source
$\mathbf{e}_j^{interf}$	:	Interference for the $j-$th source
$\mathbf{e}_j^{target}$	:	Target distortion for the $j-$th source
f	:	Perceptual score
f_L	:	Lowest frequency in critical bands
f_l	:	Lower frequency edge
f_U	:	Highest frequency in critical bands
f_u	:	Upper frequency edge
g	:	Sigmoid function
p	:	Probability
q	:	Instrumental distribution
$\mathbf{q}_j$	:	Feature vector of $j-$th source
q_j^{artif}	:	Feature for $j-$th source artifacts
q_j^{interf}	:	Feature for $j-$th source interference
$q_j^{overall}$	:	Feature for $j-$th source overall distortion
q_j^{target}	:	Feature for $j-$th source target distortion
s_b	:	Threshold index at $b-$th band
$\mathbf{s}_j$	:	$j-$th source in time domain
$\hat{\mathbf{s}}_j$	:	Estimated $j-$th source in time domain
v_p	:	Output weight of sigmoid p
w_a	:	Sine analysis window
$\mathbf{w}_j^c$	:	Mixing weights
$\mathbf{w}_p$	:	Input weight vector of sigmoid p
A	:	Normalizing factor
$\mathbf{A}$	:	Matrix of amplitude envelopes
$\mathbf{A}^\phi$	:	ϕ-th encoding matrix
A_{dB}	:	Response of ear
$\mathscr{A}$	:	Entropy
$A_{\mathscr{G}}$	:	Entropy of Gamma distribution
$A_{\mathscr{M}}$	:	Entropy of multinomial distribution
B_{bi}^s	:	Normalization factor

$\mathscr{B}$: Bound
C : Number of NTF layers
C_{IS} : IS divergence
C_{KL} : KL divergence
C_W : Weighted divergence
$\mathbf{D}^{IS}$: IS divergence in linear-frequency
$\mathbb{D}^{IS}$: IS divergence in log-frequency
$\mathbf{D}^{KL}$: KL divergence in linear-frequency
$\mathbb{D}^{KL}$: KL divergence in log-frequency
E_0 : Energy constant
E_{bi} : Pitch pattern in $b-$th critical band of $i-$th frame with internal noise
E^a_{bi} : Energy in $b-$th critical band of $i-$th frame
E^b_{bi} : Energy in $b-$th critical band of $i-$th frame after energy correction
E^f_{bi} : Energy in $b-$th critical band of $i-$th frame after time-domain spreading
E^s_{bi} : Spread bark domain energy in $b-$th critical band of $i-$th frame
$\tilde{E}^s_{bi}$: Excitation patterns in $b-$th critical band of $i-$th frame
E^{IN}_b : Internal noise
E^t_b : Excitation threshold
E_{min} : Minimum energy
$\overset{\rightarrow\tau}{\mathbf{E}}_A$: Matrix with elements $\langle A_{r,i-\tau}\rangle$
$\mathbf{E}^\tau_S$: Matrix with elements $\langle S^\tau_{kr}\rangle$
$\mathbf{F}$: Complex spectrogram
$\hat{\mathbf{F}}_j$: Estimated complex spectrogram of the $j-$th source
$\hat{\mathbf{F}}^c_j$: Estimated complex spectrogram of the $j-$th source at the $c-$th resolution
F_s : Sampling frequency
$\mathbf{G}$: Divergence matrix in critical bands
$\mathscr{G}$: Gamma distribution
H : Number of time frames in the grid
I : Number of time frames
J : Number of sources
K : Number of frequency components
$\mathbf{K}^c_j$: Kurtosis for the $j-$th source at the $c-$th resolution
$\mathbf{L}$: Transform matrix (from linear to Bark)
$\mathbf{L}^F$: Transform matrix (from linear to log-frequency)
$\mathbb{L}$: Transform matrix (from log-frequency to Bark)
$\mathscr{L}$: Marginal log-likelihood
$\overset{\rightarrow\tau}{\mathbf{L}}_A$: Matrix with elements $\exp(\langle \log A_{r,i-\tau}\rangle)$
$\mathbf{L}^\tau_S$: Matrix with elements $\exp(\langle \log S^\tau_{kr}\rangle)$
M : Number of log-frequency components
$\mathscr{M}$: Multinomial function
N : Step size of FIR filter
N_c : Number of critical bands

N_F	:	Frame length
$\mathbf{P}$	:	Weighting score matrix in critical bands
$\mathscr{PO}$	:	Poisson disribution
$\mathbf{Q}$	:	Gain matrix
R	:	Rank of factorization
$\mathbb{R}$	:	Set of real numbers
S	:	Spreading function
$\mathbf{S}^F$	:	Basis matrix in linear-frequency domain
$\mathbf{S}^\tau$	:	τ-th basis matrix
$\mathbb{S}$	:	Matrix of delayed source signals
T	:	Number of basis matrices
U	:	Number of frequency components in the grid
U_{bk}	:	Contribution of energy in $k-$th FFT bin for $b-$th band
V_k	:	Ear filtering response
$\mathbb{X}$	:	Log-frequency magnitude spectrogram
$\hat{\mathbb{X}}$	:	Estimated log-magnitude spectrogram
$\hat{\mathbb{X}}_j$	:	Estimated log-magnitude spectrogram for the $j-$th source
$\mathbf{X}$	:	Magnitude spectrogram matrix / tensor
$\mathbf{X}_c$	:	Magnitude spectrogram at $c-$th resolution
$\hat{\mathbf{X}}$	:	Estimated magnitude spectrogram
$\hat{\mathbf{X}}_j$	:	Estimated magnitude spectrogram for the $j-$th source
$\mathbf{X}_w$	:	Magnitude spectrogram after ear filtering
$\mathbf{W}$	:	Weighting score matrix in log-frequency domain
$\mathbf{W}^F$	:	Weighting score matrix in linear-frequency domain
$\mathbf{Y}_r$	:	Latent sources
$\hat{\mathbf{Y}}_r$	:	Estimated latent sources
Z_{bark}	:	Bark value
α_b	:	Constant for controlling decaying energy
$\alpha_{jbi,h}$	:	Distance between j-th and $h-$th sources at $b-th$ band, $i-$th frame
β	:	Mean of Gamma distribution
δ	:	Dirac-delta function
Δt	:	Time delay
ℓ	:	FIR filter length
$\eta_{A_{ir}}$	:	Step size of gradient descent algorithm for updating A_{ir}
$\eta_{Q_{cr}}$	:	Step size of gradient descent algorithm for updating Q_{cr}
$\eta_{S_{kr}}$	:	Step size of gradient descent algorithm for updating S_{kr}
Γ	:	Gamma function
Λ	:	Ratio of the original and estimated magnitude spectrograms
ν	:	Sparsity parameter for Gamma distribution
Ω	:	Grid
Φ	:	Joint distribution
Ψ	:	Digamma function

Σ_A : Matrix with elements $\sum_\tau \sum_k \langle Y^\tau_{r,k,i+\tau} \rangle$
Σ^τ_S : Matrix with elements $\sum_k Y^\tau_{rki}$
Θ : Number of encoding matrices
θ : Hyperparameters
θ^A : Parameters for **A**
θ^S : Parameters for **S**

LIST OF TABLES

Page

LIST OF FIGURES

Page

PERCEPTUAL AUDIO SOURCE SEPARATION BY SUBSPACE LEARNING

SUMMARY

Single-channel audio source separation problem occurs when a single observation of the mixture of a number of sources is available. In this type of problems, the aim is to estimate the individual sources constituting the mixture. In practice, it is very common to have multiple sources being active at the same time in a single recording. In such a situation, human listener has the ability to keep the attention to a single audio source in an adverse acoustical condition. However, the problem of automatically estimating several sources from one input signal is an under-determined and ill-posed challenging problem for the researchers.

Since the available information is not adequate to reconstruct the sources completely, solution of the single channel audio source separation problem relies on making appropriate assumptions about the sources. In this thesis, we develop models and algorithms that provide a framework to separate audio signals from single observation based on the subspace learning methods namely, Non-negative Matrix Factorization (NMF) and Non-negative Tensor Factorization (NTF) where the sources are not necessarily assumed as independent.

First, we introduce the perceptually weighted Non-negative Matrix Factor 2-D Deconvolution (PW-NMF2D) and the perceptually weighted Clustered NMF (PW-CNMF) methods to separate musical instruments in polyphonic music mixtures. Our approaches integrate the human perception into source separation, that allow to improve the perceptual quality of the separated sources. We weight the cost functions of the NMF and NMF2D methods measured by Kullback-Leibler (KL) and Itakura-Saito (IS) divergences, which are the special cases of $\beta-$divergence using a perceptual weighting score matrix. The weighting score matrix assigns a loudness sensation value per each time-frequency component based on the Perceptual Evaluation of Audio Quality (PEAQ) model defined in ITU-R BS. 1387. Thus, the contribution of the element-wise divergence to the cost function is increased/decreased using a high/low weighting score for the perceptually important/unimportant components. These perceptually enhanced PW-NMF2D and PW-CNMF methods constitute blind source separation methods which enhance the quality of the separated sources as perceived by humans.

Second, we investigate an adaptive time-frequency resolution based sound source separation method to separate either music or speech mixtures. In the literature, the NMF based source separation algorithms work at a fixed time-frequency resolution based Short Time Fourier Transform (STFT) magnitude or power spectrogram. Both speech and music signals have stationary and transients parts in different parts of

a single recording and they should be analyzed using appropriate windows. It is known that,the transients and percussives should be analyzed using a short window, whereas the stationary signals should be analyzed using long windows. If the signal is analyzed using a fixed time-frequency resolution, smearing occurs either in time or frequency resulting a poor separation of sources. In order to minimize the smearing caused by a fixed time-frequency resolution based analysis, we propose to separate the signals in several time-frequency resolutions under a supervised approach where the source bases are learned in advance from the training data. The separation is performed using Non-negative Tensor Factorization (NTF) where each layer of the tensor represents the same single channel mixture in various resolutions. The separated signals obtained from each resolution are then fused adaptively based on the maximal energy compaction principle method. This supervised multiresolution separation method estimates the sparsity of the sources obtained from different time-frequency resolutions and fuses them accordingly. This method is named as MultiResolution NTF (MR-NTF) and it increases the separability of sources by minimizing the smearing both in time and frequency.

Third, we investigate the clustering problem encountered in NMF based separation problems. It is known that the representation capability of the NMF algorithms is enhanced as the rank of the factorization is increased. If the rank is selected as greater than the number of sources, the order of the bases become random. Thus, it is required to cluster the bases into the sources. In this thesis, we overcome the clustering problem encountered in musical source separation problems by using two unsupervised approaches. The first unsupervised approach is based on the assumption that the timbre of a note played by an instrument is constant for the entire range of pitch. Using PW-NMF2D method, a single basis vector is extracted for a single instrument and shifted in frequency to approximate the bases of the other notes played by the same instrument. This assumption works only if the input signal is represented using a log-frequency magnitude spectrogram. The proposed PW-NMF2D method can capture both the temporal structure and the pitch change which occurs when an instrument plays different notes. In the second approach, in order to maximize the representation capability, we first extract the NMF bases with a high rank value, i.e.13, using perceptually enhanced NMF. The extracted bases are clustered into the same source using NMF2D if they share the same timbre at a different pitch.

Fourth, we focus on incorporating prior information into the separation scheme. Prior information about the sources can be integrated into the separation method using statistical methods. In our Bayesian Non-negative Matrix Factor Deconvolution approach (BNMFD), the original non-negative update equations of NMFD are obtained using an Expectation-Maximization (EM) algorithm for the Maximum Likelihood (ML) estimation of a conditionally Poisson model through data augmentation. The proposed BNMFD approach retains the attractive features of conventional NMFD such as easy implementation and monotonic convergence while opens up the way to develop more powerful models by incorporating the available prior information into the decomposition algorithm.

In MR-NTF method, we also incorporate the prior information available about the sources into the separation scheme in a supervised approach. This is achieved by learning the bases of the sources from the available training data of each source by

applying NMF on the training data at various time-frequency resolutions a prior to separation of the mixture. The learned source bases are fixed in the separation of the mixture signal, thus only the gains and the amplitude envelopes of the bases are updated iteratively through multiplicative update rules. This enables us to perform separation through a high-rank factorization by omitting the clustering problem.

We present conventional and perceptual evaluation of the proposed approaches on an extended dataset and compare the results to either the NMF, the CNMF and the NMF2D methods. We observe that the proposed models improve the quality of the sources both in terms of the conventional and the perceptual measures.

ALTUZAY ÖĞRENME İLE ALGISAL SES KAYNAK AYRIŞTIRMA

ÖZET

Birden fazla kaynaktan oluşan tek bir gözlem işaretinin mevcut olduğu kaynak ayrıştırma problemine, tek-kanaldan kaynak ayrıştırma problemi denilmektedir. Bu tür problemlerde amaç, karışımı oluşturan kaynakların tek tek elde edilmesidir. Pratikte, tek bir ses kaydında birden fazla kaynağın aynı anda etkin olması çok sık rastlanan bir durumdur. Böyle bir durumda, insan, ters akustik bir koşul içinde dikkatini tek bir kaynağa odaklayabilme yeteneğine sahiptir. Ancak, tek bir gözlemden birden fazla kaynağın otomatik olarak kestirilmesi problemi, "eksik tanımlanmış" ve kötü-konumlanmış bir problem olup, araştırmacıları uğraştırmaya devam etmektedir.

Bu tür "eksik tanımlanmış" problemlerde mevcut olan bilgi, kaynakları elde etmek için yeterli olmadığından, problemin çözümü kaynaklarla ilgili uygun varsayımlar yapılmasına dayanmaktadır. Bu tez çalışmasında, tek bir gözlem işaretinden karışımı oluşturan ses kaynaklarını ayrıştırmak için bir çerçeve sağlayan Negatif Olmayan Matris Ayrıştırma (NOMA) ve Negatif Olmayan Tensör Ayrıştırma (NOTA) yöntemlerine dayalı modeller ve algoritmalar geliştirilmektedir. NOMA ve NOTA, negatif olmayan verinin yaklaşık olarak ayrıştırılmasına olanak sağlayan parçalara-dayalı bir boyut düşürme tekniği olduğu için tercih edilmektedir.

İlk olarak, polifonik müzik karışımlarından müzik aletlerini ayrıştırmak için algısal olarak ağırlıklandırılmış Negatif Olmayan Çarpan 2-B Ters Evrişim (AA-NOÇ2BTE) ve algısal olarak ağırlıklandırılmış Öbeklenmiş Negatif Olmayan Matris Ayrıştırma (AA-ÖNOMA) önermekteyiz. Önerdiğimiz yöntemler, insan algısını kaynak ayrıştırma probleminde kullanarak, ayrıştırılan kaynak işaretlerinin algısal kalitesini artırmayı hedeflemektedir. NOMA ve NOÇ2BTE yöntemlerinin, $\beta-$ıraksayının özel halleri olan Kullback-Leibler (KL) ve Itakura-Saito (IS) ıraksayları ile ölçülen maliyet fonksiyonlarını ağırlıklandırmaktayız. Ağırlık matrisi, ITU-R BS. 1387 önerisinde tanımlanan Ses Kalitesinin Algısal Değerlendirilmesi (SKAD) modeline dayalı olarak hesaplanmakta olup, her zaman-frekans bileşenine bir seslilik duyumsama değeri atamaktadır. Böylelikle, ıraksayın her bir teriminin maliyet fonksiyonuna katkısı algısal olarak önemli/önemsiz bileşenler için yüksek/düşük ağırlık katsayıları kullanılarak artırılmaktadır/azaltılmaktadır. Algısal olarak ağırlıklandırılmış AA-NOÇ2BTE ve AA-ÖNOMA yöntemleri, ayrıştırılan kaynakların insan algısına göre kalitesini artıran gözükapalı kaynak ayrıştırma yöntemleri olarak önerilmektedir.

İkinci olarak, müzik ya da konuşma işaretlerinden oluşan karışımları, kaynaklarına ayırmak için uyarlamalı zaman-frekans çözünürlüğüne dayalı bir kaynak ayrıştırma yöntemi önermekteyiz. Bilimsel yazında, NOMA tabanlı ses kaynak ayrıştırma yöntemleri Kısa Zamanlı Fourier Dönüşümüne (KZFD) dayalı sabit zaman-frekans çözünürlüğünde çalışmaktadır. Tek bir konuşma ya da müzik kaydının farklı

kısımlarında geçiş bölümleri ve durağan bölümler bulunmaktadır ve bu bölümler uygun çerçeveler seçilerek incelenmelidirler. Bilinmektedir ki, geçiş kısımları ve vurmalı çalgılar kısa pencereler kullanılarak, durağan işaretler ise uzun pencereler kullanılarak analiz edilmelidir. Eğer, işaret sabit zaman-frekans çözünürlüğü kullanılarak incelenirse, zaman ya da frekansta yayılma oluşmakta ve kaynak ayrıştırma başarımı yetersiz olmaktadır. Sabit zaman-frekans çözünürlüğüne dayalı analiz kullanımına bağlı olarak oluşan yayılmayı en küçüklemek için, ayrıştırmayı birden fazla çözünürlükte gerçekleştirmeyi önermekteyiz. Önerilen yöntem, ayrıştırma öncesinde kaynak bazlarını eğitim kümesinden öğrenilmekte olduğundan eğiticili bir yaklaşımdır. Ayrıştırma, tensörün her bir katmanı farklı zaman-frekans çözünürlüğünde aynı karışım işaretini temsil etmek üzere Negatif Olmayan Tensör Ayrıştırma (NOTA) yöntemi ile gerçekleştirilmektedir. Her bir çözünürlükte ayrıştırılan kaynak işaretleri, büyükçe enerji sıkıştırma ilkesi yöntemine dayalı olarak uyarlamalı bir şekilde birleştirilmektedir. Önerilen eğiticili çok çözünürlüklü ayrıştırma yöntemi, farklı zaman-frekans çözünürlüklerinden kestirilen kaynak işaretlerinin seyrekliklerini hesaplayarak, seyrekliklerine göre farklı çözünürlük sonuçlarını birleştirmektedir. Bu yönteme Çok Çözünürlüklü NOTA (ÇÇ-NOTA) yöntemi ismi verilmektedir. ÇÇ-NOTA yöntemi, zaman ve frekansta yayılımı en küçükleyerek, kaynakların ayrılabilirliğini artırmaktadır.

Üçüncü olarak, NOMA tabanlı ayrıştırma problemlerinde ortaya çıkan öbekleme sorununu incelemekteyiz. Bilinmektedir ki, ayrıştırma düzeyi (rank) artırıldıkça, NOMA yönteminin temsil etme gücü artmaktadır. Eğer düzey, kaynak sayısından büyük seçilirse elde edilen bazların sıralanışının rastgele olduğu görülmektedir. Bu sebeple, elde edilen bazların kaynaklara öbeklenmesi gerekmektedir. Bu tez çalışmasında, müzik kaynak ayrıştırma problemlerindeki öbekleme problemi iki farklı yöntem kullanılarak ortadan kaldırılmaktadır. İlk gözükapalı öbekleme yöntemi, bir müzik aleti tarafından çalınan bir notanın tınısının, perdenin tüm eriminde sabit olduğu varsayımına dayanmaktadır. AA-NOÇ2BTE yöntemi kullanıldığında, her bir müzik aleti için bir baz vektörü elde edilmektedir. Bu baz vektörü, frekansta ötelenerek aynı müzik aleti tarafından çalınan diğer notaların bazları yaklaşık olarak elde edilmektedir. Bu varsayım, sadece giriş işaretinin log-frekans genlik spektrogramı kullanılarak temsil edildiği durumlarda geçerli olmaktadır. Önerilen AA-NOÇ2BTE yöntemi, bir müzik aletinin farklı notalar çalması durumunda, zamansal yapıyı ve perde değişimini yakalayabilmektedir. İkinci öbekleme yaklaşımında ise, NOMA yönteminin temsil etme gücünü artırmak için 13 gibi yüksek bir düzey seçilmekte ve algısal olarak güçlendirilmiş NOMA kullanılarak kaynakların bazları elde edilmektedir. Elde edilen bazlar, ikinci bir adımda NOÇ2BTE kullanılarak, eğer farklı perdede aynı tınıya sahiplerse aynı kaynağa öbeklenmektedirler.

Dördüncü olarak, ayrıştırma yöntemine önsel bilginin eklenmesi üzerinde durulmaktadır. Kaynaklar hakkında mevcut olan önsel bilgi, ayrıştırma yöntemine istatistiksel yaklaşımlarla eklenebilmektedir. Önerdiğimiz Bayesçi Negatif Olmayan Matris Çarpan Ters Evrişimi (BNOMÇTE) yaklaşımında, Negatif Olmayan Matris Çarpan Ters Evrişimi (NOMÇTE) yönteminin orijinal negatif olmayan güncelleme formülleri, veri artımı aracılığıyla koşullu Poisson modelinin En Büyük Olabilirlik Kestirimi (EBOK) için beklentiyi en büyükleyen bir algoritma kullanılarak elde edilmektedir. Önerilen BNOMÇTE yaklaşımı, NOMÇTE yönteminin kolay uygulanabilirlik ve

tekdüze yakınsaması gibi çekici özelliklerini korurken, ayrıştırma algoritmasına mevcut olan önsel bilginin eklenmesiyle geliştirilebilecek daha kuvvetli modeller için öncülük yapmaktadır.

ÇÇ-NOTA yaklaşımında da, kaynaklar hakkında var olan önsel bilgi eğiticili bir şekilde kaynak ayrıştırma yöntemine eklenmektedir. Bunun için, ÇÇ-NOTA kullanılarak her kaynak için mevcut olan eğitim verisinden, kaynakların bazları öğrenilmektedir. Öğrenilen bazlar karışım işaretinin ayrıştırılması esnasında sabitlenerek, kazançlar ve bazların genlik zarfları çarpımsal güncelleme kuralları aracılığıyla yinelemeli olarak güncellenmektedir. Bu yaklaşım, öbekleme problemini ortadan kaldırarak yüksek-düzeyli bir ayrıştırma yapma olanağı sağlamaktadır.

Genişletilmiş bir veri kümesinde, önerilen yöntemlerin geleneksel ve algısal değerlendirilmesi yapılmakta ve önerilen yöntemlerin başarımı NOMA, ÖNOMA ve NOÇ2BTE yöntemlerinin başarımıyla kıyaslanmaktadır. Önerilen modellerin, kaynakların kalitesini geleneksel ve algısal ölçütlere göre artırdığı gözlenmektedir.

1. INTRODUCTION

The term “audio source” is used to refer to an individual physical source or to an entity that humans perceive individually. While humans have advanced skills in “hearing out” individual sources from complex mixtures even in noisy conditions; the computer-based modeling of this process is a very difficult problem.

The process of estimating the individual sources from the mixture signal is called sound source separation. Sound source separation has many applications including music transcription, remixing, chord estimation, pitch modification, rendering of stereo CDs on multichannel devices, robust speech recognition, speaker separation from recorded meeting and video conferencing –“Cocktail party” problem. When the number of observations is less than the number of sources, the problem is called *under-determined*. The single-channel source separation problem is the most difficult case among the under-determined separation problems in which only a single observation exists. Moreover, according to the information used, the sound source separation methods are referred to as *blind* when any prior information about the sources and the mixing matrix is not available.

The focus of this thesis is to develop models and algorithms that provide a framework to separate audio signals from a single observation. In single-channel source separation, a human listener has the ability to keep the attention to a single audio source in an adverse acoustical condition. Although the available information in this situation is not sufficient for exact source reconstruction, it is possible to estimate or to obtain the approximate solutions by utilizing some information about the problem.

Many algorithms have been proposed for solving the source separation problem. These algorithms can be grouped into three main categories: the methods based on Computational Auditory Scene Analysis (CASA), statistical spatial methods and statistical spectral methods.

The CASA methods separate the mixed sources by grouping the mixture elements represented in time-frequency domain into the individual sources using psychoacoustical cues [1]. Conventionally, the grouping cues such as common onset/offset, modulation, harmonicity and spatial location are used. Periodicity is the main feature used by the CASA systems, thus the CASA systems cannot handle aperiodic signals.

Statistical spatial methods commonly use simple probabilistic source models for extracting the underlying sources of multi-channel recordings. Independent Component Analysis (ICA) [2], which remains in this group, has been successfully used to solve source separation problems in several application areas. ICA attempts to estimate the independent components by maximizing the statistical independence of the estimated components. However, it requires at least as many observations as the number of sources. Other statistical models aim to separate a larger number of sources using time-frequency binary masks under the assumption that they are disjoint in the time-frequency plane. Interchannel Intensity Difference (IID) and Interchannel Time Difference (ITD) are commonly used as time-frequency features to estimate the source azimuths and to derive optimal masks [3]. However, these methods fail in time-frequency regions where the sources highly overlap because the used binary masks generate burbling noise [3].

In the literature, statistical spectral methods have also been proposed for the separation of single channel mixtures. A favorite statistical spectral model for blind source separation is Independent Subspace Analysis (ISA) proposed by Hyvärinen and Hoyer [4]. After the work of Casey and Westner [5], ISA has been commonly used to refer the techniques which apply ICA to factor the spectrogram of a single channel audio signal to separate sound sources. When magnitude or power spectrograms are used as the input mixture signal, the extracted basis vectors represent the magnitudes or the powers of the important frequency components which are non-negative by definition. Thus, it may be advantageous to restrict the basis vectors and the gains to be entry-wise non-negative. The standard ISA algorithm, which is used for factorization, does not satisfy the non-negativity constraints. Moreover, it is shown that the non-negativity

restrictions alone are sufficient for the separation of the sources, without the explicit assumption of statistical independence [6].

Non-negative Matrix Factorization (NMF) [7] can be used as a simple but efficient spectral model for factorizing the input data into a linear combination of basis vectors with non-negativity constraints for both the basis and the encoding matrices. It has been extensively used in audio source separation, where the mixture is represented in the form of spectrogram. In [8], an algorithm based on factorizing the magnitude spectrogram of an input signal into sum of components with temporal continuity and sparseness criteria is proposed. Conventionally, NMF does not take into account the relative positions of each frequency basis vector, thus discards the temporal information. Smaragdis [9] introduces the Non-negative Matrix Factor Deconvolution (NMFD) algorithm, in which each instrument is modeled by a time-frequency signature that varies in intensity over time.

In NMF based source separation methods, the representation capability of the factorization is improved if the rank of the factorization is selected high. However, when the rank is higher than the number of sources, the bases extracted in random order should be clustered into sources. This problem is known as the permutation problem in the literature. In polyphonic music source separation, the convolutive NMF approaches like Shifted Non-negative Matrix Factorization (SNMF) [10] and Non-negative Matrix Factor 2-D Deconvolution (NMF2D) [11] avoid the need for clustering under the assumption that the notes belonging to a single source consist of translated versions of a single basis. Thus, SNMF and NMF2D generate a single basis for each instrument and it captures the other notes played by the same instrument by translating it in frequency. In another approach [12], the shift-invariance property is used in a computationally more effective two-step separation process which factorizes the magnitude spectrogram of the mixture into notes using a high rank approximation and clusters the notes into sources by applying the SNMF method in the log-frequency domain as an additional step. The method is referred as Clustered NMF (CNMF).

In all the mentioned methods, the perceptuality has been mostly ignored. Only a few works incorporate the psychoacoustics into the separation scheme. Among these, Virtanen [13] performs a perceptually weighted NMF algorithm for single channel

source separation which enhances the quality of the separated sources. However, this method does not consider the harmonicity of the notes which prevents separating mixtures of pitched instruments composed of several notes. In [14], a perceptually motivated Frequency-Domain Independent Component Analysis (FDICA) scheme is proposed in order to solve the permutation problem by filtering the perceptually irrelevant frequency components based on the masking properties of speech. The work proposed in [14] deals with the permutation problem rather than the separation quality. In [15], the NMF2D is directly applied on an auditory representation (sonogram) of mixed audio sources. The method significantly decreases the size of the data thus the computational complexity while improving the separation quality. However, clustering of the bases is performed using K-means, which complicates the source separation.

Another shortcoming of NMF based separation methods occur in the representation of the input mixture signal. The music and speech signals have both transient and stationary parts in a single recording. For an accurate analysis, each part should be analyzed using appropriate window lengths. However, most of the source separation methods including NMF work at a fixed time-frequency resolution. This causes pre-echoes at the transient parts and insufficient frequency resolution at the stationary parts [16]. There is no analytical way to find an optimal fixed analysis window length which can minimize these artifacts. In the literature, some methods have been proposed to enclose multiresolution approaches into source separation. In [17], it is stated that larger filter sizes are required to achieve high separation performance whereas drawbacks of the permutation problem can be reduced by using short filter sizes. In the literature, Adaptive Decorrelation Filtering (ADF) is proposed as a multistage processing scheme in which the filter size is increased at each stage in order to once self-align the permutations in the early stages with a small filter size and keep them aligned for increasing filter sizes [17]. In [18], the separability of the original signals is increased by adapting the time-frequency resolution of the Short Time Fourier Transform (STFT) to a specific mixture's characteristics. Hence, the smearing is minimized based on a transient detection measure for each time frame, but the proposed method ignores the smearing around each time-frequency component. In [19], the speakers learned at different time-frequency resolutions are separated by

applying NMF at each resolution and the separated sources are adaptively fused based on maximum energy compaction principle [16]. However, when the data is separately analyzed at each resolution as it is performed in [19], the information gained from the covariance of components extracted at different resolutions cannot be effectively integrated [20]. The method introduced in this thesis combines the parallel NMF factorizations used in [19] into a single NTF scheme, thus performs a joint optimization by fusing the information from various time-frequency resolutions. It is shown that the fusion of the available information in a joint optimization scheme enhances the quality of the separated sources.

Another shortcoming of the NMF based factorization techniques is that a single basis vector is typically far from being sufficient to represent a single audio source. On the other hand, factorization with a higher rank requires a method for clustering the basis vectors. However, it is difficult to develop a general clustering method which provides high performance for different contents. In [12, 21], the bases corresponding to the sources of pitched instruments are clustered using SNMF under the assumption that the notes belonging to a single instrument consist of translated versions of a single frequency basis vector which represents the typical frequency spectrum of any note played by the same instrument. Moreover, in [22], the bases of each speaker are learned in advance and the learned basis matrix is fixed throughout the NMFD iterations where the corresponding gains are updated iteratively. In [23], the instrumental sources are extracted from the mixed signal by clustering the harmonic structures obtained at different frames under the assumption that only a single source occurs at one time. To our knowledge, the unsupervised clustering methods have been usually proposed for pitched instruments and an unsupervised clustering method that works for all types of audio signals does not exist.

In the literature, performance of the audio source separation is conventionally reported in terms of distortion measures. Commonly used measures are the Signal-to-Distortion-Ratio (SDR), the Signal-to-Interference-Ratio (SIR) and the Signal-to-Artifacts-Ratio (SAR) [24]. These three measures are calculated based on the time-domain differences of the estimated and the original signals [24]. Although the quality of the separated sources can be evaluated as satisfactory in terms of

these distortion measures, they may sound perceptually annoying. In order to measure the perceptual quality of the separated sources, recently the quality metrics Overall Perceptual Score (OPS), Interference-related Perceptual Score (IPS) and Artifacts-related Perceptual Score (APS) are introduced [25]. Recent works focus on the development of new methods to incorporate the perceptuality both in the separation and the evaluation stages.

1.1 Main Contributions of the Thesis

In this thesis, we focus on four different problems encountered in single channel audio source separation. The first problem involved in this thesis is the incorporation of perceptuality into source separation. Existing separation algorithms may perform satisfactory based on some distortion measures although the separated sources may sound annoying to listeners. To our knowledge, perceptuality has been mostly ignored in audio source separation. Thus, it is aimed to develop new subspace learning based audio source separation methods that ensure perceptual quality of the separated sources. In the proposed blind source separation approaches, the objective function is formulated as a weighted divergence measured by the Kullback-Leibler (KL) and the Itakura-Saito (IS) divergences, which are special cases of the $\beta-$divergence [26]. A weighting score matrix is introduced in which a loudness sensation value is assigned to each critical band of each time frame based on the Perceptual Evaluation of Audio Quality (PEAQ) model defined in ITU-R recommendations BS. 1387 [27]. The psychoacoustic model in BS.1387 is preferred because it has been considered as the most effective objective measurement scheme of perceptual audio quality that takes into account the perceptual properties of the human ear more comprehensively [27]. Consequently, weighting the divergence in each critical band of each frame enables a subspace learning that mimics the human perception. The proposed factorization methods are formulated as two audio source separation schemes: Perceptually Weighted NMF2D (PW-NMF2D) and Perceptually Weighted Clustered NMF (PW-CNMF). The PW-NMF2D model extracts a single time-frequency signature for each source which is shifted in time and frequency to capture the time and frequency dependencies of the notes played by the same instrument. The PW-CNMF

method extracts several bases for each source using the perceptually weighted NMF method and then clusters the bases to the sources at an additional step. The performances of the introduced methods are compared to the existing NMF based methods. It is shown in [21,28] that the proposed methods outperform the conventional NMF based algorithms including SNMF [10], NMF2D [11] and CNMF [12] in audio quality by an amount of 0.5-2 dB based on the conventional measures SDR, SIR, SAR and 1-6 points based on the perceptual measures OPS, IPS and APS. Chapter 3 presents the details of the proposed PW-NMF2D and PW-CNMF schemes.

In the context of the thesis, the second problem involved in is audio source separation at adaptively tuned time-frequency resolutions. In the literature, most of the spectral algorithms perform separation using fixed time-frequency resolution based STFT. It is known that, the transients and the percussives should be analyzed using a short window, whereas the stationary signals should be analyzed using long windows. Otherwise, smearing occurs either in time or frequency domain resulting in a poor separation quality. In order to alleviate the artifacts caused by a fixed time-frequency based STFT, the separation is achieved in several resolutions by fusing the information in a joint optimization scheme using Non-negative Tensor Factorization (NTF) based on the KL divergence. The separated signals estimated at different resolutions are fused adaptively using the maximal energy compaction principle method proposed in [16]. The motivation of the information fusion after tensor factorization is to perform the separation in an adaptive time-frequency resolution which minimizes the smearing both in time and frequency [19]. This supervised multiresolution separation method is named as MultiResolution NTF (MR-NTF). It is shown by the performance evaluation on a large dataset that the proposed MR-NTF scheme enhances the separability of the sources, thus decreases the interference by minimizing the smearing. It improves the separation quality compared to the fixed time-frequency resolution based approaches by an amount of 1-4 dB based on the conventional measures SDR, SIR and 5-30 points based on the perceptual measures OPS and IPS.

In order to take advantages of the high rank factorization, solutions to the clustering problem are also investigated in the context of the thesis. If the rank of the factorization is selected as two in the NMF based approaches, each column of the basis matrix

corresponds to the basis vector of a single source. However, it is hard to capture all the characteristics of a source using a single basis. If the rank is increased, the order of the bases become random hence they must be clustered into individual sources. In the proposed methods, the clustering problem is solved by using two unsupervised approaches. An unsupervised clustering method is developed for audio source separation that combines the proposed perceptual weighting algorithm with the clustering techniques proposed in [12] and [29]. The first unsupervised approach PW-NMF2D [15,21] is performed based on the SNMF method proposed by FitzGerald *et al.* and the NMF2D method proposed by Schmidt *et al.* [29] under the assumption that the timbre of a note played by an instrument is constant for the entire range of the pitch. Thus, a single basis function for a single instrument can be obtained and shifted in frequency to approximate the bases of the other notes played by the same instrument. This assumption holds only if a logarithmic frequency resolution of the spectrogram is used to represent the input signal. The NMF2D method proposed in [29] is an extension of SNMF [30] method in time which performs the source separation on the log-frequency spectrogram that can capture both the temporal structure and the pitch change which occurs when an instrument plays different notes. In both [29] and [30], the log-frequency spectrogram is used since a pitch change corresponds to a displacement on the frequency axis. In order to maximize the representation capability, in PW-CNMF scheme, the NMF bases are extracted with a high rank value, i.e. 13 (the number of notes in an octave) using perceptually enhanced NMF. Then, an unsupervised clustering method which basically applies NMF2D, with a rank value equal to the number of sources, on the decomposed data is performed for reconstructing the individual audio sources. This clustering approach is based on the CNMF method proposed in [12]. The developed PW-CNMF approach [28] improves the quality of the separated signals significantly for music data since it clusters the decomposed bases into the same source if they share the same timbre at a different pitch. The clustering problem and the solutions to this problem are discussed in detail in Chapter 3.

The fourth and the last problem focused on this thesis is the incorporation of the available prior information into the decomposition algorithm. To achieve this, an

NMFD algorithm is investigated from a statistical perspective. In the proposed Bayesian Non-negative Matrix Factor Deconvolution (BNMFD) approach [31], the multiplicative update rules for NMFD under the KL divergence are derived using an Expectation-Maximization (EM) algorithm for Maximum Likelihood (ML) estimation of a conditionally Poisson model through data augmentation. It is shown that the proposed BNMFD scheme retains the attractive features of the conventional NMFD such as easy implementation and monotonic convergence while enabling to develop more powerful models by incorporating the available prior information into the decomposition algorithm. The proposed Bayesian approach to NMFD is presented in Chapter 5. The available prior information is also incorporated into the separation scheme by learning the bases of the sources from the available training data of each source. In the MR-NTF method, the pre-learned source bases are fixed during the separation of the mixture signal, thus only the gains and the amplitude envelopes for the basis vectors are updated iteratively. This enables to perform the separation through a high-rank factorization by omitting the clustering problem. This approach is named as informed factorization in the latest works [32]. The learning scheme for the bases are described in Chapter 4.

Chapter 2 presents the theoretical background by briefly describing the baseline NMF / NTF model, the signal resynthesis method and the performance evaluation criteria. In Chapter 3, the proposed PW-NMF2D and PW-CNMF algorithms are presented. The proposed multiresolution audio source separation framework MR-NTF is introduced in Chapter 4. Chapter 5 introduces the variational approach to NMFD method. Chapter 6 concludes the thesis.

2. THEORETICAL BACKGROUND

In this chapter, the mathematical background on the NMF, the NMF2D and the NTF models are presented. Following the basic NMF-based separation approaches, the resynthesis method used in the proposed approaches is described. Finally, the conventional as well as the perceptual performance measures are described.

2.1 Sound Source Separation by NMF Based Methods

Single-channel sound source separation techniques frequently use a time-frequency representation of the signal such as spectrogram derived by the STFT. In this section, a brief description of the NMF and the NMF2D methods which are widely used for single channel sound source separation are presented. Then, the NTF algorithm which is an extension of the NMF model and widely used for multi-channel separation applications is described.

2.1.1 Non-Negative Matrix Factorization

NMF [8] can be effectively used in the decomposition of a magnitude/power spectrogram $\mathbf{X} \in \mathbb{R}^{K \times I}$ into factors due to its ability to give an additive parts-based representation by factorizing $\mathbf{X}$ as

$$\mathbf{X} \approx \hat{\mathbf{X}} = \mathbf{S}^F \mathbf{A}, \tag{2.1}$$

where $\mathbf{S}^F \in \mathbb{R}^{K \times R}$ and $\mathbf{A} \in \mathbb{R}^{R \times I}$ are the non-negative basis and the encoding matrices, respectively. K, I and R are the number of frequency components, the number of time frames, and the rank of the factorization, respectively. The rank R is usually chosen such that $KR + RI \ll KI$, thus reducing the dimension of the data.

Figure 2.1 illustrates the factorization schematically. Each element X_{ki} of the observation matrix $\mathbf{X}$ is represented as the multiplication of the $k-$th row of the basis matrix $\mathbf{S}^F$ and the $i-$th column of the encoding matrix $\mathbf{A}$.

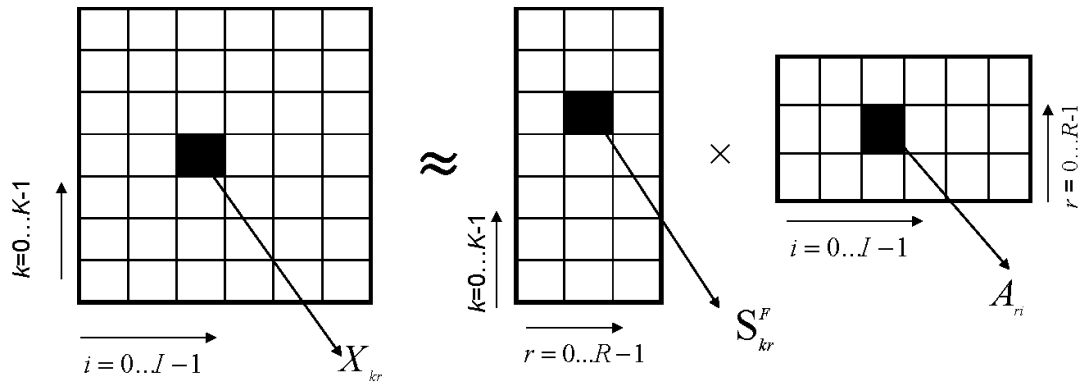


Figure 2.1: Source decomposition by NMF.

In the literature, the majority of the NMF based sound source separation methods perform factorization by minimizing a generalized KL divergence [8, 9]. The KL divergence is defined as:

$$C_{KL} = \sum_k \sum_i X_{ki} \log \frac{X_{ki}}{\hat{X}_{ki}} - X_{ki} + \hat{X}_{ki}, \tag{2.2}$$

where X_{ki} and $\hat{X}_{ki}$, $(k = 1 \ldots K, i = 1 \ldots I)$ denote the $k-$th frequency component of the $i-$th frame of the mixture magnitude spectrogram and the estimated mixture magnitude spectrogram, respectively.

The IS divergence [33] is also suggested due to the fact that its scale invariant formulation allows low and high energy components to have the same relative importance in the factorization. The IS divergence is defined as:

$$C_{IS} = \sum_k \sum_i \frac{X_{ki}}{\hat{X}_{ki}} - \log \frac{X_{ki}}{\hat{X}_{ki}} - 1. \tag{2.3}$$

It is shown that both the KL and the IS divergences are special cases of the $\beta-$ divergence [26]. The optimal parameters of the $\beta-$divergence depends on the statistics of the data being investigated. In [34], it is shown that the IS divergence outperforms the KL divergence on modeling the percussive components while the KL divergence has better performance on modeling the pitched instruments.

By applying gradient descent optimization procedure on the KL divergence, the update rules for $\mathbf{A}$ and $\mathbf{S}^F$ are obtained as [7]:

$$A_{ri} \leftarrow A_{ri} \frac{\sum_k S^F_{kr} X_{ki} / \hat{X}_{ki}}{\sum_k S^F_{kr}} \tag{2.4}$$

$$S^F_{kr} \leftarrow S^F_{kr} \frac{\sum_i A_{ri} X_{ki} / \hat{X}_{ki}}{\sum_i A_{ri}}. \tag{2.5}$$

Similarly, the update rules for $\mathbf{A}$ and $\mathbf{S}^F$ are obtained for the IS divergence based cost function as [33]:

$$A_{ri} \leftarrow A_{ri} \frac{\sum_k S^F_{kr} X_{ki} / \hat{X}^2_{ki}}{\sum_k S^F_{kr} / \hat{X}_{ki}} \tag{2.6}$$

$$S^F_{kr} \leftarrow S^F_{kr} \frac{\sum_i A_{ri} X_{ki} / \hat{X}^2_{ki}}{\sum_i A_{ri} / \hat{X}_{ki}}. \tag{2.7}$$

Some efforts have been made to improve the subspace representation provided by the basis vectors and/or time varying gains by imposing further constraints on the decomposition, such as sparsity [8, 35], spatial localization [35], and smoothness [8, 35]. Alternatively, the NMF methods can be derived in a probabilistic setting, based on the distribution of the data [31, 36]. In [36], Maximum-a-Posteriori (MAP) estimation of the basis ($\mathbf{S}^F$) and the gain ($\mathbf{A}$) matrices are derived under the assumption that $\mathbf{S}^F$ and $\mathbf{A}$ are independently determined by a Gaussian process connected by a link function. In our approach [31], a probabilistic interpretation and a full Bayesian inference are proposed for the NMFD model which also facilitate automatic model selection and determination of the sparsity criteria.

NMF provides a useful tool for analyzing data. However, it ignores the potential dependencies across successive columns or rows of its input $\mathbf{X}$. A regularly repeating pattern that spans multiple columns (rows) of $\mathbf{X}$ would have to be represented by NMF using multiple bases (gains) that describe the entire sequence.

2.1.2 Non-Negative Matrix Factor 2-D Deconvolution

Let an instrument be modeled by a specific time-frequency signature modulating the sound of the instrument over time τ. When an instrument plays a note at a certain time τ with a certain pitch ϕ, it corresponds to displacing the time-frequency signature on the log-frequency axis. This forms the basic idea of the NMF2D model proposed by Schmidt *et al.* [11], [37].

In this model, the log-magnitude spectrogram is used to represent the observation signal. To obtain the log-magnitude spectrogram, each frame is transformed into frequency domain by taking the Discrete Fourier Transform (DFT) and then grouping the spectrogram bins into the logarithmically spaced frequency bins. Only

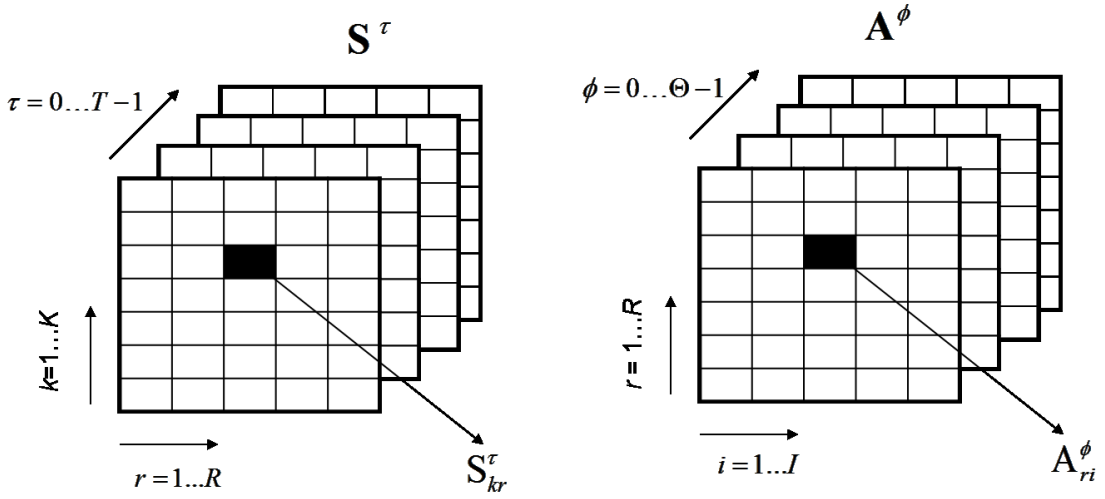


Figure 2.2: Source decomposition by NMF2D.

positive frequencies are retained by taking the absolutive values of the log-frequency spectrogram resulting in $\mathbb{X}_{mi}$, where $m = 0 \ldots M-1$ is the log-frequency index and $i = 0 \ldots I-1$ is the frame index .

In order to model both time and frequency characteristics of an audio spectrogram efficiently, the conventional NMF model is extended to be a 2-D convolution of $\mathbf{S}^{\tau}$, which depends on time τ, and $\mathbf{A}^{\phi}$, which depends on pitch ϕ. This forms the NMF2D model proposed in [11]:

$$\mathbb{X} \approx \hat{\mathbb{X}} = \sum_{\tau} \sum_{\phi} \mathbf{S}^{\tau} \mathbf{A}^{\phi}, \tag{2.8}$$

where $\downarrow \phi$ denotes the downward shift operator which moves each element in the matrix ϕ rows down, and $\rightarrow \tau$ denotes the right shift operator which moves each element in the matrix τ columns to the right [11]. In this model, $\mathbb{X}$ is the log-frequency magnitude spectrogram obtained through multiplication of the magnitude spectrogram $\mathbf{X}$ with a transform matrix $\mathbf{L}^{F} \in \mathbb{R}^{M \times K}$ which yields $\mathbb{X} = \mathbf{L}^{F}\mathbf{X}$ [38].

Each element in $\hat{\mathbb{X}}$ is defined as:

$$\hat{\mathbb{X}}_{mi} = \sum_{\tau} \sum_{\phi} \sum_{r} S^{\tau}_{m-\phi,r} A^{\phi}_{r,i-\tau}, \tag{2.9}$$

where r denotes the component index; ϕ shifts each element in the matrix ϕ rows down, and τ shifts each element in the matrix τ columns to the right. The leftmost columns of $A^{\phi}_{r,i-\tau}$ and the uppermost rows of $S^{\tau}_{k-\phi,r}$ are set to zero so as to maintain the original size of the input.

Effectively, what happens is that the set of $r-$th columns of $\mathbf{S}^{\tau}$ defines a two-dimensional basis. This basis is shifted and scaled by convolution across the axis of τ with the $r-$th row of $\mathbf{A}^{\phi}$. The resulting reconstruction is a summation of all the basis convolution results for each of the R bases. Figure 2.2 explains the notation.

In terms of computational complexity, the NMF2D technique depends mostly on the parameters $\tau = 0 \cdots T-1$ and $\phi = 0 \cdots \Theta - 1$. If $\Theta = 1$, then it reduces to the NMFD method proposed in [9], and if $\Theta = 1, T = 1$ it reduces to the conventional NMF method. Otherwise it is burdened with extra matrix updates equivalent to one NMF per unit of $T \times \Theta$.

In the literature, the NMF2D based algorithms are used to estimate the non-negative basis vectors and mixing matrices iteratively based on the minimization of a cost function such as the KL and the IS divergences.

By applying gradient descent optimization procedure on the KL divergence, the update rules for $\mathbf{A}^{\phi}$ and $\mathbf{S}^{\tau}$ are obtained as:

$$A^{\phi}_{ri} \leftarrow A^{\phi}_{ri} \frac{\sum_m \sum_\tau \frac{\mathbb{X}_{m,i+\tau}}{\hat{\mathbb{X}}_{m,i+\tau}} S^{\tau}_{m-\phi,r}}{\sum_m \sum_\tau S^{\tau}_{m-\phi,r}} \tag{2.10}$$

$$S^{\tau}_{mr} \leftarrow S^{\tau}_{mr} \frac{\sum_\phi \sum_i \frac{\mathbb{X}_{m+\phi,i}}{\hat{\mathbb{X}}_{m+\phi,i}} A^{\phi}_{r,i-\tau}}{\sum_\phi \sum_i A^{\phi}_{r,i-\tau}}. \tag{2.11}$$

Similarly, the update rules based on the IS divergence are obtained as:

$$A^{\phi}_{ri} \leftarrow A^{\phi}_{ri} \frac{\sum_m \sum_\tau \frac{\mathbb{X}_{m,i+\tau}}{\hat{\mathbb{X}}^2_{m,i+\tau}} S^{\tau}_{m-\phi,r}}{\sum_m \sum_\tau \frac{S^{\tau}_{m-\phi,r}}{\hat{\mathbb{X}}_{m,i+\tau}}} \tag{2.12}$$

$$S^{\tau}_{mr} \leftarrow S^{\tau}_{mr} \frac{\sum_\phi \sum_i \frac{\mathbb{X}_{m+\phi,i}}{\hat{\mathbb{X}}^2_{m+\phi,i}} A^{\phi}_{r,i-\tau}}{\sum_\phi \sum_i \frac{A^{\phi}_{r,i-\tau}}{\hat{\mathbb{X}}_{m+\phi,i}}}. \tag{2.13}$$

These updates are applied iteratively until the two factors converge. The derivation of the update rules can be found in Appendix A.1 and Appendix A.2.

To give an idea, the source matrices $\mathbf{S}^{\tau}$ and the basis matrices $\mathbf{A}^{\phi}$ obtained by minimizing the KL divergence for each factor are displayed in Figure 2.3 along with the mixture spectrogram. The R columns of $\mathbf{S}^{\tau}$ obtained by NMF2D represent the

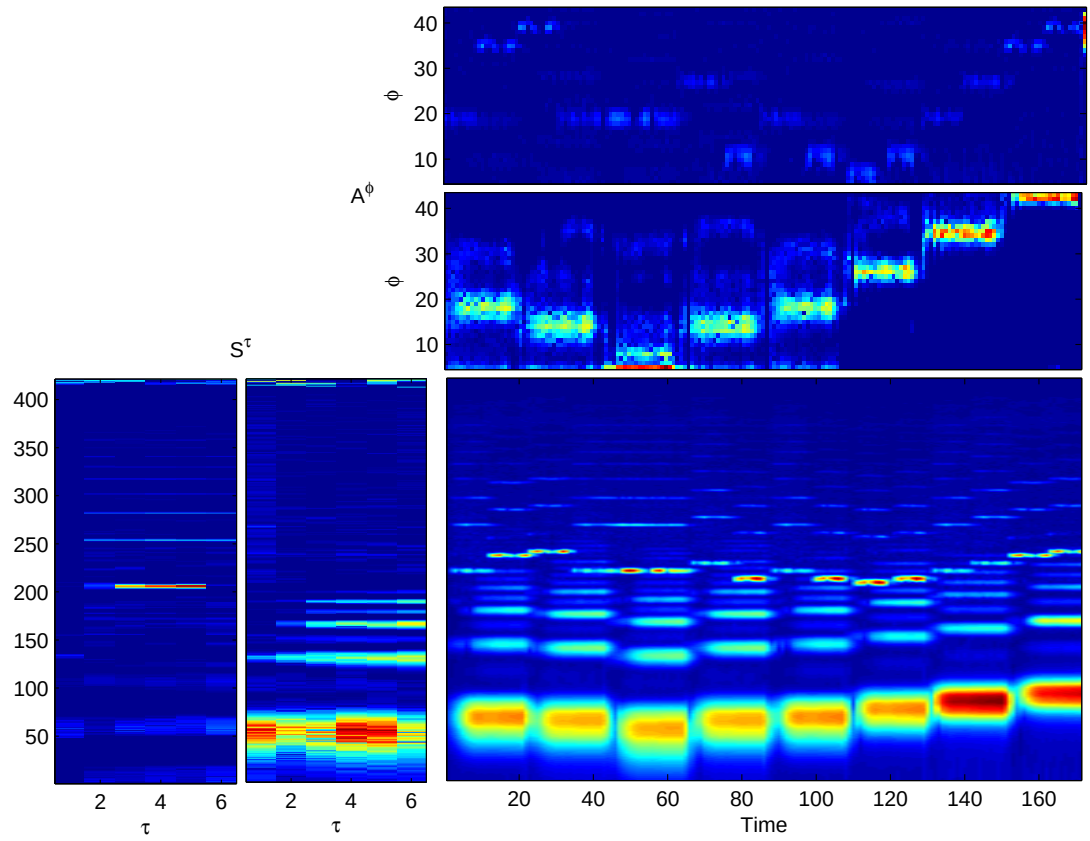


Figure 2.3: Factorization of a piece of music using NMF2D. The two time-frequency plots on the left are $\mathbf{S}^{\tau}$ for each factor, i.e. the time-frequency signature of the two sources. The two time-pitch plots on the top are $\mathbf{A}^{\phi}$ for each factor showing how the sources are placed in time and pitch.

dominant spectral patterns contained in the input whereas their weights $\mathbf{A}^{\phi}$ correspond to their temporal profiles.

After convergence, the log-frequency magnitude spectrogram $\hat{\mathbb{X}}_r$ for the $r-$th component is estimated as:

$$\hat{\mathbb{X}}_r \approx \sum_{\tau}\sum_{\phi} S^{\tau}_{m-\phi,r} A^{\phi}_{r,i-\tau}. \tag{2.14}$$

In the NMF2D model, each component r corresponds to a source j. These estimated log-frequency magnitude spectrograms are mapped back to linear frequency domain by $\hat{\mathbf{X}}_j = \mathbf{L}^{F\top}\hat{\mathbb{X}}_j$ [38].

In [9], it is shown that this type of analysis is efficient for finding the salient spectral sequences contained in auditory scenes and can be further employed to extract them.

In Chapter 3, the performance of NMF2D is evaluated as applied on perceptual audio source separation.

2.1.3 Non-Negative Tensor Factorization

A model which imposes non-negativity on factor matrices and the input tensor is called the Non-negative Tensor Factorization (NTF). The Parallel Factor Analysis (PARAFAC) algorithms decompose the input tensor into a sum of multi-linear terms in a way analogous to the bilinear matrix decomposition [20]. PARAFAC usually does not impose any orthogonality constraints. A non-negative version of PARAFAC was first introduced by Carroll et al. [39].

In this thesis, we use an algorithm which optimizes a generalized KL divergence between the observed data and the NTF model:

$$D(\mathbf{X}||\hat{\mathbf{X}}) = \sum_{c=1}^{C}\sum_{k=1}^{K}\sum_{i=1}^{I}\left(X_{cki}\log\frac{X_{cki}}{\hat{X}_{cki}} - X_{cki} + \hat{X}_{cki}\right), \tag{2.15}$$

where c,k,i are indices over the channel, the frequency bin and the time frame, respectively. $\mathbf{X} \in \mathbb{R}^{C\times K\times I}$ is a 3-way tensor containing the C magnitude spectrograms of the mixture and $\hat{\mathbf{X}}$ is an approximation to $\mathbf{X}$. Note that, NMF is a particular case of NTF, when the tensor has a single layer ($C = 1$). Besides enabling to factorize higher-order matrices, the main advantage of NTF over NMF is that a tensor factorization is unique [40].

The signal model in (**2.15**) can be written as:

$$X_{cki} \approx \hat{X}_{cki} = \sum_{r=1}^{R} Q_{cr}S_{kr}A_{ir}, \tag{2.16}$$

where $\mathbf{Q} \in \mathbb{R}^{C\times R}$ is a matrix containing the gains of each component in each layer, $\mathbf{S} \in \mathbb{R}^{K\times R}$ is the basis matrix containing the frequency basis vectors and $\mathbf{A} \in \mathbb{R}^{I\times R}$ is the corresponding amplitude envelopes for the frequency basis vectors. NTF is widely used in multichannel source separation applications where each tensor slice $\mathbf{X}_c$ represents the magnitude spectrogram of the mixture in the $c-$th channel. Each rank-1 spectrogram $\mathbf{S}_r\mathbf{A}_r^\top$ defines a component with its own spatial que $\mathbf{Q}_r$. Our approach differs from the methods in the literature by representing the single observation in a tensor structure where each layer $\mathbf{X}_c$ of the tensor is the magnitude spectrogram of

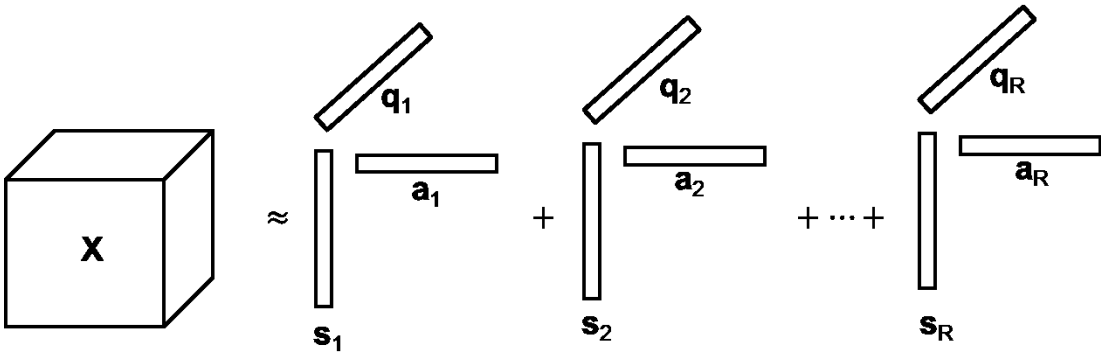


Figure 2.4: Non-negative Tensor Factorization of a 3-way array.

the same mixture in a different time-frequency resolution. It is proposed to learn a single set of frequency basis vectors **S** which can be used to describe each resolution of the input signal, a corresponding set of amplitude basis vectors **A**, and a set of corresponding gains **Q** which decides in which proportion a given pair of frequency and amplitude basis vectors exist in each resolution.

In this work, the most used tensor model, PARAFAC structure is used. The PARAFAC structure allows representing a tensor of order *n* by means of *n* matrices, called matrix factors which is illustrated in Figure 2.4. The factor matrices refer to the combination of the vectors from the rank-one components, i.e., $\mathbf{S} = [\mathbf{s}_1 \quad \mathbf{s}_2 \cdots \mathbf{s}_R]$ and likewise for **Q** and **A**.

Eliminating the terms in the cost function (**2.15**) which are constant yields:

$$D(\mathbf{X}||\hat{\mathbf{X}}) = \sum_{c=1}^{C}\sum_{k=1}^{K}\sum_{i=1}^{I}\left(-X_{cki}\log\hat{X}_{cki} + \hat{X}_{cki}\right) \tag{2.17}$$

$$= \sum_{c=1}^{C}\sum_{k=1}^{K}\sum_{i=1}^{I}\left(-X_{cki}\log\Big(\sum_{r=1}^{R}Q_{cr}S_{kr}A_{ir}\Big) + \sum_{r=1}^{R}Q_{cr}S_{kr}A_{ir}\right). \tag{2.18}$$

Taking the gradient of the cost function in (**2.18**) with respect to **Q** yields the following update equation:

$$Q_{cr} = Q_{cr} + \eta_{Q_{cr}}\left[\sum_{k=1}^{K}S_{kr}\sum_{i=1}^{I}\Lambda_{cki}A_{ir} - \sum_{k=1}^{K}S_{kr}\sum_{i=1}^{I}A_{ir}\right], \tag{2.19}$$

where

$$\Lambda_{cki} = X_{cki}/\hat{X}_{cki}. \tag{2.20}$$

Equation (**2.19**) can be converted into a multiplicative update rule by setting η_Q as:

$$\eta_{Q_{cr}} = Q_{cr}/\Big(\sum_{k=1}^{K}S_{kr}\sum_{i=1}^{I}A_{ir}\Big). \tag{2.21}$$

The multiplicative update rule for **Q** is then given by:

$$Q_{cr} \leftarrow Q_{cr}\left(\sum_{k=1}^{K} S_{kr}\sum_{i=1}^{I}\Lambda_{cki}A_{ir}\right)/\left(\sum_{k=1}^{K} S_{kr}\sum_{i=1}^{I}A_{ir}\right). \tag{2.22}$$

The update equations for **S** and **A** can be derived in a similar manner. Taking the gradient of the cost function in (**2.18**) with respect to **S** yields the following update equation:

$$S_{kr} = S_{kr} + \eta_{Skr}\left[\sum_{c=1}^{C} Q_{cr}\sum_{i=1}^{I}\Lambda_{cki}A_{ir} - \sum_{c=1}^{C} Q_{cr}\sum_{i=1}^{I}A_{ir}\right]. \tag{2.23}$$

Equation (**2.23**) can be converted into a multiplicative update rule by setting η_S as:

$$\eta_{Skr} = S_{kr}/\big(\sum_{c=1}^{C} Q_{cr}\sum_{i=1}^{I}A_{ir}\big). \tag{2.24}$$

The multiplicative update rule for **S** is then given by:

$$S_{kr} \leftarrow S_{kr}\left(\sum_{c=1}^{C} Q_{cr}\sum_{i=1}^{I}\Lambda_{cki}A_{ir}\right)/\left(\sum_{c=1}^{C} Q_{cr}\sum_{i=1}^{I}A_{ir}\right). \tag{2.25}$$

Taking the gradient of the cost function in (**2.18**) with respect to **A** yields the following update equation:

$$A_{ir} = A_{ir} + \eta_{Air}\left[\sum_{c=1}^{C} Q_{cr}\sum_{k=1}^{K}\Lambda_{cki}S_{kr} - \sum_{c=1}^{C} Q_{cr}\sum_{k=1}^{K}S_{kr}\right]. \tag{2.26}$$

Equation (**2.26**) can be converted into a multiplicative update rule by setting η_A as:

$$\eta_{Air} = A_{ir}/\big(\sum_{c=1}^{C} Q_{cr}\sum_{k=1}^{K}S_{kr}\big). \tag{2.27}$$

The multiplicative update rule for **A** is then given by:

$$A_{ir} \leftarrow A_{ir}\left(\sum_{c=1}^{C} Q_{cr}\sum_{k=1}^{K}\Lambda_{cki}S_{kr}\right)/\left(\sum_{c=1}^{C} Q_{cr}\sum_{k=1}^{K}S_{kr}\right). \tag{2.28}$$

The use of the multiplicative updates ensures that once $\mathbf{Q},\mathbf{S}$ and **A** are randomly initialized to non-negative values, the resulting matrices will only contain non-negative values.

Upon convergence, the magnitude spectrogram of each source j is obtained as:

$$\hat{X}_{jki} = \sum_{c=1}^{C}\sum_{r=1}^{R} Q_{jcr}S_{jkr}A_{jir} \tag{2.29}$$

where $\mathbf{Q}_j$, $\mathbf{S}_j$ and $\mathbf{A}_j$ stand for the gains, the bases and the basis envelopes for the $j-$th source, respectively.

2.2 Resynthesis of Audio Sources

Several methods are proposed for resynthesis using the recovered source spectrogram which are compared in [41]. Commonly used one applies the original phase information to $\hat{\mathbf{X}}_j$ hence inverts the resultant spectrogram to the time domain. This requires a prior knowledge of the sources. However, this information is not provided in blind approaches. Another method is the binary masking method used by Schmidt *et al.* in [11] where spectrogram masks are constructed for each instrument by assigning each time-frequency component to the instrument with the highest power at that bin [11]. This assumption does not hold well for musical signals in which the instruments usually play in harmony with one and other, resulting in overlapping partials. In the third method, the recovered source spectrogram $\hat{\mathbf{X}}_j$ is used to filter the original spectrogram $\mathbf{F}$. This method has an advantage over the binary masking method [11] hence gives better results when dealing with musical signals [41]. Another method estimates the spectrogram of the estimated source by filtering the original spectrogram with the element-wise ratio of the estimated magnitude spectrogram of the source to the sum of the estimated magnitude spectrograms of all sources. This method works better than the previously mentioned methods by assigning the proportional quantity of amplitude to each time-frequency cell to the source in relation to the total amplitude of all sources in this cell.

In the proposed approaches, based on the reconstructed individual magnitude source spectra $\hat{\mathbf{X}}_j$, spectrogram of each source is constructed by using the resynthesis method introduced in [41, 42]

$$\hat{F}_{jki} = F_{ki}\frac{\hat{X}_{jki}^2}{\sum_j \hat{X}_{jki}^2}, \tag{2.30}$$

where F_{ki} is the $k-$th frequency component of the $i-$th frame of the mixture spectrogram, $\hat{X}_{jki}$ is the $k-$th frequency component of the $i-$th frame of the estimated magnitude spectrogram of the $j-$th source. (**2.30**) represents an adapted Wiener filtering approach proposed by Benaroya *et al.* in [42] for stationary Gaussian sources. While music and speech signals can be considered as approximately stationary on a frame by frame basis, this approach can be applied as a spectral filter ensuring the unaccounted energy in the input mixture to be redistributed to the resulting sources in proportion. In [41], it is shown that the Wiener filtering approach results in better

perceptual quality compared to the previously mentioned methods. Finally, the time domain estimates $\hat{\mathbf{s}}_j$ of the sources are obtained by applying Inverse STFT (ISTFT) on the estimated coefficients $\hat{F}_{jki}$.

2.3 Performance Evaluation Measures

In order to evaluate the perceived quality of the separated sound sources obtained using the proposed algorithms, the conventional measures proposed in [24] are used. Most of the existing methods report the performance using the BSSEVAL toolbox [43] in terms of SDR, SIR and SAR [24]. Therefore, the performance is reported using these measures in order to make comparison with the conventional methods tractable. However, since our objective is the evaluation of perceptual quality of the separated sources, the performance is also reported in terms of perceptual measures recently described in [25]. The measures quantify the performance with a scoring mechanism where the scores are named as OPS, IPS and APS. Similar to the conventional measures, the perceptual quality measures are also designed to evaluate the performance of audio separation in terms of interference, artifacts, and total distortion but as perceived by humans. The PEASS Toolkit [44] is used to examine the perceptual quality of the separated sources.

The computation of these measures are defined step by step in detail starting from the reconstructed signal. The reconstructed $j-$th source signal denoted by $\hat{\mathbf{s}}_j$ can be computed as:

$$\hat{\mathbf{s}}_j = \mathbf{s}_j + \mathbf{e}_j^{target} + \mathbf{e}_j^{interf} + \mathbf{e}_j^{artif}, \tag{2.31}$$

where $\mathbf{s}_j$ is the original source signal; $\mathbf{e}_j^{target}$, $\mathbf{e}_j^{interf}$ and $\mathbf{e}_j^{artif}$ account for the target distortion, the interferences of the unwanted sources and the deformations induced by the separation algorithm that are not allowed (artifacts), respectively. In [24], it is assumed that these components are linearly distorted versions of the true source signals, where distortion is modeled through time-invariant Finite Impulse Response (FIR) filters. The coefficients of these filters are computed by two nested least-square projections: first, the distortion signal is projected onto the subspace spanned by delayed versions of all source signals $s_h(t-\Delta t), 1 \leq h \leq J, 0 \leq \Delta t \leq \ell - 1$ (J is the number of sources), so as to obtain $e_j^{target}(t) + e_j^{interf}(t)$ at time t. Then it is further

projected on the smaller subspace spanned by delayed versions of the target signal $s_j(t-\Delta t), 0 \leq \Delta t \leq \ell - 1$, so as to obtain $e_j^{target}(t)$ alone. At the last step, $e_j^{artif}(t)$ is defined as the residual [24].

If these components are assumed to be linearly distorted versions of the original source signals [15,24,31], they cannot effectively estimate the expected components perceived by humans. The method proposed in [25] proposes to use the auditory time-frequency resolution instead of the linear frequency resolution to calculate the distortions. For this issue, the estimated source signal $\hat{s}_j(t)$ and all original source signals $s_h(t), 1 \leq h \leq J$ are first split into subband signals $\hat{s}_{jb}(t)$ and $s_{hb}(t)$ indexed by b using a bank of 4th-order gammatone filters [25]. The center frequencies are linearly spaced on the auditory-motivated Equivalent Rectangular Bandwidth (ERB) scale from 20 Hz to the Nyquist frequency.

In each subband b, the estimated source signal $\hat{s}_{jb}(t)$ and the delayed true source signals $s_{hb}(t-\Delta t), 1 \leq h \leq J$, $-\ell/2 \leq \Delta t \leq \ell/2$ (ℓ is the filter length), are partitioned into overlapping time frames indexed by i via

$$\hat{s}_{jbi}(t) = w_a(t)\hat{s}_{jb}(t - iN) \quad \textbf{(2.32)}$$

$$s_{hbi}^{\Delta t}(t) = w_a(t)s_{hb}(t - iN - \Delta t), \quad \textbf{(2.33)}$$

where w_a denotes the sine analysis window with length $4N$ and step size N.

The distortion components are estimated by an additional filtering in each subband of each time frame via multichannel time-invariant FIR filtering of the target source signal and the interfering source signals using:

$$e_{jbi}^{target}(t) = \sum_{\Delta t=-\ell/2}^{\ell/2} \alpha_{jbi,j}(\Delta t)s_{jbi}^{\Delta t}(t) \quad \textbf{(2.34)}$$

$$e_{jbi}^{interf}(t) = \sum_{h \neq j} \sum_{\Delta t=-\ell/2}^{\ell/2} \alpha_{jbi,h}(\Delta t)s_{hbi}^{\Delta t}(t) \quad \textbf{(2.35)}$$

$$e_{jbi}^{artif}(t) = \hat{s}_{jbi}(t) - s_{jbi}(t) - e_{jbi}^{target}(t) - e_{jbi}^{interf}(t). \quad \textbf{(2.36)}$$

Unlike [24], centered FIR filters are used and the interference component explicitly excludes the target source j. The filter coefficients are computed by the least-squares projection of the distortion $\hat{s}_{jbi}(t) - s_{jbi}(t)$ onto the subspace spanned by the delayed

source signals $s_{hbi}^{\Delta t}(t), 1 \leq h \leq J, -\ell/2 \leq \Delta t \leq \ell/2$. The vector of coefficients α_{jbi} is defined as $\alpha_{jbi} = \mathbb{S}_{bi}^{+}(\hat{\mathbf{s}}_{jbi} - \mathbf{s}_{jbi})$ where $\hat{\mathbf{s}}_{jbi}$ and $\mathbf{s}_{jbi}$ are, respectively, $4N \times 1$ vectors of the estimated and the original source signals, $\mathbb{S}_{bi}$ is the $4N \times (\ell+1)J$ matrix of delayed original source signals and $^{+}$ denotes matrix pseudo-inversion.

Full-duration distortion components are reconstructed from the time-localized components in each subband using Overlap-Add (OLA):

$$e_{jb}^{target}(t) = \sum_i w_s(t - iN) e_{jbi}^{target}(t - iN) \quad \textbf{(2.37)}$$

$$e_{jb}^{interf}(t) = \sum_i w_s(t - iN) e_{jbi}^{interf}(t - iN) \quad \textbf{(2.38)}$$

$$e_{jb}^{artif}(t) = \sum_i w_s(t - iN) e_{jbi}^{artif}(t - iN), \quad \textbf{(2.39)}$$

where $\mathbf{w}_s$ is a sine synthesis window of length $4N$. The full band distortion components $\mathbf{e}_j^{target}, \mathbf{e}_j^{interf}$ and $\mathbf{e}_j^{artif}$ are obtained using the gammatone synthesis filters [25]. As shown in [25], using the auditory filterbanks removes the sources from the artifacts components whereas it can still be heard with the state-of-the-art one [24]. It also enhances the relevance of the target distortion and the interference components.

The state-of-the-art distortion measures are defined as [24, 25]:

$$SDR \equiv 10 \log_{10} \frac{\|\mathbf{s}_j\|^2}{\|\hat{\mathbf{s}}_j - \mathbf{s}_j\|^2}, \quad \textbf{(2.40)}$$

$$SIR \equiv 10 \log_{10} \frac{\|\mathbf{s}_j + \mathbf{e}_j^{target}\|^2}{\|\mathbf{e}_j^{interf}\|^2}, \quad \textbf{(2.41)}$$

$$SAR \equiv 10 \log_{10} \frac{\|\mathbf{s}_j + \mathbf{e}_j^{target} + \mathbf{e}_j^{interf}\|^2}{\|\mathbf{e}_j^{artif}\|^2}. \quad \textbf{(2.42)}$$

As it is stated in [25], these energy ratios do not always fit the perceptual salience of each component within the estimated source. One of these issues is that, the low frequency components affect the energy ratios more than perception. Another issue is that the auditory masking of soft distortion components by the target signal or by louder distortion components is not taken into account. In order to overcome these issues, a two-step approach is proposed in [25]. First, the salience of each distortion component is assessed using auditory model-based metrics and then the resulting features are combined by nonlinear mapping which produces the objective measures:

- The Overall Perceptual Score (OPS)

- Interference-related Perceptual Score (IPS)
- Artifacts-related Perceptual Score (APS).

The Perceptual Similarity Measure (PSM) provided by the PErception MOdel based Quality (PEMO-Q) auditory model [45] is employed to calculate the perceptual salience of the overall distortion and of each specific distortion component by comparing the estimated source signal $\hat{\mathbf{s}}_j$ with itself minus the distortion under consideration. This yields four features:

$$q_j^{overall} = \text{PSM}(\hat{\mathbf{s}}_j, \mathbf{s}_j) \quad \textbf{(2.43)}$$

$$q_j^{target} = \text{PSM}(\hat{\mathbf{s}}_j, \mathbf{s}_j \mathbf{e}_j^{target}) \quad \textbf{(2.44)}$$

$$q_j^{interf} = \text{PSM}(\hat{\mathbf{s}}_j, \mathbf{s}_j \mathbf{e}_j^{interf}) \quad \textbf{(2.45)}$$

$$q_j^{artif} = \text{PSM}(\hat{\mathbf{s}}_j, \mathbf{s}_j \mathbf{e}_j^{artif}). \quad \textbf{(2.46)}$$

A nonlinear mapping is applied to combine these features into a scalar measure for each grading task and to adapt the feature scale to the subjective grading scale. For a given task, the feature vector $\mathbf{q}_j$ involves either the four features $\mathbf{q}_j = [q_j^{overall}, q_j^{target}, q_j^{interf}, q_j^{artif}]$ or a subset of these features.

A one-hidden-layer feedforward neural network composed of P sigmoids is employed to map each feature vector $\mathbf{q}_j$ into an OPS, IPS or APS score $f(\mathbf{q}_j)$ via the function

$$f(\mathbf{q}_j) = \sum_{p=1}^{P} v_p g(\mathbf{w}_p^\top \mathbf{q}_j + b_p) \quad \textbf{(2.47)}$$

where $g(x) = 1/(1+\exp(-x))$ is the sigmoid function and v_p, $\mathbf{w}_p$ and b_p denote the output weight, the vector of input weights and the bias of sigmoid p.

In order to measure the perceptual quality of the reconstructed sources, the PEASS Toolkit [44] is used and the results are reported in terms of OPS, IPS, APS.

3. PERCEPTUALLY ENHANCED BLIND SINGLE-CHANNEL MUSIC SOURCE SEPARATION BY NMF

In this chapter, it is focused on blind single channel separation of pitched instruments by defining the problem as a constrained optimization problem where the perceptual quality of the separated sources constitutes the constraint.

In music source separation, NMF is capable of modeling each note played by a given instrument with a single basis vector [10]. On the other hand, since a music signal is composed of several notes, the separation generates multiple bases for each instrument. This requires the clustering of the components into sources and it is a common drawback of the source separation methods where the number of extracted components is greater than the number of sources. FitzGerald *et al.* proposed SNMF [10] which avoids the need for clustering under the assumption that the notes belonging to a single source consist of the translated versions of a single basis vector provided that a log-frequency scale is used to represent the bases. Similarly, the NMF2D algorithm is proposed by Schmidt *et al.* [11] for separating the instruments in polyphonic music by representing each instrument using a single time-frequency profile convolved in both time and frequency in a log-frequency magnitude spectrogram. Thus, NMF2D method performs both the time and the frequency translations encountered, respectively, in NMFD [9] and SNMF [10]. The advantage of convolutive NMF methods is that they perform separation by considering the time-frequency relations of different notes generated by an instrument without requiring a clustering scheme. However, the expense is the high computational complexity. Another deficit of the SNMF [10] and the NMF2D [11] methods is that, there is no true inverse to a log-frequency spectrogram which affect the quality of the separated sources. In [10], an efficient Constant-Q-Transform introduced in [46] is used which allows an improved inverse transform with an increase in redundancy by a factor around four or five.

Another important issue in source separation is the quality of the separated sources. In order to improve the sound quality, it is clear that the human perception has to be included into the separation scheme. Knowing that the sound quality is related to the human perception, incorporating the loudness perception model of the Human Auditory System (HAS) into the NMF-based single channel audio source separation is proposed. This is achieved by formulating a weighted divergence in which the objective function is minimized within perceptual critical bands for each source. The weights are selected based on the PEAQ model defined in ITU-R BS. 1387 [27].

In the literature, a number of work utilizes the framework of psychoacoustics [47] in subspace learning. Among these, Virtanen [13] performs a perceptually weighted NMF algorithm for single channel source separation by assigning a weighting coefficient for each critical band of each frame in order to model the loudness perception of the HAS. In [13], NMF factorization is performed within 24 critical bands defined in [47] and improvement achieved by including the perceptual frequency masking into factorization is reported. Each separated time-domain component is mapped to the time domain and clustered into the sources by a supervised approach which minimizes the residual-to-signal-ratio. Moreover, this method does not consider the harmonicity of the notes. Thus, it is not capable of robustly separating the mixtures of the pitched instruments composed of more than a single note.

In [14], a perceptually motivated FDICA scheme is proposed which filters the frequency components that are perceptually irrelevant by exploiting the masking properties of speech. The clustering problem is solved by utilizing a similarity measure among spectral envelopes of the separated frequency components and the coherency of perceptually masked mixing matrices in several adjacent frequencies. In performance evaluations, the permutation capability is investigated rather than the separation quality. In our previous work [15], the NMF2D is directly performed on an auditory representation of mixed audio sources which decreases the size of the data thus the computational complexity significantly while improving the separation quality. Furthermore, a K-means based clustering is performed on the separated bases that complicates the source separation. In [48], the NMF is used for audio content representation and its reconstruction capability is evaluated based on the perceptual

quality measure referred as Noise-to-Mask-Ratio (NMR) [27]. A following work by Nikunen & Virtanen [49], minimizes the perceptual distortion of NMF based audio representation using the NMR as objective criteria [48]. Both [48] and [49] perform perceptually weighted NMF for signal reconstruction not for source separation.

The introduced single channel music source separation method in this thesis integrates the human perception into the source separation and solves the clustering problem using two blind approaches [28]. A weighted optimization function based on the KL and the IS divergences is formulated that adopts the PEAQ auditory model defined in ITU-R BS.1387 [27] into the separation scheme through a weighting score matrix. The ITU-R BS.1387 is considered as the most effective objective measurement scheme of perceptual audio quality. It introduces 109 critical bands which enables to analyze the signals in a higher frequency resolution and also formulates time spreading which has not been considered before [27]. To our knowledge, no source separation method exists integrating the standardized PEAQ model [27] into the source separation scheme before [15, 21]. Moreover, our preliminary algorithms [15, 21] are the first works that integrate human perception into shift-invariant model NMF2D. These constitute the main differences between the proposed algorithms and the weighted NMF method presented in [13]. In the proposed PW-NMF2D approach [21, 28] as well as our previous work [15], each source is represented using a time-frequency signature convolved in time and frequency to capture the different notes of each instrument. Inspiring from [12], the proposed second approach, inserts the weighting scheme in a similar two step algorithm and it is named as PW-CNMF [28]. In PW-CNMF, a high rank perceptually weighted NMF factorization is performed in order to obtain a single basis for each note. These bases are then clustered into sources using NMF2D in an additional step. Thus, the components of the notes belonging to the same source are clustered into the same group and then resynhtesized back to the time domain using the encoding matrix obtained from the mixture. Note that, both of the proposed methods perform clustering without requiring any information about the sources. The proposed methods are compared with the SNMF [10], NMF2D [11] and CNMF [12] schemes on a large test set. It is shown that the use of perceptually motivated factorization increases the perceptual quality of the separated sources. Moreover, it is concluded

that using temporal masking along with the frequency masking in the calculation of the weighting matrix enables us efficiently modeling temporal changes of the sources.

The rest of this chapter is organized as follows. The introduced perceptually enhanced models are described in Section 3.1. The experimental results obtained by the proposed methods are reported in Section 3.2. This chapter is concluded with some discussions in Section 3.3.

3.1 Proposed Perceptually Weighted NMF

In Section 3.1.1, the perceptually weighted cost functions are formulated for the proposed PW-NMF2D and PW-CNMF methods. Section 3.1.2 gives the details of perceptual weighting.

3.1.1 Weighted objective function

In the proposed blind source separation method, the objective function is formulated as a weighted divergence by

$$C_W = \sum_{i=1}^{I} \sum_{b=1}^{N_c} P_{bi} G_{bi}, \tag{3.1}$$

where P_{bi} is the weighting score assigned to the $b-$th critical band of the $i-$th time frame of the observed mixture spectrogram matrix; I and N_c are the total number of the time frames and the critical bands, respectively. The critical bands represent the frequency resolution of the auditory system and they are spaced linearly below 500 Hz and logarithmically above 500 Hz. This scale is named as the Bark scale, a frequency scale on which equal distances correspond to perceptually equal distances, and the critical bands have equal bandwidth in terms of bark [47]. The matrix $\mathbf{G}$ shown in (**3.1**) holds the element-wise divergence. Note that the cost function formulated by (**3.1**) reduces to the objective function of the conventional NMF where $P_{bi} = 1, \forall b, i$.

Weighting the objective function of decomposition in each critical band enables a subspace learning that mimics the human perception. This yields a perceptual distortion measure based on an auditory model that forces the source separation to be optimized in terms of perceptual quality. This is achieved by increasing/decreasing the cost term corresponding to the critical bands adaptively. Specifically in (**3.1**), the

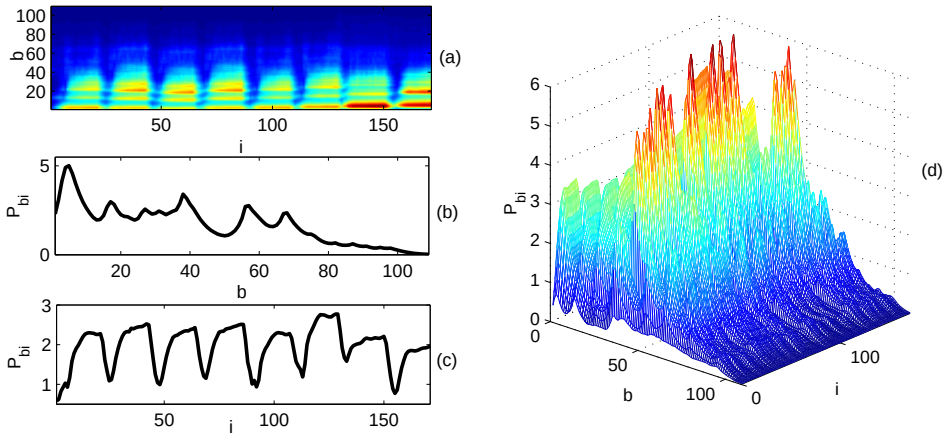


Figure 3.1: Weighting scores P_{bi} in (a) each critical band of each audio frame, (b) each critical band of $i = 117$th frame, (c) $b = 31$st critical band of all time frames, (d) in each critical band of each audio frame as a mesh plot.

weighting score matrix controls the additive penalty term derived by the perceptually masked loudness levels. The perceptual masking is applied based on the perceptual model of the PEAQ defined in ITU-R BS.1387 [27] for extracting the weighting score matrix. The aim is to assign a "perceptual significance " weight P_{bi} for each critical band b in each audio frame i. In Figure 3.1, the weighting scores calculated (a) in each critical band of each frame of the input signal is illustrated with the P_{bi}, $b = 1 \cdots N_c$, $i = 1 \cdots I$; (b) for a single frame $i = 117$ of the signal with P_{bi}, $b = 1 \cdots N_c$; (c) for a specific critical band $b = 31$ over all frames with P_{bi}, $i = 1 \cdots I$. The total number of critical bands is $N_c = 109$ as it is defined in [27]. It is observed that, the weighting scores are adaptively adjusted through the critical bands. Moreover, it can be observed that the distribution is capable of tracking the temporal changes of the spectrogram. The overall weighting matrix is displayed as a mesh plot in Figure 3.1 (d).

Note that the factorization is performed based on the cost function defined in bark scale as in (**3.1**). Perceptually weighting the cost function of NMF is first proposed by Virtanen [13]. Unlike [13], the proposed perceptual model is integrated into the shift-invariant models NMF2D [11] and CNMF [12], and the weighting coefficients are calculated based on the PEAQ model defined in [27]. Thus, the clustering problem encountered in separation is eliminated using these two different state-of-the-art methods. The proposed perceptually weighted NMF2D and CNMF methods are named as PW-NMF2D and PW-CNMF, respectively.

3.1.1.1 Source separation by perceptually weighted NMF2D

The NMF2D model assumes that the frequency signature of a specific note played by an instrument has a fixed temporal pattern (echo); and different notes played by the same instrument has same time-log-frequency signature but varying in fundamental frequency (shift) where the frequency is defined in log-scale [11]. It is also shown that individual notes can be effectively tracked in the log-frequency domain with a higher resolution [10, 11]. Therefore, the bases of the NMF2D are initialized in log-scale using $M = 421$ frequency bands. To formulate the relation between the bark scale and the higher resolution log-frequency scale, $\mathbf{G}$ is mapped by $\mathbb{D} = \mathbb{L}\mathbf{G}$ into the log-frequency domain using a transform matrix $\mathbb{L} \in \mathbb{R}^{M \times B}$, where each column of $\mathbb{L}$ contains the power response of a bark band for all log-frequency indices. The matrix $\mathbb{D}$ holds the element-wise components of the optimization function. The use of both the KL and the IS divergences are investigated for optimizing the factorization. The element-wise KL divergence can be defined as

$$\mathbb{D}_{mi}^{KL} = \mathbb{X}_{mi} \log \frac{\mathbb{X}_{mi}}{\hat{\mathbb{X}}_{mi}} - \mathbb{X}_{mi} + \hat{\mathbb{X}}_{mi}, \tag{3.2}$$

where $m = 1 \ldots M$ is the log-frequency index, $i = 1 \ldots I$ is the frame index. Similarly, the element-wise IS divergence can be defined as

$$\mathbb{D}_{mi}^{IS} = \frac{\mathbb{X}_{mi}}{\hat{\mathbb{X}}_{mi}} - \log \frac{\mathbb{X}_{mi}}{\hat{\mathbb{X}}_{mi}} - 1. \tag{3.3}$$

Consequently, the objective function in (**3.1**) can be rewritten in the higher resolution log-frequency scale as in (**3.4**)

$$C_W = \sum_i \sum_b \sum_m P_{bi} \mathbb{L}_{bm}^{\top} \mathbb{D}_{mi} = \sum_i \sum_m W_{mi} \mathbb{D}_{mi}, \tag{3.4}$$

where $\mathbf{W} = \mathbb{L}\mathbf{P}$.

Essentially, the weighting scores are calculated in the critical bands [27] and in order to be consistent with the signal representation; the weighting score matrix is mapped into the logarithmically spaced frequency bins. Thus, the weighting is effectively performed in perceptual bands of the HAS. Hence the weighted KL divergence update

rules for $\mathbf{A}^{\phi}$ and $\mathbf{S}^{\tau}$ are derived as:

$$A^{\phi}_{ri} \leftarrow A^{\phi}_{ri} \frac{\sum_m \sum_\tau W_{m,i+\tau} \frac{\mathbb{X}_{m,i+\tau}}{\hat{\mathbb{X}}_{m,i+\tau}} S^{\tau}_{m-\phi,r}}{\sum_m \sum_\tau S^{\tau}_{m-\phi,r} W_{m,i+\tau}} \tag{3.5}$$

$$S^{\tau}_{mr} \leftarrow S^{\tau}_{mr} \frac{\sum_\phi \sum_i W_{m+\phi,i} \frac{\mathbb{X}_{m+\phi,i}}{\hat{\mathbb{X}}_{m+\phi,i}} A^{\phi}_{r,i-\tau}}{\sum_\phi \sum_i A^{\phi}_{r,i-\tau} W_{m+\phi,i}}. \tag{3.6}$$

Similarly, the weighted IS divergence update rules are obtained as:

$$A^{\phi}_{ri} \leftarrow A^{\phi}_{ri} \frac{\sum_m \sum_\tau W_{m,i+\tau} \frac{\mathbb{X}_{m,i+\tau}}{\hat{\mathbb{X}}^2_{m,i+\tau}} S^{\tau}_{m-\phi,r}}{\sum_m \sum_\tau S^{\tau}_{m-\phi,r} \frac{W_{m,i+\tau}}{\hat{\mathbb{X}}_{m,i+\tau}}} \tag{3.7}$$

$$S^{\tau}_{mr} \leftarrow S^{\tau}_{mr} \frac{\sum_\phi \sum_i W_{m+\phi,i} \frac{\mathbb{X}_{m+\phi,i}}{\hat{\mathbb{X}}^2_{m+\phi,i}} A^{\phi}_{r,i-\tau}}{\sum_\phi \sum_i A^{\phi}_{r,i-\tau} \frac{W_{m+\phi,i}}{\hat{\mathbb{X}}_{m+\phi,i}}}. \tag{3.8}$$

In order to obtain the separated sources, both of these updates are applied iteratively until the two factors converge. In Appendix A.1 and Appendix A.2, derivation of the multiplicative update equations are presented for the basis and the gain matrices based on the KL and the IS divergences, respectively.

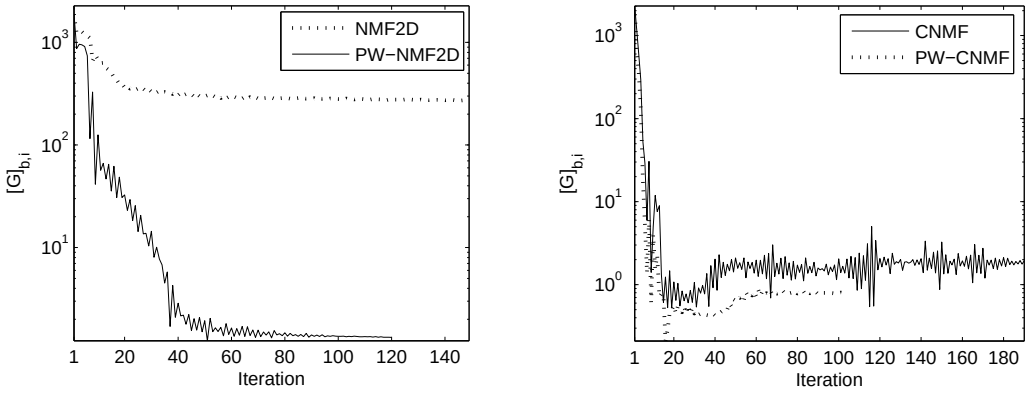
After convergence, the PW-NMF2D algorithm yields the log-frequency magnitude spectrogram for each component r,

$$\hat{\mathbb{X}}_{rmi} \quad = \quad \sum_\tau \sum_\phi S^{\tau}_{m-\phi,r} A^{\phi}_{r,i-\tau}, \quad r = 1 \cdots R, \tag{3.9}$$

where m and i are the indexes for the log-frequency bin and the time frame, respectively. Note that each $r-$th component represents a single source. Thus, the component index r can be replaced with the source index j. The estimated log-frequency magnitude spectrograms $\hat{\mathbb{X}}_j$ are mapped to the linear-frequency domain by $\hat{\mathbf{X}}_j = \mathbf{L}^{F\top} \hat{\mathbb{X}}_j$.

The proposed PW-NMF2D algorithms based on the KL and the IS divergences are summarized in Algorithm 1 and Algorithm 2, respectively.

The change of the element-wise KL divergence for a single time-frequency component is plotted in Figure 3.2(a) for NMF2D and PW-NMF2D methods. For this particular time-frequency component, the value of the weighting score is high, hence the contribution of the element-wise KL divergence in this component is increased in the



(a) (b)

Figure 3.2: The change of the element-wise KL divergence G_{bi} obtained in the $b =$ 31st critical band of $i = 117$th time frame at each iteration by (a) the NMF2D and the proposed PW-NMF2D methods, (b) the CNMF and the proposed PW-CNMF methods.

Algorithm 1 The PW-NMF2D algorithm based on the KL divergence.

1: Calculate perceptual weighting score matrix $\mathbf{P}$

2: Map $\mathbf{P}$ to log-frequency domain $\mathbf{W} = \mathbb{L}\mathbf{P}$

3: Initialize $\mathbf{A}^{\phi}$ and $\mathbf{S}^{\tau}$ randomly

4: for $n = 1 \cdots$ MAXITER do

5:

$$\hat{\mathbb{X}}_{mi} = \sum_{\tau}\sum_{\phi}\sum_{r} S^{\tau}_{m-\phi,r} A^{\phi}_{r,i-\tau}$$

6:

$$A^{\phi}_{ri} \leftarrow A^{\phi}_{ri} \frac{\sum_m \sum_\tau W_{m,i+\tau} \frac{\mathbb{X}_{m,i+\tau}}{\hat{\mathbb{X}}_{m,i+\tau}} S^{\tau}_{m-\phi,r}}{\sum_m \sum_\tau S^{\tau}_{m-\phi,r} W_{m,i+\tau}}$$

7:

$$\hat{\mathbb{X}}_{mi} = \sum_{\tau}\sum_{\phi}\sum_{r} S^{\tau}_{m-\phi,r} A^{\phi}_{r,i-\tau}$$

8:

$$S^{\tau}_{mr} \leftarrow S^{\tau}_{mr} \frac{\sum_\phi \sum_i W_{m+\phi,i} \frac{\mathbb{X}_{m+\phi,i}}{\hat{\mathbb{X}}_{m+\phi,i}} A^{\phi}_{r,i-\tau}}{\sum_\phi \sum_i A^{\phi}_{r,i-\tau} W_{m+\phi,i}}$$

9: end for

10: Calculate magnitude log-frequency spectrogram: $\hat{\mathbb{X}}_{jmi} = \sum_\tau \sum_\phi S^{\tau}_{m-\phi,j} A^{\phi}_{j,i-\tau}$

11: Map to linear-frequency domain: $\hat{\mathbf{X}}_j = {\mathbf{L}^F}^{\top} \hat{\mathbb{X}}_j$

12: Calculate complex frequency spectrogram: $\hat{F}_{jki} = F_{ki} \frac{\hat{X}_{jki}}{\sum_j \hat{X}_{jki}}$

13: Calculate time domain estimate: $\hat{\mathbf{s}}_j(t) \leftarrow \mathrm{ISTFT}(\hat{\mathbf{F}}_j)$

Algorithm 2 The PW-NMF2D algorithm based on the IS divergence.

1: Calculate perceptual weighting score matrix $\mathbf{P}$
2: Map $\mathbf{P}$ to log-frequency domain $\mathbf{W} = \mathbb{L}\mathbf{P}$
3: Initialize $\mathbf{A}^{\phi}$ and $\mathbf{S}^{\tau}$ randomly
4: for $n = 1 \cdots$ MAXITER do
5:

$$\hat{\mathbb{X}}_{mi} = \sum_{\tau}\sum_{\phi}\sum_{r} S^{\tau}_{m-\phi,r} A^{\phi}_{r,i-\tau}, \quad m = 1 \cdots M, i = 1 \cdots I$$

6:

$$A^{\phi}_{ri} \leftarrow A^{\phi}_{ri} \frac{\sum_m \sum_\tau W_{m,i+\tau} \frac{\mathbb{X}_{m,i+\tau}}{\hat{\mathbb{X}}^2_{m,i+\tau}} S^{\tau}_{m-\phi,r}}{\sum_m \sum_\tau S^{\tau}_{m-\phi,r} \frac{W_{m,i+\tau}}{\hat{\mathbb{X}}_{m,i+\tau}}}$$

7:

$$\hat{\mathbb{X}}_{mi} = \sum_{\tau}\sum_{\phi}\sum_{r} S^{\tau}_{m-\phi,r} A^{\phi}_{r,i-\tau}, \quad m = 1 \cdots M, i = 1 \cdots I$$

8:

$$S^{\tau}_{mr} \leftarrow S^{\tau}_{mr} \frac{\sum_\phi \sum_i W_{m+\phi,i} \frac{\mathbb{X}_{m+\phi,i}}{\hat{\mathbb{X}}^2_{m+\phi,i}} A^{\phi}_{r,i-\tau}}{\sum_\phi \sum_i A^{\phi}_{r,i-\tau} \frac{W_{m+\phi,i}}{\hat{\mathbb{X}}_{m+\phi,i}}}$$

9: end for
10: Calculate magnitude log-frequency spectrogram: $\hat{\mathbb{X}}_{jmi} = \sum_\tau \sum_\phi S^{\tau}_{m-\phi,j} A^{\phi}_{j,i-\tau}$
11: Map to linear-frequency domain: $\hat{\mathbf{X}}_j = \mathbf{L}^{F\top} \hat{\mathbb{X}}_j$
12: Calculate complex frequency spectrogram: $\hat{F}_{jki} = F_{ki} \frac{\hat{X}_{jki}}{\sum_j \hat{X}_{jki}}$
13: Calculate time domain estimate: $\hat{\mathbf{s}}_j(t) \leftarrow \mathrm{ISTFT}(\hat{\mathbf{F}}_j)$

total additive error given in (**3.1**). This results in forcing the KL divergence component which is more critical to human hearing to be reduced to a lower value. The weighting scheme also causes the algorithm to converge in less steps.

3.1.1.2 Source separation by perceptually weighted CNMF

As a second method, the perceptual weighting scheme is integrated into the Clustered NMF (CNMF) method proposed in [12]. The CNMF scheme [12] divides the single step algorithm NMF2D into a two-step approach and performs the separation in linear frequency domain resulting in an improved separation quality. In the first step, the bases of each note are extracted using a high rank NMF. In the second step, these bases are clustered into sources using NMF2D. To formulate the relation between the bark scale and the linear-frequency scale, the optimization function in (**3.1**) is derived by replacing $\mathbf{G}$ with $\mathbf{L}^{\top}\mathbf{D}$ in which $\mathbf{L} \in \mathbb{R}^{B \times K}$ is used to map the magnitude spectrogram

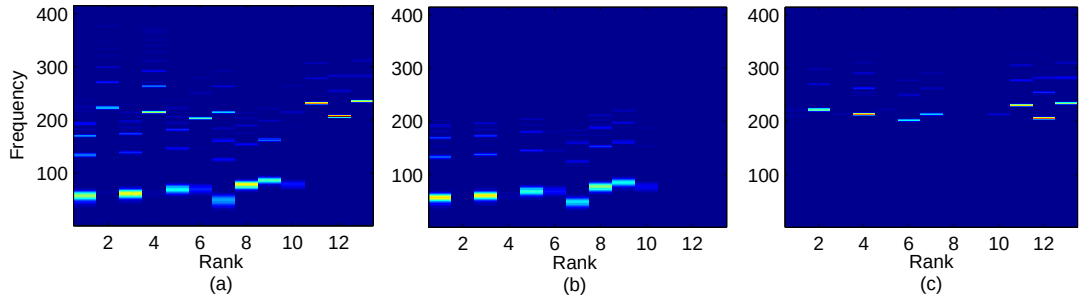


Figure 3.3: Decomposed bases using PW-CNMF: (a) bases in log-frequency domain (b) clustered bases for source 1 (c) clustered bases for source 2.

into bark scale. Furthermore, the difference matrix $\mathbf{D}$ represents the element-wise KL divergence

$$D_{ki}^{KL} = X_{ki} \log \frac{X_{ki}}{\hat{X}_{ki}} - X_{ki} + \hat{X}_{ki}, \tag{3.10}$$

or element-wise IS divergence

$$D_{ki}^{IS} = \frac{X_{ki}}{\hat{X}_{ki}} - \log \frac{X_{ki}}{\hat{X}_{ki}} - 1, \tag{3.11}$$

where $k = 1 \ldots K$ is the linear-frequency index, $i = 1 \ldots I$ is the frame index. $\mathbf{X}$ and $\hat{\mathbf{X}}$ represent the mixture magnitude spectrogram and the NMF model, respectively. Each element in $\hat{\mathbf{X}}$ is defined as:

$$\hat{X}_{ki} = \sum_{r=1}^{R} S_{kr}^{F} A_{ri}, \tag{3.12}$$

where r denotes the component index and $R >= J$. This method calculates a single basis for each note.

In the proposed PW-CNMF method, the objective function defined in (**3.1**) is written in linear-frequency scale as:

$$C_W = \sum_i \sum_b \sum_k P_{bi} L_{bk}^{\top} D_{ki} = \sum_i \sum_k W_{ki}^{F} D_{ki}, \tag{3.13}$$

where $\mathbf{W}^F = \mathbf{LP}$. Essentially, the weighting score matrix calculated in the critical bands [27] is mapped into the linearly spaced frequency bins in order to be consistent with the signal representation.

Hence the weighted KL divergence and IS divergence update rules for $\mathbf{A}$ and $\mathbf{S}^F$ are derived as in (**3.5**-**3.8**) with $\Theta = 1$ and $T = 1$.

The bases $\mathbf{S}^F$ obtained by PW-CNMF method is then mapped to the log-frequency $\mathbf{S} = \mathbf{L}^F \mathbf{S}^F$ in order to use the frequency shifting property of the NMF2D algorithm which

Algorithm 3 The PW-CNMF algorithm based on the KL divergence.

1: Calculate perceptual weighting score matrix $\mathbf{P}$
2: Map $\mathbf{P}$ to linear frequency domain $\mathbf{W}^F = \mathbf{LP}$
3: Initialize $\mathbf{A}$ and $\mathbf{S}^F$ randomly
4: for $n = 1 \cdots$ MAXITER do
5:
$$\hat{X}_{ki} = \sum_r S^F_{kr} A_{ri}$$
6:
$$A_{ri} \leftarrow A_{ri} \frac{\sum_k W^F_{ki} \frac{X_{ki}}{\hat{X}_{ki}} S^F_{kr}}{\sum_k S^F_{kr} W^F_{ki}}$$
7:
$$\hat{X}_{ki} = \sum_r S^F_{kr} A_{ri}$$
8:
$$S^F_{kr} \leftarrow S^F_{kr} \frac{\sum_i W^F_{ki} \frac{X_{ki}}{\hat{X}_{ki}} A_{ri}}{\sum_i A_{ri} W^F_{ki}}$$
9: end for
10: Map $\mathbf{S}^F$ to log-frequency domain: $\mathbf{S} = \mathbf{L}^F \mathbf{S}^F$
11: Cluster $\mathbf{S}$ into $\mathbf{S}_j, j = 1 \cdots J$ via NMF2D
12: Map $\mathbf{S}_j$ to linear frequency domain $\mathbf{S}^F_j = {\mathbf{L}^F}^\top \mathbf{S}_j$
13: Calculate magnitude frequency spectrogram: $\hat{\mathbf{X}}^F_j = \mathbf{S}^F_j \mathbf{A}$
14: Calculate complex frequency spectrogram: $\hat{F}_{jki} = F_{ki} \frac{\hat{X}^F_{jki}}{\sum_r \hat{X}^F_{jki}}$
15: Calculate time domain estimate: $\hat{\mathbf{s}}_j(t) \leftarrow \mathrm{ISTFT}(\hat{\mathbf{F}}_j)$

is applicable only in a log-frequency scale. The bases $\mathbf{S}$ obtained after the mapping is illustrated in Figure 3.3(a). Clustering of the bases into J sources $\mathbf{S}_j, j = 1 \cdots J$ is achieved using NMF2D by only considering the frequency shifts and ignoring the time dependencies. The clustered bases $\mathbf{S}_j$ are illustrated in Figure 3.3 (b) and (c) for each source.

These bases $\mathbf{S}_j$ are mapped back to the linear frequency domain $\mathbf{S}^F_j = {\mathbf{L}^F}^\top \mathbf{S}_j$ before resynthesis. Then, the magnitude spectrogram for each source j is calculated as,

$$\hat{\mathbf{X}}_j \quad = \quad \mathbf{S}^F_j \mathbf{A}, \quad j = 1 \cdots J. \tag{3.14}$$

The proposed PW-CNMF algorithms based on the KL and the IS divergences are summarized in Algorithm 3 and Algorithm 4, respectively.

Algorithm 4 The PW-CNMF algorithm based on the IS divergence.

1: Calculate perceptual weighting score matrix $\mathbf{P}$
2: Map $\mathbf{P}$ to linear frequency domain $\mathbf{W}^F = \mathbf{L}\mathbf{P}$
3: Initialize $\mathbf{A}$ and $\mathbf{S}^F$ randomly
4: for $n = 1 \cdots$ MAXITER do
5:

$$\hat{X}_{ki} = \sum_r S^F_{kr} A_{ri}$$

6:

$$A_{ri} \leftarrow A_{ri} \frac{\sum_k W^F_{ki} \frac{X_{ki}}{\hat{X}^2_{ki}} S^F_{kr}}{\sum_k S^F_{kr} \frac{W^F_{ki}}{\hat{X}_{ki}}}$$

7:

$$\hat{X}_{ki} = \sum_r S^F_{kr} A_{ri}$$

8:

$$S^F_{kr} \leftarrow S^F_{kr} \frac{\sum_i W^F_{ki} \frac{X_{ki}}{\hat{X}^2_{ki}} A_{ri}}{\sum_i A_{ri} \frac{W^F_{ki}}{\hat{X}_{ki}}}$$

9: end for
10: Map $\mathbf{S}^F$ to log-frequency domain: $\mathbf{S} = \mathbf{L}^F \mathbf{S}^F$
11: Cluster $\mathbf{S}$ into $\mathbf{S}_j, j = 1 \cdots J$ via NMF2D
12: Map $\mathbf{S}_j$ to linear frequency domain $\mathbf{S}^F_j = {\mathbf{L}^F}^\top \mathbf{S}_j$
13: Calculate magnitude frequency spectrogram: $\hat{\mathbf{X}}^F_j = \mathbf{S}^F_j \mathbf{A}$
14: Calculate complex frequency spectrogram: $\hat{F}_{jki} = F_{ki} \frac{\hat{X}^F_{jki}}{\sum_r \hat{X}^F_{jki}}$
15: Calculate time domain estimate: $\hat{\mathbf{s}}_j(t) \leftarrow \text{ISTFT}(\hat{\mathbf{F}}_j)$

The element-wise KL divergence is plotted for a single time-frequency component in Figure 3.2(b) for the CNMF [12] and PW-CNMF methods. Similar to Figure 3.2 (a), the PW-CNMF algorithm converges in less steps compared to CNMF and the KL divergence in this component which is more critical to human hearing is reduced by PW-CNMF to a lower value.

The spectrograms for the original source and the mixture are illustrated together with the estimated spectrograms of the sources obtained by the proposed two approaches in Figure 3.4. We observe that, both the PW-NMF2D (c) and the PW-CNMF (d) methods remove the components belonging to the interference source successfully. The advantage of PW-NMF2D over the PW-CNMF method is that it considers the time shifts thus captures the temporal structure. However, PW-CNMF method suppresses the components belonging to the interference better. These figures are illustrated

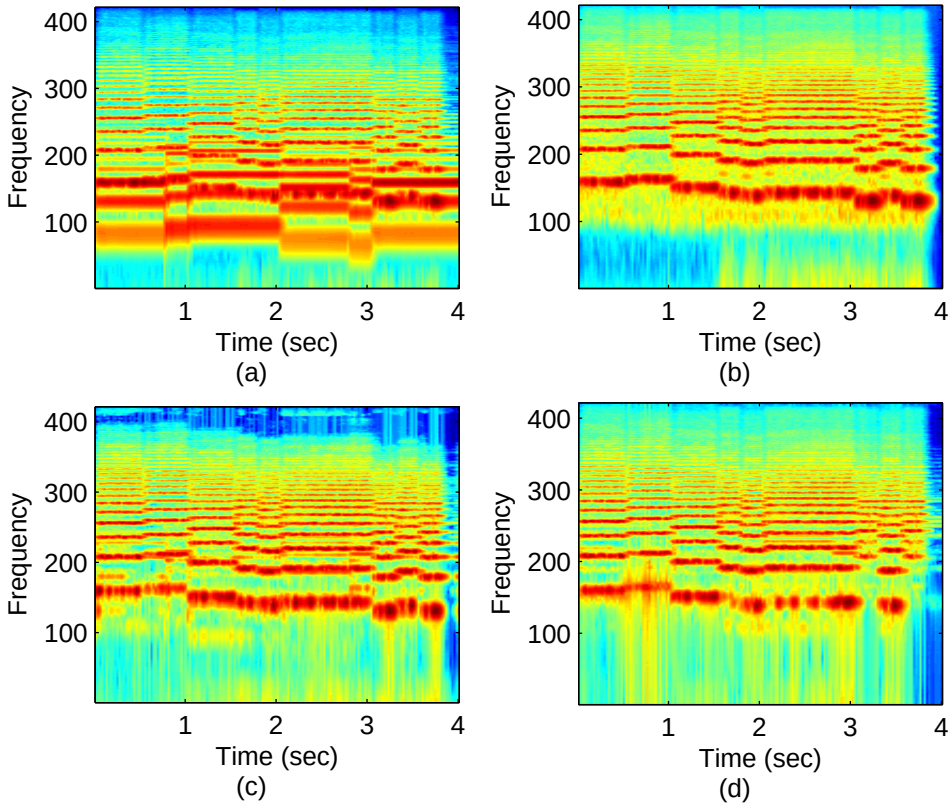


Figure 3.4: Log-frequency spectrograms of (a) mixture signal, (b) the original source, (c) extracted source using PW-NMF2D, (d) extracted source using PW-CNMF.

in log-frequency domain in order to have a more detailed illustration of the lower frequency components. In linear frequency scale, all the lower frequency components look scrunched together and the relationship between different pitches cannot be observed clearly.

3.1.2 Extracting the weighting score matrix

The objective of proposing to apply a psychoacoustic preprocessing on the observed audio mixture before source separation is to enhance the information which is critical to human hearing while suppressing the non-critical components. It is known that the quantitative significance of an auditory object within a mixture can be measured by its perceived loudness through the HAS [27] which is the sensory system for the sense of hearing. Knowing this, the loudness of one frame is modeled by calculating the excitation using perceptually motivated frequency scale (Bark scale) and critical bandwidth [27], compressing the excitation, and integrating over frequency [13, 27].

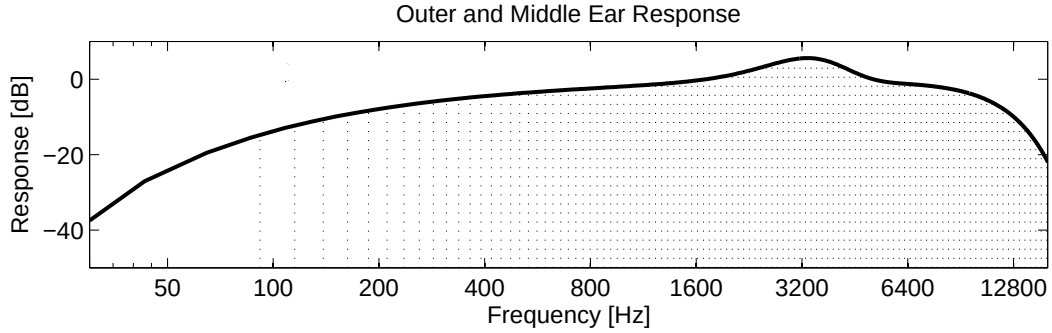
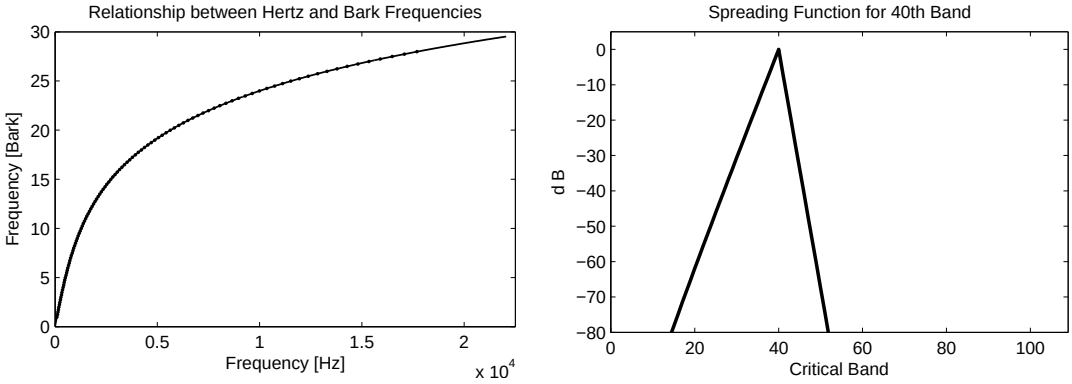


Figure 3.5: The response of outer and middle ear model based on ITU-R BS. 1387 (solid). The dotted lines mark the center frequencies of the critical-bands.



(a) (b)

Figure 3.6: (a) The relationship between the Hertz and the Bark frequencies. (b)The spreading function for $b = 40$th critical band given in ITU-R BS 1387.

Thus, the loudness can be estimated individually for each critical band. In our system, $N_c = 109$ separate bands are spaced uniformly on the Bark scale. Note that the loudness is a subjective measure describing the strength of the ear's perception of a sound and true loudness is affected by parameters other than sound pressure, including frequency and duration.

In the following, the main steps used for deriving the weighting matrix are briefly presented.

(1) Let $\mathbf{X}$ denote the magnitude spectrogram of the mixed audio signal $\mathbf{x}$ calculated using the STFT. First, a weighting is applied on the observed mixture spectrogram reflecting the filtering of the outer and middle ear

$$X_{ki}^{w2} = V_k^2 {X_{ki}}^2, \qquad 0 \leq k \leq \frac{N_F}{2}, \tag{3.15}$$

where $\mathbf{X}^w$ represents the weighted spectrogram; N_F is the length of one time frame; k and i are the frequency and time indexes, respectively. Each frequency is weighted

according to the human ear filtering [27] using:

$$V_k = 10^{A_{dB}(\frac{kF_S}{N_F})/20}, \tag{3.16}$$

where F_s is the sampling frequency of the signal. The response of the outer and the middle ear for the frequency f (in Hz) is modeled as:

$$A_{dB}(f_{Hz}) = -2.184(\frac{f}{1000})^{-0.8} + 6.5e^{-0.6(\frac{f}{1000}-3.3)^2} - 0.001(\frac{f}{1000})^{3.6}. \tag{3.17}$$

The frequency response of the outer and the middle ear is plotted in Figure 3.5. The main characteristic of this function is to emphasize the influence of the frequencies around 3–4 kHz while reducing the influence of very high and low frequencies. The peak value of 5.6 dB occurs near 3.3 kHz.

(2) The frequency bins of $\mathbf{X}^w$ are grouped into $N_c = 109$ critical bands in bark scale as it is defined in PEAQ auditory model [27]. The conversion between the bark and the frequency scale can be computed with

$$Z_{\text{bark}}(f_{Hz}) = 7\text{a}\sinh(f/650). \tag{3.18}$$

The frequency bands start at $f_L = 80$ Hz and stop at $f_U = 18$kHz. All bands but the last have the same width in Barks. The relationship between the Hertz and the Bark frequencies are plotted in Figure 3.6 (a).

The energy $\mathbf{E}^a$ [27] is calculated in each critical band b of each frame i as:

$$E^a_{bi} = \sum_{k=k_l(b)}^{k_u(b)} U_{bk} X^{w2}_{ki}, \tag{3.19}$$

where U_{bk} is the contribution from the energy in DFT bin k for the $b-$ th frequency band. U_{bk} is non-zero over the interval $k_l(b) \leq k \leq k_u(b)$ and defined as

$$U_{bk} = \frac{\max\left[0, \{\min(f_u(b), \frac{2k+1}{2}\frac{F_s}{N_F}) - \max(f_l(b), \frac{2k-1}{2}\frac{F_s}{N_F})\}\right]}{\left(\frac{F_s}{N_F}\right)}, \tag{3.20}$$

where $f_l(b)$ and $f_u(b)$ are the lower and the upper frequency edges for band b, respectively [27]. The energy is set to $E_{min} = 1 \times 10^{-12}$ if it is less than E^a_{bi} [27],

$$E^b_{bi} = \max(E^a_{bi}, E_{min}). \tag{3.21}$$

An offset E_b^{IN} is added to the band energies to compensate for the internal noise generated in the ear [27],

$$E_{bi} = E_{bi}^{b} + E_b^{IN}. \tag{3.22}$$

The energies E_{bi} are referred to the "pitch patterns" [27]. The internal noise E_b^{IN} in (**3.22**) is modeled at the central frequency $f_c(b)$ of each band b as [27]

$$E_b^{IN} = 10^{1.456(f_c(b)/1000)^{-0.8}/10}. \tag{3.23}$$

(3) The spread Bark-domain energy response is calculated as:

$$E_{bi}^{s} = \frac{1}{B_{bi}^{s}} \left(\sum_{p=0}^{N_c-1} \left(E_{pi} S(b,p,E_{pi}) \right)^{0.4} \right)^{\frac{1}{0.4}}, \tag{3.24}$$

where E_{pi} is the energy component at band p and frame i; $B_{bi}^{s} = (\sum_{p=0}^{N_c-1} S(i,p,1)^{0.4})^{\frac{1}{0.4}}$ is the normalization factor. Note that the normalization factor does not depend on data [27]. $S(b,p,E_{pi})$ is the spreading function of band b for an energy component E_{pi} at band p and defined as

$$S(b,p,E_{pi}) = \begin{cases} \frac{1}{A(p,E_{pi})}(10^{2.7\Delta z})^{b-l}, & b \le p, \\ \frac{1}{A(p,E_{pi})}\left[(10^{-2.4\Delta z})(10^{-23\Delta z/f_c(p)})(E_{pi}^{0.2\Delta z})\right]^{b-l}, & b \ge p, \end{cases} \tag{3.25}$$

where $\Delta z = 1/4$ and the normalizing term $A(p,E_{pi})$ is chosen as equal to $\sum_b S(b,p,E_{pi})$ to give a unit area for each center frequency p. The distribution of the spreading function for the critical band $i = 40$ is illustrated in Figure 3.6 (b) and it has a similar characteristic for all bands.

(4) In addition to the frequency masking, the HAS applies the pre- and post- masking effects in time domain, while such temporal masking effects are involved neither in [13, 50] nor in [15]. Time-domain spreading concerns temporal masking effects that enables efficiently modeling the temporal changes of the sources. In this model, temporal masking effect is considered by means of a first-order smooth filtering. Time domain spreading E_{bi}^{f} is calculated for each band b of each frame i as [27]:

$$E_{bi}^{f} = \alpha_b E_{b,i-1}^{f} + (1-\alpha_b) E_{bi}^{s}, \tag{3.26}$$

where α_b controls the time constant of the averaging for decaying energies [27]. $\tilde{E}_{bi}^{s}$'s are called *excitation patterns* and defined as:

$$\tilde{E}_{bi}^{s} = \max(E_{bi}^{f}, E_{bi}^{s}). \tag{3.27}$$

Note that if $\alpha_b = 0$, then $E^f_{bi} = E^s_{bi}$ and $\tilde{E}^s_{bi} = E^s_{bi}$ that corresponds a memoryless model.

(5) Finally the weighting score P_{bi}, corresponding to the loudness index of a frame i within a band b is calculated as [47]

$$P_{bi} = c\Big(\frac{E^t_b}{s_b E_0}\Big)^{0.23}\left[\Big(1 - s_b + \frac{s_b \tilde{E}^s_{bi}}{E^t_b}\Big)^{0.23} - 1\right], \tag{3.28}$$

where s_b is the threshold index, E^t_b is the excitation threshold and $\tilde{E}^s_{bi}$'s are the *excitation patterns*. [27]. The loudness P_{bi} is calculated in units of sone. Thus, if the measured loudness value is doubled in terms of sone, the sound is perceived by the human ear twice as loud. In order to get this relationship, E_0 is set to 10^4 (40 dB relative to 0 dB SPL) and c is set to 1.07664 in [27]. As a result, a sonogram representation reflects the specific loudness sensation of the HAS [15, 51]. The threshold index s_b is defined as

$$s_b = 10^{[-2-2.05\arctan(\frac{f_c(b)}{4000})-0.75\arctan\left((\frac{f_c(b)}{1600})^2\right)]/10}, \tag{3.29}$$

and the excitation threshold E^t_b is defined as

$$E^t_b = 10^{3.64(\frac{f_c(b)}{1000})^{-0.8}/10}. \tag{3.30}$$

The threshold index defined in (**3.29**) is the ratio of the intensity of a just-audible test tone and the intensity of the internal noise within a critical band [52]. The excitation threshold in quiet is the low frequency part of the outer and the middle ear filtering and the internal noise terms.

Figure 3.7 illustrates the spectrums derived through the processing steps. The time domain of the signal is displayed in (a). The spectrogram and the spectrogram after ear filtering are illustrated in (b) and (c), respectively. In Bark-scale (starting from (d)), the information that is least critical to our hearing sensation is removed while the most critical parts are retained without perceptual loss. Figure 3.7 (f) represents the weighting score matrix derived from the input signal.

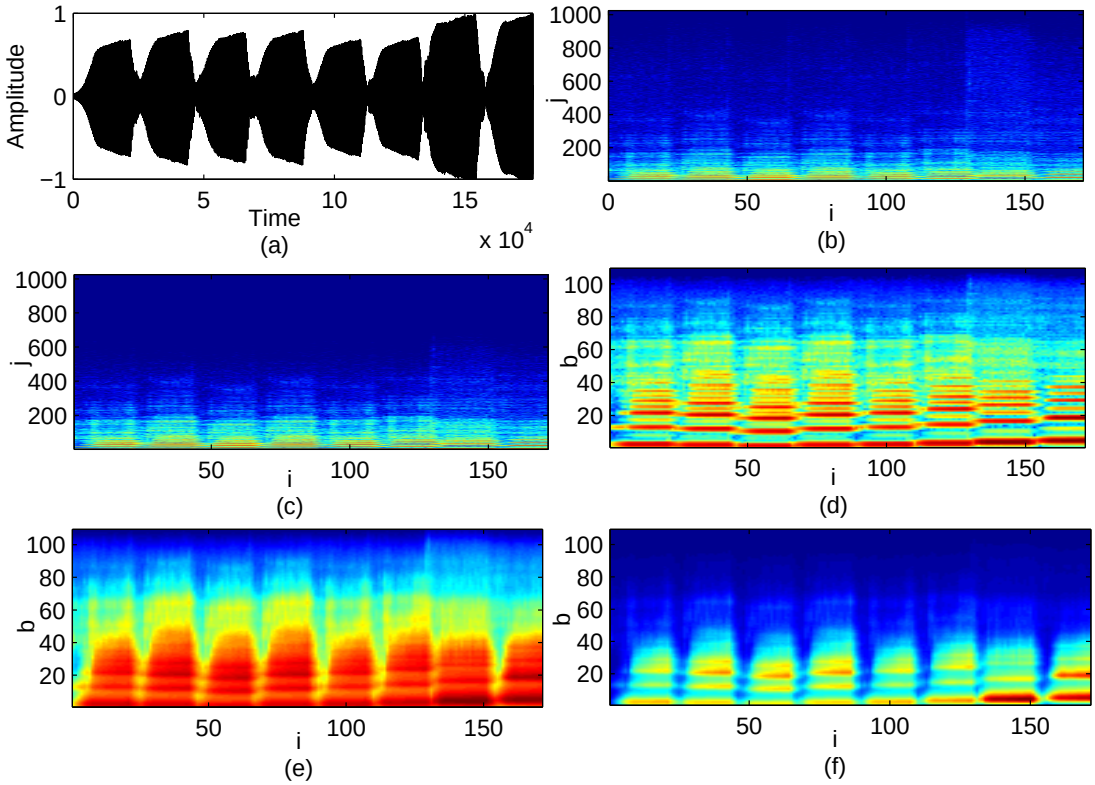


Figure 3.7: The characteristics of the weighting score matrix can be displayed for a particular audio signal: (a) time-domain, (b) spectrogram, (c) spectrogram after ear filtering, (d) spectrogram in bark scale, (e) bark spectrogram after time and frequency masking, (f) proposed weighting score matrix based on ITU-R BS. 1387.

3.2 Test Results

In this section, the test data is described and then the evaluations of the experiments are reported. The PW-NMF2D algorithm is implemented in Matlab by incorporating the proposed weighting scheme into the NMF2D code provided by the author of [11]. The PW-CNMF algorithm is implemented in Matlab by incorporating the weighting scheme into the high-rank NMF factorization. The clustering of PW-CNMF bases are performed using NMF2D by considering the convolution in frequency.

3.2.1 Datasets

In order to evaluate the separation performance of the proposed method with respect to the existing methods [10–12], two different datasets of pitched instruments are used. Dataset 1 includes 15 different orchestral instruments which are downloaded

from [53]. 25 monophonic mixtures of two instruments and 25 monophonic mixtures of three instruments are generated by mixing with unity gain. The input mixtures are of 4 to 8 seconds in length and sampled at a rate of $F_S = 44.1$ kHz. In mixture signals, a wide range of pitches are covered from 87.31 Hz to 1567.98 Hz and the melodies played by the individual instruments in each signal are in harmony. Dataset 2 includes six mixtures of two instruments from a total of six different instruments labeled "nodrums" and downloaded from the SISEC database [54]. The source files are real recordings of 11 seconds in length and sampled at $F_S = 16$ kHz. For both of the datasets, the generated mixtures highly overlap in frequency which enables us to investigate the capability of the algorithms for discriminating notes of the same pitch played by different instruments.

3.2.2 Objective tests

In this section, the performance improvement achieved by incorporating the perceptual weighting scheme into the source separation using both the KL and the IS divergences is investigated. The performance evaluation is performed under two main test cases. First, the effect of the convolutive NMF structures in audio source separation is evaluated. Therefore, a set of comparative tests are made to compare the NMF2D algorithm with the SNMF [10] and the CNMF [12] algorithms. The comparison with SNMF [10] is performed because SNMF is also a convolutive NMF approach that does not require an additional step for clustering the bases into the sources. Unlike NMF2D, SNMF performs only the frequency shifts to track the shift-invariant property of the basis vectors. CNMF [12] is a two-step approach that clusters the bases factorized through NMF into sources using SNMF. After specifying the appropriate parameter set at the first case, in the second test case, the improvement achieved by the proposed perceptual weighting is examined for a number of audio mixtures.

3.2.2.1 Performance based on factorization parameters

First, the effect of the convolution parameters T (in time) and Θ (in frequency) is evaluated on the separation quality. A number of T = [1 2 6 10] and Θ= [1 6 13 19 26 32 39] values are used to track the time and the frequency components, respectively. Note that, when the convolution parameters T and Θ are adjusted accordingly, NMF

($T = 1, \Theta = 1$), NMFD [9] ($\Theta = 1$) and SNMF [10] ($T = 1$) can be obtained as a special case of NMF2D. The rank is fixed to the number of sources for convolutive NMF approaches, SNMF [10] and NMF2D [11]. The separation of two sources from a single mixture is performed using all combinations of these parameters on our dataset using two different divergences which amounted to 56 experiments for 31 mixtures (25 mixtures from Dataset 1 and six mixtures from Dataset 2), repeated five times for a total of 8680 experiments. In order to evaluate the separation performance of the proposed methods for more than two sources, separation of three sources from a single mixture is performed using all combinations of these parameters on Dataset 1 using the KL divergence, which amounts to 28 experiments for 25 mixtures, repeated five times for a total of 3500 experiments. The means of the performance measures over all the experiments are reported to examine the effect of various parameters. Since the characteristics of two datasets are different, the results are reported for each dataset separately.

In order to incorporate the PEAQ model into the NMF2D method, the music signals are analyzed by a $N_F = 2048$ point Hanning windowed STFT with 50% overlap. $N_F/2 + 1 = 1025$ STFT slices are obtained. The linearly spaced frequency bins are grouped into $M = \lfloor \log \frac{F_S/2}{50} / \log 2^{(1/48)} \rfloor$ ($M = 421$ bins for Dataset 1, $M = 351$ bins for Dataset 2) logarithmically spaced frequency bins in the range of 50 Hz to $F_S/2$ with 48 bins per octave, which corresponds to four times the resolution of the equal tempered musical scale. The perceptual weighting matrix is calculated within $N_c = 109$ critical bands defined in the PEAQ model of ITU-R BS. 1387 [27].

The results obtained for different values of Θ by fixing the parameter T are reported to evaluate the effect of the frequency shifting on the separation performance. Similarly, the role of the time shifting is examined on the separation quality for different values of T at a fixed Θ. According to the results of extensive tests on Dataset 1, it is concluded that the highest separation quality is achieved for $T = 2$ and $\Theta = 39$.

Source separation performance obtained by NMF2D and PW-NMF2D based on the KL divergence for separating two sources from a single mixture from Dataset 1 are illustrated in Figure 3.8- 3.11. Figure 3.8 displays the separation performance in terms of SDR, SIR and SAR for different values of Θ, when $T = 2$. It can be seen that

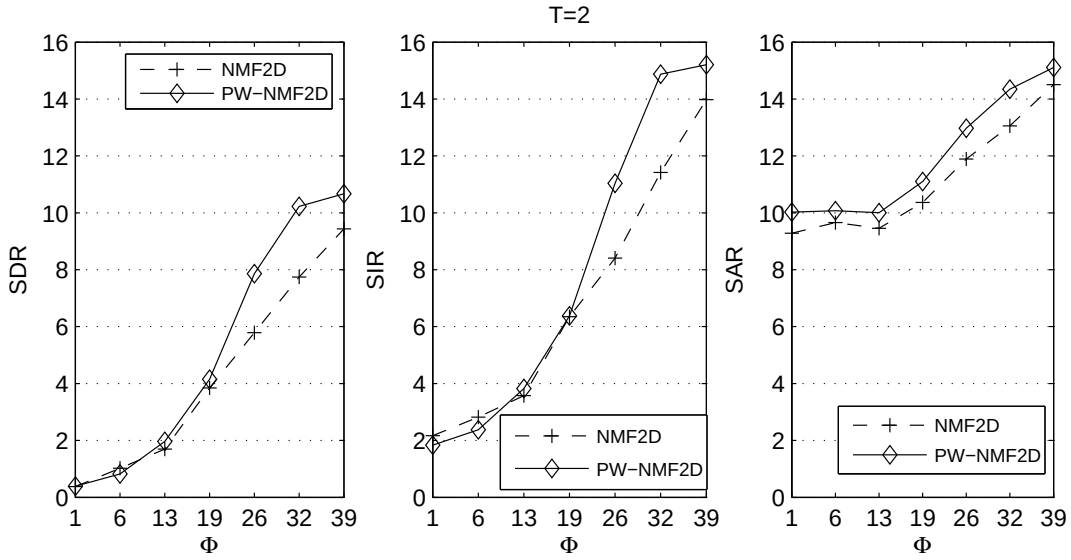


Figure 3.8: The mean SDR, SIR and SAR values obtained by NMF2D and PW-NMF2D for separating two sources from a single mixture on Dataset 1 for various number of frequency shifts $\Theta = \{1, 6, 13, 18, 26, 32, 39\}$ where time shift is fixed as $T = 2$.

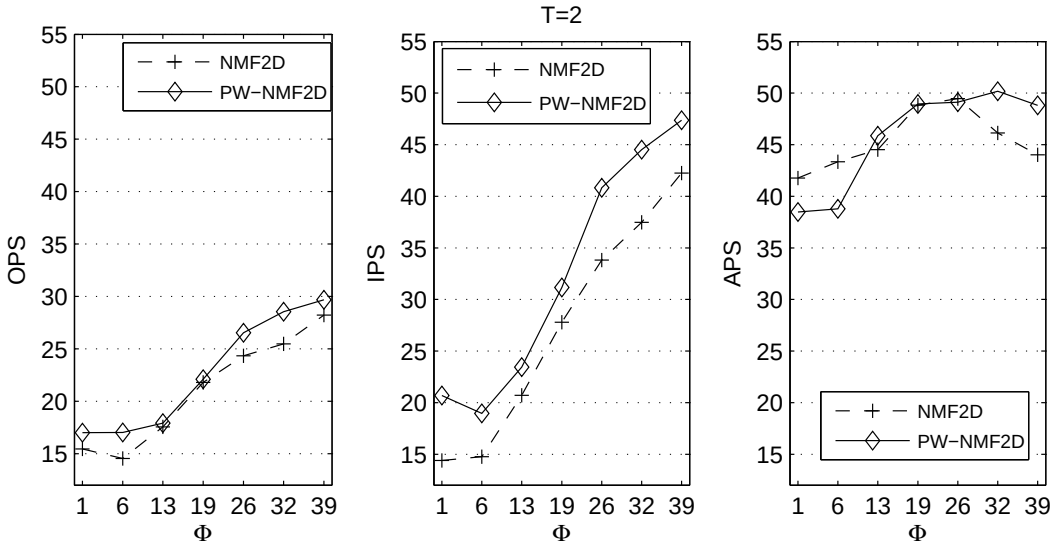


Figure 3.9: The mean OPS, IPS and APS values obtained by NMF2D and PW-NMF2D for separating two sources from a single mixture on Dataset 1 for various number of frequency shifts $\Theta = \{1, 6, 13, 18, 26, 32, 39\}$ where time shift is fixed as $T = 2$.

performance of the NMF2D increases with Θ, although the gain is not significant after $\Theta = 39$. The similar results are observed for the perceptual measures in terms of OPS, IPS and APS as it is displayed in Figure 3.9. Thus, $\Theta = 39$ is fixed to avoid extra computational complexity. To see the gain provided by convolution in the direction of time specified by the parameter T, $\Theta = 39$ is fixed and separation is evaluated for

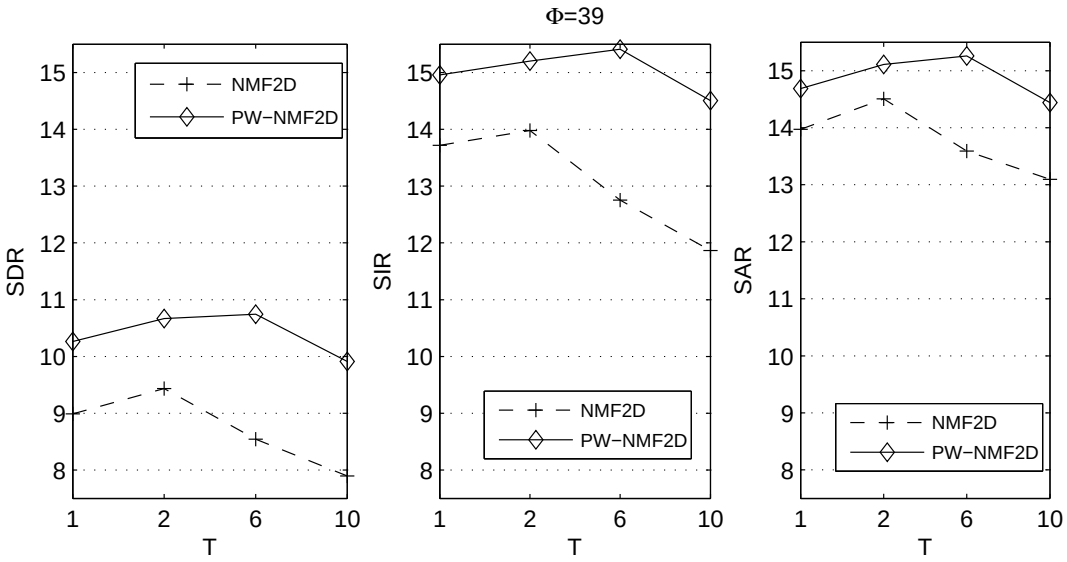


Figure 3.10: The mean SDR, SIR and SAR values obtained by NMF2D and PW-NMF2D for separating two sources from a single mixture on Dataset 1 for various number of time shifts $T = \{1, 2, 6, 10\}$ where frequency shift is fixed as $\Theta = 39$.

different values of T. The results plotted in Figure 3.10 and 3.11 illustrate that the performance improvement obtained by increasing T is not as significant as in the case of Θ, both in terms of the conventional measures SDR, SIR, SAR and the perceptual measures OPS, IPS, APS, respectively. Therefore $T = 2$ is selected which gives a good compromise between the separation quality and the computational complexity. For Dataset 2, the optimal parameters are determined as $T = 6, \Theta = 39$. The overall results of the objective tests obtained by the perceptually enhanced PW-NMF2D algorithm based on the KL divergence on Dataset 1 are shown in Figure 3.8-3.11. It can be concluded that the same parameter set ($T = 2$, $\Theta = 39$) can be chosen as optimal for the weighted factorization. If the performances of NMF2D and PW-NMF2D are compared for mean values obtained at optimal parameters $T = 2, \Theta = 39$, it is observed that PW-NMF2D outperforms NMF2D by around 2, 5 and 5 in terms of OPS, IPS and APS, respectively. Similarly, the improvement obtained by perceptual weighting is around 1 dB in terms of SDR, SIR and SAR.

Similar tests are performed to investigate the performance of PW-CNMF with respect to CNMF with *rank* = [2 13 26 39] and Θ = [0 13 26 39]. The highest separation quality is obtained for *rank* $= 13, \Theta = 39$ on both Dataset 1 and Dataset 2.

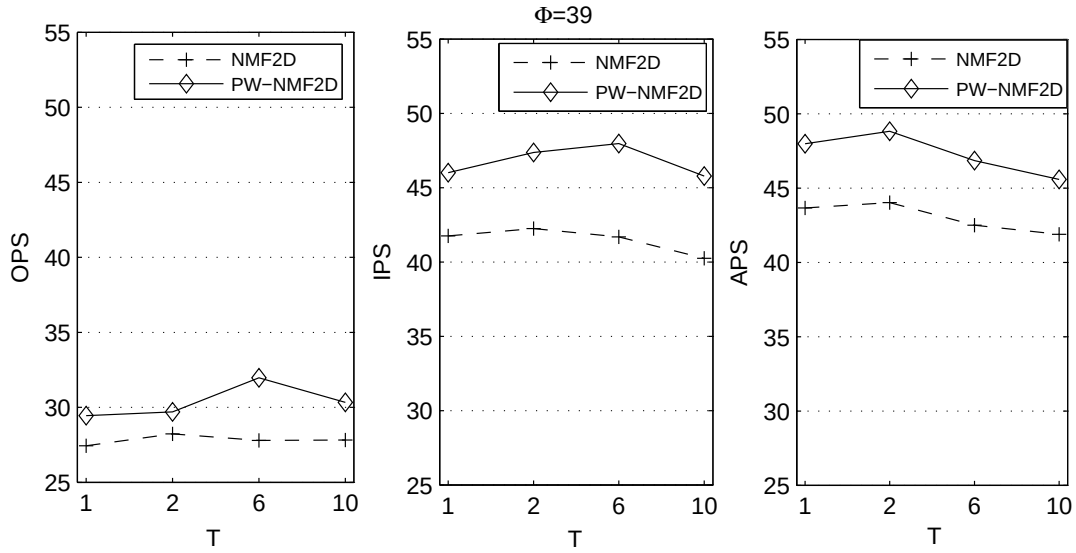


Figure 3.11: The mean OPS, IPS and APS values obtained by NMF2D and PW-NMF2D for separating two sources from a single mixture on Dataset 1 for various number of time shifts $T = \{1, 2, 6, 10\}$ where frequency shift is fixed as $\Theta = 39$.

3.2.2.2 Effect of perceptual weighting

After determining the convolution parameters, the effect of the proposed weighted factorization is investigated by comparing the quality of the separated sources obtained by the proposed PW-NMF2D and PW-CNMF algorithms with the NMF, the SNMF [10], the NMF2D [11] and the CNMF [12] algorithms. Table 3.1 and Table 3.2 report the mean OPS, IPS, APS, SDR, SIR and SAR values obtained by the SNMF, the NMF2D, the CNMF as well as the proposed weighted PW-NMF2D and PW-CNMF methods obtained for separating two sources from a single mixture on Dataset 1 and Dataset 2, respectively. The source separation performance achieved based on the KL and the IS divergences are reported at the upper and the lower parts of the tables, respectively. The SNMF results are obtained by using the code provided by the author of [10] with the same parameter set. The results of the CNMF method proposed in [12] for clustering the NMF bases are reported as CNMF. The PW-CNMF results are obtained using perceptually enhanced NMF algorithm followed by NMF2D clustering.

In Table 3.1, it is observed that NMF is not capable of extracting the sources using a single basis for each source. On the upper part of Table 3.1 (KL divergence), in terms of OPS and IPS, the performance of SNMF [10], CNMF [12] and NMF2D [11]

Table 3.1: Performance in terms of mean OPS, IPS, APS, SDR, SIR and SAR values obtained for separating two sources from Dataset 1 by NMF, SNMF, NMF2D, CNMF, PW-NMF2D, and PW-CNMF.

Methods	OPS	IPS	APS	SDR	SIR	SAR
	KL Divergence					
NMF	15.02	14.02	43.61	0.33	2.10	9.18
SNMF [10]	29.53	47.69	42.20	8.78	12.91	14.95
NMF2D [11]	28.23	42.25	44.02	9.44	13.98	14.69
CNMF [12]	27.74	45.44	38.58	10.33	16.60	14.04
PW-NMF2D	**29.69**	47.37	**48.82**	**10.67**	15.20	**15.11**
PW-CNMF	27.90	**51.75**	43.67	10.61	**17.06**	14.68
	IS Divergence					
NMF	17.53	20.52	38.67	0.52	2.97	9.44
SNMF [10]	29.30	47.46	43.20	6.47	11.52	11.76
NMF2D [11]	30.10	48.65	45.09	6.76	11.84	11.83
CNMF [12]	27.81	**53.09**	40.74	6.98	12.60	12.37
PW-NMF2D	**30.50**	49.09	**45.87**	**8.08**	**13.22**	**12.63**
PW-CNMF	27.93	52.79	40.71	7.19	12.87	12.36

are comparable. On the other hand, the convolutive approaches SNMF and NMF2D outperform CNMF in terms of APS. In terms of BSSEval measures, it can be seen that SNMF and NMF2D have comparable performance while CNMF outperforms SNMF and NMF2D in terms of SDR and SIR. If PW-CNMF and PW-NMF2D are compared, it is observed that PW-NMF2D outperforms PW-CNMF in terms of OPS and APS; while PW-CNMF outperforms PW-NMF2D in terms of IPS. Note that, SNMF [10] and CNMF [12] perform the separation of instruments based on the frequency relations of the notes played by the same instrument but they discard the temporal relations. NMF2D [11] is an extended model thus considers both the time and frequency relations and achieves a better separation if the parameters are selected appropriately. If the performances of NMF2D [11] and PW-NMF2D are compared, it is observed that the perceptual weighting improves the performance by 1-2 dB in terms of SDR, SIR and SAR while an improvement of 1-5 units is observed in terms of perceptual measures. The increase obtained by perceptual weighting is less significant if it is compared to the results of CNMF [12] and PW-CNMF. Although the results based on the KL divergence outperforms the results based on the IS divergence, the performance improvement obtained by perceptual weighting is similar in both cases.

Table 3.2: Performance in terms of mean OPS, IPS, APS, SDR, SIR and SAR values obtained for separating two sources from Dataset 2 by NMF, SNMF, NMF2D CNMF, PW-NMF2D, and PW-CNMF.

Methods	OPS	IPS	APS	SDR	SIR	SAR
	KL Divergence					
NMF	16.55	23.07	38.06	0.14	1.49	10.34
SNMF [10]	17.04	22.53	40.96	11.30	14.38	17.76
NMF2D [11]	27.14	44.96	48.52	11.75	14.61	18.05
CNMF [12]	32.50	57.74	39.75	12.72	16.53	20.57
PW-NMF2D	26.75	42.33	**50.56**	12.38	15.88	18.45
PW-CNMF	**32.96**	**58.14**	41.19	**13.27**	**17.25**	**20.73**
	IS Divergence					
NMF	16.19	26.56	38.26	0.66	2.35	10.16
SNMF [10]	15.52	27.72	39.34	8.53	11.04	14.65
NMF2D [11]	19.96	35.25	42.47	8.58	11.59	14.65
CNMF [12]	19.96	40.35	34.14	9.60	13.48	17.25
PW-NMF2D	**23.57**	33.89	**44.21**	10.12	13.51	17.85
PW-CNMF	21.47	**44.68**	36.06	**10.42**	**13.94**	**18.42**

Moreover, the IS divergence based CNMF [12] outperforms the KL-based methods in terms of IPS and the IS-based PW-NMF2D outperforms the KL-based methods in terms of OPS.

Table 3.2 reports the separation performance obtained for separating two sources from a single mixture from Dataset 2. Similarly, it is observed that the results based on the KL divergence outperforms the results based on the IS divergence and the weighting has a similar effect for the IS divergence case. However, the improvement obtained by perceptual weighting on the KL divergence based NMF2D [11] is less significant on Dataset 2 with an amount of 1 dB in terms of SDR, SIR and SAR. The highlighted results show the highest scores in terms of OPS, IPS, APS, SDR, SIR and SAR obtained using the separation schemes. Although the highest IPS values are obtained by CNMF [12] on Dataset 1, it is only slightly higher than the values obtained by PW-CNMF and PW-NMF2D. It can be concluded that, the weighting scheme noticeably improves the separation performance of NMF2D [11] and a slight improvement in the performance of CNMF [12] is observed. It is also observed that the highest quality in terms of the overall distortion (OPS, SDR), and artifacts (APS, SAR) is achieved by PW-NMF2D; while the PW-CNMF performs the highest

Table 3.3: Performance in terms of mean OPS, IPS, APS, SDR, SIR and SAR values obtained for separation of three sources from a single observation by NMF, SNMF, NMF2D, CNMF, PW-NMF2D, and PW-CNMF with KL divergence on Dataset 1.

Methods	OPS	IPS	APS	SDR	SIR	SAR
NMF	16.9	20.8	20.0	-2.1	-0.3	6.5
SNMF [10]	26.5	40.0	31.8	3.8	6.4	10.5
NMF2D [11]	25.6	37.8	31.0	3.6	6.3	10.1
CNMF [12]	25.1	43.1	27.6	**4.7**	**7.5**	**12.0**
PW-NMF2D	**26.1**	37.9	**31.4**	4.3	6.7	11.2
PW-CNMF	25.0	**44.2**	27.2	4.5	7.5	11.9

quality in terms of the interference distortion (IPS, SIR) on Dataset 1. On Dataset 2, PW-CNMF performs the highest quality in terms of OPS, IPS, SDR,SIR and SAR while PW-NMF2D has the highest performance in terms of APS.

Table 3.3 reports the mean PEASS (OPS, IPS and APS) [25] and BSSEval (SDR, SIR and SAR) [24] measures obtained by the SNMF, the NMF2D, the CNMF as well as the proposed weighted PW-NMF2D and PW-CNMF methods with the KL divergence obtained for separating three sources from a single mixture on Dataset 1. It is observed that the results of SNMF and NMF2D are comparable both in terms of the BSSEval measures and the PEASS measures. PW-NMF2D achieves the separation with 0.5-1 dB improvement in terms of SDR, SIR and SAR and a slight improvement in terms of OPS, IPS and APS compared to NMF2D. If the results obtained by CNMF and PW-CNMF are compared, it is observed that the perceptual weighting enhances the separation performance by 1 unit in terms of IPS while the OPS and APS results are comparable to the ones obtained by CNMF. Although the highest SDR, SIR and SAR results are obtained by CNMF, they are only slightly higher than the ones obtained by PW-CNMF. Similar to the two-source case, it is observed that PW-CNMF outperforms PW-NMF2D in terms of IPS while the OPS and APS results obtained by PW-NMF2D are higher than the ones obtained by PW-CNMF.

If the separation quality achieved for three sources reported on Table 3.3 is compared with the performance reported on Table 3.1 for two sources (note that the same sources are used for the mixtures), it can be seen that the OPS measures are comparable for two-source mixtures and three-source mixtures, while IPS and APS

values decrease for three-source mixtures. It is also observed that the SDR and SIR values obtained for the three-source mixtures are around half of the SDR and SIR values obtained for two-source mixtures. The decrease in the SAR and IPS values are less significant compared to the results obtained for two-source mixtures. The difference observed in the BSSEval measures and the PEASS measures obtained for two-source and three-source mixtures show that the PEASS measures coincide more with the evaluation by listening based on the separated examples [55].

If the results on both datasets are summarized, it can be concluded that if the overall quality is more important depending on the application, PW-NMF2D should be preferred. On the other hand, if the interference is required to be minimized, PW-CNMF can be preferred. A deficit of the CNMF method over NMF2D is that, it applies the separation twice: first performs NMF on the mixture spectrogram and then runs NMF2D on the separated bases for clustering. The difficulty encountered in the NMF2D algorithm is that there is not a direct inverse transform from the log-magnitude frequency spectrogram to the time-domain. Thus, the estimated spectrograms are first mapped to the linear frequency domain and then inverted to the time-domain. During this mapping, some information is lost due to the conversions.

3.2.3 Computational complexity

In this subsection, the computational complexity of the proposed PW-NMF2D and PW-CNMF methods is discussed with respect to the conventional NMF2D [11], SNMF [10] and CNMF [12] methods.

The NMF-based source separation methods contain three main computational steps: calculating the mixture spectrogram, applying source separation on the mixture spectrogram and inversion of the estimated source spectrograms to the time-domain signal. As the first and the last step are common to all methods, we exclude these steps in our complexity measurements.

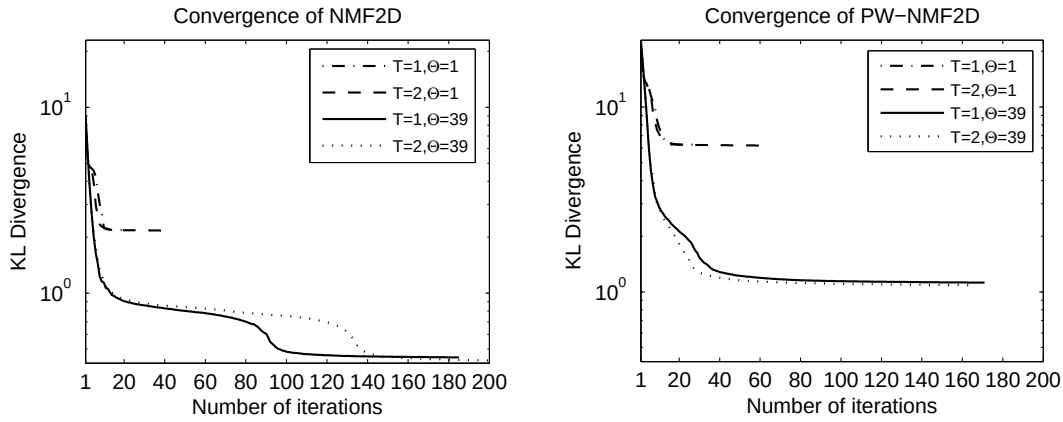
In order to measure the computational complexity of the methods, a 4 sec audio mixture is separated on a PC equipped with a 2.4 GHz Intel® Core™2 Quad processor. The separation time is measured while excluding the file read/write operations and the spectrogram calculation/inversion stages. In Table 3.4, the running

Table 3.4: Run time (in sec.), iteration steps and cost D for a typical run of KL-divergence based SNMF, NMF2D, CNMF, PW-NMF2D and PW-CNMF on a 4 sec mixture.

Methods	Run time (sec).	Iteration steps	C_{KL}
SNMF [10]	63.745	50	0.055
NMF2D [11]	27.631	262	0.429
CNMF [12]	1.156	300	0.471
PW-NMF2D	17.204	165	0.428
PW-CNMF	4.799	90	0.470

time in seconds, the iteration steps and the cost function C_{KL} are listed which are obtained by all considered methods based on the KL divergence and under the same starting conditions. It is observed that the run time of each method varies significantly. This huge difference comes from the difference in the dimensions of the input matrices. The input of the SNMF [10], the CNMF [12], the PW-CNMF and the NMF2D [11] algorithms are Constant-Q-Transform (CQT) spectrogram of length 216×10798, spectrogram of length 2049×169, spectrogram of length 1025×171 and log-frequency spectrogram of length 421×171, respectively. The CQT transform used in [10] is a very redundant transform proposed to give a high quality inverse transform. This redundancy increases the computational complexity of the method significantly. However, compared to the NMF2D method [11], its separation quality does not improve much as it can be seen from Table 3.1 and 3.2. Although the proposed PW-CNMF method runs slower than CNMF [12], it is almost real-time. The increase in the computational complexity of PW-CNMF comes from the difference in the resolution of the CQT transform and the shift parameter used in this work and the work proposed in [12]. When NMF2D [11] is used, the weighting scheme incorporated into the separation algorithm speeds up the convergence time by decreasing the number of iterations required for convergence.

C_{KL} versus iteration is plotted for NMF2D [11] and PW-NMF2D using various convolution parameters in Figure 3.12 (a) and (b), respectively. The results show that when $T = 1, \Theta = 1$ (NMF), the algorithm converges in less steps but to a much higher



(a) (a)

Figure 3.12: (a) Iteration steps vs. objective function on a 4 sec mixture for a typical run of (a) NMF2D, (b) PW-NMF2D using various convolution parameters T and Θ.

KL divergence. When $\Theta = 1, T = 2$, the algorithm converges to a slightly lower value. When we increase Θ, the algorithm converges around 200 steps.

3.3 Conclusion

In this work, the human perception is integrated into single channel blind music source separation by NMF based convex optimization using the KL and the IS divergences. It is shown that the proposed PW-NMF2D and PW-CNMF [28] are capable of successfully separating the sources when the number of sources is known. The best average results are obtained by using the KL divergence as the cost function. Both of the proposed methods extract a single basis for each note and solves the problem of clustering the notes into sources. PW-NMF2D solves the clustering problem under the assumption that all notes for an instrument can be represented by shifting the pitch of the time-frequency signature of a single note. On the other hand, the PW-CNMF method extracts a single basis vector for each note of the mixture using a high rank NMF. Then each note is factorized into sources using NMF2D [11] by capturing the spectral envelope of a note which is assumed to change with pitch. Although PW-CNMF operates in two steps, its complexity is lower. This is because the clustering in the second step is performed on the basis matrix which is much smaller than the spectrogram.

The proposed perceptually enhanced NMF based methods can be used in a wide range of applications including music information retrieval, music transcription, auditory scene analysis where the perceptual quality is important. It can be concluded that PW-NMF2D achieves a higher sound quality in terms of overall distortion and artifacts while PW-CNMF suppresses the interference of the unwanted sources better than PW-NMF2D on Dataset 1, while PW-CNMF outperforms PW-NMF2D in terms of OPS, IPS, SDR, SIR and SAR on Dataset 2. Thus, depending on the application, nature of the data and the requirements, either PW-CNMF or PW-NMF2D can be applied for source separation. In order to extend the model to cover more than two sources, the rank of the factorization and the convolution parameters should be adjusted accordingly. Obviously, increasing the number of sources decreases the separation performance specifically for highly correlated sources. However, including a prior information about the sources should be helpful to improve the separation quality.

4. AN ADAPTIVE TIME-FREQUENCY RESOLUTION FRAMEWORK FOR SINGLE CHANNEL SOURCE SEPARATION BASED ON NTF

The NMF based source separation methods presented in the previous chapters have a shortcoming in that they work at a fixed time-frequency resolution STFT. The STFT of an audio signal can be visualized as a 2D-image known as spectrogram, where the dimensions correspond to time and frequency. The main problem with the fixed time-frequency resolution STFT is the smearing of the signal energy in either direction. Smearing in time causes artifacts such as pre- or post-echoes around the transients which lead to incorrect detection of the temporal changes. On the other hand, smearing of the signal energy in the frequency domain prevents distinguishing the closely spaced harmonics. In particular, shorter windows are chosen around time-localized transients such as plosives in speech and percussive instruments in music, since this produces the most concentrated energy distribution of the STFT coefficients. On the other hand, vowels and voiced consonants in speech, which are oscillatory in nature or harmonics in music, tend to be spread over time but localized in frequency. Thus, the energy concentration is maximized when spectrally concentrated and temporally broad windows are utilized.

In order to overcome the drawbacks arising from working at a single resolution, an adaptive short-time analysis-synthesis scheme is proposed that yields a source separation method capable of supervised adaptive resolution tuning. In the proposed method, the multiresolution time-frequency representation of the observed signal is represented as an “n-way array” or in other terms as a “tensor”, where each layer of the tensor denotes the magnitude spectrogram of the observed mixture obtained at a different time-frequency resolution. The bases **S** are learned for each source from various time-frequency resolution based representations to describe each source in various resolutions. The convergence of the factorization yields the amplitude envelopes **A** and the gain matrix **Q** representing the contribution of each component in each resolution (layer). Since the input is represented as a multidimensional

matrix, NTF which is a natural generalization of NMF in higher dimensional spaces is preferred. The convergence of the NTF algorithm yields the separated sources in each time-frequency resolution. After reconstructing the sources in various resolutions, the adaptation is performed based on a measure of local time-frequency concentration first proposed in [56]. This process yields a sparser representation of the sources by adaptively fusing the information obtained at various resolutions. Moreover, a learning approach proposed in [19, 22] is used in order to extract each source's data outside of the input samples and resolves the clustering problem.

Various works are proposed which deal with the limitations of fixed time-frequency resolution short-time analysis. One group of works propose adaptive STFT schemes [16, 56]. In [56], a simple and efficient technique is designed for adapting the time-frequency and the time-scale representations over time for analysis purposes. A concentration based adaptive procedure is performed which maximizes the kurtosis of short-time time-frequency concentration [56]. However, this work only deals with the analysis and does not integrate the adaptive resolution results for resynthesis of the signal which is required in source separation. More recent work focuses on adaptive time-frequency resolution both for analysis and synthesis. In the method proposed by Basu et al [57], the time-frequency resolution is adapted based on the measure used in [56] but the analysis method is altered thus it enables perfect reconstruction of the signal using a modified overlap-add (OLA) procedure. Their following work [58] introduces a broad family of adaptive, linear time-frequency representations termed superposition frames via two signal adaptation schemes based on greedy selection and dynamic programming. A given discrete Gabor frame is adapted to an observed signal via superposition of neighboring translates of a single window function to yield the superposition frames for adaptive time-frequency analysis and fast reconstruction. However, it does not address the window requirements to ensure the existence of upper superposition frame bounds in infinite-dimensional setting or characterize the structure of canonical superposition duals in such cases.

There are some works combining multiresolution approach in source separation. In [17], a gradient algorithm for the MultiResolution Frequency Domain (MRFD) Adaptive Decorrelation Filtering (ADF) approach is derived. It is denoted that a

large filter size is required for good separation performance but the filter size has to be limited to the order of 10^2 in order to minimize the permutation problem. In the MRFD algorithm, ADF is implemented in a multistage process with increasing filter size at each stage, thus once self-aligning the permutations in the early stages with small filter size they tend to remain aligned for increasing filter size. In [18], an NMF based separation is performed which increases the capability of separating the original signals from the mixture by adapting the time-frequency resolution of the STFT to the mixture's characteristics. The mixing is performed based on a transient detection measure and an energy concentration measure for each time frame. Thus, it does not consider the smearing around each time-frequency component but in each time frame. In [59], Wavelet Packet (WP) transform which is a multiscale transform is used to decompose signals into sets of local features with various degrees of sparsity. First, an overcomplete set of WP features of mixtures is constructed with various generating functions and various families of wavelets. Then, the best subset is chosen with respect to the estimation of the separation error and used for separation. This work enhances the quality of the separated sources and investigates how the separation error is affected by the sparsity of decomposition coefficients and by the misfit between the probabilistic model of these coefficients and their actual distribution. In our previous work [19], the learned speakers are separated at different time-frequency resolutions by applying NMF separately at each resolution and adaptively combining the separated sources based on maximum energy compaction principle [16]. The method proposed in this chapter combines the parallel NMF factorizations introduced in [19] into a single NTF scheme, thus performs a joint optimization by fusing the information coming from various time-frequency resolutions. It is shown that the fusion of the available information in a joint optimization scheme enhances the quality of the separated sources.

NTF has been widely used to separate multichannel recordings [60, 61] where the number of observations is more than one. CANDECOMP/PARAFAC (CP) and Tucker are special cases of NTF and widely used in various applications [62]. In [60], NTF is performed on two channel mixtures of three instruments by optimizing a KL divergence based on a PARAFAC model. In [61], a Shifted NTF (SNTF) method

is proposed which extends the SNMF to multichannel case for separating harmonic instruments from multichannel mixtures. In [34], a statistical NTF framework of the power and magnitude spectrograms is proposed for separating multichannel mixtures, based on the IS and the KL divergences, respectively. The model also allows to cluster the source components within the decomposition. In [63], under-determined blind separation of convolutive speech mixtures is performed by combining PARAFAC analysis with Capon beamforming in order to reduce the crosstalk. This work results in a low complexity adaptive blind source separation algorithm that can track the changes in the mixing environment by batch computation of the PARAFAC decomposition for each frequency bin. In [64], a feature extraction method based on Gabor filtering of primary cortical representation of speech signals and tensor factorization is introduced. Two-dimensional Gabor functions with different scales and directions are employed to analyze the localized patches of the power spectrogram. The Non-negative Tensor Principal Component Analysis (NTPCA) with sparse constraints is developed for multifactor analysis of speech by maximizing the covariance of data samples on the tensor structure. This model explores both the spectral and the temporal structures simultaneously within one model and the energy of speech signal is concentrated on a few components thus the consistent statistical characters are reserved while the noise components are suppressed. All the mentioned NTF based separation approaches use the information from different observations for extracting the bases and the gains by joint optimization. Different from the separation methods based on NTF, the proposed scheme uses only a single observation represented at various time-frequency resolutions and enhances the quality of the separated sources compared to the NMF based methods which use the information from a single observation represented at a fixed time-frequency resolution.

It is known that, the representation capability of the NMF based methods increase as the rank of the factorization is increased. However, when the rank is greater than the number of sources, the extracted bases place in a random order in the basis matrix. This requires clustering of the bases into sources by making some assumptions about the sources [10–12,21,28] or using some prior information [19,22]. To our knowledge, the unsupervised clustering methods are usually proposed for pitched instruments and

no unsupervised clustering method exists which works for all kinds of audio signals. Thus, in this work, it is preferred to learn the bases from the training data of each source which is not included in the test data with a high rank multiresolution based factorization. The source bases are fixed; the amplitude envelopes and the gains are extracted from the mixture signal by updating iteratively. The learning of the source bases is based on the work proposed in [22] and solves the clustering problem.

The rest of this chapter is organized as follows. The proposed adaptive time-frequency resolution based supervised single channel source separation method is introduced in Section 4.1. The experimental results obtained by the proposed methods are compared to the single resolution based NMF method and the MR-NMF method proposed in [19] and reported in Section 4.2. It is concluded with some discussions in Section 4.3.

4.1 Proposed Approach

The proposed MR-NTF approach aims to optimize a generalized KL divergence defined in (**2.15**). This process involves learning codebooks **S** of sources from the time-frequency representations of their training signals via NMF at various resolutions as it is proposed in [19]. By fixing the frequency bases **S**; **Q** and **A** are updated iteratively through the multiplicative update rules given in (**2.19**) and (**2.26**), respectively. The $\mathbf{Q}, \mathbf{S}$ and **A** factors are then applied on the magnitude spectrogram tensor **X** of the mixture to extract source estimates. Finally, the output from multiple resolutions are fused to obtain more robust estimates of the sources using the maximal energy compaction principle method used in [16], [19] for varying the time-frequency resolution adaptively. This approach estimates the sparsity of different time-frequency resolutions and mixes them accordingly so as to obtain minimal smearing both in time and frequency directions.

4.1.1 Dictionary learning

In order to eliminate the permutation problem, a pre-learning is applied on a training set of the original sources. Once the basis vectors of the sources are learned separately, they can be used to estimate each source from a monophonic mixture. Hence, supervised separation of two sources from a single observation is performed, where

Algorithm 5 Dictionary learning.

1: Input: $\mathbf{s}_j(t), \quad j = 1 \cdots J$
2: $\mathbf{X}_{jc} \leftarrow STFT_c(\mathbf{s}_j), \quad j = 1 \cdots J, c = 1 \cdots C$
3: Initialize $\mathbf{A}_{jc}$ and $\mathbf{S}_{jc}$ randomly
4: for $n = 1 \cdots$ MAXITER do
5:

$$\hat{X}_{jcki} = \sum_r S_{jckr} A_{jcri}$$

6:

$$A_{jcri} \leftarrow A_{jcri} \frac{\sum_k \frac{X_{jcki}}{\hat{X}_{jcki}} S_{jckr}}{\sum_k S_{jckr}}$$

7:

$$\hat{X}_{jcki} = \sum_r S_{jckr} A_{jcri}$$

8:

$$S_{jckr} \leftarrow S_{jckr} \frac{\sum_i \frac{X_{jcki}}{\hat{X}_{jcki}} A_{jcri}}{\sum_i A_{jcri}}$$

9: end for
10: Set $\mathbf{S} = \{S_{jckr}\}$, $j = 1 \cdots J, c = 1 \cdots C, k = 1 \cdots K, r = 1 \cdots R/C$.

the source signals can be either music or speech. The supervised or informed source separation is commonly used in the literature to separate the sources with the help of a prior information supplied as in the form of a prior distribution or pre-learned sources [22, 32, 65].

In particular, the MR-NTF bases of the $j-$th source ($j = 1 \cdots J$) are learned from the corresponding magnitude spectrograms $\mathbf{X}_{jc}$ at various resolutions $c = 1 \cdots C$ in the Dictionary Learning block of Figure 4.1. Here, NMF is used to learn the sound dictionaries for each source from a limited training set of signals from each source and then this information is used as prior knowledge for the separation. The bases extracted at different resolutions are combined into a matrix $\mathbf{S} = [\mathbf{S}_1 \cdots \mathbf{S}_{JR}]$ which is of size $K \times JR$, where R/C is the number of bases learned from each resolution $c = 1 \cdots C$ for each source $j = 1 \cdots J$. Note that, the observed mixture signal $\mathbf{x}(t)$ does not include the source signals used for learning.

The dictionary learning algorithm is summarized in Algorithm 5.

The MR-NTF is applied on the magnitude spectrogram tensor $\mathbf{X}$, which is constructed by tensorizing the magnitude spectrograms of the observed mixture at various resolutions, by fixing the bases to $\mathbf{S} \in \mathbb{R}^{K \times JR}$ and learning their amplitudes $\mathbf{A} \in \mathbb{R}^{I \times JR}$ and the gains $\mathbf{Q} \in \mathbb{R}^{C \times JR}$ of each factor in each resolution. The extraction of the sources is described in the following subsection.

4.1.2 Fusing the information from various time-frequency resolutions by MR-NTF

Building filter banks with variable time-frequency resolutions is commonly addressed for audio compression methods [66]. However, these approaches are limited by the fact that the compression requires to keep the size of the data representing the signal at a minimum. On the other hand, the audio processing methods such as source separation can benefit from redundancy which leads to multi-resolution framework presented here. Moreover, fusing the information from various time-frequency resolution based representations to extract the sources results in a better separation quality compared to using the information from different resolutions separately as it is proposed in [19].

Figure 4.1 illustrates the main steps of the NTF-based source separation algorithm running on the multiresolution representation of the input mixture. In the figure, only two resolutions are depicted for clarity, but the framework can be extended to any number. First, the magnitude spectrograms $\mathbf{X}_c, c = 1 \cdots C$ at different time-frequency resolutions are obtained. The hop size and the frequency grids should be equal for all C STFT resolutions in order to ensure that all the STFT magnitudes $\mathbf{X}_c$ are calculated in the same grid of the time-frequency locations. In order to achieve that, the smaller STFT windows are zero padded to ensure that all of the STFTs have the same number of frequencies. The $(k,i)-$th time-frequency component of the mixture magnitude spectrogram obtained at two different resolutions are represented as X_{cki} where $c = \{1,2\}$ in Figure 4.1. The magnitude spectrograms are combined into a tensor $\mathbf{X}$, where the $c-$th layer of $\mathbf{X}$ represents the mixture magnitude spectrogram $\mathbf{X}_c$ calculated based on the $c-$th resolution. Then, the bases learned at the dictionary learning block are fixed as $\mathbf{S} = [\mathbf{S}_1 \cdots \mathbf{S}_{JR}]$ where each column of $\mathbf{S}$ has K frequency components. MR-NTF is performed on the tensor $\mathbf{X} \in \mathbb{R}^{C \times K \times I}$ by fixing $\mathbf{S} \in \mathbb{R}^{K \times JR}$ and updating

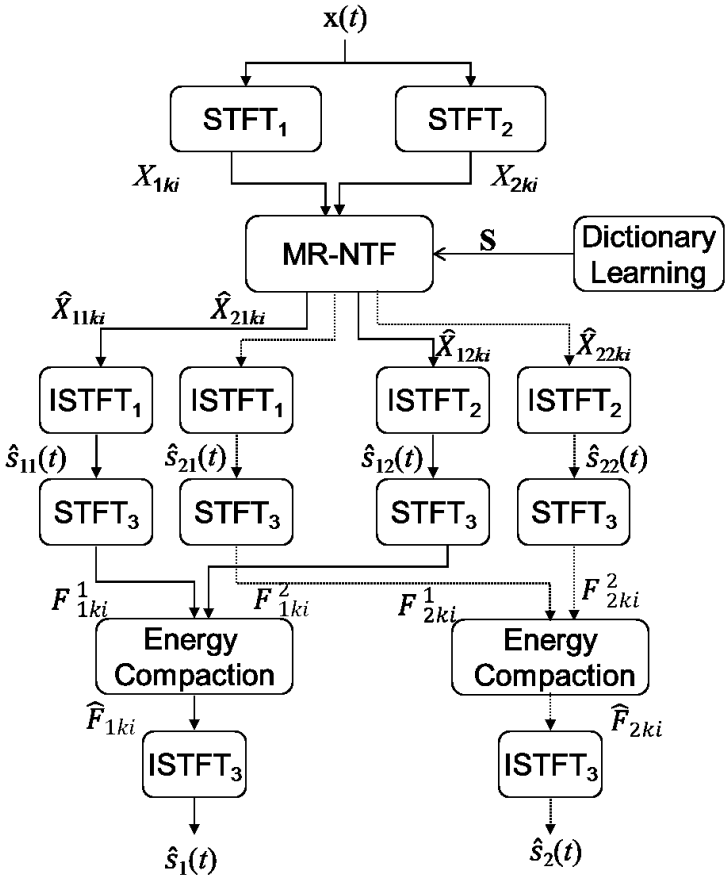


Figure 4.1: General scheme for the proposed source separation method using adaptive MR-NTF.

$\mathbf{A} \in \mathbb{R}^{I \times JR}$ and $\mathbf{Q} \in \mathbb{R}^{C \times JR}$ at each iteration. Upon convergence, $\mathbf{A} = [\mathbf{A}_1 \cdots \mathbf{A}_{JR}]$ and $\mathbf{Q} = [\mathbf{Q}_1 \cdots \mathbf{Q}_{JR}]$ are segmented into sources, each $\mathbf{A}_j, \mathbf{Q}_j$, $j = 1 \cdots J$ corresponding to one source. After obtaining the $\mathbf{Q}$ and $\mathbf{A}$ factors, the contribution of each source in the mixture magnitude spectrogram at the $c-$th resolution are estimated from:

$$\hat{X}_{jcki} = \sum_{r=1}^{R} Q_{jcr} S_{jkr} A_{jir}, \tag{4.1}$$

where S_{jkr} represents the $k-$th frequency component of the $r-$th basis vector of the $j-$th source; Q_{jcr} represents the gain of the $r-$th component of the $j-$th source in the $c-$th resolution; A_{jir} represents the time envelope for the $r-$th component of the $j-$th source in the $i-$th time frame. In Figure 4.1, the $k-$th frequency component of the $i-$th time frame of the estimated magnitude spectrograms for the $j-$th source obtained at the $c-$th resolution are represented as $\hat{X}_{jcki}$, where $j = \{1,2\}$ and $c = \{1,2\}$.

In order to estimate the sources in time domain, the complex source spectrograms should be estimated from the magnitude spectrograms. The complex source

spectrograms are directly retrieved using Wiener filter [41]

$$\hat{F}_{jcki} = F_{cki} \frac{\hat{X}_{jcki}}{\sum_j \hat{X}_{jcki}}, \tag{4.2}$$

where F_{cki} represents the $k-$th frequency component of the $i-$th time frame for the mixture signal represented at the $c-$th resolution. Time-domain estimate $\hat{s}_{jc}(t)$ of the $j-$th source at the $c-$th resolution is obtained by applying ISTFT on the estimated coefficients $\hat{F}_{jcki}$. In Figure 4.1, $\hat{s}_{jc}(t)$ represents the estimated $j-$th source at the $c-$th resolution, where $j = \{1,2\}$ and $c = \{1,2\}$.

4.1.3 Enhanced sparsity by maximal energy compaction

The aim is to combine the source signals from different resolutions in order to achieve an optimally compact representation in each part of the time-frequency plane. As it is described in Section 4.1.2, fusion of the information is performed in the separation step in order to use the information from different time-frequency resolution based representations. After resynthesis of the sources in various resolutions, another fusion is performed using the maximal energy compaction principle as in [16] by an additional filter bank, with a fixed time-frequency resolution that transforms the resulting separated source signals into the time-frequency coefficients on the same time-frequency grid as in the analysis steps. This is achieved by first transforming the estimated time-domain sources $\hat{s}_{jc}(t)$ into the time-frequency domain by using a fixed time-frequency resolution. The resulting STFTs $F^c_{jki}, j = 1 \cdots J$ correspond to the time-frequency representations of the sources obtained from MR-NTF in the $c-$th layer. In Figure 4.1, F^c_{jki} is represented for two sources $\{j = 1,2\}$ at two different resolutions $c = \{1,2\}$.

The fusion of the sound signals extracted at different layers of MR-NTF is performed for each source to obtain a sparser representation. In order to fuse the information efficiently at every time-frequency bin (k,i), a rectangular area Ω is considered around this point. There is a trade-off between selecting a small or a large area. If the area is small, there won't be enough coefficients to calculate a robust estimate of the energy smearing. If it is too big, it will not be a local estimate. Less smearing around a time-frequency component yields a sparser representation, thus maximizes the energy compaction. This process results in obtaining the separated sources with

less interference. In order to estimate the sparsity in a rectangular grid $\Omega = H \times U$, a kurtosis [16], [19] based method is used:

$$K^c_{jki} = \frac{\frac{1}{HU}\sum_{k',i' \in \Omega}(|F^c_{jk'i'}|^2 - |\bar{F}^c_j|)^4}{\left(\frac{1}{HU}\sum_{k',i' \in \Omega}(|F^c_{jk'i'}|^2 - |\bar{F}^c_j|)^2\right)^2}, \tag{4.3}$$

where K^c_{jki} is the kurtosis calculated for the $k-$th frequency component of the $i-$th frame of the $j-$th source in the $c-$th resolution; $|\bar{F}^c_j|$ is the sample mean of the squared STFT magnitudes $|F^c_{jki}|^2$ in the grid Ω. Kurtosis is widely used for measuring the nongaussianity of a distribution. If large number of components are inactive or almost inactive, the value of the kurtosis gets larger and the distribution gets more peaky [18, 67]. Therefore, kurtosis is widely used as a measure of sparsity [19, 56, 57] . In [57], it is denoted that in cases where the spectral peaks are rapidly varying across time, due to variation in vocal tract configuration, the spectral kurtosis measure tends to choose windows so that formant tracks are nearly flat within them.

In order to avoid hard switching from one resolution to another, the squared magnitude coefficients from different resolutions are fused adaptively. The fusion is performed by a weighted sum of the squared magnitude spectrogram coefficients:

$$|F_{jki}|^2 = \sum_{c=1}^{C} w^c_{jki} |F^c_{jki}|^2, \tag{4.4}$$

where F^c_{jki} is the $k-$th frequency component at the $i-$th time frame of the $j-$th source obtained for the $c-$th fixed time-frequency resolution and the mixing weights are calculated as:

$$w^c_{jki} = \frac{K^c_{jki}}{\sum_{c=1}^{C} K^c_{jki}}. \tag{4.5}$$

This step yields a single power-magnitude spectrogram $|\mathbf{F}_j|^2$ for each source. In (**4.4**), the estimated sources from different time-frequency resolutions are combined in an adaptive way, such that the smearing in both time and frequency is minimized.

In (**4.4**), the adaptive power magnitudes are calculated for each time-frequency component. Thus, the adaptive representation requires the phase information in order to estimate the time-domain sources. Wiener filtering of the mixture spectrogram $\mathbf{F}_j$ is performed such as

$$\hat{F}_{jki} = F_{ki} \frac{|F_{jki}|^2}{\sum_{j=1}^{J} |F_{jki}|^2}. \tag{4.6}$$

Algorithm 6 The MR-NTF algorithm based on the KL divergence.

1: Input: $\mathbf{x}(t)$
2: $\mathbf{X}_c \leftarrow STFT_c(\mathbf{x}), \quad c = 1 \cdots C$
3: Initialize $\mathbf{Q}$ and $\mathbf{A}$ randomly
4: for $n = 1 \cdots$ MAXITER do
5:
$$\hat{X}_{cki} = \sum_r Q_{jcr} S_{jkr} A_{jir}$$
6:
$$Q_{cr} \leftarrow Q_{cr} \left(\sum_{k=1}^{K} S_{kr} \sum_{i=1}^{I} \Lambda_{cki} A_{ir} \right) / \left(\sum_{k=1}^{K} S_{kr} \sum_{i=1}^{I} A_{ir} \right)$$
7:
$$\hat{X}_{cki} = \sum_r Q_{jcr} S_{jkr} A_{jir}$$
8:
$$A_{ir} \leftarrow A_{ir} \left(\sum_{c=1}^{C} Q_{cr} \sum_{k=1}^{K} \Lambda_{cki} S_{kr} \right) / \left(\sum_{c=1}^{C} Q_{cr} \sum_{k=1}^{K} S_{kr} \right)$$
9: end for
10: Calculate magnitude frequency spectrogram: $\hat{X}_{jcki} = \sum_r Q_{jcr} S_{jkr} A_{jir}$
11: Calculate complex frequency spectrogram: $\hat{F}_{jcki} = F_{cki} \frac{\hat{X}_{jcki}}{\sum_r \hat{X}_{cjki}}$
12: Calculate time domain estimate: $\hat{\mathbf{s}}_{jc}(t) \leftarrow \text{ISTFT}(\hat{\mathbf{F}}_{jc})$
13: $\mathbf{F}_j^c \leftarrow STFT_3(\hat{\mathbf{s}}_{jc})$
14: Calculate sparsity: $K_{jki}^c = \frac{\frac{1}{HU} \sum_{k',i' \in \Omega} (|F_{jk'i'}^c|^2 - |\bar{F}_j^c|)^4}{\left(\frac{1}{HU} \sum_{k',i' \in \Omega} (|F_{jk'i'}^c|^2 - |\bar{F}_j^c|)^2 \right)^2}$
15: Calculate fusing weights: $w_{jki}^c = \frac{K_{jki}^c}{\sum_c K_{jki}^c}$
16: Fuse the signals: $|F_{jki}| = \sum_c w_{jki}^c |F_{jki}^c|$
17: Estimate complex source spectrogram: $\hat{F}_{jki} = F_{ki} \frac{|F_{jki}|^2}{\sum_j |F_{jki}|^2}$
18: $\hat{\mathbf{s}}_j(t) \leftarrow \text{ISTFT}_3(\hat{\mathbf{F}}_j)$

The ISTFT is then applied on the complex STFT coefficients $\hat{F}_{jki}$ in order to transform the extracted sources $\hat{F}_{jki}$ to time domain signals $\hat{s}_j(t)$.

The proposed MR-NTF algorithm is summarized in Algorithm 6.

4.2 Performance Evaluation

In this section, we first describe the test data and then proceed with experiments.

4.2.1 Datasets

Three audio datasets are considered and described below.

- **Dataset A** consists of ten monophonic mixtures synthetically generated by summing $J = 2$ different but equal length sentences uttered by male and female speakers from the TIMIT database. A training data of length 21 to 33 sec is used for each speaker in order to learn the bases of each speaker. The length of the evaluation sentences are 2 to 3 sec long. All the audio files are sampled at 16 kHz.

- **Dataset B** consists of six synthetic monophonic mixtures of $J = 2$ musical sources (drums, cello, acoustic guitar, electric guitar, piano, bass, lead vocals) created using 5 sec excerpts of original separated tracks from the song "Sunrise" and "We are in Love" by S. Hurley, available under a Creative Commons License at [68]. A training data of length 30 sec is used for each source in order to learn the bases of each source. All audio files are downsampled to 16 kHz.

- **Dataset C** consists of three synthetic monophonic mixtures of $J = 2$ speech and music sources (cello, drums and piano). The speech and the musical sources are selected randomly from Dataset A and Dataset B, respectively. Training data is of length 21 to 33 sec while the length of the evaluation data is 2 to 3 sec.

Note that, all the test signals consist of two sources created by linear mixing of individual source signals with unity gain and the evaluation data is not included in the training set.

4.2.2 Model parameters

The effect of various parameters is investigated on the separation quality. The investigated model parameters are discussed in the following.

4.2.2.1 STFT parameters

The data is analyzed using a Hanning windowed STFT of $N_F = \{256, 512, 1024, 2048, 4096\}$ samples. The choice of the STFT window size is important and there is a trade-off between the time resolution and the frequency resolution. Although a long window size gives a good frequency resolution but poor time resolution, a short window size gives a good time resolution but a poor frequency resolution. The choice of the window length depends on the signal characteristics. Stationary signal analysis requires a good frequency resolution, while the transient signals or percussives can be analyzed better using a good time resolution. Thus, various STFT lengths $N_F = \{256, 512, 1024, 2048, 4096\}$ are used and the STFT magnitudes calculated using different STFT lengths are combined in a single tensor. The time-frequency magnitudes are calculated on the same grid by zero padding the windowed signals and using equal STFT analysis hops for every resolution which is equal to the half of the smallest STFT length. The estimated signals obtained from various resolutions are analyzed using a Hanning windowed STFT of length $N_F = 1024$ with 25% overlap.

4.2.2.2 NTF parameters

In our case, the NTF parameters consist of the rank of the factorization R and the number of the NTF layers C. The number of the NTF layers is actually determined by the number of the time-frequency resolutions which is directly related to STFT. The choice of the number of bases R per each source is important. If a high value is chosen, the model can represent the diversities in the source; e.g. using a single basis for each note or each elementary sound object. However, it may also cause the model to overfit the data. On the other hand, one or few components per source are not enough to represent all the diversities in the data. In the experiments of Section 4.2.3, the effect of the rank of the factorization is investigated by selecting the training number of bases for each source as $R = \{1, 10, 20, 50, 100\}$.

4.2.2.3 Grid size for sparsity

Another important parameter is related to the adaptive fusing. The adaptive fusing is performed based on a sparsity measure calculated over a time-frequency grid Ω around each time-frequency component with various sizes. The length of the rectangular

grid is selected as the combinations of $H = \{3,5,9\}$ time frames and $U = \{3,5,9\}$ frequency bins around the point of interest. The tradeoff here is that small Ωs do not allow to form a robust estimate of the energy smearing as there are too few STFT coefficients inside, and large Ωs are not local enough for fine control of the resolution.

4.2.3 Separation results

In this section, the performance improvement achieved by the joint optimization of the single channel source separation based on a multiresolution representation is investigated. The proposed approach fuses the separated signals from each layer of the tensor in an adaptive way, where each layer corresponds to the representation of the mixture signal with a different time-frequency resolution. The MR-NTF algorithm is run for fifty iterations. The separation is performed using all combinations of the parameters mentioned in Section 4.2.2 on our dataset which amounted to 45 experiments (five rank values, nine different grid size) for 19 test signals using seven different resolution set represented by $C = \{2,3,4,5\}$ layered tensors. First, the separation is evaluated using a $C = 5$ layered tensor where each layer corresponds to the magnitude spectrogram of the mixture signal obtained with STFT lengths of $N_F = \{256,512,1024,2048,4096\}$. The performance is also evaluated using $C = 2, C = 3$ and $C = 4$ layered tensors for representing the input mixture signal, where $C = 2, C = 3$ and $C = 4$ different STFT lengths out of $N_F = \{256,512,1024,2048,4096\}$ are selected. The performance measures are averaged for all these experiments separately on three datasets and the effect of the parameters mentioned in Section 4.2.2 is analyzed.

In these experiments, the aim is to investigate the improvement in the separation quality of the estimated sources obtained by adaptive time-frequency resolution and compare it with the results obtained at a single fixed time-frequency resolution. In the proposed MR-NTF approach, each layer of the input tensor represents a fixed time-frequency resolution based representation of the mixture signal. Thus, the output of the separation gives the separated sources in each resolution. In each row of Figure 4.2, the magnitude spectrograms of the mixture signal are illustrated at two different resolutions (a, b), the original signal at two different resolutions (c, d), the estimated signal from each

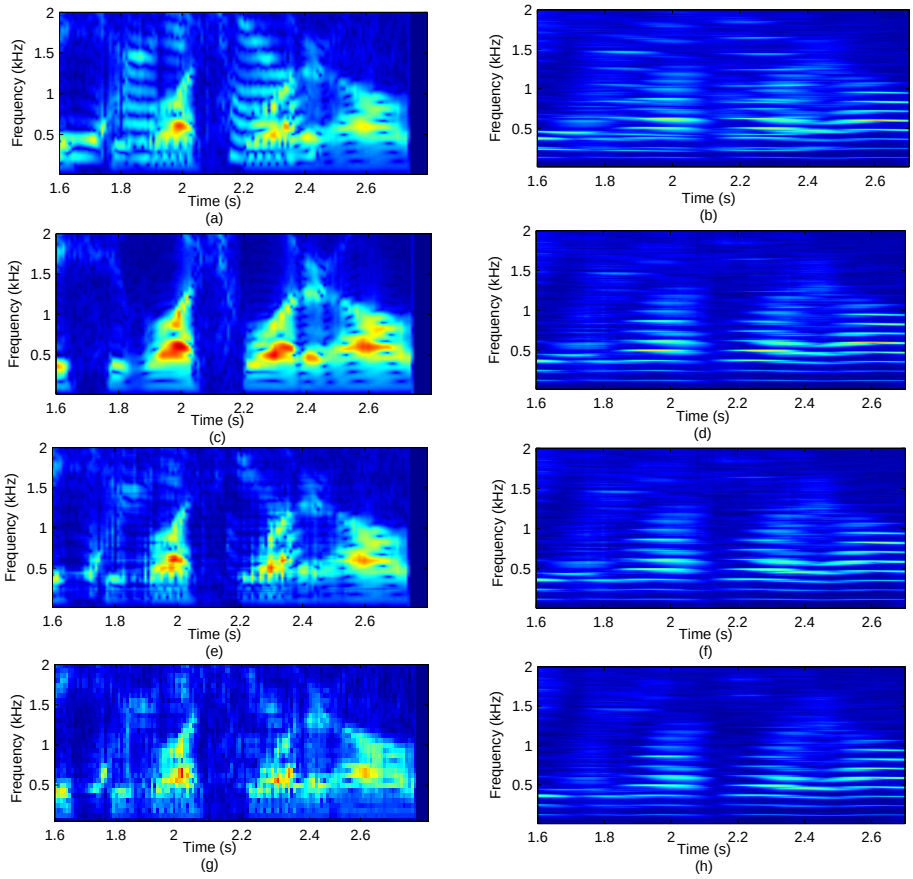


Figure 4.2: The magnitude spectrograms for (a) the mixture signal at $c = 1$st resolution (layer), (b) the mixture signal at $c = 2$nd resolution (layer), (c) the original source at $c = 1$st resolution (layer), (d) the original source at $c = 2$nd resolution (layer), (e) the estimated source from the $c = 1$st resolution (layer) where $N_F = 256$, (f) the estimated source from $c = 2$nd resolution (layer) where $N_F = 4096$, (g) the estimated source by NMF where $N_F = 256$, (h) the estimated source by NMF where $N_F = 4096$.

layer of the MR-NTF (e, f), and the estimated signal obtained by NMF(g, h), where $N_F = 256$ and $N_F = 4096$ for a test signal from Dataset A. These figures are cropped for the 1.6-2.8 seconds of the signal between the 0-2 kHz frequency range for a detailed illustration. In Figure 4.2, it can be seen that the frequency resolution of the first column is not enough to separate the harmonics and also smearing appears in frequency. In the second column of the figure, the frequency resolution of the representation is high enough and the harmonics are well separated in the frequency domain. However, the time resolution is low and smearing occurs in time. This smearing can be observed significantly between the 2 seconds and the 2.2 seconds of the signal in the second column of Figure 4.2. If the sources obtained by MR-NTF in the first layer (e) and NMF using a single resolution (g) are compared, it can be seen

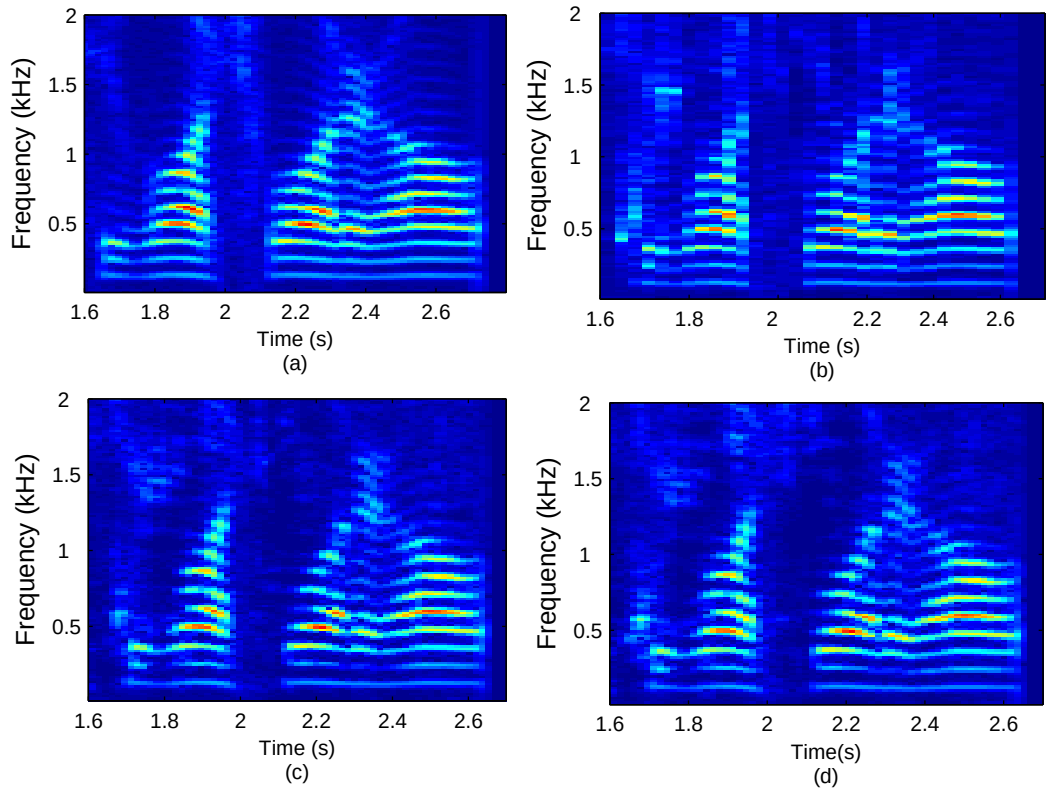


Figure 4.3: The magnitude spectrograms for (a)the original source, (b) the extracted source from a fixed resolution, (c) the extracted source by using MR-NMF, (d) the extracted source by using the proposed MR-NTF.

that fusing the information from different resolutions during optimization enhances the quality of the separated sources.

The proposed work introduces a method which fuses the separated sources observed in Figure 4.2 (e) and (f) and generates a signal which has adaptive time-frequency resolution. Since these signals have different resolutions, they cannot be fused and transformed to the time domain. Thus, each of the separated signal is first transformed into the time domain by Wiener filtering of the mixture spectrogram with the same resolution. Then, the signals are transformed back to time-frequency domain using the same STFT parameters for analysis. In Figure 4.3 (a), the magnitude spectrogram of the original source is illustrated in the new resolution where the STFT is applied using $N_F = 1024$ length Hanning window with 25% overlap. Figure 4.3 (b) illustrates the separated signal obtained using NMF at a fixed time-frequency resolution. The signals obtained using the adaptive time-frequency resolution with MR-NMF [19] and the proposed MR-NTF are represented in (c) and (d), respectively. It can be seen

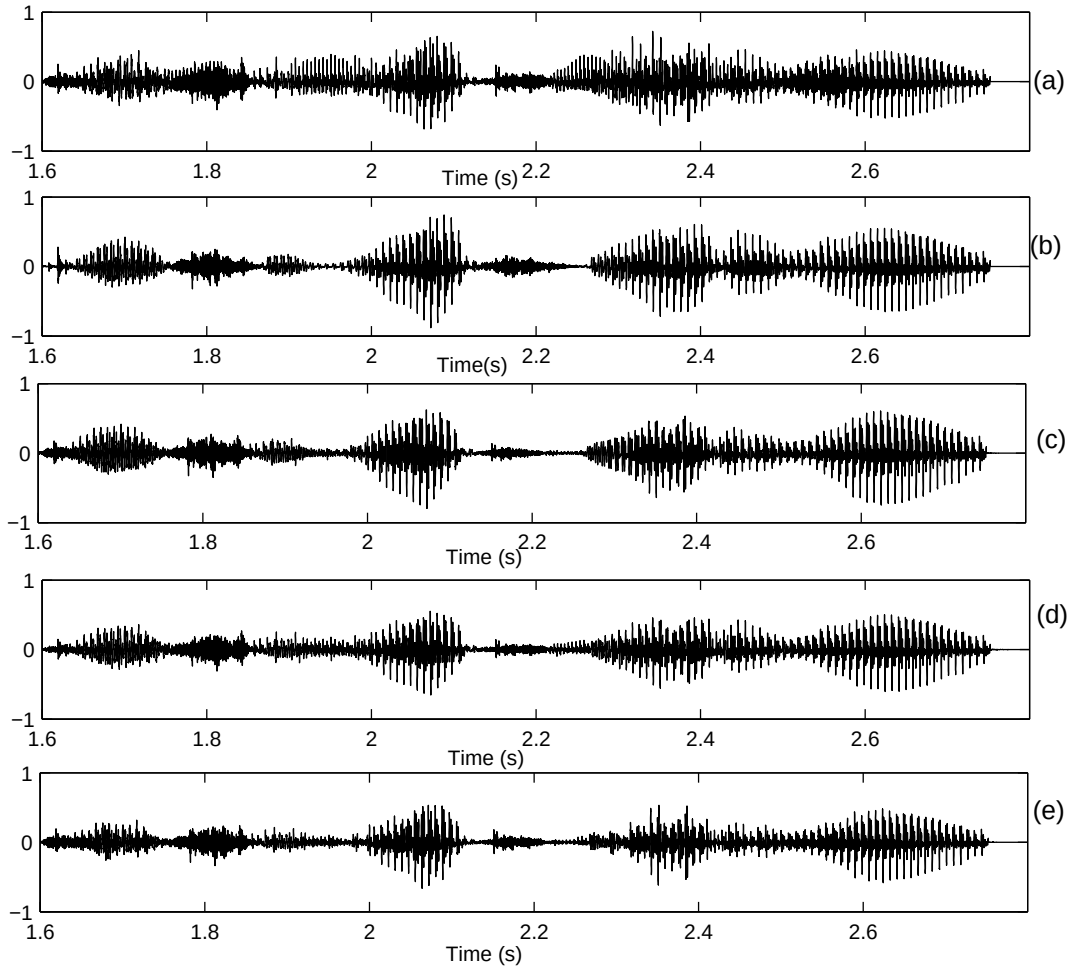


Figure 4.4: The time domain illustrations for (a) the mixture signal, (b) the original source signal, (c) the estimated source signal separated from the mixture signal based on the adaptive resolution, (d) the estimated source from the $c = 1-$st resolution (layer) where $N_F = 256$, (e) the estimated source from the $c = 2-$nd resolution (layer) where $N_F = 4096$.

that the smearing in time and frequency is minimized in adaptive resolution based representations.

The time-domain estimates of the signals illustrated in Figure 4.3 are shown in Figure 4.4 for the same time interval. It represents (a) the mixture signal (b) the original source signal, (c) the estimated signal separated from the mixture signal based on the adaptive resolution, (d) the separated signal from the $c = 1$st layer and (e) the separated signal from the $c = 2$nd layer. In Figure 4.4, more interference occurs from the unwanted source obtained from the $c = 1$st resolution illustrated in (d). The separated signal from the $c = 2$nd resolution has degradation of the signal in some parts which can be observed around the 2.4 seconds of the signal. Thus, the estimated signal from the

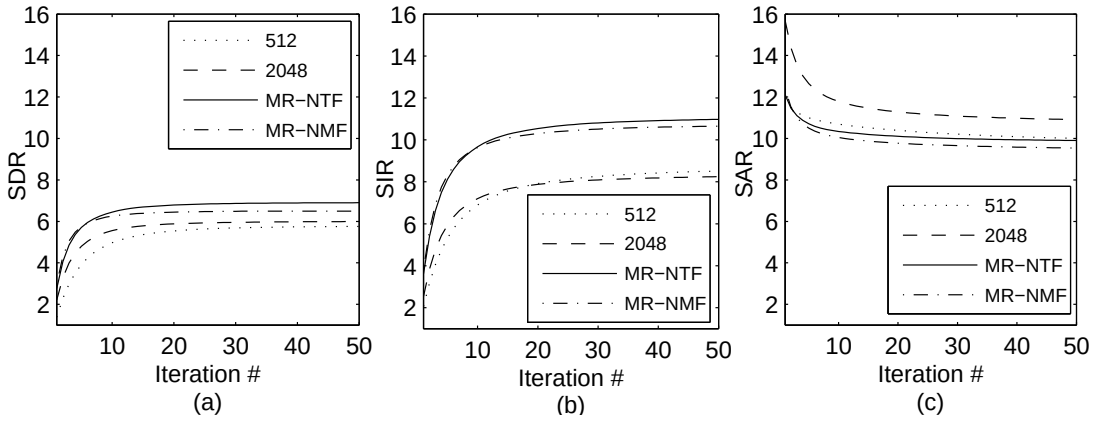


Figure 4.5: The mean SDR, SIR and SAR values with respect to iteration number obtained on Dataset A using fixed time-frequency resolution NMF, adaptive MR-NMF and MR-NTF.

adaptive resolution has less energy degradation and less smearing compared to single resolution results.

In the remaining part of this section, the effect of the parameters are investigated on the separation performance. The separation is performed at various resolutions with $C = \{2,3,4,5\}$ on Datasets A, B and C and the results are reported for each dataset separately.

The consistency of optimization of the proposed approach is checked with respect to source separation performance improvement as measured by SDR, SIR and SAR. Ten speech + speech mixtures of Dataset A are used and the MR-NTF algorithm is run for 50 iterations. $R = 20$ components are extracted for each source. Figure 4.5 displays the mean SDR, SIR and SAR values with respect to iterations obtained on ten mixtures using single resolution NMF with STFT lengths $N_F = 512$, $N_F = 2048$; MR-NMF [19] and the proposed MR-NTF. The results show that the source separation performance improves consistently in terms of SDR and SIR, while it decreases in terms of SAR in fixed and adaptive resolution based schemes. It is observed that both the MR-NMF and the MR-NTF methods improve the quality of the sources compared to the fixed resolution NMF results and MR-NTF outperforms MR-NMF in terms of SDR and SIR. Although running the algorithm for 30 iterations is enough for convergence, the MR-NTF scheme is run for 50 iterations for all test files.

Table 4.1: Performance in terms of SDR, SIR, SAR, OPS, IPS and APS values obtained by the MR-NTF method on a mixture of female and male speakers from Dataset A, where $C = 2$ is used with $N_F = \{512, 2048\}$. The measures are reported for the female (Fml) speaker, the male (Ml) speaker and their average (Avg).

	SDR			SIR			SAR		
	Fml	Ml	Avg	Fml	Ml	Avg	Fml	Ml	Avg
$c = 1$	8.5	8.9	8.7	11.3	13.3	12.3	12.0	11.0	11.5
$c = 2$	9.0	9.7	9.4	11.5	13.7	12.6	**13.0**	**12.1**	**12.6**
MR-NTF	**9.8**	**10.1**	**10.0**	**14.1**	**19.0**	**16.6**	12.0	10.8	11.4
	OPS			IPS			APS		
$c = 1$	32.2	47.7	40.0	37.8	63.8	50.8	60.2	43.1	51.7
$c = 2$	19.9	29.1	24.5	25.0	51.9	38.5	**71.3**	**75.9**	**73.6**
MR-NTF	**41.3**	**48.7**	**45.0**	**52.3**	**67.3**	**59.8**	53.9	38.1	46.0

After determining the iteration number, the separation results are first evaluated for ten speech + speech mixtures on Dataset A. In Table 4.1, the quality of the separated sources is reported in terms of mean SDR, SIR and SAR and their perceptual correspondences OPS, IPS and APS obtained from one speech mixture where rank is selected as $R = 20$ for each source. The best performance is obtained when $C = 2$ where $N_F = 512$ and $N_F = 2048$ are used for representing the mixture signal in different layers of the input tensor. Thus, only the results obtained on a $C = 2$ layered tensor where $N_F = \{512, 2048\}$ are used to represent the input mixture are reported. The results obtained on each layer are reported in the third and fourth rows of the table in terms of SDR, SIR and SAR. The adaptive results obtained by our MR-NTF method are reported on the fifth row where $U = 3$ and $H = 3$ is selected for the grid width (frequency components) and height (time frames) to calculate the kurtosis around each time-frequency component. It is observed that, the effect of the grid size for kurtosis analysis is not significant. Thus, in order to decrease the computational complexity, it is selected as $U = 3, H = 3$. It can be seen that, the proposed adaptive MR-NTF scheme improves the quality of the separated sources by 1-2 dB in terms of SDR and 3-6 dB in terms of SIR for both speaker compared to fixed-resolution results. The artifacts obtained for the adaptive resolution scheme and the fixed-resolution scheme are similar. The quality in terms of perceptual measures are reported in the lower part of Table 4.1. It is observed that an improvement of 1-22 and 4-27 points in terms

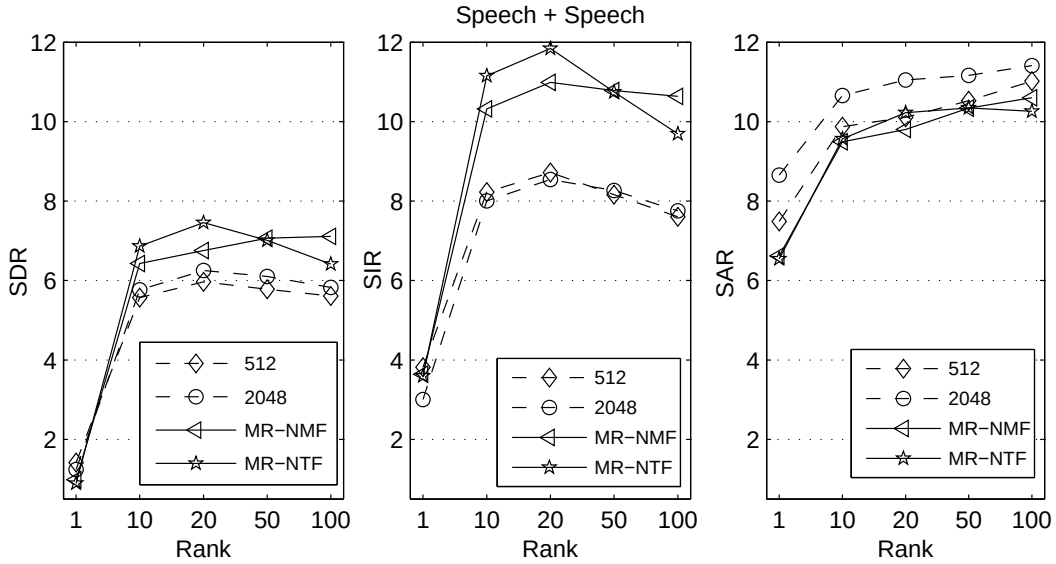


Figure 4.6: The mean SDR, SIR and SAR values obtained for various number of bases per each source obtained on Dataset A using a fixed time-frequency resolution NMF, MR-NMF and MR-NTF.

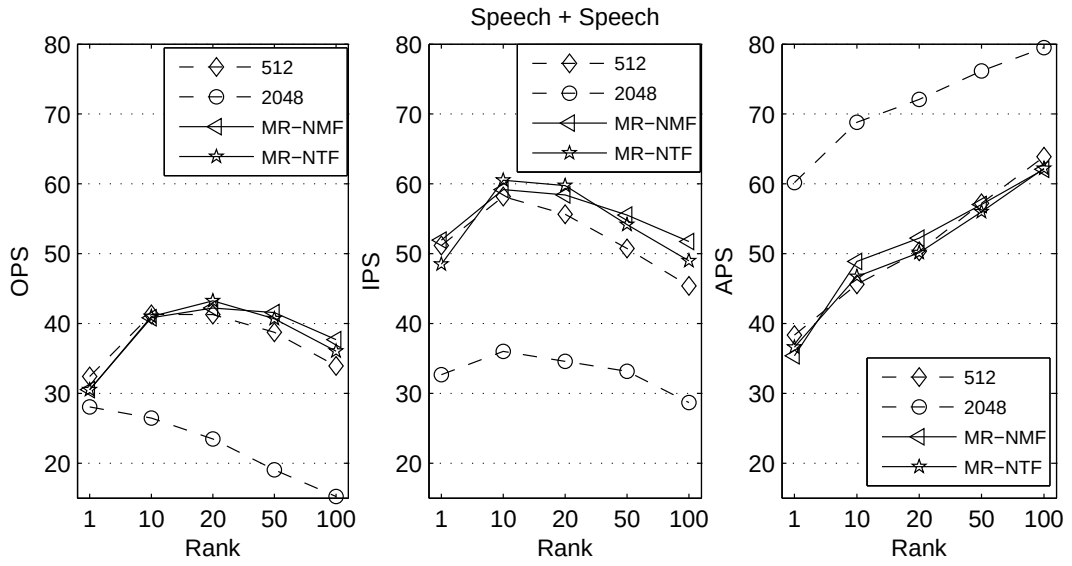


Figure 4.7: The mean OPS, IPS and APS values obtained for various number of bases per each source obtained on Dataset A using a fixed time-frequency resolution NMF, MR-NMF and MR-NTF.

of OPS and IPS are obtained by the proposed approach, respectively. The artifact is increased in the proposed approach by an amount of 5-43 points in terms of APS. Moreover, the results obtained in each layer outperforms the NMF factorization in a single time-frequency resolution slightly.

In Figure 4.6 and 4.7, the effect of the rank parameter is investigated on the separation performance. Thus, $R = \{1, 10, 20, 50, 100\}$ bases are learned for each source in the

training and used for separating the sources. The separation results are averaged over a set of ten pairs of male/female speakers from the Dataset A. The horizontal axes in Figure 4.6 and 4.7 display the number of bases per each source. In the figures, the results obtained by fixed time-frequency resolutions from each tensor layer are also plotted to show the improvement by mixing the fixed time-frequency resolution results in an adaptive way. As it is seen from Figure 4.6, the proposed method increases the separation quality by around 2 dB and 3 dB in terms of SDR and SIR while the SAR is decreased. In terms of the perceptual measures, an improvement is achieved by the proposed method in terms of OPS and IPS around 2-5 and 10, respectively. Moreover, learning $R = 20$ bases for each source gives the optimum results in terms of all measures. In order to compare the performance of the proposed MR-NTF method with the previous work [19] named as MR-NMF, the same tests are performed with same parameters. In MR-NMF, the multiple NMF instances are run in parallel and the sources obtained from different resolutions are merged as it is described in Section 4.1.3. If MR-NMF and MR-NTF results are compared, it can be seen that both methods outperform the results obtained from a fixed resolution. MR-NTF outperforms MR-NMF by 1-2 dB in terms of SDR and SIR for rank values of 10 and 20. It can be concluded that, choosing a rank value of 20 is enough for extracting the sources successfully which gives the opportunity to separate the sources with a lower computational complexity.

In Table 4.2, the quality of the separated sources are reported in terms of SDR, SIR, SAR and their perceptual correspondences OPS, IPS, APS obtained from a music mixture consisting of a drum and an electric guitar signal where the rank is selected as $R = 20$ per each source. The results are obtained on a $C = 4$ layered tensor where $c = \{1,2,3,4\}$ represent the time-frequency resolution for STFT lengths $N_F = \{512,1024,2048,4096\}$, respectively. The results obtained on each layer are reported in the 3-6th rows of the table in terms of SDR, SIR and SAR, followed by the adaptive MR-NTF results. In order to see the effect of perceptual weighting in the separation, the separation is also performed by integrating the perceptual weighting proposed in [28] into the MR-NTF approach. The results obtained by perceptually weighted MR-NTF are represented as PWMR-NTF in Table 4.2. It is observed that,

Table 4.2: Performance in terms of SDR, SIR, SAR, OPS, IPS and APS values obtained by the MR-NTF method on a mixture of drums and electric guitar excerpts from Dataset B, where $C = 4$ is used with $N_F = \{512, 1024, 2048, 4096\}$. The measures are reported for the drums (Drms), the electric guitar (El.Gtr) and their average (Avg).

	SDR			SIR			SAR		
	Drms	El.Gtr	Avg	Drms	El.Gtr	Avg	Drms	El.Gtr	Avg
$c = 1$	14.5	15.4	14.9	16.5	19.4	17.9	18.8	17.6	18.2
$c = 2$	14.4	16.5	15.4	16.0	21.7	18.9	19.4	18.0	18.7
$c = 3$	15.7	17.2	16.4	17.9	23.4	20.6	19.7	18.4	19.1
$c = 4$	15.6	**17.5**	16.6	17.7	**23.6**	20.6	**20.0**	**18.7**	**19.4**
MR-NTF	16.5	16.5	16.5	**20.3**	21.6	21.0	18.9	18.2	18.5
PWMR-NTF	**16.6**	16.8	**16.7**	20.3	22.4	**21.3**	19.0	18.2	18.6
	OPS			IPS			APS		
$c = 1$	45.5	45.7	45.6	66.3	63.7	65.0	45.1	48.1	46.6
$c = 2$	**46.3**	43.6	45.0	62.7	58.9	60.8	50.8	53.6	52.2
$c = 3$	41.9	34.4	39.2	55.2	44.7	49.9	60.6	63.6	62.1
$c = 4$	23.9	19.3	21.6	40.8	33.0	36.9	**73.8**	**68.8**	**71.3**
MR-NTF	44.2	**47.1**	**45.6**	69.2	65.1	67.2	48.6	53.9	51.3
PWMR-NTF	43.2	44.0	43.6	**69.4**	**68.1**	**68.8**	47.3	46.3	46.8

the proposed adaptive MR-NTF scheme improves the quality of the separated sources by 1-3 dB in terms of SDR and 1-3 dB in terms of SIR for both sources compared to the fixed-resolution results. Although the SIR value obtained on the $c = 4$th layer is slightly higher than the SIR value obtained by the proposed MR-NTF method, the average SIR value is increased by the proposed method. The artifacts obtained for the adaptive resolution scheme and the fixed-resolution are similar. The quality of the separated sources in terms of perceptual measures is reported in the lower part of Table 4.2. The difference between the OPS values obtained in different layers of NTF are very high, and the OPS value of the proposed MR-NTF approach is very similar to the highest scores obtained for each source in different layers. The IPS value is increased by 2-33 points and it is significantly higher than the IPS values obtained on a single layer. Again, APS is decreased by the proposed method. If the effect of the perceptual weighting is investigated, it is observed that the PWMR-NTF method enhances the quality of the separated sources slightly compared to MR-NTF scheme.

In Figure 4.8 and 4.9, the effect of the rank parameter is investigated for separation of two instruments from Dataset B. The results are averaged over a set of six mixtures

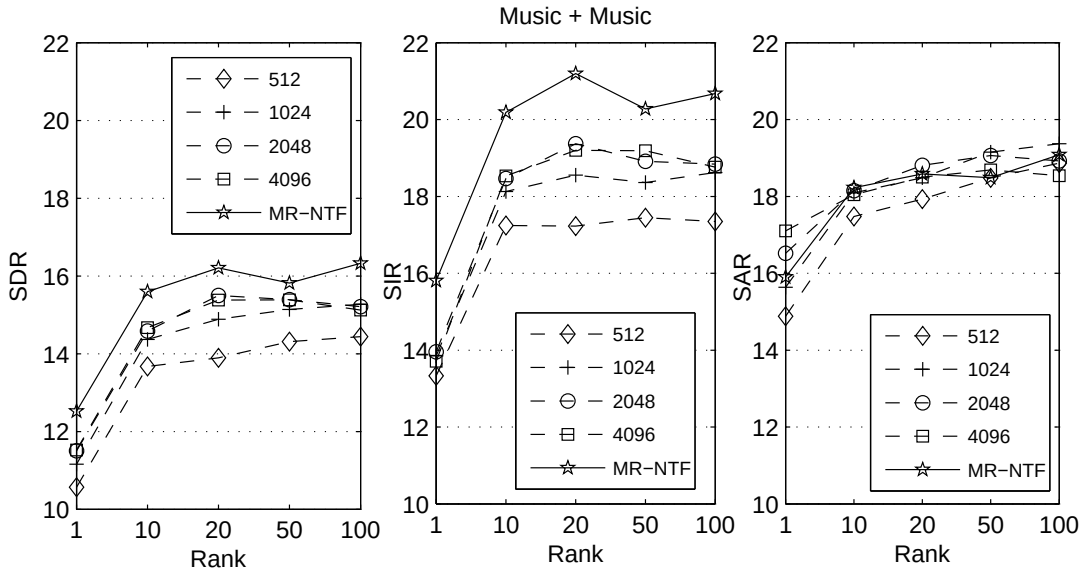


Figure 4.8: The mean separation results in terms of SDR, SIR and SAR for various number of bases per each source obtained on Dataset B.

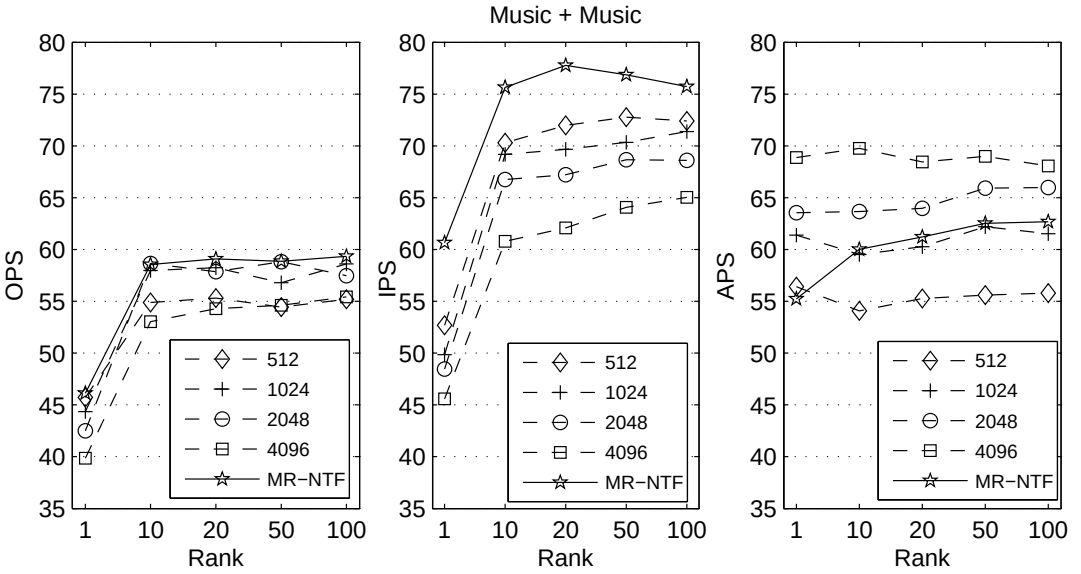


Figure 4.9: The mean separation results in terms of OPS, IPS and APS for various number of bases per each source obtained on Dataset B.

from Dataset B. It is observed that, the quality in terms of total distortion (SDR and OPS) and interference (SIR and IPS) is increased by the proposed MR-NTF method. On the other hand, the artifacts obtained by the proposed method seems to be around the average of the artifact distortions (APS and SAR) obtained by the fixed time-frequency resolutions. Also, learning $R = 20$ bases for each source gives the optimum results in terms of all measures.

Table 4.3: Performance in terms of SDR, SIR, SAR, OPS, IPS and APS values obtained by the MR-NTF method on a mixture of a male speaker and a piano excerpt from Dataset C, where $C = 3$ is used with $N_F = \{512, 1024, 2048\}$. The measures are reported for the male (Male) speaker, the piano (Piano) and their average (Avg).

	SDR			SIR			SAR		
	Male	Piano	Avg	Male	Piano	Avg	Male	Piano	Avg
$c = 1$	4.2	12.1	8.2	7.2	14.3	10.8	8.0	16.3	12.2
$c = 2$	7.6	14.2	10.9	12.2	16.7	14.5	9.7	18.0	13.9
$c = 3$	8.5	**15.3**	**11.9**	13.3	**18.4**	15.9	**10.5**	**18.2**	**14.3**
MR-NTF	**8.6**	14.3	11.5	**17.5**	16.9	**17.2**	9.3	17.9	13.6
	OPS			IPS			APS		
$c = 1$	45.0	**42.7**	**43.8**	63.0	**60.6**	61.8	36.8	66.54	51.7
$c = 2$	51.9	27.1	39.5	69.0	57.8	**63.4**	46.9	**71.4**	59.2
$c = 3$	**54.1**	18.8	36.5	69.7	54.8	62.2	**57.6**	69.1	**63.36**
MR-NTF	42.3	34.4	38.3	**73.4**	51.8	62.6	37.5	67.4	54.4

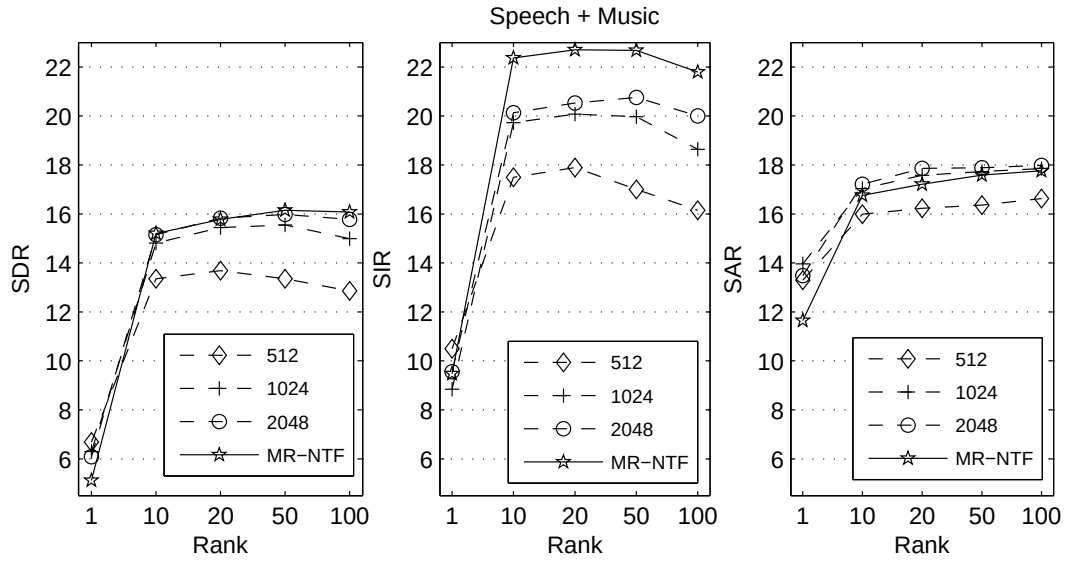


Figure 4.10: The mean SDR, SIR and SAR values obtained for various number of bases per each source obtained on Dataset C.

On Dataset C, the highest performance is obtained when MR-NTF is applied on a three layered tensor, where the tensor layers represent the STFT magnitudes with window lengths $N_F = \{512, 1024, 2048\}$, respectively. The improvement obtained by adaptive time-frequency resolution is less significant on Dataset C as it can be seen from the results reported in Table 4.3 for a mixture of a male speaker and a piano excerpt. Again, the highest separation performance is obtained when $R = 20$ bases are learned for each source as it can be seen from Figure 4.10 and Figure 4.11.

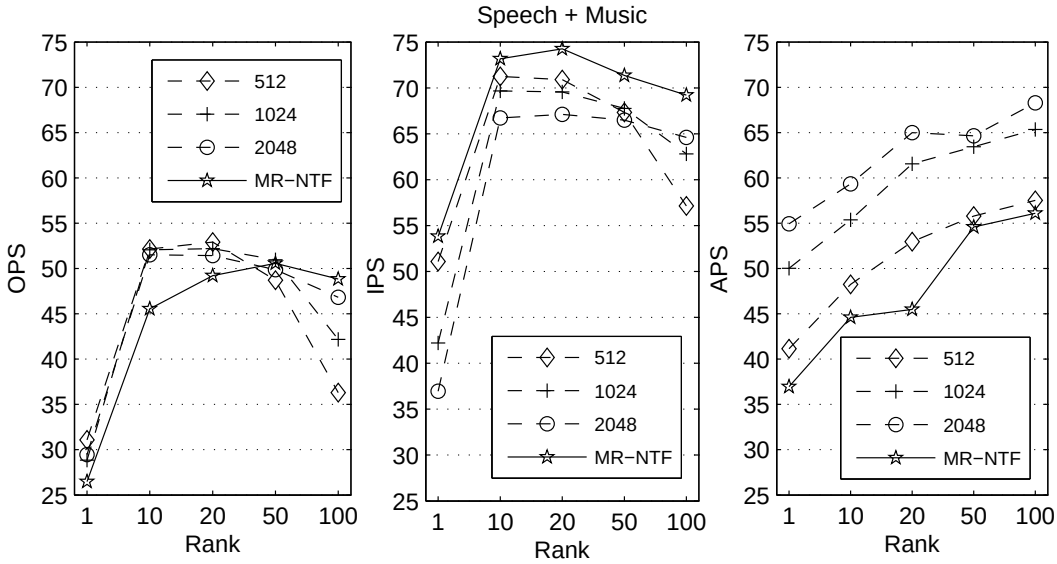


Figure 4.11: The mean OPS, IPS and APS values obtained for various number of bases per each source obtained on Dataset C.

Table 4.4: Running time in seconds for a typical run of KL-divergence based MR-NMF and MR-NTF on a 3 sec mixture.

Methods	MR-NMF		MR-NTF	
Rank	Train	Test	Train	Test
1	15.30	1.07	15.58	1.03
10	15.74	0.98	15.93	1.02
20	15.55	1.02	16.13	1.07
50	16.36	1.24	16.90	1.25
100	17.53	1.63	18.06	1.63

4.2.4 Computational complexity

In this subsection, the computational complexity of the proposed MR-NTF and MR-NMF methods is discussed.

In both MR-NTF and MR-NMF, the fusing scheme is common. Thus, the fusing step is excluded in the complexity measure.

In order to measure the computational complexity of the methods, a 3 sec audio mixture is separated on a PC equipped with a 2.4 GHz Intel® Core™2 Quad processor. The separation time is measured by excluding the file read/write operations and the spectrogram calculation/inversion stages. In Table 4.4, the running time is listed in seconds obtained by the MR-NMF and the MR-NTF methods based on the

KL divergence and under the same starting conditions. It is observed that the running time for MR-NMF and MR-NTF is similar. Although the running time required for learning the bases in the training step is high and can be performed offline prior to the separation, both MR-NMF and MR-NTF schemes work real-time.

4.3 Conclusion

In this chapter, an adaptive time-frequency resolution based supervised method has been presented for separating known types of sounds from a single observation. The single mixture signal has been represented in various time-frequency resolutions in different layers of the tensor. Then, the proposed MR-NTF approach has been performed in order to extract the sources. The decomposition algorithm has been applied on single channel mixtures of two sources where the sources consist of two musical signals, two speech signals or a music and a speech signal. The clustering problem has been resolved by learning the bases of the sources and fixing the bases during the factorization.

An improvement of 1-4 dB has been observed in terms of SDR and SIR relative to the fixed time-frequency resolution separation results. In terms of perceptual measures OPS and IPS, the improvement has been around 5-30. The proposed algorithm has been shown to outperform the single channel NMF algorithm and successfully separate the sources if the rank of the factorization is selected appropriately. The NTF algorithm is unique, thus the proposed method is not sensitive to the initialization.

5. BAYESIAN INFERENCE FOR NON-NEGATIVE MATRIX FACTOR DECONVOLUTION MODELS

In most of the audio source separation problems, it is difficult to have complete information about the sources even the number of sources. Hence it is common to make an initial assumption about the number of sources before applying source separation. Similarly we simplify the separation problem with extra constraints depending on assumed type of the sources whether they are music, speech or both. In these cases, in order to find a solution, an inductive reasoning based procedure is required. Recent work on audio source separation deals with Bayesian methods that can be effectively used to solve these problems that prevent to deduce a unique solution.

In Bayesian methods, probabilistic generative models of the source signals are built and the source signals are estimated in a minimum mean squared error (MMSE) or Maximum a Posteriori (MAP) estimation from the mixture. Generally, the model consists of the latent variables and prior conditional distributions between the variables which describe the mixing matrix, the set of source signals or giving more details about the relative positions of the sources. After constructing the model, the probability of the particular values of the parameters for describing the situation accurately are calculated based on the observations and the prior information.

There are several works in the literature that combines the NMF based audio source separation with Bayesian methods. In the early works, ICA is investigated under a Bayesian approach which also relaxes the hypotheses of ICA such as square mixing matrix and independent and identically distributed sources [69]. This approach also enables to incorporate prior information about the mixing matrix into the separation scheme. In underdetermined cases, the problem is ill-posed because the mixing matrix is not invertible and the prior information about the sources is important for separation. The Bayesian approach allows to incorporate the available information about the unknowns into the separation through the priors.

The statistical approaches offer both a strong theoretical framework and have the advantage to make the assumptions and manage the constraints through models and priors. Another advantage of the statistical approaches is the possibility to switch from the ML estimation to the MAP estimation by using the Bayes rule [70]. Thus, choosing adequate prior distributions $p(\mathbf{S})$ and $p(\mathbf{A})$ enable to introduce the desired properties into the decomposition. Furthermore, the statistical framework provides a strong theoretical basis and efficient algorithms with proven convergence, like the EM algorithm and its variants to estimate the NMF factors [70].

Theoretically, the NMF based algorithms can be interpreted as computing a ML or a MAP estimate of the non-negative factorizing matrices under some assumptions on the distribution of the data and the factors. The main limitation of conventional subspace learning methods is the deficiency of incorporating the prior information into the model explicitly while proving convergence. In the NMF based separation scheme introduced in Chapter 3, no prior information is used about the sources and the separation of the musical sources is performed based on the shift-invariance assumption. Chapter 4 introduces a supervised approach which proposes to learn the frequency bases of the sources from the available training data prior to the separation of the mixture. Recently, informed source separation [32] methods are proposed, in which the available prior information about the sources are incorporated into the separation scheme through a source model.

Probabilistic decomposition enables controlling the form of the learned bases and gains through the imposition of prior probabilities and incorporating constraints through these priors. In [71, 72], it is proven that the multiplicative update rules of $\mathbf{S}$ and $\mathbf{A}$ obtained for NMF with the KL divergence is equivalent to the ML estimation of the parameters $\{\mathbf{S}, \mathbf{A}\}$ under some assumptions. In this framework, the original nonnegative multiplicative update equations of NMF appear as an EM algorithm for ML estimation of a conditionally Poisson model via data augmentation. Starting from this view, it is possible to do Bayesian model selection that automatically adjust the sparseness criteria from the data. Inference in the resulting models can be carried out easily using variational (structured mean field) Bayes (VB). The resulting algorithm opens up the way for a full Bayesian treatment for model selection via computation of

the marginal likelihoods (the evidence), such as estimating the dimensions of the basis matrix. In [70], a Bayesian approach of NMF is introduced which allows to enforce harmonicity in the columns of the basis matrix **S** and temporal smoothness in the rows of the encoding matrix **A** which are the desired properties for the musical sources.

Knowing that the multiplicative update rules of NMF is equivalent to ML estimation of the model parameters under a Bayesian approach [72], in this chapter, we extend this framework to NMFD [9] in order to take into account the temporal information in source separation. It is shown that the proposed BNMFD retains the attractive features of the conventional NMFD such as easy implementation and monotonic convergence while enabling to develop more powerful models by incorporating the available prior information into the decomposition algorithm. Furthermore, it can be considered as an attempt to automatic selection of rank parameter without knowing the number of sources.

5.1 The Statistical Perspective

In NMFD, the non-negative observation data matrix $\mathbf{X} \in \mathbb{R}^{K \times I}$ is factorized to be a convolution of non-negative matrices $\mathbf{A}$ and $\mathbf{S}^{\tau}$ which depends on time τ:

$$X_{ki} \approx \hat{X}_{ki} = \sum_{r=1}^{R} \sum_{\tau=0}^{T-1} S_{kr}^{\tau} A_{r,i-\tau} \tag{5.1}$$

where R is the number of sources. The matrices $\mathbf{S} \in \mathbb{R}^{K \times R \times T}$ and $\mathbf{A} \in \mathbb{R}^{R \times I}$ are referred to as the basis matrix and the gain matrix, respectively. It is also possible to include a "2-D" convolution in the k dimension as well [11]; this is straightforward to include in our framework but is omitted for simplicity.

Subject to the positivity constraints, $(\mathbf{S}, \mathbf{A})$ are extracted which minimizes the KL divergence defined in (**2.2**). Obviously, as NMFD adds new dimensions to the basic NMF model, the number of free parameters increases drastically. Hence, for prediction, interpolation or separation tasks it is important to incorporate a form of regularization. The standard approach to regularization is to make use of sparsity, i.e. enforcing most of the model parameters to be zero or close to zero. However, it is not always clear how to choose the optimal criteria. In this chapter, an approach based on a full Bayesian treatment is investigated, where the sparseness criteria can be selected

automatically from data. Therefore, first a statistical perspective is needed for the basic NMFD model.

In this subsection, NMFD is described from a statistical perspective. This view opens up the way for developing extensions that facilitate more sophisticated inference as well as more realistic and flexible modeling. The first step is the derivation of the KL divergence from a ML principle. The following hierarchical model is considered for the basis and the gain matrices:

$$\mathbf{S} \sim p(\mathbf{S}|\theta^S) \tag{5.2}$$

$$\mathbf{A} \sim p(\mathbf{A}|\theta^A), \tag{5.3}$$

where $p(\mathbf{S}|\theta^S)$ and $p(\mathbf{A}|\theta^A)$ are the prior distributions of $\mathbf{S}$ and $\mathbf{A}$ with hyperparameters θ^S and θ^A, respectively. Depending on the degree of prior information about $\mathbf{S}$ and $\mathbf{A}$, the hyperparameters can be either fixed or they can be estimated from the observed data $\mathbf{X}$.

The total magnitude of the observed signal X_{ki}, in each time-frequency point is the sum of the magnitudes of individual sources, i.e.

$$X_{ki} = \sum_{\tau}\sum_{r} Y_{rki}^{\tau}. \tag{5.4}$$

It is assumed that the magnitude at each time-frequency component Y_{rki}^{τ} produced by the $r-$th source is Poisson distributed:

$$Y_{rki}^{\tau} \sim \mathscr{P}\mathscr{O}(Y_{rki}^{\tau}; \hat{Y}_{rki}^{\tau}) \tag{5.5}$$

where $\hat{Y}_{rki}^{\tau} = S_{kr}^{\tau} A_{r,i-\tau}$ and $\mathscr{P}\mathscr{O}$ denotes the Poisson distribution [72]. It is shown in [72] that the KL minimization formulation of the NMF algorithm can be derived from a probabilistic model where the observations are superposition of R independent Poisson distributed latent sources.

The variables $\mathbf{Y}_r = \{Y_{rki}^{\tau}\}$ are called as latent sources and they can be analytically marginalized out to obtain the marginal likelihood

$$\begin{aligned} \log p(\mathbf{X}|\mathbf{S},\mathbf{A}) &= \log \sum_{\mathbf{Y}} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}|\mathbf{S},\mathbf{A}) \\ &= \log \prod_{k}\prod_{i} \mathscr{P}\mathscr{O}(X_{ki}; \sum_{r}\sum_{\tau} \hat{X}_{rki}^{\tau}), \end{aligned} \tag{5.6}$$

which results from the well known superposition property of Poisson random variables. The maximization of this objective in **S** and **A** is equivalent to the minimization of the KL divergence in (**2.2**). In the derivation of original NMFD in [9], this objective is stated first; the **Y** variables are introduced implicitly later during the optimization on **S** and **A**. This algorithm is actually equivalent to EM, ignoring the priors $p(\mathbf{S}|.)$ and $p(\mathbf{A}|.)$.

Given the probabilistic interpretation, it is possible to propose various hierarchical prior structures to fit the requirements of an application. Here, a simple choice of a conjugate prior is described as

$$\begin{aligned} S^{\tau}_{kr} &\sim \mathcal{G}(S^{\tau}_{kr}; a^{\tau S}_{kr}, b^{\tau S}_{kr}/a^{\tau S}_{kr}) \\ A_{ri} &\sim \mathcal{G}(A_{ri}; a^{A}_{ri}, b^{A}_{ri}/a^{A}_{ri}), \end{aligned} \tag{5.7}$$

where, $\mathcal{G}$ denotes the density of a gamma random variable. The main reason for choosing a Gamma distribution is its computational convenience: Gamma distribution is the conjugate prior to Poisson intensity. Qualitatively, the shape parameter a controls the sparsity of the representation. Remember that $\mathcal{G}(x; a, b/a)$ has the mean b and standard deviation $b/\sqrt{a}$. Hence, for large a, all coefficients have more or less the same magnitude b and typical representations are full. In contrast, for small a, most of the coefficients are very close to zero and only very few dominates, hence favoring a sparse representation.

The hierarchical model in (**5.7**) is potentially more flexible than the basic model given in (**5.4**) and (**5.5**), in that it allows a lot of freedom for more realistic modeling. First of all, the hyperparameters can be estimated from examples of a certain class of source to capture the invariant features. Another possibility is Bayesian model selection, where alternative models can be compared in terms of their marginal likelihoods. This enables to estimate the model order, for example, the optimum number of bases R or the optimum number of convolution order T to represent a source.

5.2 Full Bayesian Inference using Variational Bayes

In source separation with NMFD, the posterior of the bases **S** and the gains **A** are needed to be calculated given data and hyperparameters $\theta \equiv (\theta^S, \theta^A)$. Another

important quantity is the marginal likelihood (also known as the evidence) which can be written as:

$$p(\mathbf{X}|\theta) = \int d\mathbf{S} d\mathbf{A} \sum_{\mathbf{Y}} p(\mathbf{X}|\mathbf{Y}) p(\mathbf{Y}|\mathbf{S},\mathbf{A}) p(\mathbf{S},\mathbf{A}|\theta). \tag{5.8}$$

The marginal likelihood can be used to estimate the hyperparameters θ, given examples of a source class

$$\theta^* = \arg\max_{\theta} p(\mathbf{X}|\theta), \tag{5.9}$$

or to compare two given models via *Bayes factors*

$$l(\theta^S, \theta^A) = \frac{p(\mathbf{X}|\theta^S)}{p(\mathbf{X}|\theta^{\mathbf{A}})}. \tag{5.10}$$

This latter quantity is especially useful for comparing different classes of models. However, the integrations can not be obtained in closed form. In order to approximate inference strategies, VB is described in this thesis.

The VB [73] method is sketched here to bound the marginal loglikelihood as

$$\begin{aligned} \mathscr{L}_{\mathbf{X}} \equiv \log p(\mathbf{X}|\theta) &\geq \sum_{\mathbf{Y}} \int d\mathbf{S} d\mathbf{A} q \log \frac{p(\mathbf{X},\mathbf{Y},\mathbf{S},\mathbf{A}|\theta)}{q} \\ &= \langle \log p(\mathbf{X},\mathbf{Y},\mathbf{S},\mathbf{A}|\theta) \rangle_q + \mathscr{A}[q] \equiv \mathscr{B}_{VB}[q], \end{aligned} \tag{5.11}$$

where, $q = q(\mathbf{Y},\mathbf{S},\mathbf{A})$ is an instrumental distribution and $\mathscr{A}[q]$ is its entropy. The bound is tight for the exact posterior, but as this distribution is complex, a factorized form for the instrumental distribution is assumed by ignoring some of the couplings present in the exact posterior

$$\begin{aligned} q(\mathbf{Y},\mathbf{S},\mathbf{A}) &= q(\mathbf{Y}) q(\mathbf{S}) q(\mathbf{A}) \\ &= \big(\prod_{ki} q(Y_{r=1:R,ki}^{\tau=1:T})\big)\big(\prod_{ki} q(S_{ki}^{\tau=1:T})\big)\big(\prod_{ri} q(A_{ri})\big) \equiv \prod_{\alpha\in\mathscr{C}} q_\alpha, \end{aligned} \tag{5.12}$$

where $\alpha \in \mathscr{C} = \{\{\mathbf{Y}\},\{\mathbf{S}\},\{\mathbf{A}\}\}$. Hence, attaining the exact marginal likelihood $\mathscr{L}_{\mathbf{X}}(\theta)$ is not guaranteed. Yet, the bound property is preserved and the aim of VB is to optimize the bound. Although it is not possible to obtain the best q distribution respecting the factorization in closed form, a local optimum can be attained by the following fixed point iteration:

$$q_\alpha^{(n+1)} \propto \exp\big(\langle \log p(\mathbf{X},\mathbf{Y},\mathbf{S},\mathbf{A}|\theta) \rangle_{q_{\neg\alpha}^{(n)}}\big), \tag{5.13}$$

where $q_{\neg\alpha} = q/q_\alpha$. This iteration monotonically improves each factor of the q distribution, i.e. $\mathscr{B}[q^{(n)}] \leq \mathscr{B}[q^{(n+1)}]$ for $n = 1,2,\cdots$ given an initialization $q^{(0)}$. The order of the factors is not important for the convergence and the blocks can be visited in arbitrary order. On the other hand, generally the attained fixed point depends upon the order of the updates as well as the starting point $q^{(0)}(.)$. The following update order is chosen in our derivations,

$$q(\mathbf{Y})^{(n+1)} \propto \exp\left(\langle \log p(\mathbf{X},\mathbf{Y},\mathbf{S},\mathbf{A}|\theta)\rangle_{q(\mathbf{S})^{(n)}q(\mathbf{A})^{(n)}}\right) \quad \textbf{(5.14)}$$

$$q(\mathbf{S})^{(n+1)} \propto \exp\left(\langle \log p(\mathbf{X},\mathbf{Y},\mathbf{S},\mathbf{A}|\theta)\rangle_{q(\mathbf{Y})^{(n+1)}q(\mathbf{A})^{(n)}}\right) \quad \textbf{(5.15)}$$

$$q(\mathbf{A})^{(n+1)} \propto \exp\left(\langle \log p(\mathbf{X},\mathbf{Y},\mathbf{S},\mathbf{A}|\theta)\rangle_{q(\mathbf{Y})^{(n+1)}q(\mathbf{S})^{(n+1)}}\right). \quad \textbf{(5.16)}$$

As it is denoted in **(5.14)**, first, the factor $q(\mathbf{Y})^{(n+1)}$ at the $(n+1)$th iteration is updated by fixing the factors $q(\mathbf{S})^{(n)}$ and $q(\mathbf{A})^{(n)}$ obtained at the (n)th iteration. In **(5.15)**, $q(\mathbf{S})^{(n+1)}$ is updated by fixing $q(\mathbf{Y})^{(n+1)}$ obtained at the $(n+1)$th iteration and $q(\mathbf{A})^{(n)}$ obtained at the (n)th iteration. Finally, $q(\mathbf{A})^{(n+1)}$ is updated by fixing $q(\mathbf{Y})^{(n+1)}$ and $q(\mathbf{S})^{(n+1)}$ obtained at the $(n+1)$th iteration as denoted in **(5.16)**.

5.2.1 Variational Update Equations and Sufficient Statistics

The expectations of $\langle \log p(\mathbf{X},\mathbf{Y},\mathbf{S},\mathbf{A}|\theta)\rangle$ are functions of the sufficient statistics of q. The expressions and the derivations are presented in the following. First, the full joint distribution is expressed as:

$$\Phi = p(\mathbf{X},\mathbf{Y},\mathbf{S},\mathbf{A}|\theta) = p(\mathbf{X}|\mathbf{Y})p(\mathbf{Y}|\mathbf{S},\mathbf{A})p(\mathbf{S}|\theta^S)p(\mathbf{A}|\theta^A). \quad \textbf{(5.17)}$$

Then, the logarithm of the joint distribution can be written as:

$$\begin{aligned}
\log\Phi &= \log p(\mathbf{X}|\mathbf{Y}) + \log p(\mathbf{Y}|\mathbf{S},\mathbf{A}) + \log p(\mathbf{S}|\theta^S) + \log p(\mathbf{A}|\theta^A) \\
&= \sum_k\sum_i \log\delta(X_{ki} - \sum_r\sum_\tau Y^\tau_{rki}) \\
&\quad + \sum_k\sum_i\sum_r\sum_\tau \left(-S^\tau_{kr}A_{r,i-\tau} + Y^\tau_{rki}\log(S^\tau_{kr}A_{r,i-\tau}) - \log\Gamma(Y^\tau_{rki}+1)\right) \\
&\quad + \sum_k\sum_r\sum_\tau (a^{\tau S}_{kr}-1)\log S^\tau_{kr} - \frac{a^{\tau S}_{kr}}{b^{\tau S}_{kr}}S^\tau_{kr} - \log\Gamma(a^{\tau S}_{kr}) - a^{\tau S}_{kr}\log\frac{b^{\tau S}_{kr}}{a^{\tau S}_{kr}} \\
&\quad + \sum_r\sum_i (a^A_{ri}-1)\log A_{ri} - \frac{a^A_{ri}}{b^A_{ri}}A_{ri} - \log\Gamma(a^A_{ri}) - a^A_{ri}\log\frac{b^A_{ri}}{a^A_{ri}}. \quad \textbf{(5.18)}
\end{aligned}$$

The update equation for the latent sources Y^{τ}_{kri} can be written as:

$$\begin{aligned} q(Y^{\tau=1:T}_{r=1:R,ki}) &\propto \exp\left(\sum_r\sum_\tau \left(Y^{\tau}_{rki}(\langle\log(S^{\tau}_{kr}\rangle + \langle\log A_{r,i-\tau}\rangle)\right) - \log\Gamma(Y^{\tau}_{rki}+1))\right) \\ &\quad \delta(X_{ki} - \sum_r\sum_\tau Y^{\tau}_{rki}) \\ &\quad \propto \mathscr{M}(Y^{\tau=1:T}_{r=1:R,ki}; X_{ki}, p^{\tau=1:T}_{k,1:R,i}), \end{aligned} \tag{5.19}$$

where

$$p^{\tau}_{rki} = \exp(\langle\log S^{\tau}_{kr}\rangle + \langle\log A_{r,i-\tau}\rangle) / \sum_r\sum_\tau \exp(\langle\log S^{\tau}_{kr}\rangle + \langle\log A_{r,i-\tau}\rangle) \tag{5.20}$$

$$\langle Y^{\tau}_{rki}\rangle = X_{ki} p^{\tau}_{rki}. \tag{5.21}$$

The update equation for the latent sources $\mathbf{Y}$ leads to the following:

$$q(Y^{\tau=1:T}_{r=1:R,ki}) \propto \mathscr{M}(Y^{\tau=1:T}_{r=1:R,ki}; X_{ki}, p^{\tau=1:T}_{r=1:R,ki}) \tag{5.22}$$

where

$$\begin{aligned} p^{\tau}_{rki} &= \frac{\exp(\langle\log S^{\tau}_{kr}\rangle + \langle\log A_{r,i-\tau}\rangle)}{\sum_r\sum_\tau \exp(\langle\log S^{\tau}_{kr}\rangle + \langle\log A_{r,i-\tau}\rangle)} \\ \langle Y^{\tau}_{rki}\rangle &= X_{ki} p^{\tau}_{rki}. \end{aligned} \tag{5.23}$$

The update equation for the bases S^{τ}_{kr} can be written as:

$$\begin{aligned} q(S^{\tau}_{kr}) &\propto \exp\left((a^{\tau S}_{kr} + \sum_i \langle Y^{\tau}_{rki}\rangle - 1)\log S^{\tau}_{kr} - (\frac{a^{\tau S}_{kr}}{b^{\tau S}_{kr}} + \sum_i \langle A_{r,i-\tau}\rangle) S^{\tau}_{kr}\right) \\ &\propto \mathscr{G}(S^{\tau}_{kr}; \nu^{\tau S}_{kr}, \beta^{\tau S}_{kr}), \end{aligned} \tag{5.24}$$

where

$$\begin{aligned} \nu^{\tau S}_{kr} &\equiv a^{\tau S}_{kr} + \sum_i \langle Y^{\tau}_{rki}\rangle \\ \beta^{\tau S}_{kr} &\equiv \left(\frac{a^{\tau S}_{kr}}{b^{\tau S}_{kr}} + \sum_i \langle A_{r,i-\tau}\rangle\right)^{-1}. \end{aligned} \tag{5.25}$$

Then,

$$\begin{aligned} \exp(\langle\log S^{\tau}_{kr}\rangle) &= \exp\left(\Psi(\nu^{\tau S}_{kr})\right)\beta^{\tau S}_{kr} \\ \langle S^{\tau}_{kr}\rangle &= \nu^{\tau S}_{kr}\beta^{\tau S}_{kr}. \end{aligned} \tag{5.26}$$

In (**5.26**), Ψ denotes the digamma function defined as $\Psi(a) \equiv d\log\Gamma(a)/da$.

After assigning $m = i - \tau$, the update equation for the gains A_{rm} can be written as:

$$q(A_{rm}) \propto \exp\left((a^A_{rm} + \sum_\tau \sum_k \langle Y^\tau_{rk,m+\tau} \rangle - 1) \log A_{rm} - (\frac{a^A_{rm}}{b^A_{rm}} + \sum_\tau \sum_k \langle S^\tau_{kr} \rangle) A^\tau_{rm} \right)$$
$$\propto \mathcal{G}(A_{rm}; v^A_{rm}, \beta^A_{rm}), \tag{5.27}$$

where

$$v^A_{rm} \equiv a^A_{rm} + \sum_\tau \sum_k \langle Y^\tau_{rk,m+\tau} \rangle$$
$$\beta^A_{rm} \equiv \left(\frac{a^A_{rm}}{b^A_{rm}} + \sum_\tau \sum_k \langle S^\tau_{kr} \rangle \right)^{-1}. \tag{5.28}$$

Then,

$$\exp(\langle \log A_{rm} \rangle) = \exp\left(\Psi(v^A_{rm})\right) \beta^A_{rm}$$
$$\langle A_{rm} \rangle = v^A_{rm} \beta^A_{rm}. \tag{5.29}$$

One of the attractive features of NMFD is easy and efficient implementation. In this section, the update equations of [9] are derived in compact matrix notation to illustrate that these attractive properties are retained for the full Bayesian treatment. A subtle but key point in the efficiency of the algorithm is that explicitly storing and computing the $T \times R \times I$ object $\mathbf{Y}^\tau$ is avoided, as only the marginal statistics are needed during optimization.

5.2.2 The Variational Bound

As it is denoted in Section 5.2, the marginal likelihood cannot be obtained in closed form. The variational bound is an adequate approximation to the marginal likelihood and can be written as

$$\mathcal{B}_{VB} = \langle \log \Phi \rangle_q + \mathcal{A}[q], \tag{5.30}$$

where the energy term is given by the expectation of the expression in (**5.18**) as

$$\langle \log \Phi \rangle_q = \sum_k \sum_i \log \delta(X_{ki} - \sum_r \sum_\tau \langle Y^\tau_{rki} \rangle)$$
$$+ \sum_k \sum_i \sum_r \sum_\tau \left(- \langle S^\tau_{kr} \rangle \langle A_{r,i-\tau} \rangle + \langle Y^\tau_{rki} \rangle \langle \log(S^\tau_{kr} A_{r,i-\tau}) \rangle - \langle \log \Gamma(Y^\tau_{rki} + 1) \rangle \right)$$
$$+ \sum_k \sum_r \sum_\tau (a^{\tau S}_{kr} - 1) \langle \log S^\tau_{kr} \rangle - \frac{a^{\tau S}_{kr}}{b^{\tau S}_{kr}} \langle S^\tau_{kr} \rangle - \log \Gamma(a^{\tau S}_{kr}) - a^{\tau S}_{kr} \log \frac{b^{\tau S}_{kr}}{a^{\tau S}_{kr}}$$
$$+ \sum_r \sum_i (a^A_{ri} - 1) \langle \log A_{ri} \rangle - \frac{a^A_{ri}}{b^A_{ri}} \langle A_{ri} \rangle - \log \Gamma(a^A_{ri}) - a^A_{ri} \log \frac{b^A_{ri}}{a^A_{ri}}, \tag{5.31}$$

and $\mathscr{A}[q]$ denotes the entropy of the variational approximation distribution q:

$$\mathscr{A}[q] = -\langle \log q \rangle = \sum_k \sum_i A_{\mathscr{M}}[Y^{\tau=1:T}_{r=1:R,ki}] + \sum_k \sum_r A_{\mathscr{G}}[S^{\tau=1:T}_{kr}] + \sum_r \sum_i A_{\mathscr{G}}[A_{ri}]. \quad \textbf{(5.32)}$$

The individual entropies are defined as:

$$\begin{aligned}
A_{\mathscr{M}}[Y^{\tau=1:T}_{r=1:R,ki}] &= -\log\Gamma(X_{ki}+1) - \sum_\tau \sum_r \langle Y^\tau_{rki}\rangle \log p^\tau_{rki} + \sum_\tau \sum_r \langle \log\Gamma(Y^\tau_{rki}+1)\rangle \\
&\quad - \langle \log\delta(X_{ki} - \sum_\tau \sum_r Y^\tau_{rki})\rangle \\
&= -\log\Gamma(X_{ki}+1) + \sum_\tau \sum_r \langle \log\Gamma(Y^\tau_{rki}+1)\rangle - \langle \log\delta(X_{ki} - \sum_\tau \sum_r Y^\tau_{rki})\rangle \\
&\quad - \sum_\tau \sum_r \langle Y^\tau_{rki}\rangle \big((\langle \log S^\tau_{kr}\rangle + \langle \log A_{r,i-\tau}\rangle) - \sum_r \sum_\tau (\langle \log S^\tau_{kr}\rangle + \langle \log A_{r,i-\tau}\rangle)\big) \\
A_{\mathscr{G}}[S^\tau_{kr}] &= -(\nu^{\tau S}_{kr} - 1)\Psi(\nu^{\tau S}_{kr}) + \log\beta^{\tau S}_{kr} + \nu^{\tau S}_{kr} + \log\Gamma(\nu^{\tau S}_{kr}) \\
A_{\mathscr{G}}[A_{ri}] &= -(\nu^A_{ri} - 1)\Psi(\nu^A_{ri}) + \log\beta^A_{ri} + \nu^A_{ri} + \log\Gamma(\nu_{ri}A).
\end{aligned}$$

If the individual entropies which are expressed above are written into the bound equation given in (**5.30**), the bound takes the form:

$$\begin{aligned}
\mathscr{B}_{VB} = &- \sum_k \sum_i \sum_r \sum_\tau \langle S^\tau_{kr}\rangle \langle A_{r,i-\tau}\rangle \\
&+ \sum_k \sum_r \sum_\tau \langle \log S^\tau_{kr}\rangle \left(a^{\tau S}_{kr} - 1 + \sum_i \langle Y^\tau_{rki}\rangle \right) \\
&+ \sum_r \sum_i \langle \log A_{ri}\rangle \left(a^A_{ri} - 1 + \sum_k \sum_\tau \langle Y^\tau_{rki}\rangle \right) \\
&+ \sum_k \sum_r \sum_\tau -\frac{a^{\tau S}_{kr}}{b^{\tau S}_{kr}} \langle S^\tau_{kr}\rangle - \log\Gamma(a^{\tau S}_{kr}) - a^{\tau S}_{kr} \log \frac{b^{\tau S}_{kr}}{a^{\tau S}_{kr}} \\
&+ \sum_r \sum_i -\frac{a^A_{ri}}{b^A_{ri}} \langle A_{ri}\rangle - \log\Gamma(a^A_{ri}) - a^A_{ri} \log \frac{b^A_{ri}}{a^A_{ri}} \\
&+ \sum_k \sum_i \left(-\log\Gamma(X_{ki}+1) - \sum_\tau \sum_r \langle Y^\tau_{rki}\rangle \log p^\tau_{rki} \right) \\
&+ \sum_k \sum_r \sum_\tau \left(-(\nu^{\tau S}_{kr} - 1)\Psi(\nu^{\tau S}_{kr}) + \log\beta^{\tau S}_{kr} + \nu^{\tau S}_{kr} + \log\Gamma(\nu^{\tau S}_{kr}) \right) \\
&+ \sum_r \sum_i \left(-(\nu^A_{ri} - 1)\Psi(\nu^A_{ri}) + \log\beta^A_{ri} + \nu^A_{ri} + \log\Gamma(\nu^A_{ri}) \right). \quad \textbf{(5.33)}
\end{aligned}$$

The required expectations are obtained as:

$$\begin{aligned}
\sum_\tau \sum_r \langle Y^\tau_{rki}\rangle \log p^\tau_{rki} &= X_{ki} \left(\sum_\tau \sum_r p^\tau_{rki} \log p^\tau_{rki} \right) \\
&= X_{ki} \left(\sum_\tau \sum_r \frac{\exp(\langle \log S^\tau_{kr}\rangle + \langle \log A_{r,i-\tau}\rangle)}{\sum_{r'} \sum_{\tau'} \exp(\langle \log S^{\tau'}_{kr'}\rangle + \langle \log A_{r',i-\tau'}\rangle)} \right.
\end{aligned}$$

$$
\begin{aligned}
&\big(\langle \log S^{\tau}_{kr}\rangle + \langle \log A_{r,i-\tau}\rangle \\
&\quad - \log\big(\sum_{r'}\sum_{\tau'} \exp(\langle \log S^{\tau}_{kr}\rangle + \langle \log A_{r,i-\tau}\rangle)\big)\big)\Big) \\
&= X_{ki}\Bigg(\sum_{\tau}\sum_{r} \frac{\exp(\langle \log S^{\tau}_{kr}\rangle)\exp(\langle \log A_{r,i-\tau}\rangle)}{\sum_{r'}\sum_{\tau'}\exp(\langle \log S^{\tau'}_{kr'}\rangle)\exp(\langle \log A_{r',i-\tau'}\rangle)} \\
&\quad \big(\langle \log S^{\tau}_{kr}\rangle + \langle \log A_{r,i-\tau}\rangle \\
&\quad - \log\big(\sum_{r'}\sum_{\tau'} \exp(\langle \log S^{\tau'}_{kr'}\rangle)\exp(\langle \log A_{r,i-\tau'}\rangle)\big)\big)\Bigg) \\
&= X_{ki}\Bigg(\sum_{\tau}\sum_{r} \frac{\exp(\langle \log S^{\tau}_{kr}\rangle)\exp(\langle \log A_{r,i-\tau}\rangle)}{\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau'}]_{ki}}\langle \log S^{\tau}_{kr}\rangle\Bigg) \\
&\quad + X_{ki}\Bigg(\sum_{\tau}\sum_{r} \frac{\exp(\langle \log S^{\tau}_{kr}\rangle)\exp(\langle \log A_{r,i-\tau}\rangle)}{\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau'}]_{ki}}\langle \log A_{r,i-\tau}\rangle\Bigg) \\
&\quad - X_{ki}\Bigg(\sum_{\tau}\sum_{r} \frac{\exp(\langle \log S^{\tau}_{kr}\rangle)\exp(\langle \log A_{r,i-\tau}\rangle)}{\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau'}]_{ki}}\log\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau'}]_{ki}\big)\Bigg) \\
&= X_{ki}\Bigg(\sum_{\tau}\sum_{r} \frac{\exp(\langle \log S^{\tau}_{kr}\rangle)\exp(\langle \log A_{r,i-\tau}\rangle)}{\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau'}]_{ki}}\langle \log S^{\tau}_{kr}\rangle\Bigg) \\
&\quad + X_{ki}\Bigg(\sum_{\tau}\sum_{r} \frac{\exp(\langle \log S^{\tau}_{kr}\rangle)\exp(\langle \log A_{r,i-\tau}\rangle)}{\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau'}]_{ki}}\langle \log A_{r,i-\tau}\rangle\Bigg) \\
&\quad - X_{ki}\Bigg(\sum_{\tau} \frac{[\mathbf{L}_S^{\tau}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau}]_{ki}}{\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau'}]_{ki}}\log\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau'}]_{ki}\big)\Bigg) \\
&= X_{ki}\Bigg(\sum_{\tau}\sum_{r} \frac{\exp(\langle \log S^{\tau}_{kr}\rangle)\exp(\langle \log A_{r,i-\tau}\rangle)}{\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau'}]_{ki}}\langle \log S^{\tau}_{kr}\rangle\Bigg) \\
&\quad + X_{ki}\Bigg(\sum_{\tau}\sum_{r} \frac{\exp(\langle \log S^{\tau}_{kr}\rangle)\exp(\langle \log A_{r,i-\tau}\rangle)}{\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau'}]_{ki}}\langle \log A_{r,i-\tau}\rangle\Bigg) \\
&\quad - X_{ki}\log\big(\sum_{\tau}[\mathbf{L}_S^{\tau}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau}]_{ki}\big). \qquad \textbf{(5.34)}
\end{aligned}
$$

If summation is performed over the k and i dimensions:

$$
\begin{aligned}
\sum_{k}\sum_{i}\sum_{\tau}\sum_{r}\langle Y^{\tau}_{rki}\rangle \log p^{\tau}_{rki} &= \sum_{k}\sum_{i}\mathbf{X}.*\Bigg(\sum_{\tau}\big((\mathbf{L}_S^{\tau}.*\log(\mathbf{L}_S^{\tau}))\overrightarrow{\mathbf{L}_A}^{\tau}\big)./\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau'}]\big)\Bigg) \\
&\quad + \sum_{k}\sum_{i}\mathbf{X}.*\Bigg(\sum_{\tau}\big(\mathbf{L}_S^{\tau}(\overrightarrow{\mathbf{L}_A}^{\tau}.*\log(\overrightarrow{\mathbf{L}_A}^{\tau}))\big)./\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau'}]\big)\Bigg) \\
&\quad - \sum_{k}\sum_{i}\mathbf{X}.*\Bigg(\sum_{\tau}(\mathbf{L}_S^{\tau}\overrightarrow{\mathbf{L}_A}^{\tau}).*\log\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau'}]\big)./\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_{\mathbf{A}}}^{\tau'}]\big)\Bigg) \\
&= \sum_{k}\sum_{i}\mathbf{X}.*\Bigg(\sum_{\tau}\Bigg(\big((\mathbf{L}_S^{\tau}.*\log(\mathbf{L}_S^{\tau}))\overrightarrow{\mathbf{L}_A}^{\tau}\big) + \big(\mathbf{L}_S^{\tau}(\overrightarrow{\mathbf{L}_A}^{\tau}.*\log(\overrightarrow{\mathbf{L}_A}^{\tau}))\big)
\end{aligned}
$$

$$
\begin{aligned}
&\quad - (\mathbf{L}_S^{\tau}\overrightarrow{\mathbf{L}_A^{\tau}}).*\log\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_\mathbf{A}^{\tau'}}]\big)\Big)./\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_\mathbf{A}^{\tau'}}]\big)\Big) \\
&= \sum_k\sum_i \mathbf{X}.*\Bigg(\sum_\tau\big(\mathbf{L}_S^{\tau}.*\log(\mathbf{L}_S^{\tau})\overrightarrow{\mathbf{L}_A^{\tau}} + \mathbf{L}_S^{\tau}(\overrightarrow{\mathbf{L}_A^{\tau}}.*\log(\overrightarrow{\mathbf{L}_A^{\tau}}))\big) \\
&\quad ./\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_\mathbf{A}^{\tau'}}]\big) - \log\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_\mathbf{A}^{\tau'}}]\big)\Bigg). \qquad \textbf{(5.35)}
\end{aligned}
$$

If these expectations are written into the variational bound expression:

$$
\begin{aligned}
\mathscr{B}_{VB} = &-\sum_k\sum_i\sum_\tau \mathbf{E}_S^{\tau}\overrightarrow{\mathbf{E}_A^{\tau}} \\
&+\sum_k\sum_r\sum_\tau(\log\mathbf{L}_S^{\tau}).*(\nu_S^{\tau}-1) \\
&+\sum_r\sum_i(\log\mathbf{L}_A).*(\nu_A-1) \\
&+\sum_k\sum_r\sum_\tau -(\mathbf{a}_S^{\tau}./\mathbf{b}_S^{\tau}).*\mathbf{E}_S^{\tau} - \log\Gamma(\mathbf{a}_S^{\tau}) + \mathbf{a}_S^{\tau}.*\log(\mathbf{a}_S^{\tau}./\mathbf{b}_S^{\tau}) \\
&+\sum_r\sum_i -(\mathbf{a}_A./\mathbf{b}_A).*\mathbf{E}_A - \log\Gamma(\mathbf{a}_A) + \mathbf{a}_A.*\log(\mathbf{a}_A./\mathbf{b}_A) \\
&+\sum_k\sum_i -\log\Gamma(\mathbf{X}+1) \\
&-\sum_k\sum_i \mathbf{X}.*\Bigg(\sum_\tau\big((\mathbf{L}_S^{\tau}.*\log(\mathbf{L}_S^{\tau}))\overrightarrow{\mathbf{L}_A^{\tau}} + \mathbf{L}_S^{\tau}(\overrightarrow{\mathbf{L}_A^{\tau}}.*\log(\overrightarrow{\mathbf{L}_A^{\tau}}))\big) \\
&\quad ./\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_\mathbf{A}^{\tau'}}]\big) - \log\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_\mathbf{A}^{\tau'}}]\big)\Bigg) \\
&+\sum_k\sum_r\sum_\tau\Big(-(\nu_S^{\tau}-1).*\Psi(\nu_S^{\tau}) + \log\beta_S^{\tau} + \nu_S^{\tau} + \log\Gamma(\nu_S^{\tau})\Big) \\
&+\sum_r\sum_i\Big(-(\nu_A-1).*\Psi(\nu_A) + \log\beta_A + \nu_A + \log\Gamma(\nu_A)\Big). \qquad \textbf{(5.36)}
\end{aligned}
$$

After replacing the $\Psi(\nu_S^{\tau})$ and $\Psi(\nu_A)$ with $\log(\mathbf{L}_S^{\tau})$ and $\log(\mathbf{L}_A)$:

$$
\begin{aligned}
\mathscr{B}_{VB} = &-\sum_k\sum_i\sum_\tau \mathbf{E}_S^{\tau}\overrightarrow{\mathbf{E}_A^{\tau}} \\
&+\sum_k\sum_r\sum_\tau(\log\mathbf{L}_S^{\tau}).*(\nu_S^{\tau}-1) \\
&+\sum_r\sum_i(\log\mathbf{L}_A).*(\nu_A-1) \\
&+\sum_k\sum_r\sum_\tau -(\mathbf{A}_S^{\tau}./\mathbf{B}_S^{\tau}).*\mathbf{E}_S^{\tau} - \log\Gamma(\mathbf{A}_S^{\tau}) + \mathbf{A}_S^{\tau}.*\log(\mathbf{A}_S^{\tau}./\mathbf{B}_S^{\tau}) \\
&+\sum_r\sum_i -(\mathbf{A}_A./\mathbf{B}_A).*\mathbf{E}_A - \log\Gamma(\mathbf{A}_A) + \mathbf{A}_A.*\log(\mathbf{A}_A./\mathbf{B}_A) \\
&+\sum_k\sum_i -\log\Gamma(\mathbf{X}+1) \\
&-\sum_k\sum_i \mathbf{X}.*\Bigg(\sum_\tau\big((\mathbf{L}_S^{\tau}.*\log(\mathbf{L}_S^{\tau}))\overrightarrow{\mathbf{L}_A^{\tau}} + \mathbf{L}_S^{\tau}(\overrightarrow{\mathbf{L}_A^{\tau}}.*\log(\overrightarrow{\mathbf{L}_A^{\tau}}))\big)
\end{aligned}
$$

$$
\begin{aligned}
&./\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_\mathbf{A}^{\tau'}}]\big)-\log\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_\mathbf{A}^{\tau'}}]\big)\Big)\\
&+\sum_k\sum_r\sum_\tau\Big(-(\nu_S^\tau-1).*\big(\log(\mathbf{L}_S^\tau)-\log(\beta_S^\tau)\big)+\log\beta_S^\tau+\nu_S^\tau+\log\Gamma(\nu_S^\tau)\Big)\\
&+\sum_r\sum_i\Big(-(\nu_A-1).*\big(\log(\mathbf{L}_A)-\log(\beta_A)\big)+\log\beta_A+\nu_A+\log\Gamma(\nu_A)\Big).
\end{aligned}
\tag{5.37}
$$

After removing the terms which cancel each other, the variational bound is obtained as:

$$
\begin{aligned}
\mathscr{B}_{VB}=&\sum_k\sum_i\Big((\sum_\tau-\mathbf{E}_S^\tau\overrightarrow{\mathbf{E}_A^\tau})-\log\Gamma(\mathbf{X}+1)\Big)\\
&+\sum_k\sum_r\sum_\tau-(\mathbf{A}_S^\tau./\mathbf{B}_S^\tau).*\mathbf{E}_S^\tau-\log\Gamma(\mathbf{A}_S^\tau)+\mathbf{A}_S^\tau.*\log(\mathbf{A}_S^\tau./\mathbf{B}_S^\tau)\\
&+\sum_r\sum_i-(\mathbf{A}_A./\mathbf{B}_A).*\mathbf{E}_A-\log\Gamma(\mathbf{A}_A)+\mathbf{A}_A.*\log(\mathbf{A}_A./\mathbf{B}_A)\\
&-\sum_k\sum_i\mathbf{X}.*\Big(\sum_\tau\big((\mathbf{L}_S^\tau.*\log(\mathbf{L}_S^\tau))\overrightarrow{\mathbf{L}_A^\tau}+\mathbf{L}_S^\tau(\overrightarrow{\mathbf{L}_A^\tau}.*\log(\overrightarrow{\mathbf{L}_A^\tau}))\big)\\
&\quad./\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_\mathbf{A}^{\tau'}}]\big)-\log\big(\sum_{\tau'}[\mathbf{L}_S^{\tau'}\overrightarrow{\mathbf{L}_\mathbf{A}^{\tau'}}]\big)\Big)\\
&+\sum_k\sum_r\sum_\tau\nu_S^\tau.*\big(\log\beta_S^\tau+1\big)+\log\Gamma(\nu_S^\tau)\\
&+\sum_r\sum_i\nu_A.*\big(\log\beta_A+1\big)+\log\Gamma(\nu_A).
\end{aligned}
\tag{5.38}
$$

5.2.3 Hyperparameter Estimation

The hierarchical model given by **(5.7)** allows to estimate the hyperparameters $\theta=(\theta^S,\theta^A)$ from the examples of a certain class of source to capture the invariant features. Thus, $\theta=(\theta^S,\theta^A)$ can be estimated by maximizing the bound given in **(5.30)** with respect to $a_{kr}^{\tau S}$

$$
\begin{aligned}
\frac{\partial\mathscr{B}_{VB}}{\partial a_{kr}^{\tau S}}&=\langle\log S_{kr}^\tau\rangle-\frac{1}{b_{kr}^{\tau S}}\langle S_{kr}^\tau\rangle-\Psi(a_{kr}^{\tau S})-\log b_{kr}^{\tau S}+\log a_{kr}^{\tau S}+1=0\\
c_{kr}^\tau&=\log a_{kr}^{\tau S}-\Psi(a_{kr}^{\tau S})+1=0\\
c_{kr}^\tau&\equiv\frac{\langle S_{kr}^\tau\rangle}{b_{kr}^{\tau S}}-\big(\langle\log S_{kr}^\tau\rangle-\log b_{kr}^{\tau S}\big),
\end{aligned}
$$

and with respect to $b_{kr}^{\tau S}$

$$
\frac{\partial\mathscr{B}_{VB}}{\partial b_{kr}^{\tau S}}=\frac{a_{kr}^{\tau S}}{(b_{kr}^{\tau S})^2}\langle S_{kr}^\tau\rangle-a_{kr}^{\tau S}\frac{1}{b_{kr}^{\tau S}}=0
$$

$$b_{kr}^{\tau S} = \langle S_{kr}^{\tau} \rangle.$$

Tying parameters across k as $a_r^{\tau S} = a_{kr}^{\tau S}$ and $b_r^{\tau S} = b_{kr}^{\tau S}$ yields

$$\begin{aligned}
\frac{\partial \mathscr{B}_{VB}}{\partial a_r^{\tau S}} &= \sum_k \langle \log S_{kr}^{\tau} \rangle - \sum_k \frac{1}{b_{kr}^{\tau S}} \langle S_{kr}^{\tau} \rangle - K\Psi(a_{kr}^{\tau S}) - \sum_k \log b_{kr}^{\tau S} + \log a_{kr}^{\tau S} + K = 0 \\
c_r^{\tau} &= \log a_{kr}^{\tau S} - \Psi(a_{kr}^{\tau S}) + 1 = 0 \\
c_{kr}^{\tau} &\equiv \frac{\langle S_{kr}^{\tau} \rangle}{b_{kr}^{\tau S}} - \big(\langle \log S_{kr}^{\tau} \rangle - \log b_{kr}^{\tau S} \big).
\end{aligned}$$

Finally,

$$\begin{aligned}
\frac{\partial \mathscr{B}_{VB}}{\partial b_{kr}^{\tau S}} &= \frac{a_{kr}^{\tau S}}{(b_{kr}^{\tau S})^2} \langle S_{kr}^{\tau} \rangle - a_{kr}^{\tau S} \frac{1}{b_{kr}^{\tau S}} = 0 \\
b_{kr}^{\tau S} &= \langle S_{kr}^{\tau} \rangle.
\end{aligned}$$

5.2.4 Efficient Implementation

In this section, the update equations of Section 5.2.1 are derived in compact matrix notation in order to illustrate that the attractive properties of NMFD are satisfied with a full Bayesian approach. An important point which proves the efficiency of the proposed algorithm is that the storing and computing of the latent sources $< \mathbf{S} >$ which are of length $R \times K \times I$ are avoided since only the marginal statistics are required during optimization.

Considering (**5.23**), the expected value of the latent sources can be written as

$$\sum_i \langle Y_{rki}^{\tau} \rangle = \sum_i X_{ki} p_{rki}^{\tau} = \exp(\langle \log S_{kr}^{\tau} \rangle) * \sum_i \big(X_{ki} / \langle \Lambda_{ki} \rangle \big) \exp(\langle \log A_{r,i-\tau} \rangle) \tag{5.39}$$

$$\mathbf{\Sigma}_S^{\tau} = \mathbf{L}_S^{\tau} . * \left(\big(\mathbf{X} ./ \mathbf{L}_{\Lambda} \big) \overrightarrow{\mathbf{L}_A^{\tau}}^{\top} \right) \tag{5.40}$$

$$\begin{aligned}
\sum_{\tau} \sum_k \langle Y_{r,k,i+\tau}^{\tau} \rangle &= \sum_{\tau} \sum_k X_{k,i+\tau} p_{r,k,i+\tau}^{\tau} \\
&= \exp(\langle \log A_{ri} \rangle) * \sum_{\tau} \sum_k \left(X_{k,i+\tau} / \Lambda_{k,i+\tau} \right) \exp(\langle \log S_{kr}^{\tau} \rangle)
\end{aligned} \tag{5.41}$$

$$\mathbf{\Sigma}_A = \mathbf{L}_A . * \sum_{\tau} \left(\mathbf{L}_S^{\tau \top} \big(\overleftarrow{\mathbf{X}}^{\tau} ./ \overleftarrow{\mathbf{L}_{\Lambda}}^{\tau} \big) \right), \tag{5.42}$$

where

$$\mathbf{L}_{\Lambda} = \sum_{\tau} \mathbf{L}_S^{\tau} \overrightarrow{\mathbf{L}_A^{\tau}} = \sum_{\tau} \sum_d \exp(\langle \log S_{kr}^{\tau} \rangle) \exp(\langle \log A_{r,i-\tau} \rangle). \tag{5.43}$$

In (**5.40**) and (**5.42**), the expression is represented in compact notation where the following matrices are defined as:

$$
\begin{aligned}
\mathbf{E}_S^\tau &= \{\langle S_{kr}^\tau \rangle\}, & \overrightarrow{\mathbf{E}}_A^\tau &= \{\langle A_{r,i-\tau} \rangle\}, \\
\mathbf{L}_S^\tau &= \{\exp(\langle \log S_{kr}^\tau \rangle)\}, & \overrightarrow{\mathbf{L}}_A^\tau &= \{\exp(\langle \log A_{r,i-\tau} \rangle)\}, \\
\Sigma_S^\tau &= \{\sum_i \langle Y_{rki}^\tau \rangle\}, & \Sigma_A &= \{\sum_\tau \sum_k \langle Y_{r,k,i+\tau}^\tau \rangle\}, \\
\mathbf{a}_S^\tau &= \{a_{kr}^{\tau S}\}, & \mathbf{a}_A &= \{a_{ri}^A\}, \\
\mathbf{b}_S^\tau &= \{b_{kr}^{\tau S}\}, & \mathbf{b}_A &= \{b_{ri}^A\}, \\
\nu_S^\tau &= \{\nu_{kr}^{\tau S}\}, & \nu_A &= \{\nu_{ri}^A\}, \\
\beta_S^\tau &= \{\beta_{kr}^{\tau S}\} & \beta_A &= \{\beta_{ri}^A\}
\end{aligned}
\tag{5.44}
$$

For notational convenience, $.*$ and $./$ refer to as elementwise matrix multiplication and division, respectively. After straightforward substitutions, the variational NMFD algorithm is obtained as it can be compactly expressed in Algorithm 7.

5.3 Simulation Results and Discussion

The performance of the proposed method is illustrated in a model selection context. For this purpose, synthetic data is generated from the hierarchical model with the basis matrix $\mathbf{S} \in \mathbb{R}^{5\times 32\times 2}$ ($R = 5, K = 32, T = 2$) and the gain matrix $\mathbf{A} \in \mathbb{R}^{5\times 20}$($R = 5, I = 20$). Two inference tasks are performed. First one is to find the correct number of sources R, given $\mathbf{X}$ and T; second is to find the convolution parameter T, given $\mathbf{X}$ and R. The hyperparameters of the true model are set to at $a_{kr}^{\tau S} = a^S = 10, b_{ki}^{\tau S} = b^S = 1, a_{ri}^A = a^A = 1, b_{ri}^A = b^A = 100$. In the experiment, the hyperparameters are jointly estimated from data, using hyperparameter adaptation. The marginal likelihood for models is evaluated with the number of bases $R = 1\cdots 10$, convolution order $T = 0\cdots 10$ and a variational lower bound $\mathscr{B}_{VB}$ via variational Bayes. The variational algorithm is run until convergence of the bound or for MAXITER = 20000 steps, whichever occurs first. In Figure 5.1 a (b), the lower bound is shown as the average of five independent runs as a function of model order R (T), where for each R (T), the bound is optimized independently by jointly optimizing hyperparameters a^S, b^S, a^A, b^A in (**5.7**) using the hyperparameter update equations. It is observed that, the correct model order can be

Algorithm 7 Variational Nonnegative Matrix Factor Deconvolution

1: Initialize :

$$\mathbf{L}_S{}^{(0)} = \mathbf{E}_S{}^{(0)} \sim \mathscr{G}(\cdot; \mathbf{a}_S, \mathbf{b}_S./\mathbf{a}_S)$$
$$\mathbf{L}_A{}^{(0)} = \mathbf{E}_A{}^{(0)} \sim \mathscr{G}(\cdot; \mathbf{a}_A, \mathbf{b}_A./\mathbf{a}_A)$$
$$\mathbf{L}_\Lambda^{(0)} = \sum_\tau \mathbf{L}_S^{\tau(0)} \overrightarrow{\mathbf{L}_A}^{\tau(0)}$$

2: for $n = 1 \cdots$ MAXITER do

3: Source sufficient statistics

$$\Sigma_S^{\tau(n)} := \mathbf{L}_S^{\tau(n-1)} .* \left(\left(\mathbf{X}./\mathbf{L}_\Lambda^{(n-1)} \right) \overrightarrow{\mathbf{L}_A}^{\tau(n-1)^\top} \right)$$
$$\Sigma_A^{(n)} = \mathbf{L}_A .* \sum_\tau \left(\mathbf{L}_S^{\tau(n-1)^\top} \left(\overleftarrow{\mathbf{X}}^\tau ./ \overleftarrow{\mathbf{L}}_\Lambda^{\tau(n-1)} \right) \right)$$

4: Means

$$\mathbf{E}_S^{\tau(n)} := \nu_S{}^{\tau(n)} .* \beta_S{}^{\tau(n)}, \quad \nu_S^{\tau(n)} = \mathbf{a}_S^\tau + \Sigma_S^{\tau(n)}, \quad \beta_S^{\tau(n)} = 1./\left(\mathbf{a}_S^\tau./\mathbf{b}_S^\tau + \mathbf{1}\, \overrightarrow{\mathbf{E}_A}^{\tau(n-1)^\top} \right),$$
$$\mathbf{E}_A^{(n)} := \nu_A{}^{(n)} .* \beta_{\mathbf{A}}{}^{(n)}, \quad \nu_A{}^{(n)} = \mathbf{a}_A + \Sigma_A{}^{(n)}, \quad \beta_A{}^{(n)} = \mathbf{1}./\left(\mathbf{a}_A./\mathbf{b}_A + \left(\sum_\tau \mathbf{E}_S^{\tau(n-1)} \right)^\top \mathbf{1} \right)$$

5: Compute Bound

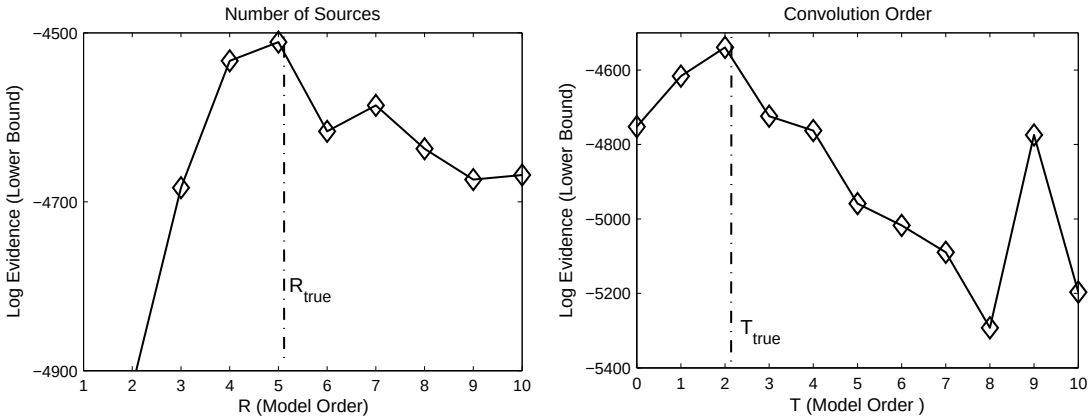
6: Means of Logs

$$\mathbf{L}_S^{\tau(n)} = \exp\left(\Psi(\nu_S^{\tau(n)}) \right) .* \beta_S^{\tau(n)} \quad \mathbf{L}_A{}^{(n)} = \exp\left(\Psi(\nu_A^{(n)}) \right) .* \beta_A^{(n)} \quad \mathbf{L}_\Lambda^{(n)} = \sum_\tau \mathbf{L}_S^{\tau(n)} \overrightarrow{\mathbf{L}_A}^{\tau(n)}$$

7: end for

inferred even when the hyperparameters are unknown a-priori. This is also useful for estimation of model order from real data.

As real data, an audio signal which is obtained by summing piano and trumpet signals is used. The signals are sampled at 11 kHz. All hyperparameters are jointly estimated. In Figure 5.2, results of model order determination for real data are shown with joint hyperparameter adaptation. Here, the variational algorithm is run for each model order $T = 0 \cdots 10$ ($R = 1 \cdots 10$) independently and the lower bound is evaluated after optimizing the hyperparameters. First, $R = 2$ is fixed and the algorithm is run for $T = 0 \cdots 10$. The log evidence is plotted in Figure 5.2 (b). The model order is determined as $T = 0$, at which the maximum lower bound is obtained. Then the convolution order is fixed as $T = 0$ and the algorithm is run for $R = 1 \cdots 10$. In

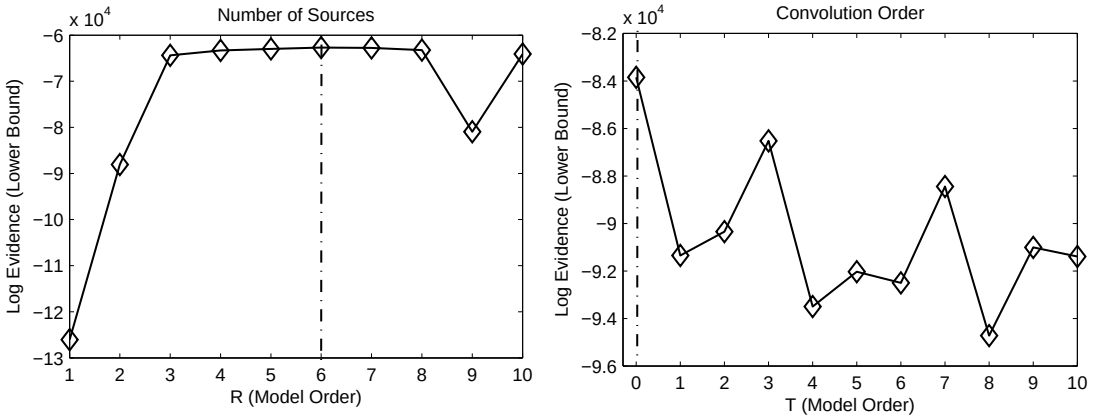


(a) (b)

Figure 5.1: Model selection by variational bound with adapted hyperparameters for synthetic data. (a) Selection of the true number of sources when the convolution order is known. (b) Selection of the true convolution order when the number of sources is is known.

this case, the model order is determined as $R = 6$ and the log evidence is plotted in Figure 5.2(a). The lower bound for the number of sources behaves as is expected from marginal likelihood, reflecting the tradeoff between too many and too few bases.

In order to see the overall performance of the proposed BNMFD method and compare with the conventional NMFD model [9], the separation quality is measured in terms of SDR, SIR, SAR [24], OPS, IPS and APS [25]. Figure 5.3 illustrates the mean SDR, SIR and SAR values obtained by the NMFD and the BNMFD methods for different convolution orders over five runs. As it is seen from the figure, the results obtained by BNMFD fluctuate more for different convolution orders. If the convolution order is set appropriately, BNMFD outperforms NMFD. Similarly, Figure 5.4 illustrates the mean OPS, IPS and APS values obtained by the NMFD and the BNMFD methods for the separated signals. It can be seen from the figure, that the perceptual scores obtained by the BNMFD method fluctuate more for different convolution orders. If the convolution order is set appropriately, BNMFD achieves a similar performance in terms of SIR and outperforms NMFD in terms of the rest of the measures. The improvement obtained by the BNMFD is more significant in terms of the artifacts distortion measured by SAR and APS.



(a) (b)

Figure 5.2: Model selection by variational bound with adapted hyperparameters for real data. (a) Selection of the true number of sources when the convolution order is fixed. (b) Selection of the true convolution order when the number of sources is fixed.

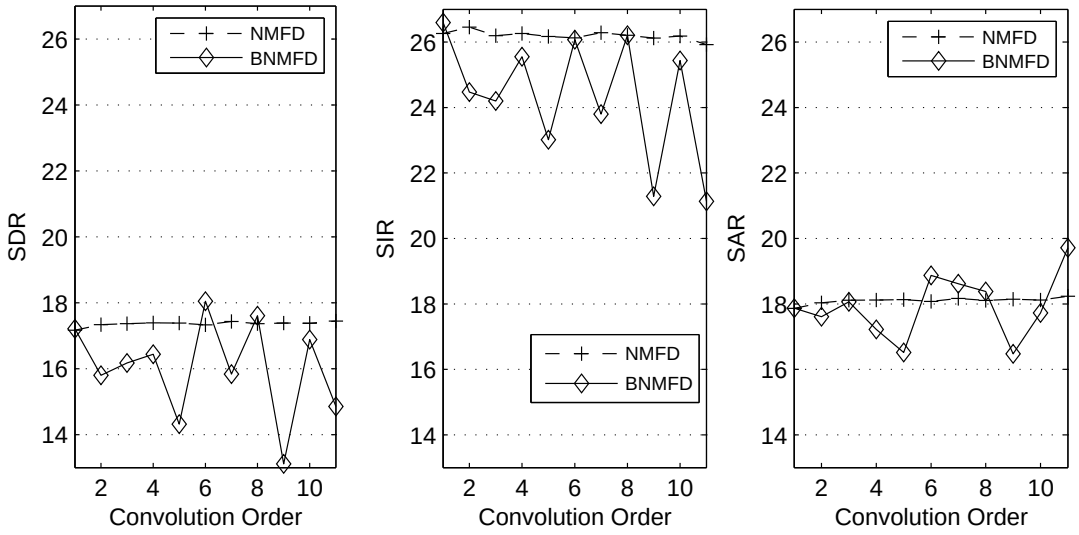


Figure 5.3: The mean SDR, SIR and SAR values obtained by the NMFD and BNMFD methods for different values of convolution order T.

5.4 Conclusion

In this chapter, NMFD has been investigated from a statistical perspective. It has been shown that the KL minimization formulation of the original algorithm can be derived from a probabilistic model where the observations are superposition of J independent Poisson distributed latent sources. Here, the basis and the gain matrices have turned out to be latent intensity parameters. The exact characterization of the

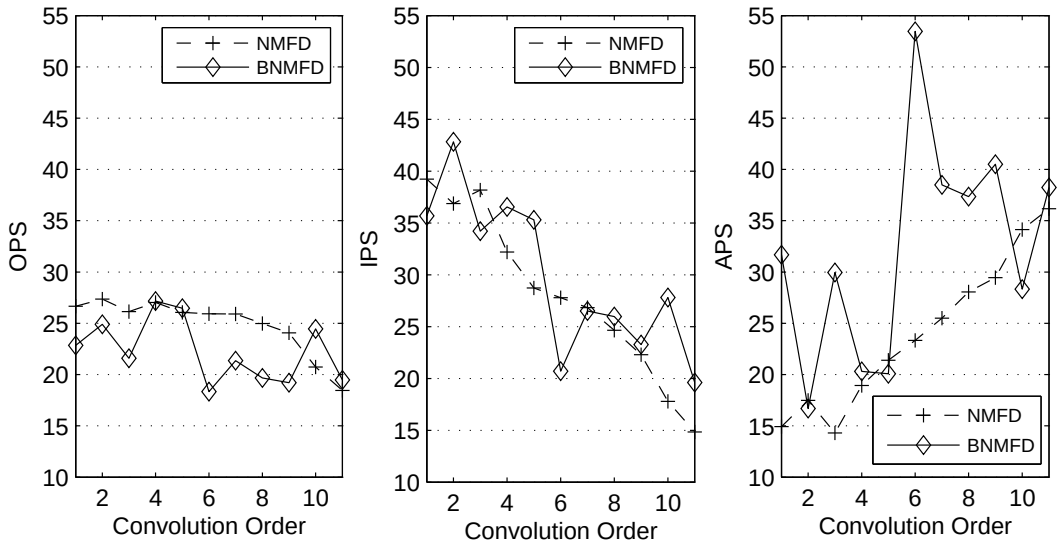


Figure 5.4: The mean OPS, IPS and APS values obtained by the NMFD and BNMFD methods for different values of convolution order T.

approximating distribution $q(\mathbf{Y})$ or full conditionals $p(\mathbf{Y}|\mathbf{X},\mathbf{S},\mathbf{A})$ as a product of multinomial distributions has leaded to a rich approximation distribution. It is known that, NMFD with the KL objective can be obtained with a full Bayesian inference. For several other distance metrics, full Bayesian inference is not as practical since $p(\mathbf{Y}|\mathbf{X},\mathbf{S},\mathbf{A})$ is not standard. It is shown in [72] that the standard NMF algorithm with multiplicative update rules is in fact an EM algorithm with data augmentation. By extending this approach, a hierarchical model has been developed with conjugate Gamma priors. A variational Bayes algorithm has been developed in this hierarchical model and also methods have been developed for estimating the marginal likelihood for model selection. The simulations suggest that the variational bound seems to be a reasonable approximation to the marginal likelihood and can guide model selection for NMFD. It has also found that both the fixed point equations as well as the bound in the variational approach can be compactly and efficiently implemented using matrix operations.

From a modeling perspective, the proposed hierarchical model provides some attractive properties. It is easy to incorporate prior knowledge about individual latent sources via hyperparameters and one can easily capture variability in the bases and gains that is potentially useful for developing robust techniques.

The main contribution proposed in this chapter is the development of a principled and practical way to estimate both the optimal sparsity criteria and model order, in terms of marginal likelihood. By maximizing the bound on marginal likelihood, all the hyperparameters can be estimated from the data, and the appropriate sparseness criteria can be found automatically. It is believed that the proposed method retains the attractive features of NMFD algorithm and enables to develop more powerful models.

6. CONCLUSIONS AND DISCUSSIONS

In this thesis, new models and algorithms have been developed for single channel audio source separation. Single-channel source separation is an under-determined problem, since the number of unknown sources to be estimated is more than the number of observed mixture signals. This causes the problem to be ill-posed since the available information is not enough to reconstruct the sources exactly. However, incorporating some assumptions or available prior information into the separation schemes makes the underdetermined problem feasible. In the context of the thesis, the focus has been made on separating sources from a single observed mixture either by making use of convex optimization under sparseness constraints or incorporating the available prior information to the separation model.

In the context of the thesis, we have mainly investigated four different aspects of single channel audio source separation. First, we have integrated the human perception model into the separation scheme which enhances the perceived quality of the separated sources. The proposed perceptually enhanced source separation scheme has been presented in Chapter 3. Secondly, we have focused on performing the separation based on an adaptive time-frequency representation which fuses the information from various time-frequency resolutions into a joint optimization scheme based on NTF. The developed model referred to as MR-NTF is described in Chapter 4. Third, we have investigated the NMFD scheme from a statistical perspective which enables to incorporate the prior information into the separation scheme. The Bayesian NMFD approach has been formulated in Chapter 5. Fourth, we have focused on the permutation problem encountered in the clustering of factorized bases into individual sources. In order to alleviate the permutation problem, two different solutions have been respectively designed by using an unsupervised approach based on shift-invariance assumption as formulated in Chapter 3 and with a supervised approach by learning the sources in advance as described in Chapter 4.

The proposed perceptual non-negative matrix factorization schemes, PW-NMF2D and PW-CNMF, employ the convolutive NMF formulation to represent the shift-invariance property of the notes which can be effectively used for separating the musical instruments in a blind manner. These methods also integrate human perception into the separation scheme, thus enhances the perceived quality of the separated sources. This has been achieved by enclosing the perceptually weighted cost terms into the optimization function of the factorization based on the KL and the IS divergences. The proposed weighted cost function has adaptively increased/decreased the contribution of each element-wise divergence to the optimization scheme by assigning a weighting score per each time-frequency component that takes into account the perceptual properties of human ear. This weighting score has enhanced the contribution of the perceptually important components while decreasing the contribution of the perceptually unimportant ones. The idea behind using the convolutive scheme is that it enables to extract the source bases under the assumption that all the notes played by an instrument can be represented by shifting the frequency signature of a specific note played by the same instrument. An instrument has a fixed temporal pattern (echo); and different notes played by the same instrument has the same time-log-frequency signature but varying in fundamental frequency (shift) where the frequency is defined in log-scale. In order to increase the separation quality, we have worked with high rank values. To eliminate the permutation problem, both PW-NMF2D and PW-CNMF use an unsupervised approach. PW-NMF2D extracts a single basis for each source which is then shifted in frequency to capture the other notes played by the same instrument under the assumption that all notes for an instrument can be represented by shifting the pitch of the time-frequency signature of a single note in log-frequency domain. The second method PW-CNMF resolves the clustering problem in an additional step where the bases obtained by a high-rank factorization are clustered into the same source if they share the same timbre at a different pitch. The evaluation results have been reported on a large dataset using the conventional SDR, SIR, SAR measures and the perceptual OPS, IPS, APS measures. Higher performance has been obtained when the cost function is in the KL-divergence form. The PW-NMF2D and the PW-CNMF methods have outperformed the existing CNMF, SNMF and NMF2D methods by an amount of 0.5-2 dB in terms of conventional measures and 1-6 points

in terms of perceptual measures. It has been shown that, the proposed perceptually enhanced PW-NMF2D and PW-CNMF methods constitute alternative blind single channel musical source separation methods which can be used in various applications where the perceptual quality is important.

The MR-NTF formulation has been introduced as an adaptive multiresolution method based on tensor factorization for supervised separation of speech or music sources from a single observation. Knowing that the transients and the stationary audio signals should be analyzed at different time-frequency resolutions to minimize the smearing, the MR-NTF represent the observed single channel mixture signal in various time-frequency resolutions at different layers of a tensor. This enables us to fuse the information from various time-frequency resolutions through a joint optimization scheme. The gains and the amplitude envelopes have been updated iteratively by applying the multiplicative rules. The source bases have been learned from the training data in advance hence the clustering problem has been resolved by fixing the learned bases during the factorization. The performance evaluations have shown an improvement of 1-4 dB in terms of SDR and SIR and 5-30 points in terms of their perceptual correspondence OPS and IPS compared to the fixed time-frequency resolution NMF results. The perceptual weighting scheme has also been integrated into the MR-NTF approach in order to enhance the perceptual quality of the separated sources.

From a statistical perspective to audio source separation, the BNMFD model has been introduced to develop a practical probabilistic approach for estimating the optimal sparsity criteria and the model order in terms of marginal likelihood. Hence, the multiplicative updates for the basis and the gain matrices under the KL divergence have been obtained from a probabilistic view, where the observations have been assumed to be superposition of J independent Poisson distributed latent sources. A hierarchical model has been designed with conjugate Gamma priors. A variational Bayes algorithm has been developed for this hierarchical model to estimate the marginal likelihood for model selection. It has been also shown that, the fixed point equations and variational bound can be implemented efficiently using matrix operations. All the hyperparameters can be estimated from the observed data and the sparseness criteria

can be obtained automatically by maximizing the bound on marginal likelihood. It has been shown that, the proposed BNMFD method retains the strong features of NMFD such as easy and efficient implementation and opens the way to develop more powerful models by incorporating prior information into the separating scheme.

The methods proposed within the thesis can be extended in several directions. In the proposed PW-CNMF method, the extracted bases of each source are clustered using shift-invariance property of the notes. In order to generalize the proposed scheme to any type of audio including speech, a general unsupervised approach can be developed.

In order to extend the proposed models, PW-NMF2D and PW-CNMF, to cover more sources, the rank of the factorization and the convolution parameters should be adjusted accordingly. Obviously, increasing the number of sources decreases the separation performance specifically for highly correlated sources. However, including a prior information about the sources should be helpful to improve the separation quality. Thus, a Bayesian approach to PW-NMF2D and PW-CNMF models can be developed by incorporating the available prior information into the separation scheme to cover the single channel mixtures of more sources.

The main difficulty in the developed models is that they require multiple forward and backward transformations from the time to the frequency domain. To alleviate this problem, the observed mixture spectrograms can be replaced by scalograms obtained based on wavelet based transformations. Furthermore, the wavelet based transformation can also be used in the representation of the mixture signals in critically spaced frequency bands. It can be shown that an optimal wavelet based filterbank under an auditory perception criterion improves the perceptual quality since it prevents the distortions arising from scale conversions.

Another future work topic can be the extension of the MR-NTF method into convolutive methods. The proposed approaches PW-NMF2D, PW-CNMF and MR-NTF can also be investigated under a Bayesian framework. From a modeling perspective, the proposed hierarchical model provides some attractive properties. It is easy to incorporate prior knowledge about individual latent sources via

hyperparameters and the variability in the bases and the gains can be easily captured that is potentially useful for developing robust techniques.

REFERENCES

[1] **Ellis, D.P.W.** (1996). Prediction-driven Computational Auditory Scene Analysis for Dense Sound Mixtures, *ESCA Workshop on the Auditory Basis of Speech Perception*, Keele, UK.

[2] **Brown, J.C. and Smaragdis, P.** (2004). Independent Component Analysis for Automatic Note Extraction from Musical Trills, *Journal of the Acoustical Society of America*, **115(5)**, 2295–2306.

[3] **Vincent, E. and Plumbley, M.D.** (2006). Single-channel Mixture Decomposition using Bayesian Harmonic Models, *6th International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, Charleston, SC, USA, pp.722–730.

[4] **Hyvärinen, A. and Hoyer, P.** (2000). Emergence of Phase and Shift Invariant Features by Decomposition of Natural Images into Independent Feature Subspaces, *Neural Computation*, **12(7)**, 1705–1720.

[5] **Casey, M.A. and Westner, A.** (2000). Separation of Mixed Audio Sources by Independent Subspace Analysis, *International Computer Music Conference*, Berlin, Germany.

[6] **Virtanen, T.** (2006). Sound Source Separation in Monaural Music Signals, *PhD Thesis*, Tampere University of Technology, Tampere, Finland.

[7] **Lee, D.D. and Seung, H.S.** (2001). Algorithms for Non-negative Matrix Factorization, *Advances in Neural Information Processing*, **13**, 556–562.

[8] **Virtanen, T.** (2007). Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria, *IEEE Transactions on Audio, Speech, and Language Processing*, **15(3)**, 1066–1074.

[9] **Smaragdis, P.** (2004). Non-negative Matrix Factor Deconvolution; Extraction of Multiple Sound Sources from Monophonic Inputs, *Lecture Notes in Computer Science 3195 Springer 2004*, Springer-Verlag Berlin Heidelberg, pp.494–499.

[10] **Jaiswal, R.**, **FitzGerald, D.**, **Coyle, E. and Rickard, S.** (2011). Shifted NMF using an Efficient Constant-Q Transform for Monaural Sound Source Separation, *22nd IET Irish Signals and Systems Conference*, Trinity College, Dublin, Ireland.

[11] **Schmidt, M.N. and Mørup, M.** (2006). Nonnegative Matrix Factor 2-D Deconvolution for Blind Single Channel Source Separation, *International Conference on Independent Component Analysis and Blind Source Separation (ICA)*, Charleston, SC, USA.

[12] **Jaiswal, R.**, **Fitzgerald, D.**, **Barry, D.**, **Coyle, E. and Rickard, S.** (2011). Clustering NMF Basis Functions using Shifted NMF for Monaural Sound Source Separation, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, pp.245–248.

[13] **Virtanen, T.O.** (2007). Monaural Sound Source Separation by Perceptually Weighted Non-Negative Matrix Factorization, Technical Report, Tampere University of Technology.

[14] **Guddeti, R. and Mulgrew, B.** (2005). Perceptually Motivated Blind Source Separation of Convolutive Mixtures, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, Philadelphia, PA, USA, pp.273–376.

[15] **Kırbız, S. and Günsel, B.** (2010). A Perpceptually Enhanced Blind Single-Channel Audio Source Separation by Non-negative Matrix Factorization, *European Signal Processing Conference (EUSIPCO)*, Aalborg, Denmark, pp.731–735.

[16] **Lukin, A. and Todd, J.** (2006). Adaptive Time-Frequency Resolution for Analysis and Processing of Audio, *120th Audio Engineering Society Convention*, Paris, France.

[17] **Chen, N.** (2001). Analysis of the Multiresolution Frequency Domain Algorithm for Blind Source Separation, *M. Sc Thesis*, Electrical Engineering New Mexico State University, New Mexico.

[18] **Craciun, A. and Spiertz, M.** (2010). Adaptive Time Frequency Resolution for Blind Source Separation, *International Student Conference on Electrical Engineering POSTER '10*, Prague, Czech Republic.

[19] **Kirbiz, S. and Smaragdis, P.** (2011). An Adaptive Time-frequency Resolution Approach for Non-negative Matrix Factorization based Single Channel Source Separation, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, pp.253–256.

[20] **Cichocki, A.**, **Zdunek, R.**, **Phan, A. and Amari, S.** (2009). Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation, Wiley, http://books.google.com.tr/books?id=8S4MRXMRdcMC.

[21] **Kırbız, S. and Günsel, B.** (2012). Perceptually Weighted Non-negative Matrix Factorization for Blind Single-channel Music Source Separation, *International Conference on Pattern Recognition (ICPR)*, Tsukuba Science City, Japan.

[22] **Smaragdis, P.** (2007). Convolutive Speech Bases and Their Application to Supervised Speech Separation, *IEEE Transactions on Audio, Speech, and Language Processing*, **15(1)**, 1–12.

[23] **Duan, Z.**, **Zhang, Y.**, **Zhang, C. and Shi, Z.** (2008). Unsupervised Single-channel Music Source Separation by Average Harmonic Structure Modeling, *IEEE Transactions on Audio, Speech, and Language Processing*, **16(4)**, 766–778.

[24] **Vincent, E.**, **Fevotte, C. and Gribonval, R.** (2006). Performance Measurement in Blind Audio Source Separation, *IEEE Transactions on Audio, Speech, and Language Processing*, **14(4)**, 1462–1469.

[25] **Emiya, V.**, **Vincent, E.**, **Harlander, N. and Hohmann, V.** (2011). Subjective and Objective Quality Assessment of Audio Source Separation, *IEEE Transactions on Audio, Speech, and Language Processing*, **19(7)**, 2046–2057.

[26] **FitzGerald, D.**, **Cranitch, M. and Coyle, E.** (2008). On the use of the Beta Divergence for Musical Source Separation, *Irish Signals and Systems Conference (ISSC)*, Galway, Ireland.

[27] **ITU-R BS.1387** (1998). Method for Objective Measurements of Perceived Audio Quality.

[28] **Kirbiz, S. and Gunsel, B.** (2013). Perceptually Enhanced Blind Single-channel Music Source Separation by Non-negative Matrix Factorization, *Digital Signal Processing*, **23(2)**, 646–658.

[29] **Mørup, M. and Schmidt, M.N.** (2006). Sparse Non-negative Matrix Factor 2-D Deconvolution, Technical Report, Technical University of Denmark.

[30] **FitzGerald, D.**, **Cranitch, M. and Coyle, E.** (2005). Shifted Non-negative Matrix Factorisation for Sound Source Separation, *IEEE Workshop of Statistical Signal Processing*, Bordeaux, France.

[31] **Kırbız, S.**, **Cemgil, A.T. and Günsel, B.** (2010). Bayesian Inference for Nonnegative Matrix Factor Deconvolution Models, *International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, pp.2812–2815.

[32] **Ozerov, A.**, **Liutkus, A.**, **Badeau, R. and Richard, G.** (2011). Informed Source Separation: Source Coding Meets Source Separation, *2011 IEEEWorkshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA.

[33] **Févotte, C.**, **Bertin, N. and Durrieu, J.L.** (2009). Nonnegative Matrix Factorization with the Itakura-Saito Divergence. With Application to Music Analysis, *Neural Computation*, **21(3)**, 793–830.

[34] **Févotte, C. and Ozerov, A.** (2011). Notes on Nonnegative Tensor Factorization of the Spectrogram for Audio Source Separation: Statistical Insights and Towards Self-clustering of the Spatial Cues, *7th international Conference on Exploring Music Contents*, CMMR'10, Springer-Verlag, Berlin, Heidelberg, pp.102–115, http://dl.acm.org/citation.cfm?id=2040789.2040799.

[35] **Chen, Z. and Cichocki, A.** (2005). Nonnegative Matrix Factorization with Temporal Smoothness and/or Spatial Decorrelation Constraints, Technical Report, Laboratory for Advanced Brain Signal Processing, RIKEN, Tokyo, Japan.

[36] **Schmidt, M.N. and Laurberg, H.** (2008). Non-negative Matrix Factorization with Gaussian Process Priors, *Computational Intelligence and Neuroscience*, **2008(Article ID 361705)**.

[37] **Schmidt, M.N. and Mørup, M.** (2005). Sparse Non-negative Matrix Factor 2-D Deconvolution for Automatic Transcription of Polyphonic Music, Technical Report, Technical University of Denmark.

[38] **Url-1** <*http://www.ee.columbia.edu/~dpwe/resources/matlab/sgram/*>, date retrieved 11.02.2012.

[39] **Carroll, J.D. De Soete, G. and Pruzansky, S.** (1989). Fitting of the latent class model via iteratively reweighted least squares CANDECOMP with nonnegativity constraints, R. Coppi and S. Bolasco, editors, *Multiway data analysis.*, Elsevier, Amsterdam, pp.463–472.

[40] **Shashua, A. and Hazan, T.** (2005). Non-negative Tensor Factorization with Applications to Statistics and Computer Vision, *International Conference on Machine Learning (ICML)*, ACM, Bonn, Germany, pp.792–799.

[41] **FitzGerald, D. and Cranitch, M.** (2007). Resynthesis Methods for Sound Source Separation using Shifted Non-negative Factorisation Models, *Irish Signals and Systems Conference (ISSC)*, Dublin, Ireland.

[42] **Benaroya, L.**, **Bimbot, F. and Gribonval, R.** (2006). Audio Source Separation with a Single Sensor, *IEEE Transactions on Audio, Speech, and Language Processing*, **14(1)**, 191–199.

[43] **Url-2** <*http://bass-db.gforge.inria.fr/bss_eval/*>, date retrieved 11.02.2012.

[44] **Url-3** <*http://bass-db.gforge.inria.fr/peass/*>, date retrieved 11.02.2012.

[45] **Huber, R. and Kollmeier, B.** (2006). PEMO-Q A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception, *IEEE Transactions on Audio, Speech, and Language Processing*, **14(6)**, 1902–1911.

[46] **Klapuri, A. and Schörkhuber, C.** (2010). Constant-Q Transform Toolbox for Music Processing, *7th Sound and Music Computing Conference*, Barcelona, Spain.

[47] **Zwicker, E. and Fastl, H.** (1999). Psychoacoustics: Facts and Models, Springer, Berlin, second edition.

[48] **O'Grady, P.D.** (2007). Sparse Separation of Under-determined Speech Mixtures, *PhD Thesis*, Department of Computer Science National University of Ireland, Maynooth, Ireland.

[49] **Nikunen, J. and Virtanen, T.** (2010). Noise-to-Mask Ratio Minimization by Weighted Non-negative Matrix Factorization, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, USA, pp.25–28.

[50] **Pampalk, E.**, **Dixon, S. and Widmer, G.** (2003). Exploring Music Collections by Browsing Different Views, *4th International Conference on Music Information Retrieval (ISMIR'03)*, Baltimore, MD, USA, pp.201–208.

[51] **Kabal, P.** (2002). An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality, Technical Report, Department of Electrical & Computer Engineering McGill University.

[52] **Thiede, T.** (1999). Perceptual Audio Quality Assessment Using a Non-Linear Filter Bank, *PhD Thesis*, Fachbereich Electrotechnik, Technical University of Berlin, Berlin, Germany.

[53] **Url-4** <*http://eleceng.dit.ie/derryfitzgerald/index.php?uid=489\&menu_id=52*>, date retrieved 11.02.2012.

[54] **SISEC**, Signal Separation Evaluation Campaign (SISEC 2008), <*http://sisec.wiki.irisa.fr*>, date retrieved 11.02.2012.

[55] **Url-5** <*http://www.mspr.itu.edu.tr/demopw.htm*>, date retrieved 11.02.2012.

[56] **Jones, D.L. and Baraniuk, R.G.** (1994). A Simple Scheme for Adapting Time-frequency Representations, *IEEE Transactions on Signal Processing*, **42(12)**, 3530–3535.

[57] **Basu, P.**, **Wolfe, P.J.**, **Rudoy, D.**, **Quatieri, T.F. and Dunn, B.** (2008). Adaptive Short-time Analysis-synthesis for Speech Enhancement, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, Nevada, USA, pp.3005–3008.

[58] **Rudoy, D.**, **Basu, P. and Wolfe, P.J.** (2010). Superposition Frames for Adaptive Time-frequency Analysis and Fast Reconstruction., *IEEE Transactions on Signal Processing*, **58(5)**, 2581–2596.

[59] **Kisilev, P.**, **Zibulevsky, M. and Zeevi, Y.Y.** (2004). Multiscale Framework For Blind Source Separation, *The Journal of Machine Learning Research*, **4(7-8)**, 1339–1364.

[60] **FitzGerald, D.**, **Cranitch, M. and Coyle, E.** (2005). Non-Negative Tensor Factorisation for Sound Source Separation, *Irish Signals and Systems Conference (ISSC)*, Dublin, Ireland.

[61] **FitzGerald, D.**, **Cranitch., M. and Coyle, E.** (2006). Sound Source Separation using Shifted Non-negative Tensor Factorisation, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, Toulouse, France, pp.653–656.

[62] **Kolda, T.G. and Bader, B.W.** (2009). Tensor Decompositions and Applications, *SIAM Review*, **51(3)**, 455–500.

[63] **Nion, D.**, **Mokios, K.N.**, **Sidiropoulos, N.D. and Potamianos, A.** (2010). Batch and Adaptive PARAFAC-based Blind Separation of Convolutive Speech Mixtures, *IEEE Transactions on Acoustics, Speech and Signal Processing*, **18(6)**, 1193–1207.

[64] **Wu, Q.**, **Zhang, L. and Shi, G.** (2011). Robust Multifactor Speech Feature Extraction Based on Gabor Analysis, *IEEE Transactions on Acoustics, Speech and Signal Processing*, **19(4)**, 927–936.

[65] **Tjoa, S.**, **Stamm, M.**, **Lin, W. and Liu, K.** (2010). Harmonic Variable-size Dictionary Learning for Music Source Separation, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, TX, USA, pp.413–416.

[66] **Painter, T. and Spanias, A.** (1997). A Review of Algorithms for Perceptual Coding of Digital Audio Signals, *13th International Conference on Digital Signal Processing*, volume 1, Santorini, Greece, pp.179–208.

[67] **Took, C.C. and Sanei, S.** (2007). Exploiting Sparsity, Sparseness and Super-Gaussianity in Underdetermined Blind Identification of Temporo-mandibular Joint Sounds, *Journal of Computers*, **2(6)**, 65–71.

[68] **Hurley, S.** Call for Remixes: Shannon Hurley, *<http://www.ccmixter.org/shannon-hurley>*, date retrieved 11.02.2012.

[69] **Knuth, K.H.** (1998). Bayesian source separation and localization, *SPIE'98: Bayesian Inference for Inverse Problems*, San Diego, pp.147–158.

[70] **Bertin, N.**, **Badeau, R. and Vincent, E.** (2010). Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription, *IEEE Transactions on Audio, Speech and Language Processing*, **18(3)**, 538–549.

[71] **Shashanka, M.**, **Raj, B. and Smaragdis, P.** (2008). Probabilistic Latent Variable Models as Non-Negative Factorizations, *Computational Intelligence and Neuroscience Journal*, **2008**.

[72] **Cemgil, A.T.** (2009). Bayesian Inference in Non-negative Matrix Factorisation Models, *Computational Intelligence and Neuroscience*, **(Article ID 785152)**.

[73] **Bishop, C.M.** (2007). Pattern Recognition and Machine Learning (Information Science and Statistics), Springer, 1 edition.

[74] **Yoo, J. and Choi, S.** (2008). Orthogonal Nonnegative Matrix Factorization: Multiplicative Updates on Stiefel Manifolds, *9th International Conference on Intelligent Data Engineering and Automated Learning*, IDEAL '08, Springer-Verlag, Berlin, Heidelberg, pp.140–147.

APPENDICES

APPENDIX A.1 : Obtaining Update Equations for PW-NMF2D under KL Divergence
APPENDIX A.2 : Obtaining Update Equations for PW-NMF2D under IS Divergence

APPENDIX A.1 :
Obtaining Update Equations for PW-NMF2D under KL Divergence

Suppose that the gradient of the objective function has a decomposition of the form

$$\frac{\partial C}{\partial A^{\phi}_{ri}} = [C^{\phi}]^{+}_{ri} - [C^{\phi}]^{-}_{ri} \tag{A.1}$$

$$\frac{\partial C}{\partial S^{\tau}_{kr}} = [C^{\tau}]^{+}_{kr} - [C^{\tau}]^{-}_{kr} \tag{A.2}$$

where $[C^{\phi}]^{+}_{ri} > 0, [C^{\phi}]^{-}_{ri} > 0, [C^{\tau}]^{+}_{kr} > 0$ and $[C^{\tau}]^{-}_{kr} > 0$ [74]. Then, multiplicative updates for A^{ϕ}_{ri} and S^{τ}_{kr} has the form:

$$A^{\phi}_{ri} \leftarrow A^{\phi}_{ri} \frac{[C^{\phi}]^{-}_{ri}}{[C^{\phi}]^{+}_{ri}} \tag{A.3}$$

$$S^{\tau}_{kr} \leftarrow S^{\tau}_{kr} \frac{[C^{\tau}]^{-}_{kr}}{[C^{\tau}]^{+}_{kr}} \tag{A.4}$$

If the objective function in (**3.4**) is chosen as KL divergence, differentiating the objective function with respect to a given element in $\mathbf{A}^{\phi}$ gives:

$$\begin{aligned} \frac{\partial C_{KL}}{\partial A^{\phi}_{rn}} &= \frac{\partial}{\partial A^{\phi}_{rn}} \sum_{k}\sum_{i} W_{k,i}\left(X_{k,i}\log\frac{X_{ki}}{\Lambda_{ki}} - X_{ki} + \Lambda_{ki}\right) \\ &= \sum_{k}\sum_{i} W_{ki}\left(1 - \frac{X_{ki}}{\Lambda_{ki}}\right)\frac{\partial \Lambda_{ki}}{\partial A^{\phi}_{rn}} \end{aligned} \tag{A.5}$$

The derivative of a given element Λ_{ki} with respect to a given element $\mathbf{A}^{\phi}$ in (**A.5**) can be derived as:

$$\begin{aligned} \frac{\partial \Lambda_{ki}}{\partial A^{\phi}_{rn}} &= \frac{\partial}{\partial A^{\phi}_{rn}}\left(\sum_{\tau}\sum_{\phi}\sum_{d} S^{\tau}_{k-\phi,d} A^{\phi}_{d,i-\tau}\right) \\ &= \frac{\partial}{\partial A^{\phi}_{rn}}\left(\sum_{\tau} S^{\tau}_{k-\phi,r} A^{\phi}_{r,i-\tau}\right) \\ &= \begin{cases} S^{\tau}_{k-\phi,r}, & \text{if } \tau = i-n \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{A.6}$$

If (**A.6**) is substituted into (**A.5**) and the index n is shanged to i,

$$\frac{\partial C_{KL}}{\partial A^{\phi}_{ri}} = [C^{\phi}]^{+}_{ri} - [C^{\phi}]^{-}_{ri} = \sum_{k}\sum_{\tau} W_{k,i+\tau}\left(1 - \frac{X_{k,i+\tau}}{\Lambda_{k,i+\tau}}\right) S^{\tau}_{k-\phi,r}. \tag{A.7}$$

Similarly, derivating the KL divergence with respect to a given element in $\mathbf{S}^{\tau}$ gives:

$$\frac{\partial C_{KL}}{\partial S^{\tau}_{m,r}} = \frac{\partial}{\partial S^{\tau}_{m,r}} \sum_{k}\sum_{i} W_{ki}\left(X_{ki}\log\frac{X_{ki}}{\Lambda_{ki}} - X_{ki} + \Lambda_{ki}\right)$$

$$= \sum_k \sum_i W_{ki} \left(1 - \frac{X_{ki}}{\Lambda_{ki}}\right) \frac{\partial \Lambda_{ki}}{\partial S^{\tau}_{m,r}} \tag{A.8}$$

The derivative of a given element Λ_{ki} with respect to a given element $\mathbf{S}^{\tau}$ in (**A.8**) can be derived as:

$$\begin{aligned} \frac{\partial \Lambda_{ki}}{\partial S^{\tau}_{m,r}} &= \frac{\partial}{\partial S^{\tau}_{m,r}} \left(\sum_{\tau} \sum_{\phi} \sum_{d} S^{\tau}_{k-\phi,d} A^{\phi}_{d,i-\tau} \right) \\ &= \frac{\partial}{\partial S^{\tau}_{m,r}} \left(\sum_{\phi} S^{\tau}_{k-\phi,r} A^{\phi}_{r,i-\tau} \right) \\ &= \begin{cases} A^{\phi}_{r,i-\tau}, & \text{if } \phi = k-m \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{A.9}$$

If (**A.9**) is substituted into (**A.8**) and the index m is changed to k,

$$\frac{\partial C_{KL}}{\partial S^{\tau}_{kr}} = [C^{\tau}]^{+}_{kr} - [C^{\tau}]^{-}_{kr} = \sum_{\phi} \sum_{i} W_{k+\phi,i} \left(1 - \frac{X_{k+\phi,i}}{\Lambda_{k+\phi,i}}\right) A^{\phi}_{r,i-\tau}. \tag{A.10}$$

The multiplicative updates become:

$$A^{\phi}_{ri} \leftarrow A^{\phi}_{ri} \frac{\sum_k \sum_{\tau} W_{k,i+\tau} \frac{X_{k,i+\tau}}{\Lambda_{k,i+\tau}} S^{\tau}_{k-\phi,r}}{\sum_k \sum_{\tau} S^{\tau}_{k-\phi,r} W_{k,i+\tau}} \tag{A.11}$$

$$S^{\tau}_{kr} \leftarrow S^{\tau}_{kr} \frac{\sum_{\phi} \sum_i W_{k+\phi,i} \frac{X_{k+\phi,i}}{\Lambda_{k+\phi,i}} A^{\phi}_{r,i-\tau}}{\sum_{\phi} \sum_i A^{\phi}_{r,i-\tau} W_{k+\phi,i}}. \tag{A.12}$$

APPENDIX A.2 :
Obtaining Update Equations for PW-NMF2D under IS Divergence

Differentiating the objective function in (**3.4**) with respect to a given element in $\mathbf{A}^{\phi}$ where **D** is the IS divergence gives:

$$\begin{aligned}\frac{\partial C_{IS}}{\partial A^{\phi}_{rn}} &= \frac{\partial}{\partial A^{\phi}_{rn}} \sum_k \sum_i W_{ki}\left(\frac{X_{ki}}{\Lambda_{ki}} - \log\frac{X_{ki}}{\Lambda_{ki}} - 1\right) \\ &= \sum_k \sum_i \frac{W_{ki}}{\Lambda_{ki}}\left(1 - \frac{X_{ki}}{\Lambda_{ki}}\right)\frac{\partial \Lambda_{ki}}{\partial A^{\phi}_{rn}} \end{aligned} \quad \textbf{(A.13)}$$

If (**A.6**) is substituted into (**A.13**) and the index n is changed to i,

$$\frac{\partial C_{IS}}{\partial A^{\phi}_{ri}} = [C^{\phi}]^{+}_{ri} - [C^{\phi}]^{-}_{ri} = \sum_k \sum_{\tau} \frac{W_{k,i+\tau}}{\Lambda_{k,i+\tau}}\left(1 - \frac{X_{k,i+\tau}}{\Lambda_{k,i+\tau}}\right) S^{\tau}_{k-\phi,r}. \quad \textbf{(A.14)}$$

Similarly, derivating the IS divergence with respect to a given element in $\mathbf{S}^{\tau}$ gives:

$$\begin{aligned}\frac{\partial C_{IS}}{\partial S^{\tau}_{m,r}} &= \frac{\partial}{\partial S^{\tau}_{m,r}} \sum_k \sum_i W_{ki}\left(\frac{X_{ki}}{\Lambda_{ki}} - \log\frac{X_{ki}}{\Lambda_{ki}} - 1\right) \\ &= \sum_k \sum_i \frac{W_{ki}}{\Lambda_{ki}}\left(1 - \frac{X_{ki}}{\Lambda_{ki}}\right)\frac{\partial \Lambda_{ki}}{\partial S^{\tau}_{m,r}} \end{aligned} \quad \textbf{(A.15)}$$

If (**A.9**) is substituted into (**A.15**) and the index m is changed to k,

$$\frac{\partial C_{IS}}{\partial S^{\tau}_{kr}} = [C^{\tau}]^{+}_{kr} - [C^{\tau}]^{-}_{kr} = \sum_{\phi} \sum_i \frac{W_{k+\phi,i}}{\Lambda_{k+\phi,i}}\left(1 - \frac{X_{k+\phi,i}}{\Lambda_{k+\phi,i}}\right) A^{\phi}_{r,i-\tau}. \quad \textbf{(A.16)}$$

The multiplicative updates become:

$$A^{\phi}_{ri} \leftarrow A^{\phi}_{ri} \frac{\sum_k \sum_{\tau} W_{k,i+\tau} \frac{X_{k,i+\tau}}{\Lambda^2_{k,i+\tau}} S^{\tau}_{k-\phi,r}}{\sum_k \sum_{\tau} S^{\tau}_{k-\phi,r} \frac{W_{k,i+\tau}}{\Lambda_{k,i+\tau}}} \quad \textbf{(A.17)}$$

$$S^{\tau}_{kr} \leftarrow S^{\tau}_{kr} \frac{\sum_{\phi} \sum_i W_{k+\phi,i} \frac{X_{k+\phi,i}}{\Lambda^2_{k+\phi,i}} A^{\phi}_{r,i-\tau}}{\sum_{\phi} \sum_i A^{\phi}_{r,i-\tau} \frac{W_{k+\phi,i}}{\Lambda_{k+\phi,i}}}. \quad \textbf{(A.18)}$$

CURRICULUM VITAE

Name Surname: Serap Kırbız

Place and Date of Birth: Elazığ, July 8th, 1979

Adress: İstanbul Teknik Üniversitesi, Elektrik-Elektronik Fakültesi, Elektronik ve Haberleşme Mühendisliği Bölümü, 34469, Maslak / İstanbul, Turkey

E-Mail: kirbiz@itu.edu.tr

B.Sc.: Mathematical Engineering, Istanbul Technical University

M.Sc.: Computer Science, Istanbul Technical University

Professional Experience and Rewards: Research Assistant

List of Publications and Patents:

▪ **Kirbiz, S.** and Gunsel B., 2013: Perceptually Enhanced Blind Single-Channel Music Source Separation by Non-Negative Matrix Factorization. *Digital Signal Processing*, vol: 23, issue: 2, pp.646-658, http://dx.doi.org/10.1016/j.dsp.2012.10.001.

▪ **Kirbiz, S.** and Gunsel B., 2012: Perceptually Weighted Non-negative Matrix Factorization for Blind Single-Channel Music Source Separation. *Proc. of International Conference on Pattern Recognition (ICPR), Nov 11-14, 2012, Tsukuba Science City, Japan.*

▪ **Kirbiz, S.** and Smaragdis, P., 2011: An Adaptive Time-Frequency Resolution Approach for Non-Negative Matrix Factorization Based Single Channel Sound Source Separation. *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 22-27, 2011, Prague, Czech Republic.

▪ **Kirbiz, S.** and Smaragdis, P., 2011: Negatif Olmayan Matris Ayrı ştırmaya Dayalı Tek-Kanaldan Ses Ayrıştırma için Uyarlamalı Zaman-Sıklık Çözünürlüğü. *IEEE 19. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SİU)* 20-22 Nisan 2011, Antalya, Türkiye.

▪ **Kirbiz, S.** and Gunsel, B., 2010: A Perceptually Enhanced Blind Single-Channel Audio Source Separation by Non-negative Matrix Factorization. *Proc. of 18th European Signal Processing Conference (EUSIPCO)*, August 23-27, 2010, Aalborg, Denmark.

▪ **Kirbiz, S.**, Cemgil, A. T. and Gunsel B., 2010: Bayesian Inference for Nonnegative Matrix Factor Deconvolution Models. *Proc. of International Conference on Pattern Recognition (ICPR)*, August 23-26, 2010, Istanbul, Turkey.

▪ **Kirbiz, S.** and Gunsel B., 2009: Negatif Olmayan Matris Ayrıştırma ile Tek-Kanaldan Algısal Ses Ayrıştırma. *IEEE 17. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SİU)*, 9-11 Nisan 2009, Antalya, Türkiye.

▪ **Kirbiz, S.**, Ulker, Y. and Gunsel, B., 2009: A Pattern Recognition Framework to Blind Audio Watermark Decoding. *AEÜ Archiv für Elektronik und Übertragungstechnik (International Journal of Electronics and Communications)*, vol: 63, issue:2, pp. 92-102, February 4, 2009.

▪ Ulker, Y., Gunsel, B. and **Kirbiz, S.**, 2008: A Multiple Model Structure for Tracking by Variable Rate Particle Filters. *Proc. of International Conference on Pattern Recognition(ICPR)*, December 8-11, 2008, Florida, USA.

▪ Lemma, A., Katzenbeisser, S., Celik, M. and **Kirbiz, S.**, 2008: Forensic Watermarking and Bit-Rate Conversion of Partially Encrypted AAC Bitstreams. *Proc. of the SPIE, Security, Forensics, Steganography and Watermarking of Multimedia Contents X*.

▪ **Kirbiz, S.**, Celik, M., Lemma, A. and Katzenbeisser, S., 2007: Decode-time Forensic Watermarking of AAC Bit-Streams. *IEEE Transactions on Information Forensics and Security*, vol: 2, No: 4, pp. 683-696, December 2007.

▪ **Kirbiz, S.**, Celik, M., Lemma, A. and Katzenbeisser, S., 2007: Forensic Watermarking during AAC Playback. *Proc. of the IEEE International Conference on Mulimedia & Expo (ICME)*, July 2-5, 2007, Beijning, China.

▪ **Kirbiz, S.**, Celik, M., Lemma, A. and Katzenbeisser, S., 2007: Adli Takip için AAC Bit-katarlarının Damgalanması. *IEEE 15. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SİU)* 11-13 Haziran 2007, Eskişehir, Türkiye.

▪ Gunsel, B., Ulker, Y. and **Kirbiz, S.**, 2006: A Statistical Framework for Audio Watermark Detection and Decoding. *Lecture Notes in Computer Science*, vol: 4105, pp: 241-248, Springer-Verlag, 2006.

▪ Gunsel, B. and **Kirbiz, S.**, 2006: Perceptual Audio Watermarking by Learning in Wavelet Domain. *Proc. of 18th International Conference on Pattern Recognition (ICPR)*, Aug 20-24, 2006, Hong Kong.

▪ **Kirbiz, S.** and Gunsel, B., 2006: Robust Audio Watermark Decoding by Supervised Learning. *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 14-19, 2006, Toulouse, France.

▪ **Kirbiz, S.** and Gunsel, B., 2006: Dalgacık Uzayında Öğrenmeye Dayalı Sayısal Ses Damgalama. *IEEE 14. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SİU)*, 17-19 Nisan 2006, Antalya, Türkiye.

▪ **Kirbiz, S.**, Yaslan, Y. and Gunsel, B., 2005: Robust Audio Watermark Decoding by Nonlinear Classification. *Proc. of 13th European Signal Processing Conference (EUSIPCO)*, Sep 4-8, 2005, Antalya, Turkiye.

PUBLICATIONS/PRESENTATIONS ON THE THESIS

▪ **Kirbiz, S.** and Gunsel B., 2013: Perceptually Enhanced Blind Single-Channel Music Source Separation by Non-Negative Matrix Factorization. *Digital Signal Processing*, vol: 23, issue: 2, pp.646-658, http://dx.doi.org/10.1016/j.dsp.2012.10.001.

▪ **Kirbiz, S.** and Gunsel B., 2012: Perceptually Weighted Non-negative Matrix Factorization for Blind Single-Channel Music Source Separation. *Proc. of International Conference on Pattern Recognition (ICPR), Nov 11-14, 2012, Tsukuba Science City, Japan.*

▪ **Kirbiz, S.** and Smaragdis, P., 2011: An Adaptive Time-Frequency Resolution Approach for Non-Negative Matrix Factorization Based Single Channel Sound Source Separation. *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 22-27, 2011, Prague, Czech Republic.

▪ **Kirbiz, S.** and Smaragdis, P., 2011: Negatif Olmayan Matris Ayrıştırmaya Dayalı Tek-Kanaldan Ses Ayrıştırma için Uyarlamalı Zaman-Sıklık Çözünürlüğü. *IEEE 19. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SİU)* 20-22 Nisan 2011, Antalya, Türkiye.

▪ **Kirbiz, S.** and Gunsel, B., 2010: A Perceptually Enhanced Blind Single-Channel Audio Source Separation by Non-negative Matrix Factorization. *Proc. of 18th European Signal Processing Conference (EUSIPCO)*, August 23-27, 2010, Aalborg, Denmark.

▪ **Kirbiz, S.**, Cemgil, A. T. and Gunsel B., 2010: Bayesian Inference for Nonnegative Matrix Factor Deconvolution Models. *Proc. of International Conference on Pattern Recognition (ICPR)*, August 23-26, 2010, Istanbul, Turkey.

▪ **Kirbiz, S.** and Gunsel B., 2009: Negatif Olmayan Matris Ayrıştırma ile Tek-Kanaldan Algısal Ses Ayrıştırma. *IEEE 17. Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SİU)*, 9-11 Nisan 2009, Antalya, Türkiye.